Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models

Yohann De Castro* Élisabeth Gassiat* Sylvain Le Corff*

April 18, 2017

Abstract

In this paper, we consider the filtering and smoothing recursions in nonparametric finite state space hidden Markov models (HMMs) when the parameters of the model are unknown and replaced by estimators. We provide an explicit and time uniform control of the filtering and smoothing errors in total variation norm as a function of the parameter estimation errors. We prove that the risk for the filtering and smoothing errors may be uniformly upper bounded by the L^1-risk of the estimators. It has been proved very recently that statistical inference for finite state space nonparametric HMMs is possible. We study how the recent spectral methods developed in the parametric setting may be extended to the nonparametric framework and we give explicit upper bounds for the L^2-risk of the nonparametric spectral estimators. In the case where the observation space is compact, this provides explicit rates for the filtering and smoothing errors in total variation norm. The performance of the spectral method is assessed with simulated data for both the estimation of the (nonparametric) conditional distribution of the observations and the estimation of the marginal smoothing distributions.

Keywords: Hidden Markov Models; Nonparametric estimation; Filtering and Smoothing; Spectral methods.

1 Introduction

Hidden Markov models are popular dynamical models applied in a variety of applications such as economics, genomics, signal processing and image analysis, ecology, environment, speech recognition, see [14] for a recent overview of HMMs. Finite state space HMMs are stochastic processes $(X_j, Y_j)_{j\geq 1}$ such that $(X_j)_{j\geq 1}$ is a Markov chain with finite state space $\mathcal{X}$ and $(Y_j)_{j\geq 1}$ are random variables with general state space $\mathcal{Y}$, independent conditionally on $(X_j)_{j\geq 1}$ and such that for all $\ell \geq 1$, the conditional distribution of Y_ℓ given $(X_j)_{j\geq 1}$ depends on X_ℓ only. The state sequence $X_{1:n} := (X_1, \cdots, X_n)$ is only partially observed through the observations $Y_{1:n} := (Y_1, \cdots, Y_n)$. The parameters of the model are the initial distribution $\pi^\star$ of the hidden chain, the transition matrix of the hidden chain $\mathbf{Q}_\star$ and the conditional distribution of Y_1 given $X_1 = x$ for all possible $x \in \mathcal{X}$ which are often called emission distributions. In many applications of finite state space HMMs (e.g. digital communication or speech recognition), it is of utmost importance to infer the sequence of hidden states. This inference task usually involves the computation of the posterior distribution of a set of hidden states $X_{k:k'}$, $1 \leq k \leq k' \leq n$, given the observations $Y_{1:s}$, $1 \leq s \leq n$. When the initial distribution of the hidden chain, its transition matrix and the conditional distribution of the observations are known, computing posterior distributions can be efficiently done using the forward-backward algorithm described in [6] and [33]. In this paper, we focus on the estimation of the filtering distributions $\mathbb{P}(X_k = x|Y_{1:k})$ and marginal smoothing distributions $\mathbb{P}(X_k = x|Y_{1:n})$ for all $1 \leq k < n$ when the parameters of the HMM are unknown and replaced by estimators. These approximations of the posterior distributions are for instance required to compute expectations of additive functionals of the hidden states given the whole set of observations $Y_{1:n}$ which appear in popular maximum likelihood inference procedures. In the case of large data sets, online variants of the Expectation Maximization (EM) algorithm which update parameter estimates as new observations are received have been proposed, [7, 8, 24]. The convergence of such online algorithms remains an open problem despite some empirical evidence highlighted in these papers. Alternatives based on the decomposition of the observations into non overlapping blocks with convergence results easier to prove have been proposed to overcome this difficulty, [27]. We believe that the results given in this paper could be useful to establish convergence properties of such online procedures

[1]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.

since they rely on the control of the smoothing error when the posterior distributions are computed with the current estimate of the parameter. The aim of our paper is twofold.

- The paper analyzes the propagation of the parameter estimation error to the estimation of filtering and smoothing distributions. Providing explicit bounds for filtering and smoothing distributions under modeling uncertainties (in our case when the parameters are replaced by estimators) is an important step for real world online learning applications, see for instance [37] for Simultaneous Localization and Mapping problems, [4] for target tracking problems or [34] for other applications in engineering, telecommunications... The ability to monitor and control such dynamic systems depends on the accuracy of the estimation of the true state of the process which may be obtained using filtering or smoothing distributions. Providing explicit bounds for filtering and smoothing errors allow to tune the algorithms to obtain a given accuracy for the parameter estimates and the required control of the posterior distributions to optimize state estimation. Although replacing parameters by their estimators to compute posterior distributions and infer the hidden states is usual in applications, there are very few theoretical results to support this practice regarding the accuracy of the estimated posterior distributions. We are only aware of [18] whose results are restricted to the filtering distribution in a parametric setting. When the parameters of the HMM are known, the forward-backward algorithm can be extended to general state space HMMs (or to finite state space HMMs when the cardinality of $\mathcal{X}$ is too large) using computational methods such as Sequential Monte Carlo methods (SMC), see [11, 15] for a review of these methods. In this context, the Forward Filtering Backward Smoothing [25, 23, 16] and Forward Filtering Backward Simulation [21] algorithms have been intensively studied, with the objective of quantifying the error made when the filtering and marginal smoothing distributions are replaced by their Monte Carlo approximations. These algorithms and some extensions have been analyzed theoretically recently, see for instance [12, 13, 17, 31]. SMC methods may also be used in algorithms when the parameters of the HMM are unknown to perform maximum likelihood parameter estimation, see [24] for on-line and off-line EM and gradient ascent based algorithms. Part of our analysis of the filtering and smoothing distributions is based on the same approach as in those papers and requires strong forgetting properties of HMMs.

- Then, the paper extends spectral methods to a nonparametric setting and provides an explicit control of the L^2-risk of the estimators. Such estimators may then be used in the computation of posterior distributions as surrogates for the true parameters and emission densities. The upper bounds obtained for the L^2-risk of the estimators are useful since asymptotic properties of estimators for finite state space HMMs have been mainly studied in the parametric case while nonparametric HMMs are used in a variety of applications with no theoretical results. Many statistical inference procedures have been proposed for nonparametric HMMs, see for instance [26] for the identification of climate states (wet and dry), [28] for automatic speech recognition, [40] for Markov chain Monte Carlo methods to identify mixtures of Dirichlet process with application to the analysis of genomic copy number variation. These nonparametric methods allow the identification of HMMs without providing any insight on their consistency or rate of convergence to establish their statistical efficiency. This is only very recently that theoretical results have been obtained for the inference of nonparametric HMM [10, 29], see also [20] for translation mixture models or [38] for Bayesian posterior consistency.

  In latent variable models such as HMMs, spectral methods are popular since they lead to algorithms that are not sensitive to a chosen initial estimate. Indeed, standard estimation methods for HMMs are based on the EM algorithm, which possesses intrinsic limitations that are hard to circumvent such as slow convergence and convergence to suboptimal local extrema. Extending spectral methods to nonparametric HMMs is thus very useful. In particular, they may be used to provide a preliminary estimator as starting point in a EM algorithm. They are also used in a refinement procedure proposed in [10]. To the best of our knowledge, the spectral method has not been extended nor studied yet in the nonparametric framework. We start from the works of Anandkumar, Hsu, Kakade and Zhang on spectral methods in the parametric setting. Their papers [22, 3] present an efficient algorithm for learning parametric HMMs or more generally finitely many linear functionals of the parameters of a HMM. Thus, it is possible to use spectral methods to estimate the projections of the emission distributions onto nested subspaces of increasing complexity. Our work brings a new quantitative insight on the tradeoff between sampling size and approximation complexity for spectral estimators. We provide a nonasymptotic precise upper bound of the risk for the variance term with respect to the number of observations and the complexity of the approximating subspace.

Section 2 provides an explicit control of the total variation filtering and smoothing errors as a function of the parameter estimation error, see Propositions 2.1 and 2.2. Application of these preliminary results to the parametric context are detailed in Theorem 2.3, and to the nonparametric context in Theorem 2.4 where it is proved that the uniform rate of convergence of the filtering and smoothing errors is driven by the L^1-risk

of the nonparametric estimator of the emission distributions. Section 3 describes how spectral methods can be extended to the nonparametric setting and provides a nonasymptotic control of the variance term in Theorem 3.1. This leads to the asymptotic behavior proved in Corollary 3.2, which may be invoked when spectral methods are used in the computation of posterior distributions, see Corollary 3.3. Finally, the results proved in the paper are illustrated in Section 4 with numerical experiments. It is shown in particular that when the number of observations increases, the errors on the filtering and marginal smoothing distributions remain bounded which illustrates our theoretical results. All detailed proofs are given in the appendices.

2 Main results

2.1 Notations and setting

In the sequel, it is assumed that the cardinality K of $\mathcal{X}$ is known (for ease of notation, $\mathcal{X}$ is set to be $\{1, \ldots, K\}$) and that $\mathcal{Y}$ is a subset of $\mathbb{R}^D$ for a positive integer D. $\mathcal{P}(\mathcal{X})$ denotes the space of probability measures on $\mathcal{X}$ and write $\mathcal{L}^D$ the Lebesgue measure on $\mathcal{Y}$. For all $n \geq 1$ and all $x \in \mathcal{X}$, the density of the conditional distribution of Y_n given $X_n = x$ with respect to $\mathcal{L}^D$ is written $f_x^\star$. Consider the following assumptions on the hidden chain.

[H1] a) The transition matrix $\mathbf{Q}_\star$ has full rank.

b) $\delta^\star := \min_{1 \leq i,j \leq K} \mathbf{Q}_\star(i,j) > 0$.

[H2] The initial distribution $\pi^\star := (\pi_1^\star, \ldots, \pi_K^\star)$ is the stationary distribution.

Remark 2.1. *Note that under* **[H1]-b)** *and* **[H2]**, *for all* $k \in \mathcal{X}$, $\pi_k^\star \geq \delta^\star > 0$.

Remark 2.2. *Assumptions* **[H1]-a)** *and* **[H2]** *appear in spectral methods, see for instance [3, 22], and in identification of HMMs, see for instance [1, 2, 19]. It is established in [20] that* **[H1]** *is sufficient to obtain identifiability of all parameters and of the number of states K in nonparametric finite translation mixtures from the joint distribution of two observations. In the special case where $K = 2$, the assumption is equivalent to require that X_1 and X_2 are not independent. [1] detailed the necessity of the full-rank assumption of $\mathbf{Q}_\star$ to identify the model when the emission densities are all distinct.*

For all $y \in \mathcal{Y}$, define $c_\star(y)$ by

$$c_\star(y) := \min_{x \in \mathcal{X}} \sum_{x' \in \mathcal{X}} \mathbf{Q}_\star(x, x') f_{x'}^\star(y) \,. \tag{1}$$

For all $y_{1:n} \in \mathcal{Y}^n$, the filtering distributions $\phi_k^\star(\cdot, y_{1:k})$ and marginal smoothing distributions $\phi_{k|n}^\star(\cdot, y_{1:n})$ may be computed explicitly for all $1 \leq k \leq n$ using the forward-backward algorithm of [6]. In the forward pass, the filtering distributions $\phi_k^\star$ are updated recursively using, for all $x \in \mathcal{X}$,

$$\phi_1^\star(x, y_1) := \frac{\pi^\star(x) f_x^\star(y_1)}{\sum_{x' \in \mathcal{X}} \pi^\star(x') f_{x'}^\star(y_1)} \quad \text{and} \quad \phi_k^\star(x, y_{1:k}) := \frac{\sum_{x' \in \mathcal{X}} \mathbf{Q}_\star(x', x) f_x^\star(y_k) \phi_{k-1}^\star(x', y_{1:k-1})}{\sum_{x', x'' \in \mathcal{X}} \mathbf{Q}_\star(x', x'') f_{x''}^\star(y_k) \phi_{k-1}^\star(x', y_{1:k-1})} \,. \tag{2}$$

Note that for all $1 \leq k \leq n$, $\phi_k^\star(x, Y_{1:k}) = \mathbb{P}(X_k = x | Y_{1:k})$. In the backward pass, the marginal smoothing distributions may be updated recursively using, for all $x \in \mathcal{X}$,

$$\phi_{n|n}^\star(x, y_{1:n}) := \phi_n^\star(x, y_{1:n}) \quad \text{and} \quad \phi_{k|n}^\star(x, y_{1:n}) := \sum_{x' \in \mathcal{X}} B_{\phi_k^\star(\cdot, y_{1:k})}^\star(x', x) \phi_{k+1|n}^\star(x', y_{1:n}) \,, \tag{3}$$

where, for all $u, v \in \mathcal{X}$ and all $1 \leq k \leq n$,

$$B_{\phi_k^\star(\cdot, y_{1:k})}^\star(u, v) := \frac{\mathbf{Q}_\star(v, u) \phi_k^\star(v, y_{1:k})}{\sum_{z \in \mathcal{X}} \mathbf{Q}_\star(z, u) \phi_k^\star(z, y_{1:k})} \,.$$

Note that for all $1 \leq k \leq n$, $\phi_{k|n}^\star(x, Y_{1:n}) = \mathbb{P}(X_k = x | Y_{1:n})$.

2.2 Preliminary results

In this paper, the parameters $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$ are unknown. Then, the recursive equations (2) and (3) may be applied replacing $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$ by some estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ to obtain approximations of the filtering and smoothing distributions. Using forgetting properties of the hidden chain, we are able to obtain an upper bound of the filtering errors and of the marginal smoothing errors involving only the estimation errors of $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$. These upper bounds are given in Propositions 2.1 and 2.2. Their proofs are postponed to Appendix A and B. Note that the upper bounds are given for any possible values $y_{1:k}$, $k \geq 1$, and may be applied to the set of observations associated with the target filtering and smoothing distributions, regardless of the set of observations used to estimate $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$. Let $\|\cdot\|_{\mathrm{tv}}$ be the total variation norm, $\|\cdot\|_2$ the Euclidian norm and $\|\cdot\|_F$ the Frobenius norm. For all $1 \leq k \leq n$, $\widehat{\phi}_k$ and $\widehat{\phi}_{k|n}$ denote the approximations of $\phi^\star_k$ and $\phi^\star_{k|n}$ obtained by replacing $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$ by the estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ in (2) and (3).

Proposition 2.1. *Assume* **[H1]-b)** *and* **[H2]** *hold. Then, for all $k \geq 1$ and all $y_{1:k} \in \mathcal{Y}^k$,*

$$\|\phi^\star_k(\cdot, y_{1:k}) - \widehat{\phi}_k(\cdot, y_{1:k})\|_{\mathrm{tv}} \leq C_\star \left(\rho_\star^{k-1} \|\pi^\star - \widehat{\pi}\|_2 / \delta^\star + \|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F / (\delta^\star(1-\rho_\star)) + \sum_{\ell=1}^{k} \rho_\star^{k-\ell} c_\star^{-1}(y_\ell) \max_{x \in \mathcal{X}} \left| f^\star_x(y_\ell) - \hat{f}_x(y_\ell) \right| \right) ,$$

where $\rho_\star := 1 - \delta^\star/(1-\delta^\star)$ and $C_\star := 4(1-\delta^\star)/\delta^\star$.

The control of the marginal smoothing distribution errors is given by the following result.

Proposition 2.2. *Assume* **[H1]-b)** *and* **[H2]** *hold. Then, for all $1 \leq k \leq n$ and all $y_{1:n} \in \mathcal{Y}^n$,*

$$\|\phi^\star_{k|n}(\cdot, y_{1:n}) - \widehat{\phi}_{k|n}(\cdot, y_{1:n})\|_{\mathrm{tv}} \leq C_\star \left(\rho_\star^{k-1} \|\pi^\star - \widehat{\pi}\|_2 / \delta^\star + [1/(1-\rho_\star) + 1/(1-\widehat{\rho})] \|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F / \delta^\star + \sum_{\ell=1}^{n} (\widehat{\rho} \vee \rho_\star)^{|\ell-k|} c_\star^{-1}(y_\ell) \max_{x \in \mathcal{X}} \left| f^\star_x(y_\ell) - \hat{f}_x(y_\ell) \right| \right) ,$$

where $\widehat{\delta} := \min_{x,x'} \widehat{\mathbf{Q}}(x,x')$ and $\widehat{\rho} := 1 - \widehat{\delta}/(1-\widehat{\delta})$.

2.3 Uniform consistency of the posterior distributions

Propositions 2.1 and 2.2 are preliminary results that can be used to understand how estimation errors on the parameters of the HMM propagate to the filtering and smoothing distributions. Assume that we are given a set of $p+n$ observations from the hidden Markov model driven by $\pi^\star$, $\mathbf{Q}_\star$ and $f^\star$. The first p observations are used to produce the estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ while filtering and smoothing are performed with the last n observations. In other words, the estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ are measurable functions of $Y_{1:p}$ and the objective is to estimate $\phi^\star_k(\cdot, Y_{p+1:p+k})$ and $\phi^\star_{k|n}(\cdot, Y_{p+k:p+n})$.

2.3.1 Parametric models

In the parametric case, the hidden Markov model depends on a parameter $\theta_\star$ which lies in a subset of $\mathbb{R}^q$ for a given $q \geq 1$. In this situation, $\theta_\star$ may be estimated by $\widehat{\theta} \in \mathbb{R}^q$ and we may write $\widehat{\pi} := \pi^{\widehat{\theta}}$, $\widehat{\mathbf{Q}} := Q_{\widehat{\theta}}$ and $\widehat{f} := f^{\widehat{\theta}}$. In the following, for any sequence of real random variables $(Z_n)_{n\geq 0}$ and any sequence $(a_n)_{n\geq 0}$ of positive real numbers, the notation $Z_n = O_{\mathbb{P}}(a_n)$ means that $(Z_n/a_n)_{n\geq 0}$ is bounded in probability i.e. for all $\varepsilon > 0$ there exists $M > 0$ such that for all $n \geq 0$, $\mathbb{P}(|Z_n|/a_n > M) < \varepsilon$.

Theorem 2.3. *Assume* **[H1]** *and* **[H2]** *hold. Assume also that for all $x, x' \in \mathcal{X}$, $\theta \mapsto Q_\theta(x,x')$ is continuously differentiable with a bounded derivative in the neighborhood of $\theta_\star$ and that for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$, $\theta \mapsto f^\theta_x(y)$ is continuously differentiable in the neighborhood of $\theta_\star$ and such that the norm of its gradient is upper bounded in this neighborhood by a function h_x such that $\int h_x(y) d\mathcal{L}^D(y) < +\infty$. Let $\widehat{\theta}$ be a consistent estimator of $\theta_\star$. Then for any $1 \leq k \leq n$,*

$$\|\phi^\star_k(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k(\cdot, Y_{p+1:p+k})\|_{\mathrm{tv}} = O_{\mathbb{P}}(\|\widehat{\theta} - \theta_\star\|_2)$$

and

$$\|\phi^\star_{k|n}(\cdot, Y_{p+1:p+n}) - \widehat{\phi}_{k|n}(\cdot, Y_{p+1:p+n})\|_{\mathrm{tv}} = O_{\mathbb{P}}(\|\widehat{\theta} - \theta_\star\|_2) .$$

The smoothness assumption in Theorem 2.3 is usual to study the asymptotic distribution of the maximum likelihood estimator in parametric HMMs. By Theorem 2.3, tight bounds on the uniform convergence rate of $\|\phi_k^\star(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k(\cdot, Y_{p+1:p+k})\|_{\mathrm{tv}}$ and of $\|\phi_{k|n}^\star(\cdot, Y_{p+1:p+n}) - \widehat{\phi}_{k|n}(\cdot, Y_{p+1:p+n})\|_{\mathrm{tv}}$ may be derived by controlling the estimation error $\|\widehat{\theta} - \theta_\star\|$. There exist several results on this error term depending on the algorithm used to obtain $\widehat{\theta}$. For instance, [36] provides explicit upper bounds for this error term in the case where $\widehat{\theta}$ is a recursive maximum likelihood estimator of $\theta_\star$, under additional assumptions on the model.

Proof. First, under **[H1]** and **[H2]**, the assumption on $\theta \mapsto Q_\theta(x, x')$ implies that $\theta \mapsto \pi_x^\theta$ is continuously differentiable with a bounded derivative in the neihgborhood of $\theta_\star$. Note also that $\sup_{k\geq 1} \rho_\star^{k-1} \leq 1$ and $\sup_{k\geq 1} \widehat{\rho}^{k-1} \leq 1$. Then, using a Taylor expansion the first two terms of the upper bound in Propositions 2.1 and 2.2 are $O_{\mathbb{P}}(\|\widehat{\theta} - \theta_\star\|_2)$. There just remains to control the last term for each of the upper bound in Propositions 2.1 and 2.2. Using a Taylor expansion, Cauchy-Schwarz inequality, and Proposition 2.1, for any $1 \leq k \leq n$,

$$\|\phi_k^\star(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k(\cdot, Y_{p+1:p+k})\|_{\mathrm{tv}} \leq O_{\mathbb{P}}(\|\widehat{\theta} - \theta_\star\|_2) + \|\widehat{\theta} - \theta_\star\|_2 \sum_{\ell=1}^{k} \rho_\star^{k-\ell} c_\star^{-1}(Y_{p+\ell}) \sum_{x\in\mathcal{X}} h_x(Y_{p+\ell})\,.$$

As the $(Y_j)_{j\geq 1}$ are stationary with distribution having probability density $\sum_{x\in\mathcal{X}} \pi_x^\star f_x^\star(y) \leq c_\star(y)/\delta^\star$, the random variable $\sum_{\ell=1}^{k} \rho_\star^{k-\ell} c_\star^{-1}(Y_{p+\ell}) \sum_{x\in\mathcal{X}} h_x(Y_{p+\ell})$ is nonnegative and has expectation upper bounded by

$$\frac{1}{\delta^\star} \sum_{\ell=1}^{k} \rho_\star^{k-\ell} \sum_{x\in\mathcal{X}} \int h_x(y) d\mathcal{L}^D(y) \leq \frac{1-\delta^\star}{(\delta^\star)^2} \sum_{x\in\mathcal{X}} \int h_x(y) d\mathcal{L}^D(y) < +\infty\,.$$

Thus, $\sum_{\ell=1}^{k} \rho_\star^{k-\ell} c_\star^{-1}(Y_{p+\ell}) \sum_{x\in\mathcal{X}} h_x(Y_{p+\ell}) = O_{\mathbb{P}}(1)$ which ends the proof of the first part of Theorem 2.3. The result for the smoothing distributions follows the same lines since, for some $\epsilon > 0$ such that $\rho_\star + \epsilon < 1$, the event $\{\widehat{\rho} \geq \rho_\star + \epsilon\}$ has probability tending to 0 as p tends to infinity when $\widehat{\theta}$ is a consistent estimator of $\theta_\star$. □

2.3.2 Nonparametric models

We first state a general theorem providing a control of the uniform consistency of the posterior distributions depending on the risk of the nonparametric estimators. This theorem also holds in the parametric context. However, the parametric literature usually focuses on the properties of the estimators distribution while nonparametric results mostly study the risk. It is known that hidden Markov model are identifiable up to permutations of the hidden states labels. Therefore, without loss of generality, the following results are stated indicating the prospective permutation of the states. Let $\mathcal{S}_K$ be the set of permutations of $\{1, \ldots, K\}$. If τ is a permutation, $\mathbb{P}_\tau$ denotes the permutation matrix associated with τ.

Theorem 2.4. *Assume* **[H1]-b)** *and* **[H2]** *hold. Then for all $n \geq 1$, for any permutation $\tau_p \in \mathcal{S}_K$,*

$$\sup_{1\leq k\leq n} \mathbb{E}\left[\|\phi_k^\star(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k^{\tau_p}(\cdot, Y_{p+1:p+k})\|_{\mathrm{tv}}\right] \leq \frac{C_\star}{(\delta^\star)^2} \left\{\mathbb{E}[\left\|\pi^\star - \mathbb{P}_{\tau_p}\widehat{\pi}_p\right\|_2] + \mathbb{E}[\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\widehat{\mathbf{Q}}_p\mathbb{P}_{\tau_p}^\top\|_F] + \sum_{x\in\mathcal{X}} \mathbb{E}[\|f_x^\star - \widehat{f}_{\tau_p(x)}\|_1]\right\}$$

and

$$\sup_{1\leq k\leq n} \mathbb{E}\left[\|\phi_{k|n}^\star(\cdot, Y_{p+1:p+n}) - \widehat{\phi}_{k|n}^{\tau_p}(\cdot, Y_{p+1:p+n})\|_{\mathrm{tv}}\right] \leq \frac{C_\star}{(\delta^\star)^2} \left\{\mathbb{E}[\left\|\pi^\star - \mathbb{P}_{\tau_p}\widehat{\pi}_p\right\|_2] + \mathbb{E}\big[\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\widehat{\mathbf{Q}}_p\mathbb{P}_{\tau_p}^\top\|_F/\widehat{\delta}\big] + \sum_{x\in\mathcal{X}} \mathbb{E}\left[\|f_x^\star - \widehat{f}_{\tau_p(x)}\|_1/\widehat{\delta}\right]\right\}.$$

Here, $\widehat{\phi}_k^{\tau_p}$ and $\widehat{\phi}_{k|n}^{\tau_p}$ are the estimations of $\phi_k^\star$ and $\phi_{k|n}^\star$ based on $\mathbb{P}_{\tau_p}\widehat{\mathbf{Q}}\,\mathbb{P}_{\tau_p}^\top$, $\mathbb{P}_{\tau_p}\widehat{\pi}$ and $\widehat{f}_{\tau_p(x)}$, for all $x \in \mathcal{X}$.

The uniform control provided by Theorem 2.4 depends explicitly on the estimation errors of all the parameters and is a theoretical guarantee that posterior distributions may be approximated consistently in nonparametric

HMMs when parameters are unknown. This result has also practical consequences. For instance, in the case of online parameter estimation procedures, new parameter estimates are computed on-the-fly as new observations are received. This parameter estimate is computed using the approximation of the posterior distributions based on previous parameter estimates and Theorem 2.4 is therefore a first step to analyze the convergence properties of such algorithms (and it may also be used to tune algorithms to obtain a required accuracy on smoothed expectations approximations).

Theorem 2.4 provides a control driven by the L^1-risk of the emission densities. Section 3 introduces a spectral method to obtain, in the nonparametric context, estimators of the transition matrix, the stationary distribution and the emission densities. The algorithm is based on projection methods which leads to controls on the L^2-risk of the emission densities. This control may be easily transformed when $\mathcal{Y}$ is a compact subset of $\mathbb{R}^D$, since in such a case there exists $C(\mathcal{Y}) > 0$ such that for any square integrable functions h_1 and h_2,

$$\|h_1 - h_2\|_1 \leq C(\mathcal{Y})\|h_1 - h_2\|_2. \tag{4}$$

Note also that very recently other methods have been proposed to control the risk of estimation procedure in nonparametric HMMs. In [10], the authors introduced a penalized least squares estimator and established an oracle inequality for the L^2-risk of the estimation of the law of three consecutive observations and a minimax rate of estimation for the emission densities. In [20], a nonparametric estimator of the unknown translated density is proposed in finite translation mixture models for which the authors proved asymptotic rates for the minimax L^1-risk.

Proof. For any $x \in \mathcal{X}$ and any $1 \leq \ell \leq n$,

$$\mathbb{E}\left[c_\star^{-1}(Y_{p+\ell})\left|f_x^\star(Y_{p+\ell}) - \hat{f}_{\tau_p(x)}(Y_{p+\ell})\right|\right] = \mathbb{E}\left[\mathbb{E}\left[c_\star^{-1}(Y_{p+\ell})\left|f_x^\star(Y_{p+\ell}) - \hat{f}_{\tau_p(x)}(Y_{p+\ell})\right|\middle|Y_{1:p+\ell-1}\right]\right],$$

with

$$\mathbb{E}\left[c_\star^{-1}(Y_{p+\ell})\left|f_x^\star(Y_{p+\ell}) - \hat{f}_{\tau_p(x)}(Y_{p+\ell})\right|\middle|Y_{1:p+\ell-1}\right] = \int \left|f_x^\star(z) - \hat{f}_{\tau_p(x)}(z)\right| c_\star^{-1}(z) g_\ell(z)\mathrm{d}z\,,$$

where $g_\ell(z) := \sum_{x_{\ell-1},x_\ell\in\mathcal{X}} \phi_{\ell-1}^\star(x_{\ell-1}, Y_{p+1:p+\ell-1})\mathbf{Q}_\star(x_{\ell-1},x_\ell) f_{x_\ell}^\star(z)$. By **[H1]-b)** and (1), $c_\star^{-1}(z)g_\ell(z) \leq (1-\delta^\star)/\delta^\star$ and

$$\mathbb{E}\left[c_\star^{-1}(Y_{p+\ell})\left|f_x^\star(Y_{p+\ell}) - \hat{f}_x(Y_{p+\ell})\right|\middle|Y_{1:p+\ell-1}\right] \leq (1-\delta^\star)\|f_x^\star - \hat{f}_{\tau_p(x)}\|_1/\delta^\star\,.$$

The result for the filtering distributions is then a consequence of the upper bound of Proposition 2.1. The proof for the smoothing distributions follows the same steps. □

3 Nonparametric spectral estimation of HMMs

3.1 Description of the spectral method

This section describes a tractable approach to get nonparametric estimators of the emission densities and the transition matrix. This procedure relies on the estimation of the projections of the emission laws onto nested subspaces of increasing complexity. This allows to illustrate the uniform consistency result provided in the previous section. Let $(M_r)_{r\geq 1}$ be an increasing sequence of integers and $(\mathfrak{P}_{M_r})_{r\geq 1}$ be a sequence of nested subspaces such that their union is dense in $\mathrm{L}^2(\mathcal{Y},\mathcal{L}^D)$. Let $\Phi_{M_r} := \{\varphi_1,\ldots,\varphi_{M_r}\}$ be an orthonormal basis of $\mathfrak{P}_{M_r}$. Note that for all $f \in \mathrm{L}^2(\mathcal{Y},\mathcal{L}^D)$,

$$\lim_{r\to\infty} \sum_{m=1}^{M_r} \langle f, \varphi_m\rangle \varphi_m = f \ \text{ in } \ \mathrm{L}^2(\mathcal{Y},\mathcal{L}^D)\,. \tag{5}$$

Note also that changing M_r may change all functions φ_r, $1 \leq m \leq M_r$ in the basis Φ_{M_r}, which will not be indicated in the notation for better clarity. We shall also drop the index r and write M instead of M_r. The following standard examples may be considered.

- **(Spline)** The space of piecewise polynomials of degree less than d_r based on the regular partition with p_r^D regular pieces on $\mathcal{Y}$. In this case, $M_r = (d_r+1)^D p_r^D$.

- **(Trig.)** The space of real trigonometric polynomials on $\mathcal{Y}$ with degree less than r. In this case, $M_r = (2r+1)^D$.

- **(Wav.)** A wavelet basis Φ_{M_r} of scale r on $\mathcal{Y}$, see [30]. In this case, $M_r = 2^{(r+1)D}$.

The functions $f^\star_{M,1}, \dots, f^\star_{M,K}$ denote the projections of the emission densities on the space $\mathfrak{P}_M$, that is, for all $x \in \mathcal{X}$,

$$f^\star_{M,x} = \sum_{m=1}^{M} \langle f^\star_x, \varphi_m \rangle \varphi_m.$$

Our approach follows the strategy described in [3] to get an estimate of the emission densities. However, the dependency on the dimension is of crucial importance in the nonparametric framework and it has not been addressed in [3]. Hence, we present in Theorem C.3 a new quantitative version of the work [3] that accounts for the dimension M. Moreover, the authors of [3] estimate the transition matrix $\mathbf{Q}_\star$ but they do not give any theoretical guarantees regarding this estimator. In this paper, we introduce a slightly different estimator that is based on a surrogate $\tilde{\pi}$ (see Step 8 of Algorithm 1) of the stationary distribution. Our estimator (see Step 9 of Algorithm 1) is then built from the "observable" operator (rather than its left singular vectors as done in [3]). Eventually, Theorem C.2 provides theoretical guarantees on our estimator of the transition matrix and its stationary distribution.

The computation of those estimators is particularly simple: it is based on one singular value decomposition, matrix inversions and one diagonalization. It is proved in Theoremn C.2 and C.3 that, with overwhelming probability, all the matrix inversions and the diagonalization can be performed safely.

For all $(p \times q)$ matrices A with $p \geq q$, $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A) \geq 0$ denote the singular values of A and $\|\cdot\|$ its operator norm. When A is invertible, let $\kappa(A) := \sigma_1(A)/\sigma_q(A)$ be its condition number. $A^\top$ is the transpose matrix of A, $A(\ell, \ell')$ its (ℓ, ℓ')th entry, $A(.\,, \ell)$ its ℓth column and $A(k, .\,)$ its kth row. When A is a $(p \times p)$ diagonalizable matrix, its eigenvalues are written $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_p(A)$. For any $1 \leq q \leq +\infty$, $\|\cdot\|_q$ is the usual L^q norm for vectors. For any row or column vector v, $\mathfrak{Diag}[v]$ denotes the diagonal matrix with diagonal entries v_i. The following vectors, matrices and tensors are used throughout the paper:

- $\mathbf{L}_M \in \mathbb{R}^M$ is the projection of the distribution of one observation on the basis Φ_M: for all $a \in \{1, \dots, M\}$, $\mathbf{L}_M(a) := \mathbb{E}[\varphi_a(Y_1)]$;

- $\mathbf{N}_M \in \mathbb{R}^{M \times M}$ is the joint distribution of two consecutive observations: for all $(a, b) \in \{1, \dots, M\}^2$, $\mathbf{N}_M(a, b) := \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)]$;

- $\mathbf{M}_M \in \mathbb{R}^{M \times M \times M}$ is the joint distribution of three consecutive observations: for all $(a, b, c) \in \{1, \dots, M\}^3$, $\mathbf{M}_M(a, b, c) := \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)]$;

- $\mathbf{O}_M \in \mathbb{R}^{M \times K}$ is the conditional distribution of one observation on the basis Φ_M: for all $(m, x) \in \{1, \dots, M\} \times \mathcal{X}$, $\mathbf{O}_M(m, x) := \mathbb{E}[\varphi_m(Y_1) | X_1 = x] = \langle f^\star_x, \varphi_m \rangle$;

- For all $x \in \mathcal{X}$, $f^\star_{M,x}$ is the projection of the emission laws on the subspace $\mathfrak{P}_M$: , $f^\star_{M,x} := \sum_{m=1}^M \mathbf{O}_M(m, x)\varphi_m$. Write $\mathbf{f}^\star_M := (f^\star_{M,1}, \dots, f^\star_{M,K})$;

- $\mathbf{P}_M \in \mathbb{R}^{M \times M}$ is the joint distribution of (Y_1, Y_3): for all $(a, c) \in \{1, \dots, M\}^2$, $\mathbf{P}_M(a, c) := \mathbb{E}[\varphi_a(Y_1)\varphi_c(Y_3)]$.

3.2 Variance of the spectral estimators

This section displays results which allow to derive the asymptotic properties of the spectral estimators. The aim of Theorem 3.1 is to provide an explicit upper bound for the variance term with respect to both p and M. Assumption **[H3]**, together with **[H1]-b)** and **[H2]**, is sufficient to obtain identifiability of nonparametric HMMs. More precisely, [19] proved that if **[H3]**, **[H1]-b)** and **[H2]** hold, the model is identifiable from the distribution of 3 consecutive observations. [1] proved that it is enough to assume that the emission densities are all distinct to prove that the parameters may be identified. However, **[H3]** is a necessary condition to apply the spectral method to obtain the nonparametric estimators of the emission densities, see for instance Lemma C.1.

[H3] The family of emission densities $\mathfrak{F}^\star := \{f^\star_1, \dots, f^\star_K\}$ is linearly independent.

Algorithm 1: Nonparametric spectral estimation of the transition matrix and the emission laws

Data: An observed chain $(Y_1, \dots, Y_{p+2})$ and a number of hidden states K.
Result: Spectral estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $(\widehat{f}_{M,x})_{x\in\mathcal{X}}$.

[Step 1] For all a, b, c in $\{1, \dots, M\}$, consider the following empirical estimators:
$\widehat{\mathbf{L}}_M(a) := \sum_{s=1}^{p} \varphi_a(Y_s)/p$, $\widehat{\mathbf{M}}_M(a,b,c) := \sum_{s=1}^{p} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})/p$,
$\widehat{\mathbf{N}}_M(a,b) := \sum_{s=1}^{p} \varphi_a(Y_s)\varphi_b(Y_{s+1})/p$ and $\widehat{\mathbf{P}}_M(a,c) := \sum_{s=1}^{p} \varphi_a(Y_s)\varphi_c(Y_{s+2})/p$.

[Step 2] Let $\widehat{\mathbf{U}}$ be the $M \times K$ matrix of orthonormal right singular vectors of $\widehat{\mathbf{P}}_M$ corresponding to its top K singular values.

[Step 3] For all $b \in \{1, \dots, M\}$, set $\widehat{\mathbf{B}}(b) := (\widehat{\mathbf{U}}^\top \widehat{\mathbf{P}}_M \widehat{\mathbf{U}})^{-1} \widehat{\mathbf{U}}^\top \widehat{\mathbf{M}}_M(.\,, b, .\,)\widehat{\mathbf{U}}$.

[Step 4] Set Θ a $(K \times K)$ unitary matrix uniformly drawn and, for all $x \in \mathcal{X}$,
$\widehat{\mathbf{C}}(x) := \sum_{b=1}^{M} (\widehat{\mathbf{U}}\Theta)(b, x)\widehat{\mathbf{B}}(b)$.

[Step 5] Compute $\widehat{\mathbf{R}}$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\widehat{\mathbf{C}}(1)$:

$$\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{C}}(1)\widehat{\mathbf{R}} = \mathfrak{Diag}[(\widehat{\Lambda}(1,1), \dots, \widehat{\Lambda}(1,K))]\,.$$

[Step 6] For all $x, x' \in \mathcal{X}$, set $\widehat{\Lambda}(x, x') := (\widehat{\mathbf{R}}^{-1}\widehat{\mathbf{C}}(x)\widehat{\mathbf{R}})(x', x')$ and $\widehat{\mathbf{O}}_M := \widehat{\mathbf{U}}\Theta\widehat{\Lambda}$.

[Step 7] Consider the estimator $(\widehat{f}_{M,x})_{x\in\mathcal{X}}$ defined by, for all $x \in \mathcal{X}$, $\widehat{f}_{M,x} := \sum_{m=1}^{M} \widehat{\mathbf{O}}_M(m, x)\varphi_m$.

[Step 8] Set $\tilde{\pi} := \left(\widehat{\mathbf{U}}^\top \widehat{\mathbf{O}}_M\right)^{-1} \widehat{\mathbf{U}}^\top \widehat{\mathbf{L}}_M$.

[Step 9] Consider the transition matrix estimator $\widehat{\mathbf{Q}} := \Pi_{\mathrm{TM}}\Big(\left(\widehat{\mathbf{U}}^\top \widehat{\mathbf{O}}_M \mathfrak{Diag}[\tilde{\pi}]\right)^{-1} \widehat{\mathbf{U}}^\top \widehat{\mathbf{N}}_M \widehat{\mathbf{U}} \left(\widehat{\mathbf{O}}_M^\top \widehat{\mathbf{U}}\right)^{-1}\Big)$ where Π_{TM} denotes the projection (with respect to the scalar product given by the Frobenius norm) onto the convex set of transition matrices, and define $\widehat{\pi}$ as the stationary distribution of $\widehat{\mathbf{Q}}$.

Finally, the following quantity is required to control the L^2-risk of the spectral estimators. For any M, define

$$\eta_3^2(\Phi_M) := \sup_{y,y' \in \mathcal{Y}^3} \sum_{a,b,c=1}^{M} \left(\varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3) - \varphi_a(y_1')\varphi_b(y_2')\varphi_c(y_3')\right)^2 . \tag{6}$$

$\eta_3(\Phi_M)$ is the only term of the upper bound of the L^2-risk involving M.
In this section, assumption **[H1]** may be replaced by the following weaker assumption **[H1']**.

[H1'] a) The transition matrix $\mathbf{Q}_\star$ has full rank.
b) $(X_n)_{n\geq 1}$ is irreducible and aperiodic.

Note that under **[H1']** and **[H2]**, there exists $\pi^\star_{\min} > 0$ such that, for all $x \in \mathcal{X}$,

$$\pi^\star_x \geq \pi^\star_{\min}\,. \tag{7}$$

Theorem 3.1 (Spectral estimators)**.** *Assume that* **[H1']** *and* **[H2]**-**[H3]** *hold. Assume also that for all $x \in \mathcal{X}$, $f^\star_x \in \mathrm{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Then, there exist positive constants $u(\mathbf{Q}^\star)$, $\mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)$ and $\mathbf{N}(\mathbf{Q}^\star, \mathfrak{F}^\star)$ such that for any $u \geq u(\mathbf{Q}^\star)$, any $\delta \in (0,1)$, any $M \geq M_{\mathfrak{F}^\star}$, there exists a permutation $\tau_M \in \mathcal{S}_K$ such that the spectral method estimators $\widehat{f}_{M,x}$, $\widehat{\pi}$ and $\widehat{\mathbf{Q}}$ (see Algorithm 1) satisfy, for any $p \geq \mathbf{N}(\mathbf{Q}^\star, \mathfrak{F}^\star)\eta_3(\Phi_M)^2 u(-\log\delta)/\delta^2$, with probability greater than $1 - 2\delta - 4e^{-u}$,*

$$\max_{x\in\mathcal{X}} \|f^\star_{M,x} - \hat{f}_{M,\tau_M(x)}\|_2 \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star) \frac{\sqrt{-\log\delta}}{\delta} \frac{\eta_3(\Phi_M)}{\sqrt{p}} \sqrt{u}\,,$$

$$\|\pi^\star - \mathbb{P}_{\tau_M}\hat{\pi}\|_2 \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star) \frac{\sqrt{-\log\delta}}{\delta} \frac{\eta_3(\Phi_M)}{\sqrt{p}} \sqrt{u}\,,$$

$$\|\mathbf{Q}^\star - \mathbb{P}_{\tau_M}\hat{\mathbf{Q}}\mathbb{P}_{\tau_M}^\top\| \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)\frac{\sqrt{-\log\delta}}{\delta}\frac{\eta_3(\Phi_M)}{\sqrt{p}}\sqrt{u}\,.$$

Corollary 3.2. *Assume that* **[H1']** *and* **[H2]-[H3]** *hold. Assume also that for all $x \in \mathcal{X}$, $f_x^\star \in \mathrm{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Let M_p be a sequence of integers tending to infinity and such that $\eta_3(\Phi_{M_p}) = o(\sqrt{p/\log p})$. For each p, define $\widehat{f}$, $\widehat{\mathbf{Q}}$ and $\widehat{\pi}$ as the estimators obtained by the spectral algorithm with this choice of M_p. Then, there exists a sequence of permutations $\tau_p \in \mathcal{S}_K$ such that*

$$\mathbb{E}\big[\max_{x\in\mathcal{X}}\|f^\star_{M_p,x} - \hat{f}_{\tau_p(x)}\|_2\big] \vee \mathbb{E}\big[\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\hat{\mathbf{Q}}\mathbb{P}_{\tau_p}^\top\|\big] \vee \mathbb{E}\big[\|\pi^\star - \mathbb{P}_{\tau_p}\hat{\pi}\|_2\big] = O\big(\eta_3(\Phi_{M_p})\sqrt{\log p/p}\big) = o(1).$$

Here, the expectations are with respect to the observations and to the random unitary matrix drawn at **[Step 4]** *of Algorithm 1.*

Proof. Apply Theorem 3.1 where, for each p, we define δ_p such that $(-\log\delta_p)/\delta_p^2 := \log p$. δ_p goes to 0 and M_p goes to infinity as p tends to infinity so that for any large enough p, $M_p \geq M_{\mathfrak{F}^\star}$. Let τ_p the permutation τ_{M_p} given by Theorem 3.1. Then, for all $p/(\mathbf{N}(\mathbf{Q}^\star, \mathfrak{F}^\star)\eta_3(\Phi_{M_p})^2\log p) \geq u \geq u(\mathbf{Q}^\star)$, with probability $1 - 4e^{-u} - 2\delta_p$,

$$\max_{x\in\mathcal{X}}\|f^\star_{M,x} - \hat{f}_{M,\tau_M(x)}\|_2 \vee \|\pi^\star - \mathbb{P}_{\tau_p}\hat{\pi}\|_2 \vee \|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\hat{\mathbf{Q}}\mathbb{P}_{\tau_p}^\top\| \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)\eta_3(\Phi_{M_p})\sqrt{\log p/p}\sqrt{u}\,.$$

It yields

$$\begin{aligned}
&\limsup_{p\to+\infty}\mathbb{E}\left[\frac{p}{\eta_3(\Phi_{M_p})^2\log p}\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\hat{\mathbf{Q}}\mathbb{P}_{\tau_p}^\top\|^2\right] \\
&\qquad \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)^2\int_0^{+\infty}\limsup_{p\to+\infty}\mathbb{P}\left(\frac{\sqrt{p}}{\mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)\eta_3(\Phi_{M_p})\sqrt{\log p}}\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p}\hat{\mathbf{Q}}\mathbb{P}_{\tau_p}^\top\| \geq \sqrt{u}\right)\mathrm{d}u\,, \\
&\qquad \leq \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)^2 u(\mathbf{Q}^\star) + \mathcal{C}(\mathbf{Q}^\star, \mathfrak{F}^\star)^2\int_{x(\mathbf{Q}^\star)}^{+\infty} 4e^{-u}\mathrm{d}u < +\infty\,.
\end{aligned}$$

The proof is similar for the other terms. □

Applying Theorem 2.4 and (4) is enough to get the following corollary whose proof is omitted. The first point is an application of Corollary 3.2 and the second is obtained following the same lines as in the proof of Corollary 3.2.

Corollary 3.3. *Assume* **[H1]-[H3]** *hold. Assume also that for all $x \in \mathcal{X}$, $f_x^\star \in \mathrm{L}^2(\mathcal{Y}, \mathcal{L}^D)$. Let M_p be a sequence of integers tending to infinity such that $\eta_3(\Phi_{M_p}) = o(\sqrt{p/\log p})$. For each p, define $\widehat{f}$, $\widehat{\mathbf{Q}}$ and $\widehat{\pi}$ as the estimators obtained by the spectral algorithm given in Section 3 with this choice of M_p. Then, there exists a sequence of permutations $\tau_p \in \mathcal{S}_K$ such that*

$$\mathbb{E}\left[\sup_{k\geq 1}\|\phi_k^\star(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k^{\tau_p}(\cdot, Y_{p+1:p+k})\|_{\mathrm{tv}}\right] = O\big(\eta_3(\Phi_{M_p})\sqrt{\log p/p} + \sum_{x\in\mathcal{X}}\|f_x^\star - f^\star_{M_p,x}\|_2\big)$$

and

$$\mathbb{E}\left[\sup_{1\leq k\leq n}\|\phi^\star_{k|n}(\cdot, Y_{p+1:p+n}) - \widehat{\phi}^{\tau_p}_{k|n}(\cdot, Y_{p+1:p+n})\|_{\mathrm{tv}}\right] = O\big(\eta_3(\Phi_{M_p})\sqrt{\log p/p} + \sum_{x\in\mathcal{X}}\|f_x^\star - f^\star_{M_p,x}\|_2\big).$$

In **(Spline)**, **(Trig.)** and **(Wav.)**, there exists a constant $C_\eta > 0$ such that $\eta_3(M) \leq C_\eta M^{3/2}$, so that the uniform rate of convergence for the posterior probabilities is $O\big({M_p}^{3/2}\sqrt{\log p/p} + \sum_{x\in\mathcal{X}}\|f_x^\star - f^\star_{M_p,x}\|_2\big)$.

4 Experimental results

This section displays several numerical experiments to assess the efficiency of our method. The $K = 2$ emission laws are beta distributions with parameters $(2, 5)$ and $(4, 3)$. In all experiments, the transition matrix $\mathbf{Q}_\star$ is

$$\mathbf{Q}_\star := \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}$$

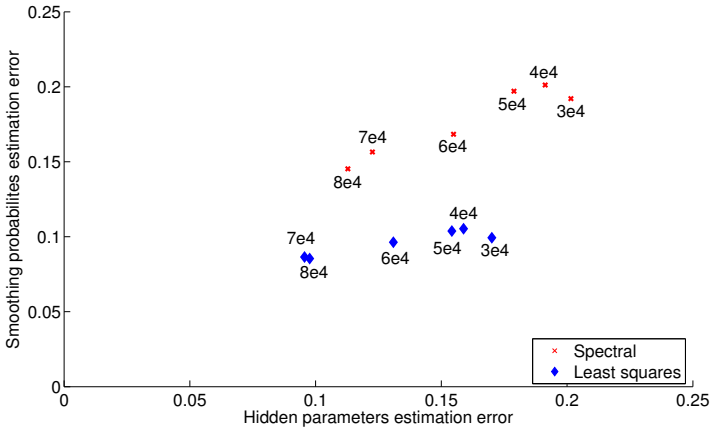

Figure 1: Illustration of Theorem 2.4: worst expected marginal smoothing probabilities obtained with the forward-backward algorithm ($n = 1e5$) combined with the spectral method or the least squares method using the projections of the emission laws on the histogram basis (left hand term in Theorem 2.4) as a function of the estimation error of the hidden parameters (right hand term in Theorem 2.4), for $p = 3e4, 4e4, 5e4, 6e4, 7e4$ and $8e4$.

and the estimation is based on the observation of a chain $(Y_i)_{i=1}^{p+n}$ of length $p + n$ with n varying from 1 to 100,000 and p varying from 30,000 to 80,000. We considered the histogram basis to build our approximation spaces, as defined in (5). The near minimax adaptive procedure described in [10]—referred to as the least-squares method—gives an estimation of $\mathbf{Q}_\star$ and of the emission laws. It is based on minimizing the empirical least squares in order to estimate the emission laws. Using the slope heuristic [5], the selected size of the model is $\hat{M}$ with $\hat{M} = 13, 14, 17, 19, 20$ and 22 for $p = 3e4, 4e4, 5e4, 6e4, 7e4$ and $8e4$ respectively. We use these values with the spectral method as well.

The Matlab codes can be found at My CoRe cloud

This section displays four numerical results:

1. The main goal is to illustrate Theorem 2.4. Expectations of the smoothing probabilities are computed taking the mean value over $\texttt{iter} = 20$ independent numerical experiments. Figure 1 displays the right hand side of Theorem 2.4—the worst expected marginal smoothing probability—as a function of the right hand term—the estimation error of the hidden parameters. It may illustrate an "at most linear" dependence between these two terms and that their ratio is bounded for small errors on the hidden parameters.

2. Figure 2 illustrates that the worst expectation of the error on the marginal smoothing probability does not explode when the chain length n goes to infinity. More precisely, the left hand side of Theorem 2.4 has been computed for $n = 1, \ldots, 100\,000$ based on an estimate of the hidden parameters built from the spectral method or the least squares method on a chain of length p varying from 30,000 to 80,000. This figure may illustrate that, for small estimation errors on the hidden parameters—large values of p, the error on the marginal smoothing probability is small and bounded whatever the chain length is.

3. Figure 3 presents a qualitative illustration of the adaptive estimation of the emission laws. Using a chain of length $p = 60{,}000$, the histogram and trigonometric bases are used as approximation spaces. Once again, the size $\hat{M}$ of the approximation space has been set using the least squares method together with the slope heuristic as in [10]. We found $\hat{M} = 19$ for the histogram basis and $\hat{M} = 18$ for the trigonometric basis. One may observe that the least squares method gives a better estimation than the spectral method.

4. Using these hidden parameter estimates, the marginal smoothing probabilities are computed using the forward-backward algorithm with a chain of length $n = 60,000$. The results are presented in Figure 4.

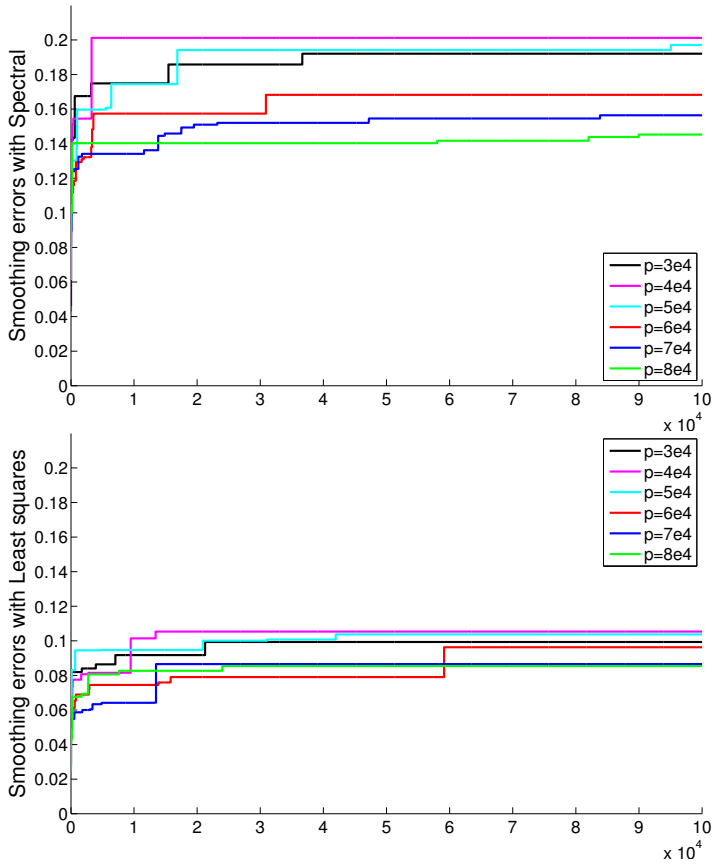

Figure 2: Illustration of Theorem 2.4: worst expected marginal smoothing probabilities obtained with the forward-backward algorithm combined with the spectral method or the least squares method using projection of the emission laws on the histogram basis (left hand term in Theorem 2.4) as a function of $n = 1, \ldots, 100\,000$.

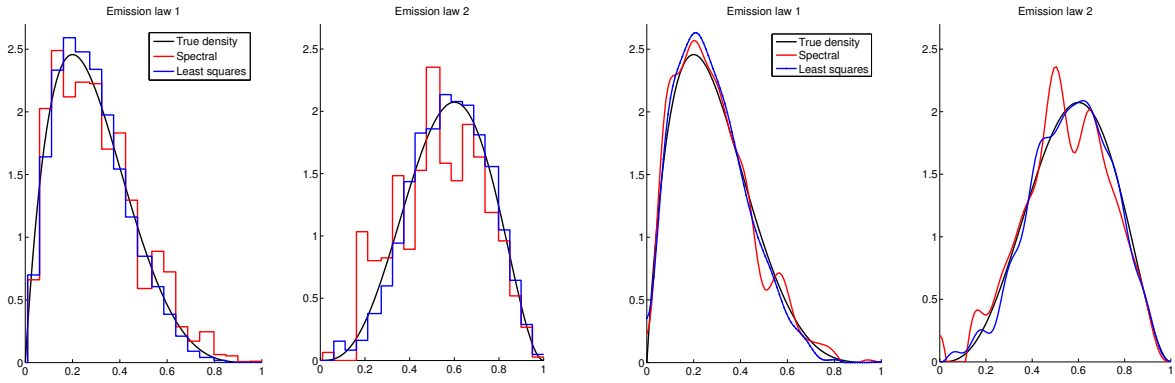

Figure 3: Estimation of beta distributions with parameters $(2, 5)$ and $(4, 3)$. The projection basis is the histogram basis ($\hat{M} = 19$) on the left and the trigonometric basis ($\hat{M} = 18$) on the right.

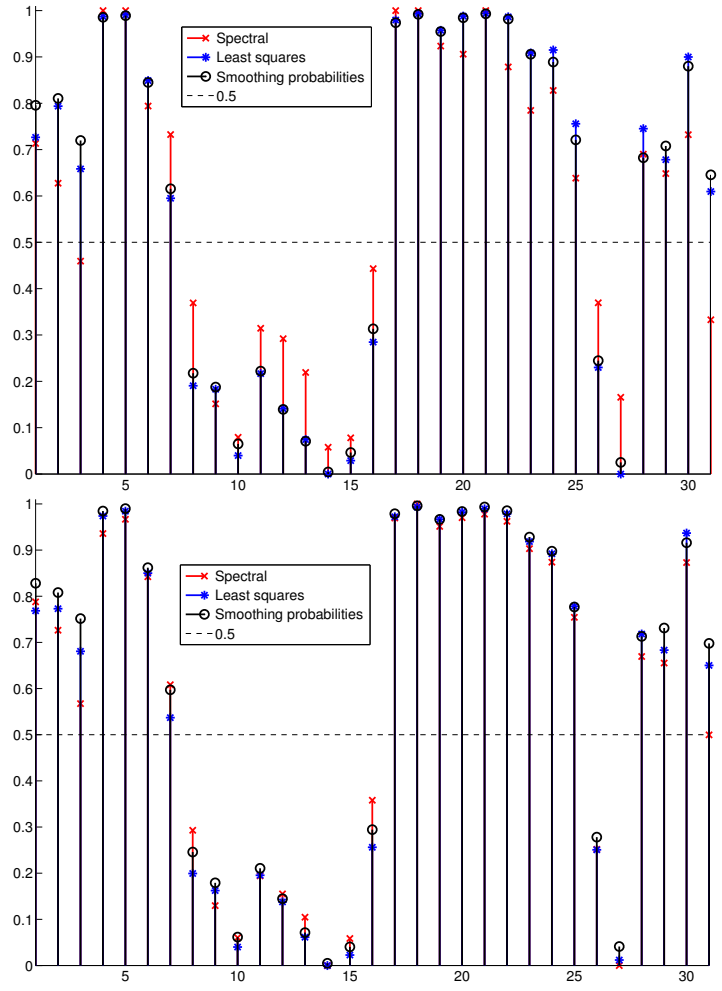

Figure 4: Marginal smoothing probabilities obtained with the forward-backward algorithm using projection of the emission laws on the histogram basis (top) or the trigonometric basis (bottom).

Conclusion and perspectives

This article focuses on the control of the estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models when the parameters are unknown. These posterior distributions are approximated using the forward-backward algorithm where parameters are replaced by any given estimators. This is the first time an explicit control of the worst expected filtering and marginal smoothing errors is established as a function of the L^1-risk of the hidden parameters. Numerical experiments assess this result by showing in particular that, for small errors on the hidden parameters, the error on the filtering and marginal smoothing distributions remains bounded when the number of observations grows.

In addition, this article introduces a new estimation procedure for nonparametric HMMs based on the spectral method and establishes upper bounds on its risk. As a byproduct of the spectral method, the algorithm does not suffer from convergence to a local minimum which leads to a reliable procedure to estimate the filtering and marginal smoothing distributions. From a computational view point, estimating the filtering and marginal smoothing requires a robust estimator of the hidden parameters and we believe that the spectral method can be efficiently used as such. Performance of this method relies heavily on the conditioning number of the empirical Gram matrix [29] of the emission densities and, hence, it requires a sufficiently large number of observations. These robustness issues are analyzed in a recent ongoing work, see [29] for a study of order estimation issues (i.e. selecting the number of hidden states) using the spectral method and the empirical least squares method. Also, interesting perspectives may include how to adapt these estimators to different regularities on the emission densities.

Appendix

A Control of the filtering error - Proof of Proposition 2.1

Let $y_{1:n} \in \mathcal{Y}^n$. The aim of this section is to establish that the total variation error between $\phi_k^\star(\cdot, y_{1:n})$ and its approximations based on $\widehat{\mathbf{Q}}$ and $\widehat{f}$ is bounded uniformly in time k. Before stating the main result, we introduce a standard decomposition of the filtering error $\phi_k^\star(\cdot, y_{1:k}) - \widehat{\phi}_k(\cdot, y_{1:k})$. For all $k \geq 1$, let $\mathrm{F}_{k,y_k}^\star$ be the forward kernel at time k and $\widehat{\mathrm{F}}_{k,y_k}$ its approximation, defined, for all $\nu \in \mathcal{P}(\mathcal{X})$, as:

$$\mathrm{F}_{k,y_k}^\star \nu(x) := \frac{\sum_{x' \in \mathcal{X}} \mathbf{Q}_\star(x', x) f_x^\star(y_k) \nu(x')}{\sum_{x', x'' \in \mathcal{X}} \mathbf{Q}_\star(x', x'') f_{x''}^\star(y_k) \nu(x')} ,$$

and

$$\widehat{\mathrm{F}}_{k,y_k} \nu(x) := \frac{\sum_{x' \in \mathcal{X}} \widehat{\mathbf{Q}}(x', x) \widehat{f}_x(y_k) \nu(x')}{\sum_{x', x'' \in \mathcal{X}} \widehat{\mathbf{Q}}(x', x'') \widehat{f}_{x''}(y_k) \nu(x')} .$$

Clearly, for all $y_{1:n} \in \mathcal{Y}^n$ and $2 \leq k \leq n$, $\phi_k^\star(\cdot, y_{1:k}) = \mathrm{F}_{k,y_k}^\star \phi_{k-1}^\star(\cdot, y_{1:k-1})$ and $\widehat{\phi}_k(\cdot, y_{1:k}) = \widehat{\mathrm{F}}_{k,y_k} \widehat{\phi}_{k-1}(\cdot, y_{1:k-1})$. The filtering error is usually written as a sum of one step errors. For all $k \geq 2$,

$$\begin{aligned} \phi_k^\star(\cdot, y_{1:k}) - \widehat{\phi}_k(\cdot, y_{1:k}) &= \mathrm{F}_{k,y_k}^\star \phi_{k-1}^\star(\cdot, y_{1:k-1}) - \widehat{\mathrm{F}}_{k,y_k} \widehat{\phi}_{k-1}(\cdot, y_{1:k-1}) \\ &= \sum_{\ell=1}^{k-1} \Delta_{k,\ell}(y_{\ell:k}) + \mathrm{F}_{k,y_k}^\star \widehat{\phi}_{k-1}(\cdot, y_{1:k-1}) - \widehat{\mathrm{F}}_{k,y_k} \widehat{\phi}_{k-1}(\cdot, y_{1:k-1}) , \end{aligned} \tag{8}$$

with $\mathrm{F}_{1,y_1}^\star \widehat{\phi}_0 = \phi_1^\star(\cdot, y_1)$ and

$$\Delta_{k,\ell}(y_{\ell:k}) := \mathrm{F}_{k,y_k}^\star \ldots \mathrm{F}_{\ell+1,y_{\ell+1}}^\star \mathrm{F}_{\ell,y_\ell}^\star \widehat{\phi}_{\ell-1}(\cdot, y_{1:\ell-1}) - \mathrm{F}_{k,y_k}^\star \ldots \mathrm{F}_{\ell+1,y_{\ell+1}}^\star \widehat{\phi}_\ell(\cdot, y_\ell) .$$

Let $\beta_{\ell|k}^\star[y_{\ell+1:k}]$ and $\mathrm{F}_{\ell|k}^\star[y_{\ell:k}]$ be the backward functions and the forward smoothing transition matrix as defined in [9, Chapter 3],

$$\beta_{\ell|k}^\star[y_{\ell+1:k}](x_\ell) := \sum_{x_{\ell+1:k}} \mathbf{Q}_\star(x_\ell, x_{\ell+1}) f_{x_{\ell+1}}^\star(y_{\ell+1}) \ldots \mathbf{Q}_\star(x_{k-1}, x_k) f_{x_k}^\star(y_k) , \tag{9}$$

$$\mathrm{F}_{\ell|k}^\star[y_{\ell:k}](x_{\ell-1}, x_\ell) := \frac{\beta_{\ell|k}^\star[y_{\ell+1:k}](x_\ell) \mathbf{Q}_\star(x_{\ell-1}, x_\ell) f_{x_\ell}^\star(y_\ell)}{\sum_{x \in \mathcal{X}} \beta_{\ell|k}^\star[y_{\ell+1:k}](x) \mathbf{Q}_\star(x_{\ell-1}, x) f_x^\star(y_\ell)} . \tag{10}$$

In the sequel, the dependency on the observations may be dropped to simplify notations. By [9, Chapter 4], for any probability distribution ν, $\mathrm{F}_k^\star \ldots \mathrm{F}_{\ell+1}^\star \nu = \nu_{\ell|k} \mathrm{F}_{\ell+1|k}^\star \ldots \mathrm{F}_{k|k}^\star$, where $\nu_{\ell|k} \propto \beta_{\ell|k}^\star \nu$. Therefore, the filtering error (8) is given by:

$$\phi_k^\star - \widehat{\phi}_k = \sum_{\ell=1}^{k-1} \left(\mu_{\ell|k}^\star \mathrm{F}_{\ell+1|k}^\star \ldots \mathrm{F}_{k|k}^\star - \widehat{\mu}_{\ell|k} \mathrm{F}_{\ell+1|k}^\star \ldots \mathrm{F}_{k|k}^\star \right) + \mathrm{F}_k^\star \widehat{\phi}_{k-1} - \widehat{\mathrm{F}}_k \widehat{\phi}_{k-1} , \tag{11}$$

where $\mu_{\ell|k}^\star \propto \beta_{\ell|k}^\star \mathrm{F}_\ell^\star \widehat{\phi}_{\ell-1}$ and $\widehat{\mu}_{\ell|k} \propto \beta_{\ell|k}^\star \widehat{\phi}_\ell$. By **[H1]-b)**, the transition matrix $\mathrm{F}_{k|n}^\star$ can be lower bounded uniformly in its first component:

$$\mathrm{F}_{\ell|k}^\star(x, x') \geq \frac{\delta^\star}{1 - \delta^\star} \frac{\beta_{\ell|k}^\star[y_{\ell+1:k}](x') f_{x'}^\star(y_\ell)}{\sum_{z \in \mathcal{X}} \beta_{\ell|k}^\star[y_{\ell+1:k}](z) f_z^\star(y_\ell)} .$$

By [9, Chapter 4], this allows to write,

$$\left\| \mu_{\ell|k}^\star \mathrm{F}_{\ell+1|k}^\star \ldots \mathrm{F}_{k|k}^\star - \widehat{\mu}_{\ell|k} \mathrm{F}_{\ell+1|k}^\star \ldots \mathrm{F}_{k|k}^\star \right\|_{\mathrm{tv}} \leq \rho_\star^{k-\ell} \| \mu_{\ell|k}^\star - \widehat{\mu}_{\ell|k} \|_{\mathrm{tv}} . \tag{12}$$

Eq. (12) is the crucial step to obtain the upper bound for the filtering error stated in Proposition 2.1. By (11) and (12),

$$\| \phi_k^\star - \widehat{\phi}_k \|_{\mathrm{tv}} \leq \sum_{\ell=1}^{k-1} \rho_\star^{k-\ell} \left\| \mu_{\ell|k}^\star - \widehat{\mu}_{\ell|k} \right\|_{\mathrm{tv}} + \left\| \mathrm{F}_k^\star \widehat{\phi}_{k-1} - \widehat{\mathrm{F}}_k \widehat{\phi}_{k-1} \right\|_{\mathrm{tv}} .$$

For all $1 \leq \ell \leq k-1$ and all bounded function h on $\mathcal{X}$, $\left|\mu^\star_{\ell|k}(h) - \widehat{\mu}_{\ell|k}(h)\right| \leq T_1 + T_2$ where

$$T_1 := \left| \frac{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) h(x) \left[\mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1}(x) - \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}(x)\right]}{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) \mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1}(x)} \right| ,$$

$$T_2 := \left| \frac{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) h(x) \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}(x)}{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}(x)} \right| \cdot \left| \frac{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) \left[\mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1}(x) - \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}(x)\right]}{\sum_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x) \mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1}(x)} \right| .$$

Both T_1 and T_2 are upper bounded by the same term so that

$$T_1 + T_2 \leq 2 \frac{\|h\|_\infty \cdot \|\beta^\star_{\ell|k}[y_{\ell+1:k}]\|_\infty}{\inf_{x\in\mathcal{X}} \beta^\star_{\ell|k}[y_{\ell+1:k}](x)} \|\mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1} - \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}\|_{\mathrm{tv}} .$$

By (9), for all $x \in \mathcal{X}$, $\beta^\star_{\ell|k}[y_{\ell+1:k}](x) \leq (1-\delta^\star) \sum_{x_{k+1:n}} f^\star_{x_{k+1}}(y_{k+1}) \ldots \mathbf{Q}_\star(x_{n-1}, x_n) f^\star_{x_n}(y_n)$ and $\beta^\star_{\ell|k}[y_{\ell+1:k}](x) \geq \delta^\star \sum_{x_{k+1:n}} f^\star_{x_{k+1}}(y_{k+1}) \ldots \mathbf{Q}_\star(x_{n-1}, x_n) f^\star_{x_n}(y_n)$, showing that

$$T_1 + T_2 \leq 2\|h\|_\infty \left(\frac{1-\delta^\star}{\delta^\star}\right) \|\mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1} - \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}\|_{\mathrm{tv}} .$$

Now, for all $2 \leq \ell \leq k$ and all bounded function h on $\mathcal{X}$, $\left|\mathrm{F}^\star_\ell \widehat{\phi}_{\ell-1}(h) - \widehat{\mathrm{F}}_\ell \widehat{\phi}_{\ell-1}(h)\right| \leq R_1 + R_2$, where

$$R_1 := \left| \frac{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \left[\mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) - \widehat{\mathbf{Q}}(x,x') \widehat{f}_{x'}(y_\ell)\right] h(x')}{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell)} \right| ,$$

$$R_2 := \left| \frac{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \widehat{\mathbf{Q}}(x,x') \widehat{f}_{x'}(y_\ell) h(x')}{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \widehat{\mathbf{Q}}(x,x') \widehat{f}_{x'}(y_\ell)} \right| \times \left| \frac{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \left[\mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) - \widehat{\mathbf{Q}}(x,x') \widehat{f}_{x'}(y_\ell)\right]}{\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell)} \right| .$$

Then,

$$\begin{aligned}
R_1 &\leq \left(\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) \right)^{-1} \sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \left| \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) - \widehat{\mathbf{Q}}(x,x') \widehat{f}_{x'}(y_\ell) \right| h(x') , \\
&\leq \left(\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) \right)^{-1} \sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \left| \mathbf{Q}_\star(x,x') - \widehat{\mathbf{Q}}(x,x') \right| f^\star_{x'}(y_\ell) h(x') \\
&\quad + \left(\sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \mathbf{Q}_\star(x,x') f^\star_{x'}(y_\ell) \right)^{-1} \sum_{x,x'\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x) \widehat{\mathbf{Q}}(x,x') \left| f^\star_{x'}(y_\ell) - \widehat{f}_{x'}(y_\ell) \right| h(x') , \\
&\leq \|h\|_\infty \left[\|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F / \delta^\star + c_\star^{-1}(y_\ell) \max_{x\in\mathcal{X}} \left| f^\star_x(y_\ell) - \hat{f}_x(y_\ell) \right| \right] ,
\end{aligned}$$

where $c_\star$ is defined in (1). The same upper bound holds for R_2. In the case $\ell = 1$,

$$\left\| \mathrm{F}^\star_1 \widehat{\phi}_0 - \widehat{\phi}_1 \right\|_{\mathrm{tv}} \leq \left\| \phi^\star_1 - \widehat{\phi}_1 \right\|_{\mathrm{tv}} \leq 2 \left[\|\pi^\star - \widehat{\pi}\|_2 / \delta^\star + c_\star^{-1}(y_1) \max_{x\in\mathcal{X}} \left| f^\star_x(y_1) - \hat{f}_x(y_1) \right| \right] .$$

Therefore, the filtering error is upper bounded as follows:

$$\begin{aligned}
\|\phi^\star_k - \widehat{\phi}_k\|_{\mathrm{tv}} \leq 4 \left(\frac{1-\delta^\star}{\delta^\star}\right) \sum_{\ell=2}^{k} \rho_\star^{k-\ell} \left[\|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F / \delta^\star + c_\star^{-1}(y_\ell) \max_{x\in\mathcal{X}} \left| f^\star_x(y_\ell) - \hat{f}_x(y_\ell) \right| \right] \\
+ 4 \left(\frac{1-\delta^\star}{\delta^\star}\right) \rho_\star^{k-1} \left[\|\pi^\star - \widehat{\pi}\|_2 / \delta^\star + c_\star^{-1}(y_1) \max_{x\in\mathcal{X}} \left| f^\star_x(y_1) - \hat{f}_x(y_1) \right| \right] .
\end{aligned}$$

B Control of the marginal smoothing error - Proof of Proposition 2.2

Let $y_{1:n} \in \mathcal{Y}^n$. The aim of this section is to establish that the total variation error between $\phi^\star_{k|n}(\cdot, y_{1:n})$ and its approximations based on $\widehat{\mathbf{Q}}$ and $\widehat{f}$ is bounded uniformly in time k. Before stating the main result, we display the decomposition of the smoothing error $\phi^\star_{k|n}(\cdot, y_{1:n}) - \widehat{\phi}_{k|n}(\cdot, y_{1:n})$ depicted in [13] and used in [17] to obtain nonasymptotic upper bounds for the marginal smoothing error when $\phi^\star_{k|n}(\cdot, y_{1:n})$ is approximated using Sequential Monte Carlo methods. In the sequel, the dependency on the observations may be dropped to simplify notations. For any bounded function h on $\mathcal{X}^n$, $\phi^\star_{1:n|n}(h)$ can be written, for any $1 \leq \ell \leq n$

$$\phi^\star_{1:n|n}(h) = \frac{\phi^\star_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, h))}{\phi^\star_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, \mathbb{1}))},$$

where $\mathbb{1}$ is the constant function which equals 1 and, for all $x_{1:\ell} \in \mathcal{X}^\ell$,

$$L^\star_{\ell,n}(x_{1:\ell}, h) := \sum_{x_{\ell+1:n} \in \mathcal{X}^{n-\ell}} \prod_{u=\ell+1}^{n} \mathbf{Q}_\star(x_{u-1}, x_u) f^\star_{x_u}(y_u) h(x_{1:n}). \tag{13}$$

As for the filtering error, the smoothing error can be decomposed as a telescopic sum of one step errors:

$$\begin{aligned}\widehat{\phi}_{1:n|n}(h) - \phi^\star_{1:n|n}(h) = \sum_{\ell=2}^{n} \left(\frac{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, h))}{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, \mathbb{1}))} - \frac{\widehat{\phi}_{1:\ell-1|\ell-1}(L^\star_{\ell-1,n}(\cdot, h))}{\widehat{\phi}_{1:\ell-1|\ell-1}(L^\star_{\ell-1,n}(\cdot, \mathbb{1}))} \right) \\ + \frac{\widehat{\phi}_1(L^\star_{1,n}(\cdot, h))}{\widehat{\phi}_1(L^\star_{1,n}(\cdot, \mathbb{1}))} - \frac{\phi^\star_1(L^\star_{1,n}(\cdot, h))}{\phi^\star_1(L^\star_{1,n}(\cdot, \mathbb{1}))}.\end{aligned} \tag{14}$$

This smoothing error can be written using filtering distributions only by introducing the following backward operators:

$$\begin{aligned}\mathcal{L}^\star_{\ell,n}(x_\ell, h) &:= \sum_{x_{1:\ell-1}} B^\star_{\phi^\star_{\ell-1}}(x_\ell, x_{\ell-1}) \dots B^\star_{\phi^\star_1}(x_2, x_1) L^\star_{\ell,n}(x_{1:\ell}, h), \\ \widehat{\mathcal{L}}_{\ell,n}(x_\ell, h) &:= \sum_{x_{1:\ell-1}} \widehat{B}_{\widehat{\phi}_{\ell-1}}(x_\ell, x_{\ell-1}) \dots \widehat{B}_{\widehat{\phi}_1}(x_2, x_1) L^\star_{\ell,n}(x_{1:\ell}, h),\end{aligned}$$

where for all $\nu \in \mathcal{P}(\mathcal{X})$, B_ν is the backward smoothing kernel given by

$$B^\star_\nu(x, x') := \frac{\mathbf{Q}_\star(x', x)\nu(x')}{\sum_{z \in \mathcal{X}} \mathbf{Q}_\star(z, x)\nu(z)}.$$

Then, for all $2 \leq t \leq n$, the one step error at time ℓ is given by

$$\delta_{\ell,n}(h) := \frac{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, h))}{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, \mathbb{1}))} - \frac{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, h))}{\widehat{\phi}_{1:\ell|\ell}(L^\star_{\ell,n}(\cdot, \mathbb{1}))} = \frac{\widehat{\phi}_\ell(\widehat{\mathcal{L}}_{\ell,n}(\cdot, h))}{\widehat{\phi}_\ell(\widehat{\mathcal{L}}_{\ell,n}(\cdot, \mathbb{1}))} - \frac{\widehat{\phi}_{\ell-1}(\widehat{\mathcal{L}}_{\ell-1,n}(\cdot, h))}{\widehat{\phi}_{\ell-1}(\widehat{\mathcal{L}}_{\ell-1,n}(\cdot, \mathbb{1}))}. \tag{15}$$

This decomposition allows to obtain the upper bound for the marginal smoothing error stated in Proposition 2.2. The result is obtained by applying the decompositions (14) and (15) to a bounded function h on $\mathcal{X}^n$ which depends on x_k only: for all $(x_1, \dots, x_n) \in \mathcal{X}^n$, $h(x_1, \dots, x_n) = h(x_k)$. The one step error given by (15) is then analyzed separately wether $k \geq \ell$ or $k < \ell$.

Case $k \geq \ell$

In this case, the function $L^\star_{\ell,n}(\cdot, h)$ defined in (13) depends on x_ℓ only. Therefore, $\widehat{\mathcal{L}}_{\ell,n}(x_\ell, h) = L^\star_{\ell,n}(x_\ell, h) = \mathcal{L}^\star_{\ell,n}(x_\ell, h)$. Thus, $\widehat{\mathcal{L}}_{\ell-1,n}(x_{\ell-1}, h) = \sum_{x_\ell \in \mathcal{X}} \mathbf{Q}_\star(x_{\ell-1}, x_\ell) f^\star_{x_\ell}(y_\ell) \mathcal{L}^\star_{\ell,n}(x_\ell, h)$ and the one step error given by (15) becomes

$$\delta_{\ell,n}(h) = \frac{\widehat{\phi}_\ell(\mathcal{L}^\star_{\ell,n}(\cdot, h))}{\widehat{\phi}_\ell(\mathcal{L}^\star_{\ell,n}(\cdot, \mathbb{1}))} - \frac{\widehat{\phi}_{\ell-1}(\sum_{x_\ell \in \mathcal{X}} \mathbf{Q}_\star(\cdot, x_\ell) f^\star_{x_\ell}(y_\ell) \mathcal{L}^\star_{\ell,n}(x_\ell, h))}{\widehat{\phi}_{\ell-1}(\sum_{x_\ell \in \mathcal{X}} \mathbf{Q}_\star(\cdot, x_\ell) f^\star_{x_\ell}(y_\ell) \mathcal{L}^\star_{\ell,n}(x_\ell, \mathbb{1}))}.$$

Define the measures μ_ℓ and $\widehat{\mu}_\ell$ on $\mathcal{X}$ by $\mu_\ell(x_\ell) := \sum_{x_{\ell-1}\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x_{\ell-1})\mathbf{Q}_\star(x_{\ell-1},x_\ell)f^\star_{x_\ell}(y_\ell)$ and $\widehat{\mu}_\ell(x_\ell) := \sum_{x_{\ell-1}\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x_{\ell-1})\widehat{\mathbf{Q}}(x_{\ell-1},x_\ell)\widehat{f}_{x_\ell}(y_\ell)$. Then,

$$\delta_{\ell,n}(h) = \frac{\widehat{\mu}_\ell(\mathcal{L}^\star_{\ell,n}(\cdot,h))}{\widehat{\mu}_\ell(\mathcal{L}^\star_{\ell,n}(\cdot,\mathbb{1}))} - \frac{\mu_\ell(\mathcal{L}^\star_{\ell,n}(\cdot,h))}{\mu_\ell(\mathcal{L}^\star_{\ell,n}(\cdot,\mathbb{1}))}\,.$$

By [9, Lemma 4.3.23] and **[H1]-b)**, $|\delta_{\ell,n}(h)| \le \rho_\star^{k-\ell}(1-\delta^\star)\,\|\mu_\ell/\mu_\ell(\mathbb{1}) - \widehat{\mu}_\ell/\widehat{\mu}_\ell(\mathbb{1})\|_{\mathrm{tv}}\,\|h\|_\infty/\delta^\star$. Following the same steps as for the proof of Proposition 2.1 yields

$$\|\mu_\ell/\mu_\ell(\mathbb{1}) - \widehat{\mu}_\ell/\widehat{\mu}_\ell(\mathbb{1})\|_{\mathrm{tv}} \le 2\|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F/\delta^\star + 2c_\star^{-1}(y_\ell)\max_{x\in\mathcal{X}}\left|f^\star_x(y_\ell) - \hat{f}_x(y_\ell)\right|\,.$$

The term $\widehat{\phi}_1(L^\star_{1,n}(\cdot,h))/\widehat{\phi}_1(L^\star_{1,n}(\cdot,\mathbb{1})) - \phi^\star_1(L^\star_{1,n}(\cdot,h))/\phi^\star_1(L^\star_{1,n}(\cdot,\mathbb{1}))$ is dealt with similarly.

Case $k < \ell$

In this case, $L^\star_{\ell,n}(x_{1:\ell},h) = h(x_k)\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})$. Therefore,

$$\begin{aligned}\widehat{\mathcal{L}}_{\ell,n}(x_\ell,h) &= \sum_{x_{1:\ell-1}} \widehat{B}_{\widehat{\phi}_{\ell-1}}(x_\ell,x_{\ell-1})\dots\widehat{B}_{\widehat{\phi}_1}(x_2,x_1)h(x_k)\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})\,,\\ &= \sum_{x_{k:\ell-1}} \mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})\widehat{B}_{\widehat{\phi}_{\ell-1}}(x_\ell,x_{\ell-1})\dots\widehat{B}_{\widehat{\phi}_k}(x_{k+1},x_k)h(x_k)\,.\end{aligned}$$

On the other hand, if $\nu_\ell(x_\ell) := \sum_{x_{\ell-1}\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x_{\ell-1})\mathbf{Q}_\star(x_{\ell-1},x_\ell)f^\star_{x_\ell}(y_\ell)\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})$,

$$\widehat{\phi}_{\ell-1}(\widehat{\mathcal{L}}_{\ell-1,n}(\cdot,h)) = \sum_{x_{k:\ell}\in\mathcal{X}^{\ell-k+1}} \nu_\ell(x_\ell)\widehat{B}_{\widehat{\phi}_{\ell-1}}(x_\ell,x_{\ell-1})\dots\widehat{B}_{\widehat{\phi}_k}(x_{k+1},x_k)h(x_k)\,.$$

Define $\widehat{\nu}_\ell(x_\ell) := \widehat{\phi}_\ell(x_\ell)\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1}) = \sum_{x_{\ell-1}\in\mathcal{X}} \widehat{\phi}_{\ell-1}(x_{\ell-1})\widehat{\mathbf{Q}}(x_{\ell-1},x_\ell)\widehat{f}_{x_\ell}(y_\ell)\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})$. Then, the one step error given by (15) becomes

$$\delta_{\ell,n}(h) = \sum_{x_{k:\ell-1}}\left(\frac{\widehat{\nu}_\ell(x_\ell)}{\widehat{\nu}_\ell(\mathbb{1})} - \frac{\nu_\ell(x_\ell)}{\nu_\ell(\mathbb{1})}\right)\widehat{B}_{\widehat{\phi}_{\ell-1}}(x_\ell,x_{\ell-1})\dots\widehat{B}_{\widehat{\phi}_k}(x_{k+1},x_k)h(x_k)$$

By [9, Lemma 4.3.23] and the fact that, for all $(x,x')\in\mathcal{X}^2$, $\widehat{\mathbf{Q}}(x,x') \ge \widehat{\delta}$,

$$|\delta_{\ell,n}(h)| \le \|h\|_\infty \widehat{\rho}^{\ell-k}\left\|\frac{\widehat{\nu}_\ell(\cdot)}{\widehat{\nu}_\ell(\mathbb{1})} - \frac{\nu_\ell(\cdot)}{\nu_\ell(\mathbb{1})}\right\|_{\mathrm{tv}}\,.$$

As for all $x_\ell\in\mathcal{X}$, $\mathcal{L}^\star_{\ell,n}(x_\ell,\mathbb{1})/\|\mathcal{L}^\star_{\ell,n}(\cdot,\mathbb{1})\|_\infty \ge \delta^\star/(1-\delta^\star)$, following the same steps as for the proof of Proposition 2.1 yields

$$\left\|\frac{\widehat{\nu}_\ell(\cdot)}{\widehat{\nu}_\ell(\mathbb{1})} - \frac{\nu_\ell(\cdot)}{\nu_\ell(\mathbb{1})}\right\|_{\mathrm{tv}} \le 2\left(\frac{1-\delta^\star}{\delta^\star}\right)\left(\|\mathbf{Q}_\star - \widehat{\mathbf{Q}}\|_F/\delta^\star + c_\star^{-1}(y_\ell)\max_{x\in\mathcal{X}}\left|f^\star_x(y_\ell) - \hat{f}_x(y_\ell)\right|\right)\,.$$

C Nonparametric spectral estimators

Theorem 3.1 follows from the following more precise results proved in this section. The proofs of the intermediate lemmas require assumptions **[H1']** and **[H2]-[H3]**.

Lemma C.1. *There exist a constant* $0 < \sigma_{K,\mathfrak{F}^\star} \le 1$ *and a positive integer* $M_{\mathfrak{F}^\star}$ *such that for all* $M \ge M_{\mathfrak{F}^\star}$,

$$\sigma_K(\mathbf{O}_M) \ge \sigma_{K,\mathfrak{F}^\star} > 0\,.$$

Proof. By **[H3]**, the $(K\times K)$ Gram matrix defined by $\mathbf{O}_\star^\top\mathbf{O}_\star := (\langle f^\star_{x_1}, f^\star_{x_2}\rangle)_{x_1,x_2\in\mathcal{X}}$ is invertible. Let $\varepsilon_{\mathfrak{F}^\star,M}$ be given by:

$$\varepsilon_{\mathfrak{F}^\star,M} := \left\|\mathbf{O}_M^\top\mathbf{O}_M - \mathbf{O}_\star^\top\mathbf{O}_\star\right\| = \left\|(\langle f^\star_{M,x_1}, f^\star_{M,x_2}\rangle - \langle f^\star_{x_1}, f^\star_{x_2}\rangle)_{x_1,x_2\in\mathcal{X}}\right\|\,. \tag{16}$$

From (5), there exists $M_{\mathfrak{F}^\star} \ge 1$ such that for all $M \ge M_{\mathfrak{F}^\star}$, $\varepsilon_{\mathfrak{F}^\star,M} \le 3\lambda_K(\mathbf{O}_\star^\top\mathbf{O}_\star)/4$. By Weyl's inequality (see Theorem D.1), $\sigma_K^2(\mathbf{O}_M) = \lambda_K(\mathbf{O}_M^\top\mathbf{O}_M) \ge \lambda_K(\mathbf{O}_\star^\top\mathbf{O}_\star)/4$. If $\sigma_K(\mathbf{O}_\star) := \lambda_K^{1/2}(\mathbf{O}_\star^\top\mathbf{O}_\star)$, note that for all $M \ge M_{\mathfrak{F}^\star}$, $\sigma_K(\mathbf{O}_M) \ge \sigma_K(\mathbf{O}_\star)/2$, which concludes the proof. □

Define the *pseudo spectral gap* $\mathbb{G}_{\rm ps}$ of the Markov chain $(X_n)_{n\geq 1}$ as

$$\mathbb{G}_{\rm ps} := \max_{k\geq 1} \left\{ \mathbb{G}\left(\mathfrak{Diag}[\pi^\star]^{-1} (\mathbf{Q}_\star^\top)^k \mathfrak{Diag}[\pi^\star] \mathbf{Q}_\star^k \right) / k \right\} ,$$

where $\mathbb{G}(A)$ denotes the spectral gap of a transition matrix A defined by

$$\mathbb{G}(A) := \begin{cases} 1 - \max\{\lambda \ : \ \lambda \text{ eigenvalue of } A \,,\ \lambda \neq 1\} & \text{if eigenvalue 1 has multiplicity 1,} \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\mathbb{G}_{\rm ps}$ depends only on the transition matrix $\mathbf{Q}_\star$ which is assumed to be aperiodic and irreducible with unique stationary distribution $\pi^\star$. Perron-Frobenius theorem ensures that the spectral gap $\mathbb{G}(A)$ is well defined and such that $0 \leq \mathbb{G}(A) \leq 2$.

Remark C.1. *If $\mathbf{Q}_\star$ is aperiodic and irreducible then $\mathbb{G}_{\rm ps} > 0$. In this case, there exists k such that $\mathbf{Q}_\star^k$ is positive (entrywise) and so is $A := \mathfrak{Diag}[\pi^\star]^{-1} (\mathbf{Q}_\star^\top)^k \mathfrak{Diag}[\pi^\star] \mathbf{Q}_\star^k$. As A is a positive transition matrix, Perron-Frobenius theorem ensures that its spectral gap is positive.*

Remark C.2. *If $\mathbf{Q}_\star$ is aperiodic, irreducible and reversible then $\mathbb{G}_{\rm ps} = \mathbb{G}(\mathbf{Q}_\star)(2 - \mathbb{G}(\mathbf{Q}_\star)) > 0$, see [32] and references therein.*

Define the mixing time $\mathbb{T}_{\rm mix}$ of the Markov chain $(X_n)_{n\geq 1}$ as

$$\mathbb{T}_{\rm mix} := \frac{1 + 3\log 2 - \log \pi^\star_{\min}}{\mathbb{G}_{\rm ps}} .$$

This mixing time has a deeper interpretation in terms of convergence towards the stationary distribution in total variation norm, see [32] for instance. For any $\delta \in (0,1)$, set

$$\mathcal{C}_\star(\mathbf{Q}_\star, \delta) := \sqrt{2/\mathbb{G}_{\rm ps}} + 2\sqrt{-2\,\mathbb{T}_{\rm mix} \log \delta} \,, \tag{17}$$

which is a constant that depends only on $\mathbf{Q}_\star$ and δ.

Theorem C.2. *Assume that* **[H1']** *and* **[H2]-[H3]** *hold. Let $\delta, \delta' \in (0,1)$ then, with probability greater than $1 - 2\delta - 4\delta'$, there exists a permutation $\tau \in \mathcal{S}_K$ such that the spectral method estimators $\widehat{f}_{M,x}$, $\widehat{\pi}$ and $\widehat{\mathbf{Q}}$ (see Algorithm 1 for a definition) satisfy, for any $M \geq M_{\mathfrak{F}^\star}$,*

- *for all $p \geq \mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')$ and all $x \in \mathcal{X}$,*

$$\| f^\star_{M,x} - \widehat{f}_{M,\tau(x)} \|_2 \leq \mathcal{C}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \eta_3(\Phi_M) / \sqrt{p} \,, \tag{18}$$

- *for all $p \geq \mathbf{N}_2(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')$,*

$$\| \mathbf{Q}_\star - \mathbb{P}_\tau \widehat{\mathbf{Q}} \mathbb{P}_\tau^\top \| \leq \mathcal{D}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \eta_3(\Phi_M) / \sqrt{p} \,, \tag{19}$$

- *for all $p \geq \mathbf{N}_3(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')$,*

$$\| \pi^\star - \mathbb{P}_\tau \widehat{\pi} \|_2 \leq \mathcal{E}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \eta_3(\Phi_M) / \sqrt{p} \,, \tag{20}$$

where $\mathbb{P}_\tau$ is the permutation matrix associated with τ, and

$$\mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') := \frac{4K}{3\sigma^2_{K,\mathfrak{F}^\star}} \mathcal{C}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta)^2 \, \mathcal{C}_\star(\mathbf{Q}_\star, \delta')^2 \eta_3(\Phi_M)^2 \,,$$

$$\mathbf{N}_2(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') := \frac{4}{{\pi^\star}^2_{\min}} \mathcal{D}'_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta)^2 \, \mathcal{C}_\star(\mathbf{Q}_\star, \delta')^2 \eta_3(\Phi_M)^2 \,,$$

$$\mathbf{N}_3(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') := \frac{4}{\sigma^2_K(\mathbf{A}_{\mathbf{Q}_\star})} \mathcal{D}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta)^2 \, \mathcal{C}_\star(\mathbf{Q}_\star, \delta')^2 \eta_3(\Phi_M)^2 \,,$$

with

$$\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) := \frac{2}{\sqrt{M}}\frac{\max_{x\in\mathcal{X}}\|f_x^\star\|_2}{\sigma^2_{K,\mathfrak{F}^\star}\pi^\star_{\min}\sigma_K(\mathbf{Q}_\star{}^2)} + \left[1+\frac{\|g^\star\|_2}{\pi^\star_{\min}\sigma^2_{K,\mathfrak{F}^\star}\sigma_K(\mathbf{Q}_\star{}^2)}\frac{1}{\sqrt{M}}\right]$$
$$\times\left[\frac{13\kappa^2(\mathbf{Q}_\star)K^{1/2}}{\pi^\star_{\min}\sigma_K(\mathbf{Q}_\star{}^2)}\frac{\kappa^2_{\mathfrak{F}^\star}}{\sigma^2_{K,\mathfrak{F}^\star}} + \frac{83}{\delta}\frac{\kappa^6(\mathbf{Q}_\star)K^5}{\pi^\star_{\min}\sigma_K(\mathbf{Q}_\star{}^2)}\frac{\kappa^6_{\mathfrak{F}^\star}\max_{k\in\mathcal{X}}\|f_k^\star\|_2}{\sigma^3_{K,\mathfrak{F}^\star}}\left\{1+\left(2\log\frac{K^2}{\delta}\right)^{1/2}\right\}\right],$$

$$\mathcal{D}'_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) := \frac{2}{3\sigma^2_{K,\mathfrak{F}^\star}}\left[4\sqrt{K}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\max_{x\in\mathcal{X}}\|f_x^\star\|_2 + \frac{3\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}{M}\right],$$

$$\mathcal{D}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) := \frac{8\|f^\star_{(Y_1,Y_3)}\|_2}{3\sigma^2_{K,\mathfrak{F}^\star}\pi^{\star 2}_{\min}}\left[\mathcal{D}'_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) + 4\sqrt{3K}\pi^\star_{\min}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) + \frac{5\pi^\star_{\min}}{\|f^\star_{(Y_1,Y_3)}\|_2\sqrt{M}}\right],$$

$$\mathcal{E}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) := \frac{16\|f^\star_{(Y_1,Y_3)}\|_2}{\sigma^2_K(\mathbf{A}_{\mathbf{Q}_\star})\sigma^2_{K,\mathfrak{F}^\star}\pi^{\star 2}_{\min}}\left[\mathcal{D}'_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) + 4\sqrt{3K}\pi^\star_{\min}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta) + \frac{5\pi^\star_{\min}}{\|f^\star_{(Y_1,Y_3)}\|_2\sqrt{M}}\right],$$

where $\kappa_{\mathfrak{F}^\star}$ is given in Lemma C.4, for all $(y_1,y_2,y_3)\in\mathcal{Y}^3$,

$$g^\star(y_1,y_2,y_3) := \sum_{x_1,x_2,x_3\in\mathcal{X}}\pi^\star(x_1)\mathbf{Q}_\star(x_1,x_2)\mathbf{Q}_\star(x_2,x_3)f^\star_{x_1}(y_1)f^\star_{x_2}(y_2)f^\star_{x_3}(y_3),$$

and $\sigma^2_K(\mathbf{A}_{\mathbf{Q}_\star})$ is the K-th largest singular value of $\begin{pmatrix}\mathrm{Id}_K-(\mathbf{Q}_\star)^\top\\ \mathbb{1}_K^\top\end{pmatrix}$ which is positive, see (29).

Theorem C.2 is proved using the analysis of [3] to control the L^2-error of the estimation based on the spectral method described in Section 3.1. Establishing this control in the nonparametric framework requires to state explicitly how all constants depend on the dimension M. Therefore, Theorem C.3 recasts and optimizes the results of [3] and is proved in Appendix F. Define

$$\gamma(\mathbf{O}_M) := \min_{x_1\neq x_2}\|\mathbf{O}_M(.,x_1)-\mathbf{O}_M(.,x_2)\|_2 \tag{21}$$

and for all $A\in\mathbb{R}^{M\times M\times M}$ and all $B\in\mathbb{R}^{M\times K}$

$$\|A\|_{\infty,2} := \max_{\|v\|_2=1}\left\|\sum_{b=1}^M v_b A(.,b,.)\right\| \quad\text{and}\quad \|B\|_{2,\infty} := \max_{x\in\mathcal{X}}\|B(.,x)\|_2 . \tag{22}$$

Theorem C.3. *Let $0<\delta<1$. Assume that $3\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|\le\sigma_K(\mathbf{P}_M)$ and that*

$$8.2K^{5/2}(K-1)\frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\left[\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}+\frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\right]<1, \tag{23}$$

$$43.4K^4(K-1)\frac{\kappa^4(\mathbf{Q}_\star\mathbf{O}_M^\top)}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\left[\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}+\frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\right]\le 1, \tag{24}$$

then, with probability greater than $1-2\delta$, the matrix $\widehat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\widehat{\mathbf{U}}$ is invertible, the random matrix $\widehat{\mathbf{C}}(1)$ is diagonalisable (see Algorithm 1), and there exists a permutation $\tau\in\mathcal{S}_K$ such that for all $x\in\mathcal{X}$,

$$\|\mathbf{O}_M(.,x)-\widehat{\mathbf{O}}_M(.,\tau(x))\|_2\le\frac{2\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\|\mathbf{O}_M\|_{2,\infty}+\left[\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}+\frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\right]$$
$$\times\left[13K^{1/2}\frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{\sigma_K(\mathbf{P}_M)}+116K^5\left\{1+\left(2\log(K^2/\delta)\right)^{1/2}\right\}\frac{\kappa^6(\mathbf{Q}_\star\mathbf{O}_M^\top)\|\mathbf{O}_M\|_{2,\infty}}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\right].$$

Preliminary lemmas

Lemma C.4. *There exists a constant $\kappa_{\mathfrak{F}^\star}$ that depends only on $\mathfrak{F}^\star$ such that for all $M\ge M_{\mathfrak{F}^\star}$, $\kappa(\mathbf{O}_M)\le\kappa_{\mathfrak{F}^\star}$ where $M_{\mathfrak{F}^\star}$ is given in Lemma C.1. For all $M\ge M_{\mathfrak{F}^\star}$, $\kappa(\mathbf{Q}_\star\mathbf{O}_M^\top)\le\kappa_{\mathfrak{F}^\star}\kappa(\mathbf{Q}_\star)$.*

Proof. Note that $\mathbf{O}_\star^\top \mathbf{O}_\star$ is nonsingular. From (5) and (16) we deduce that $\mathbf{O}_M^\top \mathbf{O}_M$ tends to $\mathbf{O}_\star^\top \mathbf{O}_\star$ as M grows to infinity. This proves the first point. Recall that $\sigma_i(AB) \leq \sigma_1(A)\sigma_i(B)$ for all $i = 1, \dots, K$. Applying this identity to $A = \mathbf{Q}_\star{}^{-1}$ and $B = \mathbf{Q}_\star \mathbf{O}_M^\top$ yields $\sigma_K(\mathbf{Q}_\star)\sigma_K(\mathbf{O}_M) \leq \sigma_K(\mathbf{Q}_\star \mathbf{O}_M^\top)$. It follows that $\kappa(\mathbf{Q}_\star \mathbf{O}_M^\top) \leq \kappa(\mathbf{Q}_\star)\kappa(\mathbf{O}_M)$. The second claim follows from the first claim. □

Lemma C.5. *For all $M \geq M_{\mathfrak{F}^\star}$, $\gamma(\mathbf{O}_M) \geq \sqrt{2}\sigma_{K,\mathfrak{F}^\star}$ and $\|\mathbf{O}_M\|_{2,\infty} \leq \max_{x\in\mathcal{X}} \|f_x^\star\|_2$, where $\gamma(\mathbf{O}_M)$ and $\|\mathbf{O}_M\|_{2,\infty}$ are defined in (21) and (22).*

Proof. Observe that $\|\mathbf{O}_M v\|_2 \geq \sigma_K(\mathbf{O}_M)\|v\|_2$. With an appropriate choice of v and using Lemma C.1 this proves the first inequality. As Φ_M is an orthonormal family, $\|\mathbf{O}_M(.\,, x)\|_2 \leq \|f_x^\star\|_2$ which proves the second claim. □

Lemma C.6. *For all $M \geq 1$,*

$$\|\mathbf{M}_M\|_{\infty,2} := \max_{\|v\|_2=1} \left\| \sum_{b=1}^M v_b \mathbf{M}_M(.\,, b, .\,) \right\| \leq \|g^\star\|_2 \,,$$

where $\|\cdot\|_{\infty,2}$ is defined in (22).

Proof. As for all $x \in \mathcal{X}$, $f_x^\star \in \mathrm{L}^2(\mathcal{Y}, \mathcal{L}^D)$, $g^\star \in \mathrm{L}^2(\mathcal{Y}^3, \mathcal{L}^{D^{\otimes 3}})$. Denote by $\langle .\,, .\,\rangle_{\mathrm{L}^2(\mathcal{Y}^3, \mathcal{L}^{D^{\otimes 3}})}$ the inner product of $\mathrm{L}^2(\mathcal{Y}^3, \mathcal{L}^{D^{\otimes 3}})$. As $\varphi_{a,b,c}(y_1, y_2, y_3) := \varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3)$ is an orthonormal family of $\mathrm{L}^2(\mathcal{Y}^3, \mathcal{L}^{D^{\otimes 3}})$,

$$\begin{aligned}
\|\mathbf{M}_M\|_{\infty,2} &= \max_{\|v\|_2=1} \left\| \sum_{b=1}^M v_b \mathbf{M}_M(.\,, b, .\,) \right\| \leq \max_{\|v\|_2=1} \sum_{b=1}^M |v_b| \|\mathbf{M}_M(.\,, b, .\,)\| \,, \\
&\leq \left(\sum_{b=1}^M \|\mathbf{M}_M(.\,, b, .\,)\|^2 \right)^{1/2} \leq \left(\sum_{b=1}^M \|\mathbf{M}_M(.\,, b, .\,)\|_F^2 \right)^{1/2} \,, \\
&= \left(\sum_{a,b,c=1}^M \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]^2 \right)^{1/2} = \left(\sum_{a,b,c=1}^M \langle g^\star, \varphi_{a,b,c}\rangle^2_{\mathrm{L}^2(\mathcal{Y}^3, \mathcal{L}^{D^{\otimes 3}})} \right)^{1/2} \leq \|g^\star\|_2 \,.
\end{aligned}$$

using Cauchy-Schwarz inequality. □

Lemma C.7. *For all $M \geq 1$, $\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_{\infty,2} \leq \|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F$, where $\|\cdot\|_{\infty,2}$ is defined in (22).*

Proof. For all $M \geq 1$,

$$\begin{aligned}
\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_{\infty,2} &= \max_{\|v\|_2=1} \left\| \sum_{b=1}^M v_b (\widehat{\mathbf{M}}_M - \mathbf{M}_M)(.\,, b, .\,) \right\| \leq \max_{\|v\|_2=1} \sum_{b=1}^M |v_b| \left\| (\widehat{\mathbf{M}}_M - \mathbf{M}_M)(.\,, b, .\,) \right\| \,, \\
&\leq \left(\sum_{b=1}^M \left\| (\widehat{\mathbf{M}}_M - \mathbf{M}_M)(.\,, b, .\,) \right\|^2 \right)^{1/2} \leq \left(\sum_{b=1}^M \left\| (\widehat{\mathbf{M}}_M - \mathbf{M}_M)(.\,, b, .\,) \right\|_F^2 \right)^{1/2} \,, \\
&= \left\| \widehat{\mathbf{M}}_M - \mathbf{M}_M \right\|_F \,.
\end{aligned}$$

using Cauchy-Schwarz inequality. □

Lemma C.8. *Under* **[H1']** *and* **[H2]**, *for all $M \geq 1$, $\sigma_K(\mathbf{P}_M) \geq \pi_{\min} \sigma_K^2(\mathbf{O}_M)\sigma_K(\mathbf{Q}^2)$. If* **[H3]** *holds, then, for all $M \geq M_{\mathfrak{F}^\star}$,*

$$\sigma_K(\mathbf{P}_M) \geq \sigma_{K,\mathfrak{F}^\star}^2 \pi_{\min}^\star \sigma_K(\mathbf{Q}_\star{}^2) \,,$$

where $M_{\mathfrak{F}^\star}$ and $\sigma_{K,\mathfrak{F}^\star}$ are defined in Lemma C.1.

Proof. By Lemma F.1 and (7),

$$\begin{aligned}
\sigma_K(\mathbf{P}_M) = \sigma_K(\mathbf{U}^\top \mathbf{P}_M \mathbf{U}) &= \sigma_K((\mathbf{U}^\top \mathbf{O}_M)\mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star{}^2(\mathbf{U}^\top \mathbf{O}_M)^\top) \,, \\
&\geq \sigma_K(\mathbf{U}^\top \mathbf{O}_M)\sigma_K(\mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star{}^2(\mathbf{U}^\top \mathbf{O}_M)^\top) \,, \\
&= \sigma_K(\mathbf{O}_M)\sigma_K(\mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star{}^2(\mathbf{U}^\top \mathbf{O}_M)^\top) \,, \\
&\geq \sigma_K(\mathfrak{Diag}[\pi^\star])\sigma_K(\mathbf{O}_M)\sigma_K((\mathbf{U}^\top \mathbf{O}_M)^\top)\sigma_K(\mathbf{Q}_\star{}^2) \,, \\
&= \pi_{\min}^\star \sigma_K^2(\mathbf{O}_M)\sigma_K(\mathbf{Q}_\star{}^2) \,,
\end{aligned}$$

which concludes the proof. □

First step: Estimation of the emission laws using a spectral method

Appendix E shows that:

$$\mathbb{P}\Big[\|\widehat{\mathbf{L}}_M - \mathbf{L}_M\|_F \geq \mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_1(\Phi_M)/\sqrt{p}\Big] \leq \delta' \,, \quad \mathbb{P}\Big[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F \geq \mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_3(\Phi_M)/\sqrt{p}\Big] \leq \delta' \,,$$
$$\mathbb{P}\Big[\|\widehat{\mathbf{N}}_M - \mathbf{N}_M\|_F \geq \mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_2(\Phi_M)/\sqrt{p}\Big] \leq \delta' \,, \quad \mathbb{P}\Big[\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|_F \geq \mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_2(\Phi_M)/\sqrt{p}\Big] \leq \delta' \,.$$

Using the preliminary lemmas of Section C and the elementary fact that $M\eta_1(\Phi_M) \leq \sqrt{M}\eta_2(\Phi_M) \leq \eta_3(\Phi_M)$, (23) and (24) along with $3\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\| \leq \sigma_K(\mathbf{P}_M)$ are satisfied when $M \geq M_{\mathfrak{F}^\star}$ and $p \geq \mathbf{N}_0(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')$ where:

$$\mathbf{N}_0(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta') := \frac{942}{\delta^2}\frac{\kappa^8(\mathbf{Q}_\star)K^{10}}{\pi^{\star 2}_{\min}\sigma^2_K(\mathbf{Q}_\star{}^2)}\frac{\kappa^8_{\mathfrak{F}^\star}}{\sigma^6_{K,\mathfrak{F}^\star}}\Big(1+\frac{\|g^\star\|_2}{\pi^\star_{\min}\sigma^2_{K,\mathfrak{F}^\star}\sigma_K(\mathbf{Q}_\star{}^2)}\frac{1}{\sqrt{M}}\Big)^2\mathcal{C}_\star(\mathbf{Q}_\star,\delta')^2\eta_3(\Phi_M)^2 \,.$$

Using Theorem C.3, with probability greater than $1-2\delta-4\delta'$, there exists a permutation τ satisfying for any $M \geq M_{\mathfrak{F}^\star}$, $p \geq \mathbf{N}_0(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')$ and $x \in \mathcal{X}$,

$$\|\mathbf{O}_M(.\,,x) - \widehat{\mathbf{O}}_M(.\,,\tau(x))\|_2 \leq \mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_3(\Phi_M)/\sqrt{p} \,.$$

This proves the first part of Theorem C.2.

Second step: Preliminary estimation of the stationary density using a spectral method

For sake of readability, assume that τ is the identity permutation. Observe that:

$$\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta') \geq \mathbf{N}_0(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta') \,.$$

Recall $\tilde{\pi} := \big(\hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M\big)^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{L}}_M$ and $\pi^\star = \big(\hat{\mathbf{U}}^\top\mathbf{O}_M\big)^{-1}\hat{\mathbf{U}}^\top\mathbf{L}_M$.

Lemma C.9. *With probability greater than $1-2\delta-4\delta'$, if $p > \mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')$ then,*

$$\|\tilde{\pi}-\pi^\star\|_2 \leq \frac{2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_1(\Phi_M)}{\sqrt{p}} + \frac{2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\frac{\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}}{\sqrt{p}-\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}}\left(\max_{x\in\mathcal{X}}\|f^\star_x\|_2 + \mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_1(\Phi_M)}{\sqrt{p}}\right) \,.$$

Proof. Set $A = \hat{\mathbf{U}}^\top\mathbf{O}_M$, $\tilde{A} = \hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M$ and $B = \hat{\mathbf{U}}^\top(\mathbf{O}_M - \widehat{\mathbf{O}}_M)$. Then,

$$\|B\| \leq \|\mathbf{O}_M - \widehat{\mathbf{O}}_M\| \leq \|\mathbf{O}_M - \widehat{\mathbf{O}}_M\|_F \leq \sqrt{K}\max_{x\in\mathcal{X}}\|\mathbf{O}_M(.\,,x) - \widehat{\mathbf{O}}_M(.\,,x)\|_2 \,,$$

which gives $\|B\| \leq \sqrt{K}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\eta_3(\Phi_M)/\sqrt{p}$. Similarly, by claim (iii) of Lemma F.3:

$$\|A^{-1}B\| \leq \|A^{-1}\|\|B\| \leq \sigma_K^{-1}(A)\|B\| \leq \frac{2\sqrt{K}\max_{x\in\mathcal{X}}\|\mathbf{O}_M(.\,,x) - \widehat{\mathbf{O}}_M(.\,,x)\|_2}{\sqrt{3}\sigma_K(\mathbf{O}_M)} \,,$$

so that

$$\|A^{-1}B\| \leq \frac{2\sqrt{K}}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_3(\Phi_M)}{\sqrt{p}} \,.$$

Observe that the condition on p and M ensures that $\|A^{-1}B\| < 1$. Apply Theorem D.2 to get that:

$$\|(\hat{\mathbf{U}}^\top\mathbf{O}_M)^{-1} - (\hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M)^{-1}\| \leq \frac{2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\frac{\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}}{\sqrt{p}-\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}} \,. \tag{25}$$

Furthermore, using (25):

$$\begin{aligned}
\|\tilde{\pi}-\pi^\star\|_2 &= \|\big(\hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M\big)^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{L}}_M - \big(\hat{\mathbf{U}}^\top\mathbf{O}_M\big)^{-1}\hat{\mathbf{U}}^\top\mathbf{L}_M\|_2 \\
&= \|\big(\hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M\big)^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{L}}_M - \big(\hat{\mathbf{U}}^\top\mathbf{O}_M\big)^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{L}}_M + \big(\hat{\mathbf{U}}^\top\mathbf{O}_M\big)^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{L}}_M - \big(\hat{\mathbf{U}}^\top\mathbf{O}_M\big)^{-1}\hat{\mathbf{U}}^\top\mathbf{L}_M\|_2 \\
&\leq \|(\hat{\mathbf{U}}^\top\mathbf{O}_M)^{-1} - (\hat{\mathbf{U}}^\top\widehat{\mathbf{O}}_M)^{-1}\|\|\widehat{\mathbf{L}}_M\|_2 + \|A^{-1}\|\|\widehat{\mathbf{L}}_M - \mathbf{L}_M\|_2 \\
&\leq \frac{2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\left(\|\widehat{\mathbf{L}}_M - \mathbf{L}_M\|_2 + \frac{\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}}{\sqrt{p}-\sqrt{\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta')}}(\|\mathbf{L}_M\|_2 + \|\widehat{\mathbf{L}}_M - \mathbf{L}_M\|_2)\right) \,.
\end{aligned}$$

Write $f^\star_{Y_1} = \sum_{x_1 \in \mathcal{X}} \pi(x_1) f^\star_{k_1}(y_1)$ the density of Y_1. Observe that:

$$\|\mathbf{L}_M\|_2 = \left(\sum_{a=1}^M \mathbb{E}\left[\varphi_a(Y_1)\right]^2 \right)^{1/2} = \left(\sum_{a=1}^M \langle f^\star_{Y_1}, \varphi_a \rangle^2 \right)^{1/2} \leq \|f^\star_{Y_1}\|_2 \leq \max_{x \in \mathcal{X}} \|f^\star_x\|_2 \,,$$

which concludes the proof. □

This results allows to state that for all $p \geq 4\mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')$,

$$\|\pi^\star - \mathbb{P}_\tau \tilde{\pi}\|_2 \leq \mathcal{D}'_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \eta_3(\Phi_M)/\sqrt{p} \,. \tag{26}$$

Third step: Estimation of the transition matrix using a spectral method

Write $\tilde{\mathbf{Q}} := \left(\hat{\mathbf{U}}^\top \widehat{\mathbf{O}}_M \mathfrak{Diag}[\tilde{\pi}]\right)^{-1} \hat{\mathbf{U}}^\top \widehat{\mathbf{N}}_M \hat{\mathbf{U}} \left(\widehat{\mathbf{O}}_M^\top \hat{\mathbf{U}}\right)^{-1}$ and note that $\hat{\mathbf{Q}} = \Pi_{TM}(\tilde{\mathbf{Q}})$ and $\mathbf{Q}_\star = \Pi_{TM}(\mathbf{Q}_\star)$. Then, by non-expansivity of the projection onto convex sets, $\|\hat{\mathbf{Q}} - \mathbf{Q}_\star\|_F \leq \|\tilde{\mathbf{Q}} - \mathbf{Q}_\star\|_F$. Moreover,

$$\mathbf{N}_2(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') \geq 4\mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') \geq \mathbf{N}_0(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta') \,.$$

Lemma C.10. *With probability greater than* $1 - 2\delta - 4\delta'$*, if* $p \geq \mathbf{N}_2(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')$ *then*

$$\|\tilde{\mathbf{Q}} - \mathbf{Q}_\star\| \leq \frac{8\|f^\star_{(Y_1,Y_3)}\|_2}{3\sigma^2_{K,\mathfrak{F}^\star} {\pi^\star}^2_{\min}} \|\tilde{\pi} - \pi^\star\|_2 + \frac{2}{\pi^\star_{\min}} \tilde{\mathcal{E}}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \frac{\eta_3(\Phi_M)}{\sqrt{p}} \,,$$

where

$$\tilde{\mathcal{E}}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) := \frac{16}{\sqrt{3}\sigma^2_{K,\mathfrak{F}^\star}} \left[\sqrt{K} \mathcal{C}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \|f^\star_{(Y_1,Y_3)}\|_2 + \frac{5}{4\sqrt{3M}} \right] .$$

Proof. Observe that (20) shows that $\|\tilde{\pi} - \pi^\star\|_2 \leq \pi^\star_{\min}/2$. Then, for any $x \in \mathcal{X}$:

$$\tilde{\pi}_x \geq \frac{\pi^\star_{\min}}{2} > 0 \,. \tag{27}$$

Set $\mathbf{V} = (\hat{\mathbf{U}}^\top \mathbf{O}_M)^{-1} \hat{\mathbf{U}}^\top$ and $\hat{\mathbf{V}} = (\hat{\mathbf{U}}^\top \widehat{\mathbf{O}}_M)^{-1} \hat{\mathbf{U}}^\top$. Note $\tilde{\mathbf{Q}} = \mathfrak{Diag}[\tilde{\pi}]^{-1} \hat{\mathbf{V}} \widehat{\mathbf{N}}_M \hat{\mathbf{V}}^\top$ and:

$$\mathbf{Q} = \mathfrak{Diag}[\pi^\star]^{-1} \mathbf{V} \mathbf{N}_M \mathbf{V}^\top \,.$$

Set $E = \hat{\mathbf{V}} - \mathbf{V}$ and $F = \widehat{\mathbf{N}}_M - \mathbf{N}_M$. Using (25) yields:

$$\|E\| \leq \frac{2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}} \frac{\sqrt{\mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')}}{\sqrt{p} - \sqrt{\mathbf{N}_1(\mathbf{Q}_\star, \mathfrak{F}^\star, \Phi_M, \delta, \delta')}} \leq \frac{8\sqrt{K}}{3\sigma^2_{K,\mathfrak{F}^\star}} \mathcal{C}_M(\mathbf{Q}_\star, \mathfrak{F}^\star, \delta) \mathcal{C}_\star(\mathbf{Q}_\star, \delta') \frac{\eta_3(\Phi_M)}{\sqrt{p}} \,.$$

By claim (iii) of Lemma F.3, $\|\mathbf{V}\| \leq \sigma_K^{-1}(\hat{\mathbf{U}}^\top \mathbf{O}_M) \leq 2/(\sqrt{3}\sigma_{K,\mathfrak{F}^\star})$. Furthermore, $\varphi_{a,c}(y_1, y_3) := \varphi_a(y_1)\varphi_c(y_3)$ is an orthonormal family of $\mathrm{L}^2(\mathcal{Y}^2, \mathcal{L}^{D^{\otimes 2}})$ and

$$\|\mathbf{N}_M\|_F = \Big(\sum_{a,c=1}^M \mathbb{E}\left[\varphi_a(Y_1)\varphi_c(Y_3)\right]^2 \Big)^{1/2} = \Big(\sum_{a,c=1}^M \langle f^\star_{(Y_1,Y_3)}, \varphi_{a,c} \rangle^2_{\mathrm{L}^2(\mathcal{Y}^2, \mathcal{L}^{D^{\otimes 2}})} \Big)^{1/2} \leq \|f^\star_{(Y_1,Y_3)}\|_2 \,.$$

Then,

$$\begin{aligned}
\|\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - \hat{\mathbf{V}}\widehat{\mathbf{N}}_M\hat{\mathbf{V}}^\top\| &= \|\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - (\mathbf{V}+E)(\mathbf{N}_M+F)(\mathbf{V}+E)^\top\| \,, \\
&= \|\mathbf{V}\mathbf{N}_M E^\top + \mathbf{V}F\mathbf{V}^\top + \mathbf{V}FE^\top + E\mathbf{N}_M\mathbf{V}^\top + E\mathbf{N}_M E^\top + EF\mathbf{V}^\top + EFE^\top\| \,, \\
&\leq 2\|E\|\|\mathbf{V}\|\|\mathbf{N}_M\| + 2\|E\|\|\mathbf{V}\|\|F\| + \|E\|^2\|\mathbf{N}_M\| + \|\mathbf{V}\|^2\|F\| + \|E\|^2\|F\| \,,
\end{aligned}$$

yields

$$\|\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - \hat{\mathbf{V}}\widehat{\mathbf{N}}_M\hat{\mathbf{V}}^\top\| \leq \frac{32\sqrt{K}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\|f^\star_{(Y_1,Y_3)}\|_2}{3\sqrt{3}\sigma^3_{K,\mathfrak{F}^\star}}\Bigg[1+\frac{\mathcal{C}_\star(\mathbf{Q}_\star,\delta')}{\|f^\star_{(Y_1,Y_3)}\|_2}\frac{\eta_3(\Phi_M)}{\sqrt{pM}}$$
$$+\frac{2\sqrt{K}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}\frac{\eta_3(\Phi_M)}{\sqrt{p}}$$
$$+\frac{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}}{4\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\|f^\star_{(Y_1,Y_3)}\|_2\sqrt{K}}\frac{1}{\sqrt{M}}$$
$$+\frac{2\sqrt{K}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')^2}{\sqrt{3}\sigma_{K,\mathfrak{F}^\star}\|f^\star_{(Y_1,Y_3)}\|_2}\frac{\eta_3^2(\Phi_M)}{p\sqrt{M}}\Bigg]\frac{\eta_3(\Phi_M)}{\sqrt{p}}$$

As $p \geq \mathbf{N}_2(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta') \geq 4\mathbf{N}_1(\mathbf{Q}_\star,\mathfrak{F}^\star,\Phi_M,\delta,\delta') = \frac{16K}{3\sigma^2_{K,\mathfrak{F}^\star}}\mathcal{C}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)^2\,\mathcal{C}_\star(\mathbf{Q}_\star,\delta')^2\eta_3(\Phi_M)^2$,

$$\|\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - \hat{\mathbf{V}}\widehat{\mathbf{N}}_M\hat{\mathbf{V}}^\top\| \leq \tilde{\mathcal{E}}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_3(\Phi_M)}{\sqrt{p}}\,. \tag{28}$$

Observe that:

$$\begin{aligned}
\|\mathbf{Q}_\star - \tilde{\mathbf{Q}}\| &= \|(\mathfrak{Diag}[\pi^\star]^{-1} - \mathfrak{Diag}[\widehat{\pi}]^{-1})\mathbf{V}\mathbf{N}_M\mathbf{V}^\top + \mathfrak{Diag}[\widehat{\pi}]^{-1}(\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - \hat{\mathbf{V}}\widehat{\mathbf{N}}_M\hat{\mathbf{V}}^\top)\| \\
&\leq \|\mathfrak{Diag}[\pi^\star]^{-1} - \mathfrak{Diag}[\widehat{\pi}]^{-1}\|\|\mathbf{V}\|^2\|\mathbf{N}_M\| + \|\mathfrak{Diag}[\widehat{\pi}]^{-1}\|\|\mathbf{V}\mathbf{N}_M\mathbf{V}^\top - \hat{\mathbf{V}}\widehat{\mathbf{N}}_M\hat{\mathbf{V}}^\top\| \\
&\leq \frac{4\|f^\star_{(Y_1,Y_3)}\|_2}{3\sigma^2_{K,\mathfrak{F}^\star}}\max_{x\in\mathcal{X}}({\pi^\star}_x^{-1} - \tilde{\pi}_x^{-1}) + \max_{x\in\mathcal{X}}\widehat{\pi}_x^{-1}\tilde{\mathcal{E}}_M(\mathbf{Q}_\star,\mathfrak{F},\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_3(\Phi_M)}{\sqrt{p}} \\
&\leq \frac{8\|f^\star_{(Y_1,Y_3)}\|_2}{3\sigma^2_{K,\mathfrak{F}^\star}{\pi^\star}^2_{\min}}\|\tilde{\pi} - \pi^\star\|_2 + \frac{2}{\pi^\star_{\min}}\tilde{\mathcal{E}}_M(\mathbf{Q}_\star,\mathfrak{F}^\star,\delta)\mathcal{C}_\star(\mathbf{Q}_\star,\delta')\frac{\eta_3(\Phi_M)}{\sqrt{p}}\,,
\end{aligned}$$

using (27) and (28). □

Combining (26) and Lemma C.10 proves the second point of Theorem C.2.

Last step: Final estimation of the stationary distribution

By **[H1']**, the transition matrix $\mathbf{Q}_\star$ is irreducible and aperiodic. Perron-Frobenius theorem shows that $\mathbf{Q}_\star$ has a unique stationary distribution $\pi^\star$. More precisely,

- $\mathbb{R}\,.\,\pi^\star = \ker(\mathrm{Id}_K - (\mathbf{Q}_\star)^\top)$ so that $(\mathbb{R}\,.\,\pi^\star)^\perp = \mathrm{range}(\mathrm{Id}_K - \mathbf{Q}_\star)$,
- and $\langle \pi^\star, \mathbb{1}_K\rangle = 1$,

where $\mathbb{1}_K = (1,\dots,1) \in \mathbb{R}^K$. Then, $\mathbb{1}_K \notin \mathrm{range}(\mathrm{Id}_K - \mathbf{Q}_\star)$ and

$$\mathrm{Rank}\begin{pmatrix}\mathrm{Id}_K - (\mathbf{Q}_\star)^\top \\ \mathbb{1}_K^\top\end{pmatrix} = K\,. \tag{29}$$

Set

$$A = \begin{pmatrix}\mathrm{Id}_K - \mathbf{Q}^\top \\ \mathbb{1}_K^\top\end{pmatrix} \quad \text{and} \quad A^\star = \begin{pmatrix}\mathrm{Id}_K - (\mathbf{Q}_\star)^\top \\ \mathbb{1}_K^\top\end{pmatrix}\,.$$

Derive first an upper bound on $\|A^+ - (A^\star)^+\|$ where A^+ denotes the Moore-Penrose pseudo-inverse of A. Note that

$$A^+ - (A^\star)^+ = (A^\star)^+(A^\star - A)A^+ - (A^\star)^+(\mathrm{Id}_{K+1} - AA^+)\,. \tag{30}$$

The last term can be written as

$$(A^\star)^+(\mathrm{Id}_{K+1} - AA^+) = (A^\star)^+(A^\star(A^\star)^+)(\mathrm{Id}_{K+1} - AA^+) = (A^\star)^+P_{\mathrm{range}(A^\star)}P_{\mathrm{range}(A)^\perp}\,,$$

where $P_{\mathrm{range}(A^\star)} = A^\star(A^\star)^+$ denotes the orthogonal projection onto $\mathrm{range}(A^\star)$ and $P_{\mathrm{range}(A)^\perp} = \mathrm{Id}_{K+1} - AA^+$ denotes the orthogonal projection onto the orthogonal of $\mathrm{range}(A)$. Define

$$s(\mathbf{Q}_\star) := \sigma_K(A^\star)\,. \tag{31}$$

Lemma C.11. *If* $\|\mathbf{Q}-\mathbf{Q}_\star\| \leq s(\mathbf{Q}_\star)/2$ *then* $\mathrm{Rank}(A) = \mathrm{Rank}(A^\star) = K$ *and*

$$\|P_{\mathrm{range}(A^\star)} P_{\mathrm{range}(A)^\perp}\| \leq \frac{2\|\mathbf{Q}-\mathbf{Q}_\star\|}{s(\mathbf{Q}_\star)} .$$

Proof. The first point follows from Weyl's inequality, see Theorem D.1. By [39],

$$\|P_{\mathrm{range}(A^\star)^\perp} P_{\mathrm{range}(A)}\| = \|P_{\mathrm{range}(A)^\perp} P_{\mathrm{range}(A^\star)}\| .$$

Moreover, since projections P are orthogonal $(P_{\mathrm{range}(A)^\perp} P_{\mathrm{range}(A^\star)})^\top = P_{\mathrm{range}(A^\star)} P_{\mathrm{range}(A)^\perp}$. Using notation of [39], one may notice that $\|\sin\theta(\mathrm{range}(A), \mathrm{range}(A^\star))\| = \|P_{\mathrm{range}(A^\star)^\perp} P_{\mathrm{range}(A)}\|$. By Wedin's theorem [39], if $\sigma_K(A) \geq s(\mathbf{Q}_\star)/2$ then $\|\sin\theta(\mathrm{range}(A), \mathrm{range}(A^\star))\| \leq \frac{2\|A-A^\star\|}{\sigma_K(A^\star)}$. We conclude using Weyl's inequality, see Theorem D.1. □

Triangular inequality in (30) gives

$$\begin{aligned}\|A^+ - (A^\star)^+\| &\leq \|(A^\star)^+\| \|\mathbf{Q}-\mathbf{Q}_\star\| \Big(\|A^+\| + \frac{2}{\sigma_K(A^\star)} \Big) , \\ &\leq \frac{\|\mathbf{Q}-\mathbf{Q}_\star\|}{\sigma_K(A^\star)} \Big(\|A^+ - (A^\star)^+\| + \frac{3}{\sigma_K(A^\star)} \Big) ,\end{aligned}$$

using that $\|(A^\star)^+\| = 1/\sigma_K(A^\star)$. Deduce that if $\|\mathbf{Q}-\mathbf{Q}_\star\| \leq \sigma_K(A^\star)/2$ then $\|A^+ - (A^\star)^+\| \leq 6\|\mathbf{Q}-\mathbf{Q}_\star\|/\sigma_K^2(A^\star)$. From Weyl's inequality, if $\|\mathbf{Q}-\mathbf{Q}_\star\| \leq \sigma_K(A^\star)/2$ then $\sigma_K(A) \geq \sigma_K(A^\star)/2$. $\mathrm{Id}_K - \mathbf{Q}^\top$ has rank $K-1$ and the eigenspace $\ker(\mathrm{Id}_K - \mathbf{Q}^\top)$ has dimension 1. Thus, $\mathbf{Q}$ is an irreducible and aperiodic transition matrix, and π is the unique solution to

$$\begin{pmatrix} \mathrm{Id}_K - \mathbf{Q}^\top \\ \mathbb{1}_K^\top \end{pmatrix} \pi = \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

Now $\|\pi - \pi^\star\|_2 \leq \|A^+ - (A^\star)^+\|$ and the last part of Theorem C.2 is proved.

D Matrix perturbation

This section provides some useful results in matrix perturbation theory. Proofs of the following theorems may be found in [35] for instance.

Theorem D.1 (Weyl's inequality)**.** *Let* A, B *be* $(p \times q)$ *matrices with* $p \geq q$ *then, for all* $i = 1, \ldots, q$*,*

$$|\sigma_i(A+B) - \sigma_i(A)| \leq \sigma_1(B) .$$

Theorem D.2. *Let* A, B *be* $(p \times p)$ *matrices. If* A *is invertible and* $\|A^{-1}B\| < 1$ *then* $\tilde{A} := A + B$ *is invertible and*

$$\|\tilde{A}^{-1} - A^{-1}\| \leq \frac{\|B\| \|A^{-1}\|^2}{1 - \|A^{-1}B\|} .$$

Theorem D.3 (Bauer-Fike)**.** *Let* A, B *be* $(p \times p)$ *matrices and* $\tilde{A} := A + B$*. Assume that* A *is diagonalizable, i.e.* $X^{-1}AX = \Lambda$*, where* $\Lambda = \mathfrak{Diag}[(\lambda_1, \ldots, \lambda_p)]$*. Then,*

$$\mathrm{sv}_A(\tilde{A}) \leq \kappa(X)\|B\| , \tag{32}$$

where $\mathrm{sv}_A(\tilde{A}) := \max_j \min_i |\tilde{\lambda}_j - \lambda_i|$ *and* $\tilde{\lambda}_j$ *denotes the eigenvalues of* $\tilde{A}$*.*

Remark D.1. *If the disks* $\mathcal{D}_i := \{\xi \,:\, |\xi - \lambda_i| \leq \kappa(X)\|B\|\}$ *are isolated from the others, then* (32) *holds with the matching distance* $\mathrm{md}(A, \tilde{A}) \leq \kappa(X)\|B\|$ *where* $\mathrm{md}(A, \tilde{A}) := \min_{\tau \in \mathcal{S}_p} \max_i |\hat{\lambda}_{\tau(i)} - \lambda_i|$*. Eventually, if* Λ*,* $\tilde{A}$ *are real valued matrices then* $\tilde{A}$ *has* p *distinct real eigenvalues.*

E Concentration inequalities

Consider consecutive observations of the same hidden Markov chain $Z_s := (Y_s, Y_{s+1}, Y_{s+2})$ for $1 \leq s \leq p$,

Lemma E.1. *For any positive u, any M and any p:*

$$\mathbb{P}\Big[\|\widehat{\mathbf{L}}_M - \mathbf{L}_M\|_F \geq \frac{\sqrt{2}\eta_1(\Phi_M)}{\sqrt{p\mathbb{G}_{\mathrm{ps}}}}(1 + 2u\sqrt{1 + \log(8/\pi^\star{}_{\min})})\Big] \leq \exp(-u^2)\,,$$

$$\mathbb{P}\Big[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F \geq \frac{\sqrt{2}\eta_3(\Phi_M)}{\sqrt{p\mathbb{G}_{\mathrm{ps}}}}(1 + 2u\sqrt{1 + \log(8/\pi^\star{}_{\min})})\Big] \leq \exp(-u^2)\,,$$

$$\mathbb{P}\Big[\|\widehat{\mathbf{N}}_M - \mathbf{N}_M\|_F \geq \frac{\sqrt{2}\eta_2(\Phi_M)}{\sqrt{p\mathbb{G}_{\mathrm{ps}}}}(1 + 2u\sqrt{1 + \log(8/\pi^\star{}_{\min})})\Big] \leq \exp(-u^2)\,,$$

$$\mathbb{P}\Big[\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|_F \geq \frac{\sqrt{2}\eta_2(\Phi_M)}{\sqrt{p\mathbb{G}_{\mathrm{ps}}}}(1 + 2u\sqrt{1 + \log(8/\pi^\star{}_{\min})})\Big] \leq \exp(-u^2)\,.$$

Proof. Set $\zeta_{\mathbf{L}_M}(Z_1, \ldots, Z_p) := \|\widehat{\mathbf{L}}_M(Z_1, \ldots, Z_p) - \mathbf{L}_M\|_2$, $\zeta_{\mathbf{M}_M}(Z_1, \ldots, Z_p) := \|\widehat{\mathbf{M}}_M(Z_1, \ldots, Z_p) - \mathbf{M}_M\|_F$, $\zeta_{\mathbf{N}_M}(Z_1, \ldots, Z_p) := \|\widehat{\mathbf{N}}_M(Z_1, \ldots, Z_p) - \mathbf{N}_M\|_F$ and $\zeta_{\mathbf{P}_M}(Z_1, \ldots, Z_p) := \|\widehat{\mathbf{P}}_M(Z_1, \ldots, Z_p) - \mathbf{P}_M\|_F$ where, for instance, $\widehat{\mathbf{L}}_M(Z_1, \ldots, Z_p)$ denotes the dependence of $\widehat{\mathbf{L}}_M$ in $Z_1, \ldots, Z_p$. We begin with $\zeta_{\mathbf{M}_M}$, other cases are similar. Form the difference with respect to the coordinate i:

$$c_i := \sup_{z_j \in \mathcal{Y}^3, z_i' \in \mathcal{Y}^3} |\zeta_{\mathbf{M}_M}(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_p) - \zeta_{\mathbf{M}_M}(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_p)|\,.$$

By the triangular inequality,

$$c_i \leq \sup_{z_j \in \mathcal{Y}^3, z_i' \in \mathcal{Y}^3} \left\|\widehat{\mathbf{M}}_M(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_p) - \widehat{\mathbf{M}}_M(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_p)\right\|_F\,,$$

so that

$$c_i \leq \frac{1}{p} \sup_{z_i \in \mathcal{Y}^3, z_i' \in \mathcal{Y}^3} \left(\sum_{a,b,c} \left(\varphi_a(y_1^{(i)})\varphi_b(y_2^{(i)})\varphi_c(y_3^{(i)}) - \varphi_a(y'^{(i)}_1)\varphi_b(y'^{(i)}_2)\varphi_c(y'^{(i)}_3)\right)^2\right)^{1/2}\,.$$

Eventually, we get that $c_i \leq \eta_3(\Phi_M)/p$. By McDiarmid's inequality [32], for all $u > 0$,

$$\mathbb{P}(\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F \geq \mathbb{E}\left[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F\right] + u) \leq \exp\left(-\frac{pu^2}{8\mathbb{T}_{\mathrm{mix}}\eta_3^2(\Phi_M)}\right)\,.$$

The following lemma may be deduced from [32].

Lemma E.2. *For any $a, b, c \in \{1, \ldots, M\}$,*

$$\mathbb{E}\left[\sum_{s=1}^{p} \frac{1}{p}[\varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}) - \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]]\right]^2 \leq \frac{4}{p\mathbb{G}_{\mathrm{ps}}}\mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3) - \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]\right]^2\,.$$

Proof. Notice that $(X_1, Y_1), (X_2, Y_2), \ldots$ is homogenous, irreducible, aperiodic and stationary Markov chain on $\mathcal{X} \times \mathcal{Y}$, whose stationary distribution is $\tilde{\pi}(x, \mathrm{d}y) := \pi_x \mu_x(\mathrm{d}y)$. Observe that its transition kernel $\tilde{\mathbf{Q}}$ satisfies, for all $x, x' \in \mathcal{X}$ and all $y, y' \in \mathcal{Y}$,

$$\tilde{\mathbf{Q}}(x, y; x', \mathrm{d}y') = \mathbf{Q}_\star(x, x')\mu_{x'}(\mathrm{d}y')\,.$$

The transition kernel $\tilde{\mathbf{Q}}$ can be viewed as an operator $\mathbb{Q}$ on the Hilbert space $\mathrm{L}^2(\tilde{\pi})$ defined, for all $f \in \mathrm{L}^2(\tilde{\pi})$, by:

$$(\mathbb{Q}f)(x, y) := \mathbb{E}_{\tilde{\mathbf{Q}}(x,y;\cdot,\cdot)}(f) = \sum_{x' \in \mathcal{X}} \mathbf{Q}_\star(x, x') \int_{\mathcal{Y}} f(x', y')\mu_{x'}(\mathrm{d}y')\,.$$

Note that $\mathbb{Q}f(x,y)$ does not depend on y. Set $E := \{f(x,y) \in \mathrm{L}^2(\tilde{\pi}) \ : \ f \text{ does not depend on } y\}$. The $\mathrm{L}^2(\tilde{\pi})$-self-adjoint operator defined, for all $f \in \mathrm{L}^2(\tilde{\pi})$, by

$$(\Pi_E f)(x,y) := \int_{\mathcal{Y}} f(x,y')\mu_x(\mathrm{d}y')\,,$$

is the orthogonal projection onto E. Since $\Pi_E \mathbb{Q} \Pi_E = \mathbb{Q}$, the set of nonzero eigenvalues of $\mathbb{Q}$ is exactly the set of nonzero eigenvalues of the K dimensional linear operator $\Pi_E \mathbb{Q} \Pi_E$. Eventually, note that the matrix of $\mathbb{Q}$ in the basis $((x,y) \mapsto \mathbb{1}_{x'=x})_{x' \in \mathcal{X}}$ is $\mathbf{Q}_\star$. Then, the pseudo spectral gap of $\mathbb{Q}$ is equal to $\mathbb{G}_{\mathrm{ps}}$ (the pseudo spectral gap of $\mathbf{Q}_\star$).

Furthermore, note the same analysis can be made for $(X_1, X_2, X_3, Z_1), (X_2, X_3, X_4, Z_2), \ldots$ and its pseudo spectral gap is the pseudo spectral gap of the Markov chain $(X_1, X_2, X_3), (X_2, X_3, X_4), \ldots$ which is $\mathbb{G}_{\mathrm{ps}}$. Indeed, the set of nonzero eigenvalues of the Markov chain $(X_1, X_2, X_3), (X_2, X_3, X_4), \ldots$ is equal to the set of nonzero eigenvalues of the Markov chain $X_1, X_2, \ldots$.

Eventually, set $g(X_s, X_{s+1}, X_{s+2}, Z_s) := (1/p)\varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$ and apply Theorem 3.1 in [32] to conclude the proof. □

Then,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F\right] &\leq \mathbb{E}\left[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_F^2\right]^{1/2}\,, \\
&\leq \mathbb{E}\left[\sum_{a,b,c}\left(\frac{1}{p}\sum_{s=1}^{p}\varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}) - \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]\right)^2\right]^{1/2}\,, \\
&\leq \left[\sum_{a,b,c}\mathbb{E}\left(\sum_{s=1}^{p}\frac{1}{p}\left\{\varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}) - \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]\right\}\right)^2\right]^{1/2}\,, \\
&\leq \frac{2}{\sqrt{p\mathbb{G}_{\mathrm{ps}}}}\left[\sum_{a,b,c}\mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3) - \mathbb{E}\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]^2\right]^{1/2}\,, \\
&\leq \left(\frac{2}{p\mathbb{G}_{\mathrm{ps}}}\right)^{1/2}\left[\mathbb{E}\sum_{a,b,c}(\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3) - \varphi_a(Y'_1)\varphi_b(Y'_2)\varphi_c(Y'_3))^2\right]^{1/2}\,, \\
&\leq \left(\frac{2\eta_3^2(\Phi_M)}{p\mathbb{G}_{\mathrm{ps}}}\right)^{1/2}\,,
\end{aligned}$$

using Jensen's inequality, Lemma E.2 and then $2\mathbb{E}[U - \mathbb{E}[U]]^2 \leq \mathbb{E}[U - U']^2$ where U is any real valued random variable with finite second moment and U' an independent copy of U. The proof is similar for $\mathbf{L}_M$, $\mathbf{N}_M$ and $\mathbf{P}_M$. □

F Proof of Theorem C.3

Preliminaries lemmas

Lemma F.1. *For all $b \in \{1, \ldots, M\}$,*

$$\mathbf{M}_M(.\,, b, .\,) = \mathbf{O}_M \mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star \mathfrak{Diag}[\mathbf{O}_M(b, .\,)]\mathbf{Q}_\star \mathbf{O}_M^\top\,.$$

Similarly, $\mathbf{P}_M = \mathbf{O}_M \mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star{}^2\mathbf{O}_M^\top$.

Proof. Let $a, c \in \{1, \dots, M\}^2$ and observe that:

$$\begin{aligned}
&(\mathbf{O}_M \mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star \mathfrak{Diag}[\mathbf{O}_M(b,.)]\mathbf{Q}_\star \mathbf{O}_M^\top)(a,c) \\
&\qquad = \sum_{(x_1,x_2,x_3)\in\mathcal{X}^3} \mathbf{O}_M(a,x_1)\pi(x_1)\mathbf{Q}_\star(x_1,x_2)\mathbf{O}_M(b,x_2)\mathbf{Q}_\star(x_2,x_3)\mathbf{O}_M(c,x_3)\,, \\
&\qquad = \sum_{(x_1,x_2,x_3)\in\mathcal{X}^3} \mathbb{E}\left[\varphi_a(Y_1)|X_1=x_1\right]\mathbb{P}(X_1=x_1)\mathbb{P}(X_2=x_2|X_1=x_1) \\
&\qquad\qquad\qquad \times \mathbb{E}\left[\varphi_b(Y_2)|X_2=x_2\right]\mathbb{P}(X_3=x_3|X_2=x_2)\mathbb{E}\left[\varphi_c(Y_3)|X_3=x_3\right]\,, \\
&\qquad = \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]\,.
\end{aligned}$$

Similarly,

$$\begin{aligned}
(\mathbf{O}_M \mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star{}^2 \mathbf{O}_M^\top)(a,c) &= \sum_{(x_1,x_2,x_3)\in\mathcal{X}^3} \mathbf{O}_M(a,x_1)\pi(x_1)\mathbf{Q}_\star(x_1,x_2)\mathbf{Q}_\star(x_2,x_3)\mathbf{O}_M(c,x_3)\,, \\
&= \sum_{(x_1,x_2,x_3)\in\mathcal{X}^3} \mathbb{E}\left[\varphi_a(Y_1)|X_1=x_1\right]\mathbb{P}(X_1=x_1)\mathbb{P}(X_2=x_2|X_1=x_1) \\
&\qquad\qquad \times \mathbb{P}(X_3=x_3|X_2=x_2)\mathbb{E}\left[\varphi_c(Y_3)|X_3=x_3\right]\,, \\
&= \mathbb{E}\left[\varphi_a(Y_1)\varphi_c(Y_3)\right]\,,
\end{aligned}$$

which concludes the proof. □

Lemma F.2. *Let* $\mathbf{U}$ *be any* $(M \times K)$ *matrix such that* $\mathbf{P}_M\mathbf{U}$ *has rank* K*. Then,*

- *for all* $b \in \{1, \dots, M\}$*,*

$$\mathbf{B}(b) := (\mathbf{P}_M\mathbf{U})^\dagger \mathbf{M}_M(.,b,.)\mathbf{U} = \mathbf{R}\mathfrak{Diag}[\mathbf{O}_M(b,.)]\mathbf{R}^{-1}\,,$$

  where $\mathbf{R}^{-1} := \mathbf{Q}_\star\mathbf{O}_M^\top\mathbf{U}$ *and* $(\mathbf{P}_M\mathbf{U})^\dagger := (\mathbf{U}^\top\mathbf{P}_M^\top\mathbf{P}_M\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{P}_M^\top$ *denotes the Moore-Penrose pseudoinverse of the matrix* $\mathbf{P}_M\mathbf{U}$ *;*

- $\mathbf{U}^\top\mathbf{P}_M\mathbf{U}$ *is invertible and, for all* $b \in \{1, \dots, M\}$*,*

$$\mathbf{B}(b) = (\mathbf{U}^\top\mathbf{P}_M\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{M}_M(.,b,.)\mathbf{U} = \mathbf{R}\mathfrak{Diag}[\mathbf{O}_M(b,.)]\mathbf{R}^{-1}\,.$$

Proof. Observe that $\mathbf{M}_M(.,b,.)\mathbf{U} = \mathbf{O}_M\mathfrak{Diag}[\pi^\star]\mathbf{Q}_\star\mathfrak{Diag}[\mathbf{O}_M(b,.)]\mathbf{R}^{-1} = \mathbf{P}_M\mathbf{U}\mathbf{R}\mathfrak{Diag}[\mathbf{O}_M(b,.)]\mathbf{R}^{-1}$ as claimed. □

Lemma F.3. *Assume that* $2\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\| < \sigma_K(\mathbf{P}_M)$*, then:*

(i)

$$\varepsilon_{\mathbf{P}_M} := \frac{\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M) - \|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|} < 1\,,$$

(ii)

$$\sigma_K(\widehat{\mathbf{P}}_M) \geq \Big[\frac{\sigma_K(\mathbf{P}_M) - \|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\Big]\sigma_K(\mathbf{P}_M) > \frac{\sigma_K(\mathbf{P}_M)}{2}\,,$$

(iii) $\sigma_K(\hat{\mathbf{U}}^\top\mathbf{U}) \geq (1 - \varepsilon^2_{\mathbf{P}_M})^{1/2}$*,*

(iv) $\sigma_K(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}}) \geq (1 - \varepsilon^2_{\mathbf{P}_M})\sigma_K(\mathbf{P}_M)$*,*

(v) for all $\alpha \in \mathbb{R}^K$ *and for all* $v \in \mathrm{Range}(\mathbf{P}_M)$*,* $\|\hat{\mathbf{U}}\alpha - v\|_2^2 \leq \|\alpha - \hat{\mathbf{U}}^\top v\|_2^2 + \varepsilon^2_{\mathbf{P}_M}\|v\|_2^2$*,*

(vi) if $3\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\| \leq \sigma_K(\mathbf{P}_M)$ *then:*

$$\sigma_K(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}}) \geq \frac{\sigma_K(\mathbf{P}_M)}{3}\,,$$

(vii)

$$\|(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}-(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}\|\leq\frac{\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)(1-\varepsilon^2_{\mathbf{P}_M})((1-\varepsilon^2_{\mathbf{P}_M})\sigma_K(\mathbf{P}_M)-\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|)}\,,$$
$$\leq 3.2\frac{\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma^2_K(\mathbf{P}_M)}\,.$$

Proof. See Lemma C.1 in [3] for the first five claims. The sixth claim follows from the fourth point and Theorem D.1. The seventh point follows from the fourth claim and Theorem D.2. □

Control of the observable operator

Claim (iv) in Lemma F.3 and Lemma F.2 ensure that, for all $b\in\{1,\dots,M\}$,

$$\tilde{\mathbf{B}}(b):=(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top\mathbf{M}_M(.\,,b,.\,)\hat{\mathbf{U}}=\tilde{\mathbf{R}}\mathfrak{Diag}[\mathbf{O}_M(b,.\,)]\tilde{\mathbf{R}}^{-1}\,,$$

where $\mathbf{R}^{-1}$ may be defined as

$$\tilde{\mathbf{R}}^{-1}:=\mathfrak{Diag}[(\|(\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}})^{-1}(.\,,1)\|_2,\dots,\|(\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}})^{-1}(.\,,K)\|_2)]\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}}\,.$$

Set $\Lambda:=\Theta^\top\hat{\mathbf{U}}^\top\mathbf{O}_M$ and for all $x\in\mathcal{X}$, $\tilde{\mathbf{C}}(x):=\sum_{b=1}^M(\hat{\mathbf{U}}\Theta)(b,x)\tilde{\mathbf{B}}(b)=\tilde{\mathbf{R}}\mathfrak{Diag}[\Lambda(x,.\,)]\tilde{\mathbf{R}}^{-1}$. Note that $\tilde{\mathbf{R}}$ has unit Euclidean norm columns:

$$\tilde{\mathbf{R}}=(\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}})^{-1}\,\mathfrak{Diag}[(\|(\mathbf{Q}^\star\mathbf{O}_M^\top\hat{\mathbf{U}})^{-1}(.\,,1)\|_2,\dots,\|(\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}})^{-1}(.\,,K)\|_2)]^{-1}\,,$$

corresponding to unit Euclidean norm eigenvectors of $\tilde{\mathbf{C}}(k)$.

Lemma F.4. *Assume that* $3\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|\leq\sigma_K(\mathbf{P}_M)$*, then, for all* $b\in\{1,\dots,M\}$*,*

$$\|\hat{\mathbf{B}}(b)-\tilde{\mathbf{B}}(b)\|\leq 3.2\frac{\|\mathbf{M}_M(.\,,b,.\,)\|}{\sigma_K(\mathbf{P}_M)}\Big[\frac{\|\widehat{\mathbf{M}}_M(.\,,b,.\,)-\mathbf{M}_M(.\,,b,.\,)\|}{\|\mathbf{M}_M(.\,,b,.\,)\|}+\frac{\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\Big]\,,$$

and for all $x\in\mathcal{X}$*,*

$$\|\hat{\mathbf{C}}(x)-\tilde{\mathbf{C}}(x)\|\leq 3.2\frac{\|\mathbf{M}_M\|_{\infty,2}}{\sigma_K(\mathbf{P}_M)}\Big[\frac{\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}}{\|\mathbf{M}_M\|_{\infty,2}}+\frac{\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\Big]\,.$$

Proof. Observe that:

$$\begin{aligned}\|\hat{\mathbf{B}}(b)-\tilde{\mathbf{B}}(b)\|\leq&\|(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top\widehat{\mathbf{M}}_M(.\,,b,.\,)\hat{\mathbf{U}}-(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top\mathbf{M}_M(.\,,b,.\,)\hat{\mathbf{U}}\|\\&+\|(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top\mathbf{M}_M(.\,,b,.\,)\hat{\mathbf{U}}-(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top\mathbf{M}_M(.\,,b,.\,)\hat{\mathbf{U}}\|\,,\\\leq&\|\hat{\mathbf{U}}^\top(\widehat{\mathbf{M}}_M(.\,,b,.\,)-\mathbf{M}_M(.\,,b,.\,))\hat{\mathbf{U}}\|\|(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\|\\&+\|(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}-(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\|\|\hat{\mathbf{U}}^\top\mathbf{M}_M(.\,,b,.\,)\hat{\mathbf{U}}\|\,,\\\leq&\|\widehat{\mathbf{M}}_M(.\,,b,.\,)-\mathbf{M}_M(.\,,b,.\,)\|\sigma_K^{-1}(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})\\&+\|\mathbf{M}_M(.\,,b,.\,)\|\|(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}-(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}\|\,.\end{aligned}$$

By claims (vi) and (vii) of Lemma F.3, $3\sigma_K(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})\geq\sigma_K(\mathbf{P}_M)$ and $\|(\hat{\mathbf{U}}^\top\widehat{\mathbf{P}}_M\hat{\mathbf{U}})^{-1}-(\hat{\mathbf{U}}^\top\mathbf{P}_M\hat{\mathbf{U}})^{-1}\|\leq 3.2\frac{\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma^2_K(\mathbf{P}_M)}$. Replacing $\mathbf{M}_M(.\,,b,.\,)$ by $\sum_{b=1}^M(\hat{\mathbf{U}}\Theta)(b,k)\mathbf{M}_M(.\,,b,.\,)$ yields the same result for $\|\hat{\mathbf{C}}(x)-\tilde{\mathbf{C}}(x)\|$. □

Lemma F.5. *Assume that* $2\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|<\sigma_K(\mathbf{P}_M)$*, then,*

(i)

$$\kappa(\tilde{\mathbf{R}}):=\|\tilde{\mathbf{R}}\|\|\tilde{\mathbf{R}}^{-1}\|\leq\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top\hat{\mathbf{U}})\leq\frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{1-\varepsilon^2_{\mathbf{P}_M}}\,,$$

(ii)

$$\mathrm{sv}_{\mathbf{C}(1)}(\hat{\mathbf{C}}(1)) \leq \kappa(\tilde{\mathbf{R}})\|\hat{\mathbf{C}}(1)-\tilde{\mathbf{C}}(1)\| \leq \frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{1-\varepsilon^2_{\mathbf{P}_M}}\|\hat{\mathbf{C}}(1)-\tilde{\mathbf{C}}(1)\|\,,$$

where $\mathrm{sv}_{\mathbf{C}(1)}(\hat{\mathbf{C}}(1)) := \max_{x_1\in\mathcal{X}}\min_{x_2\in\mathcal{X}}\left|\hat{\lambda}(1,x_1)-\lambda(1,x_2)\right|$.

(iii) If in addition,

$$\frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{1-\varepsilon^2_{\mathbf{P}_M}}\|\hat{\mathbf{C}}(1)-\tilde{\mathbf{C}}(1)\| < \min_{x,x'\in\mathcal{X}}|\Lambda(1,x)-\Lambda(1,x')|/2\,,$$

then $\hat{\mathbf{C}}(1)$ *has* K *distinct real eigenvalues and:*

$$\mathrm{md}(\mathbf{C}(1),\hat{\mathbf{C}}(1)) \leq \frac{\kappa^2(\mathbf{Q}_\star\mathbf{O}_M^\top)}{1-\varepsilon^2_{\mathbf{P}_M}}\|\hat{\mathbf{C}}(1)-\tilde{\mathbf{C}}(1)\|\,,$$

where $\mathrm{md}(\mathbf{C}(1),\hat{\mathbf{C}}(1)) := \min_{\tau\in\mathcal{S}_K}\left\{\max_{x\in\mathcal{X}}\left|\hat{\Lambda}(1,\tau(x))-\Lambda(1,x)\right|\right\}$.

Proof. Observe that $\mathbf{U}$ is an orthonormal basis of range of $\mathbf{O}_M$. The first point follows from claim (iii) of Lemma F.3. The second point is derived from Theorem D.3 and the first point. The remark following Theorem D.3 proves the last point. □

Control of the spectra

Lemma F.6. *For any* $0<\delta<1$,

$$\mathbb{P}\Big[\forall x,x_1\neq x_2\,,\ |\Lambda(x,x_1)-\Lambda(x,x_2)| \geq \frac{2\delta(1-\varepsilon^2_{\mathbf{P}_M})^{1/2}}{\sqrt{e}K^{5/2}(K-1)}\gamma(\mathbf{O}_M)\Big] \geq 1-\delta\,.$$

Furthermore:

$$\mathbb{P}\Big[\|\Lambda\|_\infty \geq \frac{1+\sqrt{2\log(K^2/\delta)}}{\sqrt{K}}\|\mathbf{O}_M\|_{2,\infty}\Big] \leq \delta\,.$$

Proof. Observe that:

$$\begin{aligned}\Lambda(x,x_1)-\Lambda(x,x_2) &= \langle\Theta(.\,,x),(\hat{\mathbf{U}}^\top\mathbf{O}_M)(.\,,x_1)-(\hat{\mathbf{U}}^\top\mathbf{O}_M)(.\,,x_2)\rangle\\ &= \langle\Theta(.\,,x),\hat{\mathbf{U}}^\top(\mathbf{O}_M(.\,,x_1)-\mathbf{O}_M(.\,,x_2))\rangle\,.\end{aligned}$$

Furthermore, from (iii) in Lemma F.3, we get that:

$$\|\hat{\mathbf{U}}^\top(\mathbf{O}_M(.\,,x_1)-\mathbf{O}_M(.\,,x_2))\|_2 \geq (1-\varepsilon^2_{\mathbf{P}_M})^{1/2}\|\mathbf{O}_M(.\,,x_1)-\mathbf{O}_M(.\,,x_2)\|_2 \geq (1-\varepsilon^2_{\mathbf{P}_M})^{1/2}\gamma(\mathbf{O}_M)\,.$$

Similarly, note that:

$$\|\Lambda\|_\infty = \max_{x,x'}|\langle\Theta(.\,,x),\hat{\mathbf{U}}^\top\mathbf{O}_M(.\,,x')\rangle|\,,$$

and $\|\hat{\mathbf{U}}^\top\mathbf{O}_M(.\,,x')\|_2 \leq \|\mathbf{O}_M(.\,,x')\|_2 \leq \|\mathbf{O}_M\|_{2,\infty}$. For sake of readability, we borrow the result of Lemma F.2 and the argument of Lemma C.6 in [3] to conclude. □

Perturbation of simultaneously diagonalizable matrices

Lemma F.7. *If* $3\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\| \leq \sigma_K(\mathbf{P}_M)$ *and:*

$$8.2K^{5/2}(K-1)\frac{\kappa^2(\mathbf{Q}\mathbf{O}_M^\top)}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\Big[\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}+\frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\Big] < 1\,, \tag{33}$$

$$43.4K^{4}(K-1)\frac{\kappa^4(\mathbf{Q}\mathbf{O}_M^\top)}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\Big[\|\widehat{\mathbf{M}}_M-\mathbf{M}_M\|_{\infty,2}+\frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M-\mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\Big] \leq 1\,, \tag{34}$$

and for all $x,x_1\neq x_2$,

$$|\Lambda(x,x_1)-\Lambda(x,x_2)| \geq \frac{\sqrt{3}\delta}{\sqrt{e}K^{5/2}(K-1)}\gamma(\mathbf{O}_M)\,,$$

and:

$$\|\Lambda\|_\infty \leq \frac{1+\sqrt{2\log(K^2/\delta)}}{\sqrt{K}}\|\mathbf{O}_M\|_{2,\infty}\,,$$

then there exists $\tau \in \mathcal{S}_K$ *such that for all* $x \in \mathcal{X}$*:*

$$\|\Lambda(.\,,x) - \hat{\Lambda}(.\,,\tau(x))\|_\infty \leq \left[13\frac{\kappa^2(\mathbf{Q}\mathbf{O}_M^\top)}{\sigma_K(\mathbf{P}_M)} + 116K^{7/2}(K-1)\left\{1+\left(2\log(K^2/\delta)\right)^{1/2}\right\}\right.$$
$$\left.\times\,\frac{\kappa^6(\mathbf{Q}\mathbf{O}_M^\top)\|\mathbf{O}_M\|_{2,\infty}}{\delta\gamma(\mathbf{O}_M)\sigma_K(\mathbf{P}_M)}\right] \times \left[\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_{\infty,2} + \frac{\|\mathbf{M}_M\|_{\infty,2}\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\right].$$

Proof. Note $\varepsilon_{\mathbf{P}_M} \leq 1/2$. Invoke the last part of Claim 4 of Lemma C.4 in [3] with $\gamma_A \leftarrow \frac{\sqrt{3}\delta}{\sqrt{e}K^{\frac{5}{2}}(K-1)}\gamma(\mathbf{O}_M)$, $\kappa(R) \leftarrow \frac{4\kappa^2(\mathbf{Q}\mathbf{O}_M^\top)}{3}$, $\|\tilde{R}\|_2^2 \leftarrow \frac{4\kappa^2(\mathbf{Q}\mathbf{O}_M^\top)}{3}$, $\epsilon_A \leftarrow 3.2\frac{\|\mathbf{M}_M\|_{\infty,2}}{\sigma_K(\mathbf{P}_M)}\left[\frac{\|\widehat{\mathbf{M}}_M - \mathbf{M}_M\|_{\infty,2}}{\|\mathbf{M}_M\|_{\infty,2}} + \frac{\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{\sigma_K(\mathbf{P}_M)}\right]$ and $\lambda_{\max} \leftarrow \frac{1+\sqrt{2\log(K^2/\delta)}}{\sqrt{K}}\|\mathbf{O}_M\|_{2,\infty}$. Observe that (33) agrees with $\varepsilon_3 < 1/2$ and (34) agrees with $\varepsilon_4 \leq 1/2$. □

Since $\Theta^\top$ is an isometry, observe that:

$$\|\hat{\mathbf{U}}^\top\mathbf{O}_M(.\,,x) - \Theta\hat{\Lambda}(.\,,\tau(x))\|_2 = \|\Lambda(.\,,x) - \hat{\Lambda}(.\,,\tau(x))\|_2 \leq \sqrt{K}\|\Lambda(.\,,x) - \hat{\Lambda}(.\,,\tau(x))\|_\infty\,.$$

Claim (v) in Lemma F.3 (with $\alpha = \Theta\hat{\Lambda}(.\,,\tau(x))$ and $v = \mathbf{O}_M(.\,,x)$) give

$$\begin{aligned}\|\mathbf{O}_M(.\,,x) - \widehat{\mathbf{O}}_M(.\,,\tau(x))\|_2 &\leq \|\hat{\mathbf{U}}^\top\mathbf{O}_M(.\,,x) - \Theta\hat{\Lambda}(.\,,\tau(x))\|_2 + \frac{3\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{2\sigma_K(\mathbf{P}_M)}\|\mathbf{O}_M(.\,,x)\|_2 \\ &\leq \sqrt{K}\|\Lambda(.\,,x) - \hat{\Lambda}(.\,,\tau(x))\|_\infty + \frac{3\|\widehat{\mathbf{P}}_M - \mathbf{P}_M\|}{2\sigma_K(\mathbf{P}_M)}\|\mathbf{O}_M(.\,,x)\|_2.\end{aligned}$$

Theorem C.3 follows from Lemma F.7.

References

[1] G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

[2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. *Twenty-Fifth Annual Conference on Learning Theory*, 23:33.1–33.34, 2012.

[4] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation*. Wiley, 2001.

[5] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, 2012.

[6] L.E. Baum, T.P. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[7] O. Cappé. Online sequential Monte Carlo EM algorithm. *Proc. 15th IEEE Workshop on Statistical Signal Processing*, pages 37–40, 2009.

[8] O. Cappé. Online EM algorithm for hidden Markov models. *J. Comput. Graph. Statist.*, 20:728–749, 2011.

[9] O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[10] Y. De Castro, E. Gassiat, and C. Lacour. Minimax adaptative estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 17:1–43, 2016.

[11] P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[12] P. Del Moral, A. Doucet, and S. Singh. A backward particle interpretation of Feynman-Kac formulae. *ESAIM: M2AN*, 44(5):947–975, 2010.

[13] R. Douc, A. Garivier, É. Moulines, and J. Olsson. Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.*, 21(6):2109–2145, 2011.

[14] R. Douc, É. Moulines, and D. Stoffer. *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Chapman & Hall, 2013.

[15] A. Doucet, N. De Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[16] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, 10:197–208, 2000.

[17] C. Dubarry and S. Le Corff. Non-asymptotic deviation inequalities for smoothed additive functionals in nonlinear state-space models. *Bernoulli*, 19(5B):2222–2249, 2013.

[18] E. Even-Dar, S.M. Kakade, and Y. Mansour. The value of observation for monitoring dynamic systems. *International Joint Conference on Artificial Intelligence*, pages 2474–2479, 2007.

[19] É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26:61–71, 2016.

[20] É. Gassiat and J. Rousseau. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016.

[21] S. J. Godsill, A. Doucet, and M. West. Monte Carlo smoothing for non-linear time series. *J. Am. Statist. Assoc.*, 50:438–449, 2004.

[22] D. Hsu, S. M. Kakade, and T. Zhang. A spectral algorithm for learning hidden Markov models. *J. Comput. System Sci.*, 78(5):1460–1480, 2012.

[23] M. Hurzeler and H.R. Kusch. Monte Carlo approximations for general state-space models. *J. Comput. Graph. Statist.*, 7:175–193, 1998.

[24] N. Kantas, A. Doucet, S.S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30:328–351, 2015.

[25] G. Kitagawa. Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, 1:1–25, 1996.

[26] M.F. Lambert, J.P. Whiting, and A.V. Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences*, 7(5):652–667, 2003.

[27] S. Le Corff and G. Fort. Online expecta- tion maximization based algorithms for inference in hidden markov models. *Electron. J. Stat.*, 7:763–792, 2013.

[28] F. Lefevre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speech and Language*, 17(2-3):113–136, 2003.

[29] L. Lehéricy. Order estimation for non-parametric hidden Markov models. *arXiv preprint arXiv:1606.00622*, 2016.

[30] Y. Meyer. *Wavelets and operators*, volume 37 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1992. Translated from the 1990 French original by D. H. Salinger.

[31] J. Olsson and J. Westerborn. Efficient particle-based online smoothing in general hidden Markov models: the PaRIS algorithm. *To appear in Bernoulli, arXiv:1412.7550*, 2016.

[32] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

[33] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–285, 1989.

[34] S. Sarkka. *Bayesian filtering and smoothing*. Cambridge University Press, 2013.

[35] G. W. Stewart and J.-G. Sun. *Matrix perturbation theory*. Academic press, 1990.

[36] V.B. Tadic. Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models. *IEEE Transactions on Information Theory*, 56(12), 2010.

[37] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT Press, Cambridge, 2005.

[38] É. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9:717–752, 2015.

[39] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[40] C. Yau, O. Papaspiliopoulos, G.O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(1):37–57, 2014.