ESAIM: Probability and Statistics November 2002, Vol. 6, 211–238
URL: http://www.emath.fr/ps/ DOI: 10.1051/ps:2002012

ADAPTIVE ESTIMATION OF THE STATIONARY DENSITY OF DISCRETE AND CONTINUOUS TIME MIXING PROCESSES

Fabienne Comte[1] and Florence Merlevède[2]

Abstract. In this paper, we study the problem of non parametric estimation of the stationary marginal density f of an α or a β-mixing process, observed either in continuous time or in discrete time. We present an unified framework allowing to deal with many different cases. We consider a collection of finite dimensional linear regular spaces. We estimate f using a projection estimator built on a data driven selected linear space among the collection. This data driven choice is performed *via* the minimization of a penalized contrast. We state non asymptotic risk bounds, regarding to the integrated quadratic risk, for our estimators, in both cases of mixing. We show that they are adaptive in the minimax sense over a large class of Besov balls. In discrete time, we also provide a result for model selection among an exponentially large collection of models (non regular case).

Mathematics Subject Classification. 62G07, 62M99.

1. Introduction

Consider a strictly stationary mixing process (X_τ) observed either in continuous time for τ varying in $[0,T]$ or in discrete time for $\tau = 1,\dots,n$ and denote in both cases by f its marginal density with respect to the Lebesgue measure. In this paper, we are interested in the problem of giving non asymptotic risk bounds in term of the $\mathbb{L}_2$-integrated risk for an estimator $\hat{f}$ of f. Namely we study $\mathbb{E}\|\hat{f}-f\|^2$ where $\|t\| = (\int_A t^2(x)\mathrm{d}x)^{1/2}$ is the $\mathbb{L}_2(A)$-norm and A is a compact set. Besides, we want to provide an adaptive procedure, that is we want to reach the optimal order for the risk without any prior information on f and in particular on its regularity.

The problem of estimating the stationary density of a continuous time process has been mainly studied using kernel estimators by Banon [1], Banon and N'Guyen [2] in a context of diffusion models, by N'Guyen [31] for Markov processes. Under some mixing conditions, their pointwise non-integrated $\mathbb{L}_2$-risk reaches the standard rate of convergence $T^{-2a/(2a+1)}$ when f belongs to the Hölder class C^a and a is known. Later Castellana and Leadbetter [14] proved that, under some specific assumption on the joint density of (X_0, X_τ), the non-integrated quadratic risk could reach the parametric rate T^{-1}. They also checked their assumption for some Gaussian processes. Castellana and Leadbetter's [14] work was a key paper concerning the problem of estimating the marginal distribution of a strictly stationary continuous time process and a lot of works in this direction followed. We refer to Bosq [9, 10], Kutoyants [29], Bosq and Davydov [11], among others, for results of this kind. In the same field, Leblanc [30] studied a weaker form of Castellana and Leadbetter's [14] condition (let us call it [CL]

Keywords and phrases: Non parametric estimation, projection estimator, adaptive estimation, model selection, mixing processes, continuous time, discrete time.

[1] Université Paris V, Laboratoire MAP5, 45 rue des Saints-Pères, 75270 Paris Cedex 06, France;
e-mail: comte@biomedicale.univ-paris5.fr

[2] Université Paris VI, LSTA, 4 place Jussieu, 75252 Paris Cedex 05, France; e-mail: merleve@ccr.jussieu.fr

in the sequel, see Sect. 3.1.2) for some diffusion processes. She built a wavelet estimator of f when f belongs to some general Besov space and proved that its $\mathbb{L}_p$-integrated risk converges at rate T^{-1} as well, provided that the regularity is known and the process is geometrically α-mixing.

In discrete time, the problem of adaptive density estimation has been widely studied in the framework of independent observations. Efromovich [25] adapted to this context a thresholding procedure developed in Efromovich and Pinsker [26]. His method is adaptive over some Sobolev ellipsoids relatively to the $\mathbb{L}_2$-loss. Donoho *et al.* [23] showed then that some local procedure of wavelet thresholding could lead to an adaptive estimator (with optimal rate up to a $\ln(n)$ factor) over a large class of Besov balls. Kerkyacharian *et al.* [27] gave a procedure of global thresholding which is adaptive over the same class and relatively to $\mathbb{L}_p$-loss functions, $p \geq 2$. Then Birgé and Massart [5] proposed a selection of models procedure which allows one to recover the previous results and also to work with general bases. Lastly, Butucea [13] studied the minimax risk of pointwise adaptive estimators of the density based on kernels with automatic bandwidth selection. Again, all the previous works are in an i.i.d. set up. The main contribution on the subject in a weakly dependent framework is the thresholding procedure studied by Tribouley and Viennet [35]. Note that Clémençon [15] also studied a wavelet adaptive density estimator in a context of Markov chains.

We want to show in this paper that we can extend Birgé and Massart's [5] inequalities to a framework of weakly dependent observations. As far as we know, no such procedure has ever been studied for continuous time processes. The method leads to an estimator reaching the optimal rate over some classes of smoothness a of the density function f without requiring a to be known. Note that the present work is done under standard mixing conditions and does not assume that condition [CL] holds: therefore, the rates are not parametric. To scheme the difference between both frameworks, we can say that standard mixing assumptions are concerned with the behavior of the mixing coefficients at large lags (long term dependence) whereas condition [CL] also regulates their behavior near zero (very short term dependence). Since the two problems are essentially different, the study of assumption [CL] is relegated to an other work (Comte and Merlevède [17]), and we focus in the present paper on a framework requiring standard assumptions on the rate of decay of the mixing coefficients at large lags.

In discrete time, we recover with our methodology the results of Tribouley and Viennet [35]: we consider only an $\mathbb{L}_2$-risk whereas they study a general $\mathbb{L}_q$-risk but we work in a general context of model selection when they specifically consider an expansion of the estimator on a particular collection of wavelets bases. Moreover we also study the framework of strongly mixing processes either under some specific assumption on the joint density of (X_0, X_k) or in the general case under a particular mixing condition (namely, geometrical strong mixing, see Sect. 2.1). Finally, we study some Besov spaces requiring in general non linear estimators to reach the optimal rates.

Our results are obtained by gathering the tools for absolutely regular processes developed in Viennet [37] and deduced from Berbee's lemma [4] or the tools for strongly mixing processes developed by Rio [32], and the procedure presented in Birgé and Massart [5] based on Talagrand's [34] inequality.

The paper is organized as follows. Section 2 presents the framework, namely the mixing assumptions, some examples of mixing processes, the definition of the estimators and the procedure of estimation. A sketch of proof is given in order to describe the methodology. Section 3 provides the results when considering regular collections of models and the comments coming herewith. Section 3.1 is devoted to the adaptive estimator in discrete and continuous time under absolute regularity. Section 3.2 studies more specifically the strong mixing case. Section 4 presents some general (non regular) collections of models that allow to study the case of non linear estimators. In Section 5, we give some practical considerations about the penalty. The proofs of the main results are deferred to Section 6.

2. The framework

In all the following we aim at estimating the marginal density f of a strictly stationary discrete time process $(X_i)_{i\in\mathbb{Z}}$ or a continuous time one $(X_\tau)_{\tau\in[0,T]}$, on a given compact set A and we denote $f_A := f\mathbb{1}_A$. Throughout the paper, $[z]$ denotes the integer part of z.

2.1. **Mixing assumptions**

In order to develop our results, we recall some standard definitions (see Doukhan [24], pp. 3-4) concerning different types of dependence. Let $\mathcal{F}_u^v$ be the σ-algebra of events generated by the random variables $\{X_\tau, u \leq \tau \leq v\}$. In the case of discrete time processes, u, τ, v are integers. A strictly stationary process $\{X_\tau\}$ is called strongly mixing or α-mixing (Rosenblatt [33]) if

$$\sup_{A\in\mathcal{F}_{-\infty}^0, B\in\mathcal{F}_\tau^\infty} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| = \alpha_\tau \to 0 \text{ as } \tau \to +\infty.$$

It is said to be absolutely regular or β-mixing (Kolmogorov and Rozanov [28]) if

$$\frac{1}{2}\sup\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(U_i\cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)|\right\} = \beta_\tau \to 0 \text{ as } \tau \to +\infty,$$

where the above supremum is taken over all finite partitions $(U_i)_{1\leq i\leq I}$ and $(V_j)_{1\leq j\leq J}$ of Ω respectively $\mathcal{F}_{-\infty}^0$ and $\mathcal{F}_\tau^\infty$ measurable.

The following relation holds: $2\alpha_\tau \leq \beta_\tau \leq 1$.

In the sequel, the processes of interest are either α-mixing or β-mixing with α-mixing coefficients α_τ or β-mixing coefficients β_τ. We define A_r and B_r as:

$$A_r := \int_0^{+\infty} s^{r-2}\beta_s \mathrm{d}s, \qquad B_r := \sum_{l\in\mathbb{N}} (l+1)^{r-2}\beta_l, \tag{2.1}$$

when the integral (or the series) is convergent. Besides we consider two kinds of rates of convergence to 0 of the mixing coefficients, that is for $\gamma = \alpha$ or β:

[AR] arithmetical γ-mixing with rate θ: there exists some $\theta > 0$[3] such that $\gamma_\tau \leq (1+\tau)^{-(1+\theta)}$ for all τ in $\mathbb{N}$ or $\mathbb{R}$;

[GEO] geometrical γ-mixing with rate θ: there exists some $\theta > 0$ such that $\gamma_\tau \leq \mathrm{e}^{-\theta\tau}$ for all τ in $\mathbb{N}$ or $\mathbb{R}$.

2.2. **Examples of mixing processes**

Let us give two simple examples of processes widely considered in the literature and which are stationary and mixing.

(1) A discrete time general autoregressive model

$$X_{i+1} = g(X_i) + \varepsilon_{i+1}, \ \ \varepsilon_i \ \text{ i.i.d.}, \quad \mathbb{E}(\varepsilon_1) = 0, \ \mathbb{E}(\varepsilon_1^2) = \sigma^2,$$

admits a stationary law provided that X_0 is independent of ε_1 and there exist $b, c > 0$ and $0 < a < 1$ such that $|g(x)| \leq a|x| - b$ when $|x| > c$. This law admits a density, say f, with respect to the Lebesgue measure, as soon as ε_1 does. If the density of X_0 is f, then the process is strictly stationary and geometrically β-mixing (see Doukhan [24], p. 102).

(2) A continuous time diffusion process is defined as the solution of a homogeneous stochastic differential equation:

$$dX_\tau = m(X_\tau)\mathrm{d}\tau + \sigma(X_\tau)dW_\tau, \ \ \tau \geq 0$$

where $(W_\tau, \mathcal{A}_\tau)$ is a standard Brownian motion on some complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with an increasing family of complete σ-algebra $\mathcal{A}_\tau$. Standard conditions on m and σ ensuring that the process is strictly stationarity and geometrically β-mixing are given in Veretennikov [36]. Leblanc [30] also gives conditions for this process to be geometrically α-mixing.

[3] We take $\theta > 0$ so that A_2 and B_2 are finite.

2.3. **Collections of models**

We consider here collections of models $(S_m)_{m\in\mathcal{M}_n}$ for which we assume that the following standard assumptions are fulfilled:

[M1] for each m in $\mathcal{M}_n$, S_m is a linear subspace of $\mathbb{L}_2(A)$ with dimension D_m and $N_n = \max_{m\in\mathcal{M}_n} D_m$ satisfies $N_n \leq n$;

[M2] there exists a constant Φ_0 such that:

$$\forall m, m' \in \mathcal{M}_n, \forall t \in S_m, \forall t' \in S_{m'}, \|t + t'\|_\infty \leq \Phi_0 \sqrt{\dim(S_m + S_{m'})}\|t + t'\|;$$

[M3] for any positive a, $\sum_{m\in\mathcal{M}_n} \sqrt{D_m}\mathrm{e}^{-a\sqrt{D_m}} \leq \Sigma(a)$, where $\Sigma(a)$ denotes a finite constant depending only on a.

As a consequence of [M2], for any orthonormal basis $(\varphi_\lambda)_{\lambda\in\Lambda}$ of $S_m + S_{m'}$, we have

$$\left\|\sum_{\lambda\in\Lambda} \varphi_\lambda^2\right\|_\infty = \sup_{t\in S_m+S_{m'}, t\neq 0} \frac{\|t\|_\infty^2}{\|t\|^2} \tag{2.2}$$

(see Barron *et al.* [3], Eqs. (3.2) and (3.3)). Three examples are usually developed as fulfilling this set of assumptions:

[T] trigonometric spaces: S_m is generated by 1, $\cos(2\pi jx), \sin(2\pi jx)$ for $j = 1, \ldots, m$ and $A = [0, 1]$, $D_m = 2m + 1$ and $\mathcal{M}_n = \{1, \ldots [n/2] - 1\}$;

[P] regular piecewise polynomial spaces: S_m is generated by r polynomials of degree $0, 1, \ldots, r - 1$ on each subinterval $[(j - 1)/m, j/m]$, for $j = 1, \ldots m$, $D_m = rm$ when $A = [0, 1]$, $m \in \mathcal{M}_n = \{1, 2, \ldots, [n/r]\}$;

[W] dyadic wavelet generated spaces as described *e.g.* in Donoho and Johnstone [22], with regularity r.

These examples describe regular collections of models. For a precise description of those spaces and their properties, we refer also to Birgé and Massart [5] and to Barron *et al.* [3].

2.4. **The estimators**

The superscript ^{c} (resp. subscript $_c$) is for quantities related to the continuous time process and the superscript ^{d} (resp. subscript $_d$) for the discrete time one. We consider the following contrast functions, for t belonging to some S_m of a collection $(S_m)_{m\in\mathcal{M}_n}$ where $n = [T]$ for $(X_\tau)_{\tau\in[0,T]}$ in continuous time and n is the number of observations for $(X_i)_{1\leq i\leq n}$ in discrete time:

$$\gamma_n^c(t) = \|t\|^2 - \frac{2}{T}\int_0^T t(X_s)\mathrm{d}s, \quad \text{and} \quad \gamma_n^d(t) = \frac{1}{n}\sum_{i=1}^n \left[\|t\|^2 - 2t(X_i)\right],$$

where we recall that $\|t\|^2 = \int_A t^2(x)\mathrm{d}x$. Note that $\mathbb{E}(\gamma_n^d(t)) = \mathbb{E}(\gamma_n^c(t)) = \|t - f\|^2 - \|f\|^2 = \|t - f_A\|^2 - \|f_A\|^2$ is minimal when $t = f$. Then the estimators are built as follows. Let

$$\hat{f}_m^c = \text{ Argmin }_{t\in S_m}\gamma_n^c(t), \quad \text{or} \quad \hat{f}_m^d = \text{ Argmin }_{t\in S_m}\gamma_n^d(t)$$

be a collection of estimators of f. Then if $(\varphi_j)_{1\leq j\leq D_m}$ is an orthonormal basis of S_m, we have

$$\hat{f}_m^c = \sum_{j=1}^{D_m} \hat{a}_j^c\varphi_j \quad \text{with } \hat{a}_j^c = \frac{1}{T}\int_0^T \varphi_j(X_s)\mathrm{d}s,$$

and

$$\hat{f}_m^d = \sum_{j=1}^{D_m} \hat{a}_j^d\varphi_j \quad \text{with } \hat{a}_j^d = \frac{1}{n}\sum_{i=1}^n \varphi_j(X_i).$$

Moreover if f_m is the $\mathbb{L}_2$-orthogonal projection of f on S_m, then

$$f_m = \sum_{j=1}^{D_m} a_j(f)\varphi_j \quad \text{with } a_j(f) = \mathbb{E}(\hat{a}_j^d) = \mathbb{E}(\hat{a}_j^c) = \langle f, \varphi_j \rangle \cdot$$

Let us now define the centered empirical processes:

$$\nu_n^c(t) = \frac{1}{T}\int_0^T [t(X_s) - \langle t, f\rangle]\,\mathrm{d}s \quad \text{and} \quad \nu_n^d(t) = \frac{1}{n}\sum_{i=1}^{n} [t(X_i) - \langle t, f\rangle] \cdot$$

In the following, we do not use any superscript nor subscript when no distinction is required.

From the definition of $\hat{f}_m$, it follows that $\gamma_n\ (\hat{f}_m) \leq \gamma_n\ (f_m)$. This together with

$$\gamma_n\ (t) - \gamma_n(f_A) = \|t - f_A\|^2 - 2\nu_n\ (t - f_A) \tag{2.3}$$

entail that

$$\|f_A - \hat{f}_m\|^2 \leq \|f_A - f_m\|^2 + 2\nu_n\ (\hat{f}_m - f_m)\,.$$

Moreover

$$\nu_n(\hat{f}_m - f_m) = \sum_{j=1}^{D_m} (\hat{a}_j - a_j(f))\nu_n(\varphi_j) = \sum_{j=1}^{D_m} [\nu_n\ (\varphi_j)]^2$$

since $\hat{a}_j - a_j(f) = \nu_n(\varphi_j)$. Under some mixing conditions we obtain that $\mathbb{E}((\nu_n)^2(\varphi_j))$ has the same order as in the independent case and is less than C/n, where C is a constant. Therefore we have

$$\mathbb{E}(\|\hat{f}_m - f_A\|^2) \leq \|f_A - f_m\|^2 + \frac{2CD_m}{n}$$

and we can see that we have the standard squared bias plus variance decomposition: $\|f_A - f_m\|^2 + 2CD_m/n$, that naturally appears in both discrete and continuous time frameworks. In order to minimize the quadratic risk $\mathbb{E}(\|\hat{f}_m - f_A\|^2)$, we need to select the model $m \in \mathcal{M}_n$ that makes $\|f_A - f_m\|^2 + 2\sum_{j=1}^{D_m} \mathrm{Var}(\nu_n\ (\varphi_j))$ as small as possible. This choice is performed by the following penalization procedure:

$$\tilde{f}^c = \hat{f}^c_{\hat{m}_c} \text{ with } \hat{m}_c = \ \mathrm{Argmin}_{\ m\in\mathcal{M}_n} \left[\gamma_n^c(\hat{f}_m^c) + \ \mathrm{pen}_c(m)\right] \tag{2.4}$$

or

$$\tilde{f}^d = \hat{f}^d_{\hat{m}_d} \text{ with } \hat{m}_d = \ \mathrm{Argmin}_{\ m\in\mathcal{M}_n} \left[\gamma_n^d(\hat{f}_m^d) + \ \mathrm{pen}_d(m)\right] \tag{2.5}$$

where pen_c and pen_d are penalty functions defined in the theorems and given by the theory. The penalty function prevents from the systematic choice of the largest space S_m of the collection and ensures the automatic bias-variance compromise for the estimate.

2.5. **Sketch of proof**

From the definition of $\tilde{f}$, it follows that for all m in $\mathcal{M}_n$,

$$\gamma_n\ (\tilde{f}) + \mathrm{pen}(\hat{m}) \leq \gamma_n\ (f_m) + \mathrm{pen}(m).$$

The above inequality together with (2.3) yield

$$\|f_A - \tilde{f}\|^2 \leq \|f_A - f_m\|^2 + 2\nu_n\ (\tilde{f} - f_m) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}). \tag{2.6}$$

Note that, equation (2.6) holds for any f_m in S_m, but the relevant choice is to take the element of S_m that makes $\|f_A - f_m\|^2$ minimum. Moreover, inequality (2.6) explains why we need to study the process ν_n. More precisely, denoting by $S^*_{m'}$ the set $\{t \in S_{m'}, \|t - f_m\| \neq 0\}$ and by $B_{m,m'}(0,1) = \{t \in S_m + S_{m'} \ / \ \|t\| = 1\}$, we successively write:

$$
\begin{aligned}
& 2|\nu_n(\tilde f - f_m)| \le 2\|\tilde f - f_m\| \sup_{t\in S^*_{\hat m}} \frac{|\nu_n(t-f_m)|}{\|t-f_m\|} \\
\le\ & \frac14\|\tilde f - f_m\|^2 + 4\left[\left(\sup_{t\in S^*_{\hat m}} \frac{|\nu_n(t-f_m)|}{\|t-f_m\|}\right)^2 - p(m,\hat m)\right]_+ + 4p(m,\hat m) \\
\le\ & \frac12\|f_m - f_A\|^2 + \frac12\|f_A - \tilde f\|^2 + 4\sum_{m'\in\mathcal{M}_n}\left[\left(\sup_{t\in S^*_{m'}} \frac{|\nu_n(t-f_m)|}{\|t-f_m\|}\right)^2 - p(m,m')\right]_+ \\
& +4p(m,\hat m) \\
\le\ & \frac12\|f_m - f_A\|^2 + \frac12\|f_A - \tilde f\|^2 + 4\sum_{m'\in\mathcal{M}_n}\left[\left(\sup_{t\in B_{m,m'}(0,1)} |\nu_n(t)|\right)^2 - p(m,m')\right]_+ \\
& +4p(m,\hat m) \\
\le\ & \frac12\|f_m - f_A\|^2 + \frac12\|f_A - \tilde f\|^2 + 4\sum_{m'\in\mathcal{M}_n} W(m') + 4p(m,\hat m), \qquad (2.7)
\end{aligned}
$$

where

$$
W(m') := \left[\left(\sup_{t\in B_{m,m'}(0,1)} |\nu_n(t)|\right)^2 - p(m,m')\right]_+ . \qquad (2.8)
$$

The aim of the proofs is to find $p(m,m')$ such that

$$
\sum_{m'\in\mathcal{M}_n} \mathbb{E}(W\ (m')) \le C/n, \qquad (2.9)
$$

where C is a constant. This allows to choose the penalty such that $4p(m,\hat m) \le \mathrm{pen}(m) + \mathrm{pen}(\hat m)$ whence $4p(m,\hat m) + \mathrm{pen}(m) - \mathrm{pen}(\hat m) \le 2\mathrm{pen}(m)$. Then (2.6) and (2.7) imply that, for all m in $\mathcal{M}_n$

$$
\mathbb{E}\|f_A - \tilde f\|^2 \le 3\|f_A - f_m\|^2 + \frac{8C}{n} + 4\mathrm{pen}(m)
$$

that is

$$
\mathbb{E}\|f_A - \tilde f\|^2 \le 4 \inf_{m\in\mathcal{M}_n}\left[\|f_A - f_m\|^2 + \mathrm{pen}(m)\right] + \frac{8C}{n}\cdot \qquad (2.10)
$$

If the penalty has the standard order D_m/n for variance terms in density estimation, then equation (2.10) guarantees an automatic trade-off between the bias term $\|f_A - f_m\|^2$ and the variance term, up to some multiplicative constant.

The principle of the control of $\mathbb{E}(W\ (m'))$ in order to obtain (2.9) is the same in all cases. Talagrand's inequality [34] allows to deal with the supremum of the empirical centered process in an independent set up (see Birgé and Massart [5]). Berbee's [4] coupling lemma, Delyon's [20] covariance inequality and Bryc's [12]

construction of approximating variables allow to deal with absolute regular dependence; a construction of approximating variables due to Rio [32] allows analogously to deal with strong mixing dependence.

The difficulties arise because of the "twice" random aspect of $\tilde{f}$, *i.e.* $\tilde{f}$ admits a decomposition on an orthonormal basis, but with a random number of random coefficients.

The aim of the study below is to prove a result of type (2.10) under relevant assumptions and for a relevant choice of the penalty function.

3. Results for regular collections of models

3.1. **Absolutely regular processes**

3.1.1. *General result*

We start with the most powerful results: they are obtained in the context of absolutely regular processes, when considering regular collections of models. We emphasize that absolute regularity allows a better control of the terms than strong mixing.

Theorem 3.1. *Consider a collection of models satisfying* [M1–M3] *with* $n = [T]$ *and* $|\mathcal{M}_n| \leq n^\epsilon$, *where* ϵ *is a positive number. Assume that the process* $(X_\tau)_{\tau\in[0,T]}$ *or* $(X_k)_{1\leq k\leq n}$ *is strictly stationary and arithmetically* [AR] β*-mixing with mixing rate* θ *and that its marginal distribution admits a density* f *with respect to the Lebesgue measure on* $\mathbb{R}$, *with* $\|f_A\|_\infty < \infty$. *Then the estimator* $\tilde{f}$ *defined as*

[Cβ] $\tilde{f} = \tilde{f}^c = \hat{f}^c_{\hat{m}_c}$ *as given in (2.4) with* $\mathrm{pen}_c(m) = \kappa\Phi_0^2 A_2 D_m/n$ *where* κ *is a universal constant, provided that* $\theta > 2\epsilon + 3$,
[Dβ] $\tilde{f} = \tilde{f}^d = \hat{f}^d_{\hat{m}_d}$ *as given in (2.5) with* $\mathrm{pen}_d(m) = \kappa\Phi_0^2 B_2 D_m/n$, *where* κ *is a universal constant, provided that* $\theta > 3$,

satisfies

$$\mathbb{E}(\|\tilde{f} - f_A\|^2) \leq \inf_{m\in\mathcal{M}_n}\left(3\|f_A - f_m\|^2 + C\frac{D_m}{n}\right) \tag{3.1}$$

where C *is a constant depending on terms among* Φ_0, θ, A_2, A_3 *(or* B_2, B_3*) and* $\|f_A\|_\infty$.

Note that geometrical mixing [GEO] implies the arithmetical one [AR] with θ as large as desired. Therefore, no constraint on θ would appear in that case.

Remark 3.1. Any bound on the penalty can be taken as a penalty. In that case, equation (3.1) holds with CD_m/n replaced by the new penalty. Under the assumptions of Theorem 3.1 and in case [Cβ], we have $A_2 \leq 1/\theta < 1/3$. Therefore, the estimator $\tilde{f}^c$ based on the penalty $\mathrm{pen}_c(m) = (\kappa/3)\Phi_0^2 D_m/n$ would lead to (3.1) as well. This penalty does not depend on the mixing coefficients. This is due to the particular structure of assumption [AR] which assumes a bound on the mixing coefficients β_t for all t and not only for $t \geq t_0$. Similarly, in case [Dβ], $B_2 \leq 1 + A_2 \leq (4/3)$ leads to a penalty $\mathrm{pen}_d(m) = (4\kappa/3)\Phi_0^2 D_m/n$ which, under the assumptions of Theorem 3.1, gives an estimator $\tilde{f}^d$ satisfying (3.1).

Remark 3.2. In the examples of regular collections detailed above, the constraint $|\mathcal{M}_n| \leq n^\epsilon$ is fulfilled. More precisely, dyadic collections correspond to $|\mathcal{M}_n|$ of order $\ln(n)$, and the constraint on θ becomes $\theta > 3$ for [Cβ]. Regular collections with $|\mathcal{M}_n| = n$ give $\theta > 5$ in case [Cβ]. Note that in case [Dβ], the constraint on θ is weaker than in case [Cβ], and does not depend on $|\mathcal{M}_n|$.

Remark 3.3. In the discrete time case, under the assumptions of Theorem 3.1 and if moreover $N_n \leq n^\omega$ for some constant ω in $[0, 1]$, then (3.1) holds for the same $\tilde{f}^d$ as soon as $\theta > 2\omega + 1$. This condition is obviously implied for any $\omega \in [0, 1]$ by $\theta > 3$.

3.1.2. *About condition* [CL]

In view of the results of Theorem 3.1, we can make the following remark. Assume that one can observe a stationary mixing process either in continuous time on $[0,n]$ or in discrete time for $t=1,\dots,n$. Under our assumption, namely under β-mixing, there is no loss in the global rate when one considers only the discrete time observations, and there is no gain to use a continuous time observation. This may be surprising but can be explained as follows: the class of continuous time β-mixing processes contains the class of discrete time β-mixing processes. To see this, consider $(X_i)_{i\in\mathbb{N}}$ a discrete time β-mixing process; then we can build a continuous time β-mixing process by simply setting $X_t = X_{[t]}$ for $t\in\mathbb{R}_+$.

It follows from this remark that, if we want the continuous time process to reach a better rate, we need to introduce an assumption taking into account what happens on small intervals of time. This kind of local behavior is governed for instance by Castellana and Leadbetter's [14] condition which in the weaker version used by Leblanc [30] can be written as follows:

[CL] there exists a positive integrable and bounded function $h(.)$ (defined on $\mathbb{R}$) such that

$$\forall x\in\mathbb{R},\ \sup_{y\in\mathbb{R}}\int_0^{+\infty}|f_\tau(x,y)-f(x)f(y)|\mathrm{d}\tau\le h(x),$$

where $f_\tau(x,y)$ is the density distribution of (X_0,X_τ) with respect to the Lebesgue measure on $\mathbb{R}^2$.

This condition hides in fact two different types of control on the process. On the one hand, the convergence condition near infinity concerns the long term behavior of the process and is of the same nature as our present mixing conditions. On the other hand, the convergence of the integral near of zero, represents an assumption of locally irregular paths of the process[4], and can lead to parametric rates for continuous time processes observed in continuous time. The study of this condition, its links with mixing conditions and its implications on the rate of convergence of an estimator based on discrete time observations of the process are developed in Comte and Merlevède [17]. We also refer to Bosq [10].

3.1.3. *Adaptation to unknown smoothness*

Inequalities as (3.1) are known to lead to results of adaptation to unknown smoothness. Take $A=[0,1]$ for simplicity. We first recall that a function f belongs to the Besov space $\mathcal{B}_{a,l,\infty}([0,1])$ if it satisfies

$$|f|_{a,l}=\sup_{y>0}y^{-a}w_d(f,y)_l<+\infty,\quad d=[a]+1,$$

where $w_d(f,y)_l$ denotes the modulus of smoothness. For a precise definition of those notions we refer to DeVore and Lorentz [21] (Chap. 2, Sect. 7), where it also proved that $\mathcal{B}_{a,p,\infty}([0,1])\subset\mathcal{B}_{a,2,\infty}([0,1])$ for $p\ge 2$. This justifies that we now restrict our attention to $\mathcal{B}_{a,2,\infty}(A)$.

Proposition 3.1. *Consider the collection of models* [T], [P] *or* [W]*, with $r>a>0$ and with $n=[T]$. Assume that an estimator $\tilde f$ of f satisfies inequality (3.1). Let $L>0$ and $K>0$. Then*

$$\left(\sup_{f\in\mathbb{B}_{a,2,\infty}(L),\|f_A\|_\infty\le K}\mathbb{E}\|f_A-\tilde f\|^2\right)^{\frac12}\le C(a,L,K)n^{-\frac{a}{2a+1}}\tag{3.2}$$

where $\mathbb{B}_{a,2,\infty}(L)=\{t\in\mathcal{B}_{a,2,\infty}(A),|t|_{a,2}\le L\}$ where $C(a,L,K)$ is a constant depending on a,L,K and also on A_2, A_3 (or B_2,B_3) and θ.

Proof. The result is a straightforward consequence of the results of DeVore and Lorentz [21] and of Birgé and Massart [5] which imply that $\|f_A-f_m\|$ is of order D_m^{-a} in the three collections, for any positive a. Thus the infimum in (3.1) is reached for a model m^* with $D_{m^*}=[n^{1/(1+2a)}]$, which is less than n for $a>0$. Then we

[4]In other words, this condition means that the sample paths of the process are not smooth.

find from (3.1) the standard non parametric rate of convergence $n^{-2a/\ (1+2a)}$ for f in any $\mathcal{B}_{a,2,\infty}$ for any $a > 0$, provided that f_A is bounded. □

Remark 3.4. The supremum norm of f_A is uniformly bounded on the Besov ball $\mathbb{B}_{a,2,\infty}(L)$ if $a > 1/2$ so that the assumption $\|f_A\|_\infty < +\infty$ is automatically fulfilled if we assume $a > 1/2$.

Those rates are known to be minimax

- for continuous time estimators, at least when dealing with non-integrated mean square risk and α-mixing sequences (see Bosq [10]);
- for discrete time estimators (even) in the independent set up, with respect to the mean square integrated risk, see Donoho *et al.* [23].

Remark 3.5. Under the assumptions of Theorem 3.1, it follows from Remark 3.3 that (3.2) holds for $f \in \mathcal{B}_{a,2,\infty}(A)$ with $a > 1/2$ and for arithmetical β-mixing [AR] with $N_n \leq \sqrt{n}$ and $\theta > 2$. Indeed the selected model m^* has dimension $D_{m^*} = [n^{1/(1+2a)}]$, which is less than $\sqrt{n}$ for $a > 1/2$. Then we find the rate $n^{-a/(1+2a)}$ for arithmetical β-mixing if $\theta > 2$ ($\omega = 1/2$).

Tribouley and Viennet [35] give the same kind of result as in Proposition 3.1 but without the general bound given in Theorem 3.1. They specifically work with a dyadic collection of wavelet spaces. In that sense, our approach generalizes their work with less technical computations. On the other hand, they deal with a general $\mathbb{L}_q$-risk instead of our specific $\mathbb{L}_2$-risk. It follows from Remark 3.5 that, for $q = 2$, we reach the same mixing condition $\theta > 2$ as them.

3.2. **The strongly mixing case**

3.2.1. *A particular result under an additional assumption*

If we want to deal with the strongly mixing case in discrete time and keep standard orders, we need a further assumption:

[Lip] let $g_{|\tau-\tau'|} = f_{(X_\tau, X_{\tau'})} - f \otimes f$, $\tau \neq \tau'$. We assume that

$$|g_{|\tau-\tau'|}(z') - g_{|\tau-\tau'|}(z)| \leq \ell_{|\tau-\tau'|}\|z - z'\|_{\mathbb{R}^2}, \quad \text{for all } z, z' \in \mathbb{R}^2, \tag{3.3}$$

for some ℓ depending on $|\tau - \tau'|$ only, and that $S(\alpha, \ell) := \sum_{k=1}^{\infty} (1 + \ell_k)\alpha_k^{1/3} < +\infty$.

Here for $x = (x_1, x_2) \in \mathbb{R}^2$, we denote by $\|x\|_{\mathbb{R}^2} = \sqrt{x_1^2 + x_2^2}$ the Euclidean norm.

Assumption [Lip] is also used in Bosq [10] (see Assumption H_2, p. 43) under the stronger form where $\ell_{|\tau-\tau'|} \equiv \ell$. In that case, the condition $\sum_k (1 + \ell_k)\alpha_k^{1/3} < +\infty$ becomes $\sum_k \alpha_k^{1/3} < +\infty$, which requires $\theta > 3$ for arithmetical mixing.

As an illustration, if we consider the trivial example of a Gaussian stationary process $(X_k)_{k\in\mathbb{Z}}$, then we find $\ell_k = C/\sqrt{1 - \rho^2(k)}$ where C is a numerical constant and $\rho(k) = \mathrm{cov}(X_0, X_k)/\mathrm{var}(X_0)$ is the autocorrelation function. Therefore ℓ_k is bounded by some ℓ as soon as for instance $\rho(k) \to 0$ when $k \to +\infty$.

Proposition 3.2. *Consider a collection of models satisfying* [M1–M3] *with $n = [T]$ and $|\mathcal{M}_n| \leq n^\epsilon$, where ϵ is a positive number. Assume that $(X_k)_{1\leq k\leq n}$ is strictly stationary and arithmetically* [AR] *α-mixing with mixing rate θ and that its marginal distribution admits a density f with respect to the Lebesgue measure on $\mathbb{R}$, with $\|f_A\|_\infty < \infty$. Assume moreover that* [Lip] *is satisfied. Then, if $\theta > 2\epsilon + 5$, the estimator $\tilde{f}^d$ defined by $\tilde{f}^d = \hat{f}^d_{\hat{m}_d}$ as given in (2.5) with $\mathrm{pen}_d(m) = C(\Phi_0, S(\alpha, \ell))D_m/n$ satisfies inequality (3.1). A suitable choice of C is $\kappa[\Phi_0^2 + 2\sqrt{2}m(A)S(\alpha, \ell)]$, where κ is a numerical constant and $m(A)$ denotes the Lebesgue measure of A.*

As previously, if $|\mathcal{M}_n|$ is of order $\ln(n)$ then the constraint on θ amounts to $\theta > 5$. If $|\mathcal{M}_n|$ is of order n, the constraint can be written $\theta > 7$. In the case where $N_n \leq n^\omega$, for $\omega \in [0, 1]$, we can write $\theta > 2\epsilon + 4\omega + 1$.

Remark 3.6. Here, the constant involved in the penalty is more complicated than in the β-mixing case: in particular, it is not possible to give an upper bound on it as in Remark 3.1. Nevertheless, the strategy explained in Section 5 below may be applied.

The main problem is to know whether [Lip] is fulfilled or not. In the case where [Lip] does not hold, the next result shows that we do not necessarily keep the same rates for any type of mixing rates.

3.2.2. *The general α-mixing case*

In the α-mixing case, if assumption [Lip] is not fulfilled, the result is much less powerful and requires a stronger constraint on the rate of the mixing. We give a result in discrete time but an analog result would hold in continuous time.

Theorem 3.2. *Consider a collection of models satisfying* [M1–M3] *and $|\mathcal{M}_n| \leq n^\epsilon$ where ϵ is a positive number. Assume that $(X_k)_{1\leq k\leq n}$ is strictly stationary and geometrically* [GEO] *α-mixing with mixing rate θ and that its marginal distribution admits a density f with respect to the Lebesgue measure on $\mathbb{R}$, such that $\|f_A\|_\infty < \infty$. Then the estimator $\tilde{f}^d$ defined by (2.5) with*

$$\mathrm{pen}_d(m) = \kappa \frac{(\epsilon+3)\Phi_0^2}{\theta} \frac{\ln(n) D_m}{n}$$

where κ is a universal constant, satisfies:

$$\mathbb{E}(\|\tilde{f}^d - f_A\|^2) \leq \inf_{m\in\mathcal{M}_n} \left(3\|f_A - f_m\|^2 + C \frac{\ln(n) D_m}{n} \right) \tag{3.4}$$

where C is a constant depending on Φ_0, θ, $\|f_A\|_\infty$.

Note that this bound implies a loss of $\ln(n)$ with respect to the minimax rate for Besov spaces $\mathcal{B}_{a,2,\infty}$; namely, $n^{-2a/(2a+1)}$. More precisely, for $f \in \mathcal{B}_{a,2,\infty}(A)$, the optimal choice $D_{m^*} = [(n/\ln(n))^{1/(2a+1)}]$ gives a rate $(n/\ln(n))^{-2a/(2a+1)}$. Moreover the result holds for regular spaces only and under geometrical strong mixing condition.

4. A result for general collections of models

The previous section shows that when we consider the $\mathbb{L}_2$-risk of the adaptive projection estimator of a function f assumed to belong to some Besov space $\mathcal{B}_{a,p,\infty}$ with $a > 0$ and $p \geq 2$, then regular collections of models lead to the standard rate of convergence. It is well-known that when $1 < p < 2$, this does not remain valid. Indeed, the bias term $\|f_A - f_m\|$ does not have the right order, namely D_m^{-a}. To recover this rate, one needs to consider general collections of models (typically non regular subdivisions of the interval A; for a concise presentation of the problem, see Birgé and Massart [7] and the references therein). Unfortunately, these general collections of models do not satisfy assumption [M2] any more, but only the following:

[M'2] $\forall m, m' \in \mathcal{M}_n, \forall t \in S_m$ and $t' \in S_{m'}$, $\|t + t'\|_\infty \leq \Phi_0 \sqrt{N_n} \|t + t'\|$.

Besides [M3] does not hold either and has to be replaced by

$$\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-L_m D_m} \leq \Sigma, \tag{4.1}$$

where Σ is a finite constant and the L_m's are suitable weights.

For sake of simplicity, we consider $A = [0, 1]$ and we only describe the general collection of piecewise polynomials (up to some constants, the result would hold for general wavelets).

[GP] We characterize the linear space S_m of piecewise polynomials of degree (strictly) less than r by $m = (d, \{b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1\})$, where $d \in \{1, \ldots, J_n\}$ and the b_j's define a partition of the interval $[0,1]$ into d intervals based on dyadic knots *i.e.* for all $j \in \{1, \ldots, d-1\}$, b_j is of the form $N_j/2^{J_n}$ with $N_j \in \mathbb{N}$. We define by $\mathcal{M}_n$ the set of all possible m of this form. Therefore, for any increasing sequence of dyadic knots $b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1$ with $d \in \{1, \ldots, J_n\}$, there exists $m \in \mathcal{M}_n$ such that $m = (d, \{b_0 = 0 < b_1 < \cdots < b_{d-1} < b_d = 1\})$ and for all $t \in S_m$ we have

$$t(x) = \sum_{j=1}^{d} P_j(x) \mathbb{1}_{[b_{j-1}, b_j[}(x),$$

where the P_j's are polynomials of degree less or equal than $r-1$. Note that $\dim(S_m) = rd$. We define the linear space $\mathcal{S}_n$ by choosing $d = 2^{J_n}$ and $b_j = j/2^{J_n}$ for $j = 0, \ldots, 2^{J_n}$. Since $\dim(\mathcal{S}_n) = r2^{J_n} := N_n$, we impose the natural constraint $r2^{J_n} \leq n$. We denote by $(\varphi_\lambda)_{\lambda \in \Lambda_m}$ and $(\varphi_\lambda)_{\lambda \in \Lambda_n}$ orthonormal bases of S_m and $\mathcal{S}_n$ respectively.

Since for each $d \in \{1, \ldots, J_n\}$,

$$|\{m \in \mathcal{M}_n / \ m_1 = d\}| = C_{2^{J_n}-1}^{d-1} \leq C_{2^{J_n}}^{d}$$

it follows that

$$\begin{aligned} \sum_{m \in \mathcal{M}_n} \mathrm{e}^{-L_m D_m} &\leq \sum_{d=1}^{2^{J_n}} C_{2^{J_n}}^{d} \mathrm{e}^{-\ln(n/r)d} \leq (1 + \exp(-\ln(n/r)))^{2^{J_n}} \\ &\leq \exp(n/r \exp(-\ln(n/r))) = e \end{aligned}$$

using that $2^{J_n} \leq n/r$. Therefore, equation (4.1) holds with $L_m = \ln(n/r)/r$ and $\Sigma = e$. For comparison, in regular collections some constant weights $L_m = L$ would be enough to ensure that $\sum_m \mathrm{e}^{-L_m D_m}$ remains bounded.

Now we can state our result which is specific to discrete time processes.

Theorem 4.1. *Consider the collection of models* [GP] *with maximal dimensions* $N_n \leq n/\ln(n)^3$. *Assume that the process* $(X_i)_{1 \leq i \leq n}$ *is strictly stationary and geometrically* [GEO] β*-mixing with rate* θ *and that its marginal distribution admits a density* f *with respect to the Lebesgue measure on* $\mathbb{R}$, *with* $\|f_A\|_\infty \leq K < +\infty$. *Let* $p > 1$ *and* $r > a > (1/p - 1/2)_+$ *and assume that* f *belongs to some Besov space* $\mathcal{B}_{a,p,\infty}(A)$. *If* $\tilde{f}^d$ *is defined by (2.5) with*

$$\mathrm{pen}_d(m) = \frac{\kappa K}{\theta} \left(1 + \frac{K}{\theta}\right) \frac{\ln^2(n) D_m}{n},$$

where κ *is a numerical constant, then*

$$\left(\sup_{f \in \mathbb{B}_{a,p,\infty}(L), \|f_A\|_\infty \leq K} \mathbb{E} \|f_A - \tilde{f}^d\|^2 \right)^{\frac{1}{2}} \leq C(a, L, K, r, \theta) \left(\frac{n}{\ln^2(n)} \right)^{-\frac{a}{2a+1}} \tag{4.2}$$

where $\mathbb{B}_{a,2,\infty}(L) = \{t \in \mathcal{B}_{a,p,\infty}(A), |t|_{a,p} \leq L\}$.

The proof of the above theorem involves tools used by Birgé and Massart [5] in the framework of independent observations. Roughly speaking, the penalty required in the case of geometrically β-mixing processes amounts to the one obtained in the independent case multiplied by $\ln(n)/\theta$ (see Prop. 4 in Birgé and Massart [5]).

Note that the penalty depends on two unknown quantities, θ and $\|f_A\|_\infty$. The latter can be replaced by a suitable estimator, the penalty becomes then random: for instance, Birgé and Massart [5] propose to use the infinite norm of the projection estimator of f on a regular space S_m of the collection for a well chosen m depending on n only. Clearly, this would also suit in our case. Concerning the constant θ coming from the mixing assumption, we refer to Section 5 below.

Concerning now the rate of convergence, the following comment can be made. In the framework of independent observations, Birgé and Massart [5] obtain the rate $(n/\ln(n))^{-a/2a+1}$. But is is clear from Birgé and Massart [7] that a suitable algorithm allows to recover the standard rate $n^{-a/2a+1}$. Their strategy targets to consider a restricted collection of irregular models in order to allow for constant weights L_m's, but a collection still large enough, to keep the same quality for the approximation $\|f_A - f_m\|$. This collection of models could be used here and would lead to the rate $(n/\ln(n))^{-a/2a+1}$. But the standard rate can not be recovered because of the methodology that we use to deal with absolutely regular processes.

5. Some practical considerations about the penalty

Recall that A_2 and B_2 are defined in (2.1) and depend on the mixing coefficients. Therefore, all the penalties found in the framework of mixing processes are of interest for their orders which remain the same as in the independent set up. Nevertheless, they always depend on the mixing coefficients, or, as explained in Remark 3.1, on the fact that assumption [AR] holds for all t.

From a practical point of view, one needs to know what to do with a given data set. Here, two solutions can be found: either a substitution of the unknown deterministic term by a random one for which we can give some justifications (a complete proof is beyond the scope of the present paper) or a method which has been used is several works even when the terms in the penalty are much easier to estimate (see Birgé and Rozenholc [8], Comte and Rozenholc [18]).

Let us be more precise and start by the second (practical) solution. The standard strategy is as follows. Consider for simplicity the collection of models associated with regular histograms and a discrete time process. Then the collection of models can be parameterized by the dimension of the model: $\mathcal{M}_n = \{D = 1, \dots, n\}$, $S_m = S_D$ where $\dim S_D = D$ and $\hat{f}_m$ can simply be denoted by $\hat{f}_D$. One can then compute $F(D, c) = \gamma_n^d(\hat{f}_D) + cD/n$ and therefore $\hat{D}(c) = \operatorname{argmin}_{1 \le D \le n} F(D, c)$. It is of course observed that $\hat{D}(c)$ is a non-increasing function of c, and generally this decrease occurs as follows: first it is very slow and the selected dimension remains very high for a while, then there is a very abrupt fall down and then again the decrease is very slow. Many simulation experiments in Birgé and Rozenholc [8], Comte and Rozenholc [18][5], led to the conclusion that a relevant choice of the constant was about twice the value of c for which the downward peak starts. Besides, it is well known that it is wiser to over- than to under-estimate the penalty: indeed a too small constant in the penalty leads to choose much too high dimensions whereas a too great one gives only a small error (a little to small dimension). Whatever the dependence between the variables, one may therefore easily apply this strategy here and select the right empirical value of the constant $\hat{c}$; the penalty is then $\hat{c}D_m/n$ and the procedure can be implemented.

Another idea, depending on more theoretical considerations, is the following. The penalty is obtained as an upper bound of $\mathbb{E}[(\sup_{t \in B_m(0,1)} \nu_n^d(t))^2]$ where $B_m(0,1) = \{t \in S_m, \|t\| \le 1\}$. We consider here a case of an absolutely regular process. If the mixing rate is not easy to estimate, some other terms may be replaced by estimators, as for instance the covariances. Indeed, we can write that

$$\begin{aligned}
\mathbb{E}\left[\left(\sup_{t \in B_m(0,1)} \nu_n^d(t)\right)^2\right] &= \frac{1}{n}\sum_{\lambda \in \Lambda_m}\left[\operatorname{Var}(\varphi_\lambda(X_1) + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\operatorname{cov}(\varphi_\lambda(X_0), \varphi_\lambda(X_k))\right] \\
&= \frac{1}{n}\sum_{\lambda \in \Lambda_m}\left[\operatorname{Var}(\varphi_\lambda(X_1) + 2\sum_{k=1}^{\psi(n)-1}\left(1 - \frac{k}{n}\right)\operatorname{cov}(\varphi_\lambda(X_0), \varphi_\lambda(X_k))\right] + R_n
\end{aligned}$$

where

$$|R_n| \le 4\Phi_0^2 \frac{D_m}{n} \sum_{k \ge \psi(n)} \beta_k = o\left(\frac{D_m}{n}\right)$$

[5]Let us precise that Birgé and Rozenholc [8] work in a framework of independent variables but Comte and Rozenholc [18] also study mixing variables generated by autoregressive models.

as soon as $\psi(n)$ tends to infinity when n tends to infinity, under [AR] with $\theta > 3$. Then a natural approximation of

$$\begin{aligned}\mathrm{pen}(m) &= \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left[\mathrm{Var}(\varphi_\lambda(X_1)) + 2\sum_{k=1}^{\psi(n)-1}\left(1-\frac{k}{n}\right)\mathrm{cov}(\varphi_\lambda(X_0),\varphi_\lambda(X_k))\right]\\ &= \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left[\mathbb{E}(\varphi_\lambda^2(X_1)) + 2\sum_{k=1}^{\psi(n)-1}\left(1-\frac{k}{n}\right)\mathbb{E}(\varphi_\lambda(X_0)\varphi_\lambda(X_k)) - u_n[\mathbb{E}(\varphi_\lambda(X_1)]^2\right],\end{aligned}$$

with $u_n = 1 + 2\sum_{k=1}^{\psi(n)-1}(1-\frac{k}{n}) = 2\psi(n) - 1 - (\psi(n)-1)\psi(n)/n$, is

$$\widehat{\mathrm{pen}}(m) = \frac{1}{n}\sum_{\lambda\in\Lambda_m}\left[\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda^2(X_i) + \frac{2}{n}\sum_{k=1}^{\psi(n)-1}\sum_{i=1}^{n-k}\varphi_\lambda(X_i)\varphi_\lambda(X_{i+k}) - u_n\,(\bar\varphi_\lambda)^2\right] \tag{5.1}$$

where $\bar\varphi_\lambda = \frac{1}{n}\sum_{k=1}^{n}\varphi_\lambda(X_i)$. It is easy to prove that the approximation of a covariance by its empirical counterpart implies a mean square error of order $1/\sqrt{n}$. Therefore, the global relative error is of order $\psi(n)/\sqrt{n}$, that is:

$$\mathbb{E}^{1/2}[(\mathrm{pen}(m) - \widehat{\mathrm{pen}}(m))^2] \le C\frac{\Phi_0^2 B_2 D_m}{n}\frac{\psi(n)}{\sqrt{n}}.$$

Some choices as $\psi(n) = \ln(n)$ or $\psi(n) = n^{1/4}$ imply therefore negligible errors. We can consider that from a theoretical and asymptotical point of view (*i.e.* for great values of n), as well as from a practical point of view, the empirical term $\widehat{\mathrm{pen}}(m)$ given by (5.1) is a relevant choice. It is beyond the scope of this paper to detail a rigorous proof of it.

6. Proofs

6.1. **Two useful results**

We give a lemma straightforwardly deduced from Talagrand's [34] inequality which is as follows:

Lemma 6.1. *Let $X_1,\dots,X_n$ be i.i.d. random variables and $\nu_n(t) = (1/n)\sum_{i=1}^{n}[t(X_i) - \mathbb{E}(t(X_i))]$ for t belonging to a countable class $\mathcal{F}$ of uniformly bounded measurable functions. Then, for any $\xi > 0$,*

$$\mathbb{E}\left[\sup_{t\in\mathcal{F}}|\nu_n(t)|^2 - 2(4+\xi^2)H^2\right]_+ \le \frac{6}{K_1}\left(\frac{v}{n}\mathrm{e}^{-K_1\xi^2\frac{nH^2}{v}} + \frac{4M_1^2}{K_1 n^2}\mathrm{e}^{-\frac{K_1\xi}{\sqrt{2}}\frac{nH}{M_1}}\right), \tag{6.1}$$

where K_1 is a universal constant,

$$\sup_{t\in\mathcal{F}}\|t\|_\infty \le M_1, \quad \mathbb{E}\left(\sup_{t\in\mathcal{F}}|\nu_n(t)|\right) \le H, \quad \sup_{t\in\mathcal{F}}\mathrm{Var}(t(X_1)) \le v.$$

Birgé and Massart [6] (p. 354, proof of Prop. 3) explain why $\mathcal{F}$ can also be taken as a unit ball of a finite dimensional space. Their argument is that, since $t \mapsto \nu_n(t)$ is a continuous function of t and the supremum is taken over a subset of a finite dimensional space, the value of the supremum does not change if it is restricted to a countable and dense subset. This also explains why all the supremums involved in the paper can be considered as measurable with respect to the probability measure.

Proof of Lemma 6.1. With the above notations, a consequence of Talagrand's [34] inequality as given by (5.13) in Corollary 2 of Birgé and Massart [6] (with f replaced by $f - \mathbb{E}(f(X_1))$ and M_1 by $2M_1$) can be written:

$$\mathbb{P}\left(\sup_{t\in\mathcal{F}}|\nu_n(t)| \geq \ (1+\eta)H + \lambda\right) \leq 3\exp\left[-K_1 n\left(\frac{\lambda^2}{v} \wedge \frac{(\eta\wedge 1)\lambda}{M_1}\right)\right]$$

which implies by taking $\eta = 1$

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in\mathcal{F}}|\nu_n(t)|^2 - 2(4+\xi^2)H^2\right]_+ &\leq \int_0^{+\infty}\mathbb{P}\left(\sup_{t\in\mathcal{F}}|\nu_n(t)|^2 \geq 2(4+\xi^2)H^2 + \tau\right)\mathrm{d}\tau \\
&\leq \int_0^{+\infty}\mathbb{P}\left(\sup_{t\in\mathcal{F}}|\nu_n(t)| \geq \sqrt{8H^2 + 2(\xi^2H^2 + \tau/2)}\right)\mathrm{d}\tau \\
&\leq 2\int_0^{+\infty}\mathbb{P}\left(\sup_{t\in\mathcal{F}}|\nu_n(t)| \geq 2H + \sqrt{\xi^2H^2+\tau}\right)\mathrm{d}\tau \\
&\leq 6\left(\int_0^{+\infty}\mathrm{e}^{-\frac{K_1 n}{v}(\xi^2H^2+\tau)}\mathrm{d}\tau + \int_0^{+\infty}\mathrm{e}^{-\frac{K_1 n}{\sqrt{2}M_1}(\xi H+\sqrt{\tau})}\mathrm{d}\tau\right)
\end{aligned}$$

and the result follows using that, for any positive constant C, $\int_0^{+\infty}\mathrm{e}^{-Cx}\mathrm{d}x = 1/C$ and $\int_0^{+\infty}\mathrm{e}^{-C\sqrt{x}}\mathrm{d}x = 2/C^2$.□

Moreover, when dealing with absolutely regular variables, we use the covariance inequality of Delyon [20], successfully exploited by Viennet [37] for partial sums of strictly stationary processes. To be more precise the result involved in the present paper is the following.

Let P be the distribution of X_0 on a probability space $\mathcal{X}$, $\int h\mathrm{d}P = \mathbb{E}_P(h)$ for any function h P-integrable. For $r \geq 2$, let $\mathcal{L}(r,\beta,P)$ be the set of functions $b : \mathcal{X} \to \mathbb{R}^+$ such that

$$b = \sum_{l\geq 0}(l+1)^{r-2}b_l \text{ with } 0 \leq b_l \leq 1 \text{ and } \mathbb{E}_P(b_l) \leq \beta_l.$$

Recall that from (2.1), B_r is the bound of the series $\sum_{l\geq 0}(l+1)^{r-2}\beta_l$. Then for $1 \leq p < \infty$ and any function b in $\mathcal{L}(2,\beta,P)$,

$$\mathbb{E}_P(b^p) \leq pB_{p+1}, \tag{6.2}$$

as soon as $B_{p+1} < \infty$. The following result holds for a strictly stationary absolutely regular sequence, $(X_i)_{i\in\mathbb{Z}}$, with β-mixing coefficients $(\beta_k)_{k\geq 0}$: if $B_2 < +\infty$, there exists $b \in \mathcal{L}\,(2,\beta,\infty)$ such that for any positive integer n and any measurable function $h \in \mathbb{L}_2(P)$, we have

$$\mathrm{Var}\left(\sum_{i=1}^n h(X_i)\right) \leq 4n\mathbb{E}_P(bh^2) = 4n\int b(x)h^2(x)\mathrm{d}P(x). \tag{6.3}$$

For the continuous time case, we use similarly the following lemma:

Lemma 6.2. *Let (X_t) be a strictly stationary continuous time β-mixing process. Then there exists a nonnegative function b such that $\mathbb{E}_P(b) \leq \int_0^{+\infty}\beta_s\mathrm{d}s$, $\mathbb{E}_P(b^p) \leq p\int_0^{+\infty}s^{p-1}\beta_s\mathrm{d}s$ and for any $h \in \mathbb{L}_2(P)$,*

$$\mathrm{Var}\left(\int_0^T h(X_s)\mathrm{d}s\right) \leq 4T\mathbb{E}_P(bh^2). \tag{6.4}$$

Of course if for instance h is bounded, $\mathbb{E}_P(bh^2) \le \|h\|_\infty^2 \mathbb{E}_P(b) \le \|h\|_\infty^2 A_2$.

Proof of Lemma 6.2. Lemma 4.1 in Viennet ([37], p. 478) implies that there exists b'_s and b''_s with values in $[0,1]$ such that $\mathbb{E}(b'_s(X_0)) \le \beta_s$, $\mathbb{E}(b''_s(X_0)) \le \beta_s$ and for any function h such that $h^2(X_0)$ is integrable,

$$|\mathrm{cov}(h(X_0), h(X_s))| \le 2\mathbb{E}_P^{1/2}(b'_s h^2)\mathbb{E}_P^{1/2}(b''_s h^2).$$

Therefore

$$\begin{aligned}
\mathrm{Var}\left(\int_0^T h(X_s)\mathrm{d}s\right) &= \int_0^T\int_0^T \mathrm{cov}(h(X_s), h(X_t))\mathrm{d}s\mathrm{d}t \\
&= 2\int_0^T (T-s)\mathrm{cov}(h(X_0), h(X_s))\mathrm{d}s \\
&\le 4T\int_0^T \mathbb{E}_P^{1/2}(b'_s h^2)\mathbb{E}_P^{1/2}(b''_s h^2)\mathrm{d}s \\
&\le 4T\int_0^T \mathbb{E}_P\left(\frac{1}{2}\,(b'_s + b''_s)h^2\right)\mathrm{d}s = 4T\mathbb{E}_P(bh^2)
\end{aligned}$$

where

$$b = \frac{1}{2}\int_0^T (b'_s + b''_s)\mathrm{d}s$$

and clearly $\mathbb{E}_P(b) \le \int_0^T \beta_s \mathrm{d}s$. The bound for $\mathbb{E}_P(b^p)$ follows analogously from the proof of Lemma 4.2 in Viennet [37] (p. 481). □

6.2. **Proof of Theorem 3.1**

6.2.1. *The continuous time β-mixing case*

The following lemma ensures the first part of Theorem 3.1:

Lemma 6.3. *Under the Assumptions of Theorem 3.1 and for the choice*

$$p(m,m') = \frac{80\Phi_0^2 A_2(D_m + D_{m'})}{n},$$

we have

$$\sum_{m'\in\mathcal{M}_n} \mathbb{E}(W^c(m')) \le K/n \tag{6.5}$$

where K is a constant depending on $\|f_A\|_\infty$, Φ_0, θ, A_2.

Indeed, equations (2.6, 2.7) and Lemma 6.3 imply

$$\frac{1}{2}\mathbb{E}\|f_A - \tilde{f}^c\|^2 \le \frac{3}{2}\|f_A - f_m\|^2 + \frac{4K}{n} + \mathbb{E}\left[\frac{320\Phi_0^2 A_2(D_m + D_{\hat{m}})}{n} + \mathrm{pen}_c(m) - \mathrm{pen}_c(\hat{m})\right]$$

which gives the result (3.1) of Theorem 3.1 for the choice $\mathrm{pen}_c(m) = 320\Phi_0^2 A_2 D_m/n$. □

Proof of Lemma 6.3. For the sake of simplicity and without loss of generality, we assume that $[T] = T = n$. We consider spaces $S_m + S_{m'}$, where m is fixed and m' can vary, and we denote by $D(m') = \dim(S_m + S_{m'})$.

Let φ_j be an orthonormal basis of $S_m + S_{m'}$ and let $Z_{i,m'}(\varphi_j) := Y_i(\varphi_j) - \mathbb{E}Y_i(\varphi_j)$ where

$$Y_i(\varphi_j) = \int_{(i-1)}^{i} \varphi_j(X_s)\mathrm{d}s.$$

Since (X_t) is β-mixing with coefficients β_t, the variables $\vec{Z}_{i,m'} = (Z_{i,m'}(\varphi_1), \dots, Z_{i,m'}(\varphi_{D(m')}))$ for $i = 1, \dots, n$ are also β-mixing with mixing coefficients β_k.

By using Berbee's lemma extended to sequences (see Bryc [12]), we build approximating variables for the vectors $\vec{Z}_{i,m'}$, denoted by $\vec{Z}^*_{i,m'}$. They are such that if $n = 2p_nq_n + r_n$, $0 \le r_n < q_n$, and $\ell = 0, \dots, p_n - 1$

$$A^*_{\ell,m'} = (\vec{Z}^*_{2\ell q_n+1,m'}, \dots, \vec{Z}^*_{(2\ell+1)q_n,m'}),\ B^*_{\ell,m'} = (\vec{Z}^*_{(2\ell+1)q_n+1,m'}, \dots, \vec{Z}^*_{(2\ell+2)q_n,m'}),$$

and analog definitions without stars, then

- $A^*_{\ell,m'}$ and $A_{\ell,m'}$ have the same law;
- $\mathbb{P}(A_{\ell,m'} \neq A^*_{\ell,m'}) \le \beta_{q_n}$;
- $A^*_{\ell,m'}$ and $(A_{0,m'}, A_{1,m'}, \dots, A_{\ell-1,m'}, A^*_{0,m'}, A^*_{1,m'}, \dots, A^*_{\ell-1,m'})$ are independent.

The blocks $B^*_{\ell,m'}$ are built in the same way.

For sake of simplicity, we assume that $n = 2p_nq_n$ (that is $r_n = 0$), which can always be done by completing the sequences with some 0's. Note that, for $t \in S_m + S_{m'}$, $t = \sum_{j=1}^{D(m')} a_j\varphi_j$, we have

$$\nu^c_n(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{D(m')} a_j Z_{i,m'}\ (\varphi_j) := \nu^{c\ (1)}_n\ (t) + \nu^{c\ (2)}_n\ (t),$$

where

$$\nu^{c\ (1)}_n\ (t) := \frac{1}{n}\sum_{j=1}^{D(m')} a_j \sum_{\ell=0}^{p_n-1}\sum_{i=2\ell q_n+1}^{(2\ell+1)q_n} Z_{i,m'}(\varphi_j) \text{ and } \nu^{c\ (2)}_n\ (t) := \frac{1}{n}\sum_{j=1}^{D(m')} a_j \sum_{\ell=0}^{p_n-1}\sum_{i=(2\ell+1)q_n+1}^{(2\ell+2)q_n} Z_{i,m'}(\varphi_j).$$

We denote now

$$\nu^{c\ (1)*}_n\ (t) := \frac{1}{n}\sum_{j=1}^{D(m')} a_j \sum_{\ell=0}^{p_n-1}\sum_{i=2\ell q_n+1}^{(2\ell+1)q_n} Z^*_{i,m'}(\varphi_j)\,, \nu^{c\ (2)*}_n\ (t) := \frac{1}{n}\sum_{j=1}^{D(m')} a_j \sum_{\ell=0}^{p_n-1}\sum_{i=(2\ell+1)q_n+1}^{(2\ell+2)q_n} Z^*_{i,m'}(\varphi_j)$$

and $\nu^{c*}_n(t) := \nu^{c\ (1)*}_n\ (t) + \nu^{c\ (2)*}_n(t)$.

We easily infer that

$$W^c(m') \le 2\sup_{t\in B_{m,m'}(0,1)} (\nu^c_n(t) - \nu^{*c}_n(t))^2 + 2W^{c*}(m') \tag{6.6}$$

where $W^{c*}(m') = \left[\left(\sup_{t\in B_{m,m'}(0,1)} |\nu^{c*}_n(t)|\right)^2 - \frac{1}{2}p(m,m')\right]_+$. We study separately the two terms that appear from (6.6), namely

$$\sum_{m'\in\mathcal{M}_n} \mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)} (\nu^c_n(t) - \nu^{*c}_n(t))^2\right) \text{ and } \sum_{m'\in\mathcal{M}_n} \mathbb{E}(W^{c*}(m')).$$

For $t \in S_m + S_{m'}$, $t = \sum_{j=1}^{D(m')} a_j \varphi_j$, let us denote

$$U_{\ell,m'}(t) = \frac{p_n}{n} \sum_{i=2\ell q_n+1}^{(2\ell+1)q_n} \sum_{j=1}^{D(m')} a_j Z_{i,m'}(\varphi_j) \quad \text{and} \quad U^*_{\ell,m'}(t) = \frac{p_n}{n} \sum_{i=2\ell q_n+1}^{(2\ell+1)q_n} \sum_{j=1}^{D(m')} a_j Z^*_{i,m'}(\varphi_j). \tag{6.7}$$

Since for all t in $B_{m,m'}(0,1)$, $U_{\ell,m'}(t) \overset{\mathcal{L}}{\sim} U^*_{\ell,m'}(t)$ we have

$$\|U^*_{\ell,m'}(t)\|_\infty = \|U_{\ell,m'}(t)\|_\infty = \left\| \frac{p_n}{n} \left\{ \int_{2\ell q_n}^{(2\ell+1)q_n} t(X_s)\mathrm{d}s - \mathbb{E} \int_{2\ell q_n}^{(2\ell+1)q_n} t(X_s)\mathrm{d}s \right\} \right\|_\infty \leq \|t\|_\infty ,$$

and then [M2] implies $\|t\|_\infty \leq \Phi_0 \sqrt{2N_n}$. Therefore, we derive that for $t \in B_{m,m'}(0,1) = \{t \in S_m + S_{m'}, \|t\| = 1\}$,

$$\begin{aligned} |\nu_n^{c\ (1)}(t) - \nu_n^{c\ (1)*}(t)| = \left| \frac{1}{p_n} \sum_{\ell=0}^{p_n-1} (U_{\ell,m'}(t) - U^*_{\ell,m'}(t)) \right| &\leq 2\Phi_0 \sqrt{2N_n} \left(\frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \mathbb{1}_{\{U_{\ell,m'}(t) \neq U^*_{\ell,m'}(t)\}} \right) \\ &\leq 2\Phi_0 \sqrt{2N_n} \left(\frac{1}{p_n} \sum_{\ell=0}^{p_n-1} \mathbb{1}_{\{A_{\ell,m'} \neq A^*_{\ell,m'}\}} \right). \end{aligned}$$

This yields

$$\mathbb{E} \left(\sup_{t \in B_{m,m'}(0,1)} |\nu_n^{c\ (1)}(t) - \nu_n^{c\ (1)*}(t)| \right)^2 \leq 8\Phi_0^2 N_n \beta_{q_n} .$$

Since a similar bound obviously holds for $\mathbb{E}\left(\sup_{t \in B_{m,m'}(0,1)} |\nu_n^{c\ (2)}(t) - \nu_n^{c\ (2)*}(t)|\right)^2$, it follows that

$$\mathbb{E} \left(\sup_{t \in B_{m,m'}(0,1)} |\nu_n^{c\ (1)}(t) + \nu_n^{c\ (2)}(t) - \nu_n^{c*}(t)| \right)^2 \leq 32\Phi_0^2 N_n \beta_{q_n} .$$

We find

$$\mathbb{E} \left\{ \sum_{m' \in \mathcal{M}_n} \sup_{t \in B_{m,m'}(0,1)} (\nu_n^c(t) - \nu_n^{*c}(t))^2 \right\} \leq C\Phi_0^2 N_n |\mathcal{M}_n| \beta_{q_n}, \tag{6.8}$$

where C is a positive constant, which implies that (6.5) holds provided that

$$N_n |\mathcal{M}_n| \beta_{q_n} \leq \frac{C}{n} \text{ for some constant } C. \tag{6.9}$$

Let us study now $W^{c(1)*}(m') := \left[\left(\sup_{t \in B_{m,m'}(0,1)} |\nu_n^{c(1)*}(t)| \right)^2 - \frac{1}{4} p(m,m') \right]_+$ (the proof for $W^{c(2)*}(m') := \left[\left(\sup_{t \in B_{m,m'}(0,1)} |\nu_n^{c(2)*}(t)| \right)^2 - \frac{1}{4} p(m,m') \right]_+$ being similar). Since $\nu_n^{c(1)*}(t) = (1/p_n) \sum_{\ell=0}^{p_n-1} U^*_{\ell,m'}(t)$ and the variables $U^*_{\ell,m'}(t)$ defined by (6.7) are independent, we can apply inequality (6.1) of Lemma 6.1, if we can compute H, v and M_1, where

$$\sup_{t \in B_{m,m'}(0,1)} \mathrm{Var}(U^*_{1,m'}(t)) \leq v, \quad \sup_{t \in B_{m,m'}(0,1)} \|U^*_{1,m'}(t)\|_\infty \leq M_1$$

and

$$\frac{1}{p_n}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}\left|\sum_{\ell=0}^{p_n-1}U^*_{\ell,m'}(t)\right|\right)\le H.$$

The bound $M_1:=\Phi_0\sqrt{D_m+D_{m'}}$ follows from [M2] applied with $\|t\|=1$ as previously. Let us compute H. Cauchy–Schwarz inequality combined with independence, the fact that $U_{\ell,m'}(t)\overset{\mathcal{L}}{\sim}U^*_{\ell,m'}(t)$ and stationarity yield

$$\begin{aligned}
&\frac{1}{p_n^2}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}\left|\sum_{\ell=0}^{p_n-1}U^*_{\ell,m'}(t)\right|\right)^2=\frac{1}{p_n^2}\mathbb{E}\left(\sup_{\Sigma_j a_j^2\le 1}\left|\sum_{j=1}^{D(m')}a_j\sum_{\ell=0}^{p_n-1}U^*_{\ell,m'}(\varphi_j)\right|\right)^2\\
\le\ &\frac{1}{p_n^2}\mathbb{E}\left\{\sum_{j=1}^{D(m')}\left(\sum_{\ell=0}^{p_n-1}U^*_{\ell,m'}(\varphi_j)\right)^2\right\}=\frac{p_n^2}{p_n n^2}\sum_{j=1}^{D(m')}\mathrm{Var}\left(\sum_{i=1}^{q_n}Y_i(\varphi_j)\right)\\
\le\ &\frac{p_n}{n^2}\sum_{j=1}^{D(m')}\mathrm{Var}\left(\int_0^{q_n}\varphi_j\ (X_s)\mathrm{d}s\right)\\
\le\ &\frac{4q_np_n}{n^2}\sum_{j=1}^{D(m')}\int_A b(u)\varphi_j^2(u)f(u)\mathrm{d}u \text{ from Lemma 6.2,}\\
\le\ &\frac{2}{n}\int_A b(u)\left\|\sum_{j=1}^{D(m')}\varphi_j^2(u)\right\|_\infty f(u)\mathrm{d}u\le\frac{2\Phi_0^2D(m')}{n}\int_A b(u)f(u)\mathrm{d}u\quad\text{with (2.2),}\\
\le\ &\frac{2\Phi_0^2D(m')A_2}{n}\quad\text{with Lemma 6.2 again.}
\end{aligned}$$

This leads to

$$H^2=\frac{2\Phi_0^2A_2D}{n},$$

where $D=D_m+D_{m'}$. For any $t\in B_{m,m'}(0,1)$, the same method leads to

$$\begin{aligned}
\mathrm{Var}(U^*_{\ell,m'}(t))\ &=\ \mathrm{Var}(U_{\ell,m'}(t))=\frac{1}{q_n^2}\mathrm{Var}\left(\int_0^{q_n}t(X_s)\mathrm{d}s\right)\le\frac{4}{q_n}\int b(x)t^2(x)f(x)\mathrm{d}x\\
&\le\ \frac{4\|t\|_\infty}{q_n}\sqrt{\int b^2(x)f(x)\mathrm{d}x\int t^2(x)f(x)\mathrm{d}x}\le\frac{4\Phi_0\sqrt{D(m')}\sqrt{2\int_0^{+\infty}s\beta_s\mathrm{d}s\|f_A\|_\infty}}{q_n},
\end{aligned}$$

using Lemma 6.2. Therefore, we find

$$v=\frac{4\sqrt{2\|f_A\|_\infty}\Phi_0\sqrt{A_3D}}{q_n}:=C_0\frac{\sqrt{D}}{q_n}.$$

Plugging H,M_1,v in inequality (6.1) and setting $\frac{1}{4}p(m,m')=2(4+\xi^2)H^2$ lead to

$$\mathbb{E}(W^{c(1)*}(m'))\le C_1\frac{\sqrt{D}}{n}\mathrm{e}^{-C_2\xi^2\sqrt{D}}+C_3\frac{q_n^2D}{n^2}\mathrm{e}^{-C_4\xi\sqrt{n}/q_n}\tag{6.10}$$

with

$$C_1=\frac{48\sqrt{2\|f_A\|_\infty}\Phi_0\sqrt{A_3}}{K_1},\ \ C_2=\frac{K_1A_2\Phi_0}{4\sqrt{2\|f_A\|_\infty A_3}},\ \ C_3=\frac{96\Phi_0^2}{K_1^2},C_4=\frac{K_1\sqrt{A_2}}{2}.$$

The constants are given here to show that for arithmetical (or geometrical) mixing, they depend on θ in such a way that they increase when θ decreases.

Under [M3],

$$\sum_{m'\in\mathcal{M}_n} \frac{\sqrt{D}}{n}\mathrm{e}^{-C_2\xi^2\sqrt{D}} \le \frac{2\Sigma(C_2\xi^2)}{n} := \frac{K}{n}$$

using that $D = D_m + D_{m'}$ and

$$\sqrt{x+y}\mathrm{e}^{-\sqrt{x+y}} \;\le\; (\sqrt{x}+\sqrt{y})\mathrm{e}^{-(\sqrt{\frac{x}{2}}+\sqrt{\frac{y}{2}})} \le\; (\sqrt{x}\mathrm{e}^{-\sqrt{x/2}}+\sqrt{y})\mathrm{e}^{-\sqrt{y/2}} \le\; (1+\sqrt{y})\mathrm{e}^{-\sqrt{y/2}}.$$

For $\xi = 1$ and $|\mathcal{M}_n| \le n^\epsilon$, the sums over $\mathcal{M}_n$ of the last term of the right-hand-side of (6.10) is less than

$$C_5 q_n^2 n^{\epsilon-1}\mathrm{e}^{-K'\sqrt{n}/q_n} \tag{6.11}$$

where C_5 and K' are positive constants. On the other hand, equation (6.9) requires

$$n^{\epsilon+1}\beta_{q_n} \le \frac{C}{n}\cdot \tag{6.12}$$

Since the mixing is arithmetic[6], take $q_n = [n^c]$ with $0 < c < 1/2$, then the term in (6.11) is less than C/n for some constant C and (6.12) holds if $\theta \ge\; (\epsilon+2)/c - 1$. The optimal choice is $q_n = [K'\sqrt{n}/((1+\epsilon)\ln(n))]$ with $\theta > 2\epsilon + 3$. □

6.2.2. *The discrete time β-mixing case*

We use Bryc's [12] construction again. But here we build variables X_i^* such that if $n = 2p_nq_n + r_n$, $0 \le r_n < q_n$, and $\ell = 0, \dots, p_n - 1$

$$A_\ell^* = (X^*_{2\ell q_n+1}, \dots, X^*_{(2\ell+1)q_n}),\;\; B_\ell^* = (X^*_{(2\ell+1)q_n+1}, \dots, X^*_{(2\ell+2)q_n}),$$

and analog definitions without stars, then

- A_ℓ^* and A_ℓ have the same law;
- $\mathbb{P}(A_\ell \ne A_\ell^*) \le \beta_{q_n}$;
- A_ℓ^* and $(A_0, A_1, \dots, A_{\ell-1}, A_0^*, A_1^*, \dots, A_{\ell-1}^*)$ are independent.

The blocks B_ℓ^* are built in the same way. Again, we assume for simplicity that $r_n = 0$.

Starting from (2.6), we write

$$2|\nu_n^d(\tilde f - f_m)| \le 2|\nu_n^d(\tilde f - f_m) - \nu_n^{d*}(\tilde f - f_m)| + 2|\nu_n^{d*}(\tilde f - f_m)| \tag{6.13}$$

where ν_n^{d*} denotes the empirical contrast computed on the X_i^*. If we denote by $B_{m,m'}(0,1)$ the unit ball of the linear space $S_m + S_{m'}$, then we have, using the same method as the one leading from (2.6) to (2.7):

$$2|\nu_n^{d*}(\tilde f - f_m)| \le \frac{1}{4}\|f_m - f_A\|^2 + \frac{1}{4}\|f_A - \tilde f\|^2 + 8\sum_{m'\in\mathcal{M}_n} W^{d*}(m') + 8p(m,\hat m), \tag{6.14}$$

[6]If the mixing was geometric, the choice $q_n = (\epsilon+2)\ln(n)/\theta$ would suit for any θ.

where $W^{d*}(m')$ is defined as in (2.8) with X_i^* replacing X_i. On an other hand

$$2|\nu_n^d(\tilde f - f_m) - \nu_n^{d*}(\tilde f - f_m)| \leq \frac{2}{n}\left|\sum_{i=1}^n [(\tilde f - f_m)(X_i) - (\tilde f - f_m)(X_i^*)]\right| + \frac{2}{n}\mathbb{E}\left|\sum_{i=1}^n [(\tilde f - f_m)(X_i) - (\tilde f - f_m)(X_i^*)]\right|.$$

The ideas for dealing with both terms are the same, for instance:

$$\begin{aligned}\frac{2}{n}\left|\sum_{i=1}^n [(\tilde f - f_m)(X_i) - (\tilde f - f_m)(X_i^*)]\right| &\leq \frac{4}{n}\sum_{i=1}^n \|\tilde f - f_m\|_\infty \mathbb{1}_{X_i \neq X_i^*} \\ &\leq \frac{4\Phi_0\sqrt{2N_n}}{n}\|\tilde f - f_m\|\sum_{i=1}^n \mathbb{1}_{X_i \neq X_i^*} \\ &\leq \frac{1}{16}\|\tilde f - f_m\|^2 + 128\Phi_0^2 N_n\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \neq X_i^*}\right)^2,\end{aligned}$$

using [M2]. Thus taking the expectation leads to

$$2\mathbb{E}|\nu_n^d(\tilde f - f_m) - \nu_n^{d*}(\tilde f - f_m)| \leq \frac{1}{4}\mathbb{E}\|\tilde f - f_A\|^2 + \frac{1}{4}\|f_A - f_m\|^2 + 256\Phi_0^2 N_n \beta_{q_n}. \tag{6.15}$$

Therefore we need

$$N_n \beta_{q_n} \leq C/n. \tag{6.16}$$

Let us study now $W^{d*}(m')$. We apply inequality (6.1) again. We find if we denote by $D = \dim(S_m) + \dim(S_{m'}) = D_m + D_{m'}$,

$$M_1 = \Phi_0\sqrt{D}, \quad v = 4\sqrt{2\|f_A\|_\infty B_3}\Phi_0\frac{\sqrt{D}}{q_n}, \quad H^2 = 2\Phi_0^2 B_2\frac{D}{n},$$

where B_r is defined by (2.1). Therefore if we choose

$$\frac{1}{2}p(m,m') = 4(4+\xi^2)\Phi_0^2 B_2\frac{D_m + D_{m'}}{n}, \tag{6.17}$$

we obtain with (6.1)

$$\mathbb{E}(W^{d*}(m')) \leq C(K_1, B_3, \Phi_0, \|f_A\|_\infty)\left[\frac{\sqrt{D}}{n}\exp\left(-C_1\xi^2\sqrt{D}\right) + \frac{q_n^2 D}{n^2}\exp\left(-C_2\xi\frac{\sqrt{n}}{q_n}\right)\right],$$

with

$$C_1 = \frac{K_1\Phi_0 B_2}{4\sqrt{2\|f_A\|_\infty B_3}}, \quad C_2 = \frac{K_1\sqrt{B_2}}{2}.$$

To bound $\sum_{m'\in\mathcal{M}_n}\mathbb{E}(W^{d*}(m'))$, we proceed in the same way as at the end of the previous proof. Taking into account that the mixing is [AR], we choose $q_n = [n^c]$ with $0 < c < \frac{1}{2}$, which implies that $\sum_{m'\in\mathcal{M}_n}\mathbb{E}(W^{d*}(m'))$ remains of order $1/n$ using [M3]. Besides (6.16) requires that $\theta > (\omega+1)/c - 1$ if $N_n \leq n^\omega$, that is $\theta > 2\omega + 1$. Then gathering (6.13, 6.15), we find that (2.6) leads to

$$\frac{1}{2}\mathbb{E}(\|f_A - \tilde f\|^2) \leq \frac{3}{2}\|f_A - f_m\|^2 + \frac{C}{n} + \mathbb{E}\left[8p(m,\hat m) + \mathrm{pen}_d(m) - \mathrm{pen}_d(\hat m)\right]$$

with $p(m,m')$ defined by (6.17) with $\xi = 1$ so that the choice $\mathrm{pen}_d(m) = 320\Phi_0^2 B_2 D_m/n$ gives the inequality (3.1) of Theorem 3.1. □

6.3. **Proof of Proposition 3.2**

Here we consider a discrete time α-mixing process satisfying assumption [Lip]. The control given by (6.15) is no longer possible. We must proceed as in the continuous time case. With obvious notations, we write as in (2.7):

$$2|\nu_n^d(\tilde{f}-f_m)| \le \frac{1}{4}\|f_m - f_A\|^2 + \frac{1}{4}\|f_A - \tilde{f}\|^2 + 8\sum_{m'\in\mathcal{M}_n} W^d(m') + 8p(m,\hat{m})$$

and as in (6.6) that:

$$W^d(m') \le 2\sup_{t\in B_{m,m'}(0,1)} (\nu_n^d(t) - \nu_n^{*d}(t))^2 + 2W^{d*}(m')\,,$$

where $W^{d*}(m') = \left[\left(\sup_{t\in B_{m,m'}(0,1)} |\nu_n^{d*}(t)|\right)^2 - \frac{1}{2}p(m,m')\right]_+$.

Moreover, we need other approximation results adapted to the α-mixing case. We shall make use of the following consequence of Theorem 4 in Rio [32].

Lemma 6.4. *Let $\{\xi_n, n\ge 1\}$ be a sequence of real random variables such that, for each $n\ge 1$, $\mathbb{P}(a_n \le \xi_n \le b_n) = 1$ where $a_n \le b_n$ are real numbers. Denote by $\mathcal{F}_1^n = \sigma(\xi_1,\dots,\xi_n)$. Then, we can redefine $\{\xi_n, n\ge 1\}$ onto a richer probability space on which there exists a sequence $\{\xi_n^*, n\ge 1\}$ of independent random variables such that, for each $n\ge 1$, ξ_n and ξ_n^* have the same distribution and*

$$\mathbb{E}\left(|\xi_n - \xi_n^*|\right) \le 2\,(b_n - a_n)\alpha\left(\mathcal{F}_1^{n-1}, \sigma\,(\xi_n)\right).$$

Moreover, for every $n>1$, ξ_n^* and $(\xi_1,\dots,\xi_{n-1})$ are independent random variables.

We construct the even and odd blocks, respectively $\{A_\ell\}_{0\le\ell\le p_n-1}$ and $\{B_\ell\}_{0\le\ell\le p_n-1}$ as before. Now, we consider the sequences $\{A_\ell^*\}_{0\le\ell\le p_n-1}$ and $\{B_\ell^*\}_{0\le\ell\le p_n-1}$ of independent random variables each distributed respectively as A_ℓ and B_ℓ, and defined as in Lemma 6.4.
Then Cauchy–Schwarz inequality implies that

$$\begin{aligned}
\sum_{m'\in\mathcal{M}_n}\sup_{t\in B_{m,m'}(0,1)}(\nu_n^d(t)-\nu_n^{*d}(t))^2 &\le \sum_{m'\in\mathcal{M}_n}\sup_{\Sigma_j a_j^2\le 1}\left[\sum_j a_j\left(\nu_n^d(\varphi_j)-\nu_n^{*d}\,(\varphi_j)\right)\right]^2\\
&\le \sum_{m'\in\mathcal{M}_n}\sum_j (\nu_n^d(\varphi_j)-\nu_n^{*d}(\varphi_j))^2 \le 2\sum_{m'\in\mathcal{M}_n}\sum_j \|\nu_n^d(\varphi_j)\|_\infty\left|\nu_n^d(\varphi_j)-\nu_n^{*d}\,(\varphi_j)\right|\\
&\le \frac{4}{n}\sum_{m'\in\mathcal{M}_n}\sum_j\|\varphi_j\|_\infty\sum_{\ell=0}^{p_n-1}\left|\sum_{i=2\ell q_n+1}^{(2\ell+1)q_n}(\varphi_j(X_i)-\varphi_j(X_i^*)) + \sum_{i=(2\ell+1)q_n+1}^{(2\ell+2)q_n}(\varphi_j(X_i)-\varphi_j(X_i^*))\right|.
\end{aligned}$$

Now Lemma 6.4 entails that for each $\ell\in[0,p_n-1]$,

$$\mathbb{E}\left|\sum_{i=2\ell q_n+1}^{(2\ell+1)q_n}(\varphi_j(X_i)-\varphi_j(X_i^*))\right| \le 4q_n\alpha_{q_n}\|\varphi_j\|_\infty\,,$$

which combined with [M2] entails

$$\mathbb{E}\left\{\sum_{m'\in\mathcal{M}_n}\sup_{t\in B_{m,m'}(0,1)}(\nu_n^d(t)-\nu_n^{*d}(t))^2\right\}\le 32\Phi_0^2N_n^2|\mathcal{M}_n|\alpha_{q_n}\,.$$

This gives the condition

$$N_n^2n^\epsilon\alpha_{q_n}\le\frac{C}{n}\cdot \tag{6.18}$$

Now we first notice that

$$\begin{aligned}\nu_n^{*d}(t) &= \frac{p_n}{n}\left\{\frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\left(\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}(t(X_i^*)-\mathbb{E}t(X_i^*))+\sum_{i=q_n\ (2\ell+1)+1}^{q_n(2\ell+2)}(t(X_i^*)-\mathbb{E}t(X_i^*))\right)\right\}\\ &:= \nu_n^{*d\ (1)}(t)+\nu_n^{*d\ (2)}(t)\,.\end{aligned}$$

Since the variables $\left(\frac{p_n}{n}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}(t(X_i^*)-\mathbb{E}t(X_i^*))\right)_{0\le\ell\le p_n-1}$ are independent by construction, we are allowed to apply (6.1) to $\nu_n^{*d\ (1)}(t)$ with adequate choice of M_1, H and v. Since for all $t\in B_{m,m'}(0,1)$,

$$\left\|\frac{p_n}{n}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}(t(X_i^*)-\mathbb{E}t(X_i^*))\right\|_\infty\le\Phi_0\sqrt{D}\,,$$

we put $M_1=\Phi_0\sqrt{D}$, where $D=D_m+D_{m'}$.

From Lemma 1.3 in Bosq [10], we know that [LIP] implies that $\|g_k\|_\infty\le\ (1+\ell_k\sqrt{2})\alpha_k^{1/3}$. This result together with stationarity lead to the following estimate

$$\begin{aligned}\sup_{t\in B_{m,m'}(0,1)}\mathrm{Var}\left(\frac{p_n}{n}\sum_{i=1}^{q_n}t(X_i)\right) &= \sup_{t\in B_{m,m'}(0,1)}\left\{\frac{p_n^2}{n^2}q_n\mathbb{E}\left(t(X_1)\right)^2+2\frac{p_n^2}{n^2}\sum_{k=1}^{q_n-1}(q_n-k)\mathrm{Cov}\left(t(X_1),t(X_{k+1})\right)\right\}\\ &\le \frac{1}{4q_n}\|f_A\|_\infty+\frac{1}{2q_n}\sup_{t\in B_{m,m'}(0,1)}\sum_{k=1}^{q_n-1}\left\{\left|\int_A\int_A t(x)t(y)g_k\ (x,y)\mathrm{d}x\mathrm{d}y\right|\right\}\\ &\le \frac{1}{4q_n}\|f_A\|_\infty+\frac{m(A)}{2q_n}\sum_{k=1}^{q_n}\left(1+\sqrt{2}\ell_k\right)\alpha_k^{1/3}\end{aligned}$$

where $m(A)$ is the Lebesgue measure of A. Thus we set

$$v=\frac{1}{4q_n}\left(\|f_A\|_\infty+2m(A)\sqrt{2}\sum_{k=1}^{\infty}(1+\ell_k)\alpha_k^{1/3}\right):=\frac{C_1}{q_n}\cdot$$

Let us now determine the quantity H. Using Cauchy–Schwarz inequality, stationarity and once again assumption (3.3) together with Lemma 1.3 in Bosq [10], we derive that

$$
\begin{aligned}
& \frac{1}{p_n^2}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}\left|\frac{p_n}{n}\sum_{\ell=0}^{p_n-1}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}(t(X_i^*)-\mathbb{E}t(X_i^*))\right|\right)^2 \\
= \;& \frac{1}{n^2}\mathbb{E}\left(\sup_{\Sigma_{j=1}^{D(m')}a_j^2\le 1}\left|\sum_{j=1}^{D(m')}a_j\sum_{\ell=0}^{p_n-1}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}[\varphi_j(X_i^*)-\mathbb{E}\varphi_j(X_i^*)]\right|\right)^2 \\
\le \;& \frac{1}{n^2}\sum_{j=1}^{D(m')}\mathrm{Var}\left(\sum_{\ell=0}^{p_n-1}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}\varphi_j(X_i^*)\right)=\frac{p_n}{n^2}\sum_{j=1}^{D(m')}\mathrm{Var}\left(\sum_{i=1}^{q_n}\varphi_j(X_i)\right) \qquad (6.19)\\
\le \;& \frac{p_nq_n}{n^2}\sum_{j=1}^{D(m')}\mathrm{Var}\left(\varphi_j(X_1)\right)+\frac{2p_n}{n^2}\sum_{j=1}^{D(m')}\sum_{k=1}^{q_n-1}(q_n-k)\mathrm{Cov}\left(\varphi_j(X_1),\varphi_j(X_{k+1})\right) \\
\le \;& \frac{1}{2n}\sum_{j=1}^{D(m')}\mathbb{E}\varphi_j^2(X_1)+\frac{1}{n}\sum_{j=1}^{D(m')}\sum_{k=1}^{q_n}\left|\int_A\int_A\varphi_j(x)\varphi_j(y)g_k\ (x,y)\mathrm{d}x\mathrm{d}y\right| \\
\le \;& \frac{1}{2n}\sum_{j=1}^{D(m')}\mathbb{E}\varphi_j^2(X_1)+\frac{1}{n}\sum_{j=1}^{D(m')}\int_A\int_A\left|\varphi_j(x)\varphi_j(y)\right|\mathrm{d}x\mathrm{d}y\sum_{k=1}^{q_n}\left(1+\sqrt{2}\ell_k\right)\alpha_k^{1/3},
\end{aligned}
$$

where $D(m')=\dim(S_m+S_{m'})$. Thus setting $D=D_m+D_{m'}$ and using (2.2), we have

$$
H^2=\frac{D}{2n}\left(\Phi_0^2+2m(A)\sqrt{2}\sum_{k=1}^{\infty}(1+\ell_k)\alpha_k^{1/3}\right):=C_2\frac{D}{n}\cdot
$$

Gathering all these last considerations and applying (6.1), we infer that

$$
\mathbb{E}(W^{d*}(m'))\le K\left[\frac{1}{n}\exp\left(-\frac{K_1C_2\xi^2}{2C_1}D\right)+\frac{q_n^2D}{n^2}\exp\left(-\frac{\xi K_1\sqrt{C_2}}{2\sqrt{2}\Phi_0}\frac{\sqrt{n}}{q_n}\right)\right],
$$

where $K=K(K_1,C_1,C_2,\Phi_0)$, provided that we choose

$$
\frac{1}{4}p(m,m')=2(4+\xi^2)C_2\frac{D}{n}\cdot
$$

The conclusion is the same as previously for the bound of $\sum_{m'\in\mathcal{M}_n}\mathbb{E}(W^{d*}(m'))$ which has the right order if $q_n=n^c$ under [AR] with $0<c<1/2$. Moreover (6.18) is fulfilled for $q_n=[n^c]$ under [AR], if $n^{2\omega}n^{\epsilon}n^{-c(1+\theta)}\le C/n$ that is $\theta\ge\ (1+2\omega+\epsilon)/c-1$ when $N_n\le n^{\omega}$. □

6.4. **Proof of Theorem 3.2**

The proof is the same as above except that [Lip] no longer holds so that the application of (6.1) is now based on the bounds $v = \dfrac{\|f_A\|_\infty}{4}$ and $H^2 = \dfrac{\Phi_0^2 D}{4p_n}$ because starting from (6.19), we can write

$$\begin{aligned}
\frac{1}{p_n^2}\mathbb{E}\left(\sup_{t\in B_{m,m'}(0,1)}\left|\frac{p_n}{n}\sum_{\ell=0}^{p_n-1}\sum_{i=2q_n\ell+1}^{q_n(2\ell+1)}(t(X_i^*)-\mathbb{E}t(X_i^*))\right|\right)^2 &\le \frac{p_n}{n^2}\sum_{j=1}^{D(m')}\mathbb{E}\left[\left(\sum_{i=1}^{q_n}\varphi_j(X_i)\right)^2\right]\\
&\le \frac{q_n p_n}{n^2}\sum_{j=1}^{D(m')}\mathbb{E}\left(\sum_{i=1}^{q_n}\varphi_j^2(X_i)\right)\\
&\le \frac{q_n^2 p_n}{n^2}\sum_{j=1}^{D(m')}\mathbb{E}\left(\varphi_j^2(X_1)\right) \qquad (6.20)\\
&\le \frac{\Phi_0^2 D(m')}{4p_n}
\end{aligned}$$

using that $\int_A f(x)\mathrm{d}x \le 1$ and (2.2).

Therefore $H^2 = \dfrac{(\epsilon+3)\Phi_0^2\ln(n)D}{2n\theta}$, for $q_n = [(\epsilon+3)\ln(n)/\theta]$ under [GEO], $D = D_m + D_{m'}$ and $|\mathcal{M}_n| \le n^\epsilon$. Moreover we still have $M_1 = \Phi_0\sqrt{D}$. Therefore we must choose

$$\frac{1}{4}p(m,m') = (4+\xi^2)\frac{(\epsilon+3)\Phi_0^2}{\theta}\frac{\ln(n)D}{n}$$

and we find

$$\mathbb{E}(W^{d*}(m')) \le K\left[\frac{\ln(n)}{n}\mathrm{e}^{-C_1\xi^2 D} + \frac{D\ln^2(n)}{n}\mathrm{e}^{-C_2\xi\sqrt{n/\ln(n)}}\right]$$

where K, C_1, C_2 are some constants depending on Φ_0, ϵ, $\|f_A\|_\infty$ and θ. Under [M3], this gives for $\sum_{m\in\mathcal{M}_n}\mathbb{E}(W^{d*}(m'))$ the order $\ln(n)/n$. □

6.5. **Proof of Theorem 4.1**

We know from Birgé and Massart [5] that in the spaces [GP] there exists a real function ψ on $\mathcal{S}_n$ such that for all $t \in \mathcal{S}_n$ and $m \in \mathcal{M}_n$, $\|t_m\|_\infty \le \psi(t)$ and which satisfies

$$|\psi(\bar{f}_n) - \psi(\hat{f}_n)| \le \Phi\sqrt{N_n}\sup_{\lambda\in\Lambda_n}|\nu_n^d(\varphi_\lambda)|, \qquad (6.21)$$

where $\bar{f}_n$ is the projection of f on $\mathcal{S}_n$ and $\hat{f}_n$ the projection estimator on $\mathcal{S}_n$.

Indeed, using Inequality (8) of Birgé and Massart [5], we can simply choose $\psi(t) = r\|t\|_\infty$. Then analogously, since for all n, $\|\bar{f}_n\|_\infty \le r\|f_A\|_\infty$, we get $\sup_n \psi(\bar{f}_n) \le r\|f_A\|_\infty = \psi(f_A)$. In addition, we get

$$\begin{aligned}
|\psi(\bar{f}_n) - \psi(\hat{f}_n)| &= r|\|\bar{f}_n\|_\infty - \|\hat{f}_n\|_\infty| \le r\|\bar{f}_n - \hat{f}_n\|_\infty \\
&\le r\left\|\sum_{\lambda\in\Lambda_n}[\beta_\lambda(\bar{f}_n) - \beta_\lambda(\hat{f}_n)]\varphi_\lambda\right\|_\infty \\
&\le r\left\|\sum_{\lambda\in\Lambda_n}\left[\langle f,\varphi_\lambda\rangle - \frac{1}{n}\sum_{i=1}^n \varphi_\lambda(X_i)\right]\varphi_\lambda\right\|_\infty \\
&\le r\left\|\sum_{\lambda\in\Lambda_n}\nu_n^d(\varphi_\lambda)\varphi_\lambda\right\|_\infty \le r\sup_{\lambda\in\Lambda_n}|\nu_n^d(\varphi_\lambda)|\left\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda\right\|_\infty
\end{aligned}$$

using (8) of Birgé and Massart [5] again. Since $\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda\|_\infty \le r\|\sum_{\lambda\in\Lambda_n}\varphi_\lambda\| = r\sqrt{N_n}$, we have

$$|\psi(\bar{f}_n) - \psi(\hat{f}_n)| \le r^2\sqrt{N_n}\sup_{\lambda\in\Lambda_n}|\nu_n^d(\varphi_\lambda)|$$

which is (6.21) with $\Phi = r^2$.

Replacing as in the proof of Theorem 3.1 the variables by their block-independent approximations imply the same constraint (6.16) as before, which is satisfied for the choice $q_n = [c\ln(n)/\theta]$ for geometrical β-mixing provided that $c \ge 2$. We choose $q_n = [4\ln(n)/\theta]$ for a reason that appears later. Then we split the probability space $\Omega = \Omega_1 \cup \Omega_1^c$, with

$$\Omega_1 = \{|\psi(\hat{f}_n) - \psi(\bar{f}_n)| \le \|f_A\|_\infty\}.$$

The proof of [Dβ] of Theorem 3.1 can be developed to bound $\mathbb{E}(\|f_A - \tilde{f}\|^2\mathbb{1}_{\Omega_1})$. Since on Ω_1,

$$\begin{aligned}
\sup_{m,m'}\|f_m - \hat{f}_{m'}\|_\infty &\le \psi(\bar{f}_n) + \psi(\hat{f}_n) \le |\psi(\hat{f}_n) - \psi(\bar{f}_n)| + 2\psi(\bar{f}_n) \\
&\le \|f_A\|_\infty + 2\psi(f_A) = (2r+1)\|f_A\|_\infty := C(f_A),
\end{aligned}$$

we can replace in $W^{d*}(m')$ the supremum on $B_{m,m'}(0,1)$ by

$$\tilde{W}^{d*}(m') = \left[\left(\sum_{t\in S^*_{m'},0\ne\|t-f_m\|_\infty\le C(f_A)}\left|\nu_n^{d*}\left(\frac{t-f_m}{\|t-f_m\|\vee x(m')}\right)\right|\right)^2 - p(m,m')\right]_+,$$

where $x(m')^2 = 8\ln^2(n)(D_m + D_{m'})/n$. Then (6.14) becomes

$$2|\nu_n^{d*}(\tilde{f} - f_m)| \le \frac{1}{4}\|f_m - f_A\|^2 + \frac{1}{4}\|f_A - \tilde{f}\|^2 + \frac{1}{8}x(m')^2 + 8\sum_{m'\in\mathcal{M}_n}\tilde{W}^{d*}(m') + 8p(m,\hat{m}).$$

Therefore we apply inequality (6.1) with

$$v = \frac{\|f_A\|_\infty}{4},\quad M_1 = \frac{C(f_A)}{2x(m')},\quad H^2 = \frac{q_n(D_m + D_{m'})}{2n}\|f_A\|_\infty,$$

starting from (6.20) and bounding $\sum_j \mathbb{E}(\varphi_j^2(X_1))$ by $D(m')\|f_A\|_\infty$ since $\int\varphi_j^2(x)\mathrm{d}x = 1$. This gives the bound

$$\mathbb{E}(\tilde{W}^{d*}(m')\mathbb{1}_{\Omega_1}) \le \frac{6}{K_1}\left(C_1\frac{\ln(n)}{n}\mathrm{e}^{-C_2\xi^2 D_{m'}} + \frac{C_3}{nD}\mathrm{e}^{-C_4\xi\sqrt{\ln(n)}D_{m'}}\right),$$

where

$$C_1 = \frac{2\|f_A\|_\infty}{\theta},\ C_2 = K_1,\ C_3 = \frac{8C^2(f_A)}{K_1\theta^2},\ C_4 = \frac{K_1\sqrt{\theta\|f_A\|_\infty}}{\sqrt{2}C(f_A)}$$

reminding that $q_n = 4\ln(n)/\theta$ and the choice:

$$\frac{1}{2}p(m,m') = \frac{4}{\theta}(4+\xi^2)\|f_A\|_\infty \frac{\ln(n)(D_m + D_{m'})}{n}.$$

Since $L_m = \ln(n/r)/r \le \ln(n)/r$ and (4.1) holds, *a fortiori* $\sum_{m'\in\mathcal{M}_n} \mathrm{e}^{-\ln(n)D_{m'}/r}$ is bounded by $\Sigma = e$. Therefore, choosing $\xi^2 = K\ln(n)$ with $K \ge \max(1/(rC_2), 1/(r^2C_4^2))$, we find that all terms are of order less than $(\ln(n)/n)\mathrm{e}^{-\ln(n)D_{m'}/r}$. Consequently we find a global order less that $\ln(n)^2/n$ for a penalty

$$\mathrm{pen}_d(m) \ge \frac{\tilde\kappa\|f_A\|_\infty}{\theta}\left\{1 + \max[1/(rK_1), 2(2r+1)^2\|f_A\|_\infty/(K_1^2r^2\theta)]\ln(n)\right\}\frac{\ln(n)D_m}{n},$$

where $\tilde\kappa$ is a numerical constant. Therefore choosing

$$\mathrm{pen}_d(m) = \frac{\kappa\|f_A\|_\infty}{\theta}\left(1 + \frac{\|f_A\|_\infty}{\theta}\right)\frac{\ln^2(n)D_m}{n}$$

implies that

$$\mathbb{E}(\|\tilde f^d - f_A\|^2 \mathbb{1}_{\Omega_1}) \le C\left(\|f_A - f_m\|^2 + \frac{\ln^2(n)D_m}{n}\right) \tag{6.22}$$

where C is a constant depending on $r, \|f_A\|_\infty, \theta$.

On the complementary of Ω_1, we use relation (6.21). Since $\tilde f$ can be seen as the projection of $\hat f_n$ on $S_{\hat m}$, we have

$$\begin{aligned}\|\tilde f\| &\le \|\tilde f\|_\infty \le \psi(\hat f_n) \le \psi(\bar f_n) + |\psi(\hat f_n) - \psi(\bar f_n)| \le \psi(f_A) + \Phi\sqrt{N_n}\sup_{\lambda\in\Lambda_n}|\nu_n^d(\varphi_\lambda)| \\ &\le \psi(f_A) + \Phi\Phi_0 N_n \text{ using that } \|\varphi_\lambda\|_\infty \le \Phi_0\sqrt{N_n}\,.\end{aligned}$$

Consequently

$$\mathbb{E}\left[\|f_A - \tilde f\|^2\mathbb{1}_{\Omega_1^c}\right] \le 2\|f_A\|^2\mathbb{P}(\Omega_1^c) + 2\mathbb{E}\left[\|\tilde f\|^2\mathbb{1}_{\Omega_1^c}\right] \le \left[2\|f_A\|^2 + 4(\psi(f_A)^2 + \Phi^2\Phi_0^2N_n^2)\right]\mathbb{P}(\Omega_1^c).$$

Therefore, we need to prove that $\mathbb{P}(\Omega_1^c) \le C/n^3$ for $N_n \le n$. With obvious notations and using (6.21), we successively write

$$\begin{aligned}\mathbb{P}(\Omega_1^c) &= \mathbb{P}(|\psi(\hat f_n) - \psi(\bar f_n)| \ge \|f_A\|_\infty) \le \mathbb{P}\left(\sup_{\lambda\in\Lambda_n}|\nu_n^d(\varphi_\lambda)| \ge \frac{\|f_A\|_\infty}{\Phi\sqrt{N_n}}\right) \\ &\le \sum_{\lambda\in\Lambda_n}\mathbb{P}\left(|\nu_n^d(\varphi_\lambda)| \ge \frac{\|f_A\|_\infty}{\Phi\sqrt{N_n}}\right) \\ &\le \sum_{\lambda\in\Lambda_n}\left[\mathbb{P}\left(|\nu_n^d(\varphi_\lambda) - \nu_n^{d*}(\varphi_\lambda)| \ge \frac{\|f_A\|_\infty}{2\Phi\sqrt{N_n}}\right) + \mathbb{P}\left(|\nu_n^{*d(1)}(\varphi_\lambda)| \ge \frac{\|f_A\|_\infty}{4\Phi\sqrt{N_n}}\right)\right. \\ &\quad \left. + \mathbb{P}\left(|\nu_n^{*d(2)}(\varphi_\lambda)| \ge \frac{\|f_A\|_\infty}{4\Phi\sqrt{N_n}}\right)\right],\end{aligned}$$

where the last two terms have the same order. For the first one, we write

$$\mathbb{P}\left(|\nu_n^d(\varphi_\lambda)-\nu_n^{d*}(\varphi_\lambda)|\geq\frac{\|f_A\|_\infty}{2\Phi\sqrt{N_n}}\right) \leq \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n(\varphi_\lambda(X_i)-\varphi_\lambda(X_i^*))\right|\frac{2\Phi\sqrt{N_n}}{\|f_A\|_\infty}$$
$$\leq \frac{4\Phi\sqrt{N_n}}{\|f_A\|_\infty}\|\varphi_\lambda\|_\infty\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{X_i\neq X_i^*}\right]\leq\frac{4\Phi\Phi_0N_n}{\|f_A\|_\infty}\beta_{q_n}$$

which gives an order $1/n^3$ for geometrical mixing [GEO] by choosing $q_n=[4\ln(n)/\theta]$.

For the other terms, we use Bernstein's inequality as recalled by Birgé and Massart [6]; namely, for $S_n=\sum_{i=1}^n Z_i$ and Z_i i.i.d., $\|Z_1\|_\infty\leq B$ and $\mathrm{Var}(Z_1)=\sigma^2$, we have for all positive η,

$$\mathbb{P}(S_n-\mathbb{E}(S_n)\geq n\eta)\leq\exp\left(-\frac{n\eta^2/2}{\sigma^2+B\eta}\right).$$

Applying the above inequality with $B=\Phi_0\sqrt{N_n}$ and $\sigma^2=\|f_A\|_\infty/4$, we find

$$\mathbb{P}\left(|\nu_n^{*d(1)}(\varphi_\lambda)|\geq\frac{\|f_A\|_\infty}{4\Phi\sqrt{N_n}}\right)\leq 2\mathbb{P}\left(\left|\frac{1}{p_n}\sum_{\ell=0}^{p_n-1}\left(\frac{1}{2q_n}\sum_{i=2\ell q_n+1}^{(2\ell+1)q_n}[\varphi_\lambda(X_i^*)-\mathbb{E}(\varphi_\lambda(X_i^*))]\right)\right|\geq\frac{\|f_A\|_\infty}{4\Phi\sqrt{N_n}}\right)$$
$$\leq 2\exp\left(-\frac{\|f_A\|_\infty}{8\Phi(\Phi+\Phi_0)}\frac{p_n}{N_n}\right)$$
$$\leq 2\exp\left(-K\frac{n}{N_n\ln(n)}\right)\leq 2\exp\left(-K\ln^2(n)\right)$$

for the above choice of q_n, and $N_n\leq n/\ln(n)^3$. K is a constant depending on $\|f_A\|_\infty$, Φ, Φ_0 and the mixing constant θ. To be precise, we need $N_n\leq (K/3)(n/\ln^2(n))$. This makes this term of order $1/n^3$ as well. Gathering all terms implies that for C or n great enough, we have $\mathbb{P}(\Omega_1^c)\leq C/n^3$. This implies that $\mathbb{E}(\|\tilde{f}^d-f_A\|^2\mathbb{1}_{\Omega_1^c})\leq C/n$ where C is a constant depending on $r,\theta,\|f_A\|_\infty$. This bound gathered with (6.22) implies that for all m in $\mathcal{M}_n$,

$$\mathbb{E}(\|\tilde{f}^d-f_A\|^2)\leq C\left(\|f_A-f_m\|^2+\frac{\ln^2(n)D_m}{n}\right)$$

where C is a constant depending on $r,\|f_A\|_\infty,\theta$. Using the approximation results recalled in Birgé and Massart [7], we know that $\|f_A-f_m\|^2$ is still of order D_m^{-2a}, for f_m in a well chosen S_m among those of dimension D_m. This achieves the proof. □

We would like to thank Lucien Birgé for useful discussions. We are also grateful to the referees and to Ph. Soulier for valuable suggestions which improved the presentation of this paper. All errors remain ours.

References

[1] G. Banon, Nonparametric identification for diffusion processes. *SIAM J. Control Optim.* **16** (1978) 380-395.
[2] G. Banon and H.T. N'Guyen, Recursive estimation in diffusion model. *SIAM J. Control Optim.* **19** (1981) 676-685.
[3] A.R. Barron, L. Birgé and P. Massart, Risk bounds for model selection *via* penalization. *Probab. Theory Related Fields* **113** (1999) 301-413.
[4] H.C.P Berbee, *Random walks with stationary increments and renewal theory.* Cent. Math. Tracts, Amsterdam (1979).
[5] L. Birgé and P. Massart, From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, edited by D. Pollard, E. Torgersen and G. Yang. Springer-Verlag, New-York (1997) 55-87.
[6] L. Birgé and P. Massart, Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** (1998) 329-375.
[7] L. Birgé and P. Massart, An adaptive compression algorithm in Besov spaces. *Constr. Approx.* **16** (2000) 1-36.

[8] L. Birgé and Y. Rozenholc, *How many bins must be put in a regular histogram?* Preprint LPMA 721, `http://www.proba.jussieu.fr/mathdoc/preprints/index.html` (2002).
[9] D. Bosq, Parametric rates of nonparametric estimators and predictors for continuous time processes. *Ann. Stat.* **25** (1997) 982-1000.
[10] D. Bosq, *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*, Second Edition. Springer Verlag, New-York, *Lecture Notes in Statist.* **110** (1998).
[11] D. Bosq and Yu. Davydov, Local time and density estimation in continuous time. *Math. Methods Statist.* **8** (1999) 22-45.
[12] W. Bryc, On the approximation theorem of Berkes and Philipp. *Demonstratio Math.* **15** (1982) 807-815.
[13] C. Butucea, Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM: P&S* **5** (2001) 1-31.
[14] J.V. Castellana and M.R. Leadbetter, On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21** (1986) 179-193.
[15] S. Clémençon, Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.* **9** (2000) 323-357.
[16] A. Cohen, I. Daubechies and P. Vial, Wavelet and fast wavelet transform on an interval. *Appl. Comput. Harmon. Anal.* **1** (1993) 54-81.
[17] F. Comte and F. Merlevède, *Density estimation for a class of continuous time or discretely observed processes*. Preprint MAP5 2002-2, `http://www.math.infor.univ-paris5.fr/map5/` (2002).
[18] F. Comte and Y. Rozenholc, Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.* **97** (2002) 111-145.
[19] I. Daubechies, *Ten lectures on wavelets*. SIAM: Philadelphia (1992).
[20] B. Delyon, *Limit theorem for mixing processes*, Technical Report IRISA. Rennes (1990) 546.
[21] R.A. DeVore and G.G. Lorentz, *Constructive approximation*. Springer-Verlag (1993).
[22] D.L. Donoho and I.M. Johnstone, Minimax estimation with wavelet shrinkage. *Ann. Statist.* **26** (1998) 879-921.
[23] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, Density estimation by wavelet thresholding. *Ann. Statist.* **24** (1996) 508-539.
[24] P. Doukhan, *Mixing properties and examples*. Springer-Verlag, *Lecture Notes in Statist.* (1995).
[25] Y. Efromovich, Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** (1985) 557-661.
[26] Y. Efromovich and M.S. Pinsker, Learning algorithm for nonparametric filtering. *Automat. Remote Control* **11** (1984) 1434-1440.
[27] G. Kerkyacharian, D. Picard and K. Tribouley, $\mathbb{L}_p$adaptive density estimation. *Bernoulli* **2** (1996) 229-247.
[28] A.N. Kolmogorov and Y.A. Rozanov, On the strong mixing conditions for stationary Gaussian sequences. *Theory Probab. Appl.* **5** (1960) 204-207.
[29] Y.A. Kutoyants, Efficient density estimation for ergodic diffusion processes. *Stat. Inference Stoch. Process.* **1** (1998) 131-155.
[30] F. Leblanc, Density estimation for a class of continuous time processes. *Math. Methods Statist.* **6** (1997) 171-199.
[31] H.T. N'Guyen, Density estimation in a continuous-time stationary Markov process. *Ann. Statist.* **7** (1979) 341-348.
[32] E. Rio, The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Probab.* **23** (1995) 1188-1203.
[33] M. Rosenblatt, A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. USA* **42** (1956) 43-47.
[34] M. Talagrand, New concentration inequalities in product spaces. *Invent. Math.* **126** (1996) 505-563.
[35] K. Tribouley and G. Viennet, $\mathbb{L}_p$ adaptive density estimation in a β-mixing framework. *Ann. Inst. H. Poincaré* **34** (1998) 179-208.
[36] A.Yu. Veretennikov, On hypoellipticity conditions and estimates of the mixing rate for stochastic differential equations. *Soviet Math. Dokl.* **40** (1990) 94-97.
[37] G. Viennet, Inequalities for absolutely regular sequences: Application to density estimation. *Probab. Theory Related Fields* **107** (1997) 467-492.