Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2014, Vol. 50, No. 3, 1028–1068
DOI: 10.1214/13-AIHP551

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES

www.imstat.org/aihp

Estimation of the transition density of a Markov chain

Mathieu Sart

[a] *Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice cedex 02, France. E-mail: msart@unice.fr*

Received 19 October 2012; revised 24 January 2013; accepted 18 February 2013

Abstract. We present two data-driven procedures to estimate the transition density of an homogeneous Markov chain. The first yields a piecewise constant estimator on a suitable random partition. By using an Hellinger-type loss, we establish non-asymptotic risk bounds for our estimator when the square root of the transition density belongs to possibly inhomogeneous Besov spaces with possibly small regularity index. Some simulations are also provided. The second procedure is of theoretical interest and leads to a general model selection theorem from which we derive rates of convergence over a very wide range of possibly inhomogeneous and anisotropic Besov spaces. We also investigate the rates that can be achieved under structural assumptions on the transition density.

Résumé. Nous présentons deux procédures pour estimer la densité de transition d'une chaîne de Markov homogène. Dans la première procédure, nous construisons un estimateur constant par morceaux sur une partition aléatoire bien choisie. Nous établissons des bornes de risque non-asymptotiques pour une perte de type Hellinger lorsque la racine carrée de la densité de transition appartient à un espace de Besov inhomogène dont l'indice de régularité peut être petit. Nous illustrons ces résultats par des simulations numériques. La deuxième procédure est d'intérêt théorique. Elle permet d'obtenir un théorème de sélection de modèle à partir duquel nous déduisons des vitesses de convergence sur des espaces de Besov inhomogènes anisotropes. Nous étudions finalement les vitesses qui peuvent être atteintes sous des hypothèses structurelles sur la densité de transition.

MSC: 62M05; 62G05

Keywords: Adaptive estimation; Markov chain; Model selection; Robust tests; Transition density

1. Introduction

Consider a time-homogeneous Markov chain $(X_i)_{i\in\mathbb{N}}$ defined on an abstract probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with values in the measured space $(\mathbb{X}, \mathcal{F}, \mu)$. We assume that for each $x \in \mathbb{X}$, the conditional law $\mathcal{L}(X_{i+1} \mid X_i = x)$ admits a density $s(x, \cdot)$ with respect to μ. Our aim is to estimate the transition density $(x, y) \mapsto s(x, y)$ on a subset $A = A_1 \times A_2$ of $\mathbb{X}^2$ from the observations $X_0, \ldots, X_n$.

Many papers are devoted to this statistical setting. A popular method to build an estimator of s is to divide an estimator of the joint density of (X_i, X_{i+1}) by an estimator of the density of X_i. The resulting estimator is called a quotient estimator. Roussas [33], Athreya and Atuncar [4] considered Kernel estimators for the densities of X_i and (X_i, X_{i+1}). They proved consistence and asymptotic normality of the quotient estimator. Other properties of this estimator were established: Roussas [34], Dorea [22] showed strong consistency, Basu and Sahoo [8] proved a Berry–Essen type theorem and Doukhan and Ghindès [24] bounded from above the integrated quadratic risk under Sobolev constraints. Clémencon [18] investigated the minimax rates when $A = [0, 1]^2$, $\mathbb{X}^2 = \mathbb{R}^2$. Given two smoothness classes $\mathscr{F}_1$ and $\mathscr{F}_2$ of real valued functions on $[0, 1]^2$ and $[0, 1]$ respectively (balls of Besov spaces), he established the lower bounds over the class

$$\mathscr{F} = \left\{ \varphi, \forall x, y \in [0, 1], \varphi(x, y) = \frac{\varphi_1(x, y)}{\varphi_2(x)}, (\varphi_1, \varphi_2) \in \mathscr{F}_1 \times \mathscr{F}_2 \right\}.$$

He developed a method based on wavelet thresholding to estimate the densities of X_i and (X_i, X_{i+1}) and showed that the quotient estimator of s is quasi-optimal in the sense that the minimax rates are achieved up to possible logarithmic factors. Lacour [28,29] used model selection via penalization to construct estimates of the densities. The resulting quotient estimator reaches the minimax rates over $\mathscr{F}$ when $\mathscr{F}_1$ and $\mathscr{F}_2$ are balls of homogeneous (but possibly anisotropic) Besov spaces on $[0,1]^2$ and $[0,1]$ respectively.

The previous rates of convergence depend on the smoothness properties of the densities of X_i and (X_i, X_{i+1}). In the favourable case where $X_0, \dots, X_n$ are drawn from a stationary Markov chain (with stationary density f), the rates depend on the smoothness properties of f or more precisely on the restriction of f to A_1. This function may however be less regular than the target function s. We refer for instance to Section 5.4.1 of Clémencon [18] for an example of a Doeblin recurrent Markov chain where the stationary density f is discontinuous on $[0,1]$ although s is constant on $[0,1]^2$. Therefore, these estimators may converge slowly even if s is smooth, which is problematic.

This issue was overcome in several papers. Clémencon [18] proposed a second procedure, based on wavelets and an analogy with the regression setting. He computed the lower bounds of minimax rates when the restriction of s on $[0,1]^2$ belongs to balls of some (possibly inhomogenous) Besov spaces and proved that its estimator achieves these rates up to a possible logarithmic factor. Lacour [27] established lower bound over balls of some (homogenous but possibly anisotropic) Besov spaces. By minimizing a penalized contrast inspired from the least-squares, she obtained a model selection theorem from which she deduced that her estimator reaches the minimax rates when $A = [0,1]^2$, $\mathbb{X}^2 = \mathbb{R}^2$. With a similar procedure, Akakpo and Lacour [3] obtained the usual rates of convergence over balls of possibly anisotropic and inhomogeneous Besov spaces (when $\mathbb{X}^2 = A = [0,1]^{2d}$). Very recently, Birgé [15] proposed a procedure based on robust testing to establish a general oracle inequality. The expected rates of convergence can be deduced from this inequality when $\sqrt{s}$ belongs to balls of possibly anisotropic and inhomogeneous Besov spaces.

These authors have used different losses in order to evaluate the performance of their estimators. In each of these papers, the risk of an estimator $\hat{s}$ is of the form $\mathbb{E}[\delta^2(s\mathbb{1}_A, \hat{s})]$ where $\mathbb{1}_A$ denotes the indicator function of the subset A and δ a suitable distance. Lacour [27], Akakpo and Lacour [3] considered the space $\mathbb{L}^2(\mathbb{X}^2, M)$ of square integrable functions on $\mathbb{X}^2$ equipped with the random product measure $M = \lambda_n \otimes \mu$ where $\lambda_n = n^{-1}\sum_{i=0}^{n-1} \delta_{X_i}$ and used the distance defined for $f, f' \in \mathbb{L}^2(\mathbb{X}^2, M)$ by

$$\delta^2\bigl(f, f'\bigr) = \frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\bigl(f(X_i, y) - f'(X_i, y)\bigr)^2 \,\mathrm{d}\mu(y).$$

Birgé [15] considered the cone $\mathbb{L}^1_+(\mathbb{X}^2, \mu\otimes\mu)$ of non-negative integrable functions and used the deterministic Hellinger-type distance defined for $f, f' \in \mathbb{L}^1_+(\mathbb{X}^2, \mu\otimes\mu)$ by

$$\delta^2\bigl(f, f'\bigr) = \frac{1}{2}\int_{\mathbb{X}^2}\bigl(\sqrt{f(x,y)} - \sqrt{f'(x,y)}\bigr)^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y).$$

These approaches, which often rely on the loss that is used, require the knowledge (or at least a suitable estimation) of various quantities depending on the unknown s, such as the supremum norm of s, or on a positive lower bound, either on the stationary density, or on $k^{-1}\sum_{j=1}^{k} s^{(l+j)}$ for some $k \geq 1$, $l \geq 0$ where $s^{(l+j)}(x,\cdot)$ is the density of the conditional law $\mathcal{L}(X_{l+j} \mid X_0 = x)$. Unfortunately, these quantities not only influence the way the estimators are built but also their performances since they are involved in the risk bounds. In the present paper, we shall rather consider the distance H (corresponding to an analogue of the random $\mathbb{L}^2$ loss above) defined on the cone $\mathbb{L}^1_+(\mathbb{X}^2, M)$ of integrable and non-negative functions by

$$H^2\bigl(f, f'\bigr) = \frac{1}{2n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\bigl(\sqrt{f(X_i, y)} - \sqrt{f'(X_i, y)}\bigr)^2 \,\mathrm{d}\mu(y) \quad \text{for all } f, f' \in \mathbb{L}^1_+\bigl(\mathbb{X}^2, M\bigr).$$

For such a loss, we shall show that our estimators satisfy an oracle-type inequality under very weak assumptions on the Markov chain. A connection with the usual deterministic Hellinger-type loss will be done under a posteriori assumptions on the chain, and hence, independently of the construction of the estimator.

Our estimation strategy can be viewed as a mix between an approach based on the minimization of a contrast and an approach based on robust tests. Estimation procedures based on tests started in the seventies with Lucien Lecam

and Lucien Birgé (Le Cam [30,31], Birgé [9–11]). More recently, Birgé [12] presented a powerful device to establish general oracle inequalities from robust tests. It was used in our statistical setting in Birgé [15] and in many others in Birgé [13,14] and Sart [35]. We make two contributions to this area. Firstly, we provide a new test for our statistical setting. This test is based on a variational formula inspired from Baraud [5] and differs from the one of Birgé [15]. Secondly, we shall study procedures that are quite far from the original one of Birgé [12]. Let us explain why.

The procedure of Birgé [12] depends on a suitable net, the construction of which is usually abstract, making thus the estimator impossible to build in practice. In the favourable cases where the net can be made explicit, the procedure is anyway too complex to be implemented (see for instance Section 3.4.2 of Birgé [13]). This procedure was afterwards adapted to estimators selection in Baraud and Birgé [6] (for histogram type estimators) and in Baraud [5] (for more general estimators). The complexity of their algorithms is of order the square of the cardinality of the family and are thus implementable when this family is not too large. In particular, given a family of histogram type estimators $\{\hat{s}_m, m \in \mathcal{M}\}$, these two procedures are interesting in practice when $\mathcal{M}$ is a collection of regular partitions (namely when all its elements have same Lebesgue measure) but become unfortunately numerically intractable for richer collections. In this work, we tackle this issue by proposing a new way of selecting among a family of piecewise constant estimators when the collection $\mathcal{M}$ ensues from the adaptive approximation algorithm of DeVore and Yu [21].

We present this procedure in the first part of the paper. It yields a piecewise constant estimator on a data-driven partition that satisfies an oracle-type inequality from which we shall deduce uniform rates of convergence over balls of (possibly) inhomogeneous Besov spaces with small regularity indices. These rates coincide, up to a possible logarithmic factor to the usual ones over such classes. Finally, we carry out numerical simulations to compare our estimator with the one of Akakpo and Lacour [3].

In the second part of this paper, we are interested in obtaining stronger theoretical results for our statistical problem. We put aside the practical considerations to focus on the construction of an estimator that satisfies a general model selection theorem. Such an estimator should be considered as a benchmark for what theoretically feasible. We deduce rates of convergence over a large range of anisotropic and inhomogeneous Besov spaces on $[0,1]^{2d}$. We shall also consider other kinds of assumptions on the transition density. We shall assume that s belongs to classes of functions satisfying structural assumptions and for which faster rates of convergence can be achieved. This approach was developed by Juditsky, Lepski and Tsybakov [26] (in the Gaussian white noise model) and by Baraud and Birgé [7] (in more statistical settings) to avoid the curse of dimensionality. More precisely, Baraud and Birgé [7] showed that these rates can be deduced from a general model selection theorem, which strengthen its theoretical interest. This strategy was used in Sart [35] to establish risk bounds over many classes of functions for Poisson processes with covariates. In the present paper, we shall use these assumptions to obtain faster rates of convergence for autoregressive Markov chains (whose conditional variance may not be constant).

This paper is organized as follows. The first procedure, which selects among piecewise constant estimators is presented and theoretically studied in Section 2. In Section 3, we carry out a simulation study and compare our estimator with the one of Akakpo and Lacour [3]. The practical implementation of this procedure is quite technical and will therefore be delayed in Appendix A. In Section 4, we establish theoretical results by using our second procedure. The proofs are postponed to Appendix B.

Let us introduce some notations that will be used all along the paper. The number $x \vee y$ (respectively $x \wedge y$) stands for $\max(x, y)$ (respectively $\min(x, y)$) and x_+ stands for $x \vee 0$. We set $\mathbb{N}^\star = \mathbb{N} \setminus \{0\}$. For (E, d) a metric space, $x \in E$ and $A \subset E$, the distance between x and A is denoted by $d(x, A) = \inf_{a \in A} d(x, a)$. The indicator function of a subset A is denoted by $\mathbb{1}_A$ and the restriction of a function f to A by $f|_A$. For all real valued function f on E, $\|f\|_\infty$ stands for $\sup_{x \in E} |f(x)|$. The cardinality of a finite set A is denoted by $|A|$. The notations $C, C', C'', \ldots$ are for the constants. The constants $C, C', C'', \ldots$ may change from line to line.

2. Selecting among piecewise constant estimators

Throughout this section, we assume that $\mathbb{X} = \mathbb{R}^d$, $A = [0,1]^{2d}$, $\mu([0,1]^d) = 1$ and $n > 3$.

2.1. *Preliminary estimators*

Given a (finite) partition m of $[0,1]^{2d}$, a simple way to estimate s on $[0,1]^{2d}$ is to consider the piecewise constant estimator on the elements of m defined by

$$\hat{s}_m = \sum_{K\in m} \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_K. \tag{2.1}$$

In the above definition, the denominator $\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x)\,\mathrm{d}\mu(x)$ may be equal to 0 for some sets K, in which case the numerator $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) = 0$ as well, and we shall use the convention $0/0 = 0$.

We now bound from above the risk of this estimator. We set

$$V_m = \left\{ \sum_{K\in m} a_K \mathbb{1}_K, \forall K \in m, a_K \in [0, +\infty) \right\}$$

and prove the following.

Proposition 2.1. *For all finite partition m of $[0,1]^{2d}$,*

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}_m)\big] \le \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m)\big] + \frac{1+\log n}{n}|m|,$$

where $C = 1/(4+\log 2)$.

Up to a constant, the risk of $\hat{s}_m$ is bounded by a sum of two terms. The first one corresponds to the approximation term whereas the second one corresponds to the estimation term.

An analogue upper bound on the empirical quadratic risk of this estimator may be found in Chapter 4 of Akakpo [1]. Her bound requires several assumptions on the partition m and the Markov chain although the present one requires none. However, unlike hers, we lose a logarithmic term.

2.2. *Definition of the partitions*

In this section, we shall deal with special choice of partitions m. More precisely, we consider the family of partitions defined by using the recursive algorithm developed in DeVore and Yu [21]. For $j \in \mathbb{N}$, we consider the set

$$\mathcal{L}_j = \big\{ \mathbf{l} = (l_1, \dots, l_{2d}) \in \mathbb{N}^{2d}, 1 \le l_i \le 2^j \text{ for } 1 \le i \le 2d \big\}$$

and define for all $\mathbf{l} = (l_1, \dots, l_{2d}) \in \mathcal{L}_j$,

$$\forall i \in \{1, \dots, 2d\}, \quad I_j(l_i) = \begin{cases} \left[\frac{l_i - 1}{2^j}, \frac{l_i}{2^j}\right) & \text{if } l_i < 2^j, \\ \left[\frac{l_i - 1}{2^j}, 1\right] & \text{if } l_i = 2^j. \end{cases}$$

We then introduce the cube $K_{j,\mathbf{l}} = \prod_{i=1}^{2d} I_j(l_i)$ and set $\mathcal{K}_j = \{K_{j,\mathbf{l}}, \mathbf{l} \in \mathcal{L}_j\}$.

The algorithm starts with $[0,1]^{2d}$. At each step, it gets a partition of $[0,1]^{2d}$ into a finite family of disjoint cubes of the form $K_{j,\mathbf{l}}$. For any such cube, one decides to divide it into the 4^d elements of $\mathcal{K}_{j+1}$ which are contained in it, or not. The set of all such partitions that can be constructed in less than ℓ steps is denoted by $\mathcal{M}_\ell$. We set $\mathcal{M}_\infty = \bigcup_{\ell \ge 1} \mathcal{M}_\ell$. Two examples of partitions are illustrated in Fig. 1 (for $d = 1$).

2.3. *The selection rule*

Given $\ell \in \mathbb{N}^\star \cup \{\infty\}$, the aim of this section is to select an estimator among the family $\{\hat{s}_m, m \in \mathcal{M}_\ell\}$.

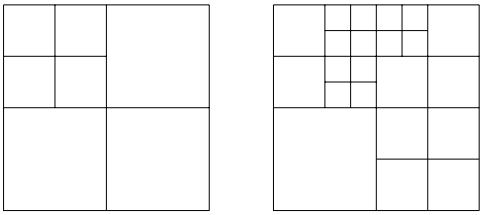


Fig. 1. Left: example of a partition of $\mathcal{M}_2$. Right: example of a partition of $\mathcal{M}_3$.

For any $K \in \bigcup_{m\in\mathcal{M}_\ell} m$ and any partition $m' \in \mathcal{M}_\ell$, let $m' \vee K$ be the partition of K defined by

$$m' \vee K = \big\{K' \cap K,\, K' \in m',\, K \cap K' \neq \varnothing\big\}.$$

Let L be a positive number and pen be the non-negative map defined by

$$\mathrm{pen}\big(m' \vee K\big) = L\frac{|m' \vee K|\log n}{n} \quad \text{for all } m' \in \mathcal{M}_\ell \text{ and } K \in \bigcup_{m\in\mathcal{M}_\ell} m.$$

This definition implies in particular that

$$\mathrm{pen}(m) = L\frac{|m|\log n}{n} \quad \text{for all partition } m \in \mathcal{M}_\ell.$$

Let us set $\alpha = (1 - 1/\sqrt{2})/2$ and for all $f, f' \in \mathbb{L}_+^1(\mathbb{X}^2, M)$,

$$\begin{aligned} T\big(f, f'\big) = {} & \frac{1}{2n\sqrt{2}}\sum_{i=0}^{n-1}\int_{\mathbb{X}} \sqrt{f(X_i, y) + f'(X_i, y)}\Big(\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}\Big)\,\mathrm{d}\mu(y) \\ & + \frac{1}{n\sqrt{2}}\sum_{i=0}^{n-1}\frac{\sqrt{f'(X_i, X_{i+1})} - \sqrt{f(X_i, X_{i+1})}}{\sqrt{f(X_i, X_{i+1}) + f'(X_i, X_{i+1})}} \\ & + \frac{1}{2n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\big(f(X_i, y) - f'(X_i, y)\big)\,\mathrm{d}\mu(y). \end{aligned}$$

We define γ for $m \in \mathcal{M}_\ell$ by

$$\gamma(m) = \left\{\sum_{K\in m}\sup_{m'\in\mathcal{M}_\ell}\Big[\alpha H^2(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'}\mathbb{1}_K) + T(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'}\mathbb{1}_K) - \mathrm{pen}\big(m' \vee K\big)\Big]\right\} + 2\,\mathrm{pen}(m).$$

Finally, we select $\hat{m}$ among $\mathcal{M}_\ell$ as any partition satisfying

$$\gamma(\hat{m}) \leq \inf_{m\in\mathcal{M}_\ell}\gamma(m) + \frac{1}{n} \tag{2.2}$$

and consider the resulting estimator $\hat{s} = \hat{s}_{\hat{m}}$.

Remarks

The estimator $\hat{s} = \hat{s}(L, \ell)$ depends on the choices of two quantities $L > 0$, $\ell \in \mathbb{N}^\star \cup \{\infty\}$. We shall see in the next section that L can be chosen as an universal numerical constant. As to ℓ, from a theoretical point of view, it can be chosen as $\ell = \infty$. In practice, we recommend to take it as large as possible. Nevertheless, the larger ℓ, the longer it takes to compute the estimator. A practical algorithm in view of computing $\hat{m}$ will be detailed in Appendix A.

The selection procedure we use may look somewhat unusual. It can be seen as a mix between a procedure based on a contrast function (which is usually easy to implement) and a procedure based on a robust test (the functional $T(f, f') = -T(f', f)$, which can be seen as a robust test between f, f', will allow us to obtain risk bounds with respect to a Hellinger-type distance). This functional is inspired from the variational formula for the Hellinger affinity described in Section 2 of Baraud [5].

In the literature, procedures based on a robust test are usually based on the minimization of a functional D known as plausibility index. In our context, D would be

$$D(m) = \sup\bigl\{H^2(\hat{s}_m, \hat{s}_{m'}), m' \in \mathcal{M}_\ell, T(\hat{s}_m, \hat{s}_{m'}) \geq \mathrm{pen}\bigl(m'\bigr) - \mathrm{pen}(m)\bigr\} \quad \text{for all partition } m \in \mathcal{M}_\ell,$$

and the estimator would be defined by $\hat{s}_{\tilde{m}}$ where $\tilde{m}$ minimizes $D(m)$ over $m \in \mathcal{M}_\ell$. The computation of $D(m)$ is unfortunately numerically intractable, which implies that $\tilde{m}$ is purely theoretical. The computation of the supremum is a constraint optimization problem and the Lagrange multipliers suggest to replace $D(m)$ by

$$\gamma_1(m) = \sup_{m' \in \mathcal{M}_\ell} \bigl[\alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \mathrm{pen}\bigl(m'\bigr)\bigr] + \mathrm{pen}(m).$$

The contrast γ can be interpreted as being a modification of γ_1 whose minimum can be found in practice. The minimums of γ, γ_1 and D may not be equal, but it can be shown that they possess similar statistical properties.

2.4. *An oracle inequality*

The main result of this section is the following.

Theorem 2.1. *There exists an universal constant $L_0 > 0$ such that, for all $L \geq L_0$, $\ell \in \mathbb{N}^\star \cup \{\infty\}$, the estimator $\hat{s} = \hat{s}(L, \ell)$ satisfies*

$$C\mathbb{E}\bigl[H^2(s\mathbb{1}_A, \hat{s})\bigr] \leq \inf_{m \in \mathcal{M}_\ell} \bigl\{\mathbb{E}\bigl[H^2(s\mathbb{1}_A, V_m)\bigr] + \mathrm{pen}(m)\bigr\}, \tag{2.3}$$

where C is an universal positive constant.

In the literature, oracle inequalities with a random quadratic loss for piecewise constant estimators have been obtained in Lacour [27] and Akakpo and Lacour [3]. Their procedures require a priori assumptions on the transition density and the Markov chain although ours requires none (except homogeneity). However, unlike theirs, our risk bound involves an extra logarithmic term. We do not know whether this term is necessary or not.

In the proof, we obtain an upper bound for L_0 which is unfortunately very rough and useless in practice. It seems difficult to obtain a sharp bound on L_0 from the theory and we have rather carried out a simulation study in order to tune L_0 (see Section 3).

2.5. *Risk bounds with respect to a deterministic loss*

Although the distance H is natural, we are interested in controlling the risk associated to a deterministic distance. To do so, we shall make a posteriori assumptions on the Markov chain.

Assumption 2.1. *The sequence $(X_i)_{i \geq 0}$ is stationary and admits a stationary density φ with respect to the Lebesgue measure μ on $\mathbb{R}^d$. There exists $\kappa_0 > 0$ such that $\varphi(x) \geq \kappa_0$ for all $x \in [0, 1]^d$.*

We introduce $\mathbb{L}_+^1([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu)$ the cone of integrable and non-negative functions on $[0, 1]^{2d}$ with respect to the product measure $(\varphi \cdot \mu) \otimes \mu$. We endow $\mathbb{L}_+^1([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu)$ with the distance h defined by

$$\forall f, f' \in \mathbb{L}_+^1\bigl([0, 1]^{2d}, (\varphi \cdot \mu) \otimes \mu\bigr), \quad h^2\bigl(f, f'\bigr) = \frac{1}{2} \int_{[0,1]^{2d}} \bigl(\sqrt{f(x, y)} - \sqrt{f'(x, y)}\bigr)^2 \varphi(x)\, \mathrm{d}x\, \mathrm{d}y.$$

In our results, we shall need the β-mixing properties of the Markov chain. We set for all $q \in \mathbb{N}^\star$

$$\beta_q = \int_{\mathbb{R}^{2d}} \big|s^{(q)}(x,y) - \varphi(y)\big|\varphi(x)\,\mathrm{d}x\,\mathrm{d}y,$$

where $s^{(q)}(x,\cdot)$ is the density of the conditional law $\mathcal{L}(X_q \mid X_0 = x)$ with respect to the Lebesgue measure. We refer to Doukhan [23] and Bradley [17] for more details on the β-mixing coefficients.

Theorem 2.2. *Under Assumption 2.1, the estimator $\hat{s}$ built in Section 2.3 with $\ell \in \mathbb{N}^\star$ and $L \geq L_0$, satisfies*

$$C\mathbb{E}\big[h^2(s\mathbb{1}_A, \hat{s})\big] \leq \inf_{m\in\mathcal{M}_\ell}\big\{h^2(s\mathbb{1}_A, V_m) + \mathrm{pen}(m)\big\} + \frac{R_n(\ell)}{n},$$

where

$$R_n(\ell) = n2^{3\ell d} \inf_{1\leq q\leq n}\left\{\exp\left(-\frac{\kappa_0}{10}\frac{n}{q2^{\ell d}}\right) + n\beta_q/q\right\} \tag{2.4}$$

and where C is an universal positive constant.

This result is interesting when the remainder term $R_n(\ell)/n$ is small enough, that is when $2^{\ell d}$ is small compared to n and when the sequence $(\beta_q)_{q\geq 1}$ goes to 0 fast enough. More precisely, $R_n(\ell)$ can be bounded independently of n, ℓ whenever ℓ, d, n and the β_q coefficients satisfy the following.

- If the chain is geometrically β-mixing, that is if there exists $b_1 > 0$ such that $\beta_q \leq \mathrm{e}^{-b_1 q}$, then

$$R_n(\ell) \leq n^2 2^{3\ell d+1}\left[\exp(-b_1 n) + \exp\left(-\frac{\kappa_0}{10}\frac{n}{2^{\ell d}}\right) + \exp\left(-\sqrt{\frac{\kappa_0 b_1}{40}\frac{n}{2^{\ell d}}}\right)\right].$$

  In particular, if ℓ, d, n are such that $2^{\ell d} \leq n/\log^3 n$, $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_1.
- If the chain is arithmetically β-mixing, that is if there exists $b_2 > 0$ such that $\beta_q \leq q^{-b_2}$, then

$$R_n(\ell) \leq \frac{C'(b_2)}{\kappa_0^{b_2+1}}\frac{2^{(4+b_2)\ell d}\log^{b_2+1}(1+\kappa_0 n/2^{\ell d})}{n^{b_2-1}},$$

  where $C'(b_2)$ depends only on b_2. Consequently, if $2^{\ell d} \leq n^{1-\zeta}/\log n$ and $b_2 \geq 5/\zeta - 4$ for $\zeta \in (0,1)$, $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_2.

2.6. *Rates of convergence*

The aim of this section is to obtain uniform risk bounds over classes of smooth transition densities for our estimator.

2.6.1. *Hölder spaces*

Given $\sigma \in (0,1]$, we say that a function f belongs to the Hölder space $\mathcal{H}^\sigma([0,1]^{2d})$ if there exists $|f|_\sigma \in \mathbb{R}_+$ such that for all $(x_1,\ldots,x_{2d}) \in [0,1]^{2d}$ and all $1 \leq j \leq 2d$, the functions $f_j(\cdot) = f(x_1,\ldots,x_{j-1},\cdot,x_{j+1},\ldots,x_{2d})$ satisfy

$$\big|f_j(x) - f_j(y)\big| \leq |f|_\sigma |x-y|^\sigma \quad \text{for all } x, y \in [0,1].$$

When the restriction of $\sqrt{s}$ to $A = [0,1]^{2d}$ is Hölderian, we deduce from (2.3) the following.

Corollary 2.1. *For all $\sigma \in (0,1]$ and $\sqrt{s}|_A \in \mathcal{H}^\sigma([0,1]^{2d})$, the estimator $\hat{s} = \hat{s}(L_0,\infty)$ satisfies*

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s})\big] \leq \big(d|\sqrt{s}|_A|_\sigma\big)^{2d/(d+\sigma)}\left(\frac{\log n}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log n}{n},$$

where C is an universal positive constant.

2.6.2. *Besov spaces*

A thinner way to measure the smoothness of the transition density is to assume that $\sqrt{s}|_A$ belongs to a Besov space. We refer to Section 3 of DeVore and Yu [21] for a definition of this space. We say that the Besov space $\mathscr{B}_q^\sigma(\mathbb{L}^p([0,1]^{2d}))$ is homogeneous when $p \geq 2$ and inhomogeneous otherwise. We set for all $p \in (1,+\infty)$ and $\sigma \in (0,1)$,

$$\mathscr{B}^\sigma\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) = \begin{cases} \mathscr{B}_p^\sigma\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) & \text{if } p \in (1,2), \\ \mathscr{B}_\infty^\sigma\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) & \text{if } p \in [2,+\infty), \end{cases}$$

and denote by $|\cdot|_{p,\sigma}$ the semi norm of $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$. We make the following assumption to deduce from (2.3) risk bounds over these spaces.

Assumption 2.2. *There exists $\kappa > 0$ such that for all $i \in \{0,\dots,n-1\}$, X_i admits a density φ_i with respect to the Lebesgue measure μ such that $\varphi_i(x) \leq \kappa$ for all $x \in [0,1]^d$.*

Note that we do not require that the chain be either stationary or mixing.

Let $(\mathbb{L}^2([0,1]^{2d}, \mu\otimes\mu), d_2)$, be the metric space of square integrable functions on $[0,1]^{2d}$ with respect to the Lebesgue measure. Under the above assumption, we deduce from (2.3) that

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})\big] \leq \inf_{m\in\mathcal{M}_\ell}\left\{\kappa d_2^2(\sqrt{s}|_A, V_m) + L_0\frac{|m|\log n}{n}\right\}.$$

When $\sqrt{s}|_A$ belongs to a Besov space, the right-hand side of this inequality can be upper bounded thanks to the approximation theorems of DeVore and Yu [21].

Corollary 2.2. *Suppose that Assumption 2.2 holds. For all $p \in (2d/(d+1), +\infty)$, $\sigma \in (2d(1/p-1/2)_+, 1)$ and $\sqrt{s}|_A \in \mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$, the estimator $\hat{s} = \hat{s}(L_0,\infty)$ satisfies*

$$C'\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s})\big] \leq |\sqrt{s}|_A|_{p,\sigma}^{2d/(d+\sigma)}\left(\frac{\log n}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log n}{n}, \tag{2.5}$$

where $C' > 0$ depends only on κ, σ, d, p.

More precisely, it is shown in the proof that the estimators $\hat{s} = \hat{s}(L_0,\ell)$ satisfy (2.5) when ℓ is large enough (when $\ell \geq d^{-1}(\log 2)^{-1}\log n$).

Rates of convergence for the deterministic loss h can be established by using Theorem 2.2 instead of Theorem 2.1. For instance, if the chain is geometrically β-mixing, we may choose ℓ the smallest integer larger than $d^{-1}(\log 2)^{-1}\log(n/\log^3 n)$, in which case the estimator $\hat{s} = \hat{s}(L_0,\ell)$ achieves the rate $(\log n/n)^{\sigma/(\sigma+d)}$ over the Besov spaces $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$, $p \in (2d/(d+1), +\infty)$, $\sigma \in (\sigma_1(p,d), 1)$ where

$$\sigma_1(p,d) = \frac{d}{4}\big(-1 + 4(1/p-1/2)_+ + \sqrt{1 + 24(1/p-1/2)_+ + 16(1/p-1/2)_+^2}\big).$$

If the chain is arithmetically β-mixing with $b_q \leq q^{-6}$, choosing ℓ the smallest integer larger than $d^{-1}(2\log 2)^{-1} \times \log(n/\log n)$ allows us to recover the same rate of convergence when $\sigma \in (\sigma_2(p,d), 1)$ where

$$\sigma_2(p,d) = d\big((1/p-1/2)_+ + \sqrt{2(1/p-1/2)_+ + (1/p-1/2)_+^2}\big).$$

We refer the reader to Section B.6 for a proof of these two results.

In the literature, Lacour [27] obtained a rate of order $n^{-\sigma/(\sigma+1)}$ over $\mathscr{B}^\sigma(\mathbb{L}^2([0,1]^2))$, which is slightly faster but her approach prevents her to deal with inhomogeneous Besov spaces and requires the prior knowledge of a suitable upper bound on the supremum norm of s. As far as we know, the rates that have been established in the other papers hold only when $\sigma > 1$.

3. Simulations

In this section, we present a simulation study to evaluate the performance of our estimator in practice. We shall simulate several Markov chains and estimate their transition densities by using our procedure.

3.1. *Examples of Markov chains*

We consider Markov chains of the form

$$X_{k+1} = F(X_k, U_k),$$

where F is some known function and where U_k is a random variable independent of $(X_0, \dots, X_k)$.

For the sake of comparison, we begin to deal with examples that have already been considered in the simulation study of Akakpo and Lacour [3]. In each of these examples, U_k is a standard Gaussian random variable.

Example 3.1. $X_{k+1} = 0.5X_k + (1 + U_k)/4$.

Example 3.2. $X_{k+1} = 12^{-1}(6 + \sin(12X_k - 6) + (\cos(X_k - 6) + 3)U_k)$.

Example 3.3.

$$X_{k+1} = \frac{1}{3}(X_k + 1) + \left(\frac{1}{9} - \frac{1}{23}\left(\frac{1}{2}\beta(5X_i/3, 4, 4) + \frac{1}{20}\beta\big((5X_i - 2)/3, 400, 400\big)\right)\right)U_k,$$

where $\beta(\cdot, a, b)$ is the density of the β distribution with parameters a and b.

Example 3.4.

$$X_{k+1} = \frac{1}{4}\big(g(X_k) + 1\big) + \frac{1}{8}U_k,$$

where g is defined by

$$g(x) = \frac{9\sqrt{2}}{4\sqrt{\pi}}\exp\big(-18(x - 1/2)^2\big) + \frac{9\sqrt{2}}{4\sqrt{\pi}}\exp\big(-162(x - 3/4)^2\big) \quad \textit{for all } x \in \mathbb{R}.$$

At first sight, Examples 3.1 and 3.2 may seem to be different than those of Akakpo and Lacour [3]. Actually, we just have rescaled the data in order to estimate on $[0, 1]^2$. The statistical problem is the same. According to Akakpo and Lacour [3], we set p large ($p = 10^4$) and we estimate the transition densities of Examples 3.1, 3.2, 3.3 and 3.4 from $(X_p, \dots, X_{n+p})$ so that the chain is approximatively stationary.

We also propose to consider the following examples. In Example 3.5, U_k is a centred Gaussian random variable with variance $1/2$, in Example 3.6, U_k admits the density

$$f(x) = \frac{5\sqrt{2}}{2\sqrt{\pi}}\big[\exp\big(-50(x - 1)^2\big) + \exp\big(-50x^2\big)\big]$$

with respect to the Lebesgue measure, and in Example 3.7, U_k is an exponential random variable with parameter 1.

Example 3.5. $X_{k+1} = 0.5X_k + (1 + U_k)/4$.

Example 3.6. $X_{k+1} = 0.5(X_k + U_k)$.

Example 3.7. $X_{k+1} = X_k/(50X_k + 1) + X_kU_k$.

We set $X_0 = 1/2$ and estimate s from $(X_0, \dots, X_n)$. These last three Markov chains are not stationary. Their transition densities are rather isotropic and inhomogeneous. The transition density of Example 3.7 is unbounded.

In what follows, our selection rule will always be applied with $L = 0.03$ (whatever, ℓ, n and the Markov chain).

3.2. *Choice of ℓ*

We discuss the choice of ℓ by simulating the preceding examples with $n = 10^3$ and by applying our selection rule for each value of $\ell \in \{1, \dots, 10\}$. The results are summarized in Table 1. When ℓ grows up, the risk of our estimator tends to decrease and then stabilize. The best choice of ℓ is obviously unknown in practice but this array shows that a good way for choosing ℓ is to take it as large as possible. This is theoretically justified by Theorem 2.1 since the right-hand side of inequality (2.3) is a non-increasing function of ℓ.

3.3. *An illustration*

We apply our procedure for Examples 3.1 and 3.6 with $n = 10^4$, $\ell = 7$. We get two estimators and draw them with the corresponding transition density in Figure 2.

This shows that the selected partition is thinner (respectively wider) to the points where the transition density is changing rapidly (respectively slower), and is thus rather well adapted to the target function s.

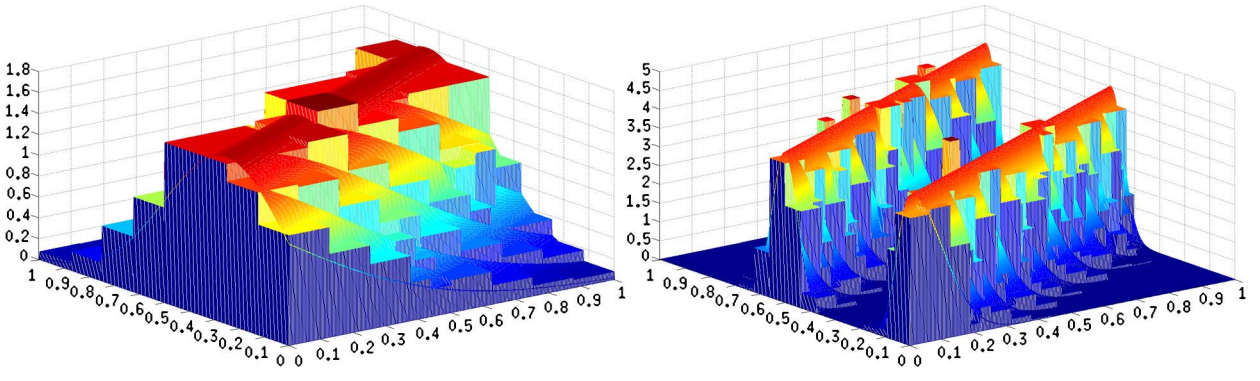
3.4. *Comparison with other procedures*

In this section, we compare our selection rule with the oracle estimator and with the piecewise constant estimator of Akakpo and Lacour [3].

The procedure of Akakpo and Lacour [3] amounts to selecting an estimator among $\{\hat{s}_m, m \in \mathcal{M}'\}$ where $\hat{s}_m$ is defined by (2.1) and where $\mathcal{M}'$ is a collection of irregular partitions on $[0, 1]^2$. Precisely, with their notations, we

Table 1
Hellinger risk $H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})$

ℓ	Ex. 3.1	Ex. 3.2	Ex. 3.3	Ex. 3.4	Ex. 3.5	Ex. 3.6	Ex. 3.7
1	0.031	0.046	0.299	0.181	0.089	0.291	0.358
2	0.011	0.015	0.087	0.107	0.024	0.170	0.241
3	0.011	0.014	0.026	0.058	0.013	0.067	0.156
4	0.011	0.018	0.026	0.035	0.015	0.046	0.113
5	0.011	0.018	0.022	0.038	0.015	0.048	0.098
6	0.011	0.018	0.022	0.038	0.015	0.048	0.065
7	0.011	0.018	0.024	0.038	0.015	0.048	0.044
8	0.011	0.018	0.024	0.038	0.015	0.048	0.040
9	0.011	0.018	0.024	0.038	0.015	0.048	0.040
10	0.011	0.018	0.024	0.038	0.015	0.048	0.040



Example 3.1 Example 3.6

Fig. 2. Estimator and transition density.

apply it with $J_\star = 5$, $\mathrm{pen}(m) = 3\|s_{|A}\|_\infty |m|/n$ and with $\mathrm{pen}(m) = 3\|\hat{s}_{m^\bullet}\|_\infty |m|/n$ where $m^\bullet$ is a partition suitably chosen (following the recommendations of Akakpo and Lacour [3], that is $J_\bullet = 3$). These two estimators are denoted by $\hat{s}^{(1)}$ and $\hat{s}^{(2)}$ respectively. Notice that these penalties, which are used in their simulation study, are not the ones prescribed by their theory. Their theoretical penalties also depend on a positive lower bound on the stationary density.

We denote by $\hat{s}^{(0)}$ the oracle estimator, that is the estimator defined as being a minimizer of the map $m \mapsto H^2(s1_{[0,1]^2}, \hat{s}_m)$ for $m \in \mathcal{M}_7$. This estimator is the best estimator of the family $\{\hat{s}_m, m \in \mathcal{M}_7\}$ and is known since the data are simulated. We consider the random variables

$$\mathcal{R}_i = \frac{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})}{H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(i)})} \quad \text{for } i = 1, 2$$

and denote by $q_0(\alpha)$ the α-quantile of $\mathcal{R}_0$. Results obtained are given in Table 2.

3.5. *Comparison with a quadratic empirical risk*

In Akakpo and Lacour [3], the risks of the estimators are evaluated with a empirical quadratic norm and we can also compare the performances of our estimator to theirs by using this risk.

To do so, let us denote by $\|\cdot\|_n$ the empirical quadratic norm defined by

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}} f^2(X_i, x)\,\mathrm{d}x \quad \text{for all } f \in \mathbb{L}^2\big(\mathbb{R}^2, M\big)$$

and set for $i \in \{1, 2\}$,

$$\mathcal{R}_i' = \frac{\|s\mathbb{1}_{[0,1]^2} - \hat{s}\|_n^2}{\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(i)}\|_n^2}.$$

The results obtained are presented in Table 3. They are very similar to those of Table 2.

4. A general procedure

In Section 2, we used our selection rule to establish the oracle inequality (2.3), from which we deduced rates of convergence over Besov spaces $\mathscr{B}^\sigma(\mathbb{L}^p([0,1]^{2d}))$ with σ lower than 1. We now aim at obtaining rates for more general spaces of functions. This includes Besov spaces with regularity index larger than 1 and spaces corresponding to structural assumptions on s. We propose a second procedure to reach this goal.

The Markov chain takes its values into $\mathbb{X}$ and we estimate s on a subset A of the form $A = A_1 \times A_2$. We always assume that $n > 3$.

Table 2
Risks for simulated data with $n = 1000$ averaged over 250 samples

	Ex. 3.1	Ex. 3.2	Ex. 3.3	Ex. 3.4	Ex. 3.5	Ex. 3.6	Ex. 3.7
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s})]$	0.011	0.017	0.022	0.038	0.018	0.052	0.049
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(0)})]$	0.007	0.011	0.015	0.028	0.012	0.037	0.041
$q_0(0.5)$	1.473	1.513	1.443	1.369	1.422	1.420	1.200
$q_0(0.75)$	1.698	1.627	1.557	1.440	1.575	1.481	1.244
$q_0(0.9)$	1.921	1.834	1.683	1.509	1.749	1.543	1.290
$q_0(0.95)$	2.113	1.965	1.770	1.558	1.839	1.590	1.317
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(1)})]$	0.017	0.018	0.028	0.058	0.024	0.103	–
$\mathbb{P}(\mathcal{R}_1 \le 1)$	0.964	0.740	0.908	1	0.984	1	–
$\mathbb{E}[H^2(s\mathbb{1}_{[0,1]^2}, \hat{s}^{(2)})]$	0.013	0.018	0.028	0.062	0.023	0.096	0.133
$\mathbb{P}(\mathcal{R}_2 \le 1)$	0.832	0.748	0.928	1	0.948	1	1

Table 3
Risks for simulated data with $n = 1000$ averaged over 250 samples

	Ex. 3.1	Ex. 3.2	Ex. 3.3	Ex. 3.4	Ex. 3.5	Ex. 3.6	Ex. 3.7
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}\|_n^2]$	0.064	0.108	0.229	0.319	0.116	0.528	2.82
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(1)}\|_n^2]$	0.147	0.133	0.257	0.423	0.205	0.743	–
$\mathbb{P}(\mathcal{R}_1' \le 1)$	0.980	0.820	0.788	0.984	0.992	1	–
$\mathbb{E}[\|s\mathbb{1}_{[0,1]^2} - \hat{s}^{(2)}\|_n^2]$	0.091	0.129	0.262	0.418	0.159	0.739	6.08
$\mathbb{P}(\mathcal{R}_2' \le 1)$	0.864	0.780	0.792	0.980	0.940	1	1

4.1. *Procedure and preliminary result*

Our second procedure is defined as follows. Let $\alpha = (1 - 1/\sqrt{2})/2$, $L > 0$, S be an at most countable set of $\mathbb{L}_+^1(\mathbb{X}^2, M)$ and $\Delta_S \ge 1$ be a map on S.

We define the application $\wp$ on S by

$$\wp(f) = \sup_{f' \in S} \left[\alpha H^2\big(f, f'\big) + T\big(f, f'\big) - L\frac{\Delta_S(f')}{n} \right] + L\frac{\Delta_S(f)}{n} \quad \text{for all } f \in S.$$

We select $\hat{s}$ among S as any element of S satisfying

$$\wp(\hat{s}) \le \inf_{f \in S} \wp(f) + \frac{1}{n}.$$

We prove the following.

Proposition 4.1. *Suppose that $f(x) = 0$ for all $f \in S$ and $x \in \mathbb{X}^2 \setminus A$ and that $\sum_{f \in S} \mathrm{e}^{-\Delta_S(f)} \le 1$. There exists an universal constant $L_0 > 0$ such that if $L \ge L_0$, the estimator $\hat{s}$ satisfies*

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s})\big] \le \mathbb{E}\left[\inf_{f \in S} \left\{ H^2(s\mathbb{1}_A, f) + L\frac{\Delta_S(f)}{n} \right\} \right], \tag{4.1}$$

where C is an universal positive constant.

4.2. *A general model selection theorem*

We shall deduce from the above proposition a model selection theorem by choosing suitably S. To do so, we consider the following assumption.

Assumption 4.1. *For all $i \in \{1, \dots, n-1\}$, X_i admits a density φ_i with respect to some known measure ν such that $\nu(A_1) = 1$. Moreover, there exists κ such that $\varphi_i(x) \le \kappa$ for all $x \in A_1$ and $i \in \{1, \dots, n-1\}$.*

We define $\mathbb{L}^2(A, \nu \otimes \mu)$ the space of square integrable functions on A with respect to the product measure $\nu \otimes \mu$, and we endow it with its natural distance

$$d^2\big(f, f'\big) = \int_A \big(f(x, y) - f'(x, y)\big)^2 \,\mathrm{d}\nu(x)\,\mathrm{d}\mu(y) \quad \text{for all } f, f' \in \mathbb{L}^2(A, \nu \otimes \mu).$$

Hereafter, a model V is a (non-trivial) finite dimensional linear space of $\mathbb{L}^2(A, \nu \otimes \mu)$.

Let us explain how to obtain a model selection theorem when Assumption 4.1 holds. Let $\mathbb{V}$ be a collection of models V and let $(\Delta(V))_{V\in\mathbb{V}}$ be a family of non-negative numbers such that $\sum_{V\in\mathbb{V}} \mathrm{e}^{-\Delta(V)} \leq 1$. For each model $V \in \mathbb{V}$, we consider an orthonormal basis $(f_1, \ldots, f_{\dim V})$ of V and set

$$T_V = \left\{ \sum_{i=1}^{\dim V} \alpha_i f_i, \alpha_i \in \frac{2}{\sqrt{n \dim V}}\mathbb{Z} \right\}.$$

We deduce from Lemma 5 of Birgé [12] that the cardinality of $S_V = \{f_+^2 \mathbb{1}_A, f \in T_V, d(f,0) \leq 2\}$ is upper bounded by $|S_V| \leq (30n)^{\dim V/2}$. We then use the above procedure with $S = \bigcup_{V\in\mathbb{V}} S_V$ and

$$\Delta_S(f) = \inf_{\substack{V\in\mathbb{V} \\ S_V \ni f}} \left\{ \Delta(V) + (\dim V)\log(30n)/2 \right\} \quad \text{for all } f \in S.$$

This yields an estimator $\hat{s}$ such that

$$C'\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ \kappa\Big(\inf_{\substack{f\in T_V \\ d(f,0)\leq 2}} d^2(\sqrt{s}|_A, f) \Big) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\},$$

where C' is an universal positive constant. Since $d(\sqrt{s}|_A, 0) \leq 1$,

$$\inf_{\substack{f\in T_V \\ d(f,0)\leq 2}} d^2(\sqrt{s}|_A, f) = d^2(\sqrt{s}|_A, T_V).$$

For all $f' \in V$, there exists $f \in T_V$ such that $d^2(f, f') \leq n^{-1}$ and thus

$$d^2(\sqrt{s}|_A, T_V) \leq 2d^2(\sqrt{s}|_A, V) + \frac{2}{n}.$$

Precisely, we have proved:

Theorem 4.1. *Suppose that Assumption 4.1 holds. Let $\mathbb{V}$ be an at most countable collection of models. Let $(\Delta(V))_{V\in\mathbb{V}}$ be a family of non-negative numbers such that*

$$\sum_{V\in\mathbb{V}} \mathrm{e}^{-\Delta(V)} \leq 1.$$

There exists an estimator $\hat{s}$ such that

$$C\mathbb{E}\left[H^2(s\mathbb{1}_A, \hat{s})\right] \leq \inf_{V\in\mathbb{V}} \left\{ d^2(\sqrt{s}|_A, V) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\},$$

where $C > 0$ depends only on κ.

The condition $\sum_{\mathbb{V}\in\mathbb{V}} \mathrm{e}^{-\Delta(V)} \leq 1$ can be interpreted as a (sub)probability on the collection $\mathbb{V}$. The more complex the family $\mathbb{V}$, the larger the weights $\Delta(V)$. When one can choose $\Delta(V)$ of order $\dim(V)$, which means that the family $\mathbb{V}$ of models does not contains too many models per dimension, the estimator $\hat{s}$ achieves the best trade-off (up to a constant) between the approximation and the variance terms.

This theorem holds under an assumption that is very mild and weaker than those of Lacour [27], Akakpo and Lacour [3] and Clémencon [18]. Birgé [15] proved a general oracle inequality when there exist integers $k \geq 1$ and $l \geq 0$ and positive numbers ρ, ϱ such that

$$\varrho \leq \frac{1}{k}\sum_{j=1}^{k} s^{(l+j)}(x, y) \leq \rho \quad \text{for all } x, y \in \mathbb{X},$$

where the parameters k, l, ϱ are known. Our assumption is then satisfied for the Markov chain $(X_{l+1}, \dots, X_n)$ with $\nu = \mu$ and $\kappa = k\rho$.

We shall consider subsets $\mathscr{F} \subset \mathbb{L}^2(A, \nu \otimes \mu)$ corresponding to smoothness or structural assumptions on $\sqrt{s}|_A$. For such an $\mathscr{F}$, we associate a collection $\mathbb{V}$ and deduce from Theorem 4.1 a risk bound for the estimator $\hat{s}$ when $\sqrt{s}|_A$ belongs to $\mathscr{F}$. This set is a generic notation and will change from section to section. In the remaining part of this paper, we shall always choose $\mathbb{X}^2 = \mathbb{R}^{2d}$, $A = [0,1]^{2d}$ and μ the Lebesgue measure.

4.3. *Smoothness assumptions*

We have introduced in Section 2.6 the isotropic Besov spaces $\mathscr{B}_q^\sigma(\mathbb{L}^p([0,1]^{2d}))$ where $\sigma \in (0,1)$. In this section, we consider the anisotropic Besov spaces $\mathscr{B}_q^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{2d})$ belongs to $(0, +\infty)^{2d}$.

Intuitively, a function f on $[0,1]^{2d}$ belongs to $\mathscr{B}_q^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ if, for all integer $j \in \{1, \dots, 2d\}$, and all $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_{2d} \in [0,1]$ the function

$$x_j \mapsto f(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_{2d})$$

belongs to $\mathscr{B}_q^{\sigma_j}(\mathbb{L}^p([0,1]))$. In particular, for all $\sigma \in (0, +\infty)$,

$$\mathscr{B}_q^\sigma\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) = \mathscr{B}_q^{(\sigma,\dots,\sigma)}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big).$$

A definition of the anisotropic Besov spaces may be found in Hochmuth [25] (for $d = 1$) and in Akakpo [1] (for larger values of d). We also consider the space $\mathcal{H}^{\boldsymbol{\sigma}}([0,1]^{2d})$ of anisotropic Hölderian functions on $[0,1]^{2d}$ with regularity $\boldsymbol{\sigma}$. A precise definition of this space may be found in Section 3.1.1 of Baraud and Birgé [7] (among other references).

For all $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{2d}) \in (0, +\infty)^{2d}$, we denote by $\bar{\boldsymbol{\sigma}}$ the harmonic mean of $\boldsymbol{\sigma}$:

$$\frac{1}{\bar{\boldsymbol{\sigma}}} = \frac{1}{2d} \sum_{i=1}^{2d} \frac{1}{\sigma_i}.$$

We set for all $p \in (0, +\infty]$,

$$\mathscr{B}^{\boldsymbol{\sigma}}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) = \begin{cases} \mathscr{B}_\infty^{\boldsymbol{\sigma}}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) & \text{if } p \in (0,1], \\ \mathscr{B}_p^{\boldsymbol{\sigma}}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) & \text{if } p \in (1,2), \\ \mathscr{B}_\infty^{\boldsymbol{\sigma}}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) & \text{if } p \in [2, +\infty), \\ \mathcal{H}^{\boldsymbol{\sigma}}\big([0,1]^{2d}\big) & \text{if } p = \infty \end{cases}$$

and denote by $|\cdot|_{p,\boldsymbol{\sigma}}$ the semi norm associated to the space $\mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$.

In this section, we are interesting in obtaining a bound risk when $\sqrt{s}|_A$ belongs to the space

$$\mathscr{B}\big([0,1]^{2d}\big) = \bigcup_{p \in (0,+\infty]} \Bigg(\bigcup_{\substack{\boldsymbol{\sigma} \in (0,+\infty)^d \\ \bar{\boldsymbol{\sigma}} > 2d(1/p - 1/2)_+}} \mathscr{B}^{\boldsymbol{\sigma}}\big(\mathbb{L}^p\big([0,1]^{2d}\big)\big) \Bigg).$$

Families of linear spaces possessing good approximation properties with respect to the elements of $\mathscr{F} = \mathscr{B}([0,1]^{2d})$ can be found in Theorem 1 of Akakpo [2]. We then deduce from Theorem 4.1,

Corollary 4.1. *Suppose that Assumption 4.1 holds with $\mathbb{X} = \mathbb{R}^d$, $A = [0,1]^{2d}$ and with $\nu \otimes \mu$ the Lebesgue measure. There exists an estimator $\hat{s}$ such that for all $\sqrt{s}|_A \in \mathscr{B}([0,1]^{2d})$,*

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s})\big] \le |\sqrt{s}|_A|_{p,\boldsymbol{\sigma}}^{2d/(d+\bar{\boldsymbol{\sigma}})} \left(\frac{\log n}{n}\right)^{\bar{\boldsymbol{\sigma}}/(\bar{\boldsymbol{\sigma}}+d)} + \frac{\log n}{n},$$

where $p \in (0, +\infty]$, $\boldsymbol{\sigma} \in (0, +\infty)^{2d}$, $\bar{\boldsymbol{\sigma}} > 2d(1/p - 1/2)_+$ are such that $\sqrt{s}|_A \in \mathscr{B}^{\boldsymbol{\sigma}}(\mathbb{L}^p([0,1]^{2d}))$ and where $C > 0$ depends only on $\kappa, d, p, \boldsymbol{\sigma}$.

To our knowledge, the only statistical procedures that can adapt both to possible inhomogeneity and anisotropy of s are those of Akakpo and Lacour [3] and Birgé [15]. The losses are different, but the rates are the same as ours (up to the logarithmic term). In view of our assumptions, we do not know if the logarithmic term can be avoided.

In the following sections, we consider classes $\mathscr{F}$ corresponding to structural assumptions on $\sqrt{s}|_A$. More precisely, rates of convergence when the chain is autoregressive with constant conditional variance (respectively non-constant conditional variance) are established in Section 4.4 (respectively Section 4.5).

4.4. *AR model*

In this section, we assume that $X_{n+1} = g(X_n) + \varepsilon_n$ where g is an unknown function and where the ε_n's are unobserved identically distributed random variables. Many papers are devoted to the estimation of the regression function g and it is beyond the scope of this paper to make an historical review for this statistical problem.

For the sake of simplicity, one shall assume throughout this section that $\mathbb{X} = \mathbb{R}$, $A = [0,1]^2$. The transition density is of the form $s(x,y) = \varphi(y - g(x))$ where φ is the density of ε_0. Since g and φ are both unknown, this suggests us to consider the class

$$\mathscr{F} = \bigcup_{\sigma>0} \big\{ f, \exists \phi \in \mathcal{H}^\sigma(\mathbb{R}), \exists g \in \mathscr{B}\big([0,1]\big), \|g\|_\infty < \infty, \forall x, y \in [0,1],\ f(x,y) = \phi\big(y - g(x)\big) \big\}.$$

A family $\mathbb{V}$ of linear spaces possessing good approximation properties with respect to the functions of $\mathscr{F}$ can be built by using Section 6.2 of Baraud and Birgé [7]. Precisely, we prove the following.

Corollary 4.2. *Suppose that Assumption 4.1 holds with $\mathbb{X} = \mathbb{R}$, $A = [0,1]^2$ and with $\nu \otimes \mu$ the Lebesgue measure on $\mathbb{R}^2$. Assume that $\sqrt{s}|_A$ belongs to $\mathscr{F}$. Let $\sigma > 0$, $p \in (0, +\infty]$, $\beta > (1/p - 1/2)_+$ be any numbers and $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $g \in \mathscr{B}^\beta(\mathbb{L}^p([0,1]))$, $\|g\|_\infty < \infty$ be any functions such that*

$$\sqrt{s(x,y)} = \phi\big(y - g(x)\big) \quad \textit{for all } x, y \in [0,1].$$

There exists two estimators $\hat{\phi} \geq 0$ and $\hat{g}$ such that the estimator $\hat{s}$ defined by

$$\hat{s}(x,y) = \big(\hat{\phi}\big(y - \hat{g}(x)\big)\big)^2 \mathbb{1}_{[0,1]^2}(x,y) \quad \textit{for all } x, y \in \mathbb{R}$$

satisfies

$$C\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s})\big] \leq C_1' \left(\frac{\log^2 n}{n}\right)^{2\beta(\sigma\wedge 1)/(2\beta(\sigma\wedge 1)+1)} + C_2' \left(\frac{\log n}{n}\right)^{2\sigma/(2\sigma+1)},$$

where $C > 0$ depends only on κ, p, σ, β, where C_1' depends only on $p, \beta, \sigma, |g|_{p,\beta}, \|g\|_\infty, |\phi|_{\infty,\sigma\wedge 1}$ and where C_2' depends only on $\sigma, \|g\|_\infty, |\phi|_{\infty,\sigma}$. Moreover, the construction of the estimators $\hat{g}$, $\hat{\phi}$ depends only on the data $X_0, \ldots, X_n$.

In particular, if ϕ is very smooth (says $\sigma \geq \beta \vee 1$), the rate of convergence corresponds to the rate of convergence for estimating g only (up to a logarithmic term).

It is interesting to compare the preceding rate to the one we would obtain under the pure smoothness assumption on $\sqrt{s}|_A$ but ignoring that $\sqrt{s}|_A$ belongs to $\mathscr{F}$. To do so, we need to specify the regularity of $\sqrt{s}|_A$, knowing that of ϕ and g. This is the purpose of the following lemma.

Lemma 4.1. *Let $\sigma, \beta > 0$, and let us define*

$$\theta(\beta,\sigma) = \begin{cases} \beta\sigma & \textit{if } \beta, \sigma \leq 1, \\ \beta \wedge \sigma & \textit{otherwise.} \end{cases}$$

Let $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $g \in \mathcal{H}^\beta([0,1])$. The function f defined by

$$f(x,y) = \phi\big(y - g(x)\big) \quad \textit{for all } x, y \in [0,1],$$

belongs to $\mathcal{H}^{(\theta(\beta,\sigma),\sigma)}([0,1]^2)$.

Moreover, for all $\sigma,\beta>0$, *there exist* $\phi\in\mathcal{H}^{\sigma}(\mathbb{R})$, $g\in\mathcal{H}^{\beta}([0,1])$ *such that the function* f *defined by*

$$f(x,y)=\phi\big(y-g(x)\big)\quad \text{for all } x,y\in[0,1],$$

belongs to $\mathcal{H}^{(a,b)}([0,1]^2)$ *if and only if* $a\le\theta(\beta,\sigma)$ *and* $b\le\sigma$.

This result says that if $\sqrt{s(x,y)}=\phi(y-g(x))$, with $\phi\in\mathcal{H}^{\sigma}(\mathbb{R})$, $g\in\mathcal{H}^{\beta}([0,1])$, then $\sqrt{s}$ is Hölderian with regularity $(\theta(\beta,\sigma),\sigma)$ on $[0,1]^2$, and this regularity cannot be improved in general except in some particular situations. Under such a smoothness assumption, the rate of estimation we would get is $(\log n/n)^{2\sigma\theta(\beta,\sigma)/(2\sigma\theta(\beta,\sigma)+\theta(\beta,\sigma)+\sigma)}$. This rate is always slower than the rate obtained under the structural assumption.

4.5. *ARCH model*

Throughout this section, we assume that $X_{n+1}=g_1(X_n)+g_2(X_n)\varepsilon_n$ where g_1,g_2 are unknown functions and where the ε_n's are unobserved identically distributed random variables. The previous model corresponded to $g_2=1$. The problem of the estimation of the mean and variance functions g_1 and g_2 was considered in several papers and we refer to Section 1.2 of Comte and Rozenholc [19] for bibliographical references.

For the sake of simplicity, one assumes that $\mathbb{X}=\mathbb{R}$ and $A=[0,1]^2$. If φ denotes the density of ε_0, the transition density s is of the form

$$s(x,y)=\big|g_2(x)\big|^{-1}\varphi\big[g_2^{-1}(x)\big(y-g_1(x)\big)\big]\quad \text{for all } x,y\in\mathbb{R}. \tag{4.2}$$

We consider thus the class

$$\begin{aligned}\mathcal{F}=\bigcup_{\sigma>0}\Big\{&f,\exists\phi\in\mathcal{H}^{\sigma}(\mathbb{R}),\exists v_1,v_2\in\mathcal{B}\big([0,1]\big),\|v_1\|_\infty<\infty,\|v_2\|_\infty<\infty,\\&\forall x,y\in[0,1],\ f(x,y)=\sqrt{\big|v_2(x)\big|}\phi\big(v_2(x)\big(y-v_1(x)\big)\big)\Big\}\end{aligned}$$

and apply Theorem 4.1 with a suitable collection $\mathbb{V}$ to obtain:

Corollary 4.3. *Suppose that Assumption 4.1 holds with* $\mathbb{X}=\mathbb{R}$, $A=[0,1]^2$ *and with* $\nu\otimes\mu$ *the Lebesgue measure on* $\mathbb{R}^2$. *Assume that* $\sqrt{s}|_A$ *belongs to* $\mathcal{F}$. *Let* $\sigma>0$, $\phi\in\mathcal{B}^{\sigma}(\mathbb{R})$ *and for all* $i\in\{1,2\}$, *let* $p_i\in(0,+\infty]$, $\beta_i>(1/p_i-1/2)_+$, $v_i\in\mathcal{B}^{\beta_i}(\mathbb{L}^{p_i}([0,1]))$, *with* $\|v_i\|_\infty<\infty$ *such that*

$$\sqrt{s(x,y)}=\sqrt{\big|v_2(x)\big|}\phi\big(v_2(x)\big(y-v_1(x)\big)\big)\quad \text{for all } x,y\in[0,1].$$

Let $p_3\in(0,+\infty]$ *and* $\beta_3>(1/p_3-1/2)_+$ *be any numbers such that* $v_3=\sqrt{|v_2|}\in\mathcal{B}^{\beta_3}(\mathbb{L}^{p_3}([0,1]))$. *There exists an estimator* $\hat{s}$ *such that*

$$C\mathbb{E}\big[H^2(s,\hat{s})\big]\le C_1'\left(\frac{\log^2 n}{n}\right)^{2\beta(\sigma\wedge1)/(2\beta(\sigma\wedge1)+1)}+C_2'\left(\frac{\log n}{n}\right)^{2\sigma/(2\sigma+1)},$$

where $\beta=\max(\beta_1,\beta_2,\beta_3)$. *The constant* $C>0$ *depends only on* $\kappa,\sigma,p_1,p_2,p_3,\beta_1,\beta_2,\beta_3,C_1'$ *depends only on* $\sigma,\|v_1\|_\infty,\|v_2\|_\infty,\|\varphi\|_\infty,|v_1|_{p_1,\beta_1},|v_2|_{p_2,\beta_2},|v_3|_{p_3,\beta_3},|\varphi|_{\infty,\sigma\wedge1}$ *and* C_2' *depends only on* $\sigma,\|v_2\|_\infty,|\varphi|_{\infty,\sigma}$. *Moreover, the construction of the estimator* $\hat{s}$ *depends only on the data* $X_0,\dots,X_n$.

If s is of the form (4.2) with φ, g_1, g_2 smooth, in the sense that $\phi=\sqrt{\varphi}\in\mathcal{H}^{\sigma}(\mathbb{R})$, $v_1=g_1\in\mathcal{B}^{\beta_1}(\mathbb{L}^{p_1}([0,1]))$, $\|v_1\|_\infty<\infty$, $v_2=g_2^{-1}\in\mathcal{B}^{\beta_2}(\mathbb{L}^{p_2}([0,1]))$, $\|v_2\|_\infty<\infty$ and $v_3=|g_2|^{-1/2}\in\mathcal{B}^{\beta_3}(\mathbb{L}^{p_3}([0,1]))$, then $\sqrt{s}|_A$ belongs to $\mathcal{F}$. If ϕ is sufficiently smooth ($\sigma\ge\beta_1\vee\beta_2\vee\beta_3\vee1$), the rate becomes

$$C''\mathbb{E}\big[H^2(s,\hat{s})\big]\le\max\left(\left(\frac{\log^2 n}{n}\right)^{2\beta_1/(2\beta_1+1)},\left(\frac{\log^2 n}{n}\right)^{2\beta_2/(2\beta_2+1)},\left(\frac{\log^2 n}{n}\right)^{2\beta_3/(2\beta_3+1)}\right).$$

Up to a logarithmic term, the first term corresponds to the bound we would get if we could estimate g_1 only. The two other terms correspond to the rate of estimation of g_2^{-1} and $|g_2|^{-1/2}$ respectively (up to a logarithmic term).

Note that if $\beta_2 \in (0,1)$, one can always choose $p_3 = 2p_2$ (with $p_3 = \infty$ if $p_2 = \infty$), $\beta_3 = \beta_2/2$, in which case the rate becomes

$$C''\mathbb{E}\big[H^2(s,\hat{s})\big] \leq \max\left(\left(\frac{\log^2 n}{n}\right)^{2\beta_1/(2\beta_1+1)}, \left(\frac{\log^2 n}{n}\right)^{\beta_2/(\beta_2+1)}\right).$$

In some situations however, β_3 can be taken larger than β_2.

As in the preceding section, we may use the lemma below to compare this rate with the one we would obtain under smoothness assumptions on $\sqrt{s}|_A$.

Lemma 4.2. *Let for all $\sigma, \beta_1, \beta_2 > 0$,*

$$\theta(\beta_1,\beta_2,\sigma) = \begin{cases} \big(2^{-1}(\beta_2 \wedge 1)\big) \wedge \sigma\beta_1 \wedge \sigma\beta_2 & \text{if } \sigma \leq 1 \text{ and } \beta_1 \wedge \beta_2 \leq 1, \\ \big(2^{-1}(\beta_2 \wedge 1)\big) \wedge \sigma \wedge \beta_1 & \text{otherwise.} \end{cases}$$

Let $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$. The function f defined by

$$f(x,y) = \sqrt{|v_2(x)|}\phi\big(v_2(x)\big(y - v_1(x)\big)\big) \quad \text{for all } x, y \in [0,1],$$

belongs to $\mathcal{H}^{(\theta(\beta_1,\beta_2,\sigma),\sigma)}([0,1]^2)$.

Moreover, there exist $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$ such that the function f defined by

$$f(x,y) = \sqrt{|v_2(x)|}\phi\big(v_2(x)\big(y - v_1(x)\big)\big) \quad \text{for all } x, y \in [0,1],$$

belongs to $\mathcal{H}^{(a,b)}([0,1]^2)$ if and only if $a \leq \theta(\beta_1,\beta_2,\sigma)$ and $b \leq \sigma$.

This proposition says that if $\sqrt{s(x,y)} = \sqrt{|v_2(x)|}\phi(v_2(x)(y - v_1(x)))$, with $\phi \in \mathcal{H}^\sigma(\mathbb{R})$, $v_1 \in \mathcal{H}^{\beta_1}([0,1])$, $v_2 \in \mathcal{H}^{\beta_2}([0,1])$, $\sqrt{s}|_A$ belongs to $\mathcal{H}^{(\theta(\beta_1,\beta_2,\sigma),\sigma)}([0,1]^2)$ and the regularity index of this space cannot be increased in general. By Corollary 4.1, we would get a rate of order $(\log n/n)^{2\theta(\beta_1,\beta_2,\sigma)\sigma/(2\theta(\beta_1,\beta_2,\sigma)\sigma+\theta(\beta_1,\beta_2,\sigma)+\sigma)}$, which is slower than the one given by Corollary 4.3.

Appendix A: Implementation of the first procedure

In this section, we explain how to construct in practice the estimator of the first procedure. This will lead to the proposition below.

Proposition A.1. *For all $L > 0$, $\ell \in \mathbb{N}^\star$, the estimator $\hat{s} = \hat{s}(L,\ell)$ of Section 2.3 can be built in less than $C(n\ell d + \ell 4^{(\ell+1)d})$ operations where C is an universal constant.*

We set for all $K \in \bigcup_{m\in\mathcal{M}_\ell} m$,

$$\hat{s}_K = \frac{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i,x)\,\mathrm{d}\mu(x)}\mathbb{1}_K$$

for all $K' \in \bigcup_{m\in\mathcal{M}_\ell} m$,

$$F_K\big(K'\big) = \alpha H^2(\hat{s}_K\mathbb{1}_{K'}, \hat{s}_{K'}\mathbb{1}_K) + T(\hat{s}_K\mathbb{1}_{K'}, \hat{s}_{K'}\mathbb{1}_K)$$

and for all $m' \in \mathcal{M}_\ell$,

$$\gamma_K\bigl(m'\bigr) = \Biggl(\sum_{K' \in m'} F_K\bigl(K'\bigr)\Biggr) - \operatorname{pen}\bigl(m' \vee K\bigr).$$

We shall find for each cube $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, a partition $m'_K \in \mathcal{M}_\ell$ such that

$$\gamma_K\bigl(m'_K\bigr) = \sup_{m'\in\mathcal{M}_\ell} \gamma_K\bigl(m'\bigr). \tag{A.1}$$

We shall compute then

$$\min_{m\in\mathcal{M}_\ell} \gamma(m) = \min_{m\in\mathcal{M}_\ell} \Biggl\{\Biggl(\sum_{K\in m} \gamma_K\bigl(m'_K\bigr)\Biggr) + 2\operatorname{pen}(m)\Biggr\}. \tag{A.2}$$

We shall find m'_K by using a slight adaptation of the procedure of Blanchard, Schäfer and Rozenholc [16]. Computing (A.2) is similar. The algorithm we propose is based on the one-to-one correspondence between $\mathcal{M}_\ell$ and the set $\mathcal{T}_\ell$ of 4^d-ary trees with depth smaller than ℓ.

Lemma A.1. *There exists ψ_ℓ a one-to-one map between $\mathcal{M}_\ell$ and $\mathcal{T}_\ell$ such that for all $m \in \mathcal{M}_\ell$, $\psi_\ell(m)$ is a tree whose leaves correspond to the elements of the partition m.*

The construction of this map may for instance be deduced from Section 3.2.4 of Baraud and Birgé [6].

We need to introduce some notations. For each tree $T \in \mathcal{T}_\ell$ and bin K'' of T, we denote by $T(K'')$ the subtree of T rooted in K''. The set of leaves of $T(K'')$ is denoted by $\mathcal{L}(T(K''))$. We set $R(K'')$ the tree reduced to its root K'' (i.e, $\mathcal{L}(R(K'')) = \{K''\}$). For all cube $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, we set

$$\mathcal{L}\bigl(T\bigl(K''\bigr)\bigr) \vee K = \bigl\{K' \cap K, K' \in \mathcal{L}\bigl(T\bigl(K''\bigr)\bigr), K' \cap K \neq \varnothing\bigr\}$$

and we define the function $\mathcal{E}$ by

$$\mathcal{E}\bigl(T\bigl(K''\bigr)\bigr) = -\bigl|\mathcal{L}\bigl(T\bigl(K''\bigr)\bigr) \vee K\bigr| + \sum_{K'\in\mathcal{L}(T(K''))} F_K\bigl(K'\bigr).$$

The key point is that computing (A.1) amounts to finding $T^\star$ such that

$$\mathcal{E}\bigl(T^\star\bigl([0,1]^{2d}\bigr)\bigr) = \sup_{T\in\mathcal{T}_\ell} \mathcal{E}\bigl(T\bigl([0,1]^{2d}\bigr)\bigr)$$

since $m'_K = \psi_\ell^{-1}(T^\star)$.

We now take advantage of the additivity of the function $\mathcal{E}$: if $T(K'')$ is not reduced to its root, and if $K''_1, \ldots, K''_{4^d}$ are the cubes of $\bigcup_{m\in\mathcal{M}_\ell} m$ such that $K'' = \bigcup_{i=1}^{4^d} K''_i$, then,

$$\mathcal{E}\bigl(T\bigl(K''\bigr)\bigr) = \sum_{i=1}^{4^d} \mathcal{E}\bigl(T\bigl(K''_i\bigr)\bigr). \tag{A.3}$$

For all cube $K'' \in \bigcup_{m\in\mathcal{M}_\ell} m$, let $T^\star(K'')$ be a tree (rooted in K'') such that

$$\mathcal{E}\bigl(T^\star\bigl(K''\bigr)\bigr) = \sup_{T\in\mathcal{T}_\ell, T\ni K''} \mathcal{E}\bigl(T\bigl(K''\bigr)\bigr).$$

Remark that if $K'' \cap K = \varnothing$, this supremum is equal to 0, in which case $T^\star(K'')$ will always stand for $R(K'')$. In general, we deduce from (A.3) that

$$\mathcal{E}\big(T^\star\big(K''\big)\big) = \max\left(\mathcal{E}\big(R\big(K''\big)\big), \sum_{i=1}^{4^d} \mathcal{E}\big(T^\star\big(K_i''\big)\big)\right). \tag{A.4}$$

Calculating (A.1) can thus be completed in that way: we start with the sets $K'' \in \bigcup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_{\ell-1}} m$ with $K'' \cap K \neq \varnothing$ for which the optimal local trees are reduced to their roots. By using relation (A.4) we find the optimal local trees $T^\star(K'')$ when $K'' \in \bigcup_{\mathcal{M}_{\ell-1}\setminus\mathcal{M}_\ell} m$, $K'' \cap K \neq \varnothing$. Proceeding recursively like this yields the optimal tree $T^\star = T^\star([0,1]^{2d})$.

Appendix B: Proofs

B.1. *Proof of Proposition 2.1*

Let us introduce the piecewise constant function

$$\bar{s}_m = \sum_{K\in m} \frac{\sum_{i=0}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]}{\sum_{i=0}^{n-1} \int_{\mathbb{X}} \mathbb{1}_K(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_K. \tag{B.1}$$

By using the triangular inequality we can decompose the risk of $\hat{s}_m$ as follows:

$$\mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}_m)\big] \leq \left(1 + \frac{2+\log 2}{2}\right)\mathbb{E}\big[H^2(s\mathbb{1}_A, \bar{s}_m)\big] + \left(1 + \frac{2}{2+\log 2}\right)\mathbb{E}\big[H^2(\bar{s}_m, \hat{s}_m)\big].$$

The first term can be bounded from above by $(4+\log 2)\mathbb{E}[H^2(s\mathbb{1}_A, V_m)]$ thanks to Lemma 2 of Baraud and Birgé [6]. For the second term, we begin to define for $K \in m$ the random variable

$$B_K = \left(\sqrt{\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{n-1} \mathbb{E}\big[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\big]}\right)^2.$$

Since $2nH^2(\hat{s}_m, \bar{s}_m) = \sum_{K\in m} B_K$, we shall bound from above the terms $\mathbb{E}[B_K]$. For this purpose, we introduce the stopping time

$$T = \inf\left\{i \geq 0, \mathbb{E}\big[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\big] \geq \frac{1}{2n}\right\} \wedge (n-1)$$

with respect to the filtration $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ generated by the random variables $X_0, \ldots, X_n$. We set $\varepsilon = 1 + \log 2 + 2\log n$ and use the algebraic inequality

$$(\sqrt{a+b} - \sqrt{c+d})^2 \leq (1+\varepsilon)(\sqrt{a} - \sqrt{c})^2 + \big(1+\varepsilon^{-1}\big)(\sqrt{b} - \sqrt{d})^2$$

to decompose $\mathbb{E}[B_K]$:

$$\begin{aligned}\mathbb{E}[B_K] \leq (1+\varepsilon)\mathbb{E}\left[\left(\sqrt{\sum_{i=0}^{T-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=0}^{T-1} \mathbb{E}\big[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\big]}\right)^2\right] \\ + \big(1+\varepsilon^{-1}\big)\mathbb{E}\left[\left(\sqrt{\sum_{i=T}^{n-1} \mathbb{1}_K(X_i, X_{i+1})} - \sqrt{\sum_{i=T}^{n-1} \mathbb{E}\big[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i\big]}\right)^2\right].\end{aligned}$$

By using $(\sqrt{a}-\sqrt{b})^2 \le (a-b)^2/b$,

$$\mathbb{E}[B_K] \le 2(1+\varepsilon)\mathbb{E}\left[\sum_{i=0}^{T-1}\mathbb{E}\big[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i\big]\right] + \big(1+\varepsilon^{-1}\big)\mathbb{E}\left[\frac{(\sum_{i=T}^{n-1}(\mathbb{1}_K(X_i,X_{i+1})-\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]))^2}{\sum_{i=T}^{n-1}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}\right]. \tag{B.2}$$

Yet,

$$\mathbb{E}\left[\sum_{i=0}^{T-1}\mathbb{E}\big[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i\big]\right] \le 1/2,$$

and we control the second term of the right-hand side of inequality (B.2), thanks to the claims below.

Claim B.1. *For all $K \in m$, $j \in \{0,\dots,n\}$, and $\mathcal{A}' \in \mathcal{F}_j = \sigma(X_0,\dots,X_j)$,*

$$\mathbb{E}\left[\frac{(\sum_{i=j}^{n-1}(\mathbb{1}_K(X_i,X_{i+1})-\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]))^2}{\sum_{i=j}^{n-1}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}\mathbb{1}_{\mathcal{A}'}\right] \le \sum_{k=j}^{n-1}\mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}_K(X_k,X_{k+1}) \mid X_k]}{\sum_{i=j}^{k}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}\mathbb{1}_{\mathcal{A}'}\right].$$

Proof. Let us define the random variables

$$Y_{n-1}(K) = \sum_{i=j}^{n-1}\big(\mathbb{1}_K(X_i,X_{i+1}) - \mathbb{E}\big[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i\big]\big) \quad \text{and} \quad Z_n(K) = \frac{Y_{n-1}^2(K)}{\sum_{i=j}^{n-1}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}.$$

We have

$$\mathbb{E}\big[Z_{n+1}(K) \mid \mathcal{F}_n\big] = \frac{\mathbb{E}([Y_{n-1}(K) + (\mathbb{1}_K(X_n,X_{n+1}) - \mathbb{E}[\mathbb{1}_K(X_n,X_{n+1}) \mid X_n])]^2 \mid \mathcal{F}_n)}{\sum_{i=j}^{n}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]} = \frac{Y_{n-1}^2(K) + \operatorname{var}(\mathbb{1}_K(X_n,X_{n+1}) \mid X_n)}{\sum_{i=j}^{n}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}.$$

Hence,

$$\mathbb{E}\big[Z_{n+1}(K) \mid \mathcal{F}_n\big] \le Z_n(K) + \frac{\mathbb{E}[\mathbb{1}_K(X_n,X_{n+1}) \mid X_n]}{\sum_{i=j}^{n}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}$$

and, since $\mathcal{A}'$ is also $\mathcal{F}_n$-measurable,

$$\mathbb{E}\big[Z_{n+1}(K)\mathbb{1}_{\mathcal{A}'}\big] \le \mathbb{E}\big[Z_n(K)\mathbb{1}_{\mathcal{A}'}\big] + \mathbb{E}\left[\frac{\mathbb{E}[\mathbb{1}_K(X_n,X_{n+1}) \mid X_n]}{\sum_{i=j}^{n}\mathbb{E}[\mathbb{1}_K(X_i,X_{i+1}) \mid X_i]}\mathbb{1}_{\mathcal{A}'}\right].$$

The result ensues from induction. □

Claim B.2. *For all sequence $(u_n)_{n\ge 0}$ in $[0,1]$, and $j \ge 0$ such that $u_j \ne 0$,*

$$\sum_{k=j}^{n-1}\frac{u_k}{\sum_{i=j}^{k}u_i} \le 1 + \log n - \log u_j.$$

Proof. Let f be any non-negative continuous function such that $u_k = \int_k^{k+1} f(t)\,\mathrm{d}t$ whatever $k \in \mathbb{N}$. Let F be the primitive of f such that $F(j) = 0$. Then,

$$\begin{aligned}\sum_{k=j}^{n-1} \frac{u_k}{\sum_{i=j}^{k} u_i} &\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(k+1)}\,\mathrm{d}t \\ &\leq 1 + \sum_{k=j+1}^{n-1} \int_k^{k+1} \frac{f(t)}{F(t)}\,\mathrm{d}t \\ &\leq 1 + \log F(n) - \log F(j+1) \\ &\leq 1 + \log\left(\sum_{k=j}^{n-1} u_k\right) - \log u_j.\end{aligned}$$

□

By using Claim B.1 with $\mathcal{A}' = [T = j]$,

$$\begin{aligned}&\mathbb{E}\left[\frac{(\sum_{i=T}^{n-1}(\mathbb{1}_K(X_i, X_{i+1}) - \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]))^2}{\sum_{i=T}^{n-1} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]}\right] \\ &\quad\leq \sum_{j=0}^{n-2} \mathbb{E}\left(\sum_{k=j}^{n-1} \frac{\mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]}{\sum_{i=j}^{k} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{T=j}\right) \\ &\qquad+ \mathbb{E}\left[\frac{(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}])^2}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1}\right].\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{E}\left[\frac{(\mathbb{1}_K(X_{n-1}, X_n) - \mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}])^2}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1}\right] &= \mathbb{E}\left[\frac{\mathrm{var}(\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1})}{\mathbb{E}[\mathbb{1}_K(X_{n-1}, X_n) \mid X_{n-1}]} \mathbb{1}_{T=n-1}\right] \\ &\leq \mathbb{P}(T = n-1).\end{aligned}$$

We then use Claim B.2 with $u_k = \mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]$ to derive

$$\begin{aligned}\sum_{j=0}^{n-2} \mathbb{E}\left(\sum_{k=j}^{n-1} \frac{\mathbb{E}[\mathbb{1}_K(X_k, X_{k+1}) \mid X_k]}{\sum_{i=j}^{k} \mathbb{E}[\mathbb{1}_K(X_i, X_{i+1}) \mid X_i]} \mathbb{1}_{T=j}\right) &\leq \sum_{j=0}^{n-2} \mathbb{E}\big[(1 + \log 2 + 2\log n)\mathbb{1}_{T=j}\big] \\ &\leq (1 + \log 2 + 2\log n)\mathbb{P}(T \neq n-1).\end{aligned}$$

Finally, $\mathbb{E}[B_K] \leq 4 + 2\log 2 + 4\log n$ and hence

$$\mathbb{E}\big[H^2(\bar{s}_m, \hat{s}_m)\big] \leq \frac{2 + \log 2 + 2\log n}{n} |m|,$$

which concludes the proof.

B.2. Proof of Theorem 2.1

When $\ell \leq n$, the result ensues from the following theorem whose proof is delayed to Section B.3. In the theorem below, the constant $L_0 = 90$ can easily be improved but it seems to be difficult to obtain the value $L_0 = 0.03$ used in practice.

Theorem B.1. *For all $L \geq 90$ and $1 \leq \ell \leq n$, the estimator $\hat{s} = \hat{s}(L, \ell)$ satisfies*

$$\forall \xi > 0, \quad \mathbb{P}\Big[C H^2(s\mathbb{1}_A, \hat{s}) \geq \inf_{m \in \mathcal{M}_\ell} \big(H^2(s\mathbb{1}_A, \hat{s}_m) + \mathrm{pen}(m) \big) + \xi \Big] \leq 3\mathrm{e}^{-n\xi}$$

where C is an universal positive constant.

By integrating the inequality above, there exists $C' > 0$ such that

$$C' \mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}) \big] \leq \inf_{m \in \mathcal{M}_\ell} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}_m) \big] + \mathrm{pen}(m) \big\}$$

and the conclusion follows from Proposition 2.1.

When ℓ is larger than n, we use the lemma below whose proof is postponed to Section B.3.3.

Lemma B.1. *For all $L \geq 15$ and $\ell \geq n+1$, $\hat{s}(L, \ell) = \hat{s}(L, n)$ and $\hat{s}(L, \infty) = \hat{s}(L, n)$.*

For $L \geq 90$, if $\ell \geq n+1$ or $\ell = \infty$, we have thus

$$C' \mathbb{E}\big[H^2(s\mathbb{1}_A, \hat{s}) \big] \leq \inf_{m \in \mathcal{M}_n} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m) \big] + \mathrm{pen}(m) \big\}.$$

Let $m^\star \in \mathcal{M}_\ell$ such that

$$2 \inf_{m \in \mathcal{M}_\ell} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m) \big] + \mathrm{pen}(m) \big\} \geq \mathbb{E}\big[H^2(s\mathbb{1}_A, V_{m^\star}) \big] + \mathrm{pen}\big(m^\star\big).$$

Since $\mathbb{E}[H^2(s\mathbb{1}_A, V_m)] \leq \mathbb{E}[H^2(s\mathbb{1}_A, 0)] \leq 1/2$ and $\{[0,1]^{2d}\} \in \mathcal{M}_\ell$,

$$\inf_{m \in \mathcal{M}_\ell} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m) \big] + \mathrm{pen}(m) \big\} \leq \frac{1}{2} + L \frac{\log n}{n}.$$

Consequently, $L|m^\star| \log(n)/n \leq 1 + 2L \log(n)/n$ and thus $|m^\star| \leq 2 + n/(L \log n) \leq n$. Remark now that the cardinality of a partition $m \in \mathcal{M}_\ell \setminus \mathcal{M}_n$ can be lower bounded by

$$|m| \geq 4^d + \big(4^d - 1\big) n \geq n + 1.$$

Consequently, $m^\star \in \mathcal{M}_n$ and hence,

$$\inf_{m \in \mathcal{M}_n} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m) \big] + \mathrm{pen}(m) \big\} \leq 2 \inf_{m \in \mathcal{M}_\ell} \big\{ \mathbb{E}\big[H^2(s\mathbb{1}_A, V_m) \big] + \mathrm{pen}(m) \big\}$$

which completes the proof.

B.3. Proof of Theorem B.1

The proof of this theorem requires the two following lemmas whose proofs are postponed to Sections B.3.1 and B.3.2.

Lemma B.2. *For all partition $m \in \mathcal{M}_\ell$,*

$$\gamma_1(m) = \sup_{m' \in \mathcal{M}_\ell} \big\{ \alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \mathrm{pen}\big(m'\big) \big\} + \mathrm{pen}(m)$$

satisfies $\gamma_1(m) \leq \gamma(m)$.

For all $m \in \mathcal{M}_\ell$, there exists a deterministic set S_m such that $\hat{s}_m \in S_m$ and such that

$$\gamma_2(m) = \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \big\{ \alpha H^2\big(\hat{s}_m, f'\big) + T\big(\hat{s}_m, f'\big) - \mathrm{pen}\big(m'\big) \big\} + 2\,\mathrm{pen}(m)$$

satisfies $\gamma(m) \leq \gamma_2(m)$.

Lemma B.3. *Set $\varepsilon = (2+3\sqrt{2})/8$. Under assumptions of Theorem B.1, for all $\xi > 0$, there exists an event Ω_ξ such that $\mathbb{P}(\Omega_\xi) \geq 1 - 3\mathrm{e}^{-n\xi}$ and on which,*

for all partition $m \in \mathcal{M}_\ell$,

$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{(1-\varepsilon)H^2\left(s\mathbb{1}_A, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}\left(m'\right)\right\} \leq (1+\varepsilon)H^2(s\mathbb{1}_A, \hat{s}_m) + \mathrm{pen}(m) + 22\xi, \tag{B.3}$$

where $S_{m'}$ is defined in Lemma B.2.

Proof of Theorem B.1. On Ω_ξ, for all $m \in \mathcal{M}_\ell$,

$$\sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{(1-\varepsilon)H^2\left(s\mathbb{1}_A, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}\left(m'\right)\right\} \leq (1+\varepsilon)H^2(s\mathbb{1}_A, \hat{s}_m) + \mathrm{pen}(m) + 22\xi.$$

If $T(\hat{s}_m, \hat{s}_{\hat{m}}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \geq 0$,

$$\begin{aligned}\alpha H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) &\leq (1-\varepsilon)H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) + T(\hat{s}_m, \hat{s}_{\hat{m}}) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m) \\ &\leq (1+\varepsilon)H^2(s\mathbb{1}_A, \hat{s}_m) + 2\,\mathrm{pen}(m) + 22\xi\end{aligned}$$

since $\alpha \leq 1-\varepsilon$ and since $\hat{s}_{\hat{m}}$ belongs to $\bigcup_{m' \in \mathcal{M}_\ell} S_{m'}$.

If $T(\hat{s}_m, \hat{s}_{\hat{m}}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) < 0$,

$$\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) \leq \alpha H^2(\hat{s}_{\hat{m}}, \hat{s}_m) + T(\hat{s}_{\hat{m}}, \hat{s}_m) - \mathrm{pen}(m) + \mathrm{pen}(\hat{m})$$

since $T(\hat{s}_{\hat{m}}, \hat{s}_m) = -T(\hat{s}_m, \hat{s}_{\hat{m}})$. Hence,

$$\begin{aligned}\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) &\leq \sup_{m' \in \mathcal{M}_\ell} \left\{\alpha H^2(\hat{s}_{\hat{m}}, \hat{s}_{m'}) + T(\hat{s}_{\hat{m}}, \hat{s}_{m'}) - \mathrm{pen}\left(m'\right)\right\} + \mathrm{pen}(\hat{m}) \\ &\leq \gamma_1(\hat{m}).\end{aligned}$$

By using Lemma B.2,

$$\gamma_1(\hat{m}) \leq \gamma(m) + \frac{1}{n} \leq \gamma_2(m) + \frac{1}{n},$$

and thus

$$\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) \leq \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{\alpha H^2\left(\hat{s}_m, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}\left(m'\right)\right\} + 2\,\mathrm{pen}(m) + \frac{1}{n}.$$

With $\upsilon = (1-\varepsilon)/\alpha - 1 > 0$,

$$\begin{aligned}\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}}) &\leq \left(1+\upsilon^{-1}\right)H^2(\hat{s}_m, s\mathbb{1}_A) \\ &\quad + \sup_{\substack{m' \in \mathcal{M}_\ell \\ f' \in S_{m'}}} \left\{(1-\varepsilon)H^2\left(s\mathbb{1}_A, f'\right) + T\left(\hat{s}_m, f'\right) - \mathrm{pen}\left(m'\right)\right\} + 2\,\mathrm{pen}(m) + \frac{1}{n} \\ &\leq \left(1+\upsilon^{-1}\right)H^2(\hat{s}_m, s\mathbb{1}_A) + \left[(1+\varepsilon)H^2(s\mathbb{1}_A, \hat{s}_m) + \mathrm{pen}(m) + 22\xi\right] + 2\,\mathrm{pen}(m) + \frac{1}{n} \\ &\leq \left(2+\varepsilon+\upsilon^{-1}\right)H^2(\hat{s}_m, s\mathbb{1}_A) + 3\,\mathrm{pen}(m) + 22\xi + \frac{1}{n}.\end{aligned}$$

This leads to

$$\alpha H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \le 2\alpha H^2(s\mathbb{1}_A, \hat{s}_m) + 2\alpha H^2(\hat{s}_m, \hat{s}_{\hat{m}})$$
$$\le 2\big(2+\alpha+\varepsilon+\upsilon^{-1}\big)H^2(\hat{s}_m, s\mathbb{1}_A) + 6\,\mathrm{pen}(m) + 44\xi + \frac{2}{n}.$$

Finally, we have proved that there exists $C>0$, such that, with probability larger than $1-3\mathrm{e}^{-n\xi}$, for all $m\in\mathcal{M}_\ell$,

$$C H^2(s\mathbb{1}_A, \hat{s}_{\hat{m}}) \le H^2(\hat{s}_m, s\mathbb{1}_A) + \mathrm{pen}(m) + \xi.$$

This concludes the proof. □

B.3.1. *Proof of Lemma* B.2

For each partition $m\in\mathcal{M}_\ell$, we define the random set $\hat{S}_m$ of functions as follows. A function $\hat{f}$ is said to belong to $\hat{S}_m$ if for each cube $K\in m$, there exists a partition $m_K\in\mathcal{M}_\ell$ such that $\hat{f}=\hat{s}_{m_K}$ on K. In other words,

$$\hat{S}_m = \Big\{ \sum_{K\in m} \hat{s}_{m_K}\mathbb{1}_K, \forall K\in m, m_K\in\mathcal{M}_\ell \Big\}.$$

For all function $\hat{f}\in\hat{S}_m$ and $K\in m$, let $m_K(\hat{f})\in\mathcal{M}_\ell$ be any partition such that

$$\big|m_K(\hat{f})\vee K\big| = \inf\big\{\big|m'\vee K\big|, m'\in\mathcal{M}_\ell, \hat{f}\mathbb{1}_K = \hat{s}_{m'}\mathbb{1}_K\big\}.$$

The function $\hat{f}\in\hat{S}_m$ is piecewise constant on the elements of the partition

$$m(\hat{f}) = \bigcup_{K\in m}\big(m_K(\hat{f})\vee K\big).$$

The whole point is that

$$\gamma(m) = \Big\{ \sum_{K\in m} \sup_{m'_K\in\mathcal{M}_\ell} \big[\alpha H^2(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'_K}\mathbb{1}_K) + T(\hat{s}_m\mathbb{1}_K, \hat{s}_{m'_K}\mathbb{1}_K) - \mathrm{pen}\big(m'_K\vee K\big)\big]\Big\} + 2\,\mathrm{pen}(m)$$
$$= \sup_{\hat{f}\in\hat{S}_m}\Big\{ \sum_{K\in m}\big[\alpha H^2(\hat{s}_m\mathbb{1}_K, \hat{f}\mathbb{1}_K) + T(\hat{s}_m\mathbb{1}_K, \hat{f}\mathbb{1}_K) - \mathrm{pen}\big(m_K(\hat{f})\vee K\big)\big]\Big\} + 2\,\mathrm{pen}(m).$$

Since $|m(\hat{f})| = \sum_{K\in m}|m_K(\hat{f})\vee K|$,

$$\gamma(m) = \sup_{\hat{f}\in\hat{S}_m}\big\{\alpha H^2(\hat{s}_m, \hat{f}) + T(\hat{s}_m, \hat{f}) - \mathrm{pen}\big(m(\hat{f})\big)\big\} + 2\,\mathrm{pen}(m). \tag{B.4}$$

We can now prove the first part of the lemma. For all partitions $m, m'\in\mathcal{M}_\ell$, the estimator $\hat{s}_{m'}$ belongs to $\hat{S}_m$. Hence,

$$\gamma(m) \ge \sup_{m'\in\mathcal{M}_\ell}\big\{\alpha H^2(\hat{s}_m, \hat{s}_{m'}) + T(\hat{s}_m, \hat{s}_{m'}) - \mathrm{pen}\big(m(\hat{s}_{m'})\big)\big\} + 2\,\mathrm{pen}(m).$$

Since $|m_K(\hat{s}_{m'})\vee K| \le |m'\vee K|$, $|m(\hat{s}_{m'})| \le \sum_{K\in m}|m'\vee K|$ and

$$\sum_{K\in m}\big|m'\vee K\big| = \big|\big\{K\cap K', \big(K, K'\big)\in m\times m', K\cap K'\neq\varnothing\big\}\big|.$$

Remark that either $K\cap K' = K$ or $K\cap K' = K'$ since K, K' are non-disjoint cubes (see Fig. 1). Hence, $|m(\hat{s}_{m'})| \le |m| + |m'|$, and $\gamma_1(m)\le\gamma(m)$ as wished.

To prove the second part of the lemma, we must define the set S_m appearing in the definition of γ_2.

Definition B.1. *Let, for all $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, $K_0, \dots, K_l$ be the cubes of $\bigcup_{m\in\mathcal{M}_\ell} m$ such that $K \subset K_i$ for all $i \in \{0, \dots, l\}$. For all $i \in \{0, \dots, l\}$, let I_i and J_i be the subsets of $[0,1]^d$ such that $K_i = I_i \times J_i$. Set*

$$S_K = \bigcup_{i=0}^{l} \left\{ \frac{a}{b\mu(J_i)} \mathbb{1}_K, a \in \{0, \dots, n\}, b \in \{1, \dots, n\} \right\}$$

with the convention $a/0 = 0$ whatever $a \in \{0, \dots, n\}$. We define S_m by

$$S_m = \left\{ \sum_{K\in m} f_K, f_K \in S_K \right\}.$$

Let us now make the link between S_m and $\hat{S}_m$. We shall show that a function $\hat{f} \in \hat{S}_m$ belongs to $S_{m(\hat{f})}$. For this purpose, let $K' \in m(\hat{f})$. By definition of $m(\hat{f})$, there exist $K \in m$, $K'' \in m_K(\hat{f})$ such that $K' = K'' \cap K$. We have $\hat{f}\mathbb{1}_K = \hat{s}_{m_K(\hat{f})}\mathbb{1}_K$ and

$$\hat{s}_{m_K(\hat{f})}\mathbb{1}_{K''} = \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K''}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_{K''}(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_{K''}.$$

Consequently,

$$\begin{aligned}
\hat{f}\mathbb{1}_{K'} &= (\hat{f}\mathbb{1}_K)\mathbb{1}_{K''} \\
&= (\hat{s}_{m_K(\hat{f})}\mathbb{1}_K)\mathbb{1}_{K''} = (\hat{s}_{m_K(\hat{f})}\mathbb{1}_{K''})\mathbb{1}_K \\
&= \frac{\sum_{i=0}^{n-1} \mathbb{1}_{K''}(X_i, X_{i+1})}{\sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_{K''}(X_i, x)\,\mathrm{d}\mu(x)} \mathbb{1}_{K'}.
\end{aligned}$$

This implies that $\hat{f}\mathbb{1}_{K'} \in S_{K'}$ and thus $\hat{f} = \sum_{K'\in m(\hat{f})} \hat{f}\mathbb{1}_{K'}$ belongs to $S_{m(\hat{f})}$.

The inequality $\gamma_2 \geq \gamma$ ensues from (B.4), $\hat{S}_m \subset \bigcup_{m'\in\mathcal{M}_\ell} S_{m'}$ and

$$\gamma_2(m) = \sup_{f'\in\bigcup_{m'\in\mathcal{M}_\ell} S_{m'}} \left\{ \alpha H^2\big(\hat{s}_m, f'\big) + T\big(\hat{s}_m, f'\big) - \Big(\inf_{\substack{m''\in\mathcal{M}_\ell \\ S_{m''}\ni f'}} \mathrm{pen}\big(m''\big) \Big) \right\} + 2\,\mathrm{pen}(m).$$

B.3.2. *Proof of Lemma* B.3
We start with the claim below.

Claim B.3. *Let ψ be the bounded function defined on $[0, +\infty)^2$ by*

$$\psi(x, y) = \frac{1}{\sqrt{2}} \frac{\sqrt{y} - \sqrt{x}}{\sqrt{x + y}} \quad \textit{for all } x, y \in [0, +\infty)$$

with the convention $0/0 = 0$.

Let, for all $f, f' \in \mathbb{L}^1_+(\mathbb{X}^2, M)$, with support included in A, $Z(f, f')$ be the random variable defined by

$$Z\big(f, f'\big) = \frac{1}{n} \sum_{i=0}^{n-1} \left(\psi\big(f(X_i, X_{i+1}), f'(X_i, X_{i+1})\big) - \int_{\mathbb{X}} \psi\big(f(X_i, y), f'(X_i, y)\big)(s\mathbb{1}_A)(X_i, y)\,\mathrm{d}\mu(y) \right).$$

Then,

$$\left(1 - \frac{1}{\sqrt{2}}\right) H^2\big(s\mathbb{1}_A, f'\big) + T\big(f, f'\big) \leq \left(1 + \frac{1}{\sqrt{2}}\right) H^2(s\mathbb{1}_A, f) + Z\big(f, f'\big) \tag{B.5}$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{X}}\psi^2\big(f(X_i,y),f'(X_i,y)\big)(s\mathbb{1}_A)(X_i,y)\,\mathrm{d}\mu(y)\le 3\big(H^2(s\mathbb{1}_A,f)+H^2\big(s\mathbb{1}_A,f'\big)\big). \tag{B.6}$$

Proof. For all $i\in\{0,\dots,n-1\}$, let

$$\begin{aligned}
T_i\big(f,f'\big)&=\frac{1}{2\sqrt{2}}\int_{\mathbb{X}}\sqrt{f(X_i,y)+f'(X_i,y)}\big(\sqrt{f'(X_i,y)}-\sqrt{f(X_i,y)}\big)\,\mathrm{d}\mu(y)\\
&\quad+\frac{1}{\sqrt{2}}\frac{\sqrt{f'(X_i,X_{i+1})}-\sqrt{f(X_i,X_{i+1})}}{\sqrt{f(X_i,X_{i+1})+f'(X_i,X_{i+1})}}+\frac{1}{2}\int_{\mathbb{X}}\big(f(X_i,y)-f'(X_i,y)\big)\,\mathrm{d}\mu(y),\\
Z_i\big(f,f'\big)&=\psi\big(f(X_i,X_{i+1}),f'(X_i,X_{i+1})\big)-\mathbb{E}\big[\psi\big(f(X_i,X_{i+1}),f'(X_i,X_{i+1})\big)\mid X_i\big],\\
H_i^2\big(f,f'\big)&=\frac{1}{2}\int_{\mathbb{X}}\big(\sqrt{f'(X_i,y)}-\sqrt{f(X_i,y)}\big)^2\,\mathrm{d}\mu(y),\\
\rho_i\big(f,f'\big)&=\int_{\mathbb{X}}\sqrt{f'(X_i,y)f(X_i,y)}\,\mathrm{d}\mu(y).
\end{aligned}$$

Since

$$T\big(f,f'\big)=\frac{1}{n}\sum_{i=0}^{n-1}T_i\big(f,f'\big),\qquad Z\big(f,f'\big)=\frac{1}{n}\sum_{i=0}^{n-1}Z_i\big(f,f'\big)\quad\text{and}\quad H^2\big(f,f'\big)=\frac{1}{n}\sum_{i=0}^{n-1}H_i^2\big(f,f'\big)$$

it is sufficient to prove that for all $i\in\{0,\dots,n-1\}$,

$$\Big(1-\frac{1}{\sqrt{2}}\Big)H_i^2\big(s\mathbb{1}_A,f'\big)+T_i\big(f,f'\big)\le\Big(1+\frac{1}{\sqrt{2}}\Big)H_i^2(s\mathbb{1}_A,f)+Z_i\big(f,f'\big) \tag{B.7}$$

and

$$\int_{\mathbb{X}}\psi^2\big(f(X_i,y),f'(X_i,y)\big)(s\mathbb{1}_A)(X_i,y)\,\mathrm{d}\mu(y)\le 3\big(H_i^2(s\mathbb{1}_A,f)+H_i^2\big(s\mathbb{1}_A,f'\big)\big). \tag{B.8}$$

These inequalities are established by using similar arguments than those developed in the proofs of Propositions 2 and 3 of Baraud [5]. Let us make them explicit. We set

$$\rho_i\big(\zeta,f,f'\big)=\frac{1}{2}\Big[\rho_i\Big(f',\frac{f+f'}{2}\Big)+\int_{\mathbb{X}}\sqrt{\frac{2f'(X_i,y)}{f(X_i,y)+f'(X_i,y)}}\,\mathrm{d}\zeta(y)\Big]\quad\text{for all measure }\zeta,$$

where the convention $0/0$ is in use. Let ζ_i be the random measure defined by $\mathrm{d}\zeta_i/\mathrm{d}\mu=(s\mathbb{1}_A)(X_i,\cdot)$. Then,

$$\mathbb{E}\big[T_i\big(f,f'\big)\mid X_i\big]=\Big[\rho_i\big(\zeta_i,f,f'\big)-\frac{1}{2}\int_{\mathbb{X}}f'(X_i,y)\,\mathrm{d}\mu(y)\Big]-\Big[\rho_i\big(\zeta_i,f',f\big)-\frac{1}{2}\int_{\mathbb{X}}f(X_i,y)\,\mathrm{d}\mu(y)\Big].$$

We deduce from relation (6) of Baraud [5],

$$\begin{aligned}
0\le\rho_i\big(\zeta_i,f,f'\big)-\rho_i\big(s\mathbb{1}_A,f'\big)&\le\frac{1}{\sqrt{2}}\big[H_i^2(s\mathbb{1}_A,f)+H_i^2\big(s\mathbb{1}_A,f'\big)\big],\\
0\le\rho_i\big(\zeta_i,f',f\big)-\rho_i(s\mathbb{1}_A,f)&\le\frac{1}{\sqrt{2}}\big[H_i^2(s\mathbb{1}_A,f)+H_i^2\big(s\mathbb{1}_A,f'\big)\big].
\end{aligned}$$

Now,

$$\begin{aligned}H_i^2\big(f', s\mathbb{1}_A\big) - H_i^2(f, s\mathbb{1}_A) &= \bigg[\rho_i(s\mathbb{1}_A, f) - \frac{1}{2}\int_{\mathbb{X}} f(X_i, y)\,\mathrm{d}\mu(y)\bigg] \\ &\quad - \bigg[\rho_i\big(s\mathbb{1}_A, f'\big) - \frac{1}{2}\int_{\mathbb{X}} f'(X_i, y)\,\mathrm{d}\mu(y)\bigg] \\ &= -\mathbb{E}\big[T_i\big(f, f'\big) \mid X_i\big] + \big[\rho_i\big(\zeta_i, f, f'\big) - \rho_i\big(s\mathbb{1}_A, f'\big)\big] \\ &\quad - \big[\rho_i\big(\zeta_i, f', f\big) - \rho_i(s\mathbb{1}_A, f)\big] \\ &\le -\mathbb{E}\big[T_i\big(f, f'\big) \mid X_i\big] + \frac{1}{\sqrt{2}}\big[H_i^2(s\mathbb{1}_A, f) + H_i^2\big(s\mathbb{1}_A, f'\big)\big].\end{aligned}$$

Inequality (B.7) ensues from the relation $Z_i(f, f') = T_i(f, f') - \mathbb{E}[T_i(f, f') \mid X_i]$. Let us now prove (B.8).

$$\begin{aligned}&2\int_{\mathbb{X}} \psi^2\big(f(X_i, y), f'(X_i, y)\big) s\mathbb{1}_A(X_i, y)\,\mathrm{d}\mu(y) \\ &\quad= \int_{\mathbb{X}} \big(\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}\big)^2 \frac{s\mathbb{1}_A(X_i, y)}{f(X_i, y) + f'(X_i, y)}\,\mathrm{d}\mu(y) \\ &\quad= \int_{\mathbb{X}} \big(\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}\big)^2 \bigg(\sqrt{\frac{s\mathbb{1}_A(X_i, y)}{f(X_i, y) + f'(X_i, y)}} - \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}\bigg)^2 \mathrm{d}\mu(y) \\ &\quad\le 2\int_{\mathbb{X}} \big(\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}\big)^2 \bigg(\sqrt{\frac{s\mathbb{1}_A(X_i, y)}{f(X_i, y) + f'(X_i, y)}} - \frac{1}{\sqrt{2}}\bigg)^2 \mathrm{d}\mu(y) \\ &\qquad+ \int_{\mathbb{X}} \big(\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}\big)^2 \mathrm{d}\mu(y) \\ &\quad\le 2\int_{\mathbb{X}} \bigg(\frac{\sqrt{f'(X_i, y)} - \sqrt{f(X_i, y)}}{\sqrt{f(X_i, y) + f'(X_i, y)}}\bigg)^2 \bigg(\sqrt{s\mathbb{1}_A(X_i, y)} - \sqrt{\frac{f(X_i, y) + f'(X_i, y)}{2}}\bigg)^2 \mathrm{d}\mu(y) \\ &\qquad+ 2H_i^2\big(f, f'\big).\end{aligned}$$

The first term of the right-hand side of this inequality can be upper bounded by $4H_i^2(s\mathbb{1}_A, \frac{f+f'}{2})$. Since the map $t \mapsto \sqrt{t}$ is concave, $2H_i^2(s\mathbb{1}_A, \frac{f+f'}{2}) \le H_i^2(s\mathbb{1}_A, f) + H_i^2(s\mathbb{1}_A, f')$. The second term is bounded from above by $H_i^2(f, f') \le 2(H_i^2(s\mathbb{1}_A, f) + H_i^2(s\mathbb{1}_A, f'))$. Finally,

$$\int_{\mathbb{X}} \psi^2\big(f(X_i, y), f'(X_i, y)\big) s\mathbb{1}_A(X_i, y)\,\mathrm{d}\mu(y) \le 3\big[H_i^2(s\mathbb{1}_A, f) + H_i^2\big(s\mathbb{1}_A, f'\big)\big]$$

as wished. □

We shall prove (B.3) by applying the following concentration inequality to the random variable $Z(f, f')$.

Claim B.4. *For all $i \le n-1$, let $\mathcal{F}_i$ be the σ-field generated by the random variables X_j for $j \in \{0, \dots, i\}$. Let $f_0, \dots, f_{n-1} \in \mathbb{L}^1(\mathbb{X}^2, M)$ such that there exists $b \in \mathbb{R}$ with $\sup_{x\in\mathbb{X}^2} |f_i(x)| \le b$ for all $i \in \{0, \dots, n-1\}$. Set*

$$S_n = \sum_{i=0}^{n-1} \big(f_i(X_i, X_{i+1}) - \mathbb{E}\big[f_i(X_i, X_{i+1}) \mid \mathcal{F}_i\big]\big)$$

and

$$V_n = \sum_{i=0}^{n-1} \mathbb{E}\big[f_i^2(X_i, X_{i+1}) \mid \mathcal{F}_i\big].$$

Then, for all $\beta > b$ and $x > 0$

$$\mathbb{P}\left[S_n \geq \frac{V_n}{2(\beta - b)} + \beta x\right] \leq \mathrm{e}^{-x}.$$

Proof. By setting $a^{-1} = 2(\beta - b)$,

$$\begin{aligned}\log \mathbb{P}[S_n \geq aV_n + \beta x] &\leq -x + \log \mathbb{E}\big[\exp\big(\beta^{-1} S_n - a\beta^{-1} V_n\big)\big] \\ &\leq -x + \log \mathbb{E}\big[\exp\big(\beta^{-1} S_{n-1} - a\beta^{-1} V_n\big)\mathbb{E}\big[\exp\big(\beta^{-1}(S_n - S_{n-1})\big) \mid \mathcal{F}_{n-1}\big]\big].\end{aligned}$$

By using Bernstein inequality (equation (2.21) of Massart [32]),

$$\mathbb{E}\big[\exp\big(\beta^{-1}(S_n - S_{n-1})\big) \mid \mathcal{F}_{n-1}\big] \leq \exp\left(\frac{\beta^{-2}(V_n - V_{n-1})}{2(1 - \beta^{-1} b)}\right)$$

and thus

$$\log \mathbb{P}[S_n \geq aV_n + \beta x] \leq -x + \log \mathbb{E}\big[\exp\big(\beta^{-1} S_{n-1} - a\beta^{-1} V_{n-1}\big)\big].$$

The result follows by induction. □

Proof of Lemma B.3. Set $z = (1 - 1/\sqrt{2})/4$, $\beta = (3/z + \sqrt{2})/2$ and for all $\xi > 0$,

$$\Omega_\xi = \left\{ \sup_{\substack{(f,f') \in S_m \times S_{m'} \\ (m,m') \in \mathcal{M}_\ell^2}} \frac{Z(f, f')}{z(H^2(f, s\mathbb{1}_A) + H^2(f', s\mathbb{1}_A)) + \mathrm{pen}(m) + \mathrm{pen}(m') + \beta\xi} < 1 \right\}.$$

On Ω_ξ, for all $m, m' \in \mathcal{M}_\ell$, $(f, f') \in S_m \times S_{m'}$,

$$Z\big(f, f'\big) \leq z\big(H^2(f, s\mathbb{1}_A) + H^2\big(f', s\mathbb{1}_A\big)\big) + \mathrm{pen}(m) + \mathrm{pen}\big(m'\big) + \beta\xi$$

and (B.3) derives from (B.5) (with $\varepsilon = 1/\sqrt{2} + z$).

It remains to prove that $\mathbb{P}(\Omega_\xi^c) \leq 3\mathrm{e}^{-n\xi}$. We have

$$\mathbb{P}\big(\Omega_\xi^c\big) \leq \sum_{\substack{(f,f') \in S_m \times S_{m'} \\ (m,m') \in \mathcal{M}_\ell^2}} \mathbb{P}\big[Z\big(f, f'\big) \geq z\big[H^2(s\mathbb{1}_A, f) + H^2\big(s\mathbb{1}_A, f'\big)\big] + \mathrm{pen}(m) + \mathrm{pen}\big(m'\big) + \beta\xi\big].$$

We apply the concentration inequality given by Claim B.4 with $f_i = \psi(f, f')$, $b = 1/\sqrt{2}$, $S_n = nZ(f, f')$ and by using relation (B.6),

$$V_n = \sum_{i=0}^{n-1} \int_{\mathbb{X}} \psi^2\big(f(X_i, y), f'(X_i, y)\big)(s\mathbb{1}_A)(X_i, y)\, \mathrm{d}\mu(y) \leq 3n\big(H^2(s\mathbb{1}_A, f) + H^2\big(s\mathbb{1}_A, f'\big)\big).$$

We obtain for all $x > 0$,

$$\mathbb{P}\left[Z\big(f, f'\big) \geq \frac{3}{\sqrt{2}(\beta\sqrt{2} - 1)}\big[H^2(s\mathbb{1}_A, f) + H^2\big(s\mathbb{1}_A, f'\big)\big] + \beta\frac{x}{n}\right] \leq \mathrm{e}^{-x}.$$

Note that $z = 3/(\sqrt{2}(\beta\sqrt{2}-1))$. By using the above inequality with

$$\beta\frac{x}{n} = \operatorname{pen}(m) + \operatorname{pen}\big(m'\big) + \beta\xi$$

we deduce that

$$\mathbb{P}\big(\Omega_\xi^c\big) \leq \sum_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} \mathrm{e}^{-n(\beta^{-1}\operatorname{pen}(m)+\beta^{-1}\operatorname{pen}(m')+\xi)}.$$

Now, the set S_K defined in Definition B.1, page 1052, satisfies $|S_K| \leq (\ell+1)n(n+1)$ which implies that $|S_m| \leq |m|^{(\ell+1)n(n+1)}$. By using $\ell \leq n$, $\log|S_m| \leq 3|m|\log(n+1)$. Since L is large enough ($L \geq 90$), $\beta^{-1}\operatorname{pen}(m) \geq (|m| + \log|S_m|)/n$ for all $m \in \mathcal{M}_\ell$. Consequently,

$$\begin{aligned}\mathbb{P}\big(\Omega_\xi^c\big) &\leq \sum_{\substack{(f,f')\in S_m\times S_{m'}\\(m,m')\in\mathcal{M}_\ell^2}} \mathrm{e}^{-(|m|+\log|S_m|+|m'|+\log|S_{m'}|+n\xi)}\\ &\leq \Bigg(\sum_{m\in\mathcal{M}_\ell}\mathrm{e}^{-|m|}\Bigg)^2\mathrm{e}^{-n\xi}.\end{aligned}$$

The conclusion follows from the inequality $\sum_{m\in\mathcal{M}_\ell}\mathrm{e}^{-|m|} \leq \sqrt{3}$ (see Section 3.2.4 of Baraud and Birgé [6]). □

B.3.3. *Proof of Lemma* B.1
The proof of this lemma is based on the two following remarks.

1. The Hellinger distance $H^2(f, f')$ and the test $T(f, f')$ are respectively upper bounded by 1 and 2 when f, f' are such that

$$\frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{R}^d} f(X_i, y)\,\mathrm{d}\mu(y) \leq 1 \quad \text{and} \quad \frac{1}{n}\sum_{i=0}^{n-1}\int_{\mathbb{R}^d} f'(X_i, y)\,\mathrm{d}\mu(y) \leq 1.$$

2. The cardinality of a partition $m \in \mathcal{M}_\ell \setminus \mathcal{M}_n$ is lower bounded by $|m| \geq n+1$ when $\ell \geq n+1$.

More precisely, the proof follows from the two claims below.

Claim B.5. *Let for each $m_1, m_2 \in \mathcal{M}_\infty$ and $K \in m_1$,*

$$\gamma_K(m_1, m_2) = \alpha H^2(\hat{s}_{m_1}\mathbb{1}_K, \hat{s}_{m_2}\mathbb{1}_K) + T(\hat{s}_{m_1}\mathbb{1}_K, \hat{s}_{m_2}\mathbb{1}_K) - \operatorname{pen}(m_2 \vee K).$$

Then, for all $\ell \in \mathbb{N}^\star$, $\ell \geq n+1$, $m_1 \in \mathcal{M}_\infty$, $K \in m_1$,

$$\sup_{m_2\in\mathcal{M}_\ell}\gamma_K(m_1, m_2) = \sup_{m_2\in\mathcal{M}_n}\gamma_K(m_1, m_2)$$

and thus

$$\sup_{m_2\in\mathcal{M}_\infty}\gamma_K(m_1, m_2) = \sup_{m_2\in\mathcal{M}_n}\gamma_K(m_1, m_2).$$

Proof. Let $m_2^\star \in \mathcal{M}_\ell$ such that $\gamma_K(m_1, m_2^\star) = \sup_{m_2\in\mathcal{M}_\ell}\gamma_K(m_1, m_2)$. In Section 2, we have defined the collection $\mathcal{M}_\ell$ of partitions of $[0,1]^{2d}$. Likewise, by using the algorithm of DeVore and Yu [21], we define the collection $\mathcal{M}_\ell(K)$

of partitions of K. Note that $m_2^\star \vee K$ belongs to $\mathcal{M}_\ell(K)$. Since $H^2(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K) \le 1$ and $|T(\hat{s}_m \mathbb{1}_K, \hat{s}_{m'} \mathbb{1}_K)| \le 2$, we have

$$\gamma_K\big(m_1, m_2^\star\big) \le 3 - L\frac{|m_2^\star \vee K| \log n}{n}.$$

Remark that

$$\gamma_K\big(m_1, m_2^\star\big) \ge \gamma_K\big(m_1, \big\{[0,1]^{2d}\big\}\big) \ge -2 - L\frac{\log n}{n}$$

which leads to

$$\big|m_2^\star \vee K\big| \le 1 + \frac{5n}{L \log n} \le n.$$

This implies that $m_2^\star \vee K$ belongs to $\mathcal{M}_n(K)$. There exists $m_2^\bullet \in \mathcal{M}_n$ such that $m_2^\bullet \vee K = m_2^\star \vee K$ and hence $\gamma_K(m_1, m_2^\bullet) = \gamma_K(m_1, m_2^\star)$ which concludes the proof. □

Claim B.6. *Set for all $m \in \mathcal{M}_\infty$ and $K \in m$,*

$$\gamma_K(m) = \sup_{m_2 \in \mathcal{M}_\ell} \gamma_K(m, m_2).$$

Then, $\gamma(m) = 2\,\mathrm{pen}(m) + \sum_{K \in m} \gamma_K(m)$ and for all $\ell \in \mathbb{N}^\star$, $\ell \ge n+1$,

$$\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m)$$

and thus

$$\inf_{m \in \mathcal{M}_\infty} \gamma(m) = \inf_{m \in \mathcal{M}_n} \gamma(m).$$

Proof. Let $m^\star \in \mathcal{M}_\ell$ such that $\inf_{m \in \mathcal{M}_\ell} \gamma(m) = \gamma(m^\star)$. By Lemma B.2,

$$\begin{aligned}
\gamma\big(m^\star\big) &\ge \sup_{m' \in \mathcal{M}_\ell} \big\{\alpha H^2(\hat{s}_{m^\star}, \hat{s}_{m'}) + T(\hat{s}_{m^\star}, \hat{s}_{m'}) - \mathrm{pen}\big(m'\big)\big\} + L\frac{|m^\star| \log n}{n} \\
&\ge \left(-2 - L\frac{\log n}{n}\right) + L\frac{|m^\star| \log n}{n} \\
&\ge -2 + L\frac{(|m^\star| - 1)\log n}{n}.
\end{aligned}$$

Now,

$$\gamma\big(m^\star\big) \le \gamma\big(\big\{[0,1]^{2d}\big\}\big) \le 3 + 2L\frac{\log n}{n}$$

which implies that

$$\big|m^\star\big| \le 3 + \frac{5n}{L \log n} \le n$$

and thus $m^\star \in \mathcal{M}_n$. □

B.4. Proof of Theorem 2.2

Consider the regular partition m_{ref} of $[0,1]^{2d}$ into cubes with side length $2^{-\ell}$, that is

$$m_{\mathrm{ref}} = \big\{K_{\ell,\mathbf{l}}, \mathbf{l} = (k,\dots,k), k \in \big\{1,\dots,2^{\ell}\big\}\big\},$$

where $K_{\ell,\mathbf{l}}$ is defined in Section 2.2. For all partition $m \in \mathcal{M}_\ell$, $V_m \subset V_{m_{\mathrm{ref}}}$. Set

$$\Omega_{\mathrm{eq}} = \big[\forall g_1, g_2 \in V_{m_{\mathrm{ref}}}, h^2(g_1,g_2) \le 11 H^2(g_1,g_2)\big]$$

and define $\bar{s}_m$ an element of V_m such that $h^2(s\mathbb{1}_A, \bar{s}_m) = h^2(s\mathbb{1}_A, V_m)$.

For all $m \in \mathcal{M}_\ell$,

$$\begin{aligned}
\mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\big] &\le \mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + \mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] \\
&\le 2\mathbb{E}\big[h^2(s\mathbb{1}_A,\bar{s}_m)\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + 2\mathbb{E}\big[h^2(\bar{s}_m,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + \mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] \\
&\le 2\mathbb{E}\big[h^2(s\mathbb{1}_A,\bar{s}_m)\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + 22\mathbb{E}\big[H^2(\bar{s}_m,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + \mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] \\
&\le 2\mathbb{E}\big[h^2(s\mathbb{1}_A,\bar{s}_m)\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + 44\mathbb{E}\big[H^2(s\mathbb{1}_A,\bar{s}_m)\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] + 44\mathbb{E}\big[H^2(\hat{s}_{\hat{m}},s\mathbb{1}_A)\mathbb{1}_{\Omega_{\mathrm{eq}}}\big] \\
&\quad + \mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big].
\end{aligned}$$

Now, $h^2(s\mathbb{1}_A,\bar{s}_m) = \mathbb{E}[H^2(s\mathbb{1}_A,\bar{s}_m)] = h^2(s\mathbb{1}_A,V_m)$ and

$$\begin{aligned}
\mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] &\le 2\mathbb{E}\big[\big(h^2(s,0) + h^2(\hat{s}_{\hat{m}},0)\big)\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] \\
&\le \mathbb{E}\Big[\Big(1 + 2\sup_{m\in\mathcal{M}_\ell} h^2(\hat{s}_m,0)\Big)\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\Big].
\end{aligned}$$

Let for all $K \in m$, I_K and J_K be the subsets of $[0,1]^d$ such that $K = I_K \times J_K$. Then,

$$2h^2(\hat{s}_m,0) = \sum_{K\in m} \frac{\sum_{i=0}^{n-1}\mathbb{1}_{I_K}(X_i)\mathbb{1}_{J_K}(X_{i+1})}{\sum_{i=0}^{n-1}\mathbb{1}_{I_K}(X_i)} \int_{I_K} \varphi(x)\,\mathrm{d}x \le |m|.$$

Since $m \subset m_{\mathrm{ref}}$, $|m| \le |m_{\mathrm{ref}}| = 4^{\ell d}$ and thus

$$\mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\mathbb{1}_{\Omega_{\mathrm{eq}}^c}\big] \le \big(1+4^{\ell d}\big)\mathbb{P}\big(\Omega_{\mathrm{eq}}^c\big) \le 2\times 4^{\ell d}\mathbb{P}\big(\Omega_{\mathrm{eq}}^c\big).$$

We have proved that there exists an universal constant $C' > 0$ such that

$$C'\mathbb{E}\big[h^2(s\mathbb{1}_A,\hat{s}_{\hat{m}})\big] \le \inf_{m\in\mathcal{M}_\ell}\big\{h^2(s\mathbb{1}_A,V_m) + \mathrm{pen}(m)\big\} + 4^{\ell d}\mathbb{P}\big(\Omega_{\mathrm{eq}}^c\big).$$

We now bound from above the term $\mathbb{P}(\Omega_{\mathrm{eq}}^c)$. We denote by $\mathbf{I}_{\mathrm{ref}}$ the regular partition of $[0,1]^d$ into cubes with side length $2^{-\ell}$. Remark that

$$\begin{aligned}
\mathbb{P}\big(\Omega_{\mathrm{eq}}^c\big) &\le \mathbb{P}\left[\exists I \in \mathbf{I}_{\mathrm{ref}}, \mathbb{P}(X_1\in I) \ge \frac{11}{n}\sum_{i=0}^{n-1}\mathbb{1}_I(X_i)\right] \\
&\le 2^{\ell d}\sup_{I\in\mathbf{I}_{\mathrm{ref}}}\mathbb{P}\left[\frac{1}{n}\sum_{i=0}^{n-1}\big(\mathbb{1}_I(X_i) - \mathbb{P}(X_i\in I)\big) \le -\frac{10}{11}\mathbb{P}(X_1\in I)\right].
\end{aligned}$$

We use the following Bennett-type inequality for β-mixing random variables (with $f = -\mathbb{1}_I$, $v = \mathbb{P}(X_1\in I)$, $c = 0$, $\xi = 10/11\mathbb{P}(X_1\in I)$).

Proposition B.1. *Let $(X_i)_{i\geq 0}$ be a stationary Markov chain with values in $\mathbb{R}^d$, and let f be a real-valued function on $\mathbb{R}^d$ upper bounded by $c \geq 0$ such that $v = \mathbb{E}[f(X_1)^2] < \infty$.*

Then, for all $q \in \{1, \ldots, n\}$ and $\xi > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=0}^{n-1}\big(f(X_i) - \mathbb{E}\big[f(X_i)\big]\big) > \xi\right) \leq 2\exp\left(-\frac{n\xi^2}{8q(v + c\xi/6)}\right) + 3n\beta_q/q.$$

We then have for all $I \in \mathbf{I}_{\mathrm{ref}}$,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=0}^{n-1}\big(\mathbb{1}_I(X_i) - \mathbb{P}(X_i \in I)\big) \leq -\frac{10}{11}\mathbb{P}(X_1 \in I)\right] &\leq 3\inf_{1\leq q\leq n}\left\{\exp\left(-\frac{25n\mathbb{P}(X_1 \in I)}{242q}\right) + n\beta_q/q\right\}\\ &\leq 3\inf_{1\leq q\leq n}\left\{\exp\left(-\frac{n\kappa_0}{10q2^{\ell d}}\right) + n\beta_q/q\right\}\end{aligned}$$

which concludes the proof.

Proof of Proposition B.1. Let l be the smallest integer larger than $n/(2q)$. We derive from Berbee's lemma and more precisely from Viennet [36] (page 484) that there exist $X_0^\star, \ldots, X_{2lq-1}^\star$ such that

- For $j = 1, \ldots, l$, the random vectors

$$\mathbf{X}_{j,1} = (X_{2(j-1)q}, \ldots, X_{2(j-1)q+q-1}) \quad \text{and} \quad \mathbf{X}_{j,1}^\star = \big(X_{2(j-1)q}^\star, \ldots, X_{2(j-1)q+q-1}^\star\big)$$

have the same distribution, and so have the random vectors

$$\mathbf{X}_{j,2} = (X_{2(j-1)q+q}, \ldots, X_{2jq-1}) \quad \text{and} \quad \mathbf{X}_{j,2}^\star = \big(X_{2(j-1)q+q}^\star, \ldots, X_{2jq-1}^\star\big).$$

- The random vectors $\mathbf{X}_{1,1}^\star, \ldots, \mathbf{X}_{l,1}^\star$ are independent. The random vectors $\mathbf{X}_{1,2}^\star, \ldots, \mathbf{X}_{l,2}^\star$ are also independent.
- The event

$$\Omega^\star = \bigcap_{1\leq j\leq l}\big(\big[\mathbf{X}_{j,1} \neq \mathbf{X}_{j,1}^\star\big] \cap \big[\mathbf{X}_{j,2} \neq \mathbf{X}_{j,2}^\star\big]\big)$$

satisfies $\mathbb{P}[(\Omega^\star)^c] \leq 2l\beta_q$.

We set $g_i(x) = f(x)$ if $i \leq n-1$ and $g_i(x) = 0$ otherwise. For $j \in \{1, \ldots, l\}$, we set

$$g'_{j,1}(x_0, \ldots, x_{q-1}) = \sum_{i=0}^{q-1} g_{2(j-1)q+i}(x_i) \quad \text{and} \quad g'_{j,2}(x_0, \ldots, x_{q-1}) = \sum_{i=0}^{q-1} g_{2(j-1)q+q+i}(x_i).$$

Then,

$$\begin{aligned}\mathbb{P}\left[\left(\frac{1}{n}\sum_{i=0}^{n-1}\big(g_i(X_i) - \mathbb{E}\big[g_i(X_i)\big]\big) > \xi\right) \cap \Omega^\star\right] &\leq \mathbb{P}\left(\sum_{j=1}^{l}\big(g'_{j,1}\big(\mathbf{X}_{j,1}^\star\big) - \mathbb{E}\big[g'_{j,1}\big(\mathbf{X}_{j,1}^\star\big)\big]\big) > n\xi/2\right)\\ &\quad + \mathbb{P}\left(\sum_{j=1}^{l}\big(g'_{j,2}\big(\mathbf{X}_{j,2}^\star\big) - \mathbb{E}\big[g'_{j,2}\big(\mathbf{X}_{j,2}^\star\big)\big]\big) > n\xi/2\right)\\ &\leq 2\exp\left(-\frac{n^2\xi^2}{8q(nv + cn\xi/6)}\right)\end{aligned}$$

by using Proposition 2.8 and inequality (2.16) of Massart [32] (in the paper of Massart [32], the Bennett inequality holds for $b = 0$ when (2.15) is replaced by (2.16)). □

B.5. Proof of Corollary 2.2

The corollary ensues from the claim below and Theorem 2 of Baraud and Birgé [6].

Claim B.7. *Under Assumption 2.2, for all $\ell \in \mathbb{N}^\star$ such that $2^{\ell d} \geq n$,*

$$\inf_{m\in\mathcal{M}_\ell}\left\{d_2^2(\sqrt{s}|_A, V_m) + \frac{|m|\log n}{n}\right\} \leq 4 \inf_{m\in\mathcal{M}_\infty}\left\{d_2^2(\sqrt{s}|_A, V_m) + \frac{|m|\log n}{n}\right\}.$$

Proof. For all partition $m \in \mathcal{M}_\infty$ and cube $K \in m$, we denote by I_K and J_K the cubes of $[0,1]^d$ such that $K = I_K \times J_K$ and set

$$\bar{s}_m = \sum_{K\in m} \frac{\int_K s(x,y)\,\mathrm{d}x\,\mathrm{d}y}{\mu\otimes\mu(K)}\mathbb{1}_K.$$

In this paper, d_2 stands for the standard euclidean distance of $\mathbb{L}^2([0,1]^{2d}, \mu\otimes\mu)$. In this proof, we make a slight abuse of notations by denoting by d_2 the standard euclidean distance of $\mathbb{L}^2(\mathbb{R}^{2d}, \mu\otimes\mu)$.

Let $m^\star$ be a partition of $\mathcal{M}_\infty$ such that

$$2\inf_{m\in\mathcal{M}_\infty}\left\{d_2^2(\sqrt{s}\mathbb{1}_A, V_m) + \frac{|m|\log n}{n}\right\} \geq d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\star}) + \frac{|m^\star|\log n}{n}.$$

Let $\mathcal{C}$ be the collection $\mathcal{C} = \{K \in m^\star, \mu(I_K) \geq 2^{-\ell d}\}$ and let $m^\bullet$ be a partition of $\mathcal{M}_\ell$ containing $\mathcal{C}$ such that

$$\left|m^\bullet\right| = \inf\left\{|m|, m \in \mathcal{M}_\ell \text{ such that } m \ni \mathcal{C}\right\}.$$

Let $A^\bullet$ be the set defined by $A^\bullet = \bigcup_{K\in\mathcal{C}} K$ and $V^\bullet_{m^\bullet} = \{f\mathbb{1}_{A^\bullet}, f \in V_{m^\bullet}\}$.

We have,

$$d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}) \leq d_2^2\left(\sqrt{s}\mathbb{1}_{A^\bullet}, V^\bullet_{m^\bullet}\right) + d_2^2(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0)$$

and

$$d_2^2\left(\sqrt{s}\mathbb{1}_{A^\bullet}, V^\bullet_{m^\bullet}\right) \leq d_2^2(\sqrt{s}\mathbb{1}_{A^\bullet}, \sqrt{\bar{s}_{m^\bullet}}\mathbb{1}_{A^\bullet}) \leq d_2^2(\sqrt{s}\mathbb{1}_A, \sqrt{\bar{s}_{m^\star}}).$$

By using Lemma 2 of Baraud and Birgé [6], $d_2^2(\sqrt{s}\mathbb{1}_A, \sqrt{\bar{s}_{m^\star}}) \leq 2d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\star})$ which shows that

$$d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}) \leq 2d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\star}) + d_2^2(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0).$$

Now,

$$\begin{aligned} d_2^2(\sqrt{s}\mathbb{1}_{A\cap(A^\bullet)^c}, 0) &\leq \sum_{K\in m^\star\setminus\mathcal{C}} \int_{I_K}\left(\int_{\mathbb{R}^d} s(x,y)\,\mathrm{d}y\right)\mathrm{d}x \\ &\leq \sum_{K\in m^\star\setminus\mathcal{C}} \mu(I_K) \leq 2^{-\ell d}|m^\star|. \end{aligned}$$

Since $|m^\bullet| \leq |m^\star|$ and $2^{-\ell d} \leq n^{-1}$, we have

$$d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\bullet}) + \frac{|m^\bullet|\log n}{n} \leq 2d_2^2(\sqrt{s}\mathbb{1}_A, V_{m^\star}) + \frac{(1+\log n)|m^\star|}{n}$$

which proves the claim. □

B.6. Rates of convergences for h

We prove the result only for geometrically β-mixing chains (the proof for arithmetically β-mixing chains being similar). We use the claim below whose proof is the same than the one of Claim B.7.

Claim B.8. *Under Assumption 2.2, for all $\ell \in \mathbb{N}^\star$ such that $2^{\ell d} \geq n/\log^3 n$,*

$$\inf_{m\in\mathcal{M}_\ell}\left\{h^2(s\mathbb{1}_A, V_m) + \frac{|m|\log n}{n}\right\} \leq 4 \inf_{m\in\mathcal{M}_\infty}\left\{h^2(s\mathbb{1}_A, V_m) + \frac{|m|\log^3 n}{n}\right\}.$$

By using this claim and Theorem 1 of Akakpo [2],

$$C\mathbb{E}\big[h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})\big] \leq |\sqrt{s}|_A|^{2d/(d+\sigma)}_{p,\sigma}\left(\frac{\log^3 n}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log^3 n}{n} + \frac{R_n(\ell)}{n} \tag{B.9}$$

and by using Theorem 2 of Akakpo [2],

$$C\mathbb{E}\big[h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})\big] \leq |\sqrt{s}|_A|^{2d/(d+\sigma)}_{p,\sigma}\left(\frac{\log n}{n} + 2^{-2\ell d\theta}\right)^{\sigma/(\sigma+d)} + \frac{\log n}{n} + \frac{R_n(\ell)}{n},$$

where $C > 0$ depends only on κ, σ, d, p and where

$$\theta = \frac{d+\sigma}{\sigma}\left(\frac{\sigma}{d} - 2\left(\frac{1}{p} - \frac{1}{2}\right)_+\right).$$

If $\sigma > \sigma_1(p,d)$ then $\theta > 1/2$. There exits thus n_0 (depending only on θ), such that if $n \geq n_0$, $2^{-2\ell\theta d} \leq \log n/n$, and hence

$$C'\mathbb{E}\big[h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})\big] \leq |\sqrt{s}|_A|^{2d/(d+\sigma)}_{p,\sigma}\left(\frac{\log n}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}.$$

If $n \leq n_0$, we deduce from (B.9),

$$\begin{aligned} C\mathbb{E}\big[h^2(s\mathbb{1}_A, \hat{s}_{\hat{m}})\big] &\leq |\sqrt{s}|_A|^{2d/(d+\sigma)}_{p,\sigma}\left(\frac{\log^3 n_0}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log^3 n_0}{n} + \frac{R_n(\ell)}{n} \\ &\leq C''\left[|\sqrt{s}|_A|^{2d/(d+\sigma)}_{p,\sigma}\left(\frac{\log n}{n}\right)^{\sigma/(\sigma+d)} + \frac{\log n}{n} + \frac{R_n(\ell)}{n}\right], \end{aligned}$$

where C'' depends only on σ, d, p. The conclusion ensues from the fact that $R_n(\ell)$ is upper bounded by a constant depending only on κ_0, b_1.

B.7. Proof of Proposition 4.1

We shall use the following lemma whose proof is similar to the one of Lemma B.3.

Lemma B.4. *Set $\varepsilon = (2+3\sqrt{2})/8$. Under assumptions of Proposition 4.1, there exists an universal constant $L_0 > 0$ such that for all $L \geq L_0$ and $\xi > 0$,*

$$\forall f, f' \in S, \quad (1-\varepsilon)H^2\big(s\mathbb{1}_A, f'\big) + T\big(f, f'\big) \leq (1+\varepsilon)H^2(s\mathbb{1}_A, f) + L\frac{\Delta_S(f) + \Delta_S(f')}{n} + 22\xi$$

with probability larger than $1 - \mathrm{e}^{-n\xi}$.

Proof of Proposition 4.1. By using the above lemma, with probability larger than $1-\mathrm{e}^{-n\xi}$, for all $f\in S$,

$$\sup_{f'\in S}\left\{(1-\varepsilon)H^2\big(s\mathbb{1}_A,f'\big)+T\big(f,f'\big)-L\frac{\Delta_S(f')}{n}\right\}\leq(1+\varepsilon)H^2(s\mathbb{1}_A,f)+L\frac{\Delta_S(f)}{n}+22\xi.$$

Thus, if $T(f,\hat f)+L\frac{\Delta_S(f)}{n}-L\frac{\Delta_S(\hat f)}{n}\geq 0$,

$$\begin{aligned}\alpha H^2(s\mathbb{1}_A,\hat f)&\leq(1-\varepsilon)H^2(s\mathbb{1}_A,\hat f)+T(f,\hat f)-L\frac{\Delta_S(\hat f)}{n}+L\frac{\Delta_S(f)}{n}\\&\leq(1+\varepsilon)H^2(s\mathbb{1}_A,f)+2L\frac{\Delta_S(f)}{n}+22\xi.\end{aligned}$$

If $T(f,\hat f)+L\frac{\Delta_S(f)}{n}-L\frac{\Delta_S(\hat f)}{n}<0$,

$$\begin{aligned}\alpha H^2(f,\hat f)&\leq\alpha H^2(\hat f,f)+T(\hat f,f)-L\frac{\Delta_S(f)}{n}+L\frac{\Delta_S(\hat f)}{n}\\&\leq\sup_{f'\in S}\left\{\alpha H^2\big(\hat f,f'\big)+T\big(\hat f,f'\big)-L\frac{\Delta_S(f')}{n}\right\}+L\frac{\Delta_S(\hat f)}{n}\\&\leq\wp(\hat f)\\&\leq\wp(f)+\frac{1}{n}\\&\leq\sup_{f'\in S}\left\{\alpha H^2\big(f,f'\big)+T\big(f,f'\big)-L\frac{\Delta_S(f')}{n}\right\}+L\frac{\Delta_S(f)}{n}+\frac{1}{n}.\end{aligned}$$

With $\upsilon=(1-\varepsilon)/\alpha-1>0$,

$$\begin{aligned}\alpha H^2(f,\hat f)&\leq\big(1+\upsilon^{-1}\big)H^2(f,s\mathbb{1}_A)\\&\quad+\sup_{f'\in S}\left\{(1-\varepsilon)H^2\big(s\mathbb{1}_A,f'\big)+T\big(f,f'\big)-L\frac{\Delta_S(f')}{n}\right\}+L\frac{\Delta_S(f)}{n}+\frac{1}{n}\\&\leq\big(1+\upsilon^{-1}\big)H^2(f,s\mathbb{1}_A)+\left[(1+\varepsilon)H^2(s\mathbb{1}_A,f)+L\frac{\Delta_S(f)}{n}+22\xi\right]+L\frac{\Delta_S(f)}{n}+\frac{1}{n}\\&\leq\big(2+\varepsilon+\upsilon^{-1}\big)H^2(f,s\mathbb{1}_A)+2L\frac{\Delta_S(f)}{n}+22\xi+\frac{1}{n}.\end{aligned}$$

This leads to

$$\begin{aligned}\alpha H^2(s\mathbb{1}_A,\hat f)&\leq 2\alpha H^2(s\mathbb{1}_A,f)+2\alpha H^2(f,\hat f)\\&\leq 2\big(2+\alpha+\varepsilon+\upsilon^{-1}\big)H^2(f,s\mathbb{1}_A)+4L\frac{\Delta_S(f)}{n}+44\xi+\frac{2}{n}.\end{aligned}$$

Finally, we have proved that there exists $C>0$, such that, with probability larger than $1-\mathrm{e}^{-n\xi}$, for all $f\in S$,

$$CH^2(s\mathbb{1}_A,\hat f)\leq H^2(f,s\mathbb{1}_A)+L\frac{\Delta_S(f)}{n}+\xi.$$

The conclusion follows. □

B.8. Proof of Corollary 4.2

Throughout this proof, the distance associated to the supremum norm $\|\cdot\|_\infty$ is denoted by d_∞. As defined page 1035, d_2 is the usual distance of the space of square integrable functions on $[0,1]^2$ with respect to the Lebesgue measure $\mu\otimes\mu$. We make a slight abuse of notation in this proof since d_2 will also stand for the distance of the space of square integrable functions on $[0,1]$ with respect to the Lebesgue measure μ.

We shall use the following lemma (the first part may be deduced from the work of Akakpo [2] whereas the second part may be deduced from results in Dahmen, DeVore, and Scherer [20]).

Lemma B.5. *There exist a collection $\mathbb{W}$ of (finite dimensional) linear spaces and a non-negative map $\Delta_{\mathbb{W}}$ on $\mathbb{W}$ such that $\sum_{W\in\mathbb{W}}\mathrm{e}^{-\Delta_{\mathbb{W}}(W)}\le 1$ and such that for all $p\in(0,+\infty]$, $\beta>(1/p-1/2)_+$ and $f\in\mathscr{B}^\beta(\mathbb{L}^p([0,1]))$, $L>0$, $\tau>0$, $\sigma>0$,*

$$C\inf_{W\in\mathbb{W}}\bigl\{L^2d_2^{2\sigma}(f,W)+\bigl(\dim W+\Delta_{\mathbb{W}}(W)\bigr)\tau\bigr\}\le\bigl(L|f|_{p,\beta}^{\sigma}\bigr)^{2/(2\sigma\beta+1)}\tau^{2\sigma\beta/(2\sigma\beta+1)}+\tau,$$

where $C>0$ depends only on p, β. Moreover, for all $\beta>0$, $f\in\mathcal{H}^\beta([0,1])$, $L>0$, $\tau>0$, $\sigma>0$,

$$C'\inf_{W\in\mathbb{W}}\bigl\{L^2d_\infty^{2\sigma}(f,W)+\bigl(\dim W+\Delta_{\mathbb{W}}(W)\bigr)\tau\bigr\}\le\bigl(L|f|_{\infty,\beta}^{\sigma}\bigr)^{2/(2\sigma\beta+1)}\tau^{2\sigma\beta/(2\sigma\beta+1)}+\tau,$$

where $C'>0$ depends only on β.

Let us define

$$u(x,y)=\frac{y-g(x)}{1+\|g\|_\infty}\quad\text{and}\quad\Phi(x)=\phi\bigl(\bigl(1+\|g\|_\infty\bigr)x\bigr)\quad\text{for all } x,y\in[0,1].$$

Let $\mathbb{W}$ be the family of linear spaces given by the above lemma. Let for all $f\in\bigcup_{W\in\mathbb{W}}W$, $a\in\mathbb{R}$, $\psi_{a,f}$ be the function defined on $[0,1]^2$ by $\psi_{a,f}(x,y)=a(y-f(x))$. Define, for all $W\in\mathbb{W}$, the linear space

$$T_W=\{\psi_{a,f},a\in\mathbb{R},f\in W\}.$$

We deduce from the proof of Theorem 2 of Baraud and Birgé [7] (with $\mathbb{F}=\mathbb{W}$, $l=1$, $\mathbb{T}_1=\{T_W,W\in\mathbb{W}\}$, $\gamma(W)=\mathrm{e}^{-\Delta_{\mathbb{W}}(W)}$, $\lambda_1(T_W)=\mathrm{e}^{-\Delta_{\mathbb{W}}(W)}$) and from relation (4.5) of Baraud and Birgé [7], that there exist an at most countable collection $\mathbb{V}$ of models and a non-negative map Δ on $\mathbb{V}$ such that $\sum_{\mathbb{V}\in\mathbb{V}}\mathrm{e}^{-\Delta(V)}\le 1$ and

$$\begin{aligned}&C\inf_{V\in\mathbb{V}}\left\{d^2(\sqrt{s}|_A,V)+\frac{\Delta(V)+\dim(V)\log n}{n}\right\}\\&\quad\le\inf_{W\in\mathbb{W}}\bigl\{|\Phi|_{\infty,\sigma\wedge1}^2d_2^{2(\sigma\wedge1)}(u,T_W)+\bigl(\dim T_W+\Delta_{\mathbb{W}}(W)\bigr)\tau_n\bigr\}\\&\qquad+\inf_{W\in\mathbb{W}}\left\{d_\infty^2(\Phi,W)+\bigl(\dim W+\Delta_{\mathbb{W}}(W)\bigr)\frac{\log n}{n}\right\},\end{aligned}$$

where $C>0$ depends only on σ and where

$$\tau_n=\bigl(\log n\vee\log\bigl(|\Phi|_{\infty,\sigma\wedge1}\bigr)\bigr)\frac{\log n}{n}.$$

Besides, for all linear space $V\in\mathbb{V}$, there exists a function $\psi\in\bigcup_{T\in\mathbb{T}_1}T$ and a linear space $W\in\mathbb{W}$ such that $V=\{f\circ\psi,f\in W\}$.

We apply Theorem 4.1 to $(\mathbb{V},\Delta)$ to construct an estimator $\hat{s}$ of the form

$$\sqrt{\hat{s}(x,y)}=\hat{\phi}\bigl(y-\hat{g}(x)\bigr)$$

such that

$$C' \inf_{V\in\mathbb{V}} \mathbb{E}\big[H^2(s\mathbb{1}_A,\hat{s})\big] \le \inf_{W\in\mathbb{W}}\big\{|\Phi|^2_{\infty,\sigma\wedge1} d_2^{2(\sigma\wedge1)}(u,T_W) + \big(\dim T_W + \Delta_{\mathbb{W}}(W)\big)\tau_n\big\} + \inf_{W\in\mathbb{W}}\left\{d^2_\infty(\Phi,W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\frac{\log n}{n}\right\},$$

where $C'>0$ depends only on σ,κ. We upper bound the two terms of the right-hand side of this inequality. We derive from

$$|\Phi|_{\infty,\sigma\wedge1} \le \big(1+\|g\|_\infty\big)^{\sigma\wedge1}|\phi|_{\infty,\sigma\wedge1} \quad\text{and}\quad d_2^{2(\sigma\wedge1)}(u,T_W) \le \frac{d_2^{2(\sigma\wedge1)}(g,W)}{(1+\|g\|_\infty)^{2(\sigma\wedge1)}}$$

that

$$\inf_{W\in\mathbb{W}}\big\{|\Phi|^2_{\infty,\sigma\wedge1} d_2^{2(\sigma\wedge1)}(u,T_W) + \big(\dim T_W + \Delta_{\mathbb{W}}(W)\big)\tau_n\big\} \le \inf_{W\in\mathbb{W}}\big\{|\phi|^2_{\infty,\sigma\wedge1} d_2^{2(\sigma\wedge1)}(g,W) + \big(\dim W + \Delta_{\mathbb{W}}(W)+1\big)\tau_n\big\}.$$

By using the above lemma,

$$C'' \inf_{W\in\mathbb{W}}\big\{|\phi|^2_{\infty,\sigma\wedge1} d_2^{2(\sigma\wedge1)}(g,W) + \big(\dim W + \Delta_{\mathbb{W}}(W)+1\big)\tau_n\big\} \le \big(|\phi|_{\infty,\sigma\wedge1}|g|^{\sigma\wedge1}_{p,\beta}\big)^{2/(2(\sigma\wedge1)\beta+1)}\tau_n^{2(\sigma\wedge1)\beta/(2(\sigma\wedge1)\beta+1)} + \tau_n \le C'''\left(\frac{\log^2 n}{n}\right)^{2(\sigma\wedge1)\beta/(2(\sigma\wedge1)\beta+1)}.$$

Similarly,

$$C'''' \inf_{W\in\mathbb{W}}\left\{d^2_\infty(\Phi,W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\frac{\log n}{n}\right\} \le \big(|\Phi|_{\infty,\sigma}\big)^{2/(2\sigma+1)}\left(\frac{\log n}{n}\right)^{2\sigma/(2\sigma+1)} + \frac{\log n}{n} \le C''''\left(\frac{\log n}{n}\right)^{2\sigma/(2\sigma+1)}.$$

B.9. Proof of Lemma 4.1

The first part of the lemma may be deduced from Proposition 4 of Baraud and Birgé [7]. For the second part, we shall build $\phi'\in\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]}\notin\bigcup_{b>\sigma}\mathcal{H}^b([0,1])$ and $g'\in\mathcal{H}^\beta([0,1])$ such that $g'(0)=0$ and

$$\phi'\circ g' \in \mathcal{H}^{\theta(\beta,\sigma)}\big([0,1]\big)\setminus\bigcup_{b>\theta(\beta,\sigma)}\mathcal{H}^b\big([0,1]\big).$$

By setting $\phi=\phi'$ and $g=-g'$, the function f defined by

$$f(x,y)=\phi'\big(y-\big(-g'(x)\big)\big) \quad \text{for all } x,y\in[0,1],$$

is suitable since $f(x,0)=\phi'\circ g'(x)$ and $f(0,y)=\phi'(y)$.

If $\sigma,\beta\le1$, we can choose $\phi'(x)=x^\sigma$ on $[0,1]$ and $g'(x)=x^\beta$. If $\beta\ge\sigma\vee1$, then choose $\phi'\in\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]}\notin\bigcup_{b>\sigma}\mathcal{H}^b([0,1])$ and $g'(x)=x$. If now, $\sigma\ge\beta\vee1$, we choose $\phi'\in\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]}\notin\bigcup_{b>\sigma}\mathcal{H}^b([0,1])$ and such that $\phi'(x)=x$ for all $x\in[0,1/2]$. We then consider $\zeta\in\mathcal{H}^\beta([0,1])\setminus\bigcup_{b>\beta}\mathcal{H}^b([0,1])$ and $g'(x)=(\zeta(x)-\zeta(0))/(2\sup_{y\in[0,1]}|\zeta(y)-\zeta(0)|)$.

B.10. *Proof of Corollary* 4.3

We shall use the distances d_2 and d_∞ that have been defined at the beginning of the proof of Corollary 4.2.

Let us define

$$\forall x, y, z \in [0,1], \quad u(x,y) = \big(u_1(x,y), u_2(x,y), u_3(x,y)\big) = \left(\frac{y - v_1(x)}{1 + \|v_1\|_\infty}, \frac{v_2(x)}{\|v_2\|_\infty}, \frac{v_3(x)}{\|v_3\|_\infty}\right),$$
$$\Phi(x,y,z) = \|v_3\|_\infty z \varphi\big(\big(1 + \|v_1\|_\infty\big)\|v_2\|_\infty xy\big).$$

Let $\mathbb{W}$ be the family of linear spaces given by Lemma B.5. Let for all $a \in \mathbb{R}$, $f \in \bigcup_{W\in\mathbb{W}} W$, $\psi_{a,f}$ be the function defined on $[0,1]^2$ by $\psi_{a,f}(x,y) = a(y - f(x))$ and g_f be the function defined on $[0,1]^3$ by $g_f(x,y,z) = zf(xy)$. For all $W \in \mathbb{W}$, we consider the linear spaces

$$T_W = \{\psi_{a,f}, a \in \mathbb{R}, f \in W\} \quad \text{and} \quad F_W = \{g_f, f \in W\}.$$

It ensues from the proof of Theorem 2 of Baraud and Birgé [7] (where $l = 3$, $\mathbb{F} = \{F_W, W \in \mathbb{W}\}$, $\mathbb{T}_1 = \{T_W, W \in \mathbb{W}\}$, $\mathbb{T}_2 = \mathbb{T}_3 = \mathbb{W}$, $\gamma(F_W) = \lambda_1(T_W) = \lambda_2(W) = \lambda_3(W) = \mathrm{e}^{-\Delta_{\mathbb{W}}(W)}$) that there exist an at most countable collection $\mathbb{V}$ of models and a non-negative map Δ on $\mathbb{V}$ such that $\sum_{\mathbb{V}\in\mathbb{V}} \mathrm{e}^{-\Delta(V)} \le 1$ and such that

$$\begin{aligned}
&C \inf_{V\in\mathbb{V}} \left\{ d^2(\sqrt{s}|_A, V) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\} \\
&\quad \le \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 \big(1 + \|v_1\|_\infty\big)^{2(\sigma\wedge 1)} \|v_2\|_\infty^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 d_2^{2(\sigma\wedge 1)}(u_1, T_W) + \big(\dim T_W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(1)} \right\} \\
&\qquad + \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 \big(1 + \|v_1\|_\infty\big)^{2(\sigma\wedge 1)} \|v_2\|_\infty^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 d_2^{2(\sigma\wedge 1)}(u_2, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(1)} \right\} \\
&\qquad + \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 \|\varphi\|_\infty^2 d_2^2(u_3, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(2)} \right\} \\
&\qquad + \inf_{W\in\mathbb{W}} \left\{ d_\infty^2(\Phi, F_W) + \big(\dim F_W + \Delta_{\mathbb{W}}(W)\big)\frac{\log n}{n} \right\},
\end{aligned}$$

where

$$\tau_n^{(1)} = \Big(\log n \vee \log\big(\|v_3\|_\infty^2 \big(1 + \|v_1\|_\infty\big)^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 \|v_2\|_\infty^{2(\sigma\wedge 1)}\big)\Big)\frac{\log n}{n},$$
$$\tau_n^{(2)} = \Big(\log n \vee \log\big(\|v_3\|_\infty^2 \|\varphi\|_\infty^2\big)\Big)\frac{\log n}{n}.$$

By applying Theorem 4.1 to $(\mathbb{V}, \Delta)$, we build an estimator $\hat{s}$ such that

$$C'\mathbb{E}\big[H^2(s,\hat{s})\big] \le \inf_{V\in\mathbb{V}} \left\{ d^2(\sqrt{s}|_A, V) + \frac{\Delta(V) + \dim(V)\log n}{n} \right\}$$

and thus

$$\begin{aligned}
C''\mathbb{E}\big[H^2(s,\hat{s})\big] \le{}& \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 |\varphi|_{\infty,\sigma}^2 \|v_2\|_\infty^{2(\sigma\wedge 1)} d_2^{2(\sigma\wedge 1)}(v_1, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(1)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 \big(1 + \|v_1\|_\infty\big)^{2(\sigma\wedge 1)} |\varphi|_{\infty,\sigma}^2 d_2^{2(1\wedge\sigma)}(v_2, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(1)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ \|\varphi\|_\infty^2 d_2^{2(1\wedge\sigma)}(v_3, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\tau_n^{(2)} \right\} \\
&+ \inf_{W\in\mathbb{W}} \left\{ \|v_3\|_\infty^2 d_\infty^2(\varphi, W) + \big(\dim W + \Delta_{\mathbb{W}}(W)\big)\frac{\log n}{n} \right\}.
\end{aligned}$$

We conclude by applying Lemma B.5 as in the end of the proof of Corollary 4.2.

B.11. *Proof of Lemma* 4.2

The first part of the lemma can be deduced from Proposition 4 of Baraud and Birgé [7]. For the second part, remark that, as in the proof of Lemma 4.1 the problem amounts to finding $\phi' \in \mathcal{H}^\sigma(\mathbb{R})$ with $\phi'|_{[0,1]} \notin \bigcup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$, $v_i' \in \mathcal{H}^{\beta_i}([0,1])$ for $i \in \{1,2\}$, $v_1'(0) = 0$, $v_2'(0) = 1$ such that

$$\sqrt{v_2'}\phi'\big(v_1'v_2'\big) \in \mathcal{H}^{\theta(\beta_1,\beta_2,\sigma)}\big([0,1]\big) \setminus \bigcup_{b>\theta(\beta_1,\beta_2,\sigma)} \mathcal{H}^b\big([0,1]\big).$$

If $\theta(\beta_1,\beta_2,\sigma) = 2^{-1}(\beta_2 \wedge 1)$, choose $v_2'(x) = (1-x)^{1\wedge\beta_2}$ and take ϕ' as being any function of $\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]} \notin \bigcup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$ and such that $\phi'(0) = 1$. If $\theta(\beta_1,\beta_2,\sigma) = \sigma$, choose $v_1'(x) = 2(\sqrt{1+x}-1)$, $v_2'(x) = 1/2(\sqrt{1+x}+1)$ and take ϕ' as being any function of $\mathcal{H}^\sigma(\mathbb{R})$ such that $\phi'|_{[0,1]} \notin \bigcup_{a>\sigma} \mathcal{H}^a(\mathbb{R})$. If $\theta(\beta_1,\beta_2,\sigma) = \sigma\beta_1$, we may assume that $\sigma \le 1$ and $\beta_1 \le 1$. We can then choose $v_1'(x) = x^{\beta_1}$, $v_2'(x) = 1$ and $\phi'(x) = x^\sigma$ for $x \in [0,1]$. If $\theta(\beta_1,\beta_2,\sigma) = \sigma\beta_2$, we may assume that $\sigma \le 1$ and $\beta_2 \le 1$ and choose $v_1'(x) = 1$ for $x \in [1/2,1]$, $v_2'(x) = 1-(1-x)^{\beta_2}$ for $x \in [1/2,1]$ and $\phi'(x) = (1-x)^\sigma$ for $x \in [0,1]$. Finally, if $\theta(\beta_1,\beta_2,\sigma) = \beta_1$, we may assume that $\beta_1 \le 1$. We can then choose $v_1'(x) = x^{\beta_1}$, $v_2'(x) = (1-x)^{1\wedge\beta_2}$ and ϕ' such that $\phi'(x) = x$ for $x \in [0,1/2]$.

B.12. *Proof of Proposition* A.1

We proceed in 3 steps.

Step 1. We associate to each cube $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, a place in the computer's memory. Then, for each $i \in \{1,\dots,n\}$ we determine the sets $K \in \bigcup_{m\in\mathcal{M}_\ell} m$ such that $\mathbb{1}_K(X_i, X_{i+1}) > 0$. There are at most ℓ such sets. This permits to store all the $\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1})$ in around $\mathcal{O}(n\ell d)$ operations. Let for all $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, I_K and J_K be the subsets of $[0,1]^d$ such that $K = I_K \times J_K$. We can store all the $\mu(J_K)$ in $\mathcal{O}(4^{\ell d})$ operations and all the $\sum_{i=0}^{n-1} \mathbb{1}_{I_K}(X_i)$ in $\mathcal{O}(n\ell d)$ operations. This permits us to store quickly

$$\sum_{i=0}^{n-1} \mathbb{1}_K(X_i, X_{i+1}) \quad \text{and} \quad \sum_{i=0}^{n-1} \int_{[0,1]^d} \mathbb{1}_K(X_i, x)\,\mathrm{d}\mu(x)$$

for all $K \in \bigcup_{m\in\mathcal{M}_\ell} m$. These values have to be calculated to know the $F_K(K')$ and thus to use the algorithm presented in Section A.

Step 2. For each $K \in \bigcup_{m\in\mathcal{M}_\ell} m$, we use the algorithm of Section A to design m'_K. Let us denote by $j \in \{0,\dots,\ell\}$ the smallest integer such that $K \in \mathcal{K}_j$ where $\mathcal{K}_j$ is defined in Section 2.2.

- To find m'_K, we begin to compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \bigcup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_{\ell-1}} m$ such that $K'' \cap K \neq \varnothing$. The complexity of this is around the number of such sets, i.e, $4^{(\ell-j)d}$.
- Next, thanks to relation (A.4) we compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \bigcup_{m\in\mathcal{M}_{\ell-1}\setminus\mathcal{M}_{\ell-2}} m$ such that $K'' \cap K \neq \varnothing$. There are $4^{(\ell-j-1)d}$ such sets. The complexity of this operation is thus $4^d \times 4^{(\ell-j-1)d}$.
- By recurrence, we compute $\mathcal{E}(T^\star(K''))$ for all $K'' \in \bigcup_{m\in\mathcal{M}_\ell\setminus\mathcal{M}_j} m$ such that $K'' \cap K \neq \varnothing$ in at most

$$4^{(\ell-j)d} + 4^d \times \sum_{k=1}^{\ell-j-1} 4^{kd} \le 3 \times 4^{(\ell-j)d}$$

operations.
- We get then $\mathcal{E}(T^\star([0,1]^d))$ in $4^d j$ additional operations.

We apply this algorithm for all $K \in \bigcup_{m\in\mathcal{M}_\ell} m$. When $K \in \mathcal{K}_j$, computing m'_K requires thus $\mathcal{O}(4^{(\ell-j)d} + 4^d j)$ operations. Since $|\mathcal{K}_j| = 4^{jd}$, computing all the m'_K requires finally

$$\sum_{j=0}^{\ell} 4^{jd}\big(4^{(\ell-j)d} + 4^d j\big) = \mathcal{O}\big(\ell 4^{(\ell+1)d}\big)$$

operations.

Step 3. Now, by slightly modifying the algorithm, we can compute (A.2) in $\mathcal{O}(4^{(\ell+1)d})$ operations.

Acknowledgments

Many thanks to Yannick Baraud for his suggestions, comments, careful reading of the paper. We are thankful to Claire Lacour for sending us the source code of the procedure of Akakpo and Lacour [3]. We gratefully acknowledge an anonymous referee for its comments and questions on an earlier draft of this paper.

References

[1] N. Akakpo. Estimation adaptative par sélection de partitions en rectangles dyadiques. Ph.D. thesis, Univ. Paris Sud, 2009.
[2] N. Akakpo. Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Statist.* **21** (2012) 1–28. MR2901269
[3] N. Akakpo and C. Lacour. Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electron. J. Statist.* **5** (2011) 1618–1653. MR2870146
[4] K. B. Athreya and G. S. Atuncar. Kernel estimation for real-valued Markov chains. *Sankhyā* **60** (1998) 1–17. MR1714774
[5] Y. Baraud. Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields* **151** (2011) 353–401. MR2834722
[6] Y. Baraud and L. Birgé. Estimating the intensity of a random measure by histogram type estimators. *Probab. Theory Related Fields* **143** (2009) 239–284. MR2449129
[7] Y. Baraud and L. Birgé. Estimating composite functions by model selection. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** (2014) 285–314. MR3161532
[8] A. K. Basu and D. K. Sahoo. On Berry–Esseen theorem for nonparametric density estimation in Markov sequences. *Bull. Inform. Cybernet.* **30** (1998) 25–39. MR1629735
[9] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Probab. Theory Related Fields* **65** (1983) 181–237. MR0722129
[10] L Birgé. Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. Henri Poincaré Probab. Stat.* **20** (1984) 201–223. MR0762855
[11] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **2** (1984) 259–282. MR0764150
[12] L. Birgé. Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **42** (2006) 273–325. MR2219712
[13] L. Birgé. Model selection for Poisson processes. In *Asymptotics: Particles, Processes and Inverse Problems* 32–64. *IMS Lecture Notes Monogr. Ser.* **55**. IMS, Beachwood, OH, 2007. MR2459930
[14] L. Birgé. Model selection for density estimation with $\mathbb{L}_2$-loss. *Probab. Theory Related Fields* **158** (2014) 533–574. MR3176358
[15] L. Birgé. Robust tests for model selection. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner* 47–64. *IMS Collections* **9**. IMS, Beachwood, OH, 2012.
[16] G. Blanchard, C. Schäfer and Y. Rozenholc. *Oracle Bounds and Exact Algorithm for Dyadic Classification Trees*. *Lecture Notes in Comput. Sci.* **3120**. Springer, Berlin, 2004. MR2177922
[17] R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** (2005) 107–144. MR2178042
[18] S. Clémencon. Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.* **9** (2000) 323–357. MR1827473
[19] F. Comte and Y. Rozenholc. Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.* **97** (2002) 111–145. MR1870963
[20] W. Dahmen, R. DeVore and K. Scherer. Multi-dimensional spline approximation. *SIAM J. Numer. Anal.* **17** (1980) 380–402. MR0581486
[21] R. DeVore and X. Yu. Degree of adaptive approximation. *Math. Comput.* **55** (1990) 625–635. MR1035930
[22] C. C. Y. Dorea. Strong consistency of kernel estimators for Markov transition densities. *Bull. Braz. Math. Soc. (N.S.)* **33** (2002) 409–418. MR1978836
[23] P. Doukhan. *Mixing: Properties and Examples*. *Lecture Notes in Statistics* **85**. Springer, New York, 1994. MR1312160
[24] 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** (1983) 271–293. MR0711186
[25] R. Hochmuth. Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.* **12** (2002) 179–208. MR1884234
[26] A. Juditsky, O. Lepski and A. Tsybakov. Nonparametric estimation of composite functions. *Ann. Statist.* **37** (2009) 1360–1404. MR2509077
[27] C. Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. Henri Poincaré Probab. Statist.* **43** (2007) 571–597. MR2347097
[28] C. Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Process. Appl.* **118** (2008) 232–260. MR2376901
[29] C. Lacour. Erratum to "Nonparametric estimation of the stationary density and the transition density of a Markov chain" [*Stochastic Process. Appl.* **118** (2008) 232–260] [MR2376901]. *Stochastic Process. Appl.* **122** (2012) 2480–2485. MR2922637
[30] L. Le Cam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** (1973) 38–53. MR0334381

[31] L. Le Cam. On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic Processes and Related Topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)* 13–54. Academic Press, New York, 1975. MR0395005

[32] P. Massart. *Concentration Inequalities and Model Selection*. *Lecture Notes in Mathematics* **1896**. Springer, Berlin, 2003. MR2319879

[33] G. G. Roussas. Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.* **21** (1969) 73–87. MR0247722

[34] G. G. Roussas. *Estimation of Transition Distribution Function and Its Quantiles in Markov Processes: Strong Consistency and Asymptotic Normality*. *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335**. Kluwer Acad. Publ., Dordrecht, 1991. MR1154345

[35] M. Sart. Model selection for poisson processes with covariates. ArXiv e-prints, 2012.

[36] G. Viennet. Inequalities for absolutely regular sequences: Application to density estimation. *Probab. Theory Related Fields* **107** (1997) 467–492. MR1440142