Nonparametric Estimation for some Specific Classes of Hidden Markov Models.

Stéphan Clémençon.
MODAL'X - Université Paris X Nanterre
Laboratoire de Probabilités et Modèles Aléatoires
UMR CNRS 7599 - Universités Paris VI et Paris VII

May 21, 2003

Abstract

We study the problem of nonparametric estimation of the stationary and transition densities of a regular Markov chain based on noisy observations when the density of the noise's terms $k(x)$ has a Fourier transform with decay of order $|w|^{-\alpha}$ as $w \to \infty$. Adopting the formalism of the wavelet-vaguelette decomposition (WVD), we propose estimation procedures based on thresholding of the WVD coefficients which are shown to be nearly optimal over a wide range of Besov classes for the variety of global $L^{p'}$ error measures, $1 \leqslant p' < \infty$.

1 Introduction.

Assume that a Markov chain $X = (X_n)_{n\in\aleph}$ with transition probability Π, and stationary probability μ, is observed through a process $Y = (Y_n)_{n\in\aleph}$ with distributions:

$$\begin{aligned}\mathcal{L}\left((Y_1, Y_2, ..., Y_n) \mid X_1, ..., X_n\right) &= \bigotimes_{i=1}^{n} \mathcal{L}(Y_i \mid X_i) \\ &= \bigotimes_{i=1}^{n} T\left(X_i, .\right),\end{aligned}$$

where T is a transition kernel from the state space of the chain X to the space in which the process Y is valued. The process $((X_n, Y_n))_{n\in\aleph}$ is still a Markov chain, but note that, in general, $(Y_n)_{n\in\aleph}$ is not. In the case when only $Y_1, ..., Y_n$ are observed, one traditionally says that the observations are partial or incomplete. The observable process $Y = (Y_n)_{n\in\aleph}$ is called *hidden Markov chain* (when $X = (X_n)_{n\in\aleph}$ may be called the *regime*).
The problem of estimating Π or μ based only on $Y_1, ..., Y_n$ is important (examples of partial observations are common in image analysis, study of DNA

sequences) and presents obvious difficulties (in particular with respect to identifiability problems). Neithertheless, in the parametric framework, several methods of statistical inference for Hidden Markov Models have been developped. Following the work of Baum & Petrie (1966) and Petrie (1969), most of them rely on the ideas of estimation by maximum-likelihood; theoretical results on the consistency of the maximum-likelihood estimator have been proved for general models (refer to Leroux (1992), Bickel & Ritov (1996), Bakry, Milhaud & Vandekerkhove (1997)), and recently practical procedures of estimation consisting in the search of minima for the contrast via recursive estimation have been proposed (see Khasminskii & Zeitouni (1996), Golubev & Khasminskii (1998)), many of them are based on stochastic algorithmic implementations (see for example Holst & Lindgren (1991) or Ryden (1997) for the use of the stochastic gradient algorithm, Vanderkerkhoeve (1996) for the use of simulated annealing, Chauveau & Vandekerkhove (1999) for the use of the Hastings-Metropolis algorithm).

In different settings like regression, spectral or probability density estimation, growing complexity of models used in physics, biology or speech processing justifies a nonparametric approach of the problem of estimation. For Hidden Markov Models, the latter may be formalized from the viewpoint of functional analysis and connected with the so-called *inverse problems theory.*

1.1 Nonparametric estimation in the case of partial observations: a linear inverse problem.

Suppose X and Y valued in subsets of $\Re$ and $\Pi(x, dy) = \pi(x, y)dy$, $\mu(dx) = f(x)dx$, as well as $T(x, dy) = t(x, y)dy$. To simplify, consider first the case of the estimation of the stationary density f. The matter is to recover f from observations $Y_1, ..., Y_n$ with common density:

$$fT(y) = \int f(x)t(x, y)dx.$$

The latter is the image of f by a linear transformation K_T. It seems impossible to solve this problem in full generality. When K_T^{-1} exists, as a bounded linear operator, it is natural to attempt to estimate f by using an estimator of the form:

$$\hat{f} = K_T^{-1}\hat{fT},$$

where $\hat{fT}$ is an estimator of fT; but unfortunately, in the cases which are of real interest, scientifically speaking, K_T is not invertible: the corresponding inverse problems are usually called "ill-posed", or more rarely "improperly-posed".

1.2 Diagonalization. Singular value decomposition methods.

Assuming further, for example, that the kernel $t(x,y)$ belongs to $L^2\left(\Re^2\right)$, the operator K_T is then a classical Hilbert-Schmidt operator of the Hilbert space $L^2\left(\Re\right)$, whose inner product and norm are denoted respectively $\langle .,. \rangle$ and $\|.\|_2$. According to very classical results in spectral analysis, K_T is a compact operator, as well as its adjoint K_T^*: it is still a Hilbert-Schmidt operator, whose kernel is simply the image of the kernel of K_T by the transformation $(x,y) \longmapsto (y,x)$. $K_T^* K_T$ is a compact autoadjoint operator, so there exists an orthonormal basis of L^2, $(e_n)_{n\in\aleph}$, which diagonalizes the latter. Denote by λ_n^2 the eigenvalue corresponding to the eigenfunction e_n. If none of them is equal to zero, from this diagonalization, we derive the reproducing formula:

$$f = \sum_n \lambda_n^{-1} \langle K_T f, h_n \rangle \, e_n, \tag{1}$$

where $h_n = \|K_T e_n\|_2^{-1} .K_T e_n$ is the renormalized image of the eigenfunction e_n. Then, one may fell an impulse to implement an estimation procedure based on the reproducing formula (1), that is to say to estimate $\langle K_T f, h_n \rangle$ by an empirical coefficient $\hat{\alpha}_n$ for all n in some finite subset S_N in order to form an estimator:

$$\hat{f}_N = \sum_{n\in S_N} \lambda_n^{-1} \hat{\alpha}_n e_n,$$

that provides a recovery whose quality can be measured by the L^2-norm for example.

The limitation of the singular value decomposition approach (SVD in abbreviated form) is obvious, it consists in the fact that there is no reason for a finite sum of weighted basis functions e_n , which derive from the operator K_T, to represent efficiently objects with inhomogeneous spatial variability, and in particular the function f whose smoothness properties do not depend on K_T a priori: this is precisely the advantage of the particular case of wavelet bases (a single orthonormal wavelet basis may serve as an unconditional basis for a wide class of Besov spaces, refer to Meyer (1990)).

1.3 The wavelet-vaguelette decomposition: an efficient method for dealing with a class of special inverse problems.

The idea to consider a decomposition of the operator in a wavelet basis instead of an eigenfunction basis had just been introduced in Donoho (1995), where the author developped a general formalism, the wavelet-vaguelette decomposition (we will write WVD), which apply to a class of specific operators and provides an efficient solution for the corresponding inverse problems. Before we show in detail how to adapt it to solve the problem of minimax estimation for a specific

class of hidden / noisy Markov models (see Section 2), let us describe shortly this method and the class of linear inverse problems to which it applies.
Let $d \in \{1,2\}$ and K be a bounded linear operator on $L^2\left(\Re^d\right)$ and $\left(\psi_\gamma\right)_{j \in Z,\ \gamma \in \Gamma_j^{(d)}}$ an orthonormal wavelet basis of $L^2(\Re^d)$ (see Daubechies (1990)). In the case when each wavelet ψ_γ belongs to the range $\mathcal{R}\left(K^*\right)$ of the adjoint of K, K^*, for each function g in the domain of K, $\mathcal{D}\left(K\right)$, we can derive the following reproducing formula from its wavelet expansion:

$$\begin{aligned} g &= \sum_{j,\ \gamma \in \Gamma_j^{(d)}} \left\langle g, \psi_\gamma \right\rangle \psi_\gamma \\ &= \sum_{j,\ \gamma \in \Gamma_j^{(d)}} \left\langle Kg, \xi_\gamma \right\rangle \psi_\gamma, \end{aligned}$$

where $\xi_\gamma \in \mathcal{D}\left(K^*\right)$ is such that $K^*\xi_\gamma = \psi_\gamma$. Further, the theory calls for the assumption that it is possible to identify constants λ_j, depending on the level index j but not on the spatial index, such that the renormalized functions

$$u_\gamma = \lambda_j.\xi_\gamma,\ \gamma \in \Gamma_j^{(d)}$$

make a set of functions with norms bounded above and below, and having the property that there exists a constant C such that for all sequence of coefficients $\left(\alpha_\gamma\right)$:

$$\left\| \sum_{j,\ \gamma \in \Gamma_j^{(d)}} \alpha_\gamma u_\gamma \right\|_2 \leqslant C \left(\sum_{j,\ \gamma \in \Gamma_j^{(d)}} \alpha_\gamma^2 \right)^{1/2}. \tag{2}$$

The latter assumption is always satisfied by some special systems, the "vaguelettes" (refer to Theorem 2 p 270 in Meyer Vol. II (1990)), so called because of the qualitative features they have in common with wavelets (localization, cancellation, smoothness), here is their definition.

Definition 1 *Continuous functions on $\Re^d$ $\left(u_{jk}\right)_{(j,k) \in Z \times Z^d}$ are called vaguelettes if there exist $\alpha > \beta > 0$ and a constant C such that*

$$\begin{aligned} \left|u_{jk}(x)\right| &\leqslant C 2^{dj/2} \left(1 + \left|2^j x - k\right|\right)^{-d-\alpha}, \\ \int u_{jk}(x)dx &= 0, \\ \left|u_{jk}(x) - u_{jk}(x')\right| &\leqslant C 2^{(d/2+\beta)j} \left|x - x'\right|^\beta. \end{aligned}$$

The property (2) is crucial for the theory as we shall see.

Hence, the theory applies if K^* is mapping vaguelettes into wavelets, and it is the case for some specific operators having in common the property of homogeneity with respect to dilatation (refer to the examples treated in Donoho

(1995), among which: Radon transform, fractional integration, special convolution operators), that is to say such that KD_a is equal to $a^{\alpha d} D_a K$ either exactly or approximately for some exponent $\alpha \geq 0$, and all a in $\Re$, D_a denoting the operator defined by $D_a f(x) = f(ax)$, and for which the constants λ_j have norms decreasing geometrically in the resolution index: $\lambda_j = 2^{-j\alpha d}$. In this particular case, if one introduces a scaling function ϕ associated to the wavelet mother ψ, from the coefficients in the wavelet expansion of ϕ_λ with $\lambda \in \Lambda_{j_1}^{(d)}$, namely

$$\phi_\lambda = \sum_{j \leqslant j_1} \sum_{\gamma \in \Gamma_j^{(d)}} p_{\lambda,\,\gamma} \psi_\gamma,$$

one may define

$$\Delta_\lambda = \sum_{j \leqslant j_1} \sum_{\gamma \in \Gamma_j^{(d)}} 2^{j\alpha d} p_{\lambda,\,\gamma} u_\gamma,$$

which satisfies to the relation

$$K^* \Delta_\lambda = \phi_\lambda.$$

As a matter of fact, from (2) and the condition $\|\phi_\lambda\|_2 = 1$, we deduce that:

$$\begin{aligned} \left\| \sum_{j \leqslant j_1} \sum_{\gamma \in \Gamma_j^{(d)}} 2^{j\alpha d} p_{\lambda,\,\gamma} u_\gamma \right\|_2 &\leqslant C 2^{j_1 \alpha} \left(\sum_{j \leqslant j_1} \sum_{\gamma \in \Gamma_j^{(d)}} p_{\lambda,\,\gamma}^2 \right)^{1/2} \\ &= C 2^{j_1 \alpha d}. \end{aligned}$$

Hence, one may write an *inhomogeneous reproducing formula:*

$$g = \sum_{\lambda \in \Lambda_{j_1}^{(d)}} \langle Kg, \Delta_\lambda \rangle \phi_\lambda + \sum_{j \geq j_1} \sum_{\gamma \in \Gamma_j^{(d)}} \lambda_j^{-1} \langle Kg, u_\gamma \rangle \psi_\gamma. \tag{3}$$

From representation (3) of the object g to recover, known to lie in a Besov class for example, it is possible to implement a procedure based on the computation of the corresponding empirical coefficients from the indirect data, which, combined with a suitable nonlinear shrinkage of these coefficients, provides an estimator with a minimax optimality. Results in that direction have been proved in Donoho (1995) in a framework assuming that data containing measurements of the function Kg contaminated with Gaussian noise are observed:

$$d(t) = Kg(t) + z(t), \qquad t \in \mathcal{T}.$$

Here, although our framework is different, since the data consist in realizations of a process (Y_n) with marginal distribution Kg, we shall show that, in some specific cases, nonlinear shrinkage of wavelet-vaguelette decomposition coefficients leads to a minimax optimal procedure too.

1.4 Contents.

The outline of the paper is as follows. In section 2 the WVD is applied to nonparametric estimation of the stationary and transition densities, $f(x)$ and $\pi(x,y)$ respectively, of a Markov chain based on noisy observations: a precise description of the statistical model we consider is given in 2.1, subsection 2.2 shows how the WVD formalism applies to the problem of nonparametric estimation for this specific model, in 2.3 we set out our proposals for adaptive estimation of $f(x)$ and $\pi(x,y)$ in this framework, and subsection 2.4 states the main results on the minimax optimality (or nearly optimality) of these procedures. Technical proofs are given in section 3.

2 Application of WVD to nonparametric estimation for Markov Chains on the basis of noisy observations.

We now turn to application of the wavelet-vaguelette decomposition to the problem of minimax nonparametric estimation of the stationary and transition densities, $f(x)$ and $\pi(x,y)$, on compact sets (on the unit interval $[0,1]$ and on the unit square $[0,1]^2$ to simplify notation) for the specific Hidden/Noisy Markov model described below. The quality of the estimators we propose will be measured by the $L^{p'}$-integrated risk.

2.1 Statistical model.

Assumptions. We consider a real valued Markov chain $(X_n)_{n\in\aleph}$, Feller, aperiodic, recurrent positive and stationary with unknown transition probability $\Pi(x,dy)=\pi(x,y)dy$ and stationarity probability $\mu(dx)=f(x)dx$ (recall that $f(y)=\int\mu(dx)\pi(x,y)$, a. s.). In addition, we consider assumptions involving criteria of speed of return times to "small sets", whose definition we recall below.

Definition 2 *For a Markov chain valued in a state space $(E,\mathcal{E})$ countably generated with transition probability Π, a set $S\in\mathcal{E}$ is said to be small if there exist an integer $m>0$, a probability measure ν supported by S, and $\delta>0$ such that*

$$\forall x\in E,\forall A\in\mathcal{E},\qquad \Pi^m(x,A)\geq\delta 1_S(x)\nu(A),$$

denoting by Π^m the m-th iterate of Π. When this holds, one says that the chain satisfies the minorization condition $\mathcal{M}(m,S,\delta,\nu)$.

Recall that accessible small sets do exist for irreducible chains (see Jain & Jamison (1967)) and that, under the assumption of positive recurrence, we have for every small set S such that $\mu(S)>0$

$$\sup_{x\in S}E_x(\tau_S)<\infty,$$

E_x denoting the expectation conditionally to $X_0 = x$ and
$\tau_S = \inf\{n \geq 1,\ X_n \in S\}$ the return time to S, and there exist some measurable test function $V : (E,\, \mathcal{E}) \to [0,\, \infty]$, μ-integrable, uniformly bounded on S (continuous when the chain is Feller), and some constant $0 < b < 1$ such that the following *drift condition towards* S, denoted by $\mathcal{D}(S, V, b)$, is satisfied

$$\forall x \in E,\ \Pi V(x) \leqslant V(x) - 1 + b1_S(x),$$

with $\Pi V(x) = \int_{y \in E} \Pi(x, dy)V(y)$. Strenghtening this property of finiteness of return times to small sets, we introduce the following notions.

Definition 3 *For a Ψ-irreducible Markov Chain $(X_n)_{n\in\aleph}$, a measurable set A is said geometrically regular if there exists $r_B > 1$ such that $\sup_{x\in A} E_x(r_B^{\tau_B}) < \infty$) for every measurable set B such that $\Psi(B) > 0$.*

Remark 4 *Recall that a small set S such that $\Psi(S) > 0$ is geometrically regular if and only if there exists $r > 1$ such that $\sup_{x\in S} E_x(r^{\tau_S}) < \infty$; if there exists such a set for a positive recurrent chain with invariant probability μ, then every other small set weighted by μ is geometrically regular, and the support of μ can be covered by a countable collection of geometrically regular sets; in such a case, one says that the chain is geometrically regular, one can refer to Chapter 5 in Nummelin (1984) and Chapter 15 in Meyn & Tweedie (1996) for further detail.*

The model. Here we focus on the hidden Markov chain $(Y_n)_{n\in\aleph}$ which is a perturbation of $(X_n)_{n\in\aleph}$ by a specific additive white noise $(\varepsilon_n)_{n\in\aleph}$:

$$Y_n = X_n + \varepsilon_n,$$

(ε_n) is assumed to be a sequence of i.i.d. random variables with known density $k(x)$, ε_n and X_n are supposed independent. Moreover, the density k is supposed bounded and obeys the conditions of high-frequency regularity:

$$\hat{k}(w) \quad \sim \quad |w|^{-\alpha},\ \text{as}\ |w| \to \infty, \tag{4}$$

$$\frac{\inf_{|w|\leqslant|\Omega|}\left|\hat{k}(w)\right|}{\left|\hat{k}(\Omega)\right|} \quad \to \quad 1,\ \text{as}\ |\Omega| \to \infty, \tag{5}$$

where $\hat{k}$ denotes the Fourier transform of k, with the normalization: $\hat{k}(w) = \int e^{-ixw}k(x)dx$. The common density of the observations $Y_1, ..., Y_n$ is the image of $f(x)$ by the convolution operator $K^{(1)}$ with kernel $k_1 = k$:

$$K^{(1)}f = k * f.$$

Similarly, the common density of the bivariate r. v. $(Y_i,\ Y_{i+1})$ is the image of $F(x, y) = f(x)\pi(x, y)$ by the convolution operator $K^{(2)}$ with kernel $k_2 = k \otimes k$.

Properties. The fact that $Z = ((X_n, Y_n))_{n\in\aleph}$ is a bivariate Markov chain is noteworthy:

$$P\left(X_{n+1} \in A, Y_{n+1} \in B \mid X_n = x\right) = \int_{(z,y)\in A\times B} \Pi(x, dz)k(y-z)dzdy, \qquad (6)$$

for all measurable sets A, B, and for all x in the state space of the regime X. From expression (6) of the transition kernel above, one can check straightforwardly that Z inherits stochastic stability properties from X : it is Feller, aperiodic, recurrent positive with stationary probability
$\mu K(dxdy) = f(x)k(y-x)dxdy$; further, observe that $S \times \Re$ is a small set for the bivariate chain Z, as soon as S is a small set for the chain X, hence if X is geometrically regular, so is Z. This latter observation is crucial, these assumptions considering criteria of speed of return times to small sets allow to make use of moment inequalities and large deviations bounds, valid for additive functionals of regular Markov chains (see Clémençon (2001)), so that risk bounds for our estimators may be calculated.

Smoothness constraints. Let $d \in \{1, 2\}$, we will use expansion of the object g to estimate on $[0,1]^d$ in an orthonormal wavelet basis on $[0,1]^d$ (refer to Cohen, Daubechies & Vial (1993)) of class $\mathcal{C}^{r+1}$, $r + 1 \geq 2^J - 1$, to cover up possible boundary effects. In order not to overcharge notation, we make no notational distinction between *edge* and *interior* wavelets (respectively, between *edge* and *interior* scaling functions), and indiscriminately denote ψ_γ, $\gamma \in \Gamma_j^{(d)} = \left\{(j,k,m),\ k \in \left\{0,...,2^{j-1}\right\}^d,\ m \in \{1,...,2d-1\}\right\}$, and ϕ_λ, $\lambda \in \Lambda_j^{(d)} = \left\{(j,k),\ k \in \left\{0,...,2^j - 1\right\}^d\right\}$ the wavelet and scaling functions at level $j \geq J$. Suppose $r + 1 > \sigma$, the object g to estimate on $[0,1]^d$ is known to belong to the ball of center 0 and radius R of the Besov space $B_{\sigma pq}\left([0,1]^d\right)$, that is to say to the set of functions defined on $[0,1]^d$ obeying

$$\|g\|_{\sigma pq}^{(d)} = \left(\sum_{\lambda\in\Lambda_J^{(d)}} |\alpha_\lambda|^p\right)^{1/p} + \left(\sum_{j\geq J} 2^{j(\sigma+d/2-d/p)q}\left(\sum_{\gamma\in\Gamma_j^{(d)}} \left|\beta_\gamma\right|^p\right)^{1/p}\right)^{1/q} \leqslant R,$$

where, denoting by $\langle .,. \rangle$ the usual inner product on $L^2\left([0,1]^d\right)$, $\alpha_\lambda = \langle g, \phi_\lambda\rangle$, $\beta_\gamma = \left\langle f, \psi_\gamma\right\rangle$. In what follows, we denote by $\mathcal{F}_{\sigma pq}^{(d)}(R)$ the set of such functions, by $\mathcal{P}_{\sigma pq}(R)$ the set of positive recurrent transition densities $\pi(x,y)$ whose stationary density $f(x)$ restricted to $[0,1]$ belongs to $\mathcal{F}_{\sigma pq}^{(1)}(R)$ and such that the bivariate density $f(x)\pi(x,y)$ restricted to $[0,1]^2$ belongs to $\mathcal{F}_{\sigma pq}^{(2)}(R)$, by $\|.\|_{p'}$ the usual $L^{p'}$-norm on $[0,1]^d$, and the notation $2^{j(n)} \simeq g(n)$ means that the sequence of integers $j(n)$ is chosen to satisfy $2^{j(n)} \leqslant g(n) < 2^{j(n)+1}$.

2.2 The WVD of the convolution operator $K^{(d)}$.

Consider an orthonormal basis of wavelets of compact support on $[0,1]^d$, (ψ_γ), deriving from "edge" and "interior" wavelet mothers $(\psi_m)_{1\leqslant m\leqslant 2d-1}$ of class $\mathcal{C}^{r+1}$ with $r+1\geq\alpha$, and let ϕ denotes the corresponding "edge" and "interior" scaling functions (of class $\mathcal{C}^{r+1}$ too).
Then, the following equalities define functions in $L^1(\Re^d)\cap L^2(\Re^d)$:

$$\xi_\gamma(x)=\frac{1}{(2\pi)^d}\int e^{ix.w}\frac{\hat{\psi}_\gamma(w)}{\hat{k}_d(-w)}dw,\ \text{for } j\geq J,\ \gamma\in\Gamma_j^{(d)},$$

as a matter of fact, they can be rewritten as follows:

$$\xi_\gamma(x)=\frac{1}{(2\pi)^d}\int e^{ix.w}\hat{\psi}_\gamma(w)\,|w|^{\alpha d}\,\frac{1}{|w|^{\alpha d}\,\hat{k}_d(w)}dw.$$

These functions are such that:

$$K^{(d)*}\xi_\gamma=\breve{k}_d*\xi_\gamma=\psi_\gamma,$$

where $K^{(d)*}$ denotes the adjoint of $K^{(d)}$, the convolution operator with kernel $\breve{k}_d(x)=k_d(-x)$. Note that, at each level of resolution j, the $\xi_{(j,\ k,\ m)}$'s corresponding to interior (respectively, edge) wavelets are deduced from each other by translation. Moreover, from Plancherel 's formula we have in the case $d=1$

$$\left\|\xi_{(j,\ k,\ 1)}\right\|_2=\frac{2^{j\alpha}}{(2\pi)^{1/2}}\left(\int_{\Re}\left|\hat{\psi}\,(w)\right|^2\frac{|w|^{2\alpha}}{|2^jw|^{2\alpha}\left|\hat{k}\,(-2^jw)\right|^2}dw\right)^{1/2},$$

and in the case $d=2$

$$\left\|\xi_{(j,\ k,\ m)}\right\|_2=\frac{2^{2j\alpha}}{2\pi}\left(\int_{\Re^2}\left|\hat{\psi}_m(w,t)\right|^2\frac{|wt|^{2\alpha}\,dwdt}{|2^{2j}wt|^{2\alpha}\left|\hat{k}(-2^jw)\hat{k}(-2^jt)\right|}\right)^{1/2},$$

with ψ_m defined by $\psi_{(j,k,m)}(x,y)=2^{2j}\psi_m(2^jx,2^jy)$.
Consequently, by virtue of (4) and (5)

$$\left\|\xi_{(j,k,m)}\right\|_2\sim 2^{j\alpha d},\ \text{as } j\to\infty.$$

Hence the norms $\left\|\xi_{(j,\ k,\ m)}\right\|_2$ scale geometrically for $j\to\infty$

(but not for $j\to-\infty$: $\left\|\xi_{(j,\ k,\ m)}\right\|_2\to 1/\left|\hat{k}(0)\right|=1$). This leads us to set

$$\begin{aligned}\lambda_j^{(d)} &= 2^{-j\alpha d}\ \text{for } j\geq 0,\\ \lambda_j^{(d)} &= 1\ \text{for } j\leqslant 0,\end{aligned}$$

and for $j \geq J$, $\gamma \in \Gamma_j^{(d)}$,

$$u_\gamma = \lambda_j^{(d)}.\xi_\gamma.$$

The set of functions (u_γ) is a system of vaguelettes, it simply arises from a classical result concerning Fourier multipliers in standard Fourier analysis (see lemma 4 in Donoho (1995)):

Lemma 5 *Suppose that $\psi : \Re^d \to \Re$ is compactly supported, is orthogonal to polynomials of degree inferior or equal to D, and has r continuous derivatives. Let $\hat{\Omega}(w)$ be homogeneous of degree 0, and $|\alpha| + D + 1 < \min(r, D)$. Define $u : \Re^d \to \Re$ by*

$$u(x) = \frac{1}{(2\pi)^d} \int_{\Re^d} e^{ix.w} \hat{\psi}(w) \hat{\Omega}(w) \left| w \right|^\alpha dw.$$

Then there exists a constant C such that:

$$\begin{aligned} |u(x)| &\leqslant C \left(1 + |x|\right)^{-(1+d)}, \; x \in \Re^d, \\ \int u(x) dx &= 0, \\ |u(x) - u(x')| &\leqslant C \left| x - x' \right|, \; (x, y) \in \Re^d \times \Re^d. \end{aligned}$$

Keeping the notations introduced in 1.3, here and throughout we continue to denote by Δ_λ the functions such that:

$$K^{(d)*} \Delta_\lambda = \phi_\lambda.$$

Example 6 *Consider the case of exponential noise: $k(x) = 1_{\{x \geq 0\}} e^{-x}$. As k is of class $\mathcal{C}^1$ on $\Re_+^*$ and $\Re_-^*$, has a jump of amplitude 1 in 0, and k' is integrable, we deduce that*

$$\hat{k}(w) \sim \frac{1}{w}, \text{ as } |w| \to \infty.$$

Hence, here the exponent is $\alpha = 1$. Moreover, by simply integrating by parts, one obtains the identity:

$$\left(I - \frac{d}{dx} \right) K^* = I,$$

which leads to take:

$$\gamma_{jk}(x) = 2^{j/2} \psi \left(2^j x - k \right) - 2^{3j/2} \psi' \left(2^j x - k \right),$$

and

$$u_{jk}(x) = 2^{-j/2} \psi \left(2^j x - k \right) - 2^{j/2} \psi^{'} \left(2^j x - k \right), \text{ for } j \geq 0.$$

Similarly

$$\theta_{jk}(x) = 2^{j/2}\phi\left(2^j x - k\right) - 2^{3j/2}\phi'\left(2^j x - k\right).$$

As another example, one can consider the case for which $k(x) = \frac{1}{2}e^{-|x|}$*. It is immediate to check that it corresponds to the case* $\alpha = 2$*, and that the following identity holds:*

$$\left(I - \frac{d^2}{dx^2}\right)K^* = I.$$

Hence in this particular case the system of vaguelettes would be constructed from:

$$\gamma_{jk}(x) = 2^{j/2}\psi\left(2^j x - k\right) - 2^{5j/2}\psi''\left(2^j x - k\right).$$

Remark 7 *Analogy with SVD. As emphasized in Donoho (1995), the WVD, in some sense similar to the SVD, offers an "almost diagonal" representation of the operator* $K^{(d)}$*. As a matter of fact, set* $v_\gamma = (\lambda_j^{(d)})^{-1}.K^{(d)}\psi_\gamma$ *for* $\gamma \in \Gamma_j^{(d)}$*, by lemma 5 again,* (v_γ) *is a system of vaguelettes too. Hence, to within scalar multipliers,* $K^{(d)}$ *is mapping wavelets into vaguelettes, when* $K^{(d)*}$ *is turning vaguelettes into wavelets and so, wavelets are "almost eigenfunctions" of* $K^{(d)*}K^{(d)}$*. As the exact eigenfunctions (globalized complex exponentials) have not the localization properties of wavelets, the WVD renounces exact diagonalization of* $K^{(d)*}K^{(d)}$*, on which the SVD is based (refer to 1.2), in order to get much better representation of the object to estimate.*

2.3 Two algorithms for adaptive estimation of stationary and transition densities based on noisy observations.

The main purpose of this paper is to develop two practical algorithms, both based on thresholding of WVD coefficients (for computational aspects of the wavelet transforms, refer to Chapter 12 in Härdle, Kerkyacharian, Picard & Tsybakov (1998) and Vidakovic (1999)), for adaptive estimation of the stationary density $f(x)$ on $[0,1]$ and the transition density $\pi(x,y)$ on $[0,1]^2$ on the basis of observations $Y_1, ..., Y_n$ drawn from the *hidden/noisy* Markov model we described in 2.1.

2.3.1 Our proposal for adaptive estimation of the stationary density.

We propose a three steps method for estimation of $f(x)$ based on n data $Y_1, ..., Y_n$, which requires no iteration. We suppose we are given an orthonormal wavelet basis on $[0,1]$. Given this tool, we construct the estimate as follows.

Algorithm 8 [1] *From the indirect data* $Y_1, ..., Y_n$*, calculate the empirical vaguelette coefficients of the density* $K^{(1)}f(x)$ *for a specific range of indices by the method of moments, yielding estimates:*

$\hat{\alpha}_\lambda = n^{-1}\sum_{i=1}^n \Delta_\lambda(Y_i)$, $\hat{\beta}_\gamma = n^{-1}\sum_{i=1}^n \xi_\gamma(Y_i)$, $j_1^{(1)} \leqslant j \leqslant j_0^{(1)}$, $\lambda \in \Lambda_{j_1^{(1)}}^{(1)}$, $\gamma \in \Gamma_j^{(1)}$.

[2] *Apply the hard-threshold nonlinearity* $\delta_h(w,t) = w1_{\{|w|>t\}}$ *to the coefficients* $\hat{\beta}_\gamma$, $\gamma \in \Gamma_j^{(1)}$, *previously calculated with threshold* $t_j^{(1)} = \mathcal{K}^{(1)}\sqrt{j2^{2j\alpha}/n}$, *getting new coefficients* $\tilde{\beta}_\gamma$.

[3] *Setting all wavelet-vaguelette coefficients equal to* 0 *for* $j > j_0^{(1)}$, *perform the reconstruction by inverting the wavelet-vaguelette transform. This yields an estimate* $\hat{f}(x)$, $x \in [0,1]$.

2.3.2 Our proposal for adaptive estimation of the transition density.

We propose a seven step method for estimation of $\pi(x,y)$ based on n data $Y_1, ..., Y_n$, with no iteration required. We suppose we are given an orthonormal wavelet basis on $[0,1]$ and a 2-dimensional orthonormal wavelet basis on $[0,1]^2$ (constructed from the latter by the method of tensorial product so as to preserve the multiresolution analysis structure).

Algorithm 9 [1], [2], [3] *Execute the three steps of Algorithm 8.*

[4] *From the indirect data* $(Y_i,\ Y_{i+1})$, $i = 1, ..., n-1$, *calculate the empirical 2-d vaguelette coefficients for some specific class of levels of resolution*: $\hat{\alpha}_\lambda = n^{-1}\sum_{i=1}^{n-1} \Delta_\lambda(Y_i, Y_{i+1})$, $\hat{\beta}_\gamma = n^{-1}\sum_{i=1}^{n-1} \xi_\gamma(Y_i, Y_{i+1})$, $j_1^{(2)} \leqslant j \leqslant j_0^{(2)}$, $\lambda \in \Lambda_{j_1^{(2)}}^{(2)}$, $\gamma \in \Gamma_j^{(2)}$.

[5] *Apply hard thresholding with threshold* $t_j^{(2)} = \mathcal{K}^{(2)}\sqrt{j2^{4j\alpha}/n}$ *to the coefficients* $\hat{\beta}_\gamma$, $\gamma \in \Gamma_j^{(2)}$, *to get the resulting coefficients* $\tilde{\beta}_\gamma$.

[6] *From the coefficients* $\hat{\alpha}_\lambda$, $\lambda \in \Lambda_{j_1^{(2)}}^{(2)}$ $\tilde{\beta}_\gamma$, $\gamma \in \Gamma_j^{(2)}$, $j_1^{(2)} \leqslant j \leqslant j_0^{(2)}$, *only, invert the inhomogeneous wavelet-vaguelette transform, producing an estimate* $\hat{F}(x,y)$, $(x,y) \in [0,1]^2$.

[7] *Divide* $\hat{F}(x,y)$ *by* $\hat{f}(x)$ *when non equal to zero, getting the estimate*

$$\hat{\pi}(x,y) = \hat{F}(x,y)/\hat{f}(x),$$

and set $\hat{\pi}(x,y) = 0$ *when* $\hat{f}(x) = 0$.

2.3.3 Insights.

Attention of the reader, who wish to get an insight into the ground for efficiency of the algorithms above without going into technical details of minimax rates of convergence, must be turned to the fact it is based on two kinds of arguments.

As pointed out in 1.2, approximation-theoretic arguments justify the use of wavelet bases in nonparametric estimation. Following the line of argument of *Computational Harmonic Analysis* (CHA), the main advantage that can be found to use wavelet expansions in statistical applications rests in the fact that a single orthonormal wavelet basis provides an unconditional basis for a large scale of Besov spaces (functions belonging to these classes are characterized by the amplitude of their wavelet coefficients only) and for this reason allows to represent these Besov classes in an "optimal" fashion: following the formalism introduced by Donoho (1993,1996), an "optimal basis" for a given function class $\mathcal{F}$ is any basis in which the coefficients of the functions lying in $\mathcal{F}$, rearranged in decreasing order of their magnitude, have fastest decay. By considering the sparsity of representation of Besov classes provided by wavelet bases, one may gain insights of the reasons why an inference method based on the estimation of a few big enough wavelet coefficients (in our case, wavelet-vaguelette coefficients) yields an estimation procedure with drastically reduced bias over these functional classes, when suitably tuned.

If the bias errors of the procedures above may be reduced thanks to the use of wavelets, the stochastic errors, due to randomness of the observations, are smallest for chains that come back infinitely often and fast enough to specific subsets, namely "small sets" (see Definition 2). Besides, it seems heuristically evident that the estimation of the conditional density $\pi(x,.)$ is possible only if the chain visits "frequently enough" the neighbourhoods of x. Hence, if one consents to view the stationary probability μ as an occupation measure (let S be any small set and A be any measurable set, recall that $\mu(A)$ is proportional to the mean of amount of time spent in A between two consecutive visits to S), one intuitively understands why Algorithm 9 successfully achieves its goal in the case when the stationary density f is bounded away from zero on $[0,1]$ (notice that, in this case, the denominator of the estimator $\hat{\pi}(x,y)$ is strictly positive with overwhelming probability).

Finally, it is fitting to notice that implementation of both algorithms does not rely on the data alone, and requires the specification of parameters $j_1^{(d)}$, $j_0^{(d)}$ controlling the expansiveness of the empirical wavelet transforms, as well as the constants $\mathcal{K}^{(d)}$ in the thresholds. Basing on explicit computation of the rates of convergence for $L^{p'}$-integrated risks of the corresponding estimators, we shall propose to select these parameters according to the maximal degree of smoothness r the object(s) to estimate a priori may have (see subsection 2.4).

2.4 Optimality of the proposals. Asymptotics of minimax risks.

Now we show that the nonlinear shrinkage of the empirical wavelet-vaguelette coefficients on which Algorithms 8 and 9 are based can be tuned to be asymptotically minimax or nearly minimax over a wide range of Besov type smoothness constraints for the variety of global $L^{p'}$ error measures, $1 \leqslant p' < \infty$.

2.4.1 Lower bounds results.

Here we state (without proof) a lower bounds result for the model described in 2.1 for comparison purposes. For notational convenience, we set

$$\begin{aligned}\sigma_d &= \sigma - d(1/p + 1/p'), \varepsilon_\alpha^{(d)} = \sigma p - d\,(p' - p)\,(\alpha + 1/2)\,, \\ \nu_\alpha^{(d)} &= \min\left(\frac{\sigma}{d + 2\,(\sigma + d\alpha)}, \frac{\sigma_d}{d + 2\,(\sigma + d\alpha - d/p)}\right).\end{aligned}$$

Notice that

$$\nu_\alpha^{(d)} = \frac{\sigma}{d + 2\,(\sigma + d\alpha)} 1_{\left\{\varepsilon_\alpha^{(d)} \geq 0\right\}} + \frac{\sigma_d}{d + 2\,(\sigma + d\alpha - d/p)} 1_{\left\{\varepsilon_\alpha^{(d)} < 0\right\}}.$$

Theorem 10 *Let $1 \leqslant p,\ q \leqslant \infty$, $\sigma > 1/p$, $p' \geq 1$. Set*

$$\begin{aligned}\mathcal{R}_n^{(1)}\,(\sigma,\ p,\ q,\ R) &= \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}_{\sigma pq}^{(1)}(R)} \left(E_f \left\|\hat{f}_n - f\right\|_{p'}^{p'}\right)^{1/p'}, \\ \mathcal{R}_n^{(2)}(\sigma,\ p,\ q,\ R) &= \inf_{\hat{\pi}_n} \sup_{\pi \in \mathcal{P}_{\sigma pq}(R)} \left(E_\pi \left\|\hat{\pi}_n - \pi\right\|_{p'}^{p'}\right)^{1/p'}\end{aligned}$$

(the infimum being taken over all estimators based on observations $Y_1, ..., Y_n$). Then, for $d \in \{1,\ 2\}$ there exists some constant C such that

$$\mathcal{R}_n^{(d)}\,(\sigma,\ p,\ q,\ R) \geq C\left(1_{\left\{\varepsilon_\alpha^{(d)} > 0\right\}} + 1_{\left\{\varepsilon_\alpha^{(d)} \leqslant 0\right\}} \log^{\nu_\alpha^{(d)}} n\right) n^{-\nu_\alpha^{(d)}}.$$

Remark 11 *Observe that the lower bounds explicitely depend on the parameter α, more precisely they are nondecreasing functions of α. This crucial fact has already been discussed in Fan (1991), which paper aimed to show why the rate in deconvolution problems should depend on the tail of the Fourier transform of the error distribution $k(x)$, or in other terms on smoothness properties of $k(x)$ (which phenomenon we can sum up this way: the smoother the error distribution is, the harder the deconvolution is).*

Remark 12 *As in many other statistical problems, the lower bounds reveal "elbows" in the rates of convergence: note that, from an observation of length n, the rate $n^{-\frac{\sigma}{d+2(\sigma+d\alpha)}}$ applies only for σ large enough, whereas the rate is $n^{-\frac{\sigma_d}{d+2(\sigma+d\alpha-d/p)}}$ in the low regularity cases (note also the* log *factor). As will be shown in 2.4.2, for geometrically regular chains, this lower bound is sharp (up to possible logarithmic factors, see Theorems 13 and 15).*

Similarly to the proof given in Clémençon (2000b) in the case of estimation from direct data, these lower bounds for optimal rates of estimation over $\mathcal{F}_{\sigma pq}^{(1)}$ (respectively, $\mathcal{P}_{\sigma pq}$) among all estimators based on the indirect observations $Y_1, ..., Y_n$ can be established by using the *method of rectangular subproblems.*

Two cases have to be considered depending on whether $\varepsilon_{\alpha}^{(d)}$ is greater than zero or not: in both cases, an appropriate orthogonal hypercube may be obtained by considering perturbations of the density $1_{[0,1]^d}$, where the perturbations are the vaguelettes $\upsilon_{\gamma} = \left(\lambda_j^{(d)}\right)^{-1} .K^{(d)}\psi_{\gamma}$. A possible proof based on this technique is developped at lenght in Clémençon (2000a).

2.4.2 Upper bounds results.

We attempt to recover the stationary density $f(x)$ on $[0,1]$ and the transition density $\pi(x,y)$ on $[0,1]^2$ as accurately as possible from indirect observations $Y_1, ..., Y_n$, by constructing estimators via *WVD shrinkage*, it simply amounts to estimate a suitable selection of WVD coefficients, and apply a nonlinear hard thresholding, depending on the level of resolution of the coefficients calculated and in addition on α, the parameter of smoothness of the error distribution $k(x)$. We claim below that, for an appropriate tuning, this allows to obtain estimators with worst case risks comparable to the rates in Theorem 10 (within to possible log factors), as soon as the regime (X_n) is assumed geometrically regular and the stationary density $f(x)$ uniformly bounded away from zero on $[0,1]$ in the case of estimation of $\pi(x,y)$.

Form of the estimators. For $d = 1, 2$, we suppose we are given an (inhomogeneous) orthonormal wavelet basis on $[0,1]^d$ of class $\mathcal{C}^{r+1}$deriving from $(\phi_{\lambda})_{\lambda\in\Lambda_{j_1}^{(d)}}$ and $\left(\psi_{\gamma}\right)_{j\geq j_1,\ \gamma\in\Gamma_j^{(d)}}$. The densities $f(x)$ and $F(x,y) = f(x)\pi(x,y)$ have formal expansions on $[0,1]$ and on $[0,1]^2$ respectively

$$\begin{aligned} f(x) &= \sum_{\lambda\in\Lambda_{j_1}^{(1)}} \alpha_{\lambda}\phi_{\lambda}(x) + \sum_{j\geq j_1}\sum_{\gamma\in\Gamma_j^{(1)}} \beta_{\gamma}\psi_{\gamma}(x), \\ F(x,y) &= \sum_{\lambda\in\Lambda_{j_1}^{(2)}} \alpha_{\lambda}\phi_{\lambda}(x,y) + \sum_{j\geq j_1}\sum_{\gamma\in\Gamma_j^{(2)}} \beta_{\gamma}\psi_{\gamma}(x,y). \end{aligned}$$

The threshold wavelet-vaguelette estimators are constructed by replacing a suitable selection of unknown coefficients α_{λ}, β_{γ} in the expansions above by the empirical estimates

$$\begin{aligned} \hat{\alpha}_{\lambda} &= n^{-1}\sum_{i=1}^{n}\Delta_{\lambda}(Y_i),\ \hat{\beta}_{\gamma} = n^{-1}\sum_{i=1}^{n}\xi_{\gamma}(Y_i) \text{ for } \lambda\in\Lambda_{j_1}^{(1)},\ \gamma\in\Gamma_j^{(1)}, \\ \hat{\alpha}_{\lambda} &= n^{-1}\sum_{i=1}^{n-1}\Delta_{\lambda}(Y_i,Y_{i+1}),\ \hat{\beta}_{\gamma} = n^{-1}\sum_{i=1}^{n-1}\xi_{\gamma}\left(Y_i,Y_{i+1}\right) \text{ for } \lambda\in\Lambda_{j_1}^{(2)},\ \gamma\in\Gamma_j^{(2)}. \end{aligned}$$

According the *WVD shrinkage* technique, we apply a nonlinear, level-dependent thresholding to the empirical wavelet-vaguelette coefficients

$$\tilde{\beta}_{\gamma} = \hat{\beta}_{\gamma}1_{\left\{\left|\hat{\beta}_{\gamma}\right|>\mathcal{K}^{(d)}\mathcal{T}_{(d)}(j)\right\}} \text{ for } \gamma\in\Gamma_j^{(d)},$$

and perform selective reconstructions using big enough wavelet-vaguelette coefficients only, in a specific range of indices $j_1^{(d)}(n) \leqslant j \leqslant j_0^{(d)}(n)$

$$\hat{f}_{n,\ j_1^{(1)},\ j_0^{(1)}}(x) \quad = \quad \sum_{\lambda\in\Lambda_{j_1}^{(1)}} \hat{\alpha}_\lambda \phi_\lambda(x) + \sum_{j=j_1^{(1)}}^{j_0^{(1)}} \sum_{\gamma\in\Gamma_j^{(1)}} \tilde{\beta}_\gamma \psi_\gamma(x), \tag{7}$$

$$\hat{F}_{n,\ j_1^{(2)},\ j_0^{(2)}}(x,y) \quad = \quad \sum_{\lambda\in\Lambda_{j_1}^{(2)}} \hat{\alpha}_\lambda \phi_\lambda(x,y) + \sum_{j=j_1^{(2)}}^{j_0^{(2)}} \sum_{\gamma\in\Gamma_j^{(2)}} \tilde{\beta}_\gamma \psi_\gamma(x,y). \tag{8}$$

The transition density estimator is obtained by forming the quotient of the estimates (7) and (8), truncated so that the estimator keeps good integrability properties

$$\hat{\pi}_n(x,y) = \frac{\hat{F}_{n,\ j_1^{(2)},\ j_0^{(2)}}(x,y)}{\hat{f}_{n,\ j_1^{(1)},\ j_0^{(1)}}(x)} 1_{\left\{\hat{f}_{n,\ j_1^{(1)},\ j_0^{(1)}}(x) \geq \chi/2\right\}}. \tag{9}$$

Now we fix $r \in \aleph$ and define the class of Besov parameters for $d = 1,\ 2:$

$$\mathcal{C}_d = \left\{(\sigma,\ p,\ q)\,;\ s < \ \sigma - d/p,\ \sigma \leqslant r,\ 1 \leqslant p,\ q \leqslant \infty\right\}.$$

Quasi-optimality of Algorithm 8. The following result combined with Theorem 10 shows that for an appropriate tuning of parameters j_1, j_0 and K, the rate of estimator (7) is optimal within possible logarithmic factors simultaneously for the range of $L^{p'}$-losses and over each Besov ball $\mathcal{F}_{\sigma pq}^{(1)}(R)$ for any parameters $(\sigma,\ p,\ q)$ in $\mathcal{C}_1$.

Theorem 13 *Assume that $X = (X_n)_{n\in\aleph}$ is a stationary, Feller, aperiodic and positive recurrent Markov Chain with stationary probability $\mu(dx) = f(x)dx$ with f belonging to some class $\mathcal{F}_{\sigma pq}^{(1)}(R)$ where $(\sigma,\ p,\ q) \in \mathcal{C}_1$. Assume further that X is geometrically regular. If parameters $j_1^1(n)$ and $j_0^1(n)$ are picked such that*

$$2^{j_1^{(1)}(n)} \simeq n^{\frac{1}{1+2(r+\alpha)}},\ 2^{j_0^{(1)}(n)} \simeq (n/\log n)^{\frac{1}{1+s+2\alpha}}\ ,$$

and if the threshold is chosen such that

$$\mathcal{T}_1\,(j) = \sqrt{j2^{2j\alpha}},$$

then, for all $p' \geq 1$ and $(\sigma,\ p,\ q) \in \mathcal{C}_1$, there exist constants C and $\mathcal{K}_0^{(1)}$ (specified after below) such that for $\mathcal{K}^{(1)} \geq \mathcal{K}_0^{(1)}$

$$\left(E_f \left\|\hat{f}_n - f\right\|_{p'}^{p'}\right)^{1/p'} \leqslant C\,(\log n)^{\delta_1} \left(\frac{\log n}{n}\right)^{\nu_\alpha^{(1)}},$$

with $\delta_1 = \max\left(1/2 - p/qp', 0\right).1_{\{2\sigma p = p' - p\}}$.

Remark 14 *As the computation in section 3 indicates, in the case when the regime X satisfies conditions $\mathcal{M}(m,\ S,\ \delta,\ \nu)$ and $\mathcal{D}(S,\ V,\ b)$ (see subsection 2.1) for explicitely known parameters, the threshold constant $\mathcal{K}_0^{(1)}$ may be practically picked as follows. Set*

$$\begin{aligned} M_1 &= (2-\delta+\frac{2}{\delta})(\sup_{x\in S}V(x)-1)+bm(1+\frac{1}{\delta})+\int V(x)\nu(dx), \\ M_2 &= \frac{2-\delta}{1-b}, \end{aligned}$$

and let C_1 be some constant such that

$$M_1+M_2(\sup_{x\in[0,1]}V(x)+\int_{x\in\Re}V(x)\mu(dx))\leqslant C_1,$$

then one may choose $\mathcal{K}_0^{(1)}=C_0^{(1)}rp'$ with a constant $C_0^{(1)}$ chosen so large that $(C_0^{(1)})^2>18c_\infty^{(1)}(8\left\|k\right\|_{L^\infty(\Re)}\delta^{-1}\mu(S)^{-1}c_1^{(1)}C_1+2C_0^{(1)}/9)\log 2$ (see lemma 20 in section 3 for the definition of constants $c_1^{(1)}$ and $c_\infty^{(1)}$).

Quasi-optimality of Algorithm 9. When the parameters which controll the expansiveness of the empirical wavelet-vaguelette transforms $j_1^{(d)}$, $j_0^{(d)}$ and $\mathcal{T}_d\left(j\right)$ are picked as in Theorem 14 below in order to minimize simultaneously bias and variance components of the risk, estimator (8) has a near minimax optimality (compare to rates in Theorem 10) simultaneously for the variety of global $L^{p'}$-error measures and over all classes $\mathcal{P}_{\sigma pq}(R)$ with $(\sigma,\ p,\ q)\in\mathcal{C}_2$.

Theorem 15 *Suppose that $X=(X_n)_{n\in\aleph}$ is a stationary, Feller, aperiodic and positive recurrent Markov chain with transition probability $\Pi(x,dy)=\pi(x,y)dy$ and stationary probability $\mu(dx)=f(x)dx$. Suppose in addition that X is geometrically regular and that there exists some constant $\chi>0$ such that $f(x)\geq\chi$ for all $x\in[0,1]$. Provided that π belongs to some class $\mathcal{P}_{\sigma pq}(R)$ where $(\sigma,\ p,\ q)\in\mathcal{C}_2$, if $p'\geq 1$ and $j_1^d(n)$, $j_0^d(n)$ and $\mathcal{T}_d(j)$ are picked so that*

$$\begin{aligned} 2^{j_1^{(d)}(n)} &\simeq n^{\frac{1}{2(1+r+2\alpha)}},\ 2^{j_0^{(d)}(n)}\simeq(n/\log n)^{\frac{1}{d(1+s+2\alpha)}} \\ \mathcal{T}_d\left(j\right) &= \sqrt{j2^{2j\alpha d}/n}, \end{aligned}$$

then, for all $(\sigma,\ p,\ q)\in\mathcal{C}_2$, there exist constants C and $\mathcal{K}_0^{(d)}$ such that for all $\mathcal{K}^{(d)}\geq\mathcal{K}_0^{(d)}$

$$\left(E\left\|\hat{\pi}_n-\pi\right\|_{p'}^{p'}\right)^{\frac{1}{p'}}\leqslant C\left(\log n\right)^{\delta_2}\left(\frac{\log n}{n}\right)^{\nu_\alpha^{(2)}},$$

with $\delta_2=\max\left(1/2-p/qp',\ 0\right).1_{\{\sigma p=p'-p\}}$

Remark 16 *Keeping the notation introduced in remark 14, we point out that a possible choice for the threshold constants is* $\mathcal{K}_0^{(d)} = C_0^{(d)} rp'$*, the constants* $C_0^{(d)}$ *being chosen so large that* $(C_0^{(d)})^2 > 18 c_\infty^{(d)} (8 \left\| k \right\|_{L^\infty(\Re)}^d \delta^{-1} \mu(S)^{-1} c_\infty^{(d)} C_1 + 2C_0^{(d)}/9) \log 2$ *(*$c_1^{(d)}$ *and* $c_\infty^{(d)}$ *being constants specified in lemma 20).*

Although the estimators specified in theorems 13 and 15 are adaptive with respect to parameters σ, p and q, their constructions require nevertheless that p', r, s, R be explicitely known.

3 Sketch of Proofs of Theorems 13 and 15.

Now we show how upper bounds results stated in 2.3 derive from the properties of wavelet shrinkage with regard to approximation over Besov spaces and the assumption of geometric regularity for the underlying chain $X = (X_n)_{n \in \aleph}$. As we shall see, this latter assumption allows to bound stochastic terms (inherent to the randomness of the observations $Y_1, ..., Y_n$) in the calculation of risk bounds so that a tuning of the procedures proposed, corresponding to an optimal balancing of bias and variance components of the risk, may be found, leading to nearly minimax methods of estimation.

We begin by stating an approximation theoretical result showing how nonlinear shrinkage of the WVD coefficients can reduce the bias of the estimators (7) and (9).

Adaptive nonlinear approximation by wavelet shrinkage. Let $d \in \{1, 2\}$, suppose that g belongs to some Besov ball $\mathcal{F}_{\sigma pq}^{(d)}(R)$ with $(\sigma,\, p,\, q) \in \mathcal{C}_d$. With the object to approximate g on $[0,1]^d$ with a least $L^{p'}$-integrated error, using a given number of non zero coefficients only, Donoho *et al.* (1996) considered a selective reconstruction of the wavelet expansion

$$g = \sum_{\lambda \in \Lambda_{j_1}^{(d)}} \alpha_\lambda \phi_\lambda + \sum_{j \geq j_1} \sum_{\gamma \in \Gamma_j^{(d)}} \beta_\gamma \psi_\gamma,$$

by keeping wavelet coefficients at levels $j_1 \leqslant j \leqslant j_0$ such that $\left| \beta_\gamma \right| \geq \mathcal{K}^{(d)} . \mathcal{T}_d(j)$, $\gamma \in \Gamma_j^{(d)}$ only:

$$\mathcal{WS}(g) = \sum_{\lambda \in \Lambda_{j_1}^{(d)}} \alpha_\lambda \phi_\lambda + \sum_{j=j_1}^{j_0} \sum_{\gamma \in \Gamma_j^{(d)}} \beta_\gamma 1_{\left\{ \left| \beta_\gamma \right| \geq \mathcal{K}^{(d)} . \mathcal{T}_d(j) \right\}} \psi_\gamma. \tag{10}$$

The following result straightforwardly follows from the precise study of the Besov modulus corresponding to the error (measured by the $L^{p'}\left([0,1]^d\right)$-norm) of approximation of g by $\mathcal{WS}(g)$ in Donoho, Johnstone, Kerkyacharian & Picard (1996a) (see section 12.4 *evaluation of the modulus of continuity*, we will use the shortening "DJKP" in what follows).

Lemma 17 *If the parameters of expansiveness of the development* (10) *are picked such that*

$$\begin{aligned} \mathcal{T}_d(j) &= \sqrt{j2^{2jd\alpha}/n}, \\ 2^{j_1(n)} &\simeq n^{\frac{1}{d+2(\alpha d+r)}},\ 2^{j_0(n)} \simeq (n/\log n)^{\frac{1}{d(1+2\alpha)}}, \end{aligned}$$

then, there exists some constant C depending only on R, σ, p and q such that

$$\left\|\mathcal{WS}(g) - g\right\|_{p'} \leqslant C\left(\frac{\log n}{n}\right)^{\nu_\alpha^{(d)}} (\log n)^{\delta_d}.$$

Moreover, if $\mathcal{N}(n)$ denotes the number of non zero coefficients in (10)*, we have*

$$\mathcal{N}(n) = \bigcirc\left(n(\log n)^{\frac{p'-p}{p}1_{\{\sigma dp \geq p'-p\}}}\right)^{\frac{1-2\nu_\alpha^{(d)}}{d(1+2\alpha)}}.$$

Remark 18 *As lemma* (19) *above shows, if logarithmic factors are left out of account, $\mathcal{N} \simeq n^{(1-2\nu_\alpha^d)/(d+2\alpha d)}$ coefficients are enough to give an approximation of g with an error of order $n^{-\nu_\alpha^d}$ by means of the wavelet shrinkage method, whereas a linear type method of approximation requires a number of coefficients of order $n^{d\nu_\alpha^d/\sigma_d}$ to achieve a comparable rate when $p' \geq p$ (refer to the discussion in DJKP (1996c))*

Moments and probability bounds. Now we establish Rosenthal type moments inequalities, classically put forward in risk computations, and large deviations inequalities, these latter allowing here to estimate the gain in terms of $L^{p'}$-error that shrinkage of the empirical wavelet-vaguelette coefficients may provide (compare to inequality (14) in DJKP (1996b)).

Lemma 19 *(Moments bounds) Let $d \in \{1,2\}$. Suppose that $X = (X_n)_{n\in\aleph}$ is geometrically regular, $|f| \leqslant R$ then, for any $m \geq 1$, there exists some constant C such that, for all $j \geq J$ such that $2^{jd} \leqslant n$,*

$$\begin{aligned} \sum_{\lambda\in\Lambda_j^{(d)}} \left|\hat{\alpha}_\lambda - \alpha_\lambda\right|^m &\leqslant C2^{jd(1+\alpha m)}n^{-m/2}, && (11) \\ \sum_{\gamma\in\Gamma_j^{(d)}} \left|\hat{\beta}_\gamma - \beta_\gamma\right|^m &\leqslant C2^{jd(1+\alpha m)}n^{-m/2}. && (12) \end{aligned}$$

Proof. First, let us turn our attention to specific properties shared by the collections of functions involved, following straightforwardly from their definitions (see subsection 2.2).

Lemma 20 *Let* (θ_{jk}) *denote either the set of functions* $(\Delta_\lambda)_{\lambda\in\Lambda_j^{(d)},\, j\geq J}$ *or* $\left(\xi_\gamma\right)_{\gamma\in\Gamma_j^{(d)},\, j\geq J}$. *There exist constants* $c_\infty^{(d)}$ *and* $c_1^{(d)}$ *such that, for* $j \geq J$

$$\begin{aligned} \|\theta_{jk}\|_\infty &\leqslant c_\infty^{(d)} 2^{jd(\alpha+1/2)}), &(13)\\ \|\theta_{jk}\|_1 &\leqslant c_1^{(d)} 2^{jd(\alpha-1/2)}), &(14) \end{aligned}$$

and we have the following localization property (see Lemma 5)

$$|\theta_{jk}(x)| \leqslant C2^{jd(\alpha+1/2)}\left(1+\left|2^j x-k\right|\right)^{-2} \qquad (15)$$

We prove inequality (11) in the case $d=1$, (12) follows from analogous arguments. In order to simplify the proof, even if it entails to replace the chain by a split chain obtained through the Nummelin splitting technique from the parameters of some condition $\mathcal{M}(m,\, S,\, \delta,\, \nu)$ (see Definition 2) fulfilled by the chain, we assume the existence of an accessible atom $A \subset \Re$ for the Markov chain $X=(X_n)_{n\in\aleph}$ (i.e. a measurable set of the state space such that $\mu\,(A)>0$ and $\forall\,(x,x')\in A^2$, $\Pi\,(x,dy)=\Pi(x',dy)$, see Nummelin (1984)).
Let P_A (respectively P_μ, resp. P_x) denote the probability measure conditionally to $X_0\in A$ (resp. such that $X_0\sim\mu$, resp. conditionally to $X_0=x$) and E_A (resp. E_μ, resp. E_x) the P_A-expectation (resp. the P_μ-expectation, resp. the P_x-expectation). As we assumed that the chain $X=(X_n)_{n\in\aleph}$ is geometrically regular and the compact interval $[0,1]$ is included in the support of μ, we have in particular $E_A\left(\tau_A^m\right)<\infty$, or equivalently $E_\mu(\tau_A^{m-1})<\infty$ (since $P_\mu(\tau_A=k)=\mu(A)P_A(\tau_A\geq k)$, see renewal theory equality in 10.17 in Meyn & Tweedie (1996)), and $\sup_{x\in[0,1]}E_x(\tau_A^m)<\infty$ (see remark 4 in subsection 2.1). In 2.1 we previously observed that $Z=((X_n,Y_n))_{n\in\aleph}$ is a Markov chain, that inherits the stochastic regularity properties from X, in particular it follows from equality (6) that Z is geometrically regular and $A\times\Re$ is an accessible atom for the latter.
Assume $m\geq 2$ from now on. As $\|\Delta_{jk}\|_\infty\leqslant c_\infty^{(1)}2^{j(\alpha+1/2)}$, the application of Proposition 8 in Clémençon (2001) to the functional $\Delta_{jk}(Y_n)$ of the geometrically regular chain Z yields:

$$\begin{aligned} E\left(\left|\hat{\alpha}_{jk}-\alpha_{jk}\right|^m\right) &\leqslant C_1\,(m)\,v_Z^m\,(\Delta_{jk})\,n^{-\frac{m}{2}}\\ &\quad +C_2(m,n,\Delta_{jk})(c_\infty^{(1)}2^{j(\alpha+\frac{1}{2})})^{m-2}n^{1-m}. \qquad (16)\end{aligned}$$

We set $\bar{\Delta}_{jk}=\Delta_{jk}-\mu K\,(\Delta_{jk})$. It is simpler to consider the case of a chain that possesses an atom, because in such a case the terms v_Z^2 and C_2 involved in the upper bound (16) may be expressed in terms of hitting time to the atom A, as follows

$$\begin{aligned}
C_2(m,n,\Delta_{jk}) &= c_m n^{-1}\{E_\mu\left(1_{\{\tau_A\leqslant n\}}\tau_A^{m-1}\sum_{i=1}^{\tau_A}\overline{\Delta}_{jk}^{2}(Y_i)\right) \\
&\quad+\sum_{1\leqslant i\leqslant k\leqslant n}E_A\left(1_{\{\tau_A>k\}}k^{-2}\overline{\Delta}_{jk}^{2}(Y_i)\right) \\
&\quad+\sum_{i=1}^{n}n^{m-1}E_\mu\left(1_{\{\tau_A>n\}}\overline{\Delta}_{jk}^{2}(Y_i)\right)\}, \qquad (17)
\end{aligned}$$

besides, θ be any bounded and measurable function defined on $\Re$, we recall the following expression of the limiting variance in the CLT for the functional $(\theta(Y_n))_{n\in\aleph}$:

$$v_Z^2(\theta) = 2\int\theta^*(x,y)\overline{\theta}(y)\mu K(dx,dy) - \int\left(\overline{\theta}(y)\right)^2(k*f)(y)dy, \qquad (18)$$

where

$$\begin{aligned}
\overline{\theta} &= \theta-\int\theta(y)(k*f)(y)dy = \theta - E_\mu\left(\theta\left(Y_1\right)\right), \\
\theta^*(x,y) &= E\left(\sum_{i=0}^{\sigma_A}\overline{\theta}(Y_i)\mid X_0=x,Y_0=y\right), \qquad (19)
\end{aligned}$$

with $\sigma_A=\inf\{n\geq 0,\ X_n\in A\}$ (see paragraph 17.4.3 p 436 in Meyn & Tweedie (1996)).
First we prove that, for some constant $\mathbf{C}_{(1)}$,

$$v_Z^2\left(\Delta_{jk}\right)\leqslant \mathbf{C}_{(1)}2^{j2\alpha}. \qquad (20)$$

From (19), we deduce that for all x in $[0,1]$, y in $\Re$

$$\left|\Delta_{jk}^*(x,y)\right| \leqslant 2\left\|\Delta_{jk}\right\|_\infty \sup_{x\in[0,1]} E_x\left(\tau_A\right) = O\left(2^{j(\alpha+1/2)}\right) \qquad (21)$$

Then, it follows from (18) that

$$v_Z^2\left(\Delta_{jk}\right)\leqslant 2\int f(x)k(y-x)\left(\Delta_{jk}^*(x,y)\left(\Delta_{jk}(y)-\mu K\left(\Delta_{jk}\right)\right)\right)dxdy.$$

We decompose the integral above as follows

$$\int f(x)k(y-x)\Delta_{jk}^*(x,y)\Delta_{jk}(y)dxdy - \mu\left(\phi_{jk}\right)\int f(x)k(y-x)\Delta_{jk}^*(x,y)dxdy. \qquad (22)$$

Provided that k is bounded on $\Re$, bound (21) combined with (14) implies that the first term in (22) is bounded by $2c_1^{(1)}c_\infty^{(1)}\sup_{x\in[0,1]}E_x\left(\tau_A\right)\|k\|_{L^\infty(\Re)}2^{j2\alpha}$.

Moreover, as $\mu\left(\phi_{jk}\right) \leqslant \|f\|_\infty \|\phi\|_{L^1(\Re)} 2^{-j/2}$, using (21) again, we deduce that the second term is $o(2^{j\alpha})$. Hence, (20) is proved with, for 2^j large enough,

$$\mathbf{C}_{(1)} = 8c_1^{(1)} c_\infty^{(1)} \|k\|_{L^\infty(\Re)} \sup_{x\in[0,1]} E_x\left(\tau_A\right). \tag{23}$$

Now we consider the term C_2. Let C denote a constant that will not necessarily be the same at each appearance. Observe that it follows from localization property (15) in lemma 20, that

$$\sup_x \sum_{k=0}^{2^j-1} \left(\Delta_{jk}(x)\right)^2 \leqslant C 2^{j(2\alpha+1)}.$$

We deduce the bound

$$\sum_{i=1}^{\tau_A} \sum_{k=0}^{2^j-1} \left(\Delta_{jk}(Y_i) - \mu K(\Delta_{jk})\right)^2 \leqslant C 2^{j(2\alpha+1)} \tau_A \qquad \text{a.s.}$$

Using this inequality in each of the three expectations in (17) and finiteness of $E_A\left(\tau_A^m\right)$ and $E_\mu\left(\tau_A^{m-1}\right)$, we obtain that

$$\sum_{k=0}^{2^j-1} C_2(2,n,A,\Delta_{jk}) = O\left(2^{j(2\alpha+1)}\right).$$

Hence, as $2^j \leqslant n$, we have

$$\begin{aligned}\sum_{k=0}^{2^j-1} \left|\hat{\alpha}_{jk} - \alpha_{jk}\right|^m &\leqslant C\left(1 + \left(\frac{2^j}{n}\right)^{(m/2-1)}\right) 2^{j(\alpha m+1)} n^{-m/2} \\ &\leqslant C 2^{j(\alpha m+1)} n^{-m/2.}\end{aligned}$$

The case $m \leqslant 2$ can be deduced by convexity, inequality (11) is so proved in the case $d = 1$, the case $d = 2$ follows from a similar reasoning. ■

Lemma 21 *(Large-deviations bounds) Let $d \in \{1,2\}$. Suppose that $(X_n)_{n\in\aleph}$ is geometrically regular. Assume in addition that $2^{j_1(d)} \leqslant 2^j \leqslant 2^{j_0(d)}$, where $2^{j_1^{(d)}}$ and $2^{j_0^{(d)}}$ are picked as in Theorem 15. Then, letting $\mathcal{K}^{(d)} = \mathbf{C}'_{(d)} \times \kappa$ with κ an arbitrary constant ($\mathbf{C}'_{(d)}$ being specified after below), the following inequality holds for all $\kappa \geq 1$, $\gamma \in \Gamma_j^{(d)}$*

$$P_\mu\left(\left|\hat{\beta}_\gamma - \beta_\gamma\right| \geq \mathcal{K}^{(d)} \sqrt{j 2^{2j\alpha d}/n}\right) \leqslant 2^{-\kappa j}. \tag{24}$$

Proof. As in the proof of lemma 19 and with the same notations, by means of the Nummelin splitting technique again, we can assume the existence of an atom A such that $\mu(A) > 0$ and $E_A\left(e^{\mathcal{L}\tau_A}\right) < \infty$ for some $\mathcal{L} > 0$ with no loss of generality. The application of Theorem 17 in the case $d = 1$ (respectively, Theorem 19 in the case $d = 2$) in Clémençon (2001) to the functional $\xi_\gamma(Y_i)$ (resp., the functional $\xi_\gamma(Y_{i-1}, Y_i)$) provides an estimate for the probability of the large-deviations event

$$\mathcal{A}_\gamma = \left\{\left|\hat{\beta}_\gamma - \beta_\gamma\right| \geq \mathcal{K}^{(d)}\sqrt{j2^{2j\alpha d}/n}\right\},\ \gamma \in \Gamma_j^{(d)},$$

namely there exists a constant C such that for arbitrary y, $\gamma \in \Gamma_j^{(d)}$:

$$\begin{aligned}P_\mu(\mathcal{A}_\gamma) \quad &\leqslant \quad C\{n(e^{-\mathcal{L}y} + e^{-\mathcal{L}\mathcal{K}^{(d)}\sqrt{\frac{j2^{2j\alpha d}}{n}}n/\left\|\xi_\gamma\right\|_\infty}) + \\ &\exp(\frac{-(\mathcal{K}^{(d)})^2 j2^{2j\alpha d}}{18(\mu(A)^{-1}v_Z^2\left(\xi_\gamma\right) + \frac{2}{9}\left\|\xi_\gamma\right\|_\infty \mathcal{K}^{(d)} y\sqrt{j2^{2j\alpha d}/n})})\}. \quad (25)\end{aligned}$$

Now, $v_Z^2\left(\xi_\gamma\right) \leqslant \mathbf{C}_{(d)}2^{2j\alpha d}$ (with $\mathbf{C}_{(d)} = 8c_1^{(d)}c_\infty^{(d)}\left\|k\right\|_{L^\infty(\Re)}^d \sup_{x\in[0,1]} E_x(\tau_A)$, see (20) in the proof of lemma 19) and $\left\|\xi_\gamma\right\|_\infty \leqslant c_\infty^{(d)}2^{j(\alpha d + d/2)}$ (see (13) in lemma 20). Hence, provided that $2^j \leqslant 2^{j_0^{(d)}} \simeq (n/\log n)^{\frac{1}{d(1+s+2\alpha)}}$, by picking $y = y_n \simeq \mathcal{C}\log n$ with a constant $\mathcal{C}$ properly chosen, we obtain that $y_n\sqrt{j2^{jd}/n} \to 0$ as $n \to \infty$. Now, let $\mathcal{K}^{(d)} = \mathbf{C}'_{(d)} \times \kappa$ with $\mathbf{C}'_{(d)}$ chosen so large that $\mathbf{C}'^2_{(d)} > 18(\mu(A)^{-1}\mathbf{C}_{(d)} + 2\mathbf{C}'_{(d)}c_\infty^{(d)}/9)\log 2$. Then, by choosing $\mathcal{C}$ large enough, (24) straightforwardly follows from (25).

When the Markov chain X satisfies conditions $\mathcal{M}(m, S, \delta, \nu)$ and $\mathcal{D}(S, V, b)$, an explicit bound proved in Fort & Moulines (2000) shows that the constants $\mathcal{K}^{(d)}$ may be practically picked according to the parameters of conditions $\mathcal{M}$ and $\mathcal{D}$ so that (24) holds. As a matter of fact, we firstly notice that $Z = (X, Y)$ satisfies then conditions $\mathcal{M} = \mathcal{M}(m, S\times\Re, \delta, \nu\otimes K^{(1)})$ and $\mathcal{D}(S\times\Re, W, b)$ with $\nu\otimes K^{(1)}(dz, dy) = \nu(dz)k(y-z)dy$ and $W(z, y) = V(z)$. Let $A_\mathcal{M}$ (respectively $\mu_\mathcal{M}$) be the atom (resp. the stationary law) of the split chain $Z^\mathcal{M}$ of the chain Z constructed from the parameters of condition $\mathcal{M}$ via the Nummelin technique. We have $\mu_\mathcal{M}(A_\mathcal{M}) = \delta\mu(S)$ and, denoting by $\tau_{A_\mathcal{M}}$ the return time to $A_\mathcal{M}$ of the split chain and by $E_x(.)$ the mean on the split space given that the original chain X starts in x, in the notation of remark 14 it straightforwardly follows from Proposition 13 in Fort & Moulines (2000) that

$$E_x(\tau_{A_\mathcal{M}}) \leqslant M_1 + M_2(V(x) + \int_{z\in\Re} V(z)\mu(dz)). \quad (26)$$

Putting together (23) and (26), we finally obtain the bounds

$$v_{Z^\mathcal{M}}^2(\xi_\gamma) \leqslant 8c_1^{(d)}c_\infty^{(d)}\left\|k\right\|_{L^\infty(\Re)}^d (M_1 + M_2(\sup_{x\in[0,1]} V(x) + \int V(z)\mu(dz)))2^{2j\alpha d},$$

which, combined with the argument above, establish the claims in remarks 14 and 16. ■

Once moments and probability inequalities for the empirical coefficients are proved, it suffices to follow line-by-line the computations developped at lenght in DJKP (1996a) (see paragraph 5.1.2). Omitting the subscripts $j_1^{(d)}$, $j_0^{(d)}$, this yields:

$$E\left(\left\|\hat{f}_n - f\right\|_{p'}^{p'}\right)^{1/p'} \leqslant C\frac{(\log n)^{\delta_1+\nu_\alpha^{(1)}}}{n^{\nu_\alpha^{(1)}}}, \tag{27}$$

$$E\left(\left\|\hat{F}_n - F\right\|_{p'}^{p'}\right)^{1/p'} \leqslant C\frac{(\log n)^{\delta_2+\nu_\alpha^{(2)}}}{n^{\nu_\alpha^{(2)}}}. \tag{28}$$

Then, by decomposing the difference $\hat{\pi}_n(x,y) - \pi(x,y)$ as follows:

$$\frac{\hat{F}_n(x,y) - F(x,y) + \pi(x,y)(f(x) - \hat{f}_n(x))}{\hat{f}_n(x)} 1_{\left\{\hat{f}_n(x)\geq\frac{\chi}{2}\right\}} - \pi(x,y)1_{\left\{\hat{f}_n(x)<\frac{\chi}{2}\right\}},$$

we obtain that

$$\begin{aligned} E\left\|\hat{\pi}_n - \pi\right\|_{p'}^{p'} &\leqslant \left(\frac{2}{\chi}\right)^{p'} 4^{p'-1}\{E\left\|\hat{F}_n - F\right\|_{p'}^{p'} + \|\pi\|_\infty^{p'} E\left\|\hat{f}_n - f\right\|_{p'}^{p'}\} \\ &\quad + \|\pi\|_\infty^{p'}\int_0^1 P\left(\hat{f}_n(x) < \chi/2\right)dx. \end{aligned} \tag{29}$$

Easy calculations allow to bound the term inherent to truncation. As we can write

$$\hat{f}_n = \hat{f}_n - f + f,$$

we have for all $x \in [0,1]$,

$$\begin{aligned} P\left(\hat{f}_n(x) < \chi/2\right) &\leqslant P\left(\left|\hat{f}_n(x) - f(x)\right| \geq \chi/2\right) \\ &\leqslant \left(\frac{2}{\chi}\right)^{p'} E\left(\left|\hat{f}_n(x) - f(x)\right|^{p'}\right), \end{aligned}$$

by using Chebyshev's inequality. We deduce that

$$\int_0^1 P\left(\hat{f}_n(x) < \chi/2\right)dx \leqslant \left(\frac{2}{\chi}\right)^{p'} E\left(\left\|\hat{f}_n - f\right\|_{p'}^{p'}\right). \tag{30}$$

Let us observe finally that $\|\pi\|_\infty \leqslant \chi^{-1} \|F\|_\infty$ and that $B_{\sigma pq}\left([0,1]^2\right)$ is continuously embedded in the space of continuous functions on $[0,1]^2$ as soon as $\sigma > 2/p$, in view of bounds (27) and (28), inequalities (29) and (30) entail:

$$E\left(\left\|\hat{\pi}_n - \pi\right\|_{p'}^{p'}\right)^{1/p'} \leqslant C\frac{(\log n)^{\nu_\alpha^{(2)}+\delta_2}}{n^{\nu_\alpha^{(2)}}},$$

for some constant C.

References

[1] Bakry, D., Milhaud, X., Vandekerkhove, P., (1997). Statistique des chaînes de Markov cachées à espace d'états fini: le cas non stationnaire. *C. R. Acad. Sci., Paris, Ser. I, Math.* **325**, No 2, 203-206.

[2] Baum, L., Petrie, T., (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **30**, 1554-1563.

[3] Bickel, P., Ritov, Y., (1996). Inference in hidden Markov models: local asymptotic normality in the stationary case. *Bernouilli*, **2**, 199-228.

[4] Chauveau, D., Vandekerkhove, P., (1999). Un algorithme de Hastings-Metropolis avec apprentissage séquentiel. *C. R. Acad. Sci., Paris, Ser. I, Math.* **329**, No 2, 173-176.

[5] Clémençon, S., (2000a). *Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov.* Thèse de doctorat de l'Université Paris VII.

[6] Clémençon, S., (2000b). Adaptive estimation of the transition density of a regular Markov chain by wavelet methods. *Math. Meth. Statist.*, **9**, No 4, 323-357.

[7] Clémençon, S., (2001). Moments and probability inequalities for sums of bounded additive functionals of a regular Markov chain via the Nummelin splitting technique. *Statist. Prob. Lett.,* **55**, 227-238.

[8] Cohen, A., Daubechies, I., Vial, P., (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comp. Harm. Analysis*, **1**, 54-81.

[9] Daubechies, I, (1990). *Ten lectures on wavelets.* SIAM, Philadelphia.

[10] Donoho, D., (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comp. Harm. Analysis*, **1**, 100-115.

[11] Donoho, D., (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comp. Harm. Analysis*, **2**, 101-126.

[12] Donoho, D., Johnstone, I., Kerkyacharian, G., Picard, D., (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journ. Roy. Soc., Series B*, No 57, 301-369.

[13] Donoho, D., Johnstone, I., Kerkyacharian, G., Picard, D., (1996a). Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 2, 508-539.

[14] Donoho, D., Johnstone, I., Kerkyacharian, G., Picard, D., (1996b). Universal near minimaxity of wavelet shrinkage. *In Festshrift for Lucien Le Cam* (D. Pollard and G. Yangs, eds). Springer, New York.

[15] Fan, J., (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.*, **19**, No 3, 1257-1272.

[16] Fort, G., Moulines, E. (2000). V-subgeometric ergodicity for a Hasting-Metropolis algorithm. *Statist. Probab. Lett.,* **49**, No 4, 401-410.

[17] Golubev, G., Khasminskii, R., (1998). Asymptotically Optimal Filtering for a Hidden Markov Model. *Math. Meth. Statist.*, **7**, No 2, 192-209.

[18] Härdle, W., Kerkyacharian, G., Picard, D., Tsybakov, A., (1998). *Wavelets, Approximation, and Statistical Applications.* Lecture Notes in Statistics, Vol. 129. Springer.

[19] Holst, U., Lindgren, G., (1991). Recursive estimation in mixture models with Markov regime. *IEEE Trans. Inform. Theory*, **37**, 1683-1690.

[20] Jain, N., Jamison, B., (1967). Contributions to Doeblin's theory of Markov processes. *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, **8**, 19-40.

[21] Khasminskii, R., Zeitouni, O., (1996). Asymptotic Filtering for finite state Markov Chain. *Stochastic Process Appl.*, **63**, 1-10.

[22] Leroux, B., (1992). Maximum likelihood estimation for hidden Markov models. *Stochastic Process. Appl.*, **40**, 127-143.

[23] Meyer, Y., (1990). *Ondelettes et opérateurs I: Ondelettes.* Hermann, Paris.

[24] Meyn, S., Tweedie, R., (1996). *Markov chains and stochastic stability.* Springer-Verlag.

[25] Nummelin, E., (1984). *General irreducible Markov chains and nonnegative operators.* Cambridge University Press.

[26] Petrie, T., (1969). Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **40**, No 1, 97-115.

[27] Ryden, T., (1994). Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.*, **22**, 1884-1895.

[28] Vandekerkhove, P., (1996). *Recuit simulé et estimation des paramètres d'une chaîne de Markov cachée.* Thèse de doctorat de l'Université Paul Sabatier, Toulouse.

[29] Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley.