About the posterior distribution in hidden Markov models with unknown number of states

ELISABETH GASSIAT¹ and JUDITH ROUSSEAU²

¹Laboratoire de Mathématiques d’Orsay UMR 8628, Université Paris-Sud, Bâtiment 425, 91405 Orsay-Cédex, France. E-mail: elisabeth.gassiat@math.u-psud.fr
²CREST-ENSAE, 3 avenue Pierre Larousse, 92245 Malakoff Cedex, France. E-mail: rousseau@ceremade.dauphine.fr

We consider finite state space stationary hidden Markov models (HMMs) in the situation where the number of hidden states is unknown. We provide a frequentist asymptotic evaluation of Bayesian analysis methods. Our main result gives posterior concentration rates for the marginal densities, that is for the density of a fixed number of consecutive observations. Using conditions on the prior, we are then able to define a consistent Bayesian estimator of the number of hidden states. It is known that the likelihood ratio test statistic for overfitted HMMs has a nonstandard behaviour and is unbounded. Our conditions on the prior may be seen as a way to penalize parameters to avoid this phenomenon. Inference of parameters is a much more difficult task than inference of marginal densities, we still provide a precise description of the situation when the observations are i.i.d. and we allow for 2 possible hidden states.

Keywords: Bayesian statistics; hidden Markov models; number of components; order selection; posterior distribution

1. Introduction

Finite state space hidden Markov models (which will be shortened to HMMs throughout the paper) are stochastic processes \((X_j, Y_j)_{j \geq 1}\) where \((X_j)_{j \geq 1}\) is a Markov chain living in a finite state space \(X\) and conditionally on \((X_j)_{j \geq 1}\) the \(Y_j\)'s are independent with a distribution depending only on \(X_j\) and living in \(Y\). HMMs are useful tools to model time series where the observed phenomenon is driven by a latent Markov chain. They have been used successfully in a variety of applications, the books MacDonald and Zucchini [14], Zucchini and MacDonald [23] and Cappé et al. [2] provide several examples of applications of HMMs and give a recent (for the latter) state of the art in the statistical analysis of HMMs. Finite state space HMMs may also be seen as a dynamic extension of finite mixture models and may be used to do unsupervised clustering. The hidden states often have a practical interpretation in the modelling of the underlying phenomenon. It is thus of importance to be able to infer both the number of hidden states (which we call the order of the HMM) from the data, and the associated parameters.

The aim of this paper is to provide a frequentist asymptotic analysis of Bayesian methods used for statistical inference in finite state space HMMs when the order is unknown. Let us first review what is known on the subject and important questions that still stay unsolved.
In the frequentist literature, penalized likelihood methods have been proposed to estimate the order of a HMM, using for instance Bayesian information criteria (BIC for short). These methods were applied for instance in Leroux and Putterman [13], Rydén et al. [21], but without theoretical consistency results. Later, it has been observed that the likelihood ratio statistics is unbounded, in the very simple situation where one wants to test between 1 or 2 hidden states, see Gassiat and Kéribin [9]. The question whether BIC penalized likelihood methods lead to consistent order estimation stayed open. Using tools borrowed from information theory, it has been possible to calibrate heavier penalties in maximum likelihood methods to obtain consistent estimators of the order, see Gassiat and Boucheron [7], Chambaz et al. [3]. The use of penalized marginal pseudo likelihood was also proved to lead to weakly consistent estimators by Gassiat [6].

On the Bayesian side, various methods were proposed to deal with an unknown number of hidden states, but no frequentist theoretical result exists for these methods. Notice though that, if the number of states is known, de Gunst and Shcherbakova [4] obtain a Bernstein–von Mises theorem for the posterior distribution, under additional (but usual) regularity conditions. When the order is unknown, reversible jump methods have been built, leading to satisfactory results on simulation and real data, see Boys and Henderson [1], Green and Richardson [12], Robert et al. [19], Spezia [22]. The ideas of variational Bayesian methods were developed in McGrory and Titterington [15]. Recently, one of the authors proposed a frequentist asymptotic analysis of the posterior distribution for overfitted mixtures when the observations are i.i.d., see Rousseau and Mengersen [20]. In this paper, it is proved that one may choose the prior in such a way that extra components are emptied, or in such a way that extra components merge with true ones. More precisely, if a Dirichlet prior \( D(\alpha_1, \ldots, \alpha_k) \) is considered on the \( k \) weights of the mixture components, small values of the \( \alpha_j \)'s imply that the posterior distribution will tend to empty the extra components of the mixture when the true distribution has a smaller number, say \( k_0 < k \) of true components. One aim of our paper is to understand if such an analysis may be extended to HMMs.

As is well known in the statistical analysis of overfitted finite mixtures, the difficulty of the problem comes from the non-identifiability of the parameters. But what is specific to HMMs is that the non-identifiability of the parameters leads to the fact that neighbourhoods of the “true” parameter values contain transition matrices arbitrarily close to non-ergodic transition matrices. To understand this on a simple example, just consider the case of HMMs with two hidden states, say \( p \) is the probability of going from state 1 to state 2 and \( q \) the probability of going from state 2 to state 1. If the observations are in fact independently distributed, their distribution may be seen as a HMM with two hidden states where \( q = 1 - p \). Neighbourhoods of the “true” values \((p, 1 - p)\) contain parameters such that \( p \) is small or \( 1 - p \) is small, leading to hidden Markov chains having mixing coefficients very close to 1. Imposing a prior condition such as \( \delta \leq p \leq 1 - \delta \) for some \( \delta > 0 \) is not satisfactory.

Our first main result Theorem 1 gives concentration rates for the posterior distribution of the marginal densities of a fixed number of consecutive observations. First, under mild assumptions on the densities and the prior, we obtain the asymptotic posterior concentration rate \( \sqrt{n} \), \( n \) the number of observations, up to a \( \log n \) factor, when the loss function is the \( L_1 \) norm between densities multiplied by some function of the ergodicity coefficient of the hidden Markov chain. Then, with more stringent assumptions on the prior, we give posterior concentration rates for the marginal densities in \( L_1 \) norm only (without the ergodicity coefficient). For instance, consider a finite state space HMM, with \( k \) states and with independent Dirichlet prior distributions
2. Finite state space hidden Markov models

2.1. Model and notations

Recall that finite state space HMMs model pairs \((X_i, Y_i)_{i \geq 1}\) where \((X_i)_{i \geq 1}\) is the unobserved Markov chain living on a finite state space \(\mathcal{X} = \{1, \ldots, k\}\) and the observations \((Y_i)_{i \geq 1}\) are conditionally independent given the \((X_i)_{i \geq 1}\). The observations take value in \(\mathcal{Y}\), which is assumed to be a Polish space endowed with its \(\sigma\)-field. Throughout the paper, we denote \(x_{1:n} = (x_1, \ldots, x_n)\).

The hidden Markov chain \((X_i)_{i \geq 1}\) has a Markov transition matrix \(Q = (q_{ij})_{1 \leq i, j \leq k}\). The conditional distribution of \(Y_i\) given \(X_i\) has a density with respect to some given measure \(\nu\) on \(\mathcal{Y}\). We denote by \(g_j(y)\), \(j = 1, \ldots, k\), the conditional density of \(Y_i\) given \(X_i = j\). Here, \(\gamma_j \in \Gamma \subset \mathbb{R}^d\) for \(j = 1, \ldots, k\), the \(\gamma_j\)'s are called the emission parameters. In the following, we parametrize the transition matrices on \(\{1, \ldots, k\}\) as \((q_{ij})_{1 \leq i \leq k, 1 \leq j \leq k - 1}\) (implying that \(q_{ik} = 1 - \sum_{j=1}^{k-1} q_{ij}\) for all \(i \leq k\)) and we denote by \(\Delta_k\) the set of probability mass functions \(\Delta_k = \{(u_1, \ldots, u_{k-1}) : u_1 \geq 0, \ldots, u_{k-1} \geq 0, \sum_{i=1}^{k-1} u_i \leq 1\}\). We shall also use the set of positive probability mass functions \(\Delta_0^k = \{(u_1, \ldots, u_{k-1}) : u_1 > 0, \ldots, u_{k-1} > 0, \sum_{i=1}^{k-1} u_i < 1\}\). Thus, we may denote the overall parameter by \(\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1; \gamma_1, \ldots, \gamma_k) \in \Theta_k\) where \(\Theta_k = \Delta_0^k \times \Gamma^k\). To alleviate notations, we will write \(\theta = (Q; \gamma_1, \ldots, \gamma_k)\), where \(Q = (q_{ij})_{1 \leq i, j \leq k}\), \(q_{ik} = 1 - \sum_{j=1}^{k-1} q_{ij}\) for all \(i \leq k\).
Throughout the paper, $\nabla_{\theta} h$ denotes the gradient vector of the function $h$ when considered as a function of $\theta$, and $D^i_{\theta} h$ its $i$th derivative operator with respect to $\theta$, for $i \geq 1$. We denote by $B_d(\gamma, \epsilon)$ the $d$ dimensional ball centered at $\gamma$ with radius $\epsilon$, when $\gamma \in \mathbb{R}^d$. The notation $a_n \gtrsim b_n$ means that $a_n$ is larger than $b_n$ up to a positive constant that is fixed throughout.

Any Markov chain on a finite state space with transition matrix $Q$ admits a stationary distribution which we denote by $\mu_Q$, if it admits more than one we choose one of them. Then for any finite state space Markov chain with transition matrix $Q$ it is possible to define real numbers $\rho_Q \geq 1$ such that, for any integer $m$, any $j \leq k$

$$
\sum_{j=1}^{k} \left| \left( Q^m \right)_{ij} - \mu_Q(j) \right| \leq \rho_Q^{-m}, \quad \rho_Q = \left( 1 - \sum_{j=1}^{k} \min_{1 \leq i \leq k} q_{ij} \right)^{-1},
$$

where $Q^m$ is the $m$-step transition matrix of the Markov chain. If $\rho_Q > 1$, the Markov chain $(X_n)_{n \geq 1}$ is uniformly geometrically ergodic and $\mu_Q$ is its unique stationary distribution. In the following, we shall also denote $\mu_{Q}$ and $\rho_{Q}$ when $\theta = (Q; \gamma_1, \ldots, \gamma_k)$.

We write $P_{\theta}$ for the probability distribution of the stationary HMM $(X_j, Y_j)_{j \geq 1}$ with parameter $\theta$. That is, for any integer $n$, any set $A$ in the Borel $\sigma$-field of $X^n \times Y^n$:

$$
P_{\theta}((X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in A) = \sum_{x_1, \ldots, x_n=1}^{k} \int_{Y^n} \mathbb{1}_A(x_1; n, y_1; n) \mu_Q(x_1) \prod_{i=1}^{n-1} q_{x_i, x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}} (y_i) \nu(dy_1) \cdots \nu(dy_n).
$$

Thus for any integer $n$, under $P_{\theta}$, $Y_{1:n} = (Y_1, \ldots, Y_n)$ has a probability density with respect to $\nu(dy_1) \cdots \nu(dy_n)$ equal to

$$
f_{n,\theta}(y_1, \ldots, y_n) = \sum_{x_1, \ldots, x_n=1}^{k} \mu_Q(x_1) \prod_{i=1}^{n-1} q_{x_i, x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}} (y_i).
$$

We note $E_{\theta}$ for the expectation under $P_{\theta}$.

We denote $\Pi_k$ the prior distribution on $\Theta_k$. As is often the case in Bayesian analysis of HMMs, instead of computing the stationary distribution $\mu_Q$ of the hidden Markov chain with transition matrix $Q$, we consider a probability distribution $\pi_X$ on the unobserved initial state $X_0$. Denote $\ell_n(\theta, x_0)$ the log-likelihood starting from $x_0$, for all $x_0 \in \{1, \ldots, k\}$, we have

$$
\ell_n(\theta, x_0) = \log \left[ \sum_{x_1, \ldots, x_n=1}^{k} \prod_{i=0}^{n-1} q_{x_i, x_{i+1}} \prod_{i=1}^{n} g_{\gamma_{x_i}} (y_i) \right].
$$

The log-likelihood starting from a probability distribution $\pi_X$ on $\mathcal{X}$ is then given by

$$
\log \left[ \sum_{x_0=1}^{k} e^{\ell_n(\theta, x_0)} \pi_X(x_0) \right].
$$

This may also be interpreted as taking a prior $\Pi = \Pi_k \otimes \pi_X$ over
\[ \Theta_k \times \{1, \ldots, k\}. \] The posterior distribution can then be written as

\[
P_{\Pi}(A|Y_{1:n}) = \frac{\sum_{x_0=1}^{k} \int_{A} e^{t_{\theta}(\theta, x_0)} \Pi_k(d\theta) \pi_{\chi}(x_0)}{\sum_{x_0=1}^{k} \int_{\Theta} e^{t_{\theta}(\theta, x_0)} \Pi_k(d\theta) \pi_{\chi}(x_0)}
\]  

(4)

for any Borel set \( A \subset \Theta_k \).

Let \( \mathcal{M}_k \) be the set of all possible probability distributions \( P_\theta \) for all \( \theta \in \Theta_k \). We say that the HMM \( P_\theta \) has order \( k_0 \) if the probability distribution of \( (Y_n)_{n \geq 1} \) under \( P_\theta \) is in \( \mathcal{M}_{k_0} \) and not in \( \mathcal{M}_k \) for all \( k < k_0 \). Notice that a HMM of order \( k_0 \) may be represented as a HMM of order \( k \) for any \( k > k_0 \). Indeed, let \( Q^0 \) be a \( k_0 \times k_0 \) transition matrix, and \( (\gamma_1^0, \ldots, \gamma_{k_0}^0) \in \Gamma^{k_0} \) be parameters that define a HMM of order \( k_0 \). Then, \( \theta = (Q; \gamma_1^0, \ldots, \gamma_{k_0}^0) \in \Theta_k \) with \( Q = (q_{ij}, 1 \leq i, j \leq k) \) such that:

\[
q_{ij} = q_{ij}^0, \quad i, j < k_0,
\]

\[
q_{ij} = q_{k_0j}^0, \quad i \geq k_0, j < k_0,
\]

\[
\sum_{l=k_0}^{k} q_{il} = q_{i0}^0, \quad i \leq k_0, \quad \text{and} \quad \sum_{l=k_0}^{k} q_{il} = q_{0k}^0, \quad i \geq k_0
\]

(5)
gives \( P_{\theta} = P_{\theta_0} \). Indeed, let \( (X_n)_{n \geq 1} \) be a Markov chain on \( \{1, \ldots, k\} \) with transition matrix \( Q \). Let \( Z \) be the function from \( \{1, \ldots, k\} \) to \( \{1, \ldots, k_0\} \) defined by \( Z(x) = x \) if \( x \leq k_0 \) and \( Z(x) = k_0 \) if \( x \geq k_0 \). Then \( (Z(X_n))_{n \geq 1} \) is a Markov chain on \( \{1, \ldots, k_0\} \) with transition matrix \( Q^0 \).

### 2.2. Posterior convergence rates for the finite marginal densities

Let \( \theta_0 = (Q^0; \gamma_1^0, \ldots, \gamma_{k_0}^0) \in \Theta_{k_0}, Q^0 = (q_{ij}^0)_{1 \leq i \leq k_0, 1 \leq j \leq k_0}, \) be the parameter of a HMM of order \( k_0 \leq k \). We now assume that \( P_{\theta_0} \) is the distribution of the observations. In this section, we fix an integer \( l \) and study the posterior distribution of the density of \( l \) consecutive observations, that is \( f_{l, \theta}, \) given by (3) with \( n = l \). We study the posterior concentration rate around \( f_{l, \theta_0} \) in terms of the \( L_1 \) loss function, when \( P_{\theta_0} \) is possibly of order \( k_0 < k \). In this case, Theorem 2.1 of de Gunst and Shcherbakova [4] does not apply and there is no result in the literature about the frequentist asymptotic properties of the posterior distribution. The interesting and difficult feature of this case is that even though \( \theta_0 \) is parameterized as an ergodic Markov chain \( Q^0 \) with \( k \) states and some identical emission parameters as described in (5), \( f_{l, \theta_0} \) can be approached by marginals \( f_{l, \rho_0 \theta} \) for which \( \rho_0 \theta \) is arbitrarily close to 1, which deteriorates the posterior concentration rate, see Theorem 1.

Let \( \pi(u_1, \ldots, u_{k-1}) \) be a prior density with respect to the Lebesgue measure on \( \Delta_k \), and let \( \omega(\gamma) \) be a prior density on \( \Gamma \) (with respect to the Lebesgue measure on \( \mathbb{R}^d \)). We consider prior distributions such that the rows of the transitions matrix \( Q \) are independently distributed from \( \pi \) and independent of the component parameters \( \gamma_i, i = 1, \ldots, k \), which are independently distributed from \( \omega \). Hence, the prior density of \( \Pi_k \) (with respect to the Lebesgue measure) is equal to \( \pi_k = \pi^{\otimes k} \otimes \omega^{\otimes k} \). We still denote by \( \pi_{\chi} \) a probability on \( \{1, \ldots, k\} \), we assume that \( \pi_{\chi}(x) > 0 \) for all \( x \in \{1, \ldots, k\} \) and set \( \Pi = \Pi_k \otimes \pi_{\chi} \). We shall use the following assumptions.
\[ a_{ij}^0 > 0, \quad 1 \leq i \leq k_0, \quad 1 \leq j \leq k_0. \]

**A0** The function \( \gamma \mapsto g_\gamma(y) \) is twice continuously differentiable in \( \Gamma \), and for any \( \gamma \in \Gamma \), there exists \( \epsilon > 0 \) such that
\[
\int \sup_{\gamma' \in B_d(\gamma, \epsilon)} \| \nabla_\gamma \log g_{\gamma'}(y) \| g_{\gamma'}(y) \nu(dy) < +\infty,
\]
\[
\int \sup_{\gamma' \in B_d(\gamma, \epsilon)} \| D^2 \gamma \log g_{\gamma'}(y) \| g_{\gamma'}(y) \nu(dy) < +\infty.
\]

**A2** There exist \( a > 0 \) and \( b > 0 \) such that
\[
\sup_{\| \gamma \| \leq nb} \| \nabla_\gamma g_{\gamma}(y) \| \leq na.
\]

**A3** \( \pi \) is continuous and positive on \( \Delta^0_k \), and there exists \( C, \alpha_1 > 0, \ldots, \alpha_k > 0 \) such that (Dirichlet type priors):
\[
\forall (u_1, \ldots, u_{k-1}) \in \Delta^0_k, \quad u_k = 1 - \sum_{i=1}^{k-1} u_i, \quad 0 < \pi(u_1, \ldots, u_{k-1}) \leq Cu_1^{\alpha_1-1} \cdots u_k^{\alpha_k-1}
\]
and \( \omega \) is continuous and positive on \( \Gamma \) and satisfies
\[
\int_{\| x \| \geq nb} \omega(x) \, dx = o\left(n^{-k(k-1+d)/2}\right), \quad (6)
\]
with \( b \) defined in assumption A2.

We will alternatively replace A3 by

**A3bis** \( \pi \) is continuous and positive on \( \Delta^0_k \), and there exists \( C \) such that (exponential type priors):
\[
\forall (u_1, \ldots, u_{k-1}) \in \Delta^0_k, \quad u_k = 1 - \sum_{i=1}^{k-1} u_i, \quad 0 < \pi(u_1, \ldots, u_{k-1}) \leq C \exp(-C/u_1) \cdots \exp(-C/u_k)
\]
and \( \omega \) is continuous and positive on \( \Gamma \) and satisfies (6).

**Theorem 1.** Assume A0–A3. Then, there exists \( K \) large enough such that
\[
P_{\Pi} \left[ \theta : \| f_{l,\theta} - f_{l,\theta_0} \|_1 (\rho_{\theta} - 1) \geq K \sqrt{\frac{\log n}{n}} \right] = o_{\mathbb{P}_{\theta_0}}(1), \quad (7)
\]
where $\rho_0 = (1 - \sum_{j=1}^k \inf_{1 \leq i \leq k} q_{ij})^{-1}$. If moreover $\bar{\alpha} := \sum_{1 \leq i \leq k} \alpha_i > k(k - 1 + d)$, then
\[
\mathbb{P}^{\Pi} [\theta : \| f_{i, \theta} - f_{i, \theta_0} \|_1 \leq 2 Kn^{-1} (\bar{\alpha} - k(k - 1 + d))/(2\bar{\alpha}) (\log n)^{-1/2} | Y_{1:n}] = o_{P_{\theta_0}}(1) \tag{(8)}
\]
If we replace $A3$ by $A3bis$, then there exists $K$ large enough such that
\[
\mathbb{P}^{\Pi} [\theta : \| f_{i, \theta} - f_{i, \theta_0} \|_1 \leq 2 Kn^{-1/2} (\log n)^{3/2} | Y_{1:n}] = o_{P_{\theta_0}}(1) \tag{(9)}
\]

Theorem 1 is proved in Section 4.1 as a consequence of Theorem 4 stated in Section 3, which gives posterior concentration rates for general HMMs.

Assumption $A0$ is the usual ergodic condition on the finite state space Markov chain. Assumptions $A1$ and $A2$ are mild usual regularity conditions on the emission densities $g_\gamma$ and hold for instance for multidimensional Gaussian distributions, Poisson distributions, or any regular exponential families. Assumption $A3$ on the prior distribution of the transition matrix $Q$ is satisfied for instance if each row of $Q$ follows a Dirichlet distribution or a mixture of Dirichlet distributions, as used in Nur et al. [16], and assumption (6) is verified for densities $\omega$ that have at most polynomial tails.

The constraint on $\bar{\alpha} = \sum_i \alpha_i$ or condition $A3bis$ are used to ensure that (8) and (9) hold respectively. The posterior concentration result (7) implies that the posterior distribution might put non-negligible mass on values of $\theta$ for which $\rho_0 - 1$ is small and $\| f_{i, \theta} - f_{i, \theta_0} \|_1$ is not. These are parameter values associated to nearly non-ergodic latent Markov chains. Since $\rho_0 - 1$ is small is equivalent to $\sum_j \min_i q_{ij}$ is small, the condition $\bar{\alpha} > k(k - 1 + d)$ prevents such pathological behaviour by ensuring that the prior mass of such sets is small enough. This condition is therefore of a different nature than Rousseau and Mengersen’s [20] condition on the prior, which characterizes the asymptotic behaviour of the posterior distribution on the parameter $\theta$. In other words, their condition allows in (static) mixture models to go from a posterior concentration result on $f_{i, \theta}$ to a posterior concentration result on $\theta$ whereas, here, the constraint on $\bar{\alpha}$ is used to obtain a posterior concentration result on $f_{i, \theta}$. Going back from $\| f_{i, \theta} - f_{i, \theta_0} \|_1$ to the parameters requires a deeper understanding of the geometry of finite HMMs, similar to the one developed in Gassiat and van Handel [8]. This will be needed to estimate the order of the HMM in Section 2.3, and fully explored when $k_0 = 1$ and $k = 2$ in Section 2.4.

For general priors, we do not know whether the $\sqrt{\log n}$ factor appearing in (7) could be replaced or not by any sequence tending to infinity. In the case where the $\alpha_i$’s are large enough (Dirichlet type priors), and when $k_0 = 1$ and $k = 2$, we obtain a concentration rate without the $\sqrt{\log n}$ factor, see Lemma 2 in Section 3. To do so, we prove Lemma 3 in Section 3 for which we need to compute explicitly the stationary distribution and the predictive probabilities to obtain a precise control of the likelihood, for $\theta$’s such that $P_{\theta}$ is near $P_{\theta_0}$, and to control local entropies of slices for $\theta$’s such that $P_{\theta}$ is near $P_{\theta_0}$ and where $\rho_0 - 1$ might be small. It is not clear to us that extending such computations to the general case is possible in a similar fashion. The log $n$ terms appearing in (8) and (9) are consequences of the $\sqrt{\log n}$ term appearing in (7).

2.3. Consistent Bayesian estimation of the number of states

To define a Bayesian estimator of the number of hidden states $k_0$, we need to decide how many states have enough probability mass, and are such that their emission parameters are
different enough. We will be able to do it under the assumptions of Theorem 1. Set $w_n = n^{-(\bar{a}-k(k+d-1))/(2\alpha)} \log n$ if A3 holds and $\tilde{a} > k(k + d - 1)$, and set $w_n = n^{-1/2}(\log n)^{3/2}$ if instead A3bis holds. Let $(u_n)_{n \geq 1}$ and $(v_n)_{n \geq 1}$ be sequences of positive real numbers tending to 0 as $n$ tends to infinity such that $w_n = o(u_n v_n)$. As in Rousseau and Mengersen [20], in the case of a misspecified model with $k_0 < k$, $f_1, \theta_0$ can be represented by merging components or by emptying extra components. For any $\theta \in \Theta_k$, we thus define $J(\theta)$ as
\[ J(\theta) = \{ j : \mathbb{P}_{\theta}(X_1 = j) \geq u_n \}, \]
that is, $J(\theta)$ corresponds to the set of non-empty components. To cluster the components that have similar emission parameters, we define for all $j \in J(\theta)$
\[ A_j(\theta) = \{ i \in J(\theta) : \| \gamma_j - \gamma_i \|^2 \leq v_n \} \]
and the clusters are defined by: for all $j_1, j_2 \in J(\theta)$, $j_1$ and $j_2$ belong to the same cluster (noted $j_1 \sim j_2$) if and only if there exist $r > 1$ and $i_1, \ldots, i_r \in J(\theta)$ with $i_1 = j_1$ and $i_r = j_2$ such that for all $1 \leq l \leq r - 1$, $A_{i_l}(\theta) \cap A_{i_{l+1}}(\theta) \neq \emptyset$. We then define the effective order of the HMM at $\theta$ as the number $L(\theta)$ of different clusters, that is, as the number of equivalent classes with respect to the equivalence relation $\sim$ defined above. By a good choice of $u_n$ and $v_n$, we construct a consistent estimator of $k_0$ by considering either the posterior mode of $L(\theta)$ or its posterior median. This is presented in Theorem 2.

To prove that this gives a consistent estimator, we need an inequality that relates the $L_1$ distance between the $l$-marginals, $\| f_{l,\theta} - f_{l,\theta_0} \|_1$, to a distance between the parameter $\theta$ and parameters $\tilde{\theta}_0$ in $\Theta_k$ such that $f_{l,\tilde{\theta}_0} = f_{l,\theta_0}$. Such an inequality will be proved in Section 4.2, under the following structural assumption.

Let $T = \{ t = (t_1, \ldots, t_k) \in \{ 1, \ldots, k \}^k : t_i < t_{i+1}, i = 0, \ldots, k - 1 \}$. If $b$ is a vector, $b^T$ denotes its transpose.

A4 For any $t = (t_1, \ldots, t_k) \in T$, any $(\pi_i)_{i=1}^{k-t_0} \in (\mathbb{R}^+)^{k-t_0}$ (if $t_0 < k$), any $(a_i)_{i=1}^{k_0}$, $(c_i)_{i=1}^{k_0} \in \mathbb{R}^{k_0}$, $(b_i)_{i=1}^{k_0} \in (\mathbb{R}^d)^{k_0}$, any $z_{i,j} \in \mathbb{R}^d$, $\alpha_{i,j} \in \mathbb{R}$, $i = 1, \ldots, k_0$, $j = 1, \ldots, t_i - t_{i-1}$ (with $t_0 = 0$), such that $\| z_{i,j} \|_1 = 1$, $\alpha_{i,j} \geq 0$ and $\sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} = 1$, for any $(\gamma_i)_{i=1}^{k-t_0}$ which belong to $\Gamma \setminus \{ \gamma_i \}$, $i = 1, \ldots, k_0$,
\begin{equation}
\sum_{i=1}^{k-t_0} \pi_i g_{\gamma_i} + \sum_{i=1}^{k_0} (a_i g_{\gamma_{i,0}} + b_i^T D^1 g_{\gamma_{i,0}}) + \sum_{i=1}^{k_0} c_i^2 \sum_{j=1}^{t_i - t_{i-1}} \alpha_{i,j} z_{i,j} D^2 g_{\gamma_{i,0}} z_{i,j} = 0,
\end{equation}
if and only if
\[ a_i = 0, \quad b_i = 0, \quad c_i = 0 \quad \forall i = 1, \ldots, k_0, \quad \pi_i = 0 \quad \forall i = 1, \ldots, k - t_0. \]

Assumption A4 is a weak identifiability condition for situations when $k_0 < k$. Notice that A4 is the same condition as in Rousseau and Mengersen [20], it is satisfied in particular for Poisson mixtures, location-scale Gaussian mixtures and any mixtures of regular exponential families.

The following theorem says that the posterior distribution of $L(\theta)$ concentrates on the true number $k_0$ of hidden states.
Theorem 2. Assume that assumptions $A_0$–$A_2$ and $A_4$ are verified. If either of the following two situations holds:

- Under assumption $A_3$ (Dirichlet type prior), if $\tilde{\alpha} > k(k + d - 1)$ and
  \[
  u_n v_n n^{(\tilde{\alpha} - k(k + d - 1)) / (2\tilde{\alpha})} \log n \to +\infty.
  \]

- Under assumption $A_3\text{bis}$ (exponential type prior), if $u_n v_n n^{1/2} / (\log n)^{3/2} \to +\infty$
  then
  \[
  P_{\theta}^{\Pi} \left[ \theta : L(\theta) \neq k_0 | Y_{1:n} \right] = o_{P_{\theta_0}} (1). \tag{11}
  \]
  If $\hat{k}_n$ is either the mode or the median of the posterior distribution of $L(\theta)$, then
  \[
  \hat{k}_n = k_0 + o_{P_{\theta_0}} (1). \tag{12}
  \]

One of the advantages of using such an estimate of the order of the HMM, is that we do not need to consider a prior on $k$ and use reversible-jump methods, see Richardson and Green [17], which can be tricky to implement. In particular, we can consider a two-stage procedure where $\hat{k}_n$ is computed based on a model with $k$ components where $k$ is a reasonable upper bound on $k_0$ and then, fixing $k = \hat{k}_n$ an empirical Bayes procedure is defined on $(Q_{i,j}, i, j \leq \hat{k}_n, \gamma_1, \ldots, \gamma_{\hat{k}_n})$.

On the event $\hat{k}_n = k_0$, which has probability going to 1 under $P_{\theta_0}$ the model is regular and using the Bernstein–von Mises theorem of de Gunst and Shcherbakova [4], we obtain that with probability going to 1, the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$ converges in distribution to the centered Gaussian with variance $V_0$, the inverse of Fisher information at parameter $\theta_0$, where $\hat{\theta}_n$ is an efficient estimator of $\theta_0$ when the order is known to be $k_0$, and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to the centered Gaussian with variance $V_0$ under $P_{\theta_0}$.

The main point in the proof of Theorem 2 is to prove an inequality that relates the $L_1$ distance between the $l$-marginals, to a distance between the parameters of the HMM. Under condition $A_4$, we prove that there exists a constant $c(\theta_0) > 0$ such that for any small enough positive $\varepsilon$,

\[
\frac{\| f_{i,\theta} - f_{i,\theta_0} \|_1}{c(\theta_0)} \geq \sum_{1 \leq j \leq k: \forall i, \| \gamma_j - \gamma_i^0 \| > \varepsilon} P_{\theta}(X_1 = j) + \sum_{i=1}^{k_0} \left| P_{\theta}(X_1 \in B(i)) - P_{\theta_0}(X_1 = i) \right| \tag{13} + \sum_{i=1}^{k_0} \left( \left. \sum_{j \in B(i)} P_{\theta}(X_1 = j)(\gamma_j - \gamma_i^0) \right\| + \frac{1}{2} \sum_{j \in B(i)} P_{\theta}(X_1 = j) \| \gamma_j - \gamma_i^0 \|^2 \right),
\]

where $B(i) = \{ j : \| \gamma_j - \gamma_i^0 \| \leq \varepsilon \}$. The above lower bound essentially corresponds to a partition of $\{1, \ldots, k\}$ into $k_0 + 1$ groups, where the first $k_0$ groups correspond to the components that are close to true distinct components in the multivariate mixture and the last corresponds to
components that are emptied. The first term on the right-hand side controls the weights of the components that are emptied (group \( k_0 + 1 \)), the second term controls the sum of the weights of the components belonging to the \( i \)th group, for \( i = 1, \ldots, k_0 \) (components merging with the true \( i \)th component), the third term controls the distance between the mean value over the group \( i \) and the true value of the \( i \)th component in the true mixture while the last term controls the distance between each parameter value in group \( i \) and the true value of the \( i \)th component. A general inequality implying (13), obtained under a weaker condition, namely \( \text{A4bis} \), holds and is stated and proved in Section 4.2.

As we have seen with Theorem 2, we can recover the true parameter \( \theta_0 \) using a two-stage procedure where first \( \hat{k}_n \) is estimated. However, it is also of interest to understand better the behaviour of the posterior distribution in the first stage procedure and see if some behaviour similar to what was observed in Rousseau and Mengersen [20] holds in the case of HMMs. From Theorem 1, it appears that HMMs present an extra difficulty due to the fact that, when the order is overestimated, the neighbourhood of \( \theta \)'s such that \( \mathbb{P}_\theta = \mathbb{P}_{\theta_0} \) contains parameters leading to non-ergodic HMMs. To have a more refined understanding of the posterior distribution, we restrict our attention in Section 2.4 to the case where \( k = 2 \) and \( k_0 = 1 \) which is still nontrivial, see also Gassiat and Kéribin [9] for the description of pathological behaviours of the likelihood in such a case.

2.4. Posterior concentration for the parameters: The case \( k_0 = 1 \) and \( k = 2 \)

In this section, we restrict our attention to the simpler case where \( k_0 = 1 \) and \( k = 2 \). In Theorem 3 below, we prove that if a Dirichlet type prior is considered on the rows of the transition matrix with parameters \( \alpha_j \)'s that are large enough the posterior distribution concentrates on the configuration where the two components (states) are merged (\( \gamma_1 \) and \( \gamma_2 \) are close to one another). When \( k = 2 \), we can parameterize \( \theta \) as \( \theta = (p, q, \gamma_1, \gamma_2) \), with \( 0 \leq p \leq 1, 0 \leq q \leq 1 \), so that

\[
Q_\theta = \left( \frac{1-p}{q}, \frac{p}{1-q} \right), \quad \mu_\theta = \left( \frac{q}{p+q}, \frac{p}{p+q} \right)
\]

when \( p \neq 0 \) or \( q \neq 0 \). If \( p = 0 \) and \( q = 0 \), set \( \mu_\theta = \left( \frac{1}{2}, \frac{1}{2} \right) \), for instance. Also, we may take

\[
\rho_\theta - 1 = (p + q) \land (2 - (p + q))
\]

When \( k_0 = 1 \), the observations are i.i.d. with distribution \( g_{\gamma_0} \, d\nu \), so that one may take \( \theta_0 = (p, 1-p, \gamma^0, \gamma^0) \) for any \( 0 < p < 1 \), or \( \theta_0 = (0, q, \gamma^0, \gamma) \) for any \( 0 < q \leq 1 \) and any \( \gamma \), or \( \theta_0 = (p, 0, \gamma, \gamma^0) \) for any \( 0 < p \leq 1 \) and any \( \gamma \). Also, for any \( x \in \mathcal{X} \), \( \mathbb{P}_{\theta_0, x} = \mathbb{P}_{\theta_0} \) and

\[
\ell_n(\theta, x) - \ell_n(\theta_0, x) = \ell_n(\theta, x) - \ell_n(\theta_0, x).
\]

We take independent Beta priors on \( (p, q) \):

\[
\Pi_2(dp, dq) = C_{\alpha, \beta} p^{\alpha - 1} (1-p)^{\beta - 1} q^{\alpha - 1} (1-q)^{\beta - 1} \mathbb{1}_{0 < p < 1} \mathbb{1}_{0 < q < 1} \, dp \, dq,
\]

thus satisfying \( \text{A3} \). Then the following holds.
Theorem 3. Assume that assumptions A0–A2 together with assumption A4 are verified and consider the prior described above with \( \omega(\cdot) \) verifying A3. Assume moreover that for all \( x, \gamma \mapsto g_\gamma(x) \) is four times continuously differentiable on \( \Gamma \), and that for any \( \gamma \in \Gamma \) there exists \( \epsilon > 0 \) such that for any \( i \leq 4 \),

\[
\int \sup_{\gamma' \in B_d(\gamma, \epsilon)} \left\| \frac{D^i g_{\gamma'}}{g_{\gamma'}} (y) \right\|^4 g_{\gamma'}(y) \nu(dy) < +\infty. \tag{14}
\]

Then, as soon as \( \alpha > 3d/4 \) and \( \beta > 3d/4 \), for any sequence \( \epsilon_n \) tending to 0,

\[
P(\gamma \leq \epsilon_n \text{ or } q \leq \epsilon_n | Y_{1:n}) = o_{\P_{\theta_0}}(1),
\]

and for any sequence \( M_n \) going to infinity,

\[
P(\|\gamma_1 - \gamma_0\| + \|\gamma_2 - \gamma_0\| \leq M_n n^{-1/4} | Y_{1:n}) = 1 + o_{\P_{\theta_0}}(1).
\]

Theorem 3 says that the extra component cannot be emptied at rate \( \epsilon_n \), where the sequence \( \epsilon_n \) can be chosen to converge to 0 as slowly as we want, so that asymptotically, under the posterior distribution neither \( p/(p+q) \) nor \( q/(p+p) \) are small, and the posterior distribution concentrates on the configuration where the components merge, with the emission parameters merging at rate \( n^{-1/4} \). Similarly in Rousseau and Mengersen [20] the authors obtain that, for independent variables, under a Dirichlet \( D(\alpha_1, \ldots, \alpha_k) \) prior on the weights of the mixture and if \( \min \alpha_i > d/2 \), the posterior distribution concentrates on configurations which do not empty the extra-components but merge them to true components. The threshold here is \( 3d/2 \) instead of \( d/2 \). This is due to the fact that there are more parameters involved in a HMM model associated to \( k \) states than in a \( k \)-components mixture model. No result is obtained here in the case where the \( \alpha_i \)'s are small. This is due to the existence of non ergodic \( \P_\theta \) in the vicinity of \( \P_{\theta_0} \) that are not penalized by the prior in such cases. Our conclusion is thus to favour large values of the \( \alpha_i \)'s.

### 3. A general theorem

In this section, we present a general theorem which is used to prove Theorem 1 but which can be of interest in more general HMMs. We assume here that the unobserved Markov chain \((X_i)_{i \geq 1}\) lives in a Polish space \( X \) and the observations \((Y_i)_{i \geq 1}\) are conditionally independent given \((X_i)_{i \geq 1}\) and live in a Polish space \( Y \). \( X, Y \) are endowed with their Borel \( \sigma \)-fields. We denote by \( \theta \in \Theta \), where \( \Theta \) is a subset of an Euclidean space, the parameter describing the distribution of the HMM, so that \( Q_\theta, \theta \in \Theta \) is the Markov kernel of \((X_i)_{i \geq 1}\) and the conditional distribution of \( Y_i \) given \( X_i \) has density with respect to some given measure \( \nu \) on \( Y \) denoted by \( g_\theta(y|x), x \in X, \theta \in \Theta \). We assume that the Markov kernels \( Q_\theta \) admit a (not necessarily unique) stationary distribution \( \mu_\theta \), for each \( \theta \in \Theta \). We still write \( \P_\theta \) for the probability distribution of the stationary HMM \((X_j, Y_j)_{j \geq 1}\) with parameter \( \theta \). That is, for any integer \( n \), any set \( A \) in the Borel
σ-field of \(X^n \times Y^n\):
\[
\mathbb{P}_\theta \left( (X_1, \ldots, X_n, Y_1, \ldots, Y_n) \in A \right) = \int_A \mu_\theta(dx_1) \prod_{i=1}^{n-1} Q_\theta(x_i, dx_{i+1}) \prod_{i=1}^n g_\theta(y_i|x_i) v(dy_1) \cdots v(dy_n).
\]
(15)

Thus for any integer \(n\), under \(P_\theta\), \(Y_1^n = (Y_1, \ldots, Y_n)\) has a probability density with respect to \(v(dy_1) \cdots v(dy_n)\) equal to
\[
f_{n,\theta}(y_1, \ldots, y_n) = \int_{X^n} \mu_\theta(dx_1) \prod_{i=1}^{n-1} Q_\theta(x_i, dx_{i+1}) \prod_{i=1}^n g_\theta(y_i|x_i).
\]
(16)

We denote by \(\Pi_{\Theta_0}\) the prior distribution on \(\Theta_0\) and by \(\pi_{X_0}\) the prior probability on the unobserved initial state, which might be different from the stationary distribution \(\mu_{\Theta_0}\). We set \(\Pi = \Pi_{\Theta_0} \otimes \pi_{X_0}\).

Similarly to before, denote \(\ell_n(\theta, x)\) the log-likelihood starting from \(x\), for all \(x \in X\).

We assume that we are given a stationary HMM \((X_j, Y_j)_{j \geq 1}\) with distribution \(P_{\theta_0}\) for some \(\theta_0 \in \Theta_0\).

For any \(\theta \in \Theta\), it is possible to define real numbers \(\rho_\theta \geq 1\) and \(0 < R_\theta \leq 2\) such that, for any integer \(m\), any \(x \in X\)
\[
\|Q_{\theta_0}^m(x, \cdot) - \mu_{\theta_0}\|_{TV} \leq R_\theta \rho_\theta^{-m},
\]
(17)
where \(\|\cdot\|_{TV}\) is the total variation norm. If it is possible to set \(\rho_\theta > 1\), the Markov chain \((X_n)_n \geq 1\) is uniformly ergodic and \(\mu_{\theta_0}\) is its unique stationary distribution. The following theorem provides a posterior concentration result in a general HMM setting, be it parametric or nonparametric and is an adaptation of Ghosal and van der Vaart [10] to the setup of HMMs. We present the assumptions needed to derive the posterior concentration rate.

**C1** There exists \(A > 0\) such that for any \((x_0, x_1) \in X^2\), \(\mathbb{P}_{\theta_0}\) almost surely, \(\forall n \in \mathbb{N}\), \(|\ell_n(\theta_0, x_0) - \ell_n(\theta_0, x_1)| \leq A\), and there exist \(S_n \subset \Theta \times X\), \(C_n > 0\) and \(\tilde{\epsilon}_n > 0\) a sequence going to 0 with \(n\tilde{\epsilon}_n^2 \rightarrow +\infty\) such that
\[
\sup_{(\theta, x) \in S_n} \mathbb{P}_{\theta_0} \left[ |\ell_n(\theta, x) - \ell_n(\theta_0, x_0)| \leq -n\tilde{\epsilon}_n^2 \right] = o(1), \quad \Pi[S_n] \geq e^{-C_n n\tilde{\epsilon}_n^2}.
\]

**C2** There exists a sequence \((F_n)_n \geq 1\) of subsets of \(\Theta\)
\[
\Pi_{\Theta_0}(F_n^c) = o(e^{-n\tilde{\epsilon}_n^2(1+C_n)}).
\]

**C3** There exists a sequence \(\epsilon_n \geq \tilde{\epsilon}_n\) going to 0, such that \((n\epsilon_n^2(1 + C_n))/(n\epsilon_n^2)\) goes to 0 and
\[
N \left( \frac{\epsilon_n}{12}, F_n, d_l(\cdot, \cdot) \right) \leq e^{(n\epsilon_n^2(\rho_{\theta_0} - 1)^2)/(16(2R_{\theta_0} + \rho_{\theta_0} - 1)^2)},
\]
where \(N(\delta, F_n, d_l(\cdot, \cdot))\) is the smallest number of \(\theta_j \in F_n\) such that for all \(\theta \in F_n\) there exists a \(\theta_j\) with \(d_l(\theta_j, \theta) \leq \delta\).
Here $d_l(\theta, \theta_j) = \| f_l, \theta - f_l, \theta_j \|_1 := \int_{\mathcal{Y}} | f_l, \theta - f_l, \theta_j | (y) \, d\nu(y)$.

**C3bis** There exists a sequence $\epsilon_n \geq \tilde{\epsilon}_n$ going to 0 such that

$$
\sum_{m \geq 1} \frac{\Pi_{\Omega}(A_{n,m}(\epsilon_n))}{\Pi(S_n)} e^{-(nm^2\epsilon_n^2)/(32l)} = o(e^{-n\epsilon_n^2})
$$

and

$$
\mathcal{N}\left( \frac{m\epsilon_n}{12}, A_{n,m}(\epsilon_n), d_l(\cdot, \cdot) \right) \leq e^{(nm^2\epsilon_n^2(\rho_\theta - 1)^2)/(16l(2R_{\theta_0} + \rho_{\theta_0} - 1)^2)},
$$

where

$$
A_{n,m}(\epsilon) = \mathcal{F}_n \cap \left\{ \theta : m\epsilon \leq \| f_l, \theta - f_l, \theta_0 \|_1 \frac{\rho_\theta - 1}{2R_{\theta} + \rho_{\theta} - 1} \leq (m + 1)\epsilon \right\}.
$$

**Theorem 4.** Assume that $\rho_{\theta_0} > 1$ and that assumptions **C1**–**C2** are satisfied, together with either assumption **C3** or **C3bis**. Then

$$
\mathbb{P}^{\Pi}\left[ \theta : \| f_l, \theta - f_l, \theta_0 \|_1 \frac{\rho_\theta - 1}{2R_{\theta} + \rho_{\theta} - 1} \geq \epsilon_n \bigg| Y_{1:n} \right] = o_{\mathbb{P}_{\theta_0}}(1).
$$

Theorem 4 gives the posterior concentration rate of $\| f_l, \theta - f_l, \theta_0 \|_1$ up to the parameter $\frac{\rho_\theta - 1}{2R_{\theta} + \rho_{\theta} - 1}$. In Ghosal and van der Vaart [10], for models of non-independent variables, the authors consider a parameter space where the mixing coefficient term (for us $\rho_\theta - 1$) is uniformly bounded from below by a positive constant over $\Theta$ (see their assumption (4.1) for the application to Markov chains or their assumption on $\mathcal{F}$ in Theorem 7 for the application to Gaussian time series), or equivalently they consider a prior whose support in $\Theta$ is included in a set where $\frac{\rho_\theta - 1}{2R_{\theta} + \rho_{\theta} - 1}$ is uniformly bounded from below, so that their posterior concentration rate is directly expressed in terms of $\| f_l, \theta - f_l, \theta_0 \|_1$. Since we do not restrict ourselves to such frameworks the penalty term $\rho_\theta - 1$ is incorporated in our result. However Theorem 4, is proved along the same lines as Theorem 1 of Ghosal and van der Vaart [10].

The assumption $\rho_{\theta_0} > 1$ implies that the hidden Markov chain $X$ is uniformly ergodic. Assumptions **C1**–**C2** and either **C3** or **C3bis** are similar in spirit to those considered in general theorems on posterior consistency or posterior convergence rates, see, for instance, Ghosh and Ramamoorthi [11] and Ghosal and van der Vaart [10]. Assumption **C3bis** is often used to eliminate some extra log $n$ term which typically appear in nonparametric posterior concentration rates and is used in particular in the proof of Theorem 3.

4. Proofs

4.1. Proof of Theorem 1

The proof consists in showing that the assumptions of Theorem 4 are satisfied.
Following the proof of Lemma 2 of Douc et al. [5] we find that, since $\rho_{\theta_0} > 1$, for any $x_0 \in \mathcal{X}$,

$$\left| \ell_n(\theta_0, x_0) - \ell_n(\theta_0, x_1) \right| \leq 2 \left( \frac{\rho_{\theta_0}}{\rho_{\theta_0} - 1} \right)^2$$

so that setting $A = 2 \left( \frac{\rho_{\theta_0}}{\rho_{\theta_0} - 1} \right)^2$ the first point of $C_1$ holds.

We shall verify assumption $C_1$ with $\tilde{\epsilon}_n = \frac{M_n}{\sqrt{n}}$ for some $M_n$ tending slowly enough to infinity and that will be chosen later. Note that the assumption $A_0$ and the construction (5) allow to define a $\tilde{\theta}_0 \in \Theta_k$ such that, writing $\tilde{\theta}_0 = (\tilde{Q}^0, \tilde{\gamma}_1^0, ..., \tilde{\gamma}_k^0)$ with $\tilde{Q}^0 = (\tilde{q}_{i,j}^0, i, j \leq k)$, if $V$ is a bounded subset of $\{\theta = (Q, \gamma_1, ..., \gamma_k); |q_{i,j} - \tilde{q}_{i,j}^0| \leq \tilde{\epsilon}_n\}$, then

$$\inf_{\theta \in V} \rho_{\theta} > 1,$$

for large enough $n$, and

$$\sup_{\theta \in V} \sup_{x, x_0 \in \mathcal{X}} \left| \ell_n(\theta, x) - \ell_n(\theta, x_0) \right| \leq 2 \sup_{\theta \in V} \left( \frac{\rho_{\theta}}{\rho_{\theta} - 1} \right)^2.$$

Following the proof of Lemma 2 of Douc et al. [5] gives that, if $A_0$ and $A_1$ hold, for all $\theta \in V$ $\mathbb{P}_{\theta_0}$-a.s.,

$$\ell_n(\theta, x_0) - \ell_n(\theta_0, x_0) = (\theta - \theta_0)^T \nabla_\theta \ell_n(\theta_0, x_0)$$

$$+ \int_0^1 (\theta - \theta_0)^T D^2_{\theta} \ell_n(\theta_0 + u(\theta - \theta_0), x_0)(\theta - \theta_0)(1 - u) \, du. \tag{19}$$

Following Theorem 2 in Douc et al. [5], $n^{-1/2} \nabla_\theta \ell_n(\theta_0, x)$ converges in distribution under $\mathbb{P}_{\theta_0}$ to $\mathcal{N}(0, V_0)$ for some positive definite matrix $V_0$, and following Theorem 3 in Douc et al. [5], we get that $\sup_{\theta \in V} n^{-1} D^2_{\theta} \ell_n(\theta, x_0)$ converges $\mathbb{P}_{\theta_0}$ a.s. to $V_0$. Thus, we may set:

$$S_n = \{\theta \in V; \|\gamma_j - \gamma_j^0\| \leq 1/\sqrt{n} \forall j \leq k\} \times \mathcal{X}$$

so that

$$\sup_{(\theta, x) \in S_n} \mathbb{P}_{\theta_0}[\ell_n(\theta, x) - \ell_n(\theta_0, x_0) < -M_n] = o(1). \tag{20}$$

Moreover, letting $D = k(k - 1 + d)$, we have $\Pi \otimes \Pi_{\mathcal{X}}(S_n) \geq n^{-D/2}$ and $C_1$ is then satisfied setting $C_n = D \log n/(2M_n^2)$.

Let now $v_n = n^{-D/(2 \min_{1 \leq \ell \leq k} a_{i\ell})} / \sqrt{\log n}$ and $u_n = n^{-D/(2 \sum_{1 \leq \ell \leq k} a_{i\ell})} / \sqrt{\log n}$, and define

$$\mathcal{F}_n = \left\{ \theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1; \gamma_1, ..., \gamma_k); q_{ij} \geq v_n, 1 \leq i \leq k, 1 \leq j \leq k, \right.$$

$$\left. \sum_{j=1}^{k} \inf_{1 \leq i \leq k} q_{ij} \geq u_n, \|\gamma_i\| \leq n^b, 1 \leq i \leq k \right\}.$$
Now, if $\theta \in \mathcal{F}_n$ then there exist $1 \leq i, j \leq k$ such that $q_{ij} \leq v_n$, or $\sum_{j=1}^{k} \inf_{1 \leq i \leq k} q_{ij} \leq u_n$, or there exists $1 \leq i \leq k$ such that $\|y_i\| \geq n^b$. Using A3 we easily obtain that for fixed $i$ and $j$, $\Pi(\{\theta: q_{ij} \leq v_n\}) = O(v_n^{-n})$ and $\Pi(\{\theta: \|y_i\| \geq n^b\}) = o(n^{-D/2})$. Also, if $\sum_{j=1}^{k} \inf_{1 \leq i \leq k} q_{ij} \leq u_n$, then there exists a function $i(\cdot)$ from $\{1, \ldots, k\}$ to $\{1, \ldots, k\}$ whose image set has cardinality at least 2 such that $\sum_{j=1}^{k} q_{i(j)j} \leq u_n$. This gives, using A3, $\Pi(\{\theta: \sum_{j=1}^{k} \inf_{1 \leq i \leq k} q_{ij} \leq u_n\}) = O(u_n^{\sum_{1 \leq i \leq k} a_i})$. Thus,

$$\Pi(\mathcal{F}_n) = O(v_n^{\min_{1 \leq i \leq k} a_i} + u_n^{\sum_{1 \leq i \leq k} a_i}) + o(n^{-D/2}).$$

We may now choose $M_n$ tending to infinity slowly enough so that $v_n^{\min_{1 \leq i \leq k} a_i} + u_n^{\sum_{1 \leq i \leq k} a_i} = o(e^{-M_n n^{-D/2}})$ and $\Pi(\mathcal{F}_n) = o(e^{-M_n n^{-D/2}})$. Then, C2 holds.

Now, using the definition of $f_{i, \theta}$, we obtain that

$$\|f_{i, \theta_1} - f_{i, \theta_2}\|_1 \leq \sum_{j=1}^{k} |\mu_{\theta_1} - \mu_{\theta_2}| + \sum_{i, j=1}^{k} |Q_{i,j}^1 - Q_{i,j}^2| + \max_{j \leq k} \|g_{\gamma_1} - g_{\gamma_2}\|_1$$

so that using Lemma 1 below, A1 and A2 we get that for some constant $B$, $\forall (\theta_1, \theta_2) \in \mathcal{F}_n$

$$\|f_{i, \theta_1} - f_{i, \theta_2}\|_1 \leq B \left( \frac{1}{v_n^2} + n^a \right) \|\theta_1 - \theta_2\|_1.$$

Thus for some other constant $\tilde{B}$,

$$N(\delta, \mathcal{F}_n, d(\cdot, \cdot)) \leq \left[ \frac{\tilde{B}}{\delta} \left( \frac{1}{v_n^2} + n^a \right) \right]^{k(k-1)+kd}$$

and C3 holds when setting $\epsilon_n = K \sqrt{\frac{\log n}{n}}$ with $K$ large enough.

We have proved that under assumptions A0, A1, A2, A3, Theorem 4 applies with $\epsilon_n = K \sqrt{\frac{\log n}{n}}$ so that

$$\mathbb{P}^\Pi \left[ \|f_{i, \theta} - f_{i, \theta_0}\|_1 (\rho_\theta - 1) \geq K \sqrt{\frac{\log n}{n}} Y_{1:n} \right] = o_{\bar{\mathbb{P}}_{\theta_0}} (1)$$

and the first part of Theorem 1 is proved. Now

$$o_{\bar{\mathbb{P}}_{\theta_0}} (1) = \mathbb{P}^\Pi \left[ \|f_{i, \theta} - f_{i, \theta_0}\|_1 (\rho_\theta - 1) \geq K \sqrt{\frac{\log n}{n}} Y_{1:n} \right]$$

$$= \mathbb{P}^\Pi \left[ \theta \in \mathcal{F}_n \text{ and } \|f_{i, \theta} - f_{i, \theta_0}\|_1 (\rho_\theta - 1) \geq K \sqrt{\frac{\log n}{n}} Y_{1:n} \right] + o_{\bar{\mathbb{P}}_{\theta_0}} (1).$$
Since $$\rho_\theta - 1 \geq \sum_{j=1}^{k} \min_{1 \leq i \leq k} q_{ij}$$, for all $$\theta \in \mathcal{F}_n$$, $$\rho_\theta - 1 \geq u_n$$,

$$\mathbb{P} \left[ \left\| f_{i,\theta} - f_{i,\theta_0} \right\|_1 (\rho_\theta - 1) \geq K \sqrt{\frac{\log n}{n}} Y_{1:n} \right] \geq \mathbb{P} \left[ \left\| f_{i,\theta} - f_{i,\theta_0} \right\|_1 \geq 2K \frac{1}{u_n} \sqrt{\frac{\log n}{n}} Y_{1:n} \right],$$

and the theorem follows when A3 holds. If now A3bis holds instead of A3, one gets, taking $$u_n = v_n = h/\log n$$, with $$h > 2C/(k + d - 1)$$

$$\Pi(\mathcal{F}_n) = O\left(v_n \exp(-C/v_n)\right) + o\left(n^{-D/2}\right) = o\left(e^{-M_n n^{-D/2}}\right)$$

by choosing $$M_n$$ increasing to infinity slowly enough so that C2 and C3 hold. The end of the proof follows similarly as before.

To finish the proof of Theorem 1, we need to prove the following lemma.

**Lemma 1.** The function $$\theta \mapsto \mu_\theta$$ is continuously differentiable in $$(\Delta_0^k)^k \times \Gamma^k$$ and there exists an integer $$c > 0$$ and a constant $$C > 0$$ such that for any $$1 \leq i \leq k, 1 \leq j \leq k - 1$$, any $$m = 1, \ldots, k$$,

$$\left| \frac{\partial \mu_\theta(m)}{\partial q_{ij}} \right| \leq \frac{C}{(\inf_{i' \neq j'} q_{i'j'})^{2c}}.$$

One may take $$c = k - 1$$.

Let $$\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1; \gamma_1, \ldots, \gamma_k)$$ be such that $$(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1) \in \Delta_0^k, Q_\theta = (q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k)$$ is a $$k \times k$$ stochastic matrix with positive entries, and $$\mu_\theta$$ is uniquely defined by the equation

$$\mu_\theta^T Q_\theta = \mu_\theta^T$$

if $$\mu_\theta$$ is the vector $$(\mu_\theta(m))_{1 \leq m \leq k}$$. This equation is solved by linear algebra as

$$\mu_\theta(m) = \frac{P_m(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1)}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1)}, \quad m = 1, \ldots, k - 1,$$

$$\mu_\theta(k) = 1 - \sum_{m=1}^{k-1} \mu_\theta(m),$$

(21)

where $$P_m, l = 1, \ldots, k - 1$$ and $$R$$ are polynomials where the coefficients are integers (bounded by $$k$$) and the monomials are all of degree $$k - 1$$, each variable $$q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1$$ appearing with power 0 or 1. Now, since the equation has a unique solution as soon as $$(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1) \in \Delta_0^k$$, then $$R$$ is never 0 on $$\Delta_0^k$$, so it may be 0 only at the boundary. Thus, as a fraction of polynomials with nonzero denominator, $$\theta \mapsto \mu_\theta$$ is infinitely differentiable in $$(\Delta_0^k)^k \times \Gamma^k$$, and the derivative has components all of form

$$\frac{P(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1)}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1)^2},$$
where again $P$ is a polynomial where the coefficients are integers (bounded by $2k$) and the monomials are all of degree $k - 1$, each variable $q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1$ appearing with power 0 or 1. Thus, since all $q_{ij}$’s are bounded by 1 there exists a constant $C$ such that for all $m = 1, \ldots, k, i = 1, \ldots, k, j = 1, \ldots, k - 1$,

$$\frac{|\partial \mu_\theta(m)|}{\partial q_{ij}} \leq \frac{C}{R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1)^2}. \tag{22}$$

We shall now prove that

$$R(q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k - 1) \geq \left( \inf_{1 \leq i \leq k, 1 \leq j \leq k, i \neq j} q_{ij} \right)^{k-1}, \tag{23}$$

which combined with (22) and (23) implies Lemma 1. Note that we can express $R$ as a polynomial function of $Q = q_{ij}, 1 \leq i \leq k, 1 \leq j \leq k, i \neq j$. Indeed, $\mu := (\mu_\theta(i))_{1 \leq i \leq k-1}$ is solution of

$$\mu^T \cdot M = V^T,$$

where $V$ is the $(k - 1)$-dimensional vector $(q_{kj})_{1 \leq j \leq k - 1}$, and $M$ is the $(k - 1) \times (k - 1)$-matrix with components $M_{i,j} = q_{kj} - q_{ij} + \mathbb{1}_{i,j}$. Since $R$ is the determinant of $M$, this leads to, for any $k \geq 2$:

$$R = \sum_{\sigma \in S_{k-1}} \varepsilon(\sigma) \prod_{1 \leq i \leq k - 1, \sigma(i) = i} \left( q_{ki} + \sum_{1 \leq j \leq k-1, j \neq i} q_{ij} \right) \prod_{1 \leq i \leq k - 1, \sigma(i) \neq i} (q_{ki} - q_{\sigma(i)i}), \tag{24}$$

where for any integer $n$, $S_n$ is the set of permutations of $\{1, \ldots, n\}$, and for each permutation $\sigma$, $\varepsilon(\sigma)$ is its signature. Thus, $R$ is a polynomial in the components of $Q$ where each monomial has integer coefficient and has $k - 1$ different factors. The possible monomials are of form

$$\beta \prod_{i \in A} q_{ki} \prod_{i \in B} q_{ij(i)},$$

where $(A, B)$ is a partition of $\{1, \ldots, k - 1\}$, and for all $i \in B, j(i) \in \{1, \ldots, k - 1\}$ and $j(i) \neq i$. In case $B = \varnothing$, the coefficient $\beta$ of the monomial is $\sum_{\sigma \in S_{k-1}} \varepsilon(\sigma) = 0$, so that we only consider partitions such that $B \neq \varnothing$. Fix such a monomial with non-null coefficient, let $(A, B)$ be the associated partition. Let $Q$ be such that, for all $i \in A, q_{ki} > 0$, for all $i \notin A, q_{ki} = 0$ and $q_{kk} > 0$ (used to handle the case $A = \varnothing$). Fix also $q_{ij(i)} = 1$ for all $i \in B$. Then, if $(A', B')$ is another partition of $\{1, \ldots, k - 1\}$ with $B' \neq \varnothing$, the monomial $\prod_{i \in A'} q_{ki} \prod_{i \in B'} q_{ij(i)} = 0$. Thus, $R(Q)$ equals $\prod_{i \in A} q_{ki} \prod_{i \in B} q_{ij(i)}$ times the coefficient of the monomial. But $R(Q) \geq 0$, so that this coefficient is a positive integer and (23) follows.

### 4.2. Proof of Theorem 2

Applying Theorem 1, we get that under the assumptions of Theorem 2, there exists $K$ such that

$$\mathbb{P}_{\theta_0}[\|f_t, \theta - f_t, \theta_0\|_1 \leq 2K w_n|Y_{1:n}) = 1 + o_{\mathbb{P}_{\theta_0}}(1).$$
But if inequality (13) holds, then as soon as
\[ \| f_1, \theta - f_{1, \theta_0} \|_1 \lesssim w_n \] (25)
we get that, for any \( j \in \{1, \ldots, k\} \), either \( \mathbb{P}_\theta(X_1 = j) \lesssim w_n \), or
\[ \exists i \in \{1, \ldots, k_0\}, \quad \mathbb{P}_\theta(X_1 = j) \| \gamma_j - \gamma_0^i \|_2^2 \lesssim w_n. \]

Let us choose \( \epsilon \leq \min_{i \neq j} \| \gamma_i^0 - \gamma_j^0 \| / 4 \) in the definition of \( B(i) \) in (13). We then obtain that for large enough \( n \), all \( j_1, j_2 \in J(\theta) \), we have \( j_1 \sim j_2 \) if and only if they belong to the same \( B(i) \), \( i = 1, \ldots, k_0 \), so that \( L(\theta) \leq k_0 \). On the other hand, \( L(\theta) < k_0 \) would mean that at least one \( B(i) \) would be empty which contradicts the fact that
\[ \left| \mathbb{P}_\theta(X_1 \in B(i)) - \mathbb{P}_{\theta_0}(X_1 = i) \right| \leq w_n. \]
Thus, for large enough \( n \), if (25) holds, then \( L(\theta) = k_0 \), so that
\[ P \left\lceil L(\theta) = k_0 \right\rceil \{ Y^n \} = 1 + o_{\theta_0}(1). \]

To finish the proof, we now prove that (13) holds under the assumptions of Theorem 2. This will follow from Proposition 1 below which is slightly more general.

An inequality that relates the \( L_1 \) distance of the \( l \)-marginals to the parameters of the HMM is proved in Gassiat and van Handel [8] for translation mixture models, with the strength of being uniform over the number (possibly infinite) of populations in the mixture. However, for our purpose, we do not need such a general result, and it is possible to obtain it for more general situations than families of translated distributions, under the structural assumption \( A_4 \). The inequality following Theorem 3.10 of Gassiat and van Handel [8] says that there exists a constant \( c(\theta_0) > 0 \) such that for any small enough positive \( \epsilon \),
\[
\frac{\| f_1, \theta - f_{1, \theta_0} \|_1}{c(\theta_0)} \geq \sum_{1 \leq j \leq k : \forall \| \gamma_j^0 - \gamma_j^0 \| > \epsilon} \mathbb{P}_\theta(X_1 = j) + \sum_{1 \leq i_1, \ldots, i_l \leq k_0} \left| \mathbb{P}_\theta(X_{1:l} \in A(i_1, \ldots, i_l)) - \mathbb{P}_{\theta_0}(X_{1:l} = i_1 \cdots i_l) \right| \]
\[
+ \left\| \sum_{(j_1, \ldots, j_l) \in A(i_1, \ldots, i_l)} \mathbb{P}_\theta(X_{1:l} = j_1 \cdots j_l) \left\{ \begin{pmatrix} \gamma_{j_1} \\ \vdots \\ \gamma_{j_l} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \vdots \\ \gamma_{i_l}^0 \end{pmatrix} \right\} \right\| 
+ \frac{1}{2} \sum_{(j_1, \ldots, j_l) \in A(i_1, \ldots, i_l)} \mathbb{P}_\theta(X_{1:l} = j_1 \cdots j_l) \left\| \begin{pmatrix} \gamma_{j_1} \\ \vdots \\ \gamma_{j_l} \end{pmatrix} - \begin{pmatrix} \gamma_{i_1}^0 \\ \vdots \\ \gamma_{i_l}^0 \end{pmatrix} \right\|_2^2, \] (26)
where \( A(i_1, \ldots, i_l) = \{(j_1, \ldots, j_l) : \|y_{j_1} - y_{i_1}^0\| \leq \varepsilon, \ldots, \|y_{j_l} - y_{i_l}^0\| \leq \varepsilon\} \). The above lower bound essentially corresponds to a partition of \( \{1, \ldots, k\} \) into \( k_0 + 1 \) groups, where the first \( k_0 \) groups correspond to the components that are close to true distinct components in the multivariate mixture and the last corresponds to components that are emptied. The first term on the right-hand side controls the weights of the components that are emptied (group \( k_0 + 1 \)), the second term controls the sum of the weights of the components belonging to the \( i \)th group, for \( i = 1, \ldots, k_0 \) (components merging with the true \( i \)th component), the third term controls the distance between the mean value over the group \( i \) and the true value of the \( i \)th component in the true mixture while the last term controls the distance between each parameter value in group \( i \) and the true value of the \( i \)th component.

Notice that (13) is a consequence of (26). We shall prove that (26) holds under an assumption slightly more general than A4. For this, we need to introduce some notations. For all \( I = (i_1, \ldots, i_l) \in \{1, \ldots, k\}^l \), define \( \gamma_I = (\gamma_{i_1}, \ldots, \gamma_{i_l}) \), \( G_{\gamma_I} = \prod_{t=1}^l g_{\gamma_t}(\gamma_t) \), \( D^1G_{\gamma_I} \) the vector of first derivatives of \( G_{\gamma_I} \) with respect to each of the distinct elements in \( \gamma_I \), note that it has dimension \( d \times |I| \), where \( |I| \) denotes the number of distinct indices in \( I \), and similarly define \( D^2G_{\gamma_I} \), the symmetric matrix in \( R^{d|I| \times d|I|} \) made of second derivatives of \( G_{\gamma_I} \) with respect to the distinct elements (indices) in \( \gamma_I \). For any \( t = (t_1, \ldots, t_{k_0}) \in T \), define for all \( i \in \{1, \ldots, k_0\} \) the set \( J(i) = \{t_{i-1} + 1, \ldots, t_i\} \), using \( t_0 = 0 \). We then consider the following condition:

**A4bis** For any \( t = (t_1, \ldots, t_{k_0}) \in T \), for all collections \((\pi_I)_I, (\gamma_I)_I, I \notin \{1, \ldots, k_0\}^l \) satisfying \( \pi_I \geq 0 \), \( \gamma_I = (\gamma_{i_1}, \ldots, \gamma_{i_l}) \) such that \( \gamma_{i_j} = \gamma_{i_j}^0 \) when \( i_j \in J(i) \) for some \( i \leq k_0 \) and \( \gamma_{i_j} \in \Gamma \setminus \{\gamma_{i_j}^0, i = 1, \ldots, k_0\} \) when \( i_j \notin \{1, \ldots, k_0\} \), for all collections \((a_I)_I, (c_I)_I, (b_I)_I, I \in \{1, \ldots, k_0\}^l \), \( a_I \in \mathbb{R}, c_I \geq 0 \) and \( b_I \in \mathbb{R}^{|I|} \), for all collection of vectors \( z_{I,J} \in \mathbb{R}^{|I|} \) with \( I \in \{1, \ldots, k_0\}^l \) and \( J \in J(i_1) \times \cdots \times J(i_l) \) satisfying \( \|z_{I,J}\| = 1 \), and all sequences \((\alpha_{I,J})_I \), satisfying \( \alpha_{I,J} \geq 0 \) and \( \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \alpha_{I,J} = 1 \),

\[
\sum_{I_1 \notin \{1, \ldots, k_0\}^l} \sum_{I \in \{1, \ldots, k_0\}^l} \pi_I G_{\gamma_I} + \sum_{I \in \{1, \ldots, k_0\}^l} \left( a_I G_{\gamma_I} + b_I^T D^1G_{\gamma_I} \right) \\
+ \sum_{I \in \{1, \ldots, k_0\}^l} c_I \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \alpha_{I,J} z_{I,J}^T D^2G_{\gamma_I} z_{I,J} = 0
\]

\[
\sum_{I \notin \{1, \ldots, k_0\}^l} \pi_I + \sum_{I \in \{1, \ldots, k_0\}^l} (|a_I| + \|b_I\| + c_I) = 0.
\]

We have the following proposition.

**Proposition 1.** Assume that the function \( \gamma \mapsto g_{\gamma}(y) \) is twice continuously differentiable in \( \Gamma \) and that for all \( y \), \( g_{\gamma}(y) \) vanishes as \( \|y\| \) tends to infinity. Then, if assumption A4bis is verified, (26) holds. Moreover, condition A4bis is verified as soon as condition A4 (corresponding to \( l = 1 \)) is verified.
Let us now prove Proposition 1. To prove the first part of the proposition, we follow the ideas of the beginning of the proof of Theorem 5.11 in Gassiat and van Handel [8]. If (26) does not hold, there exist a sequence of \( I \)-marginals \((f_l, \theta^n)_{n \geq 1}\) with parameters \((\theta^n)_{n \geq 1}\) such that for some positive sequence \(\epsilon_n\) tending to 0, \(\| f_{l, \theta^n} - f_{l, \theta^0} \|_1 / N_n(\theta^n)\) tends to 0 as \(n\) tends to infinity, with
\[
N_n(\theta) = \sum_{1 \leq j \leq d} \mathbb{P}_{\theta}(X_1 = j) + \sum_{1 \leq i_1, \ldots, i_k \leq 0} \left[ \sum_{(j_1, \ldots, j_k) \in A_n(i_1, \ldots, i_k)} \mathbb{P}_{\theta}(X_{1:l} = j_1 \cdots j_k) - \mathbb{P}_{\theta^0}(X_{1:l} = i_1 \cdots i_l) \right]
\]
with \(A_n(i_1, \ldots, i_k) = \{(j_1, \ldots, j_k) : \| \gamma_{j_1} - \gamma_{i_1}^0 \| \leq \epsilon_n, \ldots, \| \gamma_{j_k} - \gamma_{i_k}^0 \| \leq \epsilon_n\} \).

Now, \(f_{l, \theta^n} = \sum_{l \in \{1, \ldots, k\}} \mathbb{P}_{\theta^n}(X_1, \ldots, X_l = I) G_{\gamma_l^n} \) where \(\theta^n = (Q^n, (\gamma_1^n, \ldots, \gamma_k^n))\), \(Q^n\) a transition matrix on \(\{1, \ldots, k\}\). It is possible to extract a subsequence along which, for all \(i = 1, \ldots, k\), either \(\gamma_i^n\) converges to some limit \(\gamma_i\) or \(\| \gamma_i^n \|\) tends to infinity. Choose now the indexation such that for \(i = 1, \ldots, t_1\), \(\gamma_i^n\) converges to \(\gamma_1^0\), for \(i = t_1 + 1, \ldots, t_2\), \(\gamma_i^n\) converges to \(\gamma_2^0\), and so on, for \(i = t_{k_0} - 1 + 1, \ldots, t_{k_0}\), \(\gamma_i^n\) converges to \(\gamma_{k_0}^0\), and if \(t_{k_0} < k\), for some \(\tilde{k} \leq k\), for \(i = t_{k_0} + 1, \ldots, \tilde{k}\), \(\gamma_i^n\) converges to some \(\gamma_i \notin \{\gamma_1^0, \ldots, \gamma_{k_0}^0\}\), and for \(i = \tilde{k} + 1, \ldots, k\), \(\| \gamma_i^n \|\) tends to infinity. It is possible that \(\tilde{k} = t_{k_0}\) in which case no \(\gamma_i^n\) converges to some \(\gamma_i \notin \{\gamma_1^0, \ldots, \gamma_{k_0}^0\}\).

Such a \(t = (t_1, \ldots, t_{k_0}) \in T\) exists, because if \(\| f_{l, \theta^n} - f_{l, \theta^0} \|_1 / N_n(\theta^n)\) tends to 0 as \(n\) tends to infinity, \(\| f_{l, \theta^n} - f_{l, \theta^0} \|_1\), and \(N_n(\theta^n)\) tends to 0 as \(n\) tends to infinity (if it was not the case, using the regularity of \(\theta \mapsto f_{l, \theta}\) we would have a contradiction). Now along the subsequence we may write, for large enough \(n\):
\[
N_n(\theta^n) = \sum_{I \notin \{1, \ldots, k_0\}} \mathbb{P}_{\theta}(X_{1:l} = J) + \sum_{I \in \{1, \ldots, k_0\}} \left[ \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \mathbb{P}_{\theta}(X_{1:l} = J) - \mathbb{P}_{\theta^0}(X_{1:l} = I) \right]
\]
\[
\quad + \left[ \sum_{J \in J(i_1) \times \cdots \times J(i_l)} \mathbb{P}_{\theta}(X_{1:l} = J) \| \gamma_J - \gamma_J^0 \|^2 \right].
\]
We shall use a Taylor expansion till order 2. To be perfectly rigorous in the following, we need to express explicitly $I$ in terms of its distinct indices, $(\tilde{i}_1, \ldots, \tilde{i}_{|I|})$, so that $G_{\gamma I} = \prod_{i=1}^{|I|} \prod_{j: i_j = \tilde{i}_i} g_{\gamma_{i_j}}(y_j)$, but to keep notations concise we do not make such a distinction and for instance $(\gamma^n_J - \gamma^0_J)^T D^1 G_{\gamma^0_J}$ means

$$\sum_{i=1}^{|I|} (\gamma^n_{i_i} - \gamma^0_{i_i})^T \frac{\partial G_{\gamma I}}{\partial \gamma_{i_i}},$$

and similarly for the second derivatives. We have

$$f_{l, \theta^n} - f_{l, \theta^0} = \sum_{I \not\in \{1, \ldots, k_0 \}^l} \mathbb{P}_\theta(X_{1:l} = I) G^n_{\gamma I} + \sum_{I \in \{1, \ldots, k_0 \}^l} \left\{ \left[ \sum_{J \in J(i_1) \times \ldots \times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) - \mathbb{P}_{\theta^0}(X_{1:l} = I) \right] G^0_{\gamma I} + \frac{1}{2} \sum_{J \in J(i_1) \times \ldots \times J(i_l)} \mathbb{P}_\theta(X_{1:l} = J) (\gamma_J - \gamma^0_J)^T D^2 G^0_{\gamma I} (\gamma_J - \gamma^0_J) \right\}$$

with $\gamma^*_I \in (\gamma^n_I, \gamma^0_I)$. Thus, using the fact that for all $y$, $g_{\gamma}(y)$ vanishes as $\|\gamma\|$ tends to infinity, $f_{l, \theta^n} - f_{l, \theta^0}/N_n(\theta^n)$ converges pointwise along a subsequence to a function $h$ of form

$$h = \sum_{I \not\in \{1, \ldots, k_0 \}^l} \pi_I G_{\gamma I} + \sum_{I \in \{1, \ldots, k_0 \}^l} (a_I G_{\gamma^0 I} + b_I^T D^1 G_{\gamma^0 I}) + \sum_{I \in \{1, \ldots, k_0 \}^l} c_I \sum_{J \in J(i_1) \times \ldots \times J(i_l)} \alpha_{I,J} z^T_{I,J} D^2 G^0_{\gamma I} z_{I,J} = 0$$

as in condition $L(l)$, with $\sum_{I \not\in \{1, \ldots, k_0 \}^l} \pi_I + \sum_{I \in \{1, \ldots, k_0 \}^l} (\|a_I\| + \|b_I\| + c_I) = 1$. But as $\|f_{l, \theta^n} - f_{l, \theta^0}\|/N_n(\theta^n)$ tends to 0 as $n$ tends to infinity, we have $\|h\|_1 = 0$ by Fatou’s lemma, and thus $h = 0$, contradicting the assumption.

Let us now prove that $A4$ implies $A4bis$. Let

$$\sum_{I \not\in \{1, \ldots, k_0 \}^l} \pi_I G_{\gamma I} + \sum_{I \in \{1, \ldots, k_0 \}^l} (a_I G_{\gamma^0 I} + b_I^T D^1 G_{\gamma^0 I}) + \sum_{I \in \{1, \ldots, k_0 \}^l} c_I \sum_{J \in J(i_1) \times \ldots \times J(i_l)} \alpha_{I,J} z^T_{I,J} D^2 G^0_{\gamma I} z_{I,J} = 0$$

$$\Leftrightarrow \sum_{I \not\in \{1, \ldots, k_0 \}^l} \pi_I + \sum_{I \in \{1, \ldots, k_0 \}^l} (\|a_I\| + \|b_I\| + c_I) = 0$$
with \( \pi_I, a_I, b_I, \alpha_{I,J} \) and \( z_{I,J} \) be as in assumption A4bis. We group the terms depending only on \( y_1 \) and we can rewrite the equation as

\[
\sum_{i_n=0+1}^k \pi_i'(y_2, \ldots, y_l)g_{y_i}(y_1) + \sum_{i=1}^{k_0} (a_i'(y_2, \ldots, y_l)g_{y_i^0}(y_1) + b_i'T(y_2, \ldots, y_l)D^1g_{y_i^0}(y_1)) + \sum_{i=1}^{k_0} (a_i'g_{y_i^0}(y_1) + b_i'TD^2g_{y_i^0}(y_1)) = 0,
\]

where we have written

\[
z_{I,J} = (z_{I,J}(i_1), \ldots, z_{I,J}(i_l)), \quad \text{with } I = (i, i_2, \ldots, i_l), J = (j_1, \ldots, j_l), z_{I,J}(i) \in \mathbb{R}^d
\]

and

\[
c_i' = c_i \prod_{t=2}^i g_{y_i^0}(y_t).
\]

Note that if for \( i = 1, \ldots, k_0 \) and \( j = 1, \ldots, t_i - t_i' - 1 \), there exists \( w_{i,j} \in \mathbb{R}^d \) such that

\[
\sum_{i_2, \ldots, i_l=1}^{k_0} c_i' \sum_{(j_2, \ldots, j_l) \in J(i_2) \times \cdots \times J(i_l)} \alpha_{I,J}z_{I,J}(i)T D^2g_{y_i^0}(y_1)z_{I,J}(i) = w_{i,j}T D^2g_{y_i^0}(y_1)w_{i,j},
\]

where possibly \( w_{i,j} = 0 \). Let \( \alpha_{i,j} = \|w_{i,j}\|^2/(\sum_{j=1}^{t_i - t_i' - 1} \|w_{i,j}\|^2) \) if there exists \( j \) such that \( \|w_{i,j}\|^2 > 0 \) and \( c_i' = \sum_{i_2, \ldots, i_l=1}^{k_0} c_i' \sum_{j=1}^{t_i - t_i' - 1} \|w_{i,j}\|^2 \), then

\[
\sum_{j=1}^{t_i - t_i' - 1} \sum_{i_2, \ldots, i_l=1}^{k_0} c_i' \sum_{(j_2, \ldots, j_l) \in J(i_2) \times \cdots \times J(i_l)} \alpha_{I,J}z_{I,J}(i)T D^2g_{y_i^0}(y_1)z_{I,J}(i) = c_i' \sum_{j=1}^{t_i - t_i' - 1} \alpha_{i,j}w_{i,j}T D^2g_{y_i^0}(y_1)w_{i,j}
\]

and (10) implies that

\[
a_i' = c_i' = 0, \quad b_i' = 0, \quad i = 1, \ldots, k_0, \quad \pi_i' = 0, \quad i = t_k + 1, \ldots, k.
\]

Simple calculations imply that

\[
\pi_I' = \sum_{i_2, \ldots, i_l=1}^{k_0} \prod_{t=2}^l g_{y_i^0}(y_t) = 0 \quad \Leftrightarrow \quad \forall (i_2, \ldots, i_l) \in \{1, \ldots, k\}^{l-2} \pi_{i,i_2,\ldots,i_l} = 0
\]
and similarly if \( i \) is such that there exists \( j = 1, \ldots, t_i - t_{i-1} \), \( I = (i, i_2, \ldots, i_l) \) and \( J = (j, j_2, \ldots, j_l) \in J(i) \times \cdots \times J(i_l) \) such that \( c_I > 0, \alpha_J > 0 \) and \( \|z_{I, J}(i)\| > 0 \), then \( c_{i, i_2, \ldots, i_l} = 0 \) for all \( i_2, \ldots, i_l \). Else, by considering \( y_t \) for some other \( t \), we obtain that (28) implies that

\[
\pi_I = 0 \quad \forall I \notin \{1, \ldots, t_{k_0}\}^l, \quad c_I = 0 \quad \forall I \in \{1, \ldots, t_{k_0}\}^l.
\]

This leads to

\[
b_i' = \sum_{i_2, \ldots, i_l=1}^{k_0} b_I \prod_{t \geq 2} g_{\gamma_0, \theta}(y_t) = 0 \quad \forall i = 1, \ldots, k_0.
\]

A simple recursive argument implies that \( b_I = 0 \) for all \( I \in \{1, \ldots, t_{k_0}\}^l \) which in turns implies that \( a_I = 0 \) for all \( I \in \{1, \ldots, t_{k_0}\}^l \) and condition \( \textbf{A4bis} \) is verified.

### 4.3. Proof of Theorem 3

First, we obtain the following lemma.

**Lemma 2.** Under the assumptions of Theorem 3, for any sequence \( M_n \) tending to infinity,

\[
P_\Pi((p + q) \wedge (2 - (p + q)) \|f_{2, \theta} - f_{2, \theta_0}\|_1 \leq \frac{M_n}{\sqrt{n}}) = 1 + o_{P_{\theta_0}}(1).
\]

We prove Lemma 2 by applying Theorem 4, using some of the computations of the proof of Theorem 1 but verifying assumption \( \textbf{C3bis} \) instead of \( \textbf{C3} \). Set \( S_n = U_n \times \mathcal{X} \) with

\[
U_n = \left\{ \theta = (p, q, \gamma_1, \gamma_2) : \|\gamma_1 - \gamma_0\|^2 \leq \frac{1}{\sqrt{n}}, \|\gamma_2 - \gamma_0\|^2 \leq \frac{1}{\sqrt{n}}, \|q(\gamma_1 - \gamma_0) + p(\gamma_2 - \gamma_0)\| \leq \frac{1}{\sqrt{n}}, q \leq \frac{1}{2} \leq \epsilon, p \leq \frac{1}{2} \leq \epsilon \right\}
\]

for small but fixed \( \epsilon \). We shall prove later the following lemma.

**Lemma 3.** Let \( M_n \) tend to infinity. Then

\[
\sup_{(\theta, x) \in S_n} P_{\theta_0}[\ell_n(\theta, x) - \ell_n(\theta_0, x_0) < -M_n] = o(1)
\]

and

\[
\Pi(S_n) \geq n^{-3d/4}.
\]

Now we prove that assumption \( \textbf{C3bis} \) holds with \( \epsilon_n = M_n / \sqrt{n} \), which will finish the proof of Lemma 2. By Proposition 1, we obtain that there exists \( c(\theta_0) > 0 \) and \( \eta > 0 \) such that:
• If \( \|y_1 - y^0\| \leq \eta \) and \( \|y_2 - y^0\| \leq \eta \),

\[
\|f_{2,\theta} - f_{2,\theta_0}\|_1 \\
\geq c(\theta_0) \frac{1}{p + q} \left[ q (y_1 - y^0) + p (y_2 - y^0) \right] + q \|y_1 - y^0\|^2 + p \|y_2 - y^0\|^2].
\]

• If \( \|y_1 - y^0\| \leq \eta \) and \( \|y_1 - y^0\| + \|y_2 - y^0\| > 2\eta \),

\[
\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0) \left[ p \frac{p}{p + q} + q \frac{q}{p + q} \|y_1 - y^0\| \right].
\]

• If \( \|y_2 - y^0\| \leq \eta \) and \( \|y_1 - y^0\| + \|y_2 - y^0\| > 2\eta \),

\[
\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0) \left[ q \frac{q}{p + q} + p \frac{p}{p + q} \|y_2 - y^0\| \right].
\]

• If \( \|y_1 - y^0\| > \eta \) and \( \|y_2 - y^0\| > \eta \),

\[
\|f_{2,\theta} - f_{2,\theta_0}\|_1 \geq c(\theta_0).
\]

Similar upper bounds hold also by Taylor expansion. Thus, for any \( m \), \( A_{n,m}(\epsilon_n) \) is a subset of the set of \( \theta \)'s such that

\[
\min \left\{ \frac{(p + q) \wedge (2 - (p + q))}{p + q} \left[ q (y_1 - y^0) + p (y_2 - y^0) \right] + q \|y_1 - y^0\|^2 + p \|y_2 - y^0\|^2 \right\};
\]

\[
\frac{(p + q) \wedge (2 - (p + q))}{p + q} \left[ p + q \|y_1 - y^0\| \right];
\]

\[
\frac{(p + q) \wedge (2 - (p + q))}{p + q} \left[ q + p \|y_2 - y^0\| \right]; (p + q) \wedge (2 - (p + q)) \right\} \lesssim (m + 1)\epsilon_n.
\]

This leads to

\[
\Pi_2(A_{n,m}(\epsilon_n)) \lesssim [(m + 1)\epsilon_n]^{2\alpha} + [(m + 1)\epsilon_n]^{2\beta} + [(m + 1)\epsilon_n]^{\alpha + d}
\]

so that if \( \alpha, \beta > 3d/4 \) and (29) holds, there exists \( \delta > 0 \) such that

\[
\frac{\Pi_2(A_{n,m}(\epsilon_n))e^{-(nm^2\epsilon_n^2)/(32l)}}{\Pi(S_n)} \lesssim n^{-\delta}[(M_n m)^{2\alpha} + (M_n m)^{2\beta} + (M_n m)^{\alpha + d}]e^{-(M_n^2 m^2)/(32l)}.
\]

Also for all \( \epsilon > 0 \) small enough \( A_{n,m}(\epsilon) \) contains the set of \( \theta \)'s such that

\[
\max \left\{ \frac{(p + q) \wedge (2 - (p + q))}{p + q} \right\} \times \left[ q (y_1 - y^0) + p (y_2 - y^0) \right] + q \|y_1 - y^0\|^2 + p \|y_2 - y^0\|^2 \right\};
\]
Posterior distribution for HMM

\[
\frac{(p + q) \land (2 - (p + q))}{p + q} \left[ p + q \parallel \gamma_1 - \gamma^0 \parallel \right];
\]

\[
\frac{(p + q) \land (2 - (p + q))}{p + q} \left[ q + p \parallel \gamma_2 - \gamma^0 \parallel \right]; (p + q) \land (2 - (p + q)) \right\} \lesssim (m + 1) \epsilon
\]

therefore

\[
N \left( \frac{m \epsilon_n}{12}, A_{n,m}(\epsilon_n), d_l(\cdot, \cdot) \right) \lesssim m^{2+2d} \lesssim e^{(\epsilon_n^2 m^2 (\rho_0 - 1)^2)/(16l(2+\rho_0 - 1)^2)},
\]

so that assumption **C3bis** is verified.

We now prove Theorem 3. Notice first that, by setting

\[
D_n = \int_{\Theta \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \Pi_2(d\theta) \pi_\mathcal{X}(dx),
\]

as in the proof of Theorem 4 we get that for any sequence \( C_n \) tending to infinity,

\[
\mathbb{P}_{\theta_0} \left( D_n \leq C_n n^{-D/2} \right) = o(1)
\]

with \( D = d + d/2 \).

Let now \( \epsilon_n \) be any sequence going to 0 and let \( A_n = \{ -\frac{p}{p+q} \leq \epsilon_n or \frac{q}{p+q} \leq \epsilon_n \} \). For some sequence \( M_n \) going to infinity and \( \delta_n = M_n/\sqrt{n} \), let \( B_n = \{(p + q) \land (2 - (p + q)) \parallel f_{2,\theta} - f_{2,\theta_0} \parallel \leq \delta_n \} \). We then control with \( D = d + d/2 \), using Lemma 2

\[
E_{\theta_0} \left[ \mathbb{P}_{\Pi}(A_n \mid Y_{1:n}) \right] = E_{\theta_0} \left[ \mathbb{P}_{\Pi}(A_n \cap B_n \mid Y_{1:n}) \right] + o(1)
\]

\[
= E_{\theta_0} \left[ \int_{A_n \cap B_n \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \Pi_2(d\theta) \pi_\mathcal{X}(dx) \right] + o(1)
\]

\[
:= E_{\theta_0} \left[ \frac{N_n}{D_n} \right] + o(1)
\]

\[
\leq \mathbb{P}_{\theta_0} \left( D_n \leq C_n n^{-D/2} \right) + \frac{n^{D/2}}{C_n} \Pi_2(A_n \cap B_n) + o(1).
\]

Thus using (30), the first part of Theorem 3 is proved by showing that

\[
\Pi_2(A_n \cap B_n) \lesssim \delta_n^{2^a + \delta_n^{d+d/2}} \epsilon_n^{a-d/2}.
\]

Then, the second part of Theorem 3 follows from its first part and Lemma 2.

We now prove that (31) holds. Define

\[
B_n^1 = \left\{ \frac{(p + q) \land (2 - (p + q))}{p + q}
\right. \\
\times \left[ \| q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \| + q\| \gamma_1 - \gamma^0 \|^2 + p\| \gamma_2 - \gamma^0 \|^2 \right] \leq \delta_n \right\},
\]
Let us now study \( \ell_n(\theta) \)

Let us now study \( \ell_n(\theta, x) \)

Notice that on \( \theta \)

\[ x \]

not depend on \( \theta \).

\[ \lim_{n \to \infty} P(A_n \cap B_n) = \lim_{n \to \infty} P(A_n \cap B_n^1) + \lim_{n \to \infty} P(A_n \cap B_n^2) + \lim_{n \to \infty} P(A_n \cap B_n^3) + \lim_{n \to \infty} P(A_n \cap B_n^4). \]

Notice that on \( A_n \) if \( p + q \geq 1 \), then \( p \leq \epsilon_n \) and \( q \geq 1 - \epsilon_n \), or \( q \leq \epsilon_n \) and \( p \geq 1 - \epsilon_n \), so that also \( 2 - (p + q) \geq 1 - \epsilon_n \).

- On \( A_n \cap B_n^1 \), \( \max(q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)) \prec \delta_n \), \( q\|\gamma_1 - \gamma^0\|^2 \prec \delta_n \), \( p\|\gamma_2 - \gamma^0\|^2 \prec \delta_n \), and \( p \prec \epsilon_n \) or \( q \prec \epsilon_n \). This gives \( \Pi_2(A_n \cap B_n^1) \prec \delta_n^{d + d/2} \epsilon_n^{\alpha - d/2} \).
- On \( A_n \cap B_n^2 \), \( p \prec \delta_n \) and \( q\|\gamma_1 - \gamma^0\| \prec \delta_n \) in case \( p + q \leq 1 \), and \( p \leq \delta_n \), \( 1 - q \leq \delta_n \) and \( q\|\gamma_1 - \gamma^0\| \prec \delta_n \) in case \( p + q \geq 1 \), leading to \( \Pi_2(A_n \cap B_n^2) \prec \delta_n^{d + d} + \delta_n^{\alpha + \beta + d} \).
- For symmetry reasons, \( \Pi_2(A_n \cap B_n^3) = \Pi_2(A_n \cap B_n^4) \).
- On \( A_n \cap B_n^4 \), \( p \prec \delta_n \) and \( q \prec \delta_n \), so that \( \Pi_2(A_n \cap B_n^4) \prec \delta_n^{2\alpha} \).

Keeping only the leading terms, we see that (31) holds and this terminates the proof Theorem 3.

We now prove Lemma 3. We easily get \( \Pi_2(U_n) \gtrsim n^{-3d/4} \), and

\[ D_n \geq \int_{U_n \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \Pi_2(d\theta|\mathcal{X})(d\mathcal{x}). \]

Let us now study \( \ell_n(\theta, x) - \ell_n(\theta_0, x) \). First, following the proof of Lemma 2 of Douc et al. [5] we find that, for any \( \theta \in U_n \), for any \( x \),

\[ \left| \ell_n(\theta) - \ell_n(\theta, x) \right| \leq \left( \frac{1 + 2\epsilon}{1 - 2\epsilon} \right)^2, \]

where \( \ell_n(\theta) = \sum_{x=1}^k \mu_\theta(x) \ell_n(\theta, x) \). Thus, for any \( \theta \in U_n \) and any \( x \), and since \( \ell_n(\theta_0, x) \) does not depend on \( x \),

\[ \ell_n(\theta, x) - \ell_n(\theta_0, x) \geq \ell_n(\theta) - \ell_n(\theta_0) - \left( \frac{1 + 2\epsilon}{1 - 2\epsilon} \right)^2. \]

(32)

Let us now study \( \ell_n(\theta) - \ell_n(\theta_0) \).

\[ \ell_n(\theta) - \ell_n(\theta_0) = \sum_{k=1}^n \log \left[ \mathbb{P}_\theta(X_k = 1|Y_{1:k-1}) \frac{g_{\theta,1}}{g_{\gamma,0}}(Y_k) + \mathbb{P}_\theta(X_k = 2|Y_{1:k-1}) \frac{g_{\theta,2}}{g_{\gamma,0}}(Y_k) \right] \]
and we set for $k = 1$

$$
P_\theta(X_k = 1|Y_{1:k-1}) = P_\theta(X_1 = 1) = \frac{q}{p+q},
$$

$$
P_\theta(X_k = 2|Y_{1:k-1}) = P_\theta(X_1 = 2) = \frac{p}{p+q}.
$$

Denote $p_k(\theta)$ the random variable $P_\theta(X_k = 1|Y_{1:k-1})$, which is a function of $Y_{1:k-1}$ and thus independent of $Y_k$. We have the recursion

$$
p_{k+1}(\theta) = \frac{(1-p)p_k(\theta)g_{\gamma_1}(Y_k) + q(1-p_k(\theta))g_{\gamma_2}(Y_k)}{p_k(\theta)g_{\gamma_1}(Y_k) + (1-p_k(\theta))g_{\gamma_2}(Y_k)}. \quad (33)
$$

Note that, for any $p, q$ in $]0, 1[$, for any $k \geq 1$,

$$
p_k(p, q, \gamma^0, \gamma^0) = \frac{q}{p+q}.
$$

We shall denote by $D^{j}_{(\gamma_1), \ldots, (\gamma_j), (\gamma_j)}$ the $i$th partial derivative operator $j$ times with respect to $\gamma_1$ and $i-j$ times with respect to $\gamma_j$ ($0 \leq j \leq i$, the order in which derivatives are taken does not matter). Fix $\theta = (p, q, \gamma_1, \gamma_2) \in U_n$. When derivatives are taken at point $(p, q, \gamma^0, \gamma^0)$, they are written with 0 as superscript.

Using Taylor expansion till order 4, there exists $t \in [0, 1]$ such that denoting $\theta_t = t\theta + (1-t)(p, q, \gamma^0, \gamma^0)$:

$$
\ell_n(\theta) - \ell_n(\theta_0) = (\gamma_1 - \gamma^0)D_{\gamma_1}^1\ell^0_n + (\gamma_2 - \gamma^0)D_{\gamma_2}^1\ell^0_n + S_n(\theta) + T_n(\theta) + R_n(\theta, t), \quad (34)
$$

where $S_n(\theta)$ denotes the term of order 2, $T_n(\theta)$ denotes the term of order 3, and $R_n(\theta, t)$ the remainder, that is

$$
S_n(\theta) = (\gamma_1 - \gamma^0)^2D_{(\gamma_1)^2}^2\ell^0_n + 2(\gamma_1 - \gamma^0)(\gamma_2 - \gamma^0)D_{\gamma_1, \gamma_2}^2\ell^0_n + (\gamma_2 - \gamma^0)^2D_{(\gamma_2)^2}^2\ell^0_n,
$$

$$
T_n(\theta) = (\gamma_1 - \gamma^0)^3D_{(\gamma_1)^3}^3\ell^0_n + 3(\gamma_1 - \gamma^0)^2(\gamma_2 - \gamma^0)D_{\gamma_1, \gamma_2}^3\ell^0_n
$$

$$
+ 3(\gamma_1 - \gamma^0)(\gamma_2 - \gamma^0)^2D_{\gamma_1, \gamma_2}^3\ell^0_n + (\gamma_2 - \gamma^0)^3D_{(\gamma_2)^3}^3\ell^0_n
$$

and

$$
R_n(\theta, t) = \sum_{k=0}^4 \binom{k}{4} (\gamma_1 - \gamma^0)^k(\gamma_2 - \gamma^0)^{4-k}D_{(\gamma_1)^k, (\gamma_2)^{4-k}}^4\ell_n(\theta_t).
$$

Easy but tedious computations lead to the following results.

$$
(\gamma_1 - \gamma^0)D_{\gamma_1}^1\ell^0_n + (\gamma_2 - \gamma^0)D_{\gamma_2}^1\ell^0_n = \left[ \sum_{k=1}^n \frac{D_{\gamma_1}^k g_{\gamma_0}(Y_k)}{g_{\gamma_0}} \right] \left[ q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \right] / (p+q)
$$

$$
= \left[ \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{D_{\gamma_1}^k g_{\gamma_0}(Y_k)}{g_{\gamma_0}} \right] \left[ \sqrt{n} q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0) \right] / (p+q)
$$
so that

$$\sup_{\theta \in U_n} \left| (\gamma_1 - \gamma^0) D_{\gamma_1}^1 \ell^0_n + (\gamma_2 - \gamma^0) D_{\gamma_2}^1 \ell^0_n \right| = O_{P_{\theta_0}}(1).$$

(35)

Also,

$$S_n(\theta) = \left[ \frac{1}{n} \sum_{k=1}^{n} \left( \frac{D_{\gamma_1}^1 g_{\gamma_0}(Y_k)}{g_{\gamma_0}} (Y_k) \right)^2 \right]^{1/2} \left[ \frac{\sqrt{n} q (\gamma_1 - \gamma^0) + p (\gamma_2 - \gamma^0)}{p + q} \right]^2$$

$$+ \left[ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{D_{\gamma_2}^2 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right] \left[ \frac{q (n^{1/4} (\gamma_1 - \gamma^0))^2 + p (n^{1/4} (\gamma_2 - \gamma^0))^2}{p + q} \right]$$

$$+ 2(n^{1/4} (\gamma_1 - \gamma^0))^2 \left[ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (D_{\gamma_1}^1 p_k) \frac{D_{\gamma_1}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right]$$

$$- 2(n^{1/4} (\gamma_2 - \gamma^0))^2 \left[ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (D_{\gamma_2}^1 p_k) \frac{D_{\gamma_2}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right]$$

$$+ 2(n^{1/4} (\gamma_1 - \gamma^0) (\gamma_2 - \gamma^0)) \left[ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (D_{\gamma_2}^1 p_k - D_{\gamma_1}^1 p_k) \frac{D_{\gamma_1}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right].$$

Using (33) one gets that for all integer $k \geq 2$ ($D_{\gamma_1}^1 p_1^0 = 0$ and $D_{\gamma_2}^1 p_1^0 = 0$):

$$D_{\gamma_1}^1 p_k^0 = \frac{pq}{(p + q)^2} \sum_{l=1}^{k-1} (1 - p - q)^{k-l} \frac{D_{\gamma_1}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_l)$$

and

$$D_{\gamma_2}^1 p_k^0 = -D_{\gamma_1}^1 p_k^0$$

which leads to

$$E_{\theta_0} \left[ \left( \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (D_{\gamma_1}^1 p_k^0) \frac{D_{\gamma_1}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right)^2 \right] \leq \left( E_{\theta_0} \left( \frac{D_{\gamma_1}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_1) \right)^2 \right)^2$$

and

$$E_{\theta_0} \left[ \left( \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (D_{\gamma_2}^1 p_k^0) \frac{D_{\gamma_2}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_k) \right)^2 \right] \leq \left( E_{\theta_0} \left( \frac{D_{\gamma_2}^1 g_{\gamma_0}}{g_{\gamma_0}} (Y_1) \right)^2 \right)^2.$$

Thus, we obtain

$$\sup_{\theta \in U_n} \left| S_n(\theta) \right| = O_{P_{\theta_0}}(1).$$

(36)
For the order 3 term, as soon as \( \theta \in U_n \):

\[
T_n(\theta) = -\left[ \sum_{k=1}^{n} \left( \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right)^3 \right] \left[ \frac{q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)}{p+q} \right]^3 \\
+ \left[ \sum_{k=1}^{n} \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right] \left[ \frac{q}{p+q} (\gamma_1 - \gamma^0)^3 + \frac{p}{p+q} (\gamma_2 - \gamma^0)^3 \right] \\
- 3 \left[ \sum_{k=1}^{n} \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right] \left[ \frac{q}{g_{\gamma_{0}}(Y_k)} \right] \frac{g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \left[ \frac{q(\gamma_1 - \gamma^0) + p(\gamma_2 - \gamma^0)}{p+q} \right] \\
\times \left[ \frac{q}{(p+q)^2} (\gamma_1 - \gamma^0)^2 + \frac{p}{(p+q)^2} (\gamma_2 - \gamma^0)^2 \right] \\
+ O(n^{-3/4}) \left\{ \sum_{k=1}^{n} \left( \frac{D_{\gamma}^{1} p_{k}^0}{g_{\gamma_{0}}(Y_k)} \right)^2 \right. \\
+ \sum_{k=1}^{n} \left( \frac{D_{\gamma}^{1} p_{k}^0}{g_{\gamma_{0}}(Y_k)} \right) \frac{D_{\gamma}^{2} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} + \sum_{k=1}^{n} \left( \frac{D_{\gamma}^{2} p_{k}^0}{g_{\gamma_{0}}(Y_k)} \right) \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \\
\left. + \sum_{k=1}^{n} \left( \frac{D_{\gamma}^{2} p_{k}^0}{g_{\gamma_{0}}(Y_k)} \right) \frac{D_{\gamma}^{2} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right\} \\
\}
\]

so that using assumptions (14)

\[
\sup_{\theta \in U_n} |T_n(\theta)| = O_{\bar{p}}^{0} \left( n^{-1/4} \right) + O_{\bar{p}}^{0} (1) + O(n^{-1/4}) Z_n
\]

with

\[
Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left\{ \left( \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right) + \frac{D_{\gamma}^{2} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right\} D_{\gamma}^{1} p_{k}^0 \\
+ \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \left[ D_{\gamma}^{2} p_{k}^0 + D_{\gamma}^{2} p_{k}^0 + D_{\gamma,\gamma}^{2} p_{k}^0 \right] \\
\}
\]

Now using (33) one gets that for all integer \( k \geq 1 \),

\[
\frac{1}{1 - p - q} D_{\gamma,\gamma}^{2} p_{k+1}^0 = -2 \frac{pq^2}{(p+q)^3} \left( \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} \right)^2 \\
+ 2 \left( \frac{D_{\gamma}^{1} p_{k}^0}{g_{\gamma_{0}}(Y_k)} \right) \frac{D_{\gamma}^{1} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} + \frac{pq}{(p+q)^2} \frac{D_{\gamma}^{2} g_{\gamma_{0}}(Y_k)}{g_{\gamma_{0}}(Y_k)} + D_{\gamma,\gamma}^{2} p_{k}^0,
\]
\[
\frac{1}{1-p-q} D^2_{(\gamma_2)\gamma_2} p^0_{k+1} = 2 \frac{p^2 q}{(p+q)^3} \left( \frac{D^1_{\gamma} g_{\gamma_0} (Y_k)}{g_{\gamma_0}} \right)^2
\]
\[-2 \left( D^1_{\gamma_1} p^0_k \right) \frac{D^1_{\gamma} g_{\gamma_0} (Y_k)}{g_{\gamma_0}} - \frac{pq}{(p+q)^2} \frac{D^2_{\gamma_2} g_{\gamma_0} (Y_k)}{g_{\gamma_0}} + D^2_{(\gamma_2)\gamma_2} p^0_k.
\]
\[
\frac{1}{1-p-q} D^2_{(\gamma_1,\gamma_2)\gamma_2} p^0_{k+1} = 2 \frac{pq(q-p)}{(p+q)^3} \left( \frac{D^1_{\gamma} g_{\gamma_0} (Y_k)}{g_{\gamma_0}} \right)^2
\]
\[+2 \left( D^1_{\gamma_1} p^0_k \right) \frac{D^1_{\gamma} g_{\gamma_0} (Y_k)}{g_{\gamma_0}} + D^2_{(\gamma_1,\gamma_2)\gamma_2} p^0_k.
\]

and using \( D^2_{(\gamma_1)\gamma_1} p^0_1 = 0, D^2_{(\gamma_2)\gamma_2} p^0_1 = 0 \) and easy but tedious computations one gets that for some finite \( C > 0, \)
\[
E_{\theta_0} \left( Z^2_n \right) \leq CE_{\theta_0} \left( \frac{D^1_{\gamma} g_{\gamma_0} (Y_1)}{g_{\gamma_0}} \right)^2 \left[ E_{\theta_0} \left( \frac{D^1_{\gamma} g_{\gamma_0} (Y_1)}{g_{\gamma_0}} \right)^4
\]
\[+ E_{\theta_0} \left( \frac{D^2_{\gamma_1} g_{\gamma_0} (Y_1)}{g_{\gamma_0}} \right)^2 \left( E_{\theta_0} \left( \frac{D^1_{\gamma} g_{\gamma_0} (Y_1)}{g_{\gamma_0}} \right)^2 \right)^2 \right]
\]
so that we finally obtain
\[
\sup_{\theta \in U_n} |T_n(\theta)| = O_{\mathbb{P}_{\theta_0}}(1). \quad (37)
\]

Let us finally study the fourth order remainder \( R_n(\theta, t). \) We have
\[
\sup_{\theta \in U_n} |R_n(\theta, t)| \leq \frac{1}{n} \sum_{k=1}^n A_{k,n} B_{k,n},
\]
where, for big enough \( n, A_{k,n} \) is a polynomial of degree at most 4 in \( \sup_{\gamma' \in B_d(\gamma_0, \epsilon)} \| \frac{D^i_{\gamma} g_{\gamma'} (Y_k)}{g_{\gamma'}} \|, \)
and \( B_{k,n} \) is a sum of terms of form
\[
\sup_{\theta \in U_n} \left| \prod_{i=1}^4 \prod_{j=0}^{i-1} \left( D^i_{(\gamma_1)\gamma_2, (\gamma_2)^i-j} p_k(\theta) \right)^{a_{i,j}} \right|,
\]
where the \( a_{i,j} \) are non-negative integers such that \( \sum_{i=1}^4 \sum_{j=0}^{i-1} i a_{i,j} \leq 4. \)

To prove that
\[
\sup_{\theta \in U_n} |R_n(\theta, t)| = O_{\mathbb{P}_{\theta_0}}(1) \quad (39)
\]
holds, it is enough to prove that \( E_{\theta_0} \left| \sum_{k=1}^n A_{k,n} B_{k,n} \right| = O(n). \) But for each \( k, p_k(\theta) \) and its derivatives depend on \( Y_1, \ldots, Y_{k-1} \) only, so that \( A_{k,n} \) and \( B_{k,n} \) are independent random variables,
and
\[
E_{\theta_0} \left| \sum_{k=1}^{n} A_{k,n} B_{k,n} \right| \leq \sum_{k=1}^{n} E_{\theta_0} |A_{k,n}| E_{\theta_0} |B_{k,n}|
\]
\[
\leq C \max_{i=1,2,3,4} \sum_{k=1}^{n} E_{\theta_0} \left( \sup_{Y' \in B_d(y^0,\epsilon)} \left\| \frac{D_{1 i}^j g_{Y'}}{g_{Y'}} (Y_1) \right\|^4 \right)^{4}
\]
for some finite \( C > 0 \). Now, using (33) one gets that for all integer \( k \geq 1 \) and for any \( \theta \),
\[
D_{1 i}^1 p_{k+1}(\theta) = (1 - p - q) \left\{ \frac{p_k(\theta)(1 - p_k(\theta)) g_{Y_2}(Y_k) D_{1 i}^1 g_{Y_1}(Y_k) + g_{Y_1}(Y_k) g_{Y_2}(Y_k) D_{1 i}^1 p_k(\theta)}{(p_k(\theta) g_{Y_1}(Y_k) + (1 - p_k(\theta)) g_{Y_2}(Y_k))^2} \right\},
\]
\[
D_{1 i}^1 p_{k+1}(\theta) = (1 - p - q) \left\{ -p_k(\theta)(1 - p_k(\theta)) g_{Y_1}(Y_k) D_{1 i}^1 g_{Y_2}(Y_k) + g_{Y_1}(Y_k) g_{Y_2}(Y_k) D_{1 i}^1 p_k(\theta) \right\}.
\]
Notice that for any \( \theta \), any \( k \geq 2 \), \( p_k(\theta) \in (1 - p, q) \) so that for any \( \theta \in U_n \), any \( k \geq 2 \), \( p_k(\theta) \in [\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon] \). We obtain easily that for \( i = 1, 2, k \geq 2 \),
\[
\sup_{\theta \in U_n} \left| D_{1 i}^1 p_{k+1}(\theta) \right| \leq \left( \frac{2 \epsilon}{1 - 8 \epsilon} \right) \left\{ \sup_{Y' \in B_d(y^0,\epsilon)} \left\| \frac{D_{1 i}^1 g_{Y'}}{g_{Y'}} (Y_k) \right\| + \sup_{\theta \in U_n} \left| D_{1 i}^1 p_k(\theta) \right| \right\}.
\]
Using similar tricks, it is possible to get that there exists a finite constant \( C > 0 \) such that for any \( i = 1, 2, 3, 4 \), any \( j = 0, \ldots, i \), any \( k \geq 2 \),
\[
\sup_{\theta \in U_n} \left| D_{(Y_1)^i, (Y_2)^{i-j}}^i p_{k+1}(\theta) \right|
\]
\[
\leq C \epsilon \left\{ \sup_{Y' \in B_d(y^0,\epsilon)} \left\| \sum_{l=1}^{i} \frac{D_{1 i}^l g_{Y'}}{g_{Y'}} (Y_k) \right\|^{i+1-l} + \sum_{l=1}^{i} \sum_{m=0}^{i} \sup_{\theta \in U_n} \left| D_{(Y_1)^i, (Y_2)^{i-j}}^l p_k(\theta) \right|^{i+1-l} \right\}.
\]
By recursion, we obtain that there exists a finite \( C > 0 \) such that any term of form (38) has expectation uniformly bounded:
\[
E_{\theta_0} \left[ \sup_{\theta \in U_n} \prod_{i=1}^{4} \prod_{j=0}^{i} \left( D_{(Y_1)^i, (Y_2)^{i-j}}^i p_k(\theta) \right)^{a_{i,j}} \right]
\]
\[
\leq C \max_{m=1,2,3,4} \max_{r=1,2,3,4} \sup_{\theta \in U_n} \left( \sup_{Y' \in B_d(y^0,\epsilon)} \left\| \frac{D_{y^r}^m g_{Y'}}{g_{Y'}} (Y_1) \right\|^{r} \right).
\]
which concludes the proof of (39). Now, using (32), (34), (35), (36), (37) and (39), we get
\[ D_n \geq e^{-O_p(\theta_0)} \Pi_2(U_n) \]
so that (20) holds with \( S_n \) satisfying (29).

4.4. Proof of Theorem 4

The proof follows the same lines as in Ghosal and van der Vaart [10]. We write
\[
\mathbb{P} \left[ \left\| f_{l,\theta} - f_{l,\theta_0} \right\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n \left| Y_{1:n} \right| \right] = \frac{\int_{A_n \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \Pi(\theta) \pi(\mathcal{X}) (dx) \rho_\theta}{\int_{\Theta \times \mathcal{X}} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x)} \Pi(\theta) \pi(\mathcal{X}) (dx)} := \frac{N_n}{\Pi_1(S_n)}
\]
where \( A_n = \{ \theta : \left\| f_{l,\theta} - f_{l,\theta_0} \right\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n \} \). A lower bound on \( D_n \) is obtained in the following usual way. Set \( \Omega_n = \{ (\theta, x) ; \ell_n(\theta, x) - \ell_n(\theta_0, x_0), \leq -n\epsilon_n^2 \} \), which is a random subset of \( \Theta \times \mathcal{X} \) (depending on \( Y_{1:n} \)),
\[
D_n \geq \int_{\Omega_n} e^{\ell_n(\theta, x) - \ell_n(\theta_0, x_0)} \Pi(\theta) \pi(\mathcal{X}) (dx) \geq e^{-n\epsilon_n^2} \Pi(S_n \cap \Omega_n),
\]
therefore
\[
\mathbb{P}_{\theta_0} \left[ D_n < e^{-n\epsilon_n^2} \Pi(S_n)/2 \right] \leq \mathbb{P}_{\theta_0} \left[ \Pi(S_n \cap \Omega_n) \geq \Pi(S_n)/2 \right] \leq 2 \frac{\int_{\Omega_n} \Pi(\theta_0) \left( \ell_n(\theta, x) - \ell_n(\theta_0, x_0) \leq -n\epsilon_n^2 \right) \Pi(\theta) \pi(\mathcal{X}) (dx)}{\Pi(S_n)} = o(1)
\]
and
\[
\mathbb{P} \left[ \left\| f_{l,\theta} - f_{l,\theta_0} \right\|_1 \frac{\rho_\theta - 1}{2R_\theta + \rho_\theta - 1} \geq \epsilon_n \left| Y_{1:n} \right| \right] = o_{\mathbb{P}_{\theta_0}} (1) + \frac{N_n}{D_n} \frac{1}{2 \Pi_2 e^{-n\epsilon_n^2} \Pi(S_n)}.
\]
and
\[
E_{\theta_0} \left[ \int_{(A_n \cap F_n) \times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \Pi_{\Theta}(d\theta) \pi_X(dx) \right] \\
= O\left[ \Pi_{\Theta}(A_n \cap F_n) \right] = o\left( e^{-n\tilde{\epsilon}_2(C_n+1)} \right)
\]
by Fubini's theorem and assumption C2 together with the fact that \( \ell_n(\theta_0) - \ell_n(\theta_0,x_0) \) is uniformly upper bounded. This implies using assumption C1 that
\[
\mathbb{P}_{\Pi} \left[ \left\| f_{i,\theta} - f_{i,\theta_0} \right\| 1 \geq \epsilon_n \left| Y_{1:n} \right| \right] = o_{\theta_0} (1) + \frac{\tilde{N}_n}{D_n} 2D_n \geq e^{-n\tilde{\epsilon}_2} \Pi(S_n),
\]
where \( \tilde{N}_n = \int_{(A_n \cap F_n) \times \mathcal{X}} e^{\ell_n(\theta,x)-\ell_n(\theta_0,x_0)} \Pi_{\Theta}(d\theta) \pi_X(dx) \). Let now \( (\theta_j)_{j=1}^{N} = N(\delta, F_n, d_l(\cdot, \cdot)) \), be the sequence of \( \theta_j \)'s in \( F_n \) such for all \( \theta \in F_n \) there exists a \( \theta_j \) with \( d_l(\theta_j, \theta) \leq \delta \) with \( \delta = \epsilon_n/12 \). Assume for simplicity's sake and without loss of generality that \( n \) is a multiple of the integer \( l \), and define
\[
\phi_j = \frac{1}{\sum_{i=1}^{n/l} (\mathbb{1}(Y_{li+1},...,Y_{li}) \in A_j - \mathbb{P}_{\theta_0}(Y_{1},...,Y_l) \in A_j)) > t_j},
\]
where
\[
A_j = \{ (y_1, \ldots, y_l) \in \mathcal{Y}^l : f_{i,\theta_0}(y_1, \ldots, y_l) \leq f_{i,\theta_j}(y_1, \ldots, y_l) \}
\]
for some positive real number \( t_j \) to be fixed later also. Note that
\[
\mathbb{P}_{\theta_j}(Y_{1},\ldots,Y_l) \in A_j) - \mathbb{P}_{\theta_0}(Y_{1},\ldots,Y_l) \in A_j) = \frac{1}{2} \left\| f_{i,\theta_j} - f_{i,\theta_0} \right\| 1.
\]
Define also
\[
\psi_n = \max_{1 \leq j \leq N, \theta_j \in A_n} \phi_j.
\]
Then
\[
E_{\theta_0} \left( \frac{\tilde{N}_n}{D_n} \psi_n \right) \leq E_{\theta_0} \psi_n \leq N(\delta, F_n, d(\cdot, \cdot)) \max_{1 \leq j \leq N, \theta_j \in A_n} E_{\theta_0} \phi_j
\]
and
\[
E_{\theta_0}(\tilde{N}_n (1 - \psi_n)) = \int_{\mathcal{X}} E_{\theta_0,x_0}(\tilde{N}_n (1 - \psi_n)) \mu_{\theta_0}(dx_0)
= \int_{(A_n \cap F_n) \times \mathcal{X}} E_{\theta,x}(1 - \psi_n)) \Pi_{\Theta}(d\theta) \pi_X(dx).
\]
Now
\[
E_{\theta_0}[\phi_j] = \mathbb{P}_{\theta_0} \left[ \sum_{i=1}^{n/l} (\mathbb{1}(Y_{li+1},...,Y_{li}) \in A_j - \mathbb{P}_{\theta_0}(Y_{1},...,Y_l) \in A_j)) > t_j \right]
\]
and

\[ E_{\theta, x}(1 - \phi_j) \]
\[ = \mathbb{P}_{\theta, x} \left[ \sum_{i=1}^{n/l} \left( -\mathbb{1}(Y_{li-l+1}, \ldots, Y_{li}) \in A_j + \mathbb{P}_{\theta, x} \left( (Y_{li-l+1}, \ldots, Y_{li}) \in A_j \right) \right) \right] > -t_j + \sum_{i=1}^{n/l} \left( \mathbb{P}_{\theta, x} \left( (Y_{li-l+1}, \ldots, Y_{li}) \in A_j \right) - \mathbb{P}_0 \left( (Y_1, \ldots, Y_l) \in A_j \right) \right). \]

Consider the sequence \((Z_i)_{i \geq 1}\) with for all \(i \geq 1\), \(Z_i = (X_{li-l+1}, \ldots, X_{li}, Y_{li-l+1}, \ldots, Y_{li})\), which is, under \(\mathbb{P}_\theta\), a Markov chain with transition kernel \(\tilde{Q}_\theta\) given by

\[ \tilde{Q}_\theta(z, dz') = g_\theta(y'_1|x'_1) \cdots g_\theta(y'_l|x'_l) Q_\theta(x'_1, dx'_1) Q_\theta(x'_2, dx'_2) \cdots Q_\theta(x'_{l-1}, dx'_l) \mu(dy'_1) \cdots \mu(dy'_l). \]

This kernel satisfies the same uniform ergodic property as \(Q_\theta\), with the same coefficients, that is condition (17) holds with the coefficients \(R_{\theta}\) and \(\rho_{\theta}\) with the replacement of \(Q_\theta\) by \(\tilde{Q}_\theta\), and we may use Rio’s [18] exponential inequality (Corollary 1) with uniform mixing coefficients (as defined in Rio [18]) satisfying \(\phi(m) \leq R_{\theta} \rho_{\theta}^{-m}\). Indeed, by the Markov property,

\[ \phi(m) = \sup_{A \in \sigma(Z_1), B \in \sigma(Z_{m+1})} (\mathbb{P}_\theta(B) - \mathbb{P}_\theta(B|A)) \]
\[ \leq \sup_z |\mathbb{P}_\theta(Z_{m+1} \in B) - \mathbb{P}_\theta(Z_{m+1} \in B|Z_1 = z)| \]
\[ \leq R_{\theta} \rho_{\theta}^{-m}. \]

We thus obtain that, for any positive real number \(u\),

\[ \mathbb{P}_0 \left[ \sum_{i=1}^{n/l} \left( \mathbb{1}(Y_{li-l+1}, \ldots, Y_{li}) \in A_j - \mathbb{P}_0 \left( (Y_1, \ldots, Y_l) \in A_j \right) \right) > u \right] > \exp \left\{ \frac{-2lu^2(\rho_0 - 1)^2}{n(2R_{\theta} + \rho_0 - 1)^2} \right\} \quad (43) \]

and

\[ \mathbb{P}_{\theta, x} \left[ \sum_{i=1}^{n/l} \left( -\mathbb{1}(Y_{li-l+1}, \ldots, Y_{li}) \in A_j + \mathbb{P}_{\theta, x} \left( (Y_{li-l+1}, \ldots, Y_{li}) \in A_j \right) \right) > u \right] \]
\[ \leq \exp \left\{ \frac{-2lu^2(\rho_{\theta} - 1)^2}{n(2R_{\theta} + \rho_{\theta} - 1)^2} \right\}. \quad (44) \]
Set now
\[ t_j = \frac{n \| f_{i,j} - f_{i,0} \|_1}{4l}. \]

Since for any \( \theta \), \( \frac{\rho_0 - 1}{2R_0 + \rho_0 - 1} \leq 1 \) and consequently for \( \theta_j \in A_n \), \( \| f_{i,j} - f_{i,0} \|_1 \geq \epsilon_n \), we first get, using (43),
\[ E_{\theta_0}[\phi_j] \leq \exp \left\{ -\frac{n \epsilon_n^2 (\rho_{\theta_0} - 1)^2}{8l(2R_0 + \rho_{\theta_0} - 1)^2} \right\}. \tag{45} \]

Now, for any \( \theta \in A_n \),
\[ -t_j + \frac{n}{l} \sum_{i=1}^{n/l} \left( \mathbb{P}_{\theta,x}((Y_{li}, Y_{li+1}) \in A_j) - \mathbb{P}_{\theta_0}((Y_{li}, Y_{li+1}) \in A_j) \right) \]
\[ = -\frac{n \| f_{i,j} - f_{i,0} \|_1}{4l} + \frac{n}{l} \left\{ \mathbb{P}_{\theta_j}((Y_{li}, Y_{li+1}) \in A_j) - \mathbb{P}_{\theta_0}((Y_{li}, Y_{li+1}) \in A_j) \right\} \]
\[ + \frac{n}{l} \left\{ \mathbb{P}_{\theta_j}((Y_{li}, Y_{li+1}) \in A_j) - \mathbb{P}_{\theta_0}((Y_{li}, Y_{li+1}) \in A_j) \right\} \]
\[ + \sum_{i=1}^{n/l} \left( \mathbb{P}_{\theta,x}((Y_{li}, Y_{li+1}) \in A_j) - \mathbb{P}_{\theta}((Y_{li}, Y_{li+1}) \in A_j) \right) \]
\[ \geq \frac{n \| f_{i,j} - f_{i,0} \|_1}{4l} - \frac{n \| f_{i,j} - f_{i,0} \|_1}{l} - \sum_{i=1}^{n/l} R_\theta \rho_{\theta_0}^{-i} \]
\[ \geq \frac{n \| f_{i,j} - f_{i,0} \|_1}{4l} - \frac{n \| f_{i,j} - f_{i,0} \|_1}{l} - \frac{R_\theta \rho_{\theta_0}}{\rho_{\theta_0} - 1} \]
\[ \geq \frac{n}{4l} \left( 1 - \frac{5}{12} - \frac{4l}{12n \epsilon_n} \right) \| f_{i,j} - f_{i,0} \|_1 \geq \frac{n}{8l} \| f_{i,j} - f_{i,0} \|_1 \]

for large enough \( n \), using the triangular inequality and the fact that \( \| f_{i,j} - f_{i,0} \|_1 \leq \frac{\epsilon_n}{12} \leq \frac{\rho_0 - 1}{2R_0 + \rho_0 - 1} \) since \( \theta \in A_n \) and \( \frac{\rho_0 - 1}{2R_0 + \rho_0 - 1} \leq 1 \). Then for \( \theta \in A_n \) and large enough \( n \),
\[ E_{\theta,x}(1 - \phi_j) \leq \exp \left\{ -\frac{n \epsilon_n^2}{32l} \right\}. \tag{46} \]

Combining (40), with (41), (45), (42), (46) and using assumptions C1 and C3 we finally obtain for large enough \( n \)
\[ \mathbb{P}_{\theta_0} \left( \mathbb{P}^n \left[ \| f_{i,j} - f_{i,0} \|_1 \frac{\rho_0 - 1}{2R_0 + \rho_0 - 1} \geq \epsilon_n \left| Y_{1:n} \right. \right] \right) \]
\[ \leq o(1) + O \left( e^{n \epsilon_n^2 (1 + C_n)} \right) \exp \left\{ -\frac{n \epsilon_n^2}{32l} \right\} \]
\begin{align*}
&+ \exp \left\{ - \frac{n \epsilon_n^2 (\rho \theta_0 - 1)^2}{8l(2R \theta_0 + \rho \theta_0 - 1)^2} \right\} \exp \left\{ - \frac{n \epsilon_n^2 (\rho \theta_0 - 1)^2}{16l(2R \theta_0 + \rho \theta_0 - 1)^2} \right\} \\
&= o(1).
\end{align*}

Assume now that assumption **C3bis** holds. By writing \( A_n \cap \mathcal{F}_n = \bigcup_{m \geq 1} A_{n,m}(\epsilon_n) \) and using same reasoning, one gets, for some positive constant \( c \):

\begin{align*}
&\mathbb{P}_{\theta_0} \left( \mathbb{P}^\mathcal{F}_n \left[ \| f_{i,\theta_j} - f_{i,\theta_0} \|_{1, \rho_0 - 1 \over 2R \theta_0 + \rho \theta_0 - 1} \geq \epsilon_n \mid Y_{1:n} \right] \right) \\
&= o(1) + e^{n \epsilon_n^2} \sum_{m \geq 1} \frac{\prod_{i} (A_{n,m}(\epsilon_n))}{\prod_i (S_n)} \exp \left\{ - \frac{n m^2 \epsilon_n^2}{32l} \right\} \\
&+ \sum_{m \geq 1} N \left( m \epsilon_n, A_{n,m}(\epsilon_n), d_1(\cdot, \cdot) \right) \exp \left\{ - \frac{n m^2 \epsilon_n^2 (\rho \theta_0 - 1)^2}{8l(2R \theta_0 + \rho \theta_0 - 1)^2} \right\} \\
&= o(1)
\end{align*}

and the second part of Theorem 4 is proved.

**Acknowledgement**

The authors would like to thank the anonymous referees for providing constructive comments that were very helpful to improve the readability of the paper. This work was partly supported by the 2010–2014 grant ANR Banhdits AAP Blanc SIMI 1.

**References**


Posterior distribution for HMM


Received March 2013 and revised July 2013