Nonparametric regression on hidden phi-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure

Thierry Dumont, Sylvain Le Corff

Nonparametric regression on hidden Φ-mixing variables: identifiability and consistency of a pseudo-likelihood based estimation procedure

Thierry Dumont *and Sylvain Le Corff †

October 22, 2014

Abstract

This paper outlines a new nonparametric estimation procedure for unobserved Φ-mixing processes. It is assumed that the only information on the stationary hidden states $(X_k)_{k\geq 0}$ is given by the process $(Y_k)_{k\geq 0}$, where Y_k is a noisy observation of $f_\star(X_k)$. The paper introduces a maximum pseudo-likelihood procedure to estimate the function $f_\star$ and the distribution $\nu_{\star,b}$ of $(X_0,\dots,X_{b-1})$ using blocks of observations of length b. The identifiability of the model is studied in the particular cases $b=1$ and $b=2$ and the consistency of the estimators of $f_\star$ and of $\nu_{\star,b}$ as the number of observations grows to infinity is established.

1 Introduction

The model considered in this paper is made of a bivariate stochastic process $((X_k,Y_k))_{k\geq 0}$ where only the observation sequence $(Y_k)_{k\geq 0}$ is available. These observations are given by

$$Y_k = f_\star(X_k) + \epsilon_k \ ,$$

where $f_\star$ is a function defined on a space $\mathbb{X}$ and taking values in $\mathbb{R}^\ell$. The measurement noise $(\epsilon_k)_{k\geq 0}$ is an independent and identically distributed (i.i.d.) sequence of Gaussian random vectors of $\mathbb{R}^\ell$.

This paper proposes a method to estimate the function $f_\star$ and the distribution of the hidden states using only the observations $(Y_k)_{k\geq 0}$. Note that the setting introduced here encompasses the i.i.d. case and the case of hidden Markov models in which the state sequence $(X_k)_{k\geq 0}$ is a Markov chain, the observations $(Y_k)_{k\geq 0}$ are independent conditionally on $(X_k)_{k\geq 0}$ and the conditional distribution of Y_k given the state sequence depends only on X_k. These hidden models can be applied in a large variety of disciplines such as financial econometrics [24], biology [4] or speech recognition [18] (see [10] for a recent overview on these models). Such a model is used in [12] to solve a simultaneous localization and mapping problem. In this framework, the observation Y_k is the signal strength received by a mobile device from different WiFi access points and $f_\star(X_k)$ is the expected signal strength at the device position X_k. In this particular case, $(X_k)_{k\geq 0}$ is a Markov chain on a compact set $\mathbb{X}$ of $\mathbb{R}^2$ (the map) with a transition kernel that involves the distance between two consecutive positions.

It is clear that the model considered in this paper is not identifiable with no additional assumptions. For instance, if $\tilde{X}_k = \sigma(X_k)$ where $\sigma : \mathbb{X} \to \mathbb{X}$ is a bijective function, then $Y_k = f_\star \circ \sigma^{-1}(\tilde{X}_k) + \epsilon_k$. Therefore, there exist a function $\tilde{f}$ and a process $(\tilde{X}_k)_{k\geq 0}$ on $\mathbb{X}$ fully characterizing the distribution of the observation process $(Y_k)_{k\geq 0}$ and it is not possible to define a consistent estimator of $f_\star$ using the observations $(Y_k)_{k\geq 0}$ only. It is then natural to study the assumptions under which it is possible to separate the distribution of the hidden states $(X_k)_{k\geq 0}$ and the function $f_\star$ using the distribution of $(f_\star(X_k))_{k\geq 0}$. The identifiability of the model is addressed in the particular case where $\mathbb{X} \subset \mathbb{R}^m$ for some $m \geq 1$ in Proposition 4.1 and Proposition 4.3. To

*MODAL'X, Université Paris-Ouest, Nanterre, France. thierry.dumont@u-paris10.fr
†Laboratoire de Mathématiques, Université Paris-Sud and CNRS, Orsay, France. sylvain.lecorff@math.u-psud.fr

our best knowledge, these are the most general results about the identifiability of nonparametric regression models on hidden variables. It is assumed that the state-space $\mathbb{X}$ is a compact subset of $\mathbb{R}^m$ and that $f_\star$ is a $\mathcal{C}^1$-diffeomorphism. The $\mathcal{C}^1$ regularity hypothesis on the target function $f_\star$ allows to perform the estimation procedure in a Sobolev setting such as in classical regression frameworks. The invertibility of $f_\star$ is a strong assumption. Nevertheless, in the case $\ell \geq 2m+1$, this assumption is satisfied for a dense class of functions in $\mathcal{C}^1$. Moreover, only $f_\star$ is assumed to be invertible and this assumption is enough to prove identifiability results on a wider class of $\mathcal{C}^1$ candidate functions. Proposition 4.1 establishes that if $\tilde{X}_0$ has a distribution with probability density ν and if $\tilde{f} : \mathbb{X} \to \mathbb{R}^\ell$ is such that $\tilde{f}(\tilde{X}_0)$ and $f_\star(X_0)$ have the same distribution then:

(a) $\tilde{f} = f_\star \circ \phi$ with $\phi : \mathbb{X} \to \mathbb{X}$ a bijective function ;

(b) ν is obtained by a transformation of the density of X_0 involving ϕ.

As a consequence, it is shown that if X_0 is uniformly distributed on $\mathbb{X} = [0,1]$ then ϕ is an isometric transformation of $[0,1]$ ($\phi = \mathrm{id}$ or $\phi = 1 - \mathrm{id}$) and the model is almost identifiable. Proposition 4.3 states a similar result on $f_\star$ and on the distribution of (X_0, X_1) when $(\tilde{f}(\tilde{X}_0), \tilde{f}(\tilde{X}_1))$ and $(f_\star(X_0), f_\star(X_1))$ have the same distribution. As a striking consequence, Corollary 4.4 shows that if the density of the distribution of X_1 conditionally on $X_0 = x$ is of the form $q_\star(x, x') = c_\star(x)\rho_\star(\|x - x'\|)$, then $q_\star$ and the full distribution of (X_0, X_1) are identifiable. In addition, $\tilde{f} = f_\star \circ \phi$ with $\phi : \mathbb{X} \to \mathbb{X}$ an isometric function.

The paper proposes a method to estimate the function $f_\star$ and the distribution $\nu_{b,\star}$ of the hidden states $(X_0, \ldots, X_{b-1})$ for a fixed parameter b using only the observations $(Y_k)_{k\geq 0}$. Note that this nonparametric estimation problem differs from classical regression settings since the variables $(X_k)_{k\geq 0}$ are not observed. In *errors-in-variables* models, the random variables $(X_k)_{k\geq 0}$ are i.i.d. and observed through a sequence $(Z_k)_{k\geq 0}$, *i.e.* $Z_k = X_k + \eta_k$ and $Y_k = f_\star(X_k) + \epsilon_k$, where the variables $(\eta_k)_{k\geq 0}$ are i.i.d with known distribution. Many solutions have been proposed to solve this problem, see [14] and [17] for a ratio of deconvolution kernel estimators, [20] for B-splines estimators and [6] for a procedure based on the minimization of a penalized contrast. In the case where the hidden state is a Markov chain, [22] and [23] considered the following convolution model $Y_k = X_k + \epsilon_k$, where the random variables $(\epsilon_k)_{k\geq 0}$ are i.i.d. with known distribution. [22] (resp. [23]) proposed an estimate of the transition density (resp. the stationary density and the transition density) of the Markov chain $(X_k)_{k\geq 0}$ based on the minimization of a penalized L^2 contrast. However, there does not exist any result on the nonparametric estimation problem studied in this paper with unobserved states $(X_k)_{k\geq 0}$.

The estimation procedure is based on the maximization of a penalized pseudo-likelihood over a class of functions $\mathcal{F}$ and a class of densities $\mathcal{D}_b$ where the penalty term involves the "complexity" of the functions in $\mathcal{F}$. The observations are decomposed into non-overlapping blocks $(Y_{kb}, \ldots, Y_{(k+1)b-1})$ and the pseudo-loglikelihood of the observations $(Y_0, \ldots, Y_{nb-1})$ considered in this paper is given by the sum of the loglikelihood of $(Y_{kb}, \ldots, Y_{(k+1)b-1})$ for $k \in \{0, \ldots, n-1\}$. The estimator $(\widehat{f}_n, \widehat{\nu}_n)$ of $(f_\star, \nu_{b,\star})$ is defined as a maximizer of the penalized version of the pseudo-likelihood of the observations $(Y_0, \ldots, Y_{nb-1})$. This estimator of $f_\star$ can be used to define an estimator $\widehat{p}_n$ of the density of the distribution of $(Y_0, \ldots, Y_{b-1})$. Theorem 3.1 states that the Hellinger distance between $\widehat{p}_n$ and the true distribution of a block of observations vanishes as the number of observations grows to infinity. More precisely, this Hellinger distance converges at a rate which can be chosen as close as possible to $n^{-1/4}$. To establish this result, the complexity function needs only to be lower bounded by a power of the supremum norm. We believe that this rate of convergence could be improved but this would require a better understanding of the dependency between the Hellinger distance and a well chosen distance on $\mathcal{F}$. The consistency of $(\widehat{f}_n, \widehat{\nu}_n)$ follows as a consequence together with some continuity properties (see Corollary 3.3). The rate of convergence of $\widehat{f}_n$ to $f_\star$ remains an open problem and seems to be very challenging.

It is also proven that the results presented in this paper hold in the special case where the function $f_\star$ belongs to a Sobolev space. Proposition 5.1 establishes the consistency of the estimator of $f_\star$ when the penalization function is based on a Sobolev norm. An important consequence of this result is that the image $f_\star(\mathbb{X}) \subset \mathbb{R}^\ell$ of $f_\star$ (which is the compact sub-manifold of dimension m in $\mathbb{R}^\ell$ where the process $(f_\star(X_k))_{k\geq 0}$ lies) can be consistently approximated by the sub-manifold $\widehat{f}_n(\mathbb{X})$ (see Corollary 5.2).

The proof of the convergence of the Hellinger distance between $\widehat{p}_n$ and the true distribution of a block of observations relies on a concentration inequality for the empirical process of the observations. This result

is obtained by an extension of the concentration inequality for Φ-mixing processes given in [28, Theorem 3]. The inequality of [28, Theorem 3] holds for empirical processes based on uniformly bounded functions which is not the case in the model presented here but a similar control can be proven under the assumptions of this paper. Then, the control of the expectation of the supremum of the empirical process is given by a direct application of the maximal inequality for dependent processes of [11].

The theoretical results given in the paper are supported by numerical experiments. An Expectation-Maximization (see [9]) based algorithm is outlined to compute $\widehat{\nu}_n$ and $\widehat{f}_n$.

The model and the estimators are presented in Section 2. The consistency results are displayed in Section 3. The identifiability in the cases $b = 1$ and $b = 2$ is addressed in Section 4. The application to a Sobolev class of function is detailed in Section 5 and the algorithm and numerical experiments are displayed in Section 6. Section 8 gathers important proofs on the identifiability and consistency needed to state the main results.

2 Model and definitions

Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space and $(\mathbb{X}, \mathcal{X})$ be a general state-space equipped with a measure μ. Let $(X_k)_{k\geq 0}$ be a stationary process defined on Ω and taking values in $\mathbb{X}$. This process is only partially observed through the sequence $(Y_k)_{k\geq 0}$ which takes values in $\mathbb{R}^\ell$, $\ell \geq 1$. For any $k \geq 1$, the sequence $(x_1, \ldots, x_k)$ is denoted by $x_{1:k}$. The observations $(Y_k)_{k\geq 0}$ are given by

$$Y_k \stackrel{\text{def}}{=} f_\star(X_k) + \epsilon_k \;, \tag{1}$$

where $f_\star : \mathbb{X} \to \mathbb{R}^\ell$ is a measurable function and $(\epsilon_k)_{k\geq 0}$ are i.i.d. with density φ with respect to the Lebesgue measure λ of $\mathbb{R}^\ell$, given, for any $z_{1:\ell} \in \mathbb{R}^\ell$, by:

$$\varphi(z_{1:\ell}) \stackrel{\text{def}}{=} (2\pi)^{-\ell/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{\ell} z_j^2\right) \;. \tag{2}$$

By (2), the distribution of the random vector ϵ_0 is known and Gaussian with identity covariance matrix. This setting covers the case of a known and non singular covariance matrix Σ. Indeed, if $(Y_k)_{k\geq 0}$ is replaced by $(\Sigma^{-1/2} Y_k)_{k\geq 0}$, the modified noise $\Sigma^{-1/2}\epsilon_0$ is distributed according to the multivariate Gaussian distribution with identity covariance matrix.

The problem studied in the paper could be interpreted as a deconvolution problem where the complete knowledge of the noise distribution is a rather classical assumption (see for instance [3, 19, 21]). Here, the density φ is assumed to be known to simplify the proof of the identifiability of the model (Section 4). This proof only needs the characteristic function of ϵ_0 to be known and non zero. Note that the Gaussian assumption is only used to establish the consistency result (Theorem 3.1) which relies on an entropy control written for this particular choice of density function φ. A few authors have studied the deconvolution problem with unknown noise distribution. In [5], the estimation of the density of X in the model $Y = X + \epsilon$ is performed without knowing the distribution ϵ and under mild assumptions on the smoothness of the underlying densities. However, [5] only considered real valued random variables and the estimation based on Fourier transform and bandwidth selection is hardly transposable to our model. The main difference between the model studied in this paper and classical convolution models is that the random vector $f_\star(X_k)$ does not necessarily have a density with respect to the Lebesgue measure on $\mathbb{R}^\ell$. Indeed, as discussed in Section 5 (Corollary 5.2), under some assumptions on $f_\star$, if the state-space $\mathbb{X}$ is a subset of $\mathbb{R}^m$ with $m < \ell$, $f_\star(X_k)$ lies in a sub-manifold of dimension m in $\mathbb{R}^\ell$ which has a null Lebesgue measure. Therefore, classical deconvolution tools do not apply here.

One of the objectives of this paper is the estimation of the target function $f_\star \in \mathcal{F}$ where $\mathcal{F}$ is a set of functions from $\mathbb{X}$ to $\mathbb{R}^\ell$. The results presented in Sections 3 and 4 are applied in Section 5 when $\mathcal{F}$ is a Sobolev space.

Let b be a positive integer. For any sequence $(x_k)_{k\geq 0}$, define $\mathbf{x}_k \stackrel{\text{def}}{=} (x_{kb}, \ldots, x_{(k+1)b-1})$ and for any function $f : \mathbb{X} \to \mathbb{R}^\ell$, define $\mathbf{f} : \mathbb{X}^b \to \mathbb{R}^{b\ell}$ by

$$\mathbf{x} = (x_0, \ldots, x_{b-1}) \mapsto \mathbf{f}(\mathbf{x}) \stackrel{\text{def}}{=} (f(x_0), \ldots, f(x_{b-1})) \;.$$

The distribution of $\mathbf{X}_0$ is assumed to have a density $\nu_{b,\star}$ with respect to the measure $\mu^{\otimes b}$ on $\mathbb{X}^b$ which is assumed to lie in a set of probability densities $\mathcal{D}_b$. For all $f \in \mathcal{F}$ and $\nu \in \mathcal{D}_b$, let $p_{f,\nu}$ be defined, for all $\mathbf{y} \in \mathbb{R}^{b\ell}$, by

$$p_{f,\nu}(\mathbf{y}) \stackrel{\text{def}}{=} \int \nu(\mathbf{x}) \prod_{k=0}^{b-1} \varphi(y_k - f(x_k)) \mu^{\otimes b}(\mathrm{d}\mathbf{x}) \ . \tag{3}$$

Note that $p_\star \stackrel{\text{def}}{=} p_{f_\star,\nu_{b,\star}}$ is the probability density of $\mathbf{Y}_0$ defined in (1). The function

$$y_{0:nb-1} \mapsto \sum_{k=0}^{n-1} \ln p_{f,\nu}(\mathbf{y}_k)$$

is referred to as the pseudo log-likelihood of the observations up to time $nb-1$.

This paper introduces an estimation procedure based on the method of M-estimation presented in [30] and [29]. Consider a function $I : \mathcal{F} \to \mathbb{R}^+$ which characterizes the complexity of functions in $\mathcal{F}$ and let ρ_n and λ_n be some positive numbers. Define the following ρ_n-Maximum Pseudo-Likelihood Estimator (ρ_n-MPLE) of $(f_\star, \nu_{b,\star})$:

$$\left(\widehat{f}_n, \widehat{\nu}_n\right) \stackrel{\text{def}}{=} \underset{f\in\mathcal{F},\ \nu\in\mathcal{D}_b}{\text{argmax}^{\rho_n}} \left\{ \sum_{k=0}^{n-1} \ln p_{f,\nu}(\mathbf{Y}_k) - \lambda_n I(f) \right\} \ , \tag{4}$$

where $\underset{f\in\mathcal{F},\ \nu\in\mathcal{D}_b}{\text{argmax}^{\rho_n}}$ is one of the pairs (f', ν') such that

$$\sum_{k=0}^{n-1} \ln p_{f',\nu'}(\mathbf{Y}_k) - \lambda_n I(f') \geq \sup_{f\in\mathcal{F},\ \nu\in\mathcal{D}_b} \left\{ \sum_{k=0}^{n-1} \ln p_{f,\nu}(\mathbf{Y}_k) - \lambda_n I(f) \right\} - \rho_n \ .$$

The consistency of the estimators is established using a control for empirical processes associated with mixing sequences. The Φ-mixing coefficient between two σ-fields $\mathcal{U}, \mathcal{V} \subset \mathcal{E}$ is defined in [8] by

$$\Phi(\mathcal{U}, \mathcal{V}) \stackrel{\text{def}}{=} \sup_{\substack{U\in\mathcal{U}, V\in\mathcal{V}, \\ \mathbb{P}(U)>0}} \left| \frac{\mathbb{P}(U \cap V)}{\mathbb{P}(U)} - \mathbb{P}(V) \right| \ .$$

The stationary process $(X_k)_{k\geq 0}$ can be extended to a two-sided process $(X_k)_{k\in\mathbb{Z}}$ which is said to be Φ-mixing when $\lim_{i\to\infty} \Phi_i^X = 0$ where, for all $i \geq 1$,

$$\Phi_i^X \stackrel{\text{def}}{=} \Phi\left(\sigma\left(X_k \ ; \ k \leq 0\right), \sigma\left(X_k \ ; \ k \geq i\right)\right) \ , \tag{5}$$

$\sigma\left(X_k \ ; \ k \in C\right)$ being the σ-field generated by $(X_k)_{k\in C}$ for any $C \subset \mathbb{Z}$. As in [28], the required concentration inequality for the empirical process is established under the following assumption on the Φ-mixing coefficients of $(X_k)_{k\geq 0}$.

H1 The stationary process $(X_k)_{k\geq 0}$ satisfies

$$\mathbf{\Phi} \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} (\Phi_i^X)^{1/2} < \infty \ , \tag{6}$$

where Φ_i^X is given by (5).

Remark 2.1.

1. If $(X_k)_{k\geq 0}$ is i.i.d., then $\Phi_i^X = 0$ for all $i \geq 1$ and H1 is satisfied.
2. Assume $(X_k)_{k\geq 0}$ is a stationary Markov chain with transition kernel Q and stationary distribution π such that there exist $\epsilon > 0$ and a measure ϑ on $\mathbb{X}$ satisfying, for all $x \in \mathbb{X}$ and all $A \in \mathcal{X}$,

$$Q(x, A) \geq \epsilon \vartheta(A) \ .$$

Then, by [26, Theorem 16.2.4], there exists $\rho \in (0,1)$ such that, for all $x \in \mathbb{X}$ and all $A \in \mathcal{X}$,

$$|Q^n(x,A) - \pi(A)| \leq \rho^n \ .$$

Therefore, for all $n, i > 0$ and $A, B \in \mathcal{X}$ such that $\pi(A) > 0$,

$$\begin{aligned} |\mathbb{P}\left(X_{n+i} \in B | X_n \in A\right) - \mathbb{P}\left(X_{n+i} \in B\right)| &= |\mathbb{P}\left(X_{n+i} \in B | X_n \in A\right) - \pi\left(B\right)| \ , \\ &\leq \frac{1}{\pi(A)} \left| \int_A \left(Q^i(x,B) - \pi(B)\right) \pi(\mathrm{d}x) \right| \ , \\ &\leq \rho^i \ . \end{aligned}$$

The Φ-mixing coefficients associated with $(X_k)_{k\geq 0}$ decrease geometrically and H1 is satisfied.

3 General convergence results

Denote by $\widehat{p}_n$ the estimator of $p_\star$ defined by

$$\widehat{p}_n \stackrel{\text{def}}{=} p_{\widehat{f}_n, \hat{\nu}_n} \ . \tag{7}$$

The first step to prove the consistency of the estimators is to establish the convergence of $\widehat{p}_n$ to $p_\star$ using a suitable metric. This is done in Theorem 3.1 where the only assumption related to the penalization procedure is that the complexity function I is lower bounded by a power of the supremum norm. Consider the following assumptions.

H2 There exist $C > 0$ and $\upsilon > 0$ such that for all $f \in \mathcal{F}$,

$$\|f\|_\infty \leq C I(f)^\upsilon \ , \tag{8}$$

with, for any $f \in \mathcal{F}$, $\|f\|_\infty \stackrel{\text{def}}{=} \max_{1\leq j\leq \ell} \operatorname{ess\,sup}_{x\in\mathbb{X}} |f_j(x)|$.

Where ess sup denotes the essential supremum with respect to the measure μ on $\mathbb{X}$. Hence, if H2 holds, since $I : \mathcal{F} \to \mathbb{R}^+$, for all $f \in \mathcal{F}$, $\|f\|_\infty \leq C I(f)^\upsilon < \infty$.

H3 There exist $0 < \nu_- < \nu_+ < +\infty$ such that, for all $\nu \in \mathcal{D}_b$ $\nu_- \leq \nu \leq \nu_+$.

The convergence of $\widehat{p}_n$ to $p_\star$ is established using the Hellinger metric defined, for any probability densities p_1 and p_2 on $\mathbb{R}^{b\ell}$, by

$$h(p_1,p_2) \stackrel{\text{def}}{=} \left[\frac{1}{2}\int \left(p_1^{1/2}(y) - p_2^{1/2}(y)\right)^2 \mathrm{d}y\right]^{1/2} \ . \tag{9}$$

Theorem 3.1 provides a rate of convergence of $\widehat{p}_n$ to $p_\star$ and a bound for the complexity $I(\widehat{f}_n)$ of the estimator $\widehat{f}_n$.

Theorem 3.1. *Assume H1-3 hold for some υ such that $b\ell\upsilon < 1$. Assume also that λ_n and ρ_n satisfy*

$$\lambda_n n^{-1} \underset{n\to+\infty}{\longrightarrow} 0, \ \lambda_n n^{-1/2} \underset{n\to+\infty}{\longrightarrow} +\infty \text{ and } \rho_n = O\left(\frac{\lambda_n}{n}\right) \ . \tag{10}$$

Then,

$$h^2(\widehat{p}_n, p_\star) = O_{\mathbb{P}}\left(\frac{\lambda_n}{n}\right) \quad \text{and} \quad I(\widehat{f}_n) = O_{\mathbb{P}}(1) \ . \tag{11}$$

Condition (10) implies that the rate of convergence of the Hellinger distance between $\widehat{p}_n$ and the true density $p_\star$ is slower than $n^{-1/4}$. The proof of the consistency of $\widehat{p}_n$ relies on the control of the empirical process:

$$\sup_{f,\nu} \int \frac{1}{2} \ln \frac{p_{f,\nu} + p_\star}{2p_\star} \mathrm{d}\left(\mathbb{P}_n - \mathbb{P}_\star\right) \ ,$$

where $\mathbb{P}_n$ is the empirical distribution of the observations $\{\mathbf{Y}_k\}_{k=0}^{n-1}$. In Proposition 3.2, the deviation result on the empirical process is established globally on the class of functions $\{p_{f,\nu};\ f \in \mathcal{F},\ \nu \in \mathcal{D}_b\}$. A weaker condition on λ_n could be obtained with a better deviation inequality on the empirical process when p remains "close" to $p_\star$. For instance, [29, Theorem 10.6] estimated the distribution of a random variable Y using i.i.d. samples $Y_1, \ldots, Y_n$ and the penalized loglikelihood $p \mapsto \int \log p \ \mathrm{d}\mathbb{P}_n - \lambda_n I(p)$, where $I(p) = \int_{\mathbb{R}} (p^{(m)}(y))^2 \mathrm{d}y$ penalizes the m-th derivative of p. The proof of [29, equation (10.34)] established that

$$\sup_{p \in A_n(p_\star)} \frac{\int \ln \frac{p_{f,\nu}+p_\star}{2p_\star} \mathrm{d}\left(\mathbb{P}_n - \mathbb{P}_\star\right)}{1 + I(p) + I(p_\star)} = O_{\mathbb{P}}(n^{-\frac{2m}{2m+1}}) \ ,$$

where

$$A_n(p_\star) \stackrel{\text{def}}{=} \left\{p \ ;\ h(p, p_\star) \leq n^{-\frac{m}{2m+1}} \left[1 + I(p) + I(p_\star)\right]\right\}$$

to obtain $n^{-\frac{m}{2m+1}}$ as rate of convergence for $h(\widehat{p}_n, p_\star)$ that depends on the order of derivation m considered in the complexity function $I(p)$. [15] also used a localization technique to calibrate the minimal penalty which ensures the convergence of the estimate of the number of components in a general mixture model. In our case, using a localization procedure of the empirical process around the true density is complicated. We consider a general setting made of a class of functions $\mathcal{F}$, a class of densities $\mathcal{D}_b$ and a complexity function $I(f)$ that are all non specified. Therefore, Theorem 3.1 is established under the relatively mild assumptions H1-3. Hence, the rate $n^{-1/4}$ corresponds to the "worst case" rate. However, even when the model is fully specified such as in Section 5, controlling a localized version of the empirical process in order to improve the rate of convergence of $\widehat{p}_n$ remains a difficult problem.

The proof of Theorem 3.1 relies on a *basic inequality* which provides a simultaneous control of the Hellinger risk $h^2(\widehat{p}_n, p_\star)$ and of the complexity of the estimator $I(\widehat{f}_n)$. Define for any density function p on $\mathbb{Y}^{b\ell}$,

$$g_p \stackrel{\text{def}}{=} \frac{1}{2} \ln \frac{p + p_\star}{2p_\star} \ . \tag{12}$$

Let $\mathbb{P}_n$ be the empirical distribution based on the observations $\{\mathbf{Y}_k\}_{k=0}^{n-1}$, *i.e.*, for any measurable set A of $\mathbb{R}^{b\ell}$,

$$\mathbb{P}_n(A) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_A(\mathbf{Y}_k) \ .$$

By (4) and (7), following the proof of [29, Lemma 10.5], we get the basic inequality:

$$h^2(\widehat{p}_n, p_\star) + 4\lambda_n n^{-1} I(\widehat{f}_n) \leq 16 \int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) + 4\lambda_n n^{-1} I(f_\star) + \rho_n \ . \tag{13}$$

Therefore, a control of $\int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)$ in the right hand side of (13) will simultaneously provide a bound on the growth of $h^2(\widehat{p}_n, p_\star)$ and of $I(\widehat{f}_n)$. This control is given in Proposition 3.2.

Proposition 3.2. *Assume H1-3 hold. There exists a positive constant c such that, for any $\eta > 0$, there exist A and N such that for any $n \geq N$ and any $x > 0$,*

$$\mathbb{P}\left[\sup_{f \in \mathcal{F},\ \nu \in \mathcal{D}_b} \frac{\left|\int g_{p_{f,\nu}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)\right|}{1 \vee I(f)^\gamma} \geq c\mathbf{\Phi}\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{A}{\sqrt{n}}\right] \leq \frac{2e^{-\alpha x}}{1 - e^{-\alpha x}} \ ,$$

where

$$\gamma \stackrel{\text{def}}{=} b\ell\upsilon + \eta \quad \textit{and} \quad \alpha \stackrel{\text{def}}{=} \frac{\log(2)(\gamma - \upsilon)}{2^{2\gamma}} = \frac{\log(2)((b\ell - 1)\upsilon + \eta)}{2^{2(b\ell\upsilon + \eta)}} \ .$$

Proposition 3.2 is proved in Section. 8.1.

Proof of Theorem 3.1. Since $\upsilon^{-1} > b\ell$, $\eta > 0$ in Proposition 3.2 can be chosen such that $\gamma = b\ell\upsilon + \eta = 1$. For this choice of η, Proposition 3.2 implies that

$$\frac{\int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}^\star)}{1 \vee I(\widehat{f}_n)} = O_{\mathbb{P}}(n^{-\frac{1}{2}}) \ .$$

Combined with (13), this yields

$$h^2(\widehat{p}_n, p_\star) + 4\lambda_n n^{-1} I(\widehat{f}_n) \leq (1 \vee I(\widehat{f}_n)) O_{\mathbb{P}}(n^{-\frac{1}{2}}) + 4\lambda_n n^{-1} I(f_\star) + \rho_n \ . \tag{14}$$

Then, (14) directly implies that

$$4\ I(\widehat{f}_n) \leq (1 \vee I(\widehat{f}_n)) O_{\mathbb{P}}(n^{\frac{1}{2}} \lambda_n^{-1}) + 4I(f_\star) + \rho_n n \lambda_n^{-1} \ ,$$

which, together with (10), gives

$$I(\widehat{f}_n) = O_{\mathbb{P}}(1) \ .$$

Combining this result with (14) again leads to

$$h^2(\widehat{p}_n, p_\star) + O_{\mathbb{P}}(\lambda_n n^{-1}) \leq O_{\mathbb{P}}(n^{-\frac{1}{2}}) + 4\lambda_n n^{-1} I(f_\star) + \rho_n \ .$$

This concludes the proof of Theorem 3.1. □

Theorem 3.1 shows that $h^2(\widehat{p}_n, p_\star)$ vanishes as $n \to +\infty$. However, this does not imply the convergence of $(\widehat{f}_n, \widehat{\nu}_n)$ to $(f_\star, \nu_{b,\star})$. The convergence of the estimators $(\widehat{f}_n, \widehat{\nu}_n)$ is addressed in the case where the set $\mathcal{D}_b$ may be written as

$$\mathcal{D}_b = \{\nu_a;\ a \in \mathcal{A}\} \ , \tag{15}$$

where $\mathcal{A}$ is a parameter set not necessarily of finite dimension. The ρ_n-MPLE is then given by:

$$(\widehat{f}_n, \widehat{a}_n) \overset{\text{def}}{=} \underset{f \in \mathcal{F},\ a \in \mathcal{A}}{\operatorname{argmax}^{\rho_n}} \left\{ \sum_{k=0}^{n-1} \ln p_{f,\nu_a}(\mathbf{Y}_k) - \lambda_n I(f) \right\} \ .$$

Assume that $\mathcal{A}$ is equipped with a distance $d_{\mathcal{A}}$ such that $\mathcal{A}$ is compact with respect to the topology defined by $d_{\mathcal{A}}$. Assume also that $\mathcal{F}$ is equipped with a metric $d_{\mathcal{F}}$ such that $\mathcal{F}_M \overset{\text{def}}{=} \{f \in \mathcal{F};\ I(f) \leq M\}$ is compact for all $M > 0$ with respect to the topology defined by $d_{\mathcal{F}}$. Let d be the product distance on $\mathcal{F} \times \mathcal{A}$. Assume that the function $(f, a) \mapsto h^2(p_{f,\nu_a}, p_\star)$ is continuous with respect to the topology on $\mathcal{F} \times \mathcal{A}$ induced by d. Corollary 3.3 establishes the convergence of $(\widehat{f}_n, \widehat{a}_n)$ to the set $\mathcal{E}_\star$ defined as:

$$\mathcal{E}_\star \overset{\text{def}}{=} \left\{(f, a) \in \mathcal{F} \times \mathcal{A};\ h(p_{f,\nu_a}, p_{f_\star,\nu_{a_\star}}) = 0\right\} \ . \tag{16}$$

Define for all $(f, a) \in \mathcal{F} \times \mathcal{A}$,

$$d\left((f, a), \mathcal{E}_\star\right) = \inf_{(f', a') \in \mathcal{E}_\star} d\left((f, a), (f', a')\right) \ .$$

Corollary 3.3. *Assume H1-3 hold for some υ such that $\upsilon b\ell < 1$. Assume also that λ_n and ρ_n satisfy*

$$\lambda_n n^{-1} \underset{n \to +\infty}{\longrightarrow} 0, \ \lambda_n n^{-1/2} \underset{n \to +\infty}{\longrightarrow} +\infty \text{ and } \rho_n = O\left(\frac{\lambda_n}{n}\right) \ .$$

Then,

$$d\left((\widehat{f}_n, \widehat{a}_n), \mathcal{E}_\star\right) = o_{\mathbb{P}}(1) \ .$$

Corollary 3.3 is a direct consequence of Theorem 3.1 and of the properties of $d_{\mathcal{A}}$ and $d_{\mathcal{F}}$ and its proof is therefore omitted. The few assumptions on the model allow only to establish the convergence of the estimators $(\widehat{f}_n, \widehat{a}_n)$ to the set $\mathcal{E}_\star$ in Corollary 3.3.

4 Identifiability when $\mathbb{X}$ is a subset of $\mathbb{R}^m$

The aim of this section is to characterize the set $\mathcal{E}_\star$ given by (16) when $b = 1$ and when $b = 2$ (the characterization of $\mathcal{E}_\star$ when $b > 2$ follows the same lines) with some additional assumptions on the model, on $\mathcal{F}$ and on $\mathcal{D}_b$. In the sequel, $\nu_\star$ must satisfy $0 < \nu_- \leq \nu_\star \leq \nu_+$ for some constants ν_- and ν_+.

It is assumed that $\mathbb{X}$ is a subset of $\mathbb{R}^m$ for some $m \geq 1$ and that μ is the Lebesgue measure. For any subset A of $\mathbb{R}^m$, $\overset{\circ}{A}$ stands for the interior of A and $\overline{A}$ for the closure of A. Consider the following assumptions on the state-space $\mathbb{X}$.

H4 a) $\mathbb{X}$ is non empty, compact and $\overline{\overset{\circ}{\mathbb{X}}} = \mathbb{X}$,

b) $\mathbb{X}$ is arcwise and simply connected.

The compactness implies that $\mathbb{X}$ is closed and that continuous functions on $\mathbb{X}$ are bounded. By the last assumption of H4a), the interior of $\mathbb{X}$ is not empty and any element in $\mathbb{X}$ is the limit of elements of the interior of $\mathbb{X}$. Finally, $\mathbb{X}$ is arcwise and simply connected to ensure topological properties used in the proofs of the identifiability results below.

A function $f : U \to f(U) \subset \mathbb{R}^\ell$ defined on an open subset U of $\mathbb{R}^m$ is a $\mathcal{C}^1$-diffeomorphism if its differential function $x \mapsto D_x f$ is continuous (f is $\mathcal{C}^1$) and if, for all x in U, $\text{rank}(D_x f) = m$. A function $f : \mathbb{X} \to f(\mathbb{X})$ is said to be $\mathcal{C}^1$ (resp. a $\mathcal{C}^1$-diffeomorphism) if f is the restriction to $\mathbb{X}$ of a $\mathcal{C}^1$ function (resp. a $\mathcal{C}^1$-diffeomorphism) defined on an open neighborhood of $\mathbb{X}$ in $\mathbb{R}^m$.

H5 $f_\star$ is a $\mathcal{C}^1$-diffeomorphism from $\mathbb{X}$ to $f_\star(\mathbb{X})$.

H5 might be seen as a restrictive assumption. Nevertheless, when $\ell \geq 2m + 1$, H5 is satisfied for almost every continuous function form $\mathbb{X}$ to $\mathbb{R}^\ell$. Indeed, Whitney's embedding theorem ([31]) states, in this case, that any continuous function from $\mathbb{X}$ to $\mathbb{R}^\ell$ can be approximated by a smooth embedding.

In the case $b = 1$, Proposition 4.1 discusses the identifiability when $\mathcal{F}$ is a subset of $\mathcal{C}^1$. For any differential function $\phi : \mathbb{X} \to \mathbb{X}$, let J_ϕ be the determinant of the Jacobian matrix of ϕ: $J_\phi(x) = \det(D_x\phi)$.

Proposition 4.1 (b=1). *Assume that H4 and H5 hold. Let $f \in \mathcal{C}^1$ and let ν be a probability density with respect to μ such that $0 < \nu_- \leq \nu \leq \nu_+$. Then, $h(p_{f,\nu}, p_{f_\star,\nu_\star}) = 0$ if and only if $f_\star$ and f have the same image in $\mathbb{R}^\ell$, $\phi = f_\star^{-1} \circ f$ is bijective and, for μ almost every $x \in \mathbb{X}$,*

$$\nu(x) = |J_\phi(x)| \nu_\star(\phi(x)) \, .$$

The proof of Proposition 4.1 is given in Section 8.2. When $\mathcal{F} \subset \mathcal{C}^1$, Proposition 4.1 and H3 implies that the set $\mathcal{E}_\star$ defined in (16) is given by

$$\begin{aligned}(f, a) \in \mathcal{E}_\star \quad \Leftrightarrow \quad & \text{There exists a bijective function } \phi \in \mathcal{C}^1(\mathbb{X}, \mathbb{X}) \text{ such that} \\ & f = f_\star \circ \phi \text{ and } \nu_a = |J_\phi| \cdot \nu_\star \circ \phi \ \mu \text{ almost everywhere in } \mathbb{X} \, .\end{aligned}$$

Remark 4.2. Proposition 4.1 states that the candidates (f, ν) to characterize the distribution of Y_0 are necessarily related to $(f_\star, \nu_\star)$ through a state-space transformation denoted by ϕ. In the particular case where $\mathbb{X} = [0, 1]$ ($m = 1$) and $\nu_\star = 1$, Proposition 4.1 implies a sharper result. Assuming that $\nu = \nu_\star$ ($\nu_\star$ is known), Proposition 4.1 implies the existence of a $\mathcal{C}^1$ and bijective function ϕ satisfying $f = f_\star \circ \phi$ and $|J_\phi| = 1$. Therefore, $\phi : x \mapsto x$ or $\phi : x \mapsto 1 - x$ which are the two possible isometric transformations of $[0, 1]$.

Now, if $\mathbb{X} = [0, 1]$ and $\nu_\star$ is unknown and continuous we can define the uniform random variable on $[0, 1]$: $\tilde{X}_0 = F_\star(X_0)$ where $F_\star$ is the $\mathcal{C}^1$ and strictly increasing cumulative distribution function of X_0. The observation Y_0 can be written $Y_0 = \tilde{f}_\star(\tilde{X}_0) + \epsilon_0$ where $\tilde{f}_\star = f_\star \circ F_\star^{-1}$ satisfies the same hypothesis as $f_\star$. Thus, from the preceding remark, the function $\tilde{f}_\star = f_\star \circ F_\star^{-1}$ can be identified up to an isometric transformation of $[0, 1]$ from the distribution of Y_0 only.

These results cannot be extended to the case $m > 1$ where $|J_\phi| = 1$ does not necessarily imply that ϕ is isometric but only that ϕ preserves the volumes.

Proposition 4.3 discuss the identifiability when $b = 2$. In this case, $\nu_{2,\star}$ can be written $\nu_{2,\star}(x, x') = \nu_\star(x) q_\star(x, x')$ where $q_\star$ is a transition density with stationary probability density $\nu_\star$. For any transition density q on $\mathbb{X}^2$ satisfying

$$\text{for all } x, x' \in \mathbb{X}\,, 0 < q_- \leq q(x, x') \leq q_+\,, \tag{17}$$

there exists a stationary density ν associated with q satisfying, for all $x \in \mathbb{X}$, $q_- \leq \nu(x) \leq q_+$. Denote by ν_q this density.

Proposition 4.3 (b=2). *Assume that H4 and H5 hold. Let $f \in \mathcal{C}^1$ and q be a transition density satisfying (17). Let $\nu_2(x, x') = \nu_q(x) q(x, x')$. Then, $h(p_{f,\nu_2}, p_{f_\star,\nu_{\star,2}}) = 0$ if and only if $f_\star$ and f have the same image in $\mathbb{R}^\ell$, $\phi = f_\star^{-1} \circ f$ is bijective and $\mu \otimes \mu$ almost everywhere in $\mathbb{X}^2$,*

$$q(x, x') = |J_\phi(x')| q_\star(\phi(x), \phi(x'))\,. \tag{18}$$

Proposition 4.3 is proved in Section 8.3.

Corollary 4.4. *Consider the same assumptions as in Proposition 4.3. Assume in addition that $q_\star$ and q are of the form:*

$$q_\star(x, x') = c_\star(x) \rho_\star(||x - x'||)\,, \quad q(x, x') = c(x) \rho(||x - x'||)\,,$$

where ρ and $\rho_\star$ are two continuous functions defined on $\mathbb{R}_+$. Assume in addition that $\rho_\star$ is one-to-one. Then, $h(p_{f,\nu_2}, p_{f_\star,\nu_{\star,2}}) = 0$ if and only if $f_\star$ and f have the same image in $\mathbb{R}^\ell$, $\phi = f_\star^{-1} \circ f$ is an isometry on $\mathbb{X}$ and $q = q_\star$.

The proof of Corollary 4.4 is given in Section 8.3. When $\mathcal{F} \subset \mathcal{C}^1$ and for any a in $\mathcal{A}$, $\nu_a \in \mathcal{D}_2$ is of the form

$$\nu_a(x, x') = \nu_{q_a}(x) q_a(x, x') \text{ with } q_a(x, x') = c_a(x) \rho_a(||x - x'||)\,,$$

where $\rho_- \leq \rho_a \leq \rho_+$, Corollary 4.4 implies that the set $\mathcal{E}_\star$ defined in (16) is given by

$$(f, a) \in \mathcal{E}_\star \Leftrightarrow f = f_\star \circ \phi \text{ with } \phi \text{ an isometry and } q_a = q_\star$$

Finally, if the only isometry of $\mathbb{X}$ is the identity function, and if there exists a unique $a_\star$ in $\mathcal{A}$ such that $q_{a_\star} = q_\star$, then $\mathcal{E}_\star = \{(f_\star, a_\star)\}$ and the model is fully identifiable.

5 Application when $\mathcal{F}$ is a Sobolev class of functions

In this section, $\mathbb{X}$ is a subset of $\mathbb{R}^m$, $m \geq 1$ and the results of Section 3 and Section 4 are applied to a specific class of functions $\mathcal{F}$ with an example of complexity function I satisfying H2. Let $p \geq 1$, define

$$\mathrm{L}^p \stackrel{\text{def}}{=} \left\{ f : \mathbb{X} \to \mathbb{R}^\ell \;;\; \|f\|_{\mathrm{L}^p}^p = \int_{\mathbb{X}} \|f(x)\|^p \mu(\mathrm{d}x) < \infty \right\}\,.$$

For any $f : \mathbb{X} \longrightarrow \mathbb{R}^\ell$ and any $j \in \{1, \cdots, \ell\}$, the j^{th} component of f is denoted by f_j. For any vector $\alpha \stackrel{\text{def}}{=} \{\alpha_i\}_{i=1}^m$ of non-negative integers, we write $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i$ and $D^\alpha f : \mathbb{X} \to \mathbb{R}^\ell$ for the vector of partial derivatives of order α of f in the sense of distributions. Let $s \in \mathbb{N}$ and $W^{s,p}$ be the Sobolev space on $\mathbb{X}$ with parameters s and p, *i.e.*,

$$W^{s,p} \stackrel{\text{def}}{=} \{ f \in \mathrm{L}^p;\; D^\alpha f \in \mathrm{L}^p, \alpha \in \mathbb{N}^m \text{ and } |\alpha| \leq s \}\,. \tag{19}$$

$W^{s,p}$ is equipped with the norm $\|\cdot\|_{W^{s,p}}$ defined, for any $f \in W^{s,p}$, by

$$\|f\|_{W^{s,p}} \stackrel{\text{def}}{=} \left(\sum_{0 \leq |\alpha| \leq s} \|D^\alpha f\|_{\mathrm{L}^p}^p \right)^{1/p}\,. \tag{20}$$

The results of Section 3 and Section 4 can be applied to the class $\mathcal{F} = W^{s,p}$ under the following assumption.

H6 $\mathbb{X}$ has a locally Lipschitz boundary.

H6 means that all x on the boundary of $\mathbb{X}$ has a neighbourhood whose intersection with the boundary of $\mathbb{X}$ is the graph of a Lipschitz function. For any $j \in \{1, \cdots, \ell\}$ and $f \in W^{s,p}$, f_j belongs to $W^{s,p}(\mathbb{X}, \mathbb{R})$, the Sobolev space of real-valued functions with parameters s and p. Let $k \geq 0$, by [1, Theorem 6.3], if $s > m/p + k$ and if H4a) and H6 hold, $W^{s,p}(\mathbb{X}, \mathbb{R})$ is compactly embedded into $\left(\mathcal{C}^k(\mathbb{X}, \mathbb{R}), \|\cdot\|_{\mathcal{C}^k}\right)$. Arguing component by component, $W^{s,p}$ is compactly embedded into $\mathcal{C}^k \stackrel{\text{def}}{=} \mathcal{C}^k(\mathbb{X}, \mathbb{R}^\ell)$. Moreover, the identity function $id : W^{s,p} \to \mathcal{C}^k$ being linear and continuous, there exists a positive coefficient κ such that, for any $f \in W^{s,p}$,

$$\|f\|_{\mathcal{C}^k} \leq \kappa \|f\|_{W^{s,p}} \ . \tag{21}$$

Then, if $s > m/p + k$, for any $f \in \mathcal{F} = W^{s,p}$,

$$\|f\|_\infty \leq \kappa \|f\|_{W^{s,p}} \ . \tag{22}$$

In the following, $d_{\mathcal{C}^k}$ is the usual distance on $\mathcal{C}^k$ functions on $\mathbb{X}$. If the complexity function is defined by $I(f) = \|f\|^{1/\upsilon}_{W^{s,p}}$ with $\upsilon b\ell < 1$, then H2 holds and Theorem 3.1 can be applied. Moreover, by [1, Theorem 6.3], the subspace $\mathcal{F}_M$, $M \geq 1$ are quasi-compact in $\mathcal{C}^k$. Let $d_{\mathcal{A}}$ be a metric on the space $\mathcal{A}$ introduced in (15) such that $\mathcal{A}$ is compact and that, for $\mu \otimes \mu$ almost every $(x, x') \in \mathbb{X}^2$, $a \mapsto \nu_a(x, x')$ is continuous. By applications of the dominated convergence theorem, this implies the continuity of $(f, a) \mapsto h(p_{f,\nu_a}, p_\star)$. Define

$$\mathcal{F}_\star \stackrel{\text{def}}{=} \{f \in W^{s,p};\ \exists a \in \mathcal{A} \text{ such that } (f, a) \in \mathcal{E}_\star\} \ .$$

Then, Proposition 5.1 is a direct application of Corollary 3.3.

Proposition 5.1 ($\mathcal{F} = \mathbf{W^{s,p}}$, $\mathbf{s > m/p + k}$, $\mathbf{k \geq 0}$)**.** *Assume that H1, H3, H4a) and H6 hold. Assume also that* $I(f) = \|f\|^{1/\upsilon}_{W^{s,p}}$ *for some* υ *such that* $\upsilon b\ell < 1$ *and that* λ_n *and* ρ_n *satisfy*

$$\lambda_n n^{-1} \underset{n \to +\infty}{\longrightarrow} 0, \ \lambda_n n^{-1/2} \underset{n \to +\infty}{\longrightarrow} +\infty \text{ and } \rho_n = O\left(\frac{\lambda_n}{n}\right) \ .$$

Then,

$$d_{\mathcal{C}^k}\left(\widehat{f}_n, \mathcal{F}_\star\right) = o_{\mathbb{P}}(1) \ .$$

Moreover, as shown in Section 8.2, the assumption $\overline{\overset{\circ}{\mathbb{X}}} = \mathbb{X}$ together with the continuity of the functions in $\mathcal{F}$ provided by (21) imply that for any f in $\mathcal{F}_\star$, $f(\mathbb{X}) = f_\star(\mathbb{X})$. Define the Hausdorff distance $d_{\mathcal{H}}(A, B)$ between two compact subsets A and B of $\mathbb{R}^\ell$ as

$$d_{\mathcal{H}}(A, B) \stackrel{\text{def}}{=} \max\left(\sup_{a \in A} \inf_{b \in B} ||a - b||_{\mathbb{R}^\ell} \ , \ \sup_{b \in B} \inf_{a \in A} ||a - b||_{\mathbb{R}^\ell}\right) \ .$$

Proposition 5.1 implies Corollary 5.2.

Corollary 5.2 ($\mathcal{F} = \mathbf{W^{s,p}}$, $\mathbf{s > m/p}$)**.** *Assume that H1, H3, H4a) and H6 hold. Assume also that* $I(f) = \|f\|^{1/\upsilon}_{W^{s,p}}$ *for some* υ *such that* $\upsilon b\ell < 1$ *and that* λ_n *and* ρ_n *satisfy*

$$\lambda_n n^{-1} \underset{n \to +\infty}{\longrightarrow} 0, \ \lambda_n n^{-1/2} \underset{n \to +\infty}{\longrightarrow} +\infty \text{ and } \rho_n = O\left(\frac{\lambda_n}{n}\right) \ .$$

Then,

$$d_{\mathcal{H}}\left(\widehat{f}_n(\mathbb{X}), f_\star(\mathbb{X})\right) = o_{\mathbb{P}}(1) \ .$$

Corollary 5.2 establishes the consistency of the estimator $\widehat{f}_n(\mathbb{X})$ of the image of $f_\star$ in $\mathbb{R}^\ell$. This result is particularly interesting since $f_\star(\mathbb{X})$ is a manifold of dimension smaller than ℓ in $\mathbb{R}^\ell$. Thus, the proposed estimation procedure allows to approximate such manifolds, possibly of low dimensions, that are only observed with additive noise in $\mathbb{R}^\ell$. Moreover, this result holds under relatively weak assumptions on the manifold.

Since the identifiability of $f_\star$ is not necessary to have the identifiability of $f_\star(\mathbb{X})$, $f_\star$ is not assumed to be bijective to establish this result.

Proposition 5.3 below states the consistency of the estimators $(\widehat{f}_n, \widehat{a}_n)$ in the case $b = 2$ and $\mathcal{F} = W^{s,p}$. Assume that for any a in $\mathcal{A}$, $\nu_a \in \mathcal{D}_2$ is of the form

$$\nu_a(x, x') = \nu_{q_a}(x) q_a(x, x') \text{ with } q_a(x, x') = c_a(x)\rho_a(||x - x'||) \ ,$$

where $\rho_- \leq \rho_a \leq \rho_+$. It is also assumed that there exists a unique $a_\star \in \mathcal{A}$ such that $\nu_\star = \nu_{a_\star}$ and that $\rho_{a_\star}$ is one-to-one. Proposition 5.3 is a direct application of Corollary 3.3 and of Proposition 4.3 and is stated without proof.

Proposition 5.3 ($\mathcal{F} = \mathbf{W^{s,p}}$, $\mathbf{s > m/p + k}$, $\mathbf{k \geq 1}$, $\mathbf{b = 2}$). *Assume that H1 and H3-6 hold. Assume also that* $I(f) = \|f\|^{1/\upsilon}_{W^{s,p}}$ *for some* υ *such that* $2\upsilon\ell < 1$ *and that* λ_n *and* ρ_n *satisfy*

$$\lambda_n n^{-1} \underset{n\to+\infty}{\longrightarrow} 0, \ \lambda_n n^{-1/2} \underset{n\to+\infty}{\longrightarrow} +\infty \text{ and } \rho_n = O\left(\frac{\lambda_n}{n}\right) \ .$$

Then,

$$\mathcal{F}_\star = \{f_\star \circ \phi;\ \phi \text{ is an isometry of } \mathbb{X}\} \ ,$$

and

$$d_{\mathcal{C}^k}\left(\widehat{f}_n, \mathcal{F}_\star\right) = o_{\mathbb{P}}(1) \quad \text{and} \quad d_{\mathcal{A}}\left(\widehat{a}_n, a_\star\right) = o_{\mathbb{P}}(1) \ ,$$

6 Numerical experiments

6.1 Proposed Expectation Maximization algorithm

This section introduces a practical algorithm to compute the estimators defined in (4) when ρ_n is set to zero. It is assumed that the maximizer in (4) exists which is the case for instance in the Sobolev framework of Section 5 and if $\mathcal{D}_b$ is compact. This proposed Expectation-Maximization (EM) based procedure iteratively produces a sequence of estimates $\widehat{\nu}^t$, $\widehat{f}^t$, $t \geq 0$, see [9]. Assume that the current parameter estimates are given by $\widehat{\nu}^t$ and $\widehat{f}^t$. The estimates $\widehat{\nu}^{t+1}$ and $\widehat{f}^{t+1}$ are defined as one of the maximizers of the function Q:

$$(\nu, f) \mapsto Q((\nu, f), (\widehat{\nu}^t, \widehat{f}^t)) \overset{\text{def}}{=} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{\nu}^t, \widehat{f}^t}\left[\ln p_{f,\nu}\left(\mathbf{X}_k, \mathbf{Y}_k\right) \middle| \mathbf{Y}_k\right] - \lambda_n I(f) \ ,$$

where $\mathbb{E}_{\widehat{\nu}^t, \widehat{f}^t}[\cdot]$ denotes the conditional expectation under the model parameterized by $\widehat{\nu}^t$ and $\widehat{f}^t$ and where, for any $\mathbf{x} = (x_0, \ldots, x_{b-1}) \in \mathbb{X}^b$ and any $\mathbf{y} = (y_0, \ldots, y_{b-1}) \in \mathbb{R}^{\ell b}$,

$$p_{f,\nu}(\mathbf{x}, \mathbf{y}) \overset{\text{def}}{=} \nu(\mathbf{x}) \prod_{i=0}^{b-1} \varphi(y_i - f(x_i)) \ .$$

Note that the intermediate quantity $Q((\nu, f), (\widehat{\nu}^t, \widehat{f}^t))$ can be written:

$$Q((\nu, f), (\widehat{\nu}^t, \widehat{f}^t)) = Q_t^1(\nu) + Q_t^2(f) \ ,$$

where

$$Q_t^1(\nu) \overset{\text{def}}{=} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{\nu}^t, \widehat{f}^t}\left[\ln\{\nu(\mathbf{X}_k)\} \middle| \mathbf{Y}_k\right] \ , \tag{23}$$

$$Q_t^2(f) \overset{\text{def}}{=} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{\nu}^t, \widehat{f}^t}\left[\ln\left\{\prod_{i=0}^{b-1} \varphi\left(Y_{bk+i} - f(X_{bk+i})\right)\right\} \middle| \mathbf{Y}_k\right] - \lambda_n I(f) \ . \tag{24}$$

Therefore $\widehat{\nu}^{t+1}$ is obtained by maximizing the function $\nu \mapsto Q_t^1(\nu)$ and $\widehat{f}^{t+1}$ by maximizing the function $f \mapsto Q_t^2(f)$. Lemma 6.1 proves that the penalized pseudo-likelihood increases at each iteration of this EM based algorithm.

Lemma 6.1. *The sequences $\widehat{\nu}^t$ and $\widehat{f}^t$ satisfy*

$$\sum_{k=0}^{n-1} \ln p_{\widehat{f}^{t+1},\widehat{\nu}^{t+1}}\left(\mathbf{Y}_k\right) - \lambda_n I(\widehat{f}^{t+1}) \geq \sum_{k=0}^{n-1} \ln p_{\widehat{f}^{t},\widehat{\nu}^{t}}\left(\mathbf{Y}_k\right) - \lambda_n I(\widehat{f}^{t}) \; .$$

Proof. The proof follows the same lines as the one for the usual EM algorithm. For all $0 \leq k \leq n-1$, all $f \in \mathcal{F}$ and all $\nu \in \mathcal{D}_b$,

$$\begin{aligned}
\ln\left[p_{f,\nu}\left(\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n}\right] &= \ln\left[\int p_{f,\nu}\left(\mathbf{x},\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n} \mu^{\otimes b}(\mathrm{d}\mathbf{x})\right] \; , \\
&= \ln\left[\int p_{f,\nu}\left(\mathbf{x},\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n} \frac{p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right)}{p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right)} \mu^{\otimes b}(\mathrm{d}\mathbf{x})\right] \; , \\
&= \ln\left[\int p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right) \frac{p_{f,\nu}\left(\mathbf{x},\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n}}{p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right)} \mu^{\otimes b}(\mathrm{d}\mathbf{x})\right] \; , \\
&\geq \int p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right) \ln\left[\frac{p_{f,\nu}\left(\mathbf{x},\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n}}{p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{x}|\mathbf{Y}_k\right)}\right] \mu^{\otimes b}(\mathrm{d}\mathbf{x}) \; ,
\end{aligned}$$

where the last inequality comes from the concavity of $x \mapsto \log x$. Then,

$$\begin{aligned}
&\ln\left[p_{f,\nu}\left(\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(f)/n}\right] - \ln\left[p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{Y}_k\right) \mathrm{e}^{-\lambda_n I(\widehat{f}^t)/n}\right] \\
&\qquad \geq \mathbb{E}_{\widehat{\nu}^t,\widehat{f}^t}\left[\ln p_{f,\nu}\left(\mathbf{X}_k,\mathbf{Y}_k\right) - \ln p_{\widehat{f}^t,\widehat{\nu}^t}\left(\mathbf{X}_k,\mathbf{Y}_k\right) \middle| \mathbf{Y}_k\right] - \frac{\lambda_n}{n}\left(I(f) - I(\widehat{f}^t)\right) \; .
\end{aligned}$$

The proof is concluded by definition of $\widehat{\nu}^{p+1}$ and $\widehat{f}^{p+1}$. □

Remark 6.2. Like for all EM or gradient based procedures, there is no guarantee that the sequence $(\widehat{f}^t, \widehat{\nu}^t)_{t\geq 0}$ converges, when t grows to infinity, towards the target estimate:

$$(\widehat{f}_n, \widehat{\nu}_n) = \operatorname*{argmax}_{f,\nu} \left\{ \sum_{k=0}^{n-1} \ln p_{f,\nu}\left(\mathbf{Y}_k\right) - \lambda_n I(f) \right\} \; .$$

Lemma 6.1 only ensures that $(\widehat{f}^t, \widehat{\nu}^t)_{t\geq 0}$ converges towards a local maximum of the penalized pseudo likelihood. This limitation is proper to models with hidden data.

6.2 Experimental results

This section illustrates the convergence of the estimates (4) using the EM procedure of Section 6.1. The state-space is $\mathbb{X} = [0,1]$ and the unknown function $f_\star$ is given by

$$\begin{array}{rccc} f_\star & : & [0,1] & \to & \mathbb{R}^2 \\ & & x & \mapsto & (\cos(\pi x), \sin(\pi x)) \; . \end{array}$$

Therefore, throughout this section $m = 1$ and $\ell = 2$. As shown in Section 4, the identifiability of $f_\star$ up to an isometric function of $[0,1]$ can be obtained:

- In the case $b = 1$ when $\nu_\star$ is assumed to be known.
- In the case $b = 2$ when $\mathcal{D}_2$ is the set of probability densities defined on $\mathbb{X}^2$ and of the form $\nu(x,x') = \nu_1(x) \cdot c(x)\rho(|x - x'|)$.

The performance of the algorithm is assessed with two numerical experiments.

- First, $(X_k)_{k\geq 0}$ is assumed to be i.i.d. uniformly distributed on $[0,1]$ and only $f_\star$ is estimated using $b = 1$ in (4).

- Then, $(X_k)_{k\geq 0}$ is assumed to be a Markov chain with density kernel given by

$$q_\star(x,x') = q_{a_\star}(x,x') \stackrel{\text{def}}{=} C_{a_\star}(x)\exp\left(-\frac{|x'-x|}{a_\star}\right)$$

and $a_\star$ and $f_\star$ are estimated using $b=2$ in (4).

In both cases, we wish to use the Sobolev setting of Section 5 with λ_n such that $\lambda_n \propto \log(n)n^{1/2}$ and $I(f) = ||f||^{1/v}_{W^{2,2}}$ with $1/v > b\ell = 2b$ so that the hypothesis of Propositions 5.1 and 5.3 are fulfilled. However, as discussed in the next section, such a complexity function I may be intractable for the optimization problem.

6.2.1 Approximations

The computation of the intermediate quantities (23) and (24) requires an approximation of the conditional expectations $\mathbb{E}_{\widehat{\nu}^t,\widehat{f}^t}\left[h(\mathbf{X}_k,\mathbf{Y}_k)|\mathbf{Y}_k\right]$. For each $0\leq k\leq n-1$, the approximation of the distribution of $\mathbf{X}_k$ conditionally on $\mathbf{Y}_k$ when the parameters are $(\widehat{\nu}^t\widehat{f}^t)$ is dealt with Monte Carlo simulations. For each $t\geq 0$ and each $0\leq k\leq n-1$, the Monte Carlo approximation is based on a set of particles $\{\mathbf{\Xi}_k^{t,j}\}_{j=1}^{N_{mc}}$, where $\mathbf{\Xi}_k^{t,j} = (\xi_{k,0}^{t,j},\ldots,\xi_{k,b-1}^{t,j})$, associated with weights $\{\omega_k^{t,j}\}_{j=1}^{N_{mc}}$ such that for any bounded function h:

$$\mathbb{E}_{\widehat{\nu}^t,\widehat{f}^t}\left[h(\mathbf{X}_k,\mathbf{Y}_k)\middle|\mathbf{Y}_k\right] \approx \sum_{j=1}^{N_{mc}} \omega_k^{t,j} h(\mathbf{\Xi}_k^{t,j},\mathbf{Y}_k)\ .$$

Therefore, (23) and (24) are approximated by:

$$Q_t^1(\nu) \approx \sum_{k=0}^{n-1}\sum_{j=1}^{N_{mc}} \omega_k^{t,j}\ln\left\{\nu(\mathbf{\Xi}_k^{t,j})\right\}\ , \tag{25}$$

$$Q_t^2(f) \approx -\frac{1}{2}\sum_{k=0}^{n-1}\sum_{j=1}^{N_{mc}} \omega_k^{t,j}\sum_{i=0}^{b-1}\|Y_{bk+i} - f(\xi_{k,i}^{t,j})\|^2 - \lambda_n\|f\|^{1/v}_{W^{2,2}}\ . \tag{26}$$

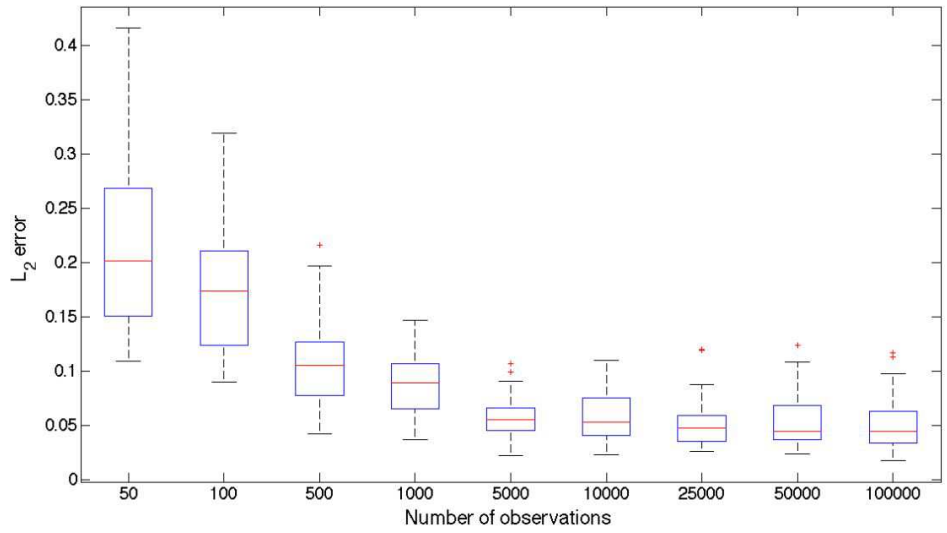
However, the maximization of (26) when $1/v > 2b$ may be complex. Relaxing the hypothesis $1/v > 2b$ by choosing $I(f) = \|f\|^2_{W^{2,2}}$ $(1/v = 2)$ allows to compute the maximizer of (26) as in [7] where the setting is similar except that $I(f) = \|f''\|^2_{L^2}$. [7] shows that the optimization problem can be written as an orthogonal projection in a Hilbert space. Nevertheless, using $1/v > 2b$ (where $2b = 2$ in the first study and $2b = 4$ in the second one) as requested by Propositions 5.1 and 5.3 leads to a much more complicated optimization problem since it can not be interpreted as an orthogonal projection in a Hilbert space. Moreover, the maximization of (26) has been widely studied when $I(f) = \|f\|^{1/v}_{W^{2,2}}$ is replaced by $I(f) = \|f''\|^2_{L^2}$. In this setting, $\widehat{f}^{p+1}$ is then a regression spline (see for instance [7, 16]). Therefore, the constraints on $I(f)$ required by Propositions 5.1 and 5.3 are relaxed in the simulations below where $I(f) = \|f''\|^2_{L^2}$ and where pre-built optimized routines are used to compute $\widehat{f}^{t+1}$ given $\widehat{f}^t$.

6.2.2 Experiment 1: $(X_k)_{k\geq 0}$ i.i.d.

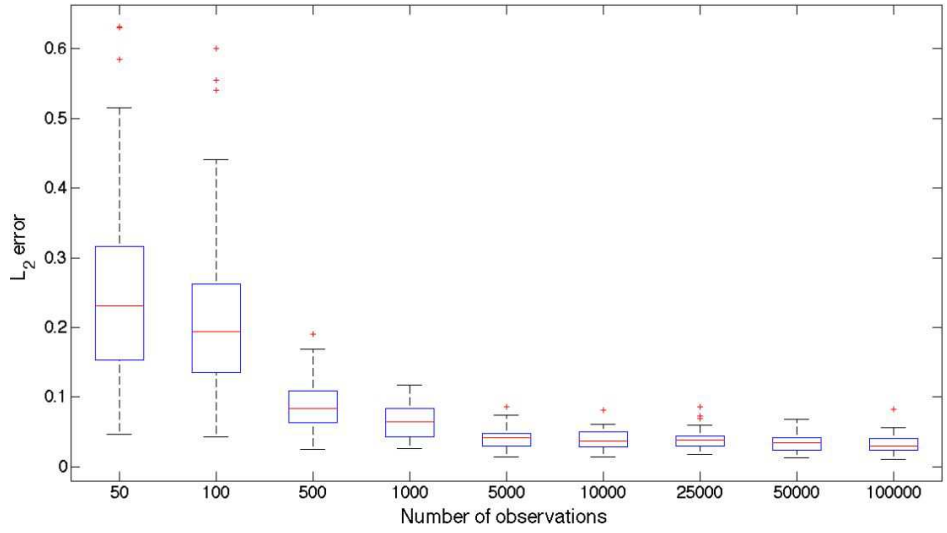
In this section, $b=1$ and $\nu_\star = 1$ is assumed to be known. The estimation of $f_\star$ is performed with $N_{mc} = 100$. In this case, for each $t\geq 0$, $0\leq k\leq n-1$ and $1\leq j\leq N_{mc}$,

$$\xi_{k,0}^{t,j} = \xi_k^{t,j} \sim \nu_\star \quad \text{and} \quad \omega_k^{t,j} \propto \varphi(Y_k - \widehat{f}^t(\xi_k^{t,j}))\ .$$

Figure 1 displays the L^2 error of the estimation of $f_\star$ after 100 iterations as a function of the number of observations. The L^2 estimation error decreases quickly for small values of n (lower than 5000) and then goes on decreasing at a lower rate as n increases. It can be seen that even with a great number of observations, a small bias still remains for both functions (with a mean a bit lower than 0.05). Indeed, there is always small errors in the estimation of $f_\star$ around $x=0$ and $x=1$.



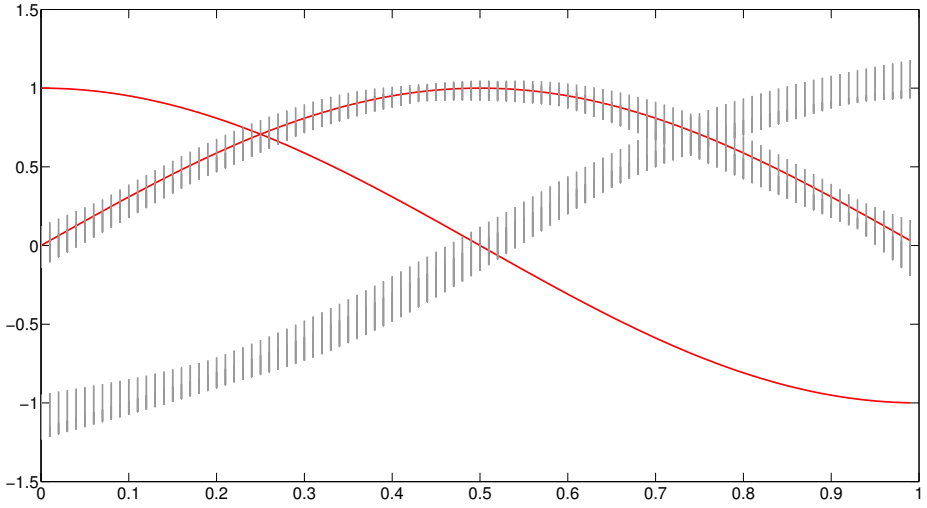
(a) f_1.



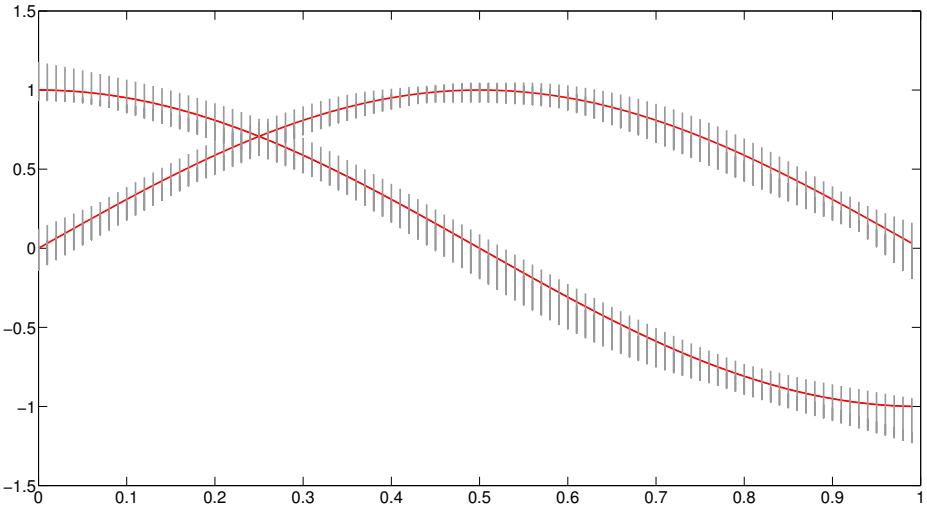
(b) f_2.

Figure 1: L^2 error after 100 iterations over 100 Monte Carlo runs.

Figure 2 shows the estimates after 100 iterations when $n = 25.000$. It can be seen that the second component of $f_\star$ is estimated with accuracy while the first component of $f_\star$ is recovered up to the isometry $x \mapsto 1 - x$ (the isometry is used in Figure 1 to compute the L^2 error). This simulation illustrates the identifiability results obtained in Section 4.



(a) With no isometry for f_1.



(b) With the isometry $x \mapsto 1 - x$ for f_1.

Figure 2: True functions (bold lines) and estimates after 100 iterations (vertical lines) over 100 Monte Carlo runs ($n = 25.000$).

6.2.3 Experiment 2: $(X_k)_{k\geq 0}$ Markov chain

In this section, $b = 2$ and $a_\star$ and $f_\star$ are estimated. Define for any $a > 0$,

$$\nu_a(x, x') = \nu_{1,a}(x) \cdot c_a(x) \exp\left(-\frac{|x - x'|}{a}\right) ,$$
$$\nu_{1,a}(x) \propto c_a^{-1}(x) = \int_{[0,1]} \exp\left(-\frac{|x - x'|}{a}\right) \mathrm{d}x' .$$

$\widehat{\nu}^{t+1}$ is given by $\nu_{\widehat{a}^{t+1}}$ where $\widehat{a}^{t+1}$ is computed by maximizing the function

$$a \mapsto \log\left(a + a^2(\exp(-1/a) - 1)\right) + \frac{1}{na} \sum_{k=0}^{n-1} \sum_{j=1}^{N_{mc}} \omega_k^{t,j} |\xi_{k,0}^{t,j} - \xi_{k,1}^{t,j}| ,$$

where, for all $0 \leq k \leq n-1$, $(\xi_{k,0}^{t,j}, \xi_{k,1}^{t,j})_{j=1}^{N_{mc}}$ are independently sampled uniformly in $[0,1]\times[0,1]$ and associated with the importance weights:

$$\omega_k^{t,j} \propto \nu_{\widehat{a}^t}(\xi_{k,0}^{t,j}) q_{\widehat{a}^t}(\xi_{k,0}^{t,j}, \xi_{k,1}^{t,j}) \varphi(Y_{2k} - \widehat{f}^t(\xi_{k,0}^{t,j})) \varphi(Y_{2k+1} - \widehat{f}^t(\xi_{k,1}^{t,j})) \; . \tag{27}$$

The Monte Carlo approximations are computed using $N_{mc} = 200$ and 20.000 observations (*i.e.* $n = 10.000$) are sampled. Figure 3 displays the estimation $a_\star$ as a function of the number of iterations of the EM algorithm over 50 independent Monte Carlo runs. The estimates converge to the true value of $a_\star$ after few iterations (about 25).

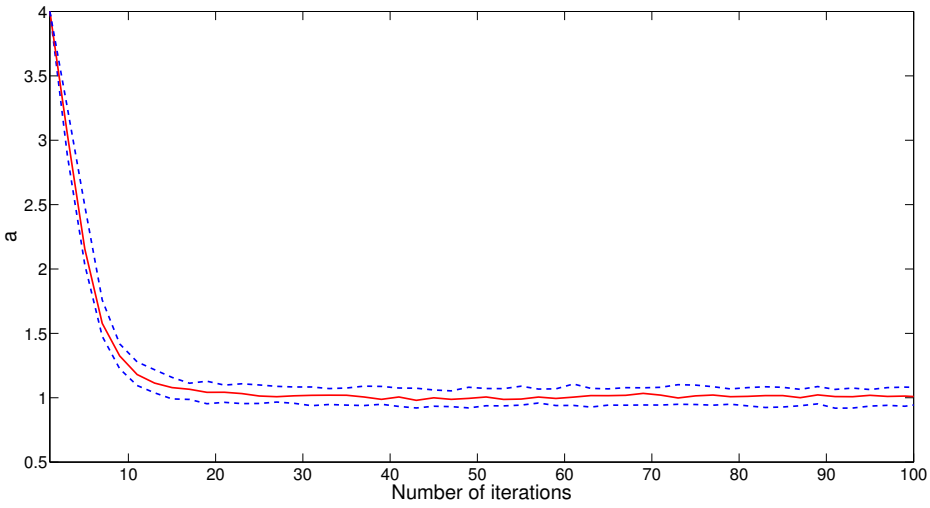


Figure 3: Estimation of $a_\star$ as a function of the number of iterations of the EM algorithm. The true value is $a_\star = 1$. Median (bold line) and upper and lower quartiles (dotted line) over 50 Monte Carlo runs.

Figure 4 illustrates Corollary 5.2. It displays the estimation of $f_\star([0,1])$ after 100 iterations for several Monte Carlo runs. It shows that despite the variability of the estimation, the image is well estimated with few observations.

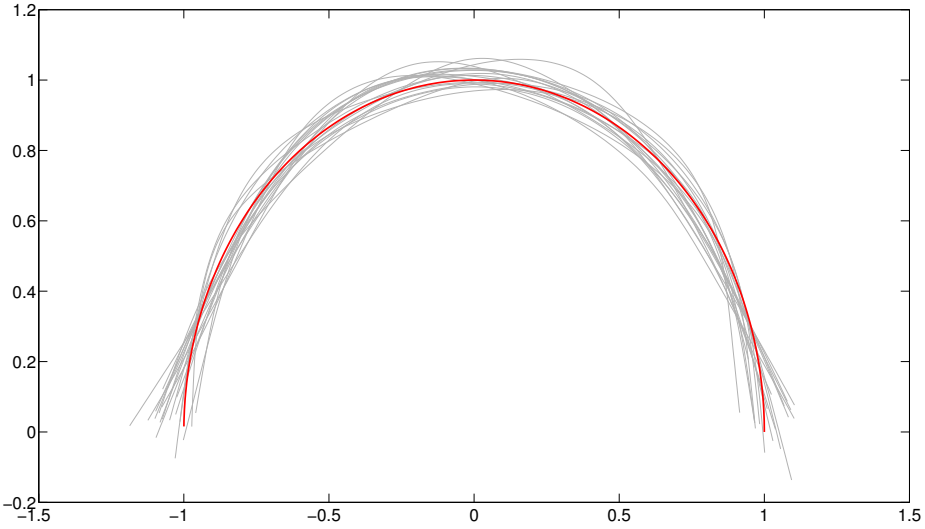


Figure 4: True image $f_\star([0,1])$ (red) and estimates after 100 iterations of the algorithm over 100 Monte Carlo runs (grey).

7 Conclusion

This paper deals with the estimation of the unknown function $f_\star$ in the model $Y_k = f_\star(X_k) + \epsilon_k$ when the regressors X_k are not observed. These regressors are assumed to be Φ-mixing which covers in particular the i.i.d. case and some Markovian cases. The consistency of a penalized pseudo-likelihood approach is first proved: the estimator of the density of a fixed number of observations $(Y_0, \dots, Y_{b-1})$ converges towards the true density in Hellinger distance. Then, it is shown that the model is not identifiable in general when the candidate functions are in $\mathcal{C}^1(\mathbb{X})$ ($\mathbb{X} \subset \mathbb{R}^m$). Nevertheless, the candidate functions and the hidden state densities are necessarily linked by a bijective function ϕ. In the special case where the conditional density of X_1 given X_0 is of the form $q_\star(x, x') = c_\star(x) q_\star(||x - x'||)$, this function ϕ is an isometry of $\mathbb{X}$. In the experimental section, the hypothesis on the penalty term assumed in the theoretical part is relaxed in order to ease the computation of the estimators. The estimation procedure seems still to provide good results. These observations might indicate that the assumptions of our main results are not optimal and that weaker penalty terms might ensure consistency. The assumptions could probably be weakened by a sharper analysis of the empirical process described in Section 8.1 and, in particular, by improving the bound on the bracketing entropy described in Section B.

8 Proofs

8.1 Proof of Proposition 3.2

The proof relies on the application of Proposition A.1 and Proposition A.2 to obtain first a concentration inequality for the class of functions $\mathcal{G}_M$, where $M \geq 1$, defined as:

$$\mathcal{G}_M \stackrel{\text{def}}{=} \left\{ g_{p_{f,\nu}} \; ; \; \nu \in \mathcal{D}_b, \; f \in \mathcal{F} \text{ and } I(f) \leq M \right\} \; ,$$

where $p_{f,\nu}$ is defined by (3) and $g_{p_{f,\nu}}$ by (12). For any $p > 0$, denote by $\mathrm{L}^p(\mathbb{P}_\star)$ the set of functions $g : \mathbb{R}^{b\ell} \to \mathbb{R}$ such that $\mathbb{E}\left[|g(\mathbf{Y}_0)|^p\right] < +\infty$. For any $\kappa > 0$ and any set $\mathcal{G}$ of functions from $\mathbb{R}^{b\ell}$ to $\mathbb{R}$, let $N(\kappa, \mathcal{G}, \|\cdot\|_{\mathrm{L}^p(\mathbb{P}_\star)})$ be the smallest integer N such that there exists a set of functions $\left\{ \left(g_i^L, g_i^U\right) \right\}_{i=1}^N$ for which:

a) $\|g_i^U - g_i^L\|_{\mathrm{L}^p(\mathbb{P}_\star)} \leq \kappa$ for all $i \in \{1, \cdots, N\}$;

b) for any g in $\mathcal{G}$, there exists $i \in \{1, \cdots, N\}$ such that

$$g_i^L \leq g \leq g_i^U \; .$$

$N(\kappa, \mathcal{G}, \|\cdot\|_{\mathrm{L}^p(\mathbb{P}_\star)})$ is the κ-number with bracketing of $\mathcal{G}$, and $H(\kappa, \mathcal{G}, \|\cdot\|_{\mathrm{L}^p(\mathbb{P}_\star)}) \stackrel{\text{def}}{=} \ln N(\kappa, \mathcal{G}, \|\cdot\|_{\mathrm{L}^p(\mathbb{P}_\star)})$ is the κ-entropy with bracketing of $\mathcal{G}$. For any bounded function g, define

$$S_n(g) \stackrel{\text{def}}{=} n \int g \, \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \; . \tag{28}$$

Application of Proposition A.1 Proposition A.1 is applied to the class of functions $\overline{\mathcal{G}}_M$ defined as

$$\overline{\mathcal{G}}_M \stackrel{\text{def}}{=} \left\{ \mathbf{g} - \mathbb{E}_\star\left[\mathbf{g}\right] ; \; g \in \mathcal{G}_M \right\} \; .$$

Since $(\boldsymbol{\epsilon}_k)_{k\geq 0}$ is i.i.d. and $(\mathbf{X}_k)_{k\geq 0}$ is Φ-mixing, $(\mathbf{Y}_k)_{k\geq 0}$ is also Φ-mixing with mixing coefficients $(\phi_i^{\mathbf{Y}})_{i\geq 0}$ satisfying, for all $i \geq 1$,

$$\phi_i^{\mathbf{Y}} \leq \phi_i^{\mathbf{X}} = \phi_{(i-1)b+1}^X \; .$$

Therefore $\mathbf{\Phi}^{\mathbf{Y}} = \sum_{i\geq 1} (\phi_i^{\mathbf{Y}})^{1/2} < \infty$. By H2, there exists $C > 0$ such that for any $i \geq 0$, and any $g \in \mathcal{G}_M$,

$$\begin{aligned} |\mathbf{g}\left(\mathbf{Y_i}\right)| \leq CM^\upsilon \left(1 + \|\mathbf{Y}_i\|\right) &\leq CM^\upsilon \left(1 + ||\mathbf{f}_\star(\mathbf{X}_i)\| + \|\boldsymbol{\epsilon}_i\|\right) \; , \\ &\leq CM^\upsilon \left(1 + \|f_\star\|_\infty + \|\boldsymbol{\epsilon}_i\|\right) \; , \\ &\leq CM^\upsilon \left(1 + \|\boldsymbol{\epsilon}_i\|\right) \; . \end{aligned}$$

Define $\mathbf{U}_i \stackrel{\text{def}}{=} CM^{\upsilon}\left(1+\|\boldsymbol{\epsilon}_i\|\right)$. Then $(\mathbf{U}_i)_{i\geq 0}$ is i.i.d., $\mathbf{g}\left(\mathbf{Y_i}\right) \leq \mathbf{U_i} + \mathbb{E}\left[\mathbf{U_0}\right]$ and there exist positive constants ν and c such that

$$\mathbb{E}\left[\left(\mathbf{U}_i + \mathbb{E}\left[\mathbf{U}_0\right]\right)^{2k}\right] \leq k! \nu c^{k-1} ,$$

where $\nu = CM^{2\upsilon}$ and $c = CM^{2\upsilon}$. Then, by Proposition A.1, there exists a positive constant c such that for any positive x,

$$\mathbb{P}\left[\sup_{g\in\mathcal{G}_M} |S_n(g)| \geq \mathbb{E}\left[\sup_{g\in\mathcal{G}_M} |S_n(g)|\right] + c\boldsymbol{\Phi}^{\mathbf{Y}}\left(\sqrt{nx}+x\right) M^{\upsilon}\right] \leq \mathrm{e}^{-x} . \tag{29}$$

Application of Proposition A.2 Proposition A.2 is used to control the inner expectation in (29). Let $r > 1$. By [25, Lemma 7.26] and since the Hellinger distance is bounded by 1, there exists a constant δ such that for any $g = g_{p_{f,\nu}} \in \mathcal{G}_M$.

$$\|g\|^{2r}_{\mathrm{L}^{2r}(\mathbb{P}_\star)} \leq Ch^2(p_{f,\nu}, p_\star) \leq \delta .$$

By Lemma B.1, for any $p' \geq 1$, and any $s' > b\ell/p'$, provided that $d > s' + b\ell(1-\frac{1}{p'})$, there exists a constant C such that, for all $u > 0$,

$$H(u, \|\cdot\|_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G}_M) \leq C\left(\frac{M^{\upsilon(s'+d+\frac{b\ell}{p'})}}{u^{2r}}\right)^{b\ell/s'} . \tag{30}$$

For any $p' \geq 1$, and any $s' > b\ell/p'$, provided that $d > s' + b\ell(1-\frac{1}{p'})$, there exists a constant C such that

$$\begin{aligned}\varphi(\delta) &\stackrel{\text{def}}{=} \int_0^\delta H^{1/2}(u, \|\cdot\|_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G}_M)\mathrm{d}u , \\ &\leq CM^{(s'+d+b\ell/p')\frac{b\ell\upsilon}{2s'}} \int_0^\delta u^{-rb\ell/s'}\mathrm{d}u .\end{aligned}$$

Choosing $d \leq s' + b\ell(1-\frac{1}{p'}) + 2$, if s' grows to $+\infty$ then the last integral is finite, and $(s'+d+b\ell/p')\frac{b\ell\upsilon}{2s'}$ tends to $b\ell\upsilon$, so that for any $\eta > 0$ there exists a positive constant C such that

$$\varphi(\delta) \leq CM^{b\ell\upsilon+\eta} .$$

Finally, by Proposition A.2 for any $\eta > 0$, there exists a constant A such that for n large enough

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_M} |S_n(g)|\right] \leq A\sqrt{n}M^{b\ell\upsilon+\eta} .$$

Then, by (29), this yields

$$\mathbb{P}\left[\sup_{g\in\mathcal{G}_M} |S_n(g)| \geq c\boldsymbol{\Phi}^{\mathbf{Y}}\left(\sqrt{nx}+x\right) M^{\upsilon} + A\sqrt{n}M^{b\ell\upsilon+\eta}\right] \leq \mathrm{e}^{-x} . \tag{31}$$

Proposition 3.2 is then proved using a peeling argument. By (28) and (31), for any $M \geq 1$, any $n \geq N$ and any $x > 0$, if $\gamma = b\ell\upsilon + \eta$,

$$\mathbb{P}\left[\sup_{g\in\mathcal{G}_M} \frac{\left|\int g \,\mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)\right|}{M^\gamma} \geq c\boldsymbol{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{A}{\sqrt{n}}\right] \leq e^{-M^{\gamma-\upsilon}x} . \tag{32}$$

We can write

$$\mathbb{P}\left[\sup_{f\in\mathcal{F},\ \nu\in\mathcal{D}_b} \frac{\left|\int g_{p_{f,\nu}} \,\mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)\right|}{1 \vee I(f)^\gamma} \geq c\boldsymbol{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}} + \frac{x}{n}\right) + \frac{(2^\gamma \vee 1)A}{\sqrt{n}}\right] \leq P_1 + \sum_{k=0}^{+\infty} T_k ,$$

where

$$P_1 \stackrel{\text{def}}{=} \mathbb{P}\left[\sup_{\substack{f\in\mathcal{F};\, I(f)\le 1,\\ \nu\in\mathcal{D}_b}} \frac{\left|\int g_{p_{f,\nu}}\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right|}{1\vee I(f)^\gamma} \ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{x}{n}\right)+\frac{(2^\gamma\vee 1)A}{\sqrt{n}}\right] ,$$

$$T_k \stackrel{\text{def}}{=} \mathbb{P}\left[\sup_{\substack{f\in\mathcal{F};\, 2^k<I(f)\le 2^{k+1},\\ \nu\in\mathcal{D}_b}} \frac{\left|\int g_{p_{f,\nu}}\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right|}{1\vee I(f)^\gamma} \ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{x}{n}\right)+\frac{(2^\gamma\vee 1)A}{\sqrt{n}}\right] .$$

By (32),

$$\begin{aligned} P_1 &\le \mathbb{P}\left[\sup_{g\in\mathcal{G}_1}\left|\int g\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right| \ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{x}{n}\right)+\frac{(2^\gamma\vee 1)A}{\sqrt{n}}\right] ,\\ &\le \mathbb{P}\left[\sup_{g\in\mathcal{G}_1}\left|\int g\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right| \ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{\sqrt{c}x}{n}\right)+\frac{A}{\sqrt{n}}\right] ,\\ &\le \mathrm{e}^{-x} \end{aligned}$$

and for all $k \ge 0$,

$$\begin{aligned} T_k &\le \mathbb{P}\left[\sup_{g\in\mathcal{G}_{2^{k+1}}} \frac{\left|\int g\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right|}{2^{\gamma(k+1)}} \ge \frac{c}{2^\gamma}\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{x}{n}\right)+\frac{(2^\gamma\vee 1)A}{2^\gamma\sqrt{n}}\right] ,\\ &\le \mathbb{P}\left[\sup_{g\in\mathcal{G}_{2^{k+1}}} \frac{\left|\int g\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right|}{2^{\gamma(k+1)}} \ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{2^{2\gamma}n}}+\frac{x}{2^{2\gamma}n}\right)+\frac{A}{\sqrt{n}}\right] ,\\ &\le \mathrm{e}^{-2^{(\gamma-\upsilon)(k+1)}x/2^{2\gamma}} . \end{aligned}$$

Using (32),

$$\begin{aligned} \mathbb{P}\Bigg[\sup_{f\in\mathcal{F},\ \nu\in\mathcal{D}_b} \frac{\left|\int g_{p_{f,\nu}}\, \mathrm{d}(\mathbb{P}_n-\mathbb{P}_\star)\right|}{1\vee I(f)^\gamma} &\ge c\mathbf{\Phi}^{\mathbf{Y}}\left(\sqrt{\frac{x}{n}}+\frac{x}{n}\right)+\frac{(2^\gamma\vee 1)A}{\sqrt{n}}\Bigg]\\ &\le e^{-x}+\sum_{k=0}^{\infty} e^{-2^{(\gamma-\upsilon)(k+1)}x/2^{2\gamma}}\\ &\le e^{-x}+\sum_{k=0}^{\infty} e^{-(k+1)x\log(2)(\gamma-\upsilon)/2^{2\gamma}}\\ &\le e^{-x}+\frac{e^{-\alpha x}}{1-e^{-\alpha x}} , \end{aligned}$$

which concludes the proof of Proposition 3.2.

8.2 Proof of Proposition 4.1

Assume that $h(p_{f,\nu}, p_{f_\star,\nu_\star}) = 0$ (the proof of the converse proposition is straightforward). Let X_0' be a random variable on $\mathbb{X}$ with distribution $\nu(x)\mu(\mathrm{d}x)$. Since ϵ_0 is a Gaussian random variable, $h(p_{f,\nu}, p_{f_\star,\nu_\star}) = 0$ implies that $f(X_0')$ has the same distribution as $f_\star(X_0)$.

Proof that f and $f_\star$ have the same image in $\mathbb{R}^\ell$. Let $y \in f(\mathbb{X})$ and $n \ge 1$. Using $\nu \ge \nu_-$, the continuity of f and $\overline{\overset{\circ}{\mathbb{X}}} = \mathbb{X}$, $f(X_0')$ has the same distribution as $f_\star(X_0)$ implies that,

$$\begin{aligned} \mathbb{P}\left\{X_0 \in f_\star^{-1}\left(B\left(y, n^{-1}\right)\right)\right\} &= \mathbb{P}\left\{X_0' \in f^{-1}\left(B\left(y, n^{-1}\right)\right)\right\} ,\\ &\ge \nu_-\, \mu\left\{f^{-1}\left(B\left(y, n^{-1}\right)\right)\right\} > 0 , \end{aligned}$$

as $f^{-1}(B(y, n^{-1}))$ is a nonempty open subset of $\mathbb{X}$. Therefore $f_\star^{-1}\left(B\left(y, n^{-1}\right)\right)$ is nonempty and for all $n \ge 1$, there exists $x_n \in \mathbb{X}$ such that $\|y - f_\star(x_n)\| < n^{-1}$. For all $n \ge 1$, $f_\star(x_n)$ is in the compact set $f_\star(\mathbb{X})$ which implies that $y \in f_\star(\mathbb{X})$. The proof of the converse inclusion follows the same lines.

Proof that ϕ is bijective. Since $f(X_0')$ has the same distribution as $f_\star(X_0)$, X_0 has the same distribution as $\phi(X_0')$ where $\phi \stackrel{\text{def}}{=} f_\star^{-1} \circ f$. By H5 ϕ exists and is $\mathcal{C}^1$. We prove that $|J_\phi| > 0$ using the following result due to [13, Theorem 2, p.99].

Lemma 8.1. *If $\phi : \mathbb{X} \to \mathbb{X}$ is Lipschitz then, for any integrable function g,*

$$\int_{\mathbb{X}} g(x)\,|J_\phi(x)|\,\mu(\mathrm{d}x) = \int_{\mathbb{X}} \sum_{x \in \phi^{-1}(\{y\})} g(x)\mu(\mathrm{d}y)\ .$$

Define $A \stackrel{\text{def}}{=} \left\{x \in \mathbb{X}\,;\ \forall x' \in \phi^{-1}(\{x\}),\ |J_\phi(x')| > 0\right\}$. Let h_1 be a bounded measurable real function on $\mathbb{X}$ and define $h \stackrel{\text{def}}{=} \mathbb{1}_A h_1$. By Lemma 8.1,

$$\begin{aligned}
\mathbb{E}\left[h \circ \phi(X_0')\right] &= \int_{\mathbb{X}} h_1(\phi(x'))\mathbb{1}_A(\phi(x'))\nu(x')\mu(\mathrm{d}x')\ , \\
&= \int_{\mathbb{X}} h_1(\phi(x'))\mathbb{1}_A(\phi(x'))\frac{\nu(x')}{|J_\phi(x')|}|J_\phi(x')|\mu(\mathrm{d}x')\ , \\
&= \int_{\mathbb{X}} h_1(x)\mathbb{1}_A(x) \sum_{x' \in \phi^{-1}(\{x\})} \frac{\nu(x')}{|J_\phi(x')|}\mu(\mathrm{d}x)\ .
\end{aligned}$$

Since X_0 has the same distribution as $\phi(X_0')$,

$$\int_{\mathbb{X}} h_1(x)\mathbb{1}_A(x)\nu_\star(x)\mu(\mathrm{d}x) = \int_{\mathbb{X}} h_1(x)\mathbb{1}_A(x) \sum_{x' \in \phi^{-1}(\{x\})} \frac{\nu(x')}{|J_\phi(x')|}\mu(\mathrm{d}x)\ .$$

Applying Lemma 8.1 with $g \stackrel{\text{def}}{=} \mathbb{1}_{|J_\phi|=0}$ implies that $\mathbb{1}_A = 1$ μ-a.s. in $\mathbb{X}$ and, μ-a.s.,

$$\nu_\star(x) = \sum_{x' \in \phi^{-1}(\{x\})} \frac{\nu(x')}{|J_\phi(x')|}\ . \tag{33}$$

Therefore, for μ almost every $x \in \mathbb{X}$ and for all $x' \in \phi^{-1}(\{x\})$,

$$|J_\phi(x')| \geq \frac{\nu_-}{\nu_+}\ .$$

By continuity of J_ϕ and using that $\overline{\overset{\circ}{\mathbb{X}}} = \mathbb{X}$, $|J_\phi(x)| > 0$ for all $x \in \mathbb{X}$. Therefore, ϕ is locally invertible and, since $\mathbb{X}$ is compact, simply connected and arcwise connected, ϕ is bijective by [2, Theorem 1.8, p.47]. Then (33) ensures that for μ almost every $x \in \mathbb{X}$,

$$\nu_\star(\phi(x)) = \frac{\nu(x)}{|J_\phi(x)|}\ ,$$

which concludes the proof of Proposition 4.1.

8.3 Proof of Proposition 4.3 and Corollary 4.4

Proof of Proposition 4.3 The proof of (18) follows the same lines as the proof of Proposition 4.1. Let (X_0', X_1') be a random variable on $\mathbb{X}^2$ with probability density $\nu(x)q(x,x')$ on $\mathbb{X}^2$. $h(p_{f,\nu^2}, p_{f_\star,\nu_\star^2}) = 0$ implies that $h(p_{f,\nu}, p_{f_\star,\nu_\star}) = 0$ and, by Proposition 4.1, $f(\mathbb{X}) = f_\star(\mathbb{X})$ and $\phi = f_\star^{-1} \circ f$ is bijective. Moreover, since (ϵ_0, ϵ_1) has a Gaussian distribution, $h(p_{f,\nu^2}, p_{f_\star,\nu_\star^2}) = 0$ implies that $(\phi(X_0'), \phi(X_1'))$ has the same distribution as (X_0, X_1) so that for any x in $\mathbb{X}$ and any bounded measurable function f on $\mathbb{X}$,

$$\mathbb{E}\left[\phi(X_1')\middle|X_0' = \phi^{-1}(x)\right] = \mathbb{E}\left[X_1 | X_0 = x\right]\ .$$

Following the proof of Proposition 4.1, this gives (18).

Proof of Corollary 4.4 Assume now that

$$q_\star(x,x') = c_\star(x)\rho_\star(||x-x'||)\ ,$$
$$q(x,x') = c(x)\rho(||x-x'||)\ .$$

We may assume, using an eventual modification of $c_\star$ and c that $\rho(0)=\rho_\star(0)=1$. By (18),

$$c(x)\rho(||x-x'||) = |J_\phi(x')|c_\star(\phi(x))\rho_\star(||\phi(x)-\phi(x')||)\ . \tag{34}$$

Applying (34) with $x=x'$ implies $|J_\phi(x)| = c(x)/c_\star(\phi(x))$. Therefore,

$$\frac{|J_\phi(x)|}{|J_\phi(x')|} = \frac{\rho_\star(||\phi(x)-\phi(x')||)}{\rho(||x-x'||)} = \frac{\rho_\star(||\phi(x')-\phi(x)||)}{\rho(||x'-x||)} = \frac{|J_\phi(x')|}{|J_\phi(x)|}$$

and then, for all $x\in\mathbb{X}$, $|J_\phi(x)|=1$.

Now (34) implies that for any x and x' in $\mathbb{X}$,

$$\rho(||x-x'||) = \rho_\star(||\phi(x)-\phi(x')||). \tag{35}$$

Let $x_0\in\overset{\circ}{\mathbb{X}}$, $y_0=\phi(x_0)$ and $d_0,d_0'>0$ be such that $B(x_0,d_0)\overset{\text{def}}{=}\{x\in\mathbb{R}^m\ ,\ ||x_0-x||<d_0\}\subset\mathbb{X}$ and $\phi(B(x_0,d_0))\subset B(y_0,d_0')$.

Let $d<d_0$ and denote by $S(x_0,d)$ the set $S(x_0,d)\overset{\text{def}}{=}\{x\in\mathbb{R}^m\ ,\ ||x_0-x||=d\}$. As $\rho_\star$ is one-to-one, write $F=\rho_\star^{-1}\circ\rho$. (35) implies that $\phi(S(x_0,d))\subset S(y_0,F(d))$. Furthermore, using the compactness and the connectivity of $S(x_0,d)$, $\phi(S(x_0,d))=S(y_0,F(d))$ which, together with the continuity of ϕ, guarantees that $\phi(B(x_0,d))=B(y_0,F(d))$. Finally, because ϕ preserves the volumes, for any $d<d_0$, $F(d)=d$ and for any $x\in\mathbb{X}$ and any $x'\in B(x,d_0)$, $||x-x'||=||\phi(x)-\phi(x')||$. The proof is concluded using the connectivity of $\mathbb{X}$.

A Concentration results for the empirical process of unbounded functions

Proposition A.1 provides a concentration inequality on the empirical process over a class of functions $\mathcal{G}$ for which $|g(Z_i)|$ can be bounded uniformly in $g\in\mathcal{G}$ by an independent process U_i with bounded moments. This unusual condition is more general than [28, Theorem 3] which considered a uniformly bounded class of functions.

Proposition A.1. *Let $(Z_n)_{n\geq 0}$ be a Φ-mixing process taking values in a set $\mathcal{Z}$. Assume that the Φ-mixing coefficients associated with $(Z_n)_{n\geq 0}$ satisfy:*

$$\mathbf{\Phi}\overset{\text{def}}{=}\sum_{i=1}^{\infty}\phi_i^{1/2}<\infty\ .$$

Let $\mathcal{G}$ be some countable class of real valued measurable functions defined on $\mathcal{Z}$. Assume that there exists a sequence of independent random variables $(U_i)_{i\geq 0}$ such that:

- for any g in $\mathcal{G}$,

$$|g(Z_i)|\leq U_i\ a.s.\ ; \tag{36}$$

- there exists some positive numbers ν and c such that, for any $k\geq 1$:

$$\sum_{i=0}^{n-1}\mathbb{E}\left[U_i^{2k}\right]\leq k!n\nu c^{k-1}\ . \tag{37}$$

Then, for any positive x,

$$\mathbb{P}\left[S_n\geq 2\mathbf{\Phi}\left(2\sqrt{n\nu x}+\sqrt{c}x\right)\right]\leq \mathrm{e}^{-x}\ ,$$

where

$$S_n=\sup_{g\in\mathcal{G}}\left|\sum_{i=0}^{n-1}g(Z_i)\right|-\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=0}^{n-1}g(Z_i)\right|\right]\ .$$

Proof. For any real valued random variable and for any real random variable X, define $\psi_X(\lambda) \stackrel{\text{def}}{=} \ln\left(\mathbb{E}\left[\exp\left(\lambda X\right)\right]\right)$, Following the proof of [28, Theorem 3] together with the discussion about the dependence structure in [28, Section 2], we have

$$\exp\left(\psi_{S_n}\left(\frac{\lambda}{4}\right)\right) \leq \mathbb{E}\left[\exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{4}V^2\right]\right]^{\frac{1}{2}} \exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{8}\mathbb{E}\left[V^2\right]\right] , \tag{38}$$

where $V^2 \stackrel{\text{def}}{=} \sum_{i=1}^n U_i^2$. Using (36) and by independence of the $(U_i)_{i\geq 0}$,

$$\begin{aligned}\exp\left(\psi_{S_n}\left(\frac{\lambda}{4}\right)\right) &\leq \mathbb{E}\left[\exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{4}\sum_{i=1}^n U_i^2\right]\right]^{\frac{1}{2}} \exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{8}\sum_{i=1}^n \mathbb{E}[U_i^2]\right] , \\ &\leq \prod_{i=1}^n \mathbb{E}\left[\exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{4}U_i^2\right]\right]^{\frac{1}{2}} \exp\left[\lambda^2\frac{\mathbf{\Phi}^2}{8}\sum_{i=1}^n \mathbb{E}[U_i^2]\right] .\end{aligned}$$

Thus,

$$\psi_{S_n}(\lambda/4) \leq \frac{1}{2}\sum_{i=1}^n \ln\left\{\mathbb{E}\left[\exp\left(\lambda^2\frac{\mathbf{\Phi}^2}{4}U_i^2\right)\right]\right\} + \lambda^2\frac{\mathbf{\Phi}^2}{8}\sum_{i=1}^n \mathbb{E}\left[U_i^2\right] .$$

Since for any $u > 0$, $\ln(u) \leq u - 1$, this yields

$$\psi_{S_n}(\lambda/4) \leq \frac{1}{2}\sum_{k=1}^\infty \frac{1}{k!}\left[\lambda^2\frac{\mathbf{\Phi}^2}{4}\right]^k \sum_{i=1}^n \mathbb{E}\left[U_i^{2k}\right] + \lambda^2\frac{\mathbf{\Phi}^2}{8}\sum_{i=1}^n \mathbb{E}\left[U_i^2\right] .$$

Then, by (37),

$$\psi_{S_n}(\lambda/4) \leq n\nu\left[\lambda^2\frac{\mathbf{\Phi}^2}{4}\right]\frac{1}{2}\sum_{k=0}^\infty \left[\lambda^2\frac{\mathbf{\Phi}^2}{4}c\right]^k + \left[\lambda^2\frac{\mathbf{\Phi}^2}{8}\nu\right] .$$

If $0 < \lambda^2\mathbf{\Phi}^2 c/4 < 1$,

$$\begin{aligned}\psi_{S_n}(\lambda/4) &\leq n\nu\lambda^2\frac{\mathbf{\Phi}^2}{8}\frac{1}{1-\lambda^2\frac{\mathbf{\Phi}^2}{4}c} + n\nu\lambda^2\frac{\mathbf{\Phi}^2}{8} , \\ &\leq n\nu\lambda^2\frac{\mathbf{\Phi}^2}{4}\frac{1}{1-\lambda^2\frac{\mathbf{\Phi}^2}{4}c} .\end{aligned}$$

Define $\nu' \stackrel{\text{def}}{=} 8n\nu\mathbf{\Phi}^2$ and $c' \stackrel{\text{def}}{=} 2\mathbf{\Phi}\sqrt{c}$. Therefore,

$$\psi_{S_n}(\lambda/4) \leq \frac{\nu'(\lambda/4)^2}{2(1-c'(\lambda/4))} . \tag{39}$$

Hence, for all $0 < \lambda < 1/c'$,

$$\psi_{S_n}(\lambda) \leq \frac{\nu'\lambda^2}{2(1-c'\lambda)} . \tag{40}$$

By the Bernstein type inequality (40), [25, Lemma 2.3] gives, for any measurable set $A \subset \Omega$ with $\mathbb{P}(A) > 0$,

$$\mathbb{E}\left[S_n|A\right] \leq \sqrt{2\nu'\ln\left(\frac{1}{\mathbb{P}(A)}\right)} + c'\ln\left(\frac{1}{\mathbb{P}(A)}\right) .$$

Hence, by [25, Lemma 2.4], for any positive x,

$$\mathbb{P}\left[S_n \geq \sqrt{2\nu' x} + c'x\right] \leq \mathrm{e}^{-x} .$$

□

Proposition A.2 below provides a control on the expectation of the empirical process. It introduces a β-mixing condition (see [8]) which is weaker than the Φ-mixing condition considered in Proposition A.1. The β-mixing coefficient between two σ-fields $\mathcal{U}, \mathcal{V} \subset \mathcal{E}$ is defined in [8] by

$$\beta(\mathcal{U}, \mathcal{V}) \stackrel{\text{def}}{=} \frac{1}{2} \sup \sum_{(i,j) \in I \times J} \left| \mathbb{P}\left(U_i \cap V_j\right) - \mathbb{P}(U_i)\mathbb{P}(V_j) \right| \ ,$$

where the supremum is taken over all finite partitions $(U_i)_{i \in I}$ and $(V_j)_{j \in J}$ respectively $\mathcal{U}$ and $\mathcal{V}$ measurable. The corresponding mixing coefficients $(\beta_i)_{i \geq 0}$ associated with a process $(X_k)_{k \geq 0}$ satisfy $\beta_i < \phi_i$ for all $i \geq 1$.

Proposition A.2. *Let $(Z_i)_{i \geq 0}$ be a stationary process taking values in a Polish space $\mathcal{Z}$. Denote by $\mathbb{P}_\star$ the distribution of Z_0 and by $\mathbb{E}_\star$ the expectation under $\mathbb{P}_\star$. Assume that $(Z_i)_{i \geq 0}$ is β-mixing with β coefficients $(\beta_i)_{i \geq 0}$ satisfying*

$$\sum_{i=1}^{\infty} \beta_i < \infty \ .$$

Let $\mathcal{G}$ be a countable class of functions on $\mathcal{Z}$. Assume that there exist $r > 1$ and $\sigma > 0$ such that for any $g \in \mathcal{G}$,

$$||g||_{\mathrm{L}^{2r}(\mathbb{P}_\star)} \stackrel{\text{def}}{=} \mathbb{E}_\star \left[g^{2r}\right]^{1/2r} \leq \delta \ .$$

Assume also that the bracketing function satisfies

$$\int_0^1 \sqrt{H(u, ||\cdot||_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G})} \mathrm{d}u < \infty \ .$$

Then,

$$\varphi(\delta) := \int_0^\delta \sqrt{H(u, ||\cdot||_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G})} \mathrm{d}u$$

is finite and there exists a constant A such that for n big enough

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} |S_n(g)|\right] \leq \sqrt{n} A \varphi(\delta) \ , \tag{41}$$

where, for all $g \in \mathcal{G}$, $S_n(g) = \sum_{i=0}^{n-1} g(Z_i) - n\mathbb{E}_\star\left[g(Z_0)\right]$.

Proof. This is a direct application of the remark following [11, Theorem 3]. □

B Entropy of the class $\mathcal{G}_M$

Lemma B.1. *For any $p' \geq 1$, any $s' > b\ell/p'$ and any even integer d, provided that $d > s' + b\ell(1 - \frac{1}{p'})$, there exists a constant C such that*

$$\forall u > 0, \ H(u, ||\cdot||_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G}_M) \leq C \left(\frac{M^{\upsilon(s' + d + \frac{b\ell}{p'})}}{u^{2r}} \right)^{b\ell/s'} \ . \tag{42}$$

Proof. By [25, Lemma 7.26], for any densities of probability p_2 and p_1 on $\mathbb{R}^{b\ell}$,

$$||g_{p_2} - g_{p_1}||^{2r}_{\mathrm{L}^{2r}(\mathbb{P}_\star)} \leq C ||\sqrt{p_2} - \sqrt{p_1}||^2_{\mathrm{L}^2(\mathbb{R}^{b\ell})} \ .$$

Since $||\sqrt{p_2} - \sqrt{p_1}||^2_{\mathrm{L}^2(\mathbb{R}^{b\ell})} \leq ||p_2 - p_1||_{\mathrm{L}^1(\mathbb{R}^{b\ell})}$, this yields, for any $u > 0$,

$$H(u, ||\cdot||_{\mathrm{L}^{2r}(\mathbb{P}_\star)}, \mathcal{G}_M) \leq H\left(\frac{u^{2r}}{C}, ||\cdot||_{\mathrm{L}^1(\mathbb{R}^{b\ell})}, \mathcal{P}_M \right) \ , \tag{43}$$

where $\mathcal{P}_M \stackrel{\text{def}}{=} \{p_{f,\nu};\ \nu \in \mathcal{D}_b,\ f \in \mathcal{F} \text{ and } I(f) \leq M\}$. Thus, it remains to bound the entropy with bracketing of the class of functions $\mathcal{P}_M$ associated with $||\cdot||_{\mathrm{L}^1(\mathbb{R}^{b\ell})}$ to control the entropy with bracketing of the class of functions $\mathcal{G}_M$ associated with $||\cdot||_{\mathrm{L}^{2r}(\mathbb{P}_\star)}$.

Define for any $p' \geq 1$ and $s' \geq 0$, the Sobolev space on $\mathbb{R}^{b\ell}$

$$W^{s',p'}\left(\mathbb{R}^{b\ell},\mathbb{R}\right) \stackrel{\text{def}}{=} \left\{h : \mathbb{R}^{b\ell} \to \mathbb{R};\ \ D^\alpha h \in \mathrm{L}^{p'}, \alpha \in \mathbb{N}^{b\ell} \text{ and } 0 \leq |\alpha| \leq s'\right\}$$

Define the polynomial weighting function $\langle \mathbf{y}\rangle^d \stackrel{\text{def}}{=} \left(1+\|\mathbf{y}\|^2\right)^{d/2}$ parametrized by d where $\mathbf{y} \in \mathbb{R}^{b\ell}$. Furthermore, define the weighted Sobolev space

$$W^{s',p'}\left(\mathbb{R}^{b\ell},\langle \mathbf{y}\rangle^d\right) \stackrel{\text{def}}{=} \left\{h;\ \ h\cdot\langle \mathbf{y}\rangle^d \in W^{s',p'}\left(\mathbb{R}^{b\ell},\mathbb{R}\right)\right\}\ .$$

Lemma B.2 ensures that, for any $p' \geq 1$, $s' > b\ell/p'$ any even integer d, the renormalized classes of functions $\mathcal{P}_M/M^{\upsilon(s'+d+\frac{b\ell}{p'})}$, $M \geq 1$ belong to the same bounded subspace of $W^{s',p'}(\mathbb{R}^{b\ell},\langle \mathbf{y}\rangle^d)$. By [27, Corollary 4], for any $p' \geq 1$, and any $s' > b\ell/p'$, provided that $d > s' + b\ell(1-\frac{1}{p'})$, there exists a constant C such that

$$\forall \epsilon > 0,\ H\left(\epsilon, \|\cdot\|_{\mathrm{L}^1(\mathbb{R}^{b\ell})}, \mathcal{P}_M/M^{\upsilon(s'+d+\frac{b\ell}{p'})}\right) \leq C\epsilon^{-b\ell/s'}\ .$$

The proof is concluded by (43). □

Lemma B.2. *Assume that H2 holds for some $\upsilon > 0$. Then, for any $p' \geq 1$, $s' > b\ell/p'$ and any even and positive number d, there exists a positive constant C such that for any $f \in \mathcal{F}$ and any $\nu \in \mathcal{D}_b$,*

$$\|p_{f,\nu}\cdot\langle \mathbf{y}\rangle^d\|_{W^{s',p'}(\mathbb{R}^{b\ell},\mathbb{R})} \leq C\kappa(\upsilon,f)^{s'+d+\frac{b\ell}{p'}}\ ,$$

where $\kappa(\upsilon,f) \stackrel{\text{def}}{=} 1 \vee I(f)^\upsilon$.

Proof. Let f be a function in $\mathcal{F}$, for any $\nu \in \mathcal{D}_b$,

$$\|p_{f,\nu}\cdot\langle \mathbf{y}\rangle^d\|^{p'}_{W^{s',p'}(\mathbb{R}^{b\ell},\mathbb{R})} = \sum_{|\alpha|\leq s'} \|D^\alpha\left(p_{f,\nu}\cdot\langle \mathbf{y}\rangle^d\right)\|^{p'}_{\mathrm{L}^{p'}}\ .$$

Applying the general Leibniz rule component by component, for any $\alpha \in \mathbb{N}^{b\ell}$,

$$D^\alpha\left(p_{f,\nu}\cdot\langle \mathbf{y}\rangle^d\right) = \sum_{\alpha'\leq\alpha}\binom{\alpha}{\alpha'} D^{\alpha'}(\langle \mathbf{y}\rangle^d) D^{\alpha-\alpha'}(p_{f,\nu})\ , \tag{44}$$

where $\binom{\alpha}{\alpha'} \stackrel{\text{def}}{=} \prod_{j=1}^{b\ell}\binom{\alpha_j}{\alpha'_j}$. Then, Lemma B.2 requires to control $\|D^{\alpha^{(1)}}(\langle \mathbf{y}\rangle^b) D^{\alpha^{(2)}}(p_{f,\nu})\|_{\mathrm{L}^{p'}}$ for any given $\alpha^{(1)}$ and $\alpha^{(2)}$ in $\mathbb{N}^{b\ell}$. For any α in $\mathbb{N}^{b\ell}$, there exists a polynomial function P_α with degree lower than $|\alpha|$ such that, for any $\mathbf{y} \in \mathbb{R}^{b\ell}$,

$$D^\alpha p_{f,\nu}(\mathbf{y}) = \int_{\mathbf{x}\in\mathbb{X}^b} P_\alpha(\mathbf{f}(\mathbf{x})-\mathbf{y}) \exp\left\{-\frac{1}{2}\|\mathbf{f}(\mathbf{x})-\mathbf{y}\|^2\right\} q_{a,b}(\mathbf{x})\mu^{\otimes b}(\mathrm{d}\mathbf{x})\ . \tag{45}$$

Moreover, since d is an even number, for any $\alpha \in \mathbb{N}^{b\ell}$ such that $|\alpha| \leq d$, $D^\alpha\langle \mathbf{y}\rangle^d$ is a polynomial function denoted by $P_{d,\alpha}$ with degree lower than $d-|\alpha|$. In the case where $|\alpha| > d$, $D^\alpha\langle \mathbf{y}\rangle^d = 0$.

By H2, there exists constant $C > 0$ such that, for any $\mathbf{x} \in \mathbb{X}^b$, $\|\mathbf{f}(\mathbf{x})\| \leq CI(f)^\upsilon \leq C\kappa(\upsilon,f)$. Since $P_{\alpha^{(2)}}$ and $P_{d,\alpha^{(1)}}$ are both polynomial functions, there exist a constant C depending on $\alpha^{(1)},\alpha^{(2)}$ and d such that, for any $\mathbf{y} \in \mathbb{R}^{b\ell}$ and any $\mathbf{x} \in \mathbb{X}^b$,

$$\left|P_{d,\alpha^{(1)}}(\mathbf{y})P_{\alpha^{(2)}}(\mathbf{f}(\mathbf{x})-\mathbf{y})\right| \leq \mathbf{1}_{|\alpha^{(1)}|\leq d}\left[C(1+\|\mathbf{y}\|)^{d-|\alpha^{(1)}|}\times(\kappa(\upsilon,f)+\|\mathbf{y}\|)^{|\alpha^{(2)}|}\right]\ .$$

Define the following subset of $\mathbb{R}^{b\ell}$

$$A_f \stackrel{\text{def}}{=} \left\{\mathbf{y} \in \mathbb{R}^{b\ell};\ \|\mathbf{y}\| \leq \kappa(\upsilon,f)\right\}\ .$$

$\|\mathbf{f}(\mathbf{x})-\mathbf{y}\|$ can be lower bounded by 0 when $\mathbf{y}$ belongs to A_f and by $|\kappa(\upsilon,f)-\|\mathbf{y}\||$ when $\mathbf{y}$ belongs to A_f^c. Therefore, uniformly in $\mathbf{x}\in\mathbb{X}^b$,

$$\exp\left\{-\frac{1}{2}\|\mathbf{f}(\mathbf{x})-\mathbf{y}\|^2\right\} \leq \mathbf{1}_{A_f}(\mathbf{y}) + \mathbf{1}_{A_f^c}(\mathbf{y}) e^{-\frac{1}{2}(\kappa(\upsilon,f)-\|y\|)^2} .$$

Thus, there exists a constant $C>0$ which dies not depend on a, such that,
for any $p'\geq 1$,

$$\|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^d) D^{\alpha^{(2)}}(p_{f,\nu})\|_{\mathrm{L}_{p'}}^{p'} \leq \mathbf{1}_{|\alpha^{(1)}|\leq d}\left[C\kappa(\upsilon,f)^{p'|\alpha^{(2)}|}\left(I_1+I_2\right)\right] ,$$

where,

$$I_1 \stackrel{\mathrm{def}}{=} \int_{A_f} (1+\|\mathbf{y}\|)^{p'(d-|\alpha^{(1)}|)} \left(1+\frac{\|\mathbf{y}\|}{\kappa(\upsilon,f)}\right)^{p'|\alpha^{(2)}|} \lambda^{\otimes b}(\mathrm{d}\mathbf{y}) ,$$

$$I_2 \stackrel{\mathrm{def}}{=} \int_{A_f^c} (1+\|\mathbf{y}\|)^{p'(d-|\alpha^{(1)}|)} \left(1+\frac{\|\mathbf{y}\|}{\kappa(\upsilon,f)}\right)^{p'|\alpha^{(2)}|} \mathrm{e}^{-\frac{p'}{2}(\kappa(\upsilon,f)-\|y\|)^2} \lambda^{\otimes b}(\mathrm{d}\mathbf{y}) .$$

By the change of variables $\mathbf{y}'=(\kappa(\upsilon,f))^{-1}\mathbf{y}$ in I_1 and I_2, there exists a constant C such that

$$\|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^d) D^{\alpha^{(2)}}(p_{f,\nu})\|_{\mathrm{L}_{p'}}^{p'} \leq C\kappa(\upsilon,f)^{p'(|\alpha^{(2)}|-|\alpha^{(1)}|+d)+b\ell} . \tag{46}$$

Using (46) in (44) with $\alpha^{(1)}=\alpha'$ and $\alpha^{(2)}=\alpha-\alpha'$ for any $|\alpha|\leq s'$ and $\alpha'\leq\alpha$ concludes the proof of Lemma B.2. □

References

[1] Adams, R. A.R. A. Fournier, J. J. F.J. J. F. (2003). Sobolev Spaces. Pure and Applied Mathematics vol. 140. Academic Press.

[2] Ambrosetti, A.A. Prodi, G.G. (1995). A primer of nonlinear analysis 34. Cambridge University Press.

[3] Carroll, R. J.R. J. Hall, P.P. (1988). Optimal rates of convergence for deconvolving a density. J. Amer. Statist. Assoc. 1184-1186.

[4] Churchill, G.G. (1992). Hidden Markov Chains and the Analysis of Genome Structure. Computers & Chemistry 16 107–115.

[5] Comte, F.F. Lacour, C.C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 601–627. 10.1111/j.1467-9868.2011.00775.x

[6] Comte, F.F. Taupin, M. L.M. L. (2007). Nonparametric estimation of the regression function in an errors-in-variables model. Statistica sinica 17 1065–1090.

[7] De Boor, C.C. Lynch, R. E.R. E. (1966). On splines and their minimum properties. J. Math. Mech 15 953–969.

[8] Dedecker, J.J., Doukhan, P.P., Lang, G.G., León, J. R.J. R., Louhichi, S.S. Prieur, C.C. (2009). Weak dependence: with examples and applications. (Lecture notes in statistics). AStA Advances in Statistical Analysis 93 119-120. 10.1007/s10182-008-0102-1

[9] Dempster, A. P.A. P., Laird, N. M.N. M. Rubin, D. B.D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39 1–38 (with discussion).

[10] Douc, R.R., Moulines, E.E. Stoffer, D. S.D. S. (2014). Nonlinear time series. Theory, methods and applications with R examples. CRC Press.

[11] Doukhan, P.P., Massart, P.P. Rio, E.E. (1995). Invariance principle for absolutely regular processes. Annales de l'Institut Henri Poincaré 31 393–427.

[12] Dumont, T.T. Le Corff, S.S. (2014). Simultaneous localization and mapping problem in wireless sensor networks. Signal Processing 101 192–203.

[13] Evans, L. C.L. C. Gariepy, R. F.R. F. (1992). Measure Theory and Fine Properties of Functions. Studies in Advanced Mathematics. CRC Press.

[14] Fan, J.J. Truong, Y. K.Y. K. (1993). Nonparametric regression with errors in variables. Ann. Statist. 21 1900–1925.

[15] Gassiat, E.E. van Handel, R.R. (2013). Consistent Order Estimation and Minimal Penalties. IEEE Transactions on Information Theory 59 1115-1128. 10.1109/TIT.2012.2221122

[16] Hastie, T. J.T. J. Tibshirani, R. J.R. J. (1990). Generalized additive models 43. CRC Press.

[17] Ioannides, D. A.D. A. Alevizos, P. D.P. D. (1997). Nonparametric regression with errors in variables and applications. Statistics and Probability Letters 32 35–43.

[18] Juang, B.B. Rabiner, L.L. (1991). Hidden Markov Models for Speech Recognition. Technometrics 33 251–272.

[19] Koo, J. Y.J. Y. (1999). Logspline Deconvolution in Besov Space. Scandinavian Journal of Statistics 26 73–86. 10.1111/1467-9469.00138

[20] Koo, J. Y.J. Y. Lee, K. W.K. W. (1998). B-spline estimation of regression functions with errors in variable. Statistics and Probability Letters 40 57–66.

[21] Lacour, C.C. (2006). Rates of convergence for nonparametric deconvolution. Comptes Rendus Mathematique 342 877–882.

[22] Lacour, C.C. (2008a). Least squares type estimation of the transition density of a particular hidden Markov model. Electron. J. Stat. 2 1–39.

[23] Lacour, C.C. (2008b). Adaptive estimation of the transition density of a particular hidden Markov chain. Journal of Multivariate Analysis 99 787–814.

[24] Mamon, R. S.R. S. Elliott, R. J.R. J. (2007). Hidden Markov models in finance. International series in operations research & management science 104. Springer.

[25] Massart, P.P. (2007). Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003. Ecole d'Eté de Probabilités de Saint-Flour vol. 1896. Springer-Verlag.

[26] Meyn, S. P.S. P. Tweedie, R. L.R. L. (1993). Markov Chains and Stochastic Stability. Communications and control engineering. Cambridge University Press.

[27] Nickel, R.R. Pötscher, B. M.B. M. (2001). Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov and Sobolev type. J. Theor. Probab. 20 177–199.

[28] Samson, P. M.P. M. (2000). Concentration of measure inequalities for Markov chains and ϕ-mixing processes. Ann. Statist. 28 416-461.

[29] Van De Geer, S. A.S. A. (2009). Empirical Processes in M-Estimation. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[30] Van Der Vaart, W.W. Wellner, A.A. (1996). Weak convergence and empirical processes. Springer.

[31] Whitney, H.H. (1986). Differentiable Manifolds. Annals of Mathematics 37 645-680.