Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models

Luc Lehéricy

Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models

Luc Lehéricy luc.lehericy@math.u-psud.fr
Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France

Abstract

We study the problem of estimating an unknown time process distribution using nonparametric hidden Markov models in the *misspecified setting*, that is when the true distribution of the process may not come from a hidden Markov model. We show that when the true distribution is exponentially mixing and satisfies a forgetting assumption, the maximum likelihood estimator recovers the best approximation of the true distribution. We prove a finite sample bound on the resulting error and show that it is optimal in the minimax sense–up to logarithmic factors–when the model is well specified.

Keywords: misspecified model, nonasymptotic bound, nonparametric statistics, maximum likelihood estimator, model selection, oracle inequality, hidden Markov model

1. Introduction

Let $(Y_1, \dots, Y_n)$ be a sample following some unknown distribution $\mathbb{P}^*$. The maximum likelihood estimator can be formalized as follows: let $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$, the *model*, be a family of possible distributions; pick a distribution $\mathbb{P}_{\hat\theta}$ of the model which maximizes the likelihood of the observed sample.

In many situations, the true distribution may not belong to the model at hand: this is the so-called *misspecified setting*. One would like the estimator to give sensible results even in this setting. This can be done by showing that the estimated distribution converges to the best approximation of the true distribution within the model. The goal of this paper is to establish a finite sample bound on the error of the maximum likelihood estimator for a large class of true distributions and a large class of nonparametric hidden Markov models.

In this paper, we consider maximum likelihood estimators (shortened MLE) based on model selection among finite state space hidden Markov models (shortened HMM). A finite state space hidden Markov model is a stochastic process $(X_t, Y_t)_t$ where only the observations $(Y_t)_t$ are observed, such that the process $(X_t)_t$ is a Markov chain taking values in a finite space and such that the Y_s are independent conditionally to $(X_t)_t$ with a distribution depending only on the corresponding X_s. The parameters of a HMM $(X_t, Y_t)_t$ are the initial distribution and the transition matrix of $(X_t)_t$ and the distributions of Y_s conditionally to X_s.

HMMs have been widely used in practice, for instance in climatology (Lambert et al., 2003), ecology (Boyd et al., 2014), voice activity detection and speech recognition (Couvreur and Couvreur, 2000; Lefèvre, 2003), biology (Yau et al., 2011; Volant et al., 2014)... One of their advantages is their ability to account for complex dependencies between the observations: despite

the seemingly simple structure of these models, the fact that the process $(X_t)_t$ is hidden makes the process $(Y_t)_t$ non-Markovian.

Up to now, most theoretical work in the literature focused on well-specified and parametric HMMs, where a smooth parametrization by a subset of $\mathbb{R}^d$ is available, see for instance Baum and Petrie (1966) for discrete state and observations spaces, Leroux (1992) for general observation spaces and Douc and Matias (2001) and Douc et al. (2011) for general state and observation spaces. Asymptotic properties for misspecified models have been studied recently by Mevel and Finesso (2004) for consistency and asymptotic normality in finite state space HMMs and Douc and Moulines (2012) for consistency in HMMs with general state space. Let us also mention Pouzo et al. (2016), who studied a generalization of hidden Markov models in a semi-misspecified setting. All these results focus on parametric models.

Few results are available on nonparametric HMMs, and all of them focus on the well-specified setting. Alexandrovich et al. (2016) prove consistency of a nonparametric maximum likelihood estimator based on finite state space hidden Markov models with nonparametric mixtures of parametric densities. Vernet (2015a,b) study the posterior consistency and concentration rates of a Bayesian nonparametric maximum likelihood estimator. Other methods have also been considered, such as spectral estimators in Anandkumar et al. (2012); Hsu et al. (2012); De Castro et al. (2017); Bonhomme et al. (2016); Lehéricy (2017) and least squares estimators in de Castro et al. (2016); Lehéricy (2017). Besides Vernet (2015b), to the best of our knowledge, there has been no result on convergence rates or finite sample error of the nonparametric maximum likelihood estimator, even in the well-specified setting.

The main result of this paper is an oracle inequality that holds as soon as the models have controlled tails. This bound is optimal when the true distribution is a HMM taking values in $\mathbb{R}$. Let us give some details about this result.

Let us start with an overview of the assumptions on the true distribution $\mathbb{P}^*$. The first assumption is that the observed process is strongly mixing. Strong mixing assumptions can be seen as a strengthened version of ergodicity. They have been widely used to extend results on independent observation to dependent processes, see for instance Bradley (2005) and Dedecker et al. (2007) for a survey on strong mixing and weak dependence conditions. The second assumption is that the process forgets its past exponentially fast. For hidden Markov models, this forgetting property is closely related to the exponential stability of the optimal filter, see for instance Le Gland and Mevel (2000); Gerencsér et al. (2007); Douc et al. (2004, 2009). The last assumption is that the likelihood of the true process has sub-polynomial tails. None of these assumptions are specific to HMMs, thus making our result applicable to the misspecified setting.

To approximate a large class of true distributions, we consider nonparametric HMMs, where the parameters are not described by a finite dimensional space. For instance, one may consider HMMs with arbitrary number of states and arbitrary emission distributions. Computing a maximizer of the likelihood directly in a nonparametric model may be hard or result in overfitting. The model selection approach offers a way to circumvent this problem. It consists in considering a countable family of parametric sets $(S_M)_{M\in\mathcal{M}}$–the *models*–and selecting one of them. The larger the union of all models, the more distributions are approximated. Several criteria can be used to select the model, such as bootstrap, cross validation (see for instance Arlot and Celisse (2010)) or penalization (see for instance

Massart (2007)). We use a penalized criterion, which consists in maximizing the function

$$(S, \theta \in S) \longmapsto \frac{1}{n} \log p_\theta(Y_1, \dots, Y_n) - \mathrm{pen}_n(S),$$

where p_θ is the density of $(Y_1, \dots, Y_n)$ under the parameter θ and the penalty pen only depends on the model S and the number of observations n.

Assume that the emission distributions of the HMMs–that is the distribution of the observations conditionally to the hidden states–are absolutely continuous with respect to some known probability measure, and call *emission densities* their densities with respect to this measure. The tail assumption ensures that the emission densities have sub-polynomial tail:

$$\forall v \geqslant e, \qquad \mathbb{P}^* \left(\sup_\gamma \gamma(Y_1) \geqslant v^{D(n)} \right) \leqslant \frac{1}{v},$$

where the supremum is taken over all emission densities γ in the models for a function $n \longmapsto D(n)$. For instance, this assumption holds when all densities are upper bounded by $e^{D(n)}$. A key remark at this point is the dependency of $D(n)$ with n: we allow the models to depend on the sample size. Typically, taking a larger sample makes it possible to consider larger models. A good choice is to take $D(n)$ proportional to $\log n$.

To stabilize the log-likelihood, we modify the models in the following way. First, only keep HMMs whose transition matrix is lower bounded by a positive function $n \longmapsto \sigma_-(n)$. We show that taking this lower bound as $(\log n)^{-1}$ is a safe choice. Then, replace the emission densities γ by a convex combination of the original emission densities and of the dominating measure λ with a weight that decreases polynomially with the sample size. In other words, replace γ by $(1 - n^{-a})\gamma + n^{-a}\lambda$ for some $a > 0$. Taking $a > 1$ ensures that the component λ is asymptotically negligible. Any $a > 0$ works, but the constants of the oracle inequality depend on it.

A simplified version of our main result (Theorem 8) is the following oracle inequality: for all $\alpha \geqslant 1$, there exists constants A and n_0 such that if the penalty is large enough, the penalized maximum likelihood estimator $\hat{\theta}_n$ satisfies for all $t \geqslant 1$, $\eta \in (0, 1)$ and $n \geqslant n_0$, with probability larger than $1 - e^{-t} - n^{-\alpha}$:

$$\mathbf{K}(\hat{\theta}_n) \leqslant (1 + \eta) \inf_{\dim(S) \leqslant n} \left\{ \inf_{\theta \in S} \mathbf{K}(\theta) + \mathrm{pen}_n(S) \right\} + \frac{A}{\eta} t \frac{(\log n)^8}{n},$$

where $\mathbf{K}(\theta)$ can be seen as a Kullback-Leibler divergence between the distributions $\mathbb{P}^*$ and $\mathbb{P}_\theta$. In other words, the estimator recovers the best approximation of the true distribution within the model, up to the penalty and the residual term.

In the case where the true distribution is a HMM, it is possible to quantify the approximation error $\inf_{\theta \in S} \mathbf{K}(\theta)$. Using the results of Kruijer et al. (2010), we show that the above oracle inequality is optimal in the minimax sense–up to logarithmic factors–for real-valued HMMs, see Corollary 12. This is done by taking HMMs whose emission densities are mixtures of exponential power distributions–which include Gaussian mixtures as a special case.

The paper is organized as follows. We detail the framework of the article in Section 2. In particular, Section 2.3 describes the assumptions on the true distribution, Section 2.4 presents the assumptions on the model and Section 2.5 introduces the Kullback Leibler

criterion used in the oracle inequality. Our main results are stated in Section 3. Section 3.1 contains the oracle inequality and Section 3.2 shows how it can be used to show minimax adaptivity for real-valued HMMs. Section 4 lists some perspectives for this work.

One may wish to relax our assumptions depending on the setting. For instance, one could want to change the dependency of the functions $B(n)$ and $\sigma_-(n)$ on n, change the tail conditions or the rate of forgetting. We give an overview of the key steps of the proof of our oracle inequality in Section 5 to make it easier to adapt our result.

Some proofs are postponed the Appendices. Appendix A contains the proof of the minimax adaptivity result and Appendix B contains the proof of the main technical lemma of Section 5.

2. Notations and assumptions

We will use the following notations:

- $a \vee b$ is the maximum of a and b, $a \wedge b$ the minimum;
- For $x \in \mathbb{R}$, we write $x^+ = x \vee 0$;
- $\mathbb{N}^* = \{1, 2, 3, \dots\}$ is the set of positive integers;
- For $K \in \mathbb{N}^*$, we write $[K] = \{1, 2, \dots, K\}$;
- Y_a^b is the vector $(Y_a, \dots, Y_b)$;
- $\mathbf{L}^2(A, \mathcal{A}, \mu)$ is the set of measurable and square integrable functions defined on the measured space $(A, \mathcal{A}, \mu)$. We write $\mathbf{L}^2(A, \mu)$ when the sigma-field is not ambiguous;
- log is the inverse function of the exponential function exp.

2.1 Hidden Markov models

Finite state space hidden Markov models (HMM in short) are stochastic processes $(X_t, Y_t)_{t \geqslant 1}$ with the following properties. The *hidden state* process $(X_t)_t$ is a Markov chain taking value in a finite set $\mathcal{X}$ (the *state space*). We denote by K the cardinality of $\mathcal{X}$, and π and $\mathbf{Q}$ the initial distribution and transition matrix of $(X_t)_t$ respectively. The *observation* process $(Y_t)_t$ takes value in a polish space $\mathcal{Y}$ (the *observation space*) endowed with a Borel probability measure λ. The observations Y_t are independent conditionally to $(X_t)_t$ with a distribution depending only on X_t. In the following, we assume that the distribution of Y_t conditionally to $\{X_t = x\}$ is absolutely continuous with respect to λ with density γ_x. We call $\gamma = (\gamma_x)_{x \in \mathcal{X}}$ the *emission densities*.

Therefore, the parameters of a HMM are its number of hidden states K, its initial distribution π (the distribution of X_1), its transition matrix $\mathbf{Q}$ and its emission densities γ. When appropriate, we write $p_{(K,\pi,\mathbf{Q},\gamma)}$ the density of the process with respect to the dominating measure under the parameters $(K, \pi, \mathbf{Q}, \gamma)$. For a sequence of observations Y_1^n, we denote by $l_n(K, \pi, \mathbf{Q}, \gamma)$ the associated log-likelihood under the parameters $(K, \pi, \mathbf{Q}, \gamma)$, defined by

$$l_n(K, \pi, \mathbf{Q}, \gamma) = \log p_{(K,\pi,\mathbf{Q},\gamma)}(Y_1^n).$$

We denote by $\mathbb{P}^*$ the true (and unknown) distribution of the process $(Y_t)_t$, $\mathbb{E}^*$ the expectation under $\mathbb{P}^*$, p^* the density of $\mathbb{P}^*$ under the dominating measure and l_n^* the log-likelihood of the observations under $\mathbb{P}^*$. Let us stress that this distribution may not be generated by a finite state space HMM.

2.2 The model selection estimator

Let $(S_{K,M,n})_{K\in\mathbb{N}^*,M\in\mathcal{M}}$ be a family of parametric models such that for all $K \in \mathbb{N}^*$ and $M \in \mathcal{M}$, the parameters $(K, \pi, \mathbf{Q}, \gamma) \in S_{K,M,n}$ correspond to HMMs with K hidden states. Note that the models $S_{K,M,n}$ may depend on the number of observations n. Let us see two ways to construct such models.

Mixture densities. Let $\{f_\xi\}_{\xi\in\Xi}$ be a parametric family of probability densities indexed by $\Xi \subset \mathbb{R}^d$. Let $\mathcal{M} \subset \mathbb{N}^*$. We choose $S_{K,M,n}$ to be the set of parameters $(K, \pi, \mathbf{Q}, \gamma)$ such that $\mathbf{Q}$ and π are uniformly lower bounded by $(\log n)^{-1}$ and for all $x \in [K]$, γ_x is a convex combination of M elements of $\{f_\xi\}_{\xi\in\Xi\cap[-n,n]^d}$.

$\mathbf{L}^2$ densities. Let $(E_M)_{M\in\mathcal{M}}$ be a family of finite dimensional subspaces of $\mathbf{L}^2(\mathcal{Y}, \lambda)$. We choose $S_{K,M,n}$ to be the set of parameters $(K, \pi, \mathbf{Q}, \gamma)$ such that $\mathbf{Q}$ and π are uniformly lower bounded by $(\log n)^{-1}$ and for all $x \in [K]$, γ_x is a probability density such that $\gamma_x = g \vee 0$ for a function $g \in E_M$ such that $\|g\|_2 \leqslant n$.

In both cases, we took a lower bound on the coefficients of the transition matrix $\mathbf{Q}$ that tends to zero when the number of observations grows. This allows to estimate parameters for which some coefficients of the transition matrix are small or zero. We prove the choice $(\log n)^{-1}$ to be a good choice in general in Theorem 8.

For all $K \in \mathbb{N}^*$ and $M \in \mathcal{M}$, we define the maximum likelihood estimator on $S_{K,M,n}$:

$$(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n}) \in \underset{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M,n}}{\arg\max} \frac{1}{n} l_n(K, \pi, \mathbf{Q}, \gamma).$$

Since the true distribution does not necessarily correspond to a parameter of $S_{K,M,n}$, taking a larger model $S_{K,M,n}$ will reduce the bias of the estimator $(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n})$. However, larger models will make the estimation more difficult, resulting in a larger variance. This means one has to perform a bias-variance tradeoff to select a model with a reasonable size. To do so, we select a number of states $\hat{K}_n$ among a set of integers $\mathcal{K}_n$ and a model index $\hat{M}_n$ among a set of indices $\mathcal{M}_n$ such that the penalized log-likelihood is maximal:

$$(\hat{K}_n, \hat{M}_n) \in \underset{K\in\mathcal{K}_n, M\in\mathcal{M}_n}{\arg\max} \left(\frac{1}{n} l_n(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n}) - \mathrm{pen}_n(K, M) \right)$$

for some penalty pen_n to be chosen.

In the following, we use the following notations.

- $\mathbf{S}_n := \bigcup_{K\in\mathcal{K}_n, M\in\mathcal{M}_n} S_{K,M,n}$ is the set of all parameters involved with the construction of the maximum likelihood estimator;
- $S^{(\gamma)}_{K,M,n} = \{\gamma \,|\, (K, \pi, \mathbf{Q}, \gamma) \in S_{K,M,n}\}$ is the set of density vectors from the model $S_{K,M,n}$. $\mathbf{S}^{(\gamma)}_n$ is defined in the same way.

2.3 Assumptions on the true distribution

In this section, we introduce the assumptions on the true distribution of the process $(Y_t)_{t\geqslant 1}$. We assume that $(Y_t)_{t\geqslant 1}$ is stationary, so that one can extend it into a process $(Y_t)_{t\in\mathbb{Z}}$.

2.3.1 Forgetting and mixing

Let us state the two assumptions on the dependency of the process $(Y_t)_t$.

[A⋆forgetting] There exists two constants $C_* > 0$ and $\rho_* \in (0,1)$ such that for all $i \in \mathbb{Z}$, for all $k, k' \in \mathbb{N}^*$ and for all $y_{i-(k\vee k')}^i \in \mathcal{Y}^{(k\vee k')+1}$,

$$|\log p^*(y_i|y_{i-k}^{i-1}) - \log p^*(y_i|y_{i-k'}^{i-1})| \leqslant C_* \rho_*^{k\wedge k'-1}$$

For the mixing assumption, let us recall the definition of the ρ-mixing coefficient. Let $(\Omega, \mathcal{F}, P)$ be a measured space and $\mathcal{A} \subset \mathcal{F}$ and $\mathcal{B} \subset \mathcal{F}$ be two sigma-fields. Let

$$\rho_{\text{mix}}(\mathcal{A}, \mathcal{B}) = \sup_{\substack{f \in \mathbf{L}^2(\Omega, \mathcal{A}, P) \\ g \in \mathbf{L}^2(\Omega, \mathcal{B}, P)}} |\text{Corr}(f, g)|.$$

The ρ-mixing coefficient of $(Y_t)_t$ is defined by

$$\rho_{\text{mix}}(n) = \rho_{\text{mix}}(\sigma(Y_i, i \geqslant n), \sigma(Y_i, i \leqslant 0)).$$

[A⋆mixing] There exists two constants $c_* > 0$ and $n_* \in \mathbb{N}^*$ such that

$$\forall n \geqslant n_*, \quad \rho_{\text{mix}}(n) \leqslant 4e^{-c_* n}.$$

[A⋆forgetting] ensures that the process forgets its initial distribution exponentially fast. This assumption is especially useful for truncating the dependencies in the likelihood. **[A⋆mixing]** is a usual mixing assumption and is used to obtain Bernstein-like concentration inequalities. Note that **[A⋆mixing]** implies that the process $(Y_t)_{t\geqslant 1}$ is ergodic.

Even if **[A⋆forgetting]** is analog to a ψ-mixing condition (see Bradley (2005) for a survey on mixing conditions) and is proved using the same tool as **[A⋆mixing]** in hidden Markov models–namely the geometric ergodicity of the hidden state process–these two assumptions are different in general. For instance, a Markov chain always satisfies **[A⋆forgetting]** but not necessarily **[A⋆mixing]**. Conversely, there exists processes satisfying **[A⋆mixing]** but not **[A⋆forgetting]**.

Lemma 1 *Assume that $(Y_t)_t$ is generated by a HMM with a compact metric state space $\mathcal{X}$ (not necessarily finite) endowed with a Borel probability measure μ. Write $\mathcal{Q}^*$ its transition kernel and assume that $\mathcal{Q}^*$ admits a density with respect to μ that is uniformly lower bounded and upper bounded by positive and finite constants σ_-^* and σ_+^*. Write $(\gamma_x^*)_{x\in\mathcal{X}}$ its emission densities and assume that they satisfy $\int \gamma_x^*(y)\mu(dx) \in (0, +\infty)$ for all $y \in \mathcal{Y}$.*

*Then **[A⋆forgetting]** and **[A⋆mixing]** hold by taking $\rho_* = 1 - \frac{\sigma_-^*}{\sigma_+^*}$, $C_* = \frac{1}{1-\rho_*}$, $c_* = \frac{-\log(1-\sigma_-^*)}{2}$ and $n_* = 1$.*

Proof This lemma follows from the geometric ergodicity of the HMM.

For **[A⋆forgetting]**, see for instance Douc et al. (2004), proof of Lemma 2.

For **[A⋆mixing]**, the Doeblin condition implies that for all distribution π and π' on $\mathcal{X}$,

$$\int |p^*(X_n = x|X_0 \sim \pi) - p^*(X_n = x|X_0 \sim \pi')|\mu(dx) \leqslant (1 - \sigma_-^*)^n \|\pi - \pi'\|_1.$$

Let $A \in \sigma(Y_t, t \geqslant k)$ and $B \in \sigma(Y_t, t \leqslant 0)$ such that $\mathbb{P}^*(B) > 0$. Taking π the stationary distribution of $(X_t)_t$ and π' the distribution of X_0 conditionally to B in the above equation implies

$$\begin{aligned}|\mathbb{P}^*(A|B) - \mathbb{P}^*(A)| &= \left| \int \mathbb{P}^*(A|X_n = x)(p^*(X_n = x) - p^*(X_n = x|B))\mu(dx) \right| \\ &\leqslant \int |p^*(X_n = x) - p^*(X_n = x|B)|\mu(dx) \\ &\leqslant 2(1 - \sigma_-^*)^n.\end{aligned}$$

Therefore, the process $(Y_t)_{t\geqslant 1}$ is ϕ-mixing with $\phi_{\text{mix}}(n) \leqslant 2(1-\sigma_-^*)^n$, so that it is ρ-mixing with $\rho_{\text{mix}}(n) \leqslant 2(\phi_{\text{mix}}(n))^{1/2} \leqslant 2\sqrt{2}(1-\sigma_-^*)^{n/2}$ (see e.g. Bradley (2005) for the definition of the ϕ-mixing coefficient and its relation to the ρ-mixing coefficient). One can check that the choice of c_* and n_* allows to obtain **[A⋆mixing]** from this inequality. ■

2.3.2 Extreme values of the true density

We need to control the probability that the true density takes extreme values.

[A⋆tail] There exists two constants $B^* \geqslant 1$ and $q \in [0, 1]$ such that

$$\forall i \in \mathbb{Z}, \quad \forall k \in \mathbb{N}, \quad \forall u \geqslant 1, \quad \mathbb{P}^*(|\log p^*(Y_i|Y_{i-k}^{i-1})| \geqslant B^* u^q) \leqslant e^{-u}.$$

In practice, only two values of q are of interest. The case $q = 0$ occurs when the densities are lower and upper bounded by positive and finite constants. If the densities are not bounded, then $q = 1$ works in most cases and corresponds to subpolynomial tails. Indeed, the lower bound on $\log p^*(Y_i|Y_{i-k}^{i-1})$ is always true when taking $q = 1$ and $B^* = 1$ by definition of the density p^*, resulting in the following equivalent assumption:

[A⋆tail'] There exists a constant $B^* \geqslant 1$ such that

$$\forall i \in \mathbb{Z}, \quad \forall k \in \mathbb{N}, \quad \forall v \geqslant e, \quad \mathbb{P}^*(p^*(Y_i|Y_{i-k}^{i-1}) \geqslant v^{B^*}) \leqslant \frac{1}{v}.$$

This can be obtained from Markov's inequality under a moment assumption, as shown in the following lemma.

Lemma 2 *Assume that there exists $\delta > 0$ such that*

$$M_\delta := \sup_{i,k} \mathbb{E}^*[(p^*(Y_i|Y_{i-k}^{i-1}))^\delta] < \infty.$$

*Then **[A⋆tail]** holds for $q = 1$ and $B^* = \frac{1+\log M_\delta}{\delta}$.*

2.4 Model assumptions

We now state the assumptions on the models. Let us recall that the distribution of the observed process is not assumed to belong to one of these models.

Consider a family of models $(S_{K,M,n})_{K\in\mathbb{N}^*,M\in\mathcal{M},n\in\mathbb{N}^*}$ such that for each K, M and n, the elements of $S_{K,M,n}$ are of the form $(K,\pi,\mathbf{Q},\gamma)$ where π is a probability density on $[K]$, $\mathbf{Q}$ is a transition matrix on $[K]$ and γ is a vector of K probability densities on $\mathcal{Y}$ with respect to λ.

2.4.1 Transition kernel

We need the following assumption on the transition matrices and initial distributions of $\mathbf{S}_n$.

[Aergodic] There exists $\sigma_-(n)\in(0,e^{-1}]$ such that for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$,

$$\inf_{x,x'\in[K]}\mathbf{Q}(x,x')\geqslant\sigma_-(n)\quad\text{and}\quad\inf_{x\in[K]}\pi(x)\geqslant\sigma_-(n).$$

[Aergodic] is standard in maximum likelihood estimation. It ensures that the process forgets the past exponentially fast, which implies that the difference between the log-likelihood $\frac{1}{n}l_n$ and its limit converges to zero with rate $1/n$ in supremum norm.

2.4.2 Tail of the emission densities

When $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$, **[Aergodic]** implies that under the parameters $(K,\pi,\mathbf{Q},\gamma)$, for all $x\in[K]$, the probability to jump to state x at time t is at least $\sigma_-(n)$, whatever the past may be. This implies that the density $p_{(K,\pi,\mathbf{Q},\gamma)}(Y_t|Y_1^{t-1})$ is lower bounded by $\sigma_-(n)\sum_x\gamma_x(Y_t)$. Furthermore, it is upper bounded by $\sum_x\gamma_x(Y_t)$. Thus, it is enough to bound this quantity to control $p_{(K,\pi,\mathbf{Q},\gamma)}$ without having to handle the time dependency.

For all $\gamma\in\mathbf{S}_n^{(\gamma)}$ and $y\in\mathcal{Y}$, let

$$b_\gamma(y)=\log\sum_x\gamma_x(y).$$

We need to control the tails of b_γ like we did for $\log p^*$ in order to get nonasymptotic bounds. This is the purpose of the following assumption.

[Atail] There exists two constants $q\in[0,1]$ and $B(n)\geqslant 1$ such that

$$\forall u\geqslant 1,\quad\mathbb{P}^*\left[\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|\geqslant B(n)u^q\right]\leqslant e^{-u}.$$

This assumption is often easy to check in practice, as shown in the following lemma.

Lemma 3 *Assume that one of the two following assumption holds:*

1. (subpolynomial tails) *there exists $D(n)\geqslant 1$ such that*

$$\forall u\geqslant 1,\quad\mathbb{P}^*\left[\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}b_\gamma(Y_1)\geqslant D(n)u\right]\leqslant e^{-u}.$$

2. (bounded densities) *there exists* $D(n) \geqslant 1$ *such that*

$$\sup_{y\in\mathcal{Y}} \sup_{\gamma\in\mathbf{S}_n^{(\gamma)}} b_\gamma(y) \leqslant D(n).$$

Consider a new model where all γ *are replaced by* $\gamma' = (1-n^{-a})\gamma + n^{-a}$ *for a fixed constant* $a > 0$*. Then* ***[Atail]*** *holds for this new model with* $q = 1$ *(resp.* $q = 0$ *with the second assumption) and* $B(n) = D(n) \vee (a \log n)$*.*

Changing the densities as in the lemma amounts to adding a mixture component (with weight n^{-a} and distribution λ) to the emission densities to make sure that they are uniformly lower bounded. We shall see in the following that if $a \geqslant 1$, then this additional component changes nothing to the approximation properties of the models, see the proof of Corollary 12. This is in agreement with the fact that this component is asymptotically never observed as soon as $a > 1$.

2.4.3 Complexity of the approximation spaces

The following assumption means that as far as the bracketing entropy is concerned, the set of emission densities of the model $S_{K,M,n}$ (without taking the hidden state into account) behaves like a parametric model with dimension m_M.

[Aentropy] There exists a function $(M,K,D,n) \longmapsto C_{\text{aux}} \geqslant 1$ and a sequence $(m_M)_{M\in\mathcal{M}} \in \mathbb{N}^{\mathcal{M}}$ such that for all $\delta > 0$, M, K and D,

$$N\left(\left\{ y \mapsto \gamma_x(y)\mathbf{1}_{\sup_{\gamma'\in\mathbf{s}_n^{(\gamma)}} |b_{\gamma'}(y)|\leqslant D} \right\}_{\gamma\in S_{K,M,n}^{(\gamma)}, x\in[K]}, d_\infty, \delta\right) \leqslant \max\left(\frac{C_{\text{aux}}}{\delta}, 1\right)^{m_M}, \quad (1)$$

where d_∞ is the distance associated with the supremum norm and $N(A,d,\epsilon)$ is the smallest number of brackets of size ϵ for the distance d needed to cover A. Let us recall that the bracket $[a;b]$ is the set of functions f such that $a(\cdot) \leqslant f(\cdot) \leqslant b(\cdot)$, and that the size of the bracket $[a;b]$ is $d(a,b)$.

Note that we allow the models to depend on the sample size n, which can make C_{aux} grow to infinity with n. To control the growth of the models, we use the following assumption.

[Agrowth] There exists $\zeta > 0$ and n_{growth} such that for all $n \geqslant n_{\text{growth}}$,

$$\sup_{K,M \text{ s.t. } K\leqslant n \text{ and } m_M\leqslant n} \log C_{\text{aux}}(M,K,B(n)(\log n)^q, n) \leqslant n^\zeta.$$

A typical way to check **[Aentropy]** is to use a parametrization of the emission densities, for instance a lipschitz application $[-1,1]^{m_M} \longrightarrow S_{K,M,n}^{(\gamma)}$. This reduces the construction of a bracket covering on $S_{K,M,n}^{(\gamma)}$ to the construction of a bracket covering of the unit ball of $\mathbb{R}^{m_M}$. In this case, C_{aux} depends on the lipschitz constant of the parametrization. An example of this approach is given in Section 3.2 for mixtures of exponential power distributions.

2.5 Limit and properties of the log-likelihood

In this section, we focus on the convergence of the log-likelihood. First, we recall results from Barron (1985) and Leroux (1992) that show the existence of its limit in a general setting. Then, we show how to control the difference between the log-likelihood and its limit using the assumptions from the previous Sections.

2.5.1 Convergence of the log-likelihood

The first result comes from Barron (1985) and shows that the true log-likelihood converges almost surely with no assumption other than the ergodicity of the process $(Y_t)_{t\geqslant 1}$.

Lemma 4 (Barron (1985)) *Assume that the process $(Y_t)_{t\geqslant 1}$ is ergodic, then there exists a quantity $l^* > -\infty$ such that*

$$\frac{1}{n} l_n^* \underset{n\to\infty}{\longrightarrow} l^* \quad a.s.$$

and

$$l^* = \lim_{n\to\infty} \mathbb{E}^*[\log p^*(Y_n | Y_1^{n-1})].$$

The second result follows from Theorem 2 of Leroux (1992). A careful reading of his proof shows that one can relax his assumptions to get the following lemma. Note that the definition of l_n extends naturally to the case where γ is not a vector of probability densities, or even a vector of integrable functions with respect to λ, through the formula

$$l_n(K, \pi, \mathbf{Q}, \gamma) = \log \sum_{x_1^n \in [K]^n} \pi(x_1) \prod_{i=1}^{n-1} Q(x_i, x_{i+1}) \prod_{i=1}^{n} \gamma_{x_i}(Y_i).$$

Lemma 5 (Leroux (1992)) *Let K be a positive integer, γ a vector of K nonnegative and measurable functions, $\mathbf{Q}$ a transition matrix of size K and π a probability measure on $[K]$.*

Assume that the process $(Y_t)_{t\geqslant 1}$ is ergodic and that $\mathbb{E}^*[(\log \gamma_x(Y_1))^+] < +\infty$ for all $x \in [K]$. Then:

1. *There exists a quantity $l(K, \mathbf{Q}, \gamma) < +\infty$ which does not depend on π such that*

$$\limsup_{n\to\infty} \frac{1}{n} l_n(K, \pi, \mathbf{Q}, \gamma) \leqslant l(K, \mathbf{Q}, \gamma) \quad \mathbb{P}^*\text{-a.s.}$$

   and such that if $\inf_{x\in[K]} \pi(x) > 0$, then

$$\frac{1}{n} l_n(K, \pi, \mathbf{Q}, \gamma) \underset{n\to\infty}{\longrightarrow} l(K, \mathbf{Q}, \gamma) \quad \mathbb{P}^*\text{-a.s.}$$

2. *Assume $l(K, \mathbf{Q}, \gamma) > -\infty$. Then the almost sure convergence also holds in $\mathbf{L}^1(\mathbb{P}^*)$.*

3. *Assume $\mathbb{E}^*|\log \gamma_x(Y_1)| < +\infty$ for all $x \in [K]$. Then $l(K, \mathbf{Q}, \gamma) > -\infty$.*

When appropriate, we define $\mathbf{K}(K, \mathbf{Q}, \gamma)$ by

$$\mathbf{K}(K, \mathbf{Q}, \gamma) := l^* - l(K, \mathbf{Q}, \gamma).$$

Note that when γ is a vector of probability densities, $\mathbf{K}(K, \mathbf{Q}, \gamma) \geqslant 0$ since it is the limit of a sequence of Kullback-Leibler divergences: under the assumptions of Lemma 5, if $\inf_{x \in [K]} \pi(x) > 0$,

$$\mathbf{K}(K, \mathbf{Q}, \gamma) = \lim_{n \to \infty} \frac{1}{n} KL(\mathbb{P}^*_{Y_1^n} \| \mathbb{P}_{Y_1^n | (K, \pi, \mathbf{Q}, \gamma)}).$$

2.5.2 Approximation of the limit

The following lemma controls the difference between the log-likelihood and its limit. When **[A⋆forgetting]** (resp. **[Aergodic]**) holds, the log-density of Y_1 conditionally to the previous observations converges exponentially fast to what can be seen as the density of Y_1 conditionally to the whole past, that is $p^*(Y_i | Y_{-\infty}^{i-1})$ (resp. $p_{(K,\mathbf{Q},\gamma)}(Y_i | Y_{-\infty}^{i-1})$). Strictly speaking, we define the limit of the log-density $L^*_{i,\infty}$ and $L_{i,\infty}(K, \mathbf{Q}, \gamma)$, which can be seen respectively as $\log p^*(Y_i | Y_{-\infty}^{i-1})$ and $\log p_{(K,\mathbf{Q},\gamma)}(Y_i | Y_{-\infty}^{i-1})$.

For all $i \in \mathbb{Z}$, $k \in \mathbb{N}^*$, let

$$L^*_{i,k} = \log p^*(Y_i | Y_{i-k}^{i-1}),$$

where the process $(Y_t)_{t \geqslant 1}$ is extended into a process $(Y_t)_{t \in \mathbb{Z}}$ by stationarity. Likewise, for all $i \in \mathbb{Z}$, $k \in \mathbb{N}^*$, $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$ and for all probability distribution μ on $[K]$, let

$$L_{i,k,\mu}(K, \mathbf{Q}, \gamma) = \log p_{(K,\mathbf{Q},\gamma)}(Y_i | Y_{i-k}^{i-1}, X_{i-k} \sim \mu),$$

where $p_{(K,\mathbf{Q},\gamma)}$ is the density of a stationary HMM with parameters $(K, \mathbf{Q}, \gamma)$. When μ is the stationary distribution of the Markov chain under the parameter $(K, \mathbf{Q}, \gamma)$, we write $L_{i,k}(K, \mathbf{Q}, \gamma)$.

Lemma 6

1. ***(Douc et al. (2004)).*** *Assume **[Aergodic]** holds. Let $\rho = 1 - \frac{\sigma_-(n)}{1 - \sigma_-(n)}$. Then for all i, k, k', μ and μ',*

$$\sup_{(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n} |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L_{i,k',\mu'}(K, \mathbf{Q}, \gamma)| \leqslant \rho^{k \wedge k' - 1} / (1 - \rho)$$

and there exists a process $(L_{i,\infty})_{i \in \mathbb{Z}}$ such that for all i and μ, $L_{i,k,\mu} \underset{k \to \infty}{\longrightarrow} L_{i,\infty}$ in supremum norm (when seen as a function of $(K, \pi, \mathbf{Q}, \gamma)$) and for all i, k and μ,

$$\sup_{(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n} |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L_{i,\infty}(K, \mathbf{Q}, \gamma)| \leqslant \rho^{k-1} / (1 - \rho).$$

2. *Assume **[A⋆forgetting]** holds, then for all i, k and k', $|L^*_{i,k} - L^*_{i,k'}| \leqslant C_* \rho_*^{k \wedge k' - 1}$ and there exists a process $(L^*_{i,\infty})_{i \in \mathbb{Z}}$ such that for all i, $L^*_{i,k} \underset{k \to \infty}{\longrightarrow} L^*_{i,\infty}$ and for all i and k,*

$$|L^*_{i,k} - L^*_{i,\infty}| \leqslant C_* \rho_*^{k-1}.$$

3. *Assume **[A⋆forgetting]** and **[Aergodic]** hold. Under $\mathbb{P}^*$, the processes $(L^*_{i,\infty})_{i\in\mathbb{Z}}$ and $(L_{i,\infty}(K,\mathbf{Q},\gamma))_{i\in\mathbb{Z}}$ are stationary for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$. Moreover, if $(Y_t)_{t\geqslant 1}$ is ergodic (for instance if **[A⋆mixing]** holds), they are ergodic and:*

   - *if **[Atail]** holds, then for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$, $l(K,\mathbf{Q},\gamma)$ exists, is finite and*
   $$l(K,\mathbf{Q},\gamma)=\mathbb{E}^*[L_{1,\infty}(K,\mathbf{Q},\gamma)];$$
   - *if **[A⋆tail]** holds, then l^* exists and is finite and*
   $$l^*=\mathbb{E}^*[L^*_{1,\infty}].$$

Proof The second point follows directly from **[A⋆forgetting]**.

The third point follows from the ergodicity of $(Y_t)_{t\geqslant 1}$ under **[A⋆mixing]**, from the integrability of $L_{i,\infty}$ and $L^*_{i,\infty}$ under **[Atail]** and **[A⋆tail]** and from Lemmas 4 and 5. ■

Note that under the assumptions of point 3 of Lemma 6, one has $\mathbf{K}(K,\mathbf{Q},\gamma)=\mathbb{E}[L^*_{1,\infty}-L_{1,\infty}(K,\mathbf{Q},\gamma)]\in[0,+\infty)$ for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$ (recall that γ is a vector of probability densities in this case), or with some notation abuses:

$$\begin{aligned}\mathbf{K}(K,\mathbf{Q},\gamma)&=\mathbb{E}^*\left[\log\left(\frac{p^*(Y_1|Y_{-\infty}^0)}{p_{(K,\mathbf{Q},\gamma)}(Y_1|Y_{-\infty}^0)}\right)\right]\\&=\mathbb{E}^*_{Y_{-\infty}^0}\left[KL(\mathbb{P}^*_{Y_1|Y_{-\infty}^0}\|\mathbb{P}_{Y_1|Y_{-\infty}^0,(K,\mathbf{Q},\gamma)})\right].\end{aligned}$$

Thus, $\mathbf{K}(K,\mathbf{Q},\gamma)$ can be seen as a Kullback Leibler divergence that measures the difference between the distribution of Y_1 conditionally to the whole past under the parameter $(K,\mathbf{Q},\gamma)$ and under the true distribution. It can be seen as the prediction error under the parameter $(K,\mathbf{Q},\gamma)$.

In the particular case where the true distribution of $(Y_t)_t$ is a finite state space hidden Markov model, $\mathbf{K}$ characterizes the true parameters, up to permutation of the hidden states, provided the emission densities are all distinct and the transition matrix is invertible, as shown in the following result.

Lemma 7 (Alexandrovich et al. (2016), Theorem 5) *Assume $(Y_t)_t$ is generated by a finite state space HMM with parameters $(K^*,\pi^*,\mathbf{Q}^*,\gamma^*)$. Assume $\mathbf{Q}^*$ is invertible and ergodic, that the emission densities $(\gamma^*_x)_{x\in[K^*]}$ are all distinct and that $\mathbb{E}^*\left[(\log\gamma^*_x(Y_1))^+\right]<\infty$ for all $x\in[K^*]$ (so that $l^*<\infty$).*

Then for all $K\in\mathbb{N}^*$, for all transition matrix $\mathbf{Q}$ of size K and for all K-uple of probability densities γ, one has $\mathbf{K}(K,\mathbf{Q},\gamma)\geqslant 0$.

In addition, if $K\leqslant K^*$, then $\mathbf{K}(K,\mathbf{Q},\gamma)=0$ if and only if $(K,\mathbf{Q},\gamma)=(K^*,\mathbf{Q}^*,\gamma^*)$ up to permutation of the hidden states.

3. Main results

3.1 Oracle inequality for the prediction error

The following theorem states an oracle inequality on the prediction error of our estimator. It shows that with high probability, our estimator performs as well as the best model of the

class in terms of Kullback Leibler divergence, up to a multiplicative constant and up to an additive term decreasing as $\frac{(\log n)^{\cdots}}{n}$, provided the penalty is large enough.

Theorem 8 *Assume **[A⋆forgetting]**, **[A⋆mixing]**, **[A⋆tail]**, **[Aergodic]**, **[Atail]**, **[Aentropy]** and **[Agrowth]** hold.*

*Let $(w_M)_{M\in\mathcal{M}}$ be a nonnegative sequence such that $\sum_{M\in\mathcal{M}} e^{-w_M} \leqslant e-1$. Assume $\sigma_-(n) = C_\sigma(\log n)^{-1}$ and $B(n) = C_B \log n$ for some constants $C_\sigma \geqslant 0$ and $C_B \geqslant 1 + \frac{\zeta}{4}$ (where ζ is defined in **[Agrowth]**). Let $\alpha \geqslant 0$. For all K and M, let*

$$(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n}) \in \underset{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M,n}}{arg\ max} \frac{1}{n} l_n(K,\pi,\mathbf{Q},\gamma),$$

$$(\hat{K}, \hat{M}) \in \underset{\substack{K\leqslant \frac{\log n}{2C_\sigma} \\ M\ s.t.\ m_M\leqslant n}}{arg\ max} \left(\frac{1}{n} l_n(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n}) - pen_n(K,M)\right)$$

and let

$$(\hat{K}, \hat{\pi}, \hat{\mathbf{Q}}, \hat{\gamma}) = (\hat{K}, \hat{\pi}_{\hat{K},\hat{M},n}, \hat{\mathbf{Q}}_{\hat{K},\hat{M},n}, \hat{\gamma}_{\hat{K},\hat{M},n})$$

be the nonparametric maximum likelihood estimator.

Then there exists constants A and C_{pen} depending only on α, C_σ, C_B, n_* and c_* and a constant n_0 depending only on α, C_σ and C_B such that for all

$$n \geqslant n_{growth} \vee n_0 \vee \exp\left(C_\sigma\left((1+C_*) \vee \frac{2-\rho_*}{1-\rho_*} \vee e^2\right)\right) \vee \exp\left(\frac{B^*}{C_B}\right) \vee \exp\sqrt{\frac{C_\sigma}{2}(n_*+1)},$$

for all $t \geqslant 1$, for all $\eta \leqslant 1$, with probability at least $1 - e^{-t} - 2n^{-\alpha}$,

$$\mathbf{K}(\hat{K}, \hat{\mathbf{Q}}, \hat{\gamma}) \leqslant (1+\eta) \underset{\substack{K\leqslant \frac{\log n}{2C_\sigma} \\ M\ s.t.\ m_M\leqslant n}}{\inf} \left\{ \inf_{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M_n}} \mathbf{K}(K,\mathbf{Q},\gamma) + 2pen_n(K,M) \right\} + \frac{A}{\eta} t \frac{(\log n)^{7+q}}{n}$$

as soon as

$$pen_n(K,M) \geqslant \frac{C_{pen}}{\eta} \frac{(\log n)^{7+q}}{n} \Big\{ w_M + (\log n)^{3+q}(m_M K + K^2 - 1) \times \left((\log n)^2 \log\log n + \log C_{aux}\right) \Big\}.$$

The proof of this theorem is presented in Section 5. Its structure and main steps are detailed in Section 5.1, and the proof of these steps are gathered in Section 5.2.

Note that this theorem is not specific to one choice of the parametric models $S_{K,M,n}$: one can choose the type of model that suits the density one wants to estimate best. In the following section, we use mixture models to estimate densities when $\mathcal{Y}$ is unbounded. If $\mathcal{Y}$ is compact, we could use $\mathbf{L}^2$ spaces and this oracle inequality would still hold.

The powers of $\log n$ in the term $(\log n)^{7+q}$ come from:

- The limitation of the dependency to the $\log n$ most recent observations, which induces a factor $(\log n)^2$;
- The dependency of $\sigma_-(n)$ and $B(n)$ on n, each of them at the root of a factor $(\log n)^2$;
- Truncating the emission densities (possible thanks to assumptions **[Atail]** and **[A⋆tail]**), which induces a factor $(\log n)^{2q}$;
- The use of a Bernstein inequality for exponentially α-mixing processes, which introduces a factor $(\log n)^2$ compared to a Bernstein inequality for independent variables. However, together with the previous point (the truncation of the emission densities), the two points only induce a factor $(\log n)^{1+q}$.

In the term $(\log n)^2 \log\log n$ of the penalty, a factor $\log n$ comes from the limitation of the dependency and a factor $\log n \log\log n$ from $\sigma_-(n)$. Finally, the term $(\log n)^{3+q}$ in the penalty comes from the dependency of $B(n)$ on n, from truncating the emission densities and from using a Bernstein inequality for exponentially α-mixing processes.

3.2 Minimax adaptive estimation using location-scale mixtures

In this section, we show that the oracle inequality of Theorem 8 allows to construct an estimator that is adaptive and minimax up to logarithmic factors when the observations are generated by a finite state space hidden Markov model. To do so, we consider models whose emission densities are finite mixtures of exponential power distributions, and use an approximation result by Kruijer et al. (2010).

Assume that $(Y_t)_{t\geqslant 1}$ is generated by a stationary HMM with parameters $(K^*, \mathbf{Q}^*, \gamma^*)$, which we call the true parameters. We consider the case $\mathcal{Y} = \mathbb{R}$ endowed with the probability λ with density $G_\lambda : y \longmapsto (\pi(1+y^2))^{-1}$ with respect to the Lebesgue measure. In order to quantify the approximation error by location-scale mixtures, we use the following assumptions from Kruijer et al. (2010).

(C1) *Smoothness.* $\log(\gamma_x^* G_\lambda)$ is locally β-Hölder with $\beta > 0$, i.e. there exists a polynomial L and a constant $R > 0$ such that if r is the largest integer smaller than β, one has

$$\forall y, y' \text{ s.t. } |y - y'| \leqslant R, \quad \left| \frac{\partial^r \log(\gamma_x^* G_\lambda)}{\partial y^r}(y) - \frac{\partial^r \log(\gamma_x^* G_\lambda)}{\partial y^r}(y') \right| \leqslant r! L(y) |x - y|^{\beta - r}.$$

(C2) *Moments.* There exists $\epsilon > 0$ such that

$$\forall j \in \{1, \dots, r\}, \quad \int \left| \frac{\partial^j \log(\gamma_x^* G_\lambda)}{\partial y^j}(y) \right|^{\frac{2\beta+\epsilon}{j}} (\gamma_x^* G_\lambda)(y) d\lambda(y) < \infty$$
$$\int L(y)^{\frac{2\beta+\epsilon}{\beta}} (\gamma_x^* G_\lambda)(y) d\lambda(y) < \infty$$

(C3) *Tail.* There exists positive constants c and τ such that

$$\gamma_x^* G_\lambda = O(e^{-c|y|^\tau}).$$

(C4) *Monotonicity.* $(\gamma_x^* G_\lambda)$ is positive and there exists $y_m < y_M$ such that $(\gamma_x^* G_\lambda)$ is nondecreasing on $(-\infty, y_m)$ and nonincreasing on $(y_M, +\infty)$.

All these assumptions refer to the functions $(\gamma_x^* G_\lambda)$, which are the densities of the emission distributions with respect to the Lebesgue measure. Hence, the choice of the dominating measure λ does not matter as far as regularity conditions are concerned.

Note that Kruijer et al. (2010) only assumed **(C3)** outside of a compact set. However, since the regularity assumption **(C1)** implies that $(\gamma_x^* G_\lambda)$ is continuous, one can assume **(C3)** for all y without loss of generality.

It is important to note that even though we require some regularity on the emission densities, for instance through the polynomial L and the constants β and τ, we do not need to know them to construct our estimator, thus making it adaptive.

We consider the following models. Let $p \geqslant 2$ be an even integer and

$$\psi(y) = \frac{1}{2\Gamma\left(1 + \frac{1}{p}\right)} e^{-y^p}.$$

Let $\mathcal{M} = \mathbb{N}^*$. We take $S_{K,M,n}$ as the set of parameters $(K, \pi, \mathbf{Q}, \gamma)$ such that

- $\inf \mathbf{Q} \geqslant \sigma_-(n) := (\log n)^{-1}$ and $\inf \pi \geqslant \sigma_-(n)$,
- For all $x \in [K]$, there exists $(s_{x,1}, \dots, s_{x,M}) \in [\frac{1}{M}; 1]^M$, $(\mu_{x,1}, \dots, \mu_{x,M}) \in [-n; n]^M$ and $w_x = (w_{x,1}, \dots, w_{x,M}) \in [0,1]^M$ such that $\sum_i w_{x,i} = 1$ and for all $y \in \mathbb{R}$,

$$\gamma_x(y) = \frac{1}{n^2} + \left(1 - \frac{1}{n^2}\right) \frac{1}{G_\lambda(y)} \sum_{i=1}^{M} w_{x,i} \frac{1}{s_{x,i}} \psi\left(\frac{y - \mu_{x,i}}{s_{x,i}}\right).$$

  In other words, the emission densities are mixtures of λ (with weight n^{-2}) and of M translations and dilatations of ψ.

Lemma 9 (Checking the assumptions) *Assume* $\inf \mathbf{Q}^* > 0$*, then:*

- ***[A⋆forgetting]*** *and* ***[A⋆mixing]*** *hold.*
- *Assume* ***(C3)****, then* ***[A⋆tail]*** *holds by taking* $B^* > \log \| \sum_x \gamma_x^* \|_\infty$ *and* $q = 1$.
- ***[Aergodic]*** *holds.*
- ***[Atail]*** *holds for all* $n \geqslant 10$ *by taking* $B(n) = 5 \log n$, $\mathcal{K}_n \subset \{K \,|\, K \leqslant n\}$ *and* $\mathcal{M}_n = \{M \,|\, m_M \leqslant n\}$ *with* $m_M = 2M$.
- ***[Aentropy]*** *and* ***[Agrowth]*** *hold for any* $\zeta > 0$ *by taking* $m_M = 2M$ *and* $C_{aux} = 4pn^3$.

Proof The first point follows from Lemma 1.

The second point follows from the fact that the densities γ_x^* are uniformly bounded under **(C3)** and by taking δ large enough in Lemma 2.

[**Aergodic**] holds by definition of the models.

See Section A.1.1 for the proof of the last two points. ■

Remark 10 *One can also take* $(s_{x,1}, \dots, s_{x,M}) \in [\frac{1}{n}; n]^M$*, in which case Lemma 9 holds by taking* $B(n) = 6 \log n$ *and* $C_{aux} = 2pn^4$*.*

The results of this section remain the same when the weight of λ *in the emission densities of* $S_{K,M,n}$ *is allowed to be larger than* n^{-2} *instead of being exactly* n^{-2}*.*

Lemma 4 from Kruijer et al. (2010) implies the following result.

Lemma 11 (Approximation rates) *Assume* ***(C1)-(C4)*** *hold. Then there exists a sequence of mixtures* $(g_{M,x})_M$ *such that* $n^{-2} + (1 - n^{-2}) g_{M,x} \in S^{(\gamma)}_{K^*,M,n}$ *for all* $n \geqslant M$ *and*

$$KL(\gamma^*_x \| g_{M,x}) = O(M^{-2\beta} (\log M)^{2\beta(1+\frac{p}{\tau})}).$$

Proof Proof in Section A.1.2. ■

Corollary 12 (Minimax adaptive estimation rates) *Assume* ***(C1)-(C4)*** *hold. Also assume that* $\inf \mathbf{Q}^* > 0$*. Then there exists a constant* $C > 0$ *such that for all* $M \geqslant 3$ *and* $n \geqslant M$*,*

$$\inf_{(K^*,\pi,\mathbf{Q},\gamma) \in S_{K^*,M,n}} \mathbf{K}(K^*, \mathbf{Q}, \gamma) \leqslant C \left(\frac{(\log n)^2}{n} + M^{-2\beta} (\log M)^{2\beta(1+\frac{p}{\tau})} (\log n)^9 \right)$$

Hence, using Theorem 8 with $pen_n(K, M) = (KM + K^2)(\log n)^{15}/n$*, there exists a constant* C *such that almost surely, there exists a (random)* n_0 *such that*

$$\begin{aligned} \forall n \geqslant n_0, \quad \mathbf{K}(\hat{K}_n, \hat{\mathbf{Q}}_n, \hat{\gamma}_n) &\leqslant C n^{\frac{-2\beta}{2\beta+1}} (\log n)^{16+\frac{p}{\tau}-\frac{7+\frac{p}{\tau}}{2\beta+1}} \\ &\leqslant C n^{\frac{-2\beta}{2\beta+1}} (\log n)^{16+\frac{p}{\tau}}. \end{aligned}$$

Proof Proof in Section A.1.3. ■

This result shows that our estimator reaches the minimax rate of convergence proved by Maugis-Rabusseau and Michel (2013) for density estimation in Hellinger distance, up to logarithmic factors. Since estimating a density is the same thing as estimating a one-state HMM, this means that our result is adaptive and minimax up to logarithmic factors when $K^* = 1$. As far as we know, knowing whether increasing the number of states makes the minimax rates of convergence better is still an open problem. It seems reasonable to think that it doesn't, which would imply that our estimator is in general adaptive and minimax.

4. Perspectives

The main result of this paper is a guarantee that maximum likelihood estimators based on nonparametric hidden Markov models give sensible results even in the misspecified setting, and that their error can be controlled nonasymptotically. Two properties of both the models and the true distributions are at the core of this result: a mixing property and a forgetting property, which can be seen as a local dependence property.

These two properties are not specific to hidden Markov models. Therefore, it is likely that our result can be generalized to many other models and distributions. To name a few, one could consider hidden Markov models with continuous state space as studied in Douc and Matias (2001) or Douc et al. (2011), or more generally partially observed Markov models, see for instance Douc et al. (2016) and reference therein. Special cases of partially observed Markov models are HMMs with autoregressive properties (Douc et al., 2004) and models with time inhomogeneous Markov regimes (Pouzo et al., 2016). One could also consider hidden Markov fields (Kunsch et al., 1995) and graphical models to generalize to more general distributions than time processes.

Another interesting approach is to consider other forgetting and mixing assumptions. For instance, Le Gland and Mevel (2000) state a more general version of the forgetting assumption where the constant is replaced by an almost surely finite random variable, and Gerencsér et al. (2007) give conditions under which the moments of this random variable are finite. Other mixing and weak dependence conditions have also been introduced in the litterature with the hope of describing more general processes, see for instance Dedecker et al. (2007).

5. Proof of the oracle inequality (Theorem 8)

5.1 Overview of the proof

By definition of $(\hat{K}, \hat{\pi}, \hat{\mathbf{Q}}, \hat{\gamma})$, one has for all $K \leqslant \frac{\log n}{2C_\sigma}$, for all M such that $m_M \leqslant n$ and for all $(K, \pi_{K,M}, \mathbf{Q}_{K,M}, \gamma_{K,M}) \in S_{K,M,n}$:

$$\frac{1}{n}l_n^* - \frac{1}{n}l_n(\hat{K}, \hat{\pi}, \hat{\mathbf{Q}}, \hat{\gamma}) \leqslant \frac{1}{n}l_n^* - \frac{1}{n}l_n(K, \pi_{K,M}, \mathbf{Q}_{K,M}, \gamma_{K,M}) + \mathrm{pen}_n(K, M) - \mathrm{pen}_n(\hat{K}, \hat{M})$$

where $\hat{K}$ and $\hat{M}$ are the selected number of hidden states and model index respectively.

Let

$$\nu(K, \pi, \mathbf{Q}, \gamma) := \left(\frac{1}{n}l_n^* - \frac{1}{n}l_n(K, \pi, \mathbf{Q}, \gamma)\right) - \mathbf{K}(K, \mathbf{Q}, \gamma),$$

then

$$\begin{aligned}\mathbf{K}(\hat{K}, \hat{\mathbf{Q}}, \hat{\gamma}) \leqslant\ & \mathbf{K}(K, \mathbf{Q}_{K,M}, \gamma_{K,M}) + 2\mathrm{pen}_n(K, M) \\ & + \nu(K, \pi_{K,M}, \mathbf{Q}_{K,M}, \gamma_{K,M}) - \mathrm{pen}_n(K, M) \\ & - \nu(\hat{K}, \hat{\pi}, \hat{\mathbf{Q}}, \hat{\gamma}) - \mathrm{pen}_n(\hat{K}, \hat{M}).\end{aligned}$$

Now, assume that with high probability, for all K, M and $(K, \pi, \mathbf{Q}, \gamma) \in S_{K,M,n}$,

$$|\nu(K, \pi, \mathbf{Q}, \gamma)| - \mathrm{pen}_n(K, M) \leqslant \eta\mathbf{K}(K, \mathbf{Q}, \gamma) + R_n \tag{2}$$

for some constant $\eta \in (0, \frac{1}{2})$, some penalty pen_n and some residual term R_n. The above inequality leads to

$$(1 - \eta)\mathbf{K}(\hat{K}, \hat{\mathbf{Q}}, \hat{\gamma}) \leqslant (1 + \eta)\mathbf{K}(K, \mathbf{Q}_{K,M}, \gamma_{K,M}) + 2\mathrm{pen}_n(K, M) + 2R_n,$$

and the oracle inequality follows by noticing that $\frac{1+\eta}{1-\eta} \leqslant 1+4\eta$ and $\frac{1}{1-\eta} \leqslant 2$ when $\eta \in (0, \frac{1}{2})$.

Let us now prove equation (2). The following remark will be useful in our proofs: since

$$p_{(K,\pi,\mathbf{Q},\gamma)}(X_k = x|Y_1^{k-1}) = \frac{\sum_{x' \in [K]} p_{(K,\pi,\mathbf{Q},\gamma)}(X_{k-1} = x'|Y_1^{k-2})\mathbf{Q}(x',x)\gamma_{x'}(Y_{k-1})}{\sum_{x' \in [K]} p_{(K,\pi,\mathbf{Q},\gamma)}(X_{k-1} = x'|Y_1^{k-2})\gamma_{x'}(Y_{k-1})}$$
$$\in [\sigma_-(n); 1] \qquad \text{using } \textbf{[Aergodic]},$$

one has for all k, μ and $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$

$$L_{i,k,\mu}(K, \mathbf{Q}, \gamma) \in \left[\log \sigma_-(n) + \log \sum_{x \in [K]} \gamma_x(Y_i); \log \sum_{x \in [K]} \gamma_x(Y_i)\right]$$
$$= [\log \sigma_-(n) + b_\gamma(Y_i); b_\gamma(Y_i)] \tag{3}$$

and finally for all $k, k' \in \mathbb{N}^*$, for all μ, μ' probability distributions and for all $(K, \pi, \mathbf{Q}, \gamma)$ and $(K', \pi', \mathbf{Q}', \gamma') \in \mathbf{S}_n$,

$$\begin{cases} |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L_{i,k',\mu'}(K', \mathbf{Q}', \gamma')| \leqslant \log \frac{1}{\sigma_-(n)} + |b_\gamma(Y_i)| + |b_{\gamma'}(Y_i)|, \\ |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L^*_{i,k'}| \leqslant \log \frac{1}{\sigma_-(n)} + |b_\gamma(Y_i)| + |L^*_{i,k'}|. \end{cases} \tag{4}$$

Approximate $\nu(K, \pi, \mathbf{Q}, \gamma)$ by the deviation

$$\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)}) := \frac{1}{n} \sum_{i=1}^{n} t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k}) - \mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})]$$

where $D > 0$ and

$$t^{(D)}_{(K,\mathbf{Q},\gamma)} : Y^0_{-k} \longmapsto (L^*_{0,k} - L_{0,k,x}(K, \mathbf{Q}, \gamma))\mathbf{1}_{|L^*_{0,k}| \vee (\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|) \leqslant D}$$

for a fixed $x \in [K]$. Note that $\|t^{(D)}_{(K,\mathbf{Q},\gamma)}\|_\infty \leqslant 2D + \log \frac{1}{\sigma_-(n)}$ thanks to equation (3).

Considering these functions $t^{(D)}_{(K,\mathbf{Q},\gamma)}$ has two advantages. The first one is to limit the time dependency to an interval of length k, which makes it possible to use the forgetting property of the process $(Y_t)_{t \in \mathbb{Z}}$. The second one is to consider bounded functionals of this process, for which one can get Bernstein-like concentration inequalities. The error of this approximation is given by the following lemma.

Lemma 13 *Assume **[Atail]**, **[Aergodic]**, **[A⋆tail]** and **[A⋆forgetting]** hold. Also assume $B(n) \geqslant B^*$ and $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*}$. Then for all $u \geqslant 1$, with probability greater than $1 - 2ne^{-u}$, for all $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$,*

$$\left|\nu(K, \pi, \mathbf{Q}, \gamma) - \bar{\nu}_k(t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)})\right| \leqslant \left(6B(n)u^q + \log \frac{1}{\sigma_-(n)}\right) e^{-u} + \frac{2}{n\rho(1-\rho)^2} + \frac{4\rho^{k-1}}{1-\rho} \tag{5}$$

where $\rho = 1 - \frac{\sigma_-(n)}{1-\sigma_-(n)}$.

Proof Proof in Section 5.2.2. ■

The following theorem is our main technical result. It shows that $\bar{\nu}_k(t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)})$ can be controlled uniformly on all models with high probability.

Theorem 14 *Assume **[Aergodic]**, **[Aentropy]** and **[A⋆mixing]**. Also assume that there exists n_1 such that for all $n \geqslant n_1$, for all $K \leqslant n$ and M such that $m_M \leqslant n$,*

$$6900\pi\,(m_M K + K^2 - 1)ke^{-4D}(\log n)^3\,(k + \log C_{aux}) \leqslant n. \tag{6}$$

Let $(w_M)_{M\in\mathcal{M}}$ be a sequence of positive numbers such that $\sum_M e^{-w_M} \leqslant e-1$. Then there exists constants C_{pen} and A depending on n_* and c_* and a numerical constant n_0 such that for all $\epsilon > 0$ and $n \geqslant n_1 \vee n_0$, the following holds.

Let pen_n be a function such that for all $K \leqslant n$ and M such that $m_M \leqslant n$,

$$\begin{aligned} pen_n(K,M) \geqslant \frac{C_{pen}}{n}(n_*+k+1)^2\Bigg[&\left(D + \log\frac{1}{\sigma_-(n)}\right)(\log n)^2(m_M K + K^2 - 1) \\ &\times\left(\frac{1}{\epsilon} \vee \frac{\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_*+k+1}\right)\left(\log n + k\log\frac{2}{\sigma_-(n)} + D + \log C_{aux}\right) \\ &+\left(\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2 + \left(\frac{1}{\epsilon}\vee\frac{\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_*+k+1}\right)\right)w_M\Bigg]. \end{aligned} \tag{7}$$

Then for all $s > 0$, with probability larger than $1 - e^{-s}$, for all $K \leqslant n \wedge \frac{1}{2\sigma_-(n)}$ and M such that $m_M \leqslant n$ and for all $(K,\pi,\mathbf{Q},\gamma) \in S_{K,M,n}$,

$$\begin{aligned} &|\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - pen_n(K,M) \leqslant \epsilon\mathbb{E}[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})^2] \\ &+ A(n_*+k+1)^2\left(\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2 + \frac{1}{\epsilon}\vee\frac{\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_*+k+1}\right)\frac{s}{n}. \end{aligned} \tag{8}$$

Proof Proof in Section B. ■

The last step is to control the variance term $\mathbb{E}[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})^2]$ by $\mathbf{K}(K,\mathbf{Q},\gamma)$.

Lemma 15 *Assume **[Atail]**, **[Aergodic]**, **[A⋆tail]** and **[A⋆forgetting]** hold. Also assume that $B(n) \geqslant B^*$ and $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*} \wedge e^{-2}$. Then for all k such that*

$$k \geqslant \frac{1}{\sigma_-(n)}\left(\log n + 2\log\frac{1}{\sigma_-(n)}\right),$$

one has for all $D > 0$, $v \geqslant \log n$ and $(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n$:

$$\frac{1}{3(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2}\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] \leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{22}{n}.$$

Proof Proof in Section 5.2.3. ■

Now that the main lemmas have been stated, let us show how the assumptions of Theorem 8 leads to the desired oracle inequality.

Let C_σ and C_B be two positive constants and let

$$\begin{cases} \sigma_-(n) = C_\sigma (\log n)^{-1} \\ B(n) = C_B \log n. \end{cases}$$

Let $\alpha \geqslant 0$. In order to have $ne^{-u} \leqslant n^{-\alpha}$, take

$$u = (1+\alpha)\log n.$$

Note that $u \geqslant 1$ for all $n \geqslant 3$. The assumptions on v and k are $v \geqslant \log n$ and $k \geqslant \frac{1}{\sigma_-(n)}\left(\log n + 2\log\frac{1}{\sigma_-(n)}\right)$ (note that the assumption on k entails $\rho^{k-1} \leqslant (1-\rho)/n$). Thus, there exists an integer n_0 depending on C_σ such that if $n \geqslant n_0$, these assumptions hold for

$$\begin{cases} k = \frac{2}{C_\sigma}(\log n)^2 \\ v = \log n \end{cases}.$$

In order to get $\epsilon \mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] \leqslant \eta \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{22\eta}{n}$ using Lemma 15, one needs

$$\frac{1}{\epsilon} \geqslant \frac{3}{\eta}\left(2C_B(\log n)^{1+q} + \log\frac{1}{C_\sigma} + \log\log n\right)^2.$$

This quantity is smaller than $\frac{48}{\eta}\left(C_B \vee \log\frac{1}{C_\sigma} \vee 1\right)^2 (\log n)^{2(1+q)}$. Let $C_\epsilon = 48(1+\alpha)^q(C_B \vee \log\frac{1}{C_\sigma} \vee 1)$ and

$$\begin{cases} \frac{1}{\epsilon} = \frac{C_\epsilon}{\eta}(\log n)^{2(1+q)} \\ D = B(n)u^q = C_B(1+\alpha)^q(\log n)^{1+q} \end{cases}.$$

There exists an integer n_0' depending only on C_σ and α such that for all $n \geqslant n_0'$,

$$\begin{aligned}\left(D + \log\frac{1}{\sigma_-(n)}\right)(\log n)^2 &= \left(2C_B(1+\alpha)^q(\log n)^{1+q} + \log\frac{1}{C_\sigma} + \log\log n\right)(\log n)^2 \\ &\leqslant \frac{(\log n)^{1-q}}{\epsilon}\end{aligned}$$

and therefore

$$\frac{1}{\epsilon} \vee \frac{\left(D + \log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_* + k + 1} \leqslant \frac{(\log n)^{1-q}}{\epsilon}.$$

Thus, there exists an integer n_0'' depending on C_σ, C_B and α such that for all $n \geqslant n_0'' \vee \exp(C_\sigma((1+C_*) \vee \frac{2-\rho_*}{1-\rho_*} \vee e^2)) \vee \exp(\frac{B^*}{C_B}) \vee \exp\sqrt{\frac{C_\sigma}{2}(n_*+1)}$ (so that $k = \frac{2}{C_\sigma}(\log n)^2 \geqslant n_* + 1$,

$B(n) \geqslant B^*$ and $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*} \wedge e^{-2}$), equation (7) is implied by

$$\begin{aligned}\mathrm{pen}_n(K,M) \geqslant \frac{2C_{\mathrm{pen}}}{n}\frac{16}{C_\sigma^2}(\log n)^4\frac{C_\epsilon}{\eta}(\log n)^{3+q} \\ \times \Bigg[w_M + 2C_B(1+\alpha)^q(\log n)^{3+q}(m_M K + K^2 - 1) \\ \times \left(\frac{2}{C_\sigma}(\log n)^2\left(\log\frac{1}{C_\sigma} + \log\log n\right) + \log C_{\mathrm{aux}}\right)\Bigg],\end{aligned}$$

such that equation (8) (combined with Lemma 15) implies

$$|\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - \mathrm{pen}_n(K,M) \leqslant \eta\mathbf{K}(K,\mathbf{Q},\gamma) + A\frac{16}{C_\sigma^2}(\log n)^4\frac{C_\epsilon}{\eta}(\log n)^{3+q}\frac{s}{n},$$

such that equation (5) implies

$$\left|\nu(K,\pi,\mathbf{Q},\gamma) - \bar{\nu}_k(t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)})\right| \leqslant \frac{12C_B(1+\alpha)^q(\log n)^{1+q}}{n^{\alpha+1}} + \frac{4(\log n)^2}{C_\sigma^2 n} + \frac{4}{n}$$

and such that when **[Agrowth]** holds and when $m_M \leqslant n$ and $K \leqslant n$, equation (6) is implied by

$$13800\pi n^2\frac{2}{C_\sigma}(\log n)^2 e^{-4(1+\alpha)^q C_B(\log n)^{1+q}}(\log n)^3 n^\zeta \leqslant n$$

for all $n \geqslant n_{\mathrm{growth}}$, which is itself implied by

$$\frac{27600\pi}{C_\sigma}n^2(\log n)^5 e^{-4C_B\log n}n^\zeta \leqslant n$$

i.e.

$$\frac{27600\pi}{C_\sigma}(\log n)^5 n^{1+\zeta-4C_B} \leqslant 1,$$

which holds for all $n \geqslant n_0''$ (up to modification of n_0'') when $C_B \geqslant 1 + \frac{\zeta}{4}$. Putting these equations together proves Theorem 8.

5.2 Proofs

5.2.1 Upper bounds for the moments of the tails

Let W be a nonnegative random variable such that for all $u \geqslant 0$, $\mathbb{P}^*(W \geqslant u^q) = e^{-u}$ (if $q > 0$; otherwise $W = 0$). Assumption **[Atail]** implies that there exists a coupling of W and $\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|$ such that on the event $\{\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)| \geqslant B(n)\}$, one has $\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)| \leqslant B(n)W$ $\mathbb{P}^*$-almost surely. Therefore, controlling the moments of W is enough to control the moments of $\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|$.

For $u \geqslant 0$, let

$$\begin{aligned}E_q(u) &= \mathbb{E}[W\mathbf{1}_{W\geqslant u}], \\ V_q(u) &= \mathbb{E}[W^2\mathbf{1}_{W\geqslant u}].\end{aligned}$$

Lemma 16 *For all $u \geqslant 1$,*

$$\begin{cases} E_q(u) \leqslant 2u^q e^{-u} \\ V_q(u) \leqslant 5u^{2q} e^{-u} \end{cases}.$$

Proof One has

$$\begin{aligned} E_q(u) &= \int_{t \geqslant 0} \mathbb{P}(W \geqslant t \vee u^q) dt \\ &= u^q e^{-u} + \int_{t \geqslant u^q} e^{-t^{1/q}} dt \\ &= u^q e^{-u} + q \int_{T \geqslant u} T^{q-1} e^{-T} dT \\ &\leqslant u^q e^{-u} + \int_{T \geqslant u} e^{-T} dT \qquad \text{since } q \leqslant 1 \\ &\leqslant 2u^q e^{-u}. \end{aligned}$$

Likewise,

$$\begin{aligned} V_q(u) &= \int_{a,b \geqslant 0} \mathbb{P}(W \geqslant a \vee b \vee u^q) dt \\ &= u^{2q} e^{-u} + 2 \int_{t \geqslant u^q} t e^{-t^{1/q}} dt \\ &= u^{2q} e^{-u} + 2q \int_{T \geqslant u} T^{2q-1} e^{-T} dT \\ &= u^{2q} e^{-u} + 2q u^{2q-1} e^{-u} + 2q(2q-1) \int_{T \geqslant u} T^{2q-2} e^{-T} dT \end{aligned}$$

by integration by parts, which is enough to conclude. ∎

5.2.2 Proof of Lemma 13 (approximating the likelihood)

Let $t_{(K,\mathbf{Q},\gamma)} : Y^0_{-k} \longmapsto L^*_{0,k} - L_{0,k,x}(K, \mathbf{Q}, \gamma)$. Then, since

$$\begin{aligned} \nu(K, \pi, \mathbf{Q}, \gamma) - \bar{\nu}_k(t_{(K,\mathbf{Q},\gamma)}) = &\frac{1}{n} \sum_{i=1}^n (L^*_{i,i-1} - L^*_{i,k}) \\ &- \frac{1}{n} \sum_{i=1}^n (L_{i,i-1,\pi}(K, \mathbf{Q}, \gamma) - L_{i,k,x}(K, \mathbf{Q}, \gamma)) \\ &- \mathbb{E}[L^*_{0,\infty} - L^*_{0,k}] + \mathbb{E}[L_{0,\infty}(K, \mathbf{Q}, \gamma) - L_{0,k,x}(K, \mathbf{Q}, \gamma)], \end{aligned}$$

one gets using Lemma 6 and [**A⋆forgetting**] that

$$\begin{aligned}|\nu(K,\pi,\mathbf{Q},\gamma)-\bar{\nu}_k(t_{(K,\mathbf{Q},\gamma)})| &\leqslant \frac{1}{n}\sum_{i=1}^{n}\frac{\rho^{(i-1)\wedge k-1}}{1-\rho}+C_*\frac{1}{n}\sum_{i=1}^{n}\rho_*^{(i-1)\wedge k-1}+\frac{\rho^{k-1}}{1-\rho}+C_*\rho_*^{k-1}\\ &\leqslant \frac{1}{n\rho(1-\rho)^2}+\frac{2\rho^{k-1}}{1-\rho}+C_*\left(\frac{1}{n\rho_*(1-\rho_*)}+2\rho_*^{k-1}\right)\\ &\leqslant \frac{2}{n\rho(1-\rho)^2}+\frac{4\rho^{k-1}}{1-\rho}\end{aligned}$$

as soon as

$$\begin{cases}\rho \geqslant \rho_* \\ \frac{1}{1-\rho} \geqslant C^*\end{cases},$$

which holds for $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*}$.

Then, note that

$$\begin{aligned}\bar{\nu}_k(t_{(K,\mathbf{Q},\gamma)})-\bar{\nu}_k(t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)}) = \frac{1}{n}\sum_{i=1}^{n} t_{(K,\mathbf{Q},\gamma)}(Y_{i-k}^i)\mathbf{1}_{|L_{i,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_i)|)>B(n)u^q}\\ -\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)\mathbf{1}_{|L_{0,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|)>B(n)u^q}].\end{aligned}$$

We restrict ourselves to the event $\bigcap_{i=1}^n\{|L_{i,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_i)|)\leqslant B(n)u^q\}$, which occurs with probability greater than $1-2ne^{-u}$ using assumptions [**Atail**] and [**A⋆tail**]. On this event,

$$\frac{1}{n}\sum_{i=1}^{n} t_{(K,\mathbf{Q},\gamma)}(Y_{i-k}^i)\mathbf{1}_{|L_{i,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_i)|)>B(n)u^q}=0.$$

Moreover,

$$|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)-t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)]| = \mathbb{E}^*[|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)|\mathbf{1}_{|L_{0,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|)>B(n)u^q}].$$

Equation (3) ensures that $|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)| \leqslant |L_{0,k}^*| + \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| + \log\frac{1}{\sigma_-(n)}$, so that

$$\begin{aligned}&|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)-t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)]|\\ &\leqslant \mathbb{E}^*\left[|L_{0,k}^*|\left(\mathbf{1}_{|L_{0,k}^*|>B(n)u^q}+\mathbf{1}_{|L_{0,k}^*|\leqslant B(n)u^q<\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|}\right)\right]\\ &\quad+\mathbb{E}^*\left[\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\left(\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)u^q}+\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\leqslant B(n)u^q<|L_{0,k}^*|}\right)\right]\\ &\quad+\mathbb{E}^*\left[\left(\log\frac{1}{\sigma_-(n)}\right)\left(\mathbf{1}_{|L_{0,k}^*|>B(n)u^q}+\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)u^q}\right)\right]\end{aligned}$$

[Atail] and **[A⋆tail]** imply that $\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} \left|b_{\gamma'}(Y_0)\right| / B(n)$ and $|L^*_{0,k}|/B^*$ can be upper bounded by the random variable W defined in Section 5.2.1, which means that for all $u \geqslant 1$,

$$\begin{aligned}
&|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y^0_{-k}) - t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})]| \\
&\qquad \leqslant B^* E_q(u) + B(n)u^q \mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| > B(n)u^q\right) \\
&\qquad\quad + B(n)E_q(u) + B(n)u^q\mathbb{P}^*(|L^*_{0,k}| > B(n)u^q) \\
&\qquad\quad + \left(\log\frac{1}{\sigma_-(n)}\right)\left(\mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| > B(n)u^q\right)\right. \\
&\qquad\qquad\qquad \left. + \mathbb{P}^*(|L^*_{0,k}| > B(n)u^q)\right) \\
&\qquad \leqslant 6B(n)u^q e^{-u} + 2\left(\log\frac{1}{\sigma_-(n)}\right)e^{-u}
\end{aligned}$$

as soon as $B(n) \geqslant B^*$, which concludes the proof.

5.2.3 Proof of Lemma 15 (controlling the variance residual)

Lemma 17 *Assume **[Atail]**, **[Aergodic]** and **[A⋆tail]** hold. Assume $\sigma_-(n) \leqslant e^{-2}$ and let*

$$\mathbf{V}(K,\mathbf{Q},\gamma) := \mathbb{E}^*\left[(L^*_{0,\infty} - L_{0,\infty}(K,\mathbf{Q},\gamma))^2\right].$$

Then for all $v \geqslant 1$,

$$\frac{1}{3(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2}\mathbf{V}(K,\mathbf{Q},\gamma) \leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{64}{3}e^{-v}.$$

Proof We need the following lemma :

Lemma 18 (Shen et al. (2013), Lemma 4) *For any two probability measures P and Q with density p and q and any $\lambda \in (0, e^{-4}]$,*

$$\mathbb{E}_P\left(\log\frac{p}{q}\right)^2 \leqslant H(P,Q)^2\left(12 + 2\left(\log\frac{1}{\lambda}\right)^2\right) + 8\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)^2 \mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right)\right]$$

where $H(P,Q)$ is the Hellinger distance between P and Q:

$$H(P,Q)^2 = -2\mathbb{E}_P[(q/p)^{1/2} - 1] = \int(\sqrt{p} - \sqrt{q})^2 d\lambda.$$

Take $P = \mathbb{P}^*_{Y_0|Y^{-1}_{-\infty}}$ and $Q = \mathbb{P}_{Y_0|Y^{-1}_{-\infty},(K,\mathbf{Q},\gamma)}$, so that $\mathbb{E}_P(\log\frac{p}{q})^2 = \mathbf{V}(K,\mathbf{Q},\gamma)$. Using equation (4), one gets

$$\begin{aligned}
\left(\log\frac{p}{q}\right)^2 &\leqslant \left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| + |L^*_{0,\infty}| + \log\frac{1}{\sigma-}\right)^2 \\
&\leqslant 2(1+\tau)\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|^2 + 2(1+\tau)|L^*_{0,\infty}|^2 + \left(1 + \frac{1}{\tau}\right)\left(\log\frac{1}{\sigma-}\right)^2
\end{aligned}$$

for any $\tau > 0$. Let v be a real number such that $2B(n)v^q = \log\frac{1}{\lambda} - \log\frac{1}{\sigma_-(n)}$, then

$$\begin{aligned}\mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right) &\leqslant \mathbf{1}\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| + |L^*_{0,\infty}| \geqslant \log\frac{1}{\lambda} - \log\frac{1}{\sigma_-(n)}\right)\\ &\leqslant \mathbf{1}\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| \vee |L^*_{0,\infty}| \geqslant B(n)v^q\right),\end{aligned}$$

so that

$$\begin{aligned}&8\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)^2 \mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right)\right]\\ &\leqslant 16(1+\tau)\mathbb{E}^*\left[|L^*_{0,\infty}|^2\left(\mathbf{1}_{|L^*_{0,\infty}|>B(n)v^q} + \mathbf{1}_{|L^*_{0,\infty}|\leqslant B(n)v^q < \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|}\right)\right]\\ &\quad + 16(1+\tau)\mathbb{E}^*\left[\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|^2\left(\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)v^q} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\leqslant B(n)v^q<|L^*_{0,\infty}|}\right)\right]\\ &\quad + 8\left(1+\frac{1}{\tau}\right)\left(\log\frac{1}{\sigma-}\right)^2\mathbb{E}^*\left[\mathbf{1}_{|L^*_{0,\infty}|>B(n)v^q} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)v^q}\right]\end{aligned}$$

[Atail] and **[A⋆tail]** imply that $\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}\left|b_{\gamma'}(Y_0)\right|/B(n)$ and $|L^*_{0,\infty}|/B^*$ can be upper bounded by the random variable W defined in Section 5.2.1, which means that for all $v \geqslant 1$,

$$\begin{aligned}&8\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)^2 \mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right)\right]\\ &\leqslant 16(1+\tau)\left((B^*)^2V_q(v) + B(n)^2v^{2q}\mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| > B(n)v^q\right)\right)\\ &\quad + 16(1+\tau)\left(B(n)^2V_q(v) + B(n)^2v^{2q}\mathbb{P}^*(|L^*_{0,k}| > B(n)v^q)\right)\\ &\quad + 8\left(1+\frac{1}{\tau}\right)\left(\log\frac{1}{\sigma_-(n)}\right)^2\left(\mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| > B(n)v^q\right) + \mathbb{P}^*(|L^*_{0,k}| > B(n)v^q)\right)\\ &\leqslant 16e^{-v}\left(\left(1+\frac{1}{\tau}\right)\left(\log\frac{1}{\sigma_-(n)}\right)^2 + 12(1+\tau)B(n)^2v^{2q}\right)\\ &\leqslant 64e^{-v}\left(\left(\log\frac{1}{\sigma_-(n)}\right)^2 + 4B(n)^2v^{2q}\right)\end{aligned}$$

as soon as $B(n) \geqslant B^*$ by taking $\tau = \frac{1}{3}$.

Therefore, for all $v \geqslant 1$ such that the real number λ defined by $2B(n)v^q = \log\frac{1}{\lambda} - \log\frac{1}{\sigma_-(n)}$ satisfies $\lambda \leqslant e^{-4}$ (i.e. $2B(n)v^q \geqslant 4 - \log\frac{1}{\sigma_-(n)}$, which holds as soon as $v \geqslant 1$ and

$\sigma_-(n) \leqslant e^{-1}$),

$$\begin{aligned}\mathbf{V}(K,\mathbf{Q},\gamma) &\leqslant \mathbb{E}^*_{Y_{-\infty}^{-1}}\left[H(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}}, \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})^2\right]\left(12 + 2\left(2B(n)v^q + \log\frac{1}{\sigma_-(n)}\right)^2\right)\\ &\qquad + 64\left(\left(\log\frac{1}{\sigma_-(n)}\right)^2 + 4B(n)^2v^{2q}\right)e^{-v}\\ &\leqslant \mathbb{E}^*_{Y_{-\infty}^{-1}}\left[KL(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}} \| \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})\right]\left(12 + 2\left(2B(n)v^q + \log\frac{1}{\sigma_-(n)}\right)^2\right)\\ &\qquad + 64\left(\log\frac{1}{\sigma_-(n)} + 2B(n)v^q\right)^2 e^{-v}\end{aligned}$$

using that the Kullback Leibler divergence is lower bounded by the Hellinger distance. The condition $2B(n)v^q > 4 - \log\frac{1}{\sigma_-(n)}$ ensures that $12 + 2(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2 \leqslant 3(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2$. Finally, using

$$\mathbb{E}^*_{Y_{-\infty}^{-1}}[KL(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}} \| \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})] = \mathbf{K}(K,\mathbf{Q},\gamma),$$

one gets

$$\mathbf{V}(K,\mathbf{Q},\gamma) \leqslant 3\left(2B(n)v^q + \log\frac{1}{\sigma_-(n)}\right)^2 \mathbf{K}(K,\mathbf{Q},\gamma) + 64\left(2B(n)v^q + \log\frac{1}{\sigma_-(n)}\right)^2 e^{-v}$$

and the lemma is proved by dividing both sides by $3\left(2B(n)v^q + \log\frac{1}{\sigma_-(n)}\right)^2$. ∎

The next step is the control of the difference between $\mathbf{V}(K,\mathbf{Q},\gamma)$ and $\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2]$. Taking $t_{(K,\mathbf{Q},\gamma)} : Y^0_{-k} \longmapsto L^*_{0,k} - L_{0,k,x}(K,\mathbf{Q},\gamma)$, one has by definition of $t^{(D)}_{(K,\mathbf{Q},\gamma)}$

$$\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] \leqslant \mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2].$$

Then,

$$\begin{aligned}|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] &- \mathbf{V}(K,\mathbf{Q},\gamma)|\\ &= \left|\mathbb{E}^*\left[(L^*_{0,k} - L_{0,k,x}(K,\mathbf{Q},\gamma))^2\right] - \mathbb{E}^*\left[(L^*_{0,\infty} - L_{0,\infty}(K,\mathbf{Q},\gamma))^2\right]\right|\\ &\leqslant \mathbb{E}^*|((L^*_{0,k} - L^*_{0,\infty}) - (L_{0,k,x} - L_{0,\infty})(K,\mathbf{Q},\gamma))\\ &\qquad \times ((L^*_{0,k} - L_{0,k,x}(K,\mathbf{Q},\gamma)) + (L^*_{0,\infty} - L_{0,\infty}(K,\mathbf{Q},\gamma)))|\\ &\leqslant 2\frac{\rho^{k-1}}{1-\rho}\left(\mathbb{E}^*\left[2\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| + |L^*_{0,k}| + |L^*_{0,\infty}|\right] + 2\log\frac{1}{\sigma_-(n)}\right)\\ &\leqslant 2\frac{\rho^{k-1}}{1-\rho}\left((2B(n) + 2B^*)(1 + E_q(1)) + 2\log\frac{1}{\sigma_-(n)}\right)\\ &\leqslant 4\frac{\rho^{k-1}}{1-\rho}\left(4B(n) + \log\frac{1}{\sigma_-(n)}\right).\end{aligned}$$

using Lemma 6, equation (4), Lemma 16, $B(n) \geqslant B^*$ and the condition on $\sigma_-(n)$ (which implies $\rho \geqslant \rho_*$ and $\frac{1}{1-\rho} \geqslant C_*$). Therefore, under the assumptions of Lemma 17, one has

$$\begin{aligned}
&\frac{1}{3(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2}\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y_{i-k}^i)^2] \\
&\leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{64}{3}e^{-v} + \frac{4\rho^{k-1}}{3(1-\rho)(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2}\left(4B(n) + \log\frac{1}{\sigma_-(n)}\right) \\
&\leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{64}{3}e^{-v} + \frac{8\rho^{k-1}}{3(1-\rho)(2B(n)v^q + \log\frac{1}{\sigma_-(n)})} \\
&\leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{64}{3}e^{-v} + \frac{2\rho^{k-1}}{3(1-\rho)}.
\end{aligned}$$

Let us take $k \geqslant -\frac{\log n}{\log\rho} + \frac{\log(1-\rho)}{\log\rho} + 1$ and $v \geqslant \log n$, so that

$$\begin{aligned}
\frac{64}{3}e^{-v} + \frac{2\rho^{k-1}}{3(1-\rho)} &\leqslant \frac{64}{3n} + \frac{2\frac{1}{n}(1-\rho)}{3(1-\rho)} \\
&\leqslant \frac{22}{n}.
\end{aligned}$$

The constant ρ is defined by $\rho = 1 - \frac{\sigma_-(n)}{1-\sigma_-(n)}$, so that $\frac{-1}{\log\rho} \leqslant \frac{1}{\sigma_-(n)}$ and $-\log(1-\rho) \leqslant \log\frac{1}{\sigma_-(n)}$. Therefore, the condition on k holds as soon as

$$k \geqslant \frac{1}{\sigma_-(n)}\left(\log n + 2\log\frac{1}{\sigma_-(n)}\right) \tag{9}$$

using that $\log\log x \leqslant (\log x)/e$ for all $x > 1$ and that $e(1-1/e) \geqslant 1$. Therefore, for all k satisfying equation (9), for all $D > 0$ and $v \geqslant \log n$,

$$\frac{1}{3(2B(n)v^q + \log\frac{1}{\sigma_-(n)})^2}\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y_{i-k}^i)^2] \leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{22}{n},$$

which concludes the proof.

Acknowledgements

I am grateful to Elisabeth Gassiat for her precious advice and insightful discussions.

References

Grigory Alexandrovich, Hajo Holzmann, and Anna Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

Animashree Anandkumar, Daniel J Hsu, and Sham M Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, volume 1, page 4, 2012.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Andrew R Barron. The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem. *The annals of Probability*, 13(4):1292–1303, 1985.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):211–229, 2016.

Charlotte Boyd, André E Punt, Henri Weimerskirch, and Sophie Bertrand. Movement models provide insights into variation in the foraging effort of central place foragers. *Ecological modelling*, 286:13–25, 2014.

Richard C Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys*, 2:107–144, 2005.

Laurent Couvreur and Christophe Couvreur. Wavelet-based non-parametric HMM's: theory and applications. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 1, pages 604–607. IEEE, 2000.

Yohann de Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *Journal of Machine Learning Research*, 17(111): 1–43, 2016.

Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Transactions on Information Theory*, 2017.

Jérôme Dedecker, Paul Doukhan, Gabriel Lang, León R José Rafael, Sana Louhichi, and Clémentine Prieur. *Weak dependence: With examples and applications*. Springer, 2007.

Randal Douc and Catherine Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7(3):381–420, 2001.

Randal Douc and Eric Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *The Annals of Statistics*, 40(5):2697–2732, 2012.

Randal Douc, Eric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *The Annals of statistics*, 32(5):2254–2304, 2004.

Randal Douc, Gersende Fort, Eric Moulines, and Pierre Priouret. Forgetting the initial distribution for hidden Markov models. *Stochastic processes and their applications*, 119 (4):1235–1256, 2009.

Randal Douc, Eric Moulines, Jimmy Olsson, and Ramon Van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *the Annals of Statistics*, 39(1):474–513, 2011.

Randal Douc, Jimmy Olsson, and Francois Roeff. Posterior consistency for partially observed Markov models. *arXiv preprint arXiv:1608.06851*, 2016.

László Gerencsér, György Michaletzky, and Gábor Molnár-Sáska. An improved bound for the exponential stability of predictive filters of hidden Markov models. *Communications in Information & Systems*, 7(2):133–152, 2007.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Willem Kruijer, Judith Rousseau, and Aad Van Der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

Hans Kunsch, Stuart Geman, Athanasios Kehagias, et al. Hidden markov random fields. *The annals of applied probability*, 5(3):577–602, 1995.

Martin F Lambert, Julian P Whiting, and Andrew V Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences Discussions*, 7(5):652–667, 2003.

François Le Gland and Laurent Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems*, 13(1):63–93, 2000.

Fabrice Lefèvre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speech & Language*, 17(2):113–136, 2003.

Luc Lehéricy. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *arXiv preprint arXiv:1706.08277*, 2017.

Brian G Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

Pascal Massart. Concentration inequalities and model selection. In *Lecture Notes in Mathematics*, volume 1896. Springer, Berlin, 2007.

Cathy Maugis-Rabusseau and Bertrand Michel. Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM: Probability and Statistics*, 17:698–724, 2013.

Florence Merlevède, Magda Peligrad, and Emmanuel Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, pages 273–292. Institute of Mathematical Statistics, 2009.

Laurent Mevel and Lorenzo Finesso. Asymptotical statistics of misspecified hidden Markov models. *IEEE Transactions on Automatic Control*, 49(7):1123–1132, 2004.

Demian Pouzo, Zacharias Psaradakis, and Martin Sola. Maximum likelihood estimation in possibly misspecified dynamic models with time inhomogeneous Markov regimes. 2016.

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Elodie Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9(1):717–752, 2015a.

Elodie Vernet. Non parametric hidden markov models with finite state space: posterior concentration rates. *arXiv preprint arXiv:1511.08624*, 2015b.

Stevenn Volant, Caroline Bérard, Marie-Laure Martin-Magniette, and Stéphane Robin. Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, 24(4):493–504, 2014.

C Yau, Omiros Papaspiliopoulos, Gareth O Roberts, and Christopher Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57, 2011.

Appendix A. Proofs for the minimax adaptive estimation

A.1 Proofs for the mixture framework

A.1.1 Proof of Lemma 9 (checking the assumptions)

Checking [Atail] By definition of the emission densities, $b_\gamma(y) \geqslant -2\log n$ for all $\gamma \in \mathbf{S}_n^{(\gamma)}$. Moreover, for all $y \in \mathcal{Y}$ and $\gamma \in S_{K,M,n}^{(\gamma)}$,

$$\begin{aligned}
b_\gamma(y) &\leqslant \log\left(\sum_{x\in[K]}\left(1 \vee \frac{\max_{\mu,s}\frac{1}{s}\psi\left(\frac{y-\mu}{s}\right)}{G_\lambda(y)}\right)\right)\\
&\leqslant \log K + 0 \vee \left(\max_{\mu,s}\log\frac{1}{s}\psi\left(\frac{y-\mu}{s}\right) - \log G_\lambda(y)\right)\\
&\leqslant \log n + 0 \vee \left(\max_{\mu,s}\left\{\log\frac{1}{s} - \left(\frac{y-\mu}{s}\right)^p\right\} + \log(1+y^2) + \log\frac{\pi}{2\Gamma(1+1/p)}\right)\\
&\leqslant \log n + 0 \vee \left(-\min_{\mu}(y-\mu)^p + \log(1+y^2) + \log M + \log\pi\right),
\end{aligned}$$

where we recall that the maximum is taken over $\mu \in [-n, n]$ and $s \in [\frac{1}{M}, 1]$. By the choice of $\mathcal{K}_n$ and $\mathcal{M}_n$, one also has $K \leqslant n$ and $m_M \leqslant n$, i.e. $M \leqslant \frac{n}{2}$.

If $y \in [-n, n]$,

$$\begin{aligned}
b_\gamma(y) &\leqslant \log n + 0 \vee \big(\log(1+y^2) + \log M + \log\pi\big)\\
&\leqslant \log n + 0 \vee \big(\log(1+n^2) + \log(n/2) + \log\pi\big)\\
&\leqslant 4\log n + \log(\pi e/2) \leqslant 5\log n
\end{aligned}$$

as soon as $n \geqslant 5$. Otherwise, one can take $y \geqslant n$ and then

$$\begin{aligned} b_\gamma(y) &\leqslant \log n + 0 \vee (-(y-n)^p + \log(1+y^2) + \log M + \log \pi) \\ &\leqslant \log n + 0 \vee (-(y-n)^p + \log(1+2(y-n)^2 + 2n^2) + \log M + \log \pi) \\ &\leqslant \log n + 0 \vee (-Y^p + \log(1+2Y^2) + \log 2n^2 + \log(n/2) + \log \pi) \end{aligned}$$

by writing $Y = y - n$ and using that $\log(a+b) \leqslant \log a + \log b$ when $a, b \geqslant 1$. Since $\max_{Y \geqslant 0}(-Y^p + \log(1+2Y^2)) \leqslant \log 3$ as soon as $p \geqslant 2$, one gets

$$b_\gamma(y) \leqslant 4 \log n + \log 3\pi \leqslant 5 \log n$$

as soons as $n \geqslant 10$.

Checking [Aentropy] and [Hgrowth] Let us first assume that there exists a constant L_p such that the function $(\mu, s) \longmapsto \frac{s^{-1}\psi(s^{-1}(y-u))}{G_\lambda(y)}$ is L_p-Lipschitz for all y (where the origin space is endowed with the supremum norm). Then a bracket covering of size ϵ of $([n,n] \times [\frac{1}{M}, 1])^M$ provides a bracket covering of $\{\gamma(\cdot|x)\}_{\gamma \in \mathbf{S}_n^{(\gamma)}, x \in [K}$ of size $L_p\epsilon$. Since there exists a bracket covering of size ϵ of $[n,n] \times [\frac{1}{M}, 1]$ for the supremum norm with less than $(\frac{4n}{\epsilon} \vee 1)^2$ brackets, one gets **[Aentropy]** by taking $C_{\text{aux}} = 4L_p n$ and $m_M = 2M$.

Let us now check that this constant L_p exists.

$$\begin{aligned} \left| \frac{\partial}{\partial \mu} \frac{\frac{1}{s}\psi\left(\frac{y-\mu}{s}\right)}{G_\lambda(y)} \right| &= \frac{1}{2\pi\Gamma(1+\frac{1}{p})(1+y^2)} \left| \frac{\partial}{\partial \mu} \frac{1}{s} \exp\left(- \left(\frac{y-\mu}{s} \right)^p \right) \right| \\ &= \frac{1}{2\pi\Gamma(1+\frac{1}{p})(1+y^2)s^2} \left| \frac{y-\mu}{s} \right|^{p-1} \exp\left(- \left(\frac{y-\mu}{s} \right)^p \right) \\ &\leqslant \frac{1}{s^2} Y^{p-1} \exp(-Y^p) \\ &\leqslant M^2 Z^{1-1/p} e^{-Z} \leqslant n^2 \end{aligned}$$

by writing $Y = |y-\mu|/s$ and $Z = Y^p$. Likewise,

$$\begin{aligned} \left| \frac{\partial}{\partial \mu} \frac{\frac{1}{s}\psi\left(\frac{y-\mu}{s}\right)}{G_\lambda(y)} \right| &= \frac{1}{2\pi\Gamma(1+\frac{1}{p})(1+y^2)} \left| -\frac{1}{s^2} + p\frac{1}{s}\frac{(y-\mu)^p}{s^{p+1}} \right| \exp\left(- \left(\frac{y-\mu}{s} \right)^p \right) \\ &\leqslant \frac{1}{s^2} |pZ - 1| e^{-Z} \\ &\leqslant n^2 \frac{p}{2} \end{aligned}$$

as soon as $p \geqslant 2$. Thus, one can take $L_p = pn^2$, which corresponds to $C_{\text{aux}} = 4pn^3$. With this C_{aux}, checking **[Hgrowth]** is straightforward for all $\zeta > 0$.

A.1.2 Proof of Lemma 11 (approximation rates)

Let $F(y) = e^{-c|y|^\tau}$. Lemma 4 of Kruijer et al. (2010) ensures that there exists $c' > 0$ and $H \geqslant 6\beta + 4p$ such that for all $x \in [K^*]$ and $u > 0$, there exists a mixture $g_{u,x}$ with

$O(u^{-1}|\log u|^{p/\tau})$ components, each with density $\frac{1}{u}\psi(\frac{\cdot-\mu}{u})$ with respect to the Lebesgue measure for some $\mu \in \{y \,|\, F(y) \geqslant c' u^H\}$, such that $g_{u,x}$ approximates the emission density γ_x^*:

$$\max_x KL(\gamma_x^* \| g_{u,x}) = O(u^{-2\beta}).$$

Take $s = u|\log u|^{-1-\frac{p}{\tau}}$. When $|\mu| \geqslant s^{-1}$, one has $F(\mu) \leqslant \exp(-cs^{-\tau}) = o(c' s^H)$. Thus, for s small enough, all translation parameters μ belong to $[-s^{-1}, s^{-1}]$. Moreover, by definition of s, the mixture $g_{u,x}$ contains fewer than s^{-1} components when s is small enough. Finally, we use that

$$s = u|\log u|^{-1-\frac{p}{\tau}} \implies u \leqslant s|\log s|^{1+\frac{p}{\tau}}.$$

Taking $s^{-1} = M$ and $g_{M,x} = g_{u,x}$ concludes the proof.

A.1.3 Proof of Corollary 12 (minimax adaptive estimation rate)

Denote by h the Hellinger distance, defined by $h(p,q)^2 = \mathbb{E}_P[(\sqrt{q/p}-1)^2]$ for all probability densities p and q associated to probability measures P and Q. Let

$$\mathbf{H}^2(K, \mathbf{Q}, \gamma) = \mathbb{E}^*_{Y^0_{-\infty}}\left[h^2(p^*_{Y_1|Y^0_{-\infty}}, p_{Y_1|Y^0_{-\infty},(K,\mathbf{Q},\gamma)})\right]$$

be the Hellinger distance between the distributions of Y_1 conditionally to $Y^0_{-\infty}$ under the true distribution and under the parameters $(K, \mathbf{Q}, \gamma)$ (see Lemma 6 for the definition of these conditional distributions).

The following lemma shows that the Kullback-Leibler divergence and the Hellinger distance are equivalent up to a logarithmic factor and a small additive term.

Lemma 19 *Assume that **[A⋆tail]**, **[A⋆forgetting]**, **[Atail]** and **[Aergodic]** hold with $B(n) = C_B \log n$ and $\sigma_-(n) = C_\sigma (\log n)^{-1}$.*

Then there exists a constant n_1 depending on C_B and C_σ such that for all $n \geqslant n_1 \vee \exp(\frac{B^*}{C_B})$, one has for all $(K, \mathbf{Q}, \gamma) \in \mathbf{S}_n$

$$\mathbf{H}^2(K, \mathbf{Q}, \gamma) \leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) \leqslant 5C_B(\log n)^2 \left(\mathbf{H}^2(K, \mathbf{Q}, \gamma) + \frac{3}{n}\right).$$

Proof The lower bound comes from the fact that the square of the Hellinger distance is smaller than the Kullback-Leibler divergence. For the upper bound, we use Lemma 4 of Shen et al. (2013): for all $v \geqslant 4$ and for all probability measures P and Q with densities p and q,

$$KL(p\|q) \leqslant h^2(p,q)\,(1+2v) + 2\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)\mathbf{1}\left\{\log\frac{p}{q} \geqslant v\right\}\right].$$

Take $p = p^*_{Y_1|Y^0_{-\infty}}$ and $q = p_{Y_1|Y^0_{-\infty},(K,\mathbf{Q},\gamma)}$. Then $\log\frac{p}{q} \leqslant |b_\gamma| + |L^*_{1,\infty}| + \log\frac{1}{\sigma_-(n)}$ and $\mathbf{1}\left\{\log\frac{p}{q} \geqslant v\right\} \leqslant \mathbf{1}\{|b_\gamma| \geqslant \frac{1}{2}(v - \log\frac{1}{\sigma_-(n)})\} \vee \mathbf{1}\{|L^*_{1,\infty}| \geqslant \frac{1}{2}(v - \log\frac{1}{\sigma_-(n)})\}$. Taking

$v = 2C_B(\log n)^2$, one gets that there exists n_1 depending only on C_B and C_σ such that for all $n \geqslant n_1$, $v + \log \sigma_-(n) \geqslant (C_B \log n)^2$ and $1 + 2v \leqslant 5C_B(\log n)^2$, so that

$$\begin{aligned}\mathbf{K}(K, \mathbf{Q}, \gamma) \leqslant & 5C_B(\log n)^2 \mathbf{H}^2(K, \mathbf{Q}, \gamma) \\ & + C_B(\log n)^2 \left\{ \mathbb{P}^*(|b_\gamma| \geqslant C_B(\log n)^2) + \mathbb{P}^*(|L^*_{1,\infty}| \geqslant C_B(\log n)^2) \right\} \\ & + 2\mathbb{E}^*[(|L^*_{1,\infty}| + |b_\gamma|) \\ & \qquad \times (\mathbf{1}\{|L^*_{1,\infty}| \geqslant C_B(\log n)^2\} \vee \mathbf{1}\{|b_\gamma| \geqslant C_B(\log n)^2\})].\end{aligned}$$

Note that **[A⋆tail]** also holds for $L^*_{1,\infty}$ using the uniform convergence of Lemma 6. This implies that $\mathbb{P}^*(|L^*_{1,\infty}| \geqslant C_B(\log n)^2) \leqslant \exp(-\log n) \leqslant n^{-1}$ since $C_B(\log n) \geqslant B^*$ for $n \geqslant \exp(\frac{B^*}{C_B})$. Likewise, **[Atail]** implies that $\mathbb{P}^*(|b_\gamma| \geqslant C_B(\log n)^2) \leqslant n^{-1}$.

The last expectation of the above equation can be written as

$$2\mathbb{E}^*[(a+b)\mathbf{1}\{a \vee b \geqslant C_B(\log n)^2\}]$$

where $a = |L^*_{1,\infty}|$ and $b = |b_\gamma|$. Then, note that

$$\begin{aligned}2\mathbb{E}^*[a\mathbf{1}\{a \vee b \geqslant C_B(\log n)^2\})] = & 2\mathbb{E}^*[a\mathbf{1}\{a \geqslant C_B(\log n)^2\})] \\ & + 2\mathbb{E}^*[a\mathbf{1}\{b \geqslant C_B(\log n)^2 > a\})] \\ \leqslant & 4C_B(\log n)^2 e^{-\log n} + 2C_B(\log n)^2 \mathbb{P}^*[b \geqslant C_B(\log n)^2] \\ \leqslant & 4C_B(\log n)^2 e^{-\log n} + 2C_B(\log n)^2 e^{-\log n} \\ \leqslant & 6C_B \frac{(\log n)^2}{n}\end{aligned}$$

using $C_B \log n \geqslant B^*$ and Lemma 16 for the first term and **[Atail]** for the second one. Likewise,

$$2\mathbb{E}^*[b\mathbf{1}\{a \vee b \geqslant C_B(\log n)^2\})] \leqslant 6C_B \frac{(\log n)^2}{n},$$

so that finally

$$\mathbf{K}(K, \mathbf{Q}, \gamma) \leqslant 5C_B(\log n)^2 \mathbf{H}^2(K, \mathbf{Q}, \gamma) + 14C_B \frac{(\log n)^2}{n},$$

which concludes the proof. ∎

Let $M \in \mathbb{N}^*$. Let $g_{M,x}$ be the approximating densities given by Lemma 11 and write $\gamma_{M,x} = n^{-2} + (1 - n^{-2})g_{M,x}$ for all $x \in [K^*]$. The following lemma controls the error $\mathbf{H}(K^*, \mathbf{Q}^*, (\gamma_{M,x})_x)$ coming from the approximation of the densities.

Lemma 20 *Assume $\sigma_-(n) \leqslant \inf \mathbf{Q}^*$, then*

$$\mathbf{H}^2(K^*, \mathbf{Q}^*, (\gamma_{M,x})_x) \leqslant \left(2 + \frac{32}{(\sigma_-(n))^3(1-\rho)^4}\right) \sum_{x \in [K^*]} h^2(\gamma^*_x, \gamma_{M,x})$$

Proof Let $p_x^* = p^*(X_1 = x|Y_{-\infty}^0)$ and $p_x = p_{(K^*,\mathbf{Q}^*,(\gamma_{M,x})_x)}(X_1 = x|Y_{-\infty}^0)$. The Cauchy-Schwarz inequality implies that $(\sqrt{\sum_x a_x} - \sqrt{\sum_x b_x})^2 \leqslant \sum_x (\sqrt{a_x} - \sqrt{b_x})^2$, so that

$$\begin{aligned}
h^2\left(\sum_x p_x^*\gamma_x^*, \sum_x p_x\gamma_{M,x}\right) &= \int \left(\sqrt{\sum_x p_x^*\gamma_x^*} - \sqrt{\sum_x p_x\gamma_{M,x}}\right)^2 d\lambda \\
&\leqslant \int \sum_x (\sqrt{p_x^*\gamma_x^*} - \sqrt{p_x\gamma_{M,x}})^2 d\lambda \\
&\leqslant 2\int \sum_x \left(p_x(\sqrt{\gamma_x^*} - \sqrt{\gamma_{M,x}})^2 + (\sqrt{p_x} - \sqrt{p_x^*})^2\gamma_x^*\right) d\lambda \\
&\leqslant 2\sum_x p_x h^2(\gamma_x^*, \gamma_{M,x}) + 2\sum_x (\sqrt{p_x^*} - \sqrt{p_x})^2 \\
&\leqslant 2\sum_x h^2(\gamma_x^*, \gamma_{M,x}) + 2\sum_x (\sqrt{p_x^*} - \sqrt{p_x})^2
\end{aligned}$$

Thus, one needs to control the expectation of the second term. Since p_x and p_x^* belong to $[\sigma_-(n); 1]$ by minoration of their transition matrices, one has

$$\sum_x (\sqrt{p_x} - \sqrt{p_x^*})^2 \in \left[\frac{1}{4}; \frac{1}{4\sigma_-(n)}\right] \sum_x (p_x - p_x^*)^2.$$

The following equation follows from a careful reading of the proof of Proposition 2.1 of De Castro et al. (2017) by noticing that the roles of γ^* and γ_M are symmetrical in their proof.

$$\sum_x |p_x - p_x^*| \leqslant \frac{4}{\sigma_-(n)(1-\rho)} \sum_{i=0}^{+\infty} \rho^i \frac{\max_x |\gamma_x^*(Y_{-i}) - \gamma_{M,x}(Y_{-i})|}{\sum_x \gamma_x^*(Y_{-i}) \vee \sum_x \gamma_{M,x}(Y_{-i})}.$$

Therefore, using the Cauchy-Schwarz inequality:

$$\begin{aligned}
\sum_x (p_x - p_x^*)^2 &\leqslant \left(\sum_x |p_x - p_x^*|\right)^2 \\
&\leqslant \frac{16}{(\sigma_-(n))^2(1-\rho)^3} \sum_{i=0}^{+\infty} \rho^i \left(\frac{\max_x |\gamma_x^*(Y_{-i}) - \gamma_{M,x}(Y_{-i})|}{\sum_x \gamma_x^*(Y_{-i}) \vee \sum_x \gamma_{M,x}(Y_{-i})}\right)^2.
\end{aligned}$$

Since $\frac{|a-b|}{2\sqrt{a\vee b}} \leqslant |\sqrt{a} - \sqrt{b}|$, one has

$$\begin{aligned}
\mathbb{E}^*\left(\frac{\max_x |\gamma_x^*(Y) - \gamma_{M,x}(Y)|}{\sum_x \gamma_x^*(Y) \vee \sum_x \gamma_{M,x}(Y)}\right)^2 &\leqslant \int \frac{\max_x (\gamma_x^*(y) - \gamma_{M,x}(y))^2}{\sum_x \gamma_x^*(y) \vee \sum_x \gamma_{M,x}(y)} d\lambda(y) \\
&\leqslant \sum_x \int \frac{(\gamma_x^*(y) - \gamma_{M,x}(y))^2}{\gamma_x^*(y) \vee \gamma_{M,x}(y)} d\lambda(y) \\
&\leqslant 4\sum_x \int \left(\sqrt{\gamma_x^*(y)} - \sqrt{\gamma_{M,x}(y)}\right)^2 d\lambda(y) \\
&= 4\sum_x h^2(\gamma_x^*, \gamma_{M,x}),
\end{aligned}$$

so that

$$\mathbb{E}^*\left[\sum_x(\sqrt{p_x^*}-\sqrt{p_x})^2\right] \leqslant \frac{1}{4\sigma_-(n)}\mathbb{E}^*\left[\sum_x(p_x-p_x^*)^2\right]$$
$$\leqslant \frac{16}{(\sigma_-(n))^3(1-\rho)^4}\sum_x h^2(\gamma_x^*,\gamma_{M,x}),$$

which concludes the proof of the lemma. ■

Finally, since $|\sqrt{a+b}-\sqrt{c}| \leqslant |\sqrt{a}-\sqrt{c}|+\sqrt{|b|}$ for all $b\in\mathbb{R}$, $a \geqslant (-b)\vee 0$ and $c\geqslant 0$, one has for all x

$$h^2(\gamma_x^*,\gamma_{M,x}) \leqslant 2h^2(\gamma_x^*,g_{M,x})+\frac{4}{n^2}$$
$$\leqslant 2KL(\gamma_x^*\|g_{M,x})+\frac{4}{n^2}.$$

Therefore,

$$\mathbf{K}(K^*,\mathbf{Q}^*,(\gamma_{M,x})_x) \leqslant 15C_B\frac{(\log n)^2}{n}$$
$$+5C_B(\log n)^2\left(2+\frac{32}{(\sigma_-(n))^3(1-\rho)^4}\right)\sum_x\left(\frac{4}{n^2}+2KL(\gamma_x^*,g_{M,x})\right).$$

Since $\sigma_-(n)=C_\sigma(\log n)^{-1}$ and $(1-\rho)^{-1}\leqslant(\sigma_-(n))^{-1}$, there exists a constant C such that for all $n\geqslant 3$,

$$\mathbf{K}(K^*,\mathbf{Q}^*,(\gamma_{M,x})_x) \leqslant C\left(\frac{(\log n)^2}{n}+M^{-2\beta}(\log M)^{2\beta(1+\frac{p}{\tau})}(\log n)^9\right)$$

by definition of the densities $g_{M,x}$.

The choice of penalty verifies the lower bound of Theorem 8. Thus, the oracle inequality of Theorem 8 with $\eta=1$, $\alpha=2$ and $t=2\log n$ entails that for n large enough and for any sequence $(M_n)_n$ such that $M_n\leqslant n/2$ for all n:

$$\mathbf{K}(\hat{K},\hat{\mathbf{Q}},\hat{\gamma}) \leqslant 2\mathbf{K}(K^*,\mathbf{Q}^*,(\gamma_{M_n,x})_x)+2\mathrm{pen}_n(K^*,M_n)+A\frac{(\log n)^9}{n}$$
$$\leqslant 2C\left(\frac{(\log n)^2}{n}+M_n^{-2\beta}(\log n)^{2\beta(1+\frac{p}{\tau})+9}\right)$$
$$+2(K^*)^2\frac{(\log n)^{15}}{n}M_n+2A\frac{(\log n)^9}{n}.$$

Taking $M_n\sim n^{\frac{1}{2\beta+1}}(\log n)^{\frac{2\beta(1+p/\tau)-6}{2\beta+1}}$, one gets the announced rate.

Appendix B. Proof of the control of $\bar{\nu}_k$ (Theorem 14)

Let us give an overview of the proof of the control of $\bar{\nu}_k$.

The first step of the proof is to obtain a Bernstein inequality on $\bar{\nu}_k(t)$ for a single function t. This is done using the mixing properties of the process $(Y_i)_i$ and by noticing that $\bar{\nu}_k(t)$ is the deviation of an empirical mean.

The second step is to transform the inequality on one function t into an inequality on the supremum over all function t belonging to a given class. This step involves the bracketing entropy of the aforementionned class. The control of this entropy is where the shape of the penalty appears.

At this stage, one is able to upper bound the supremum of $\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})$ over all parameters $(K,\pi,\mathbf{Q},\gamma) \in S_{K,M,n}$. However, this upper bound is of order $n^{-1/2}$ (up to logarithmic factors), which is suboptimal. The third step of the proof gets rid of the $n^{-1/2}$ term by considering the processes

$$W_{K,M,n} := \sup_{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M,n}} \frac{|\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})|}{\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Z_0)^2] + x^2_{K,M,n}}$$

for some constants $x_{K,M,n}$. The last step of the proof consists in taking appropriate $x_{K,M,n}$ in order to have with high probability and for all K and M

$$\begin{cases} W_{K,M,n} \leqslant \epsilon \\ W_{K,M,n} x^2_{K,M,n} \leqslant \mathrm{pen}_n(K,M) + R_n \end{cases}$$

for a residual term R_n depending on the probability, which leads to the desired inequality

$$\forall (K,\pi,\mathbf{Q},\gamma) \in S_{K,M,n}, \quad |\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - \mathrm{pen}_n(K,M) \leqslant \epsilon \mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Z_0)^2] + R_n.$$

The concentration results are stated in Section B.1. The control of the bracketing entropy is done in Section B.2. Finally, the choice of $x_{K,M,n}$ and the synthesis of the proof are done in Section B.3.

In the rest of this Section, we omit the dependency of σ_-, B, $W_{K,M}$, $x_{K,M}$ and $S_{K,M}$ on n in the notations. We also introduce the notation $\theta = (K,\pi,\mathbf{Q},\gamma)$ for $(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n$ to make the notation shorter. Given $\theta \in \mathbf{S}_n$, we write π_θ, $\mathbf{Q}_\theta$ and γ_θ its components.

B.1 Concentration inequality

First, let us introduce some notations. Let $D > 0$, $K \geqslant 1$, $M \in \mathcal{M}$ and $k \geqslant 1$. For all $i \in \mathbb{Z}$, let $Z_i = Y^i_{i-k}$. Define for all $\sigma > 0$ the sets

$$\mathbf{B}_\sigma = \{\theta \in S_{K,M} \mid \mathbb{E}^*[t^{(D)}_\theta(Z_0)^2] \leqslant \sigma^2\}.$$

Let d_k be the semi-distance defined by $d_k^2(t_1,t_2) = \mathbb{E}^*[(t_1-t_2)^2(Z_0)]$. Let $N(A,d,\epsilon) = e^{H(A,d,\epsilon)}$ denote the minimal cardinality of a covering of A by brackets of size ϵ for the semi-distance d, that is by sets $[t_1,t_2] = \{t : \mathcal{Y}^k \longmapsto \mathbb{R}\,,\, t_1(\cdot) \leqslant t(\cdot) \leqslant t_2(\cdot)\}$ such that $d(t_1,t_2) \leqslant \epsilon$. $H(A,d,\cdot)$ is called the *bracketing entropy* of A for the semi-distance d.

The first step of the proof is to obtain a Bernstein inequality for the deviations of a single $t^{(D)}(Z_i)$.

Theorem 21 *Assume [**A⋆mixing**] holds. Then there exists a constant C_{mix} depending on c_* and n_* such that the following holds.*

Let t be a real valued, measurable bounded function on $\mathcal{Y}^{k+1}$. Let $V = \mathbb{E}^*[t^2(Z_0)]$. Then for all $\lambda \in (0, \frac{1}{C_{mix}(n_*+k+1)\|t\|_\infty(\log n)^2})$ and for all $n \in \mathbb{N}$:

$$\phi(\lambda) := \log \mathbb{E}^* \exp\left[\lambda \sum_{i=1}^{n} (t(Z_i) - \mathbb{E}^* t(Z_i))\right] \leqslant \frac{C_{mix}^2(n_*+k+1)^2(nV + \|t\|_\infty^2)\lambda^2}{1 - C_{mix}(n_*+k+1)\|t\|_\infty(\log n)^2\lambda}$$

Proof The following result is a Bernstein inequality for exponentially α-mixing processes.

Lemma 22 (Merlevède et al. (2009), Theorem 2) *Let $(A_i)_{i\geqslant 1}$ be a stationary sequence of centered real-valued random variables such that $\|A_1\|_\infty \leqslant M$ and whose α-mixing coefficients satisfy, for a certain $c > 0$,*

$$\forall n \in \mathbb{N}, \qquad \alpha_{mix}(n) \leqslant e^{-2cn}.$$

Then there exists positive constants C_1 and C_2 depending on c such that for all $n \geqslant 2$ and all $\lambda \in (0, \frac{1}{C_1 M(\log n)^2})$,

$$\log \mathbb{E} \exp\left[\lambda \sum_{i=1}^{n} A_i\right] \leqslant \frac{C_2\lambda^2(nv + M^2)}{1 - C_1\lambda M(\log n)^2},$$

where v is defined by

$$v = Var(A_1) + 2\sum_{i>1} |Cov(A_1, A_i)|.$$

Assumption **[A⋆mixing]** implies that the α-mixing coefficients of $(Y_i)_i$ satisfy $\alpha_{\text{mix}}(n) \leqslant e^{-c_* n}$ for all $n \geqslant n_*$ since $4\alpha_{\text{mix}}(n) \leqslant \rho_{\text{mix}}(n)$ (see for instance Bradley (2005)). However, this is not enough to apply the previous result: one needs the inequality to hold for all n (and not for n larger than some constant) and for the process $(Z_i)_i$. To do so, we partition the process $(Z_i)_i$ into several processes for which the above result applies, and then gather the inequalities.

Consider the processes $(Z_{i(n_*+k+1)+j})_i$ with α-mixing coefficients $\alpha_{Z,j}(n)$. By construction, they satisfy $\alpha_{Z,j}(n) \leqslant e^{-c_* n_* n}$ for all $n \geqslant 1$ and $j \in \{1, \dots, n_* + k + 1\}$. Apply Lemma 22, one gets that there exists two positive constants C_1 and C_2 depending on c_* and n_* such that for all function t, all $\lambda \in (0, \frac{1}{C_1 M(\log n)^2})$ and all $n \in \mathbb{N}$:

$$\begin{aligned}\phi_j(\lambda) &:= \log \mathbb{E}^* \exp\left[\lambda \sum_{i=1}^{n} (t(Z_{i(n_*+k+1)+j}) - \mathbb{E} t(Z_{i(n_*+k+1)+j}))\right] \\ &\leqslant \frac{C_2\lambda^2(nv + \|t\|_\infty^2)}{1 - C_1\lambda\|t\|_\infty(\log n)^2}\end{aligned}$$

where, denoting $V = \mathbb{E}^* t^2(Z_0)$:

$$\begin{aligned} v &= \mathrm{Var}(t(Z_j)) + 2\sum_{i>1} |\mathrm{Cov}\left(t(Z_j), t(Z_{i(n_*+k+1)+j})\right)| \\ &\leqslant V + 2V \sum_{i>1} |\mathrm{Corr}\left(t(Z_j), t(Z_{i(n_*+k+1)+j})\right)| \\ &\leqslant V\left(1 + 8\sum_{i>1} e^{-c_* n_* i}\right) \\ &\leqslant \frac{8V}{1-e^{-c_* n_*}} \end{aligned}$$

using **[A⋆mixing]**. Finally, using that $\mathbb{E}\prod_{i=1}^k A_i \leqslant \prod_{i=1}^k (\mathbb{E}A_i^k)^{1/k}$ for any positive integer k and any positive random variable $(A_i)_{1\leqslant i\leqslant k}$, one gets

$$\phi(\lambda) \leqslant \frac{1}{n_*+k+1} \sum_{j=1}^{n_*+k+1} \phi_j((n_*+k+1)\lambda),$$

so that

$$\phi(\lambda) \leqslant \frac{\frac{8C_2}{1-e^{-c_* n_*}}(n_*+k+1)^2\lambda^2(nV + \|t\|_\infty^2)}{1 - C_1(n_*+k+1)\lambda\|t\|_\infty(\log n)^2},$$

which concludes the proof. ∎

The following result follows *mutatis mutandis* from the proof of Theorem 6.8 of Massart (2007) using the previous theorem.

Lemma 23 *Assume **[A⋆mixing]** holds. Then there exists a constant $C^* \geqslant 1$ depending on n_* and c_* such that the following holds.*

Let $\mathcal{T}$ be a class of real valued and measurable functions on $\mathcal{Y}^{k+1}$ such that $\mathcal{T}$ is separable for the supremum norm. Also assume that there exists positive numbers σ and b such that for all $t \in \mathcal{T}$, $\|t\|_\infty \leqslant b$ and $\mathbb{E}^* t^2(Z_0) \leqslant \sigma^2$ and assume that $N(\mathcal{T}, d_k, \delta)$ is finite for all $\delta > 0$.

Then for all measurable set A such that $\mathbb{P}^*(A) > 0$:

$$\mathbb{E}^*\left(\sup_{t\in\mathcal{T}} |\bar{\nu}_k(t)| \Big| A\right) \leqslant C^*(n_*+k+1)\left[\frac{E}{n} + \sigma\sqrt{\frac{1}{n}\log\left(\frac{1}{\mathbb{P}^*(A)}\right)} + \frac{b(\log n)^2}{n}\log\left(\frac{1}{\mathbb{P}^*(A)}\right)\right]$$

where

$$E = \sqrt{n}\int_0^\sigma \sqrt{H(\mathcal{T}, d_k, u) \wedge n}\, du + b(\log n)^2 H(\mathcal{T}, d_k, \sigma).$$

Now, by taking $\mathcal{T} = \{t_\theta^{(D)} \,|\theta \in \mathbf{B}_\sigma\}$ and $b = 2D + \log\frac{1}{\sigma_-}$, one gets the following lemma from Lemma 4.23 and Lemma 2.4 of Massart (2007):

Lemma 24 *Assume that there exists a function φ and constants C and $\sigma_{K,M}$ such that $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing and*

$$\forall \sigma \geqslant \sigma_{K,M} \qquad E \leqslant C\varphi(\sigma)\sqrt{n}. \tag{10}$$

Then for all $x_{K,M} \geqslant \sigma_{K,M}$ and $z > 0$, one has with probability greater than $1 - e^{-z}$:

$$W_{K,M} := \sup_{\theta \in S_{K,M}} \left| \frac{|\bar{\nu}_k(t_\theta^{(D)})|}{\mathbb{E}^*[t_\theta^{(D)}(Z_0)^2] + x_{K,M}^2} \right| \leqslant 4C^*(n_* + k + 1)\left[C\frac{\varphi(x_{K,M})}{x_{K,M}^2\sqrt{n}} + \sqrt{\frac{z}{x_{K,M}^2 n}} + \left(2D + \log\frac{1}{\sigma_-}\right)\frac{z(\log n)^2}{x_{K,M}^2 n}\right]. \tag{11}$$

The two remaining steps are the control of the bracketing entropy which will lead to equation (10) (see Section B.2) and the choice of the parameters $x_{K,M}$ and z (see Section B.3).

B.2 Control of the bracketing entropy

B.2.1 Reduction of the set

For all $\theta \in S_{K,M}$, let $\mathbf{g}_\theta = (g_{\theta,x})_{x\in[K]}$ where

$$g_{\theta,x} : y_0^k \longmapsto \begin{cases} p_\theta(X_k = x, Y_k = y_k | Y_0^{k-1} = y_0^{k-1}) & \text{if } |L_{k,k}^*| \vee \sup_{\theta' \in \mathbf{S}_n} |b_{\theta'}(y_k)| \leqslant D, \\ 0 \quad \text{otherwise.} \end{cases}$$

In order to control the bracketing entropy of $\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}$, we control the bracketing entropy of the set $\mathcal{G} := \{\mathbf{g}_\theta \mid \theta \in S_{K,M}\}$ for the distance

$$d_{\mathcal{G}}(\mathbf{g}_{\theta_1}, \mathbf{g}_{\theta_2}) = \mathbb{E}^*_{Y_0^{k-1}}\left[\sum_{x\in[K]} \int |g_{\theta_1,x}(Y_0^{k-1}, y_k) - g_{\theta_2,x}(Y_0^{k-1}, y_k)| \times \mathbf{1}_{|L_{k,k}^*| \vee \sup_{\theta'\in\mathbf{S}_n} |b_{\theta'}(y_k)| \leqslant D} d\lambda(y_k)\right].$$

Remark 25 *In the rest of Section B.2, we always assume that*

$$|L_{k,k}^*| \vee \sup_{\theta' \in \mathbf{S}_n} |b_{\theta'}(y_k)| \leqslant D \tag{12}$$

since if this is not the case, then $t_\theta^{(D)}(y_k) = t_{\theta'}^{(D)}(y_k) = 0$. This means that only the y_k satisfying equation (12) are relevant for the construction of the brackets.

For all $\theta \in S_{K,M}$, one has

$$\begin{aligned} \sum_{x\in[K]} g_{\theta,x} &= \sum_{x,x'\in[K]} p_\theta(Y_k = y_k | X_k = x)\mathbf{Q}_\theta(x,x')p_\theta(X_{k-1} = x' | Y_0^{k-1} = y_0^{k-1}) \\ &\in [\sigma_-, 1] \sum_{x,x'\in[K]} p_\theta(Y_k = y_k | X_k = x)p_\theta(X_{k-1} = x' | Y_0^{k-1} = y_0^{k-1}) \\ &= [\sigma_-, 1]e^{b_\theta(y_k)} \end{aligned}$$

so that for all $\theta \in S_{K,M}$,

$$\sigma_- e^{-D} \leqslant \sum_{x\in[K]} g_{\theta,x} \leqslant e^D.$$

Let $[a,b]$ be a bracket of size ϵ for $\mathcal{G}$ with the distance $d_{\mathcal{G}}$ such that $\sigma_- e^{-D}/2 \leqslant \sum_x a_x \leqslant \sum_x b_x \leqslant 2e^D$. Then

$$\begin{aligned}\left(\log\sum_x a_x - \log\sum_x b_x\right)^2 &\leqslant \left(\log\frac{2e^D}{\sigma_-} + \log 2e^D\right)\left|\log\sum_x a_x - \log\sum_x b_x\right| \\ &\leqslant 2\left(D + \log\frac{1}{\sigma_-}\right)\frac{2e^D}{\sigma_-}\sum_x |a_x - b_x|\end{aligned}$$

using that $\sigma_- \leqslant 1/4$ and $|\log a - \log b| \leqslant |a-b|/(a\wedge b)$.

Therefore,

$$\begin{aligned}&d_k\left(\log\sum_x a_x, \log\sum_x b_x\right)^2 \\ &= \mathbb{E}^*_{Y_0^{k-1}}\left[\int\left(\log\sum_x a_x - \log\sum_x b_x\right)^2 (Y_0^{k-1}, y_k)p^*(Y_k = y_k|Y_0^{k-1})\lambda(dy_k)\right] \\ &\leqslant 4\left(D + \log\frac{1}{\sigma_-}\right)\frac{e^D}{\sigma_-}\mathbb{E}^*_{Y_0^{k-1}}\left[\int\sum_x |a_x - b_x|(Y_0^{k-1}, y_k)L^*_{k,k}\lambda(dy_k)\right] \\ &\leqslant 4\left(D + \log\frac{1}{\sigma_-}\right)\frac{e^{2D}}{\sigma_-}\mathbb{E}^*_{Y_0^{k-1}}\left[\int\sum_x |a_x - b_x|(Y_0^{k-1}, y_k)\lambda(dy_k)\right] \\ &= 4\left(D + \log\frac{1}{\sigma_-}\right)\frac{e^{2D}}{\sigma_-}d_{\mathcal{G}}(a,b),\end{aligned}$$

so that

$$N(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, \epsilon) \leqslant \bar{N}\left(\mathcal{G}, d_{\mathcal{G}}, \left(\frac{\sigma_-\epsilon}{4(D + \log\frac{1}{\sigma_-})e^{2D}}\right)^2\right) \tag{13}$$

where $\bar{N}$ is the minimal cardinality of a bracket covering of $\mathcal{G}$ such that all brackets $[a,b]$ satisfy $\sigma_- e^{-D}/2 \leqslant \sum_x a_x \leqslant \sum_x b_x \leqslant 2e^D$.

B.2.2 Decomposition into simple sets

The aim of this section is to prove the following lemma.

Lemma 26 *Let* $\epsilon \in \left(0, \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}\right)$. *Then*

$$\bar{N}\left(\mathcal{G}, d_{\mathcal{G}}, \epsilon\right) \leqslant N\left(\{\pi_\theta\}_{\theta\in S_{K,M}}, d_\infty, \left(\frac{\sigma_-}{2}\right)^k \frac{\epsilon}{106ke^D}\right) \times N\left(\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}, d_\infty, \left(\frac{\sigma_-}{2}\right)^k \frac{\epsilon}{106ke^D}\right) \times N\left(\{\gamma_\theta\}_{\theta\in S_{K,M}}, d_\infty, \left(\frac{\sigma_-}{2}\right)^k \frac{\epsilon\, e^{-D}}{106ke^D}\right)$$

where d_∞ *is the distance of the supremum norm and where* γ_θ *denotes the function* $(x, y) \longmapsto \gamma_\theta(y|x)$.

Let:

- $[a, b]$ be a bracket of $\{\pi_\theta\}_{\theta\in S_{K,M}}$ of size ϵ for the supremum norm ;
- $[p, q]$ be a bracket of $\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}$ of size ϵ pour the supremum norm ;
- $[u, v]$ be a bracket of $\{\gamma_\theta\}_{\theta\in S_{K,M}}$ of size ϵe^{-D} for the supremum norm.

Without loss of generality, one can assume $\sigma_- \leqslant a(x) \leqslant b(x) \leqslant 1$ and $\sigma_- \leqslant p(x, x') \leqslant q(x, x') \leqslant 1$ for all $x, x' \in [K]$ since all elements of $\{\pi_\theta\}_{\theta\in S_{K,M}}$ and $\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}$ satisfy these inequalities. One can also assume that there exists $\theta \in S_{K,M}$ such that $\pi_\theta \in [a, b]$, $\mathbf{Q}_\theta \in [p, q]$ and $\gamma_\theta \in [u, v]$. Under this assumption, all brackets that we construct are non empty and for all $y \in \mathcal{Y}$, $e^{-D}(1 - K\epsilon) \leqslant \sum_x u(y|x) \leqslant \sum_x v(y|x) \leqslant e^D + K\epsilon e^{-D}$.

Using the approach of Appendix A of De Castro et al. (2017), one can write $g_{\theta,x}$ as the following product of matrices

$$g_{\theta,x}(y_0^k) = \left(\mu^\theta_{0|k-1} F^\theta_{1|k-1} \dots F^\theta_{k-1|k-1} \mathbf{Q}_\theta\right)_x \gamma_\theta(y_k|x)$$

where

$$\begin{aligned}
\beta_{i|k}(x_i) &= \sum_{x_{i+1}^k \in [K]^{k-i}} \mathbf{Q}_\theta(x_i, x_{i+1})\gamma_\theta(y_{i+1}|x_{i+1}) \dots \mathbf{Q}_\theta(x_{k-1}, x_k)\gamma_\theta(y_k|x_k),\\
\mu^\theta_{0|k}(x) &= \frac{\pi_\theta(x)\beta_{0|k}(x)}{\sum_{x'\in[K]} \pi_\theta(x')\beta_{0|k}(x')},\\
F^\theta_{i|k}(x_{i-1}, x_i) &= \frac{\beta_{i|k}(x_i)\mathbf{Q}_\theta(x_{i-1}, x_i)\gamma_\theta(y_i|x_i)}{\sum_{x\in[K]} \beta_{i|k}(x)\mathbf{Q}_\theta(x_{i-1}, x)\gamma_\theta(y_i|x)}.
\end{aligned}$$

To clarify the role of these quantities, observe that

$$\begin{aligned}
\beta_{i|k}(x_i) &= \mathbb{P}_\theta(Y_{i+1}^k | X_i = x_i),\\
\mu^\theta_{0|k}(x) &= \mathbb{P}_\theta(X_0 = x | Y_1^k),\\
F^\theta_{i|k}(x_{i-1}, x_i) &= \mathbb{P}_\theta(X_i = x_i | Y_i^k, X_{i-1} = x_{i-1}),
\end{aligned}$$

so that

$$\left(\mu^\theta_{0|k} F^\theta_{1|k} \dots F^\theta_{k|k}\right)_x = \mathbb{P}_\theta(X_k = x | Y_1^k).$$

Now, let

$$\begin{cases} \alpha_{i|k}(x_i) = \displaystyle\sum_{x_{i+1}^k \in [K]^{k-i}} p(x_i, x_{i+1}) u(y_{i+1}|x_{i+1}) \dots p(x_{k-1}, x_k) u(y_k|x_k) \\ \\ \delta_{i|k}(x_i) = \displaystyle\sum_{x_{i+1}^k \in [K]^{k-i}} q(x_i, x_{i+1}) v(y_{i+1}|x_{i+1}) \dots q(x_{k-1}, x_k) v(y_k|x_k) \end{cases},$$

$$\begin{cases} \nu(x) = \dfrac{a(x)\alpha_{0|k}(x)}{\sum_{x' \in [K]} b(x')\delta_{0|k}(x')} \\ \\ \omega(x) = \dfrac{b(x)\delta_{0|k}(x)}{\sum_{x' \in [K]} a(x')\alpha_{0|k}(x')} \end{cases},$$

and

$$\begin{cases} f_{i|k}(x_{i-1}, x_i) = \dfrac{\alpha_{i|k}(x_i) p(x_{i-1}, x_i) u(y_i|x_i)}{\sum_{x \in [K]} \delta_{i|k}(x) q(x_{i-1}, x) v(y_i|x)} \\ \\ g_{i|k}(x_{i-1}, x_i) = \dfrac{\delta_{i|k}(x_i) q(x_{i-1}, x_i) v(y_i|x_i)}{\sum_{x \in [K]} \alpha_{i|k}(x) p(x_{i-1}, x) u(y_i|x)} \end{cases}.$$

$[\nu, \omega]$ and $[f_{i|k}, g_{i|k}]$ are brackets of $\{\mu^\theta_{0|k}\}_{\theta \in S_{K,M}}$ and $\{F^\theta_{i|k}\}_{\theta \in S_{K,M}}$ for all $i \in \{1, \dots, k\}$. Moreover, if one has a bracket covering of the sets $\{\pi_\theta\}_{\theta \in S_{K,M}}$, $\{\mathbf{Q}_\theta\}_{\theta \in S_{K,M}}$ and $\{\gamma_\theta\}_{\theta \in S_{K,M}}$, then this construction gives a bracket covering of $\{\mu^\theta_{0|k}\}_{\theta \in S_{K,M}}$ and $\{F^\theta_{i|k}\}_{\theta \in S_{K,M}}$ for all $i \in \{1, \dots, k\}$.

The next step of the proof is to control the size of these new brackets.

Lemma 27 *Assume $\epsilon \leqslant \frac{1}{2K}$, then*

$$\sup_{1 \leqslant i \leqslant k} \frac{\sum_{x \in [K]} |\alpha_{i|k}(x) - \delta_{i|k}(x)|}{\sum_{x \in [K]} \alpha_{i|k}(x)} \leqslant 4 \left(\frac{2}{\sigma_-}\right)^{k-i} \epsilon.$$

and

$$\sup_{1 \leqslant i \leqslant k} \frac{\sum_{x \in [K]} |\alpha_{i|k}(x) u(y_i|x) - \delta_{i|k}(x) v(y_i|x)|}{\sum_{x \in [K]} \alpha_{i|k}(x) u(y_i|x)} \leqslant 4 \left(\frac{2}{\sigma_-}\right)^{k-i+1} \epsilon.$$

Proof Using minimalist notations, one has

$$\begin{aligned} \sum_{x \in [K]} |\alpha_{i|k}(x) - \delta_{i|k}(x)| \leqslant & \sum_{j=i+1}^{k} \sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots u_{j-1} |p_{j-1}^j - q_{j-1}^j| v_j \dots q_{k-1}^k v_k \\ & + \sum_{j=i+1}^{k} \sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots p_{j-1}^j |u_j - v_j| q_j^{j+1} \dots q_{k-1}^k v_k. \end{aligned}$$

Then, note that for all j,

$$\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots p_{j-2}^{j-1} u_{j-1} |p_{j-1}^j - q_{j-1}^j| v_j q_j^{j+1} \dots q_{k-1}^k v_k$$
$$\leqslant \epsilon \sum_{x_i^{j-1} \in [K]^{j-i}} p_i^{i+1} u_{i+1} \dots p_{j-2}^{j-1} u_{j-1} \sum_{x_j \in [K]} (u_j + \epsilon e^{-D}) \cdots \sum_{x_k \in [K]} (u_k + \epsilon e^{-D})$$

and

$$\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots p_{j-2}^{j-1} u_{j-1} p_{j-1}^j u_j p_j^{j+1} \dots p_{k-1}^k u_k$$
$$\geqslant \sigma_-^{k-j+1} \sum_{x_i^{j-1} \in [K]^{j-i}} p_i^{i+1} u_{i+1} \dots p_{j-2}^{j-1} u_{j-1} \sum_{x_j \in [K]} u_j \cdots \sum_{x_k \in [K]} u_k.$$

so that

$$\begin{aligned}
\frac{\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots u_{j-1} |p_{j-1}^j - q_{j-1}^j| v_j \dots q_{k-1}^k v_k}{\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots u_{j-1} p_{j-1}^j u_j \dots p_{k-1}^k u_k} & \\
&\leqslant \frac{\epsilon}{\sigma_-^{k-j+1}} \prod_{\ell=j}^{k} \frac{K\epsilon e^{-D} + \sum_{x_\ell} u_\ell}{\sum_{x_\ell} u_\ell} \\
&\leqslant \frac{\epsilon}{\sigma_-^{k-j+1}} \prod_{\ell=j}^{k} \left(1 + \frac{K\epsilon e^{-D}}{e^{-D}(1 - K\epsilon)}\right) \\
&\leqslant \frac{\epsilon}{\sigma_-^{k-j+1}} \left(\frac{1}{1 - K\epsilon}\right)^{k-j+1},
\end{aligned}$$

and likewise

$$\frac{\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots p_{j-1}^j |u_j - v_j| q_j^{j+1} \dots q_{k-1}^k v_k}{\sum_{x_i^k \in [K]^{k-i+1}} p_i^{i+1} u_{i+1} \dots u_{j-1} p_{j-1}^j u_j \dots p_{k-1}^k u_k} \leqslant \frac{\epsilon}{\sigma_-^{k-j+1}} \left(\frac{1}{1 - K\epsilon}\right)^{k-j}.$$

Therefore, when $\epsilon K \leqslant 1/2$, one has

$$\begin{aligned}
\frac{\sum_{x \in [K]} |\alpha_{i|k}(x) - \delta_{i|k}(x)|}{\sum_{x \in [K]} \alpha_{i|k}(x)} &\leqslant 2\epsilon \sum_{j=i+1}^{k} \left(\frac{2}{\sigma_-}\right)^{k-j+1} \\
&\leqslant 2\epsilon \sum_{a=1}^{k-i} \left(\frac{2}{\sigma_-}\right)^{a} \\
&\leqslant \frac{4\epsilon}{\sigma_-} \frac{\left(\frac{2}{\sigma_-}\right)^{k-i} - 1}{\frac{2}{\sigma_-} - 1} \\
&\leqslant \frac{4\epsilon}{2 - \sigma_-} \left(\frac{2}{\sigma_-}\right)^{k-i},
\end{aligned}$$

which gives the desired result. The proof of the second case is similar and comes from the fact that

$$\begin{aligned}
&\sum_{x\in[K]} |\alpha_{i|k}(x)u(y_i|x) - \delta_{i|k}(x)v(y_i|x)| \\
&\qquad \leqslant \sum_{j=i+1}^{k} \sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1} u_{i+1}\dots u_{j-1}|p_{j-1}^j - q_{j-1}^j|v_j\dots q_{k-1}^k v_k \\
&\qquad\quad + \sum_{j=i}^{k} \sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1} u_{i+1}\dots p_{j-1}^j|u_j - v_j|q_j^{j+1}\dots q_{k-1}^k v_k.
\end{aligned}$$

■

Lemma 28 *Assume $\epsilon \leqslant \frac{1}{2K}$, then*

$$\|\nu - \omega\|_1 \leqslant 5\left(\frac{2}{\sigma_-}\right)^{k+1}\epsilon$$

and

$$\sup_{1\leqslant i\leqslant k}\sup_{x\in[K]} \|f_{i|k}(x,\cdot) - g_{i|k}(x,\cdot)\|_1 \leqslant 5\left(\frac{2}{\sigma_-}\right)^{k-i+2}\epsilon \leqslant 5\left(\frac{2}{\sigma_-}\right)^{k+1}\epsilon. \tag{14}$$

Proof With minimalist notations, one has

$$\begin{aligned}
\sum|\nu-\omega| &= \sum\left|\frac{a\alpha}{\sum b\delta} - \frac{b\delta}{\sum a\alpha}\right| \\
&\leqslant \frac{\sum|a\alpha - b\delta|}{\sum b\delta} + \sum|b\delta|\left|\frac{1}{\sum a\alpha} - \frac{1}{\sum b\delta}\right| \\
&\leqslant \frac{\sum|a\alpha - b\delta|}{\sum b\delta} + \frac{\sum|a\alpha - b\delta|}{\sum a\alpha} \\
&\leqslant \frac{2}{\sigma_-}\frac{\sum|a-b|\alpha + \sum b|\alpha-\delta|}{\sum\alpha} \\
&\leqslant \frac{2}{\sigma_-}\left(\epsilon + 4\left(\frac{2}{\sigma_-}\right)^k\epsilon\right),
\end{aligned}$$

using that $\sigma_- \leqslant a \leqslant b \leqslant 1$.

Likewise, for all $i \in \{1, \dots, k\}$ and $x \in [K]$,

$$\begin{aligned}\sum_{x' \in [K]} |g_{i|k} - f_{i|k}|(x, x') =& \sum \left| \frac{\alpha p u}{\sum \delta q v} - \frac{\delta q v}{\sum \alpha p u} \right| \\ \leqslant& \frac{\sum |\alpha p u - \delta q v|}{\sum \delta q v} + \sum |\delta q v| \left| \frac{1}{\sum \alpha p u} - \frac{1}{\sum \delta q v} \right| \\ \leqslant& 2 \frac{\sum |\alpha p u - \delta q v|}{\sum \alpha p u} \\ \leqslant& 2 \frac{\sum |\alpha u - \delta v| q + \sum \alpha u |p - q|}{\sum \alpha p u} \\ \leqslant& \frac{2}{\sigma_-} \left(4 \left(\frac{2}{\sigma_-} \right)^{k-i+1} \epsilon + \epsilon \right).\end{aligned}$$

■

Define $\eta = 5(\frac{2}{\sigma_-})^{k+1}\epsilon$. Equation (14) implies that as soon as $\eta \leqslant 1 - K\sigma_-$ (and in particular $\eta \leqslant 1/2$ since we assume $K \leqslant \frac{1}{2\sigma_-}$), it is possible to enlarge the bracket $[f_{i|k}, g_{i|k}]$ into a bracket $[f'_{i|k}, g'_{i|k}]$ of size smaller than 3η for the norm of Lemma 28 such that $f'_{i|k}/(1-\eta)$ and $g'_{i|k}/(1+\eta)$ are transition matrices.

For instance, one can take any f' and g' such that $\sigma_- \mathbf{1}\mathbf{1}^\top \leqslant f' \leqslant f \leqslant g \leqslant g' \leqslant \mathbf{1}\mathbf{1}^\top$ coefficient-wise and such that $f'\mathbf{1} = (1-\eta)\mathbf{1}$ and $g'\mathbf{1} = (1+\eta)\mathbf{1}$ (where $\mathbf{1}$ is a vector of size K whose coefficients are all equal to 1). One can construct such a matrix f' (resp. g') by taking a suitable barycenter of the lines of $\sigma_- \mathbf{1}\mathbf{1}^\top$ and f (resp. $\mathbf{1}\mathbf{1}^\top$ and g) for the lines of f' (resp. g'). The only condition is $K\sigma_- \leqslant 1 - \eta \leqslant \max_x (f\mathbf{1})_x \leqslant \max_x (g\mathbf{1})_x \leqslant 1 + \eta \leqslant K$, which is true when $\eta \leqslant 1 - K\sigma_-$.

Let

$$\begin{cases} A_x(y_0^k) = \left(\nu f'_{1|k-1} \dots f'_{k-1|k-1} p \right)_x u(y_k|x) \\ \\ B_x(y_0^k) = \left(\omega g'_{1|k-1} \dots g'_{k-1|k-1} q \right)_x v(y_k|x) \end{cases}.$$

$[A, B]$ is a bracket of $\mathcal{G}$, and this construction gives a bracket covering of $\mathcal{G}$.

Lemma 29 *Assume $\epsilon \leqslant \frac{1}{2K} \wedge \frac{1}{10k}(\frac{\sigma_-}{2})^{k+1}$. Then for all y_0^k,*

$$\sum_{x \in [K]} |(\nu f'_{1|k} \dots f'_{k|k})_x - (\omega g'_{1|k} \dots g'_{k|k})_x| \leqslant 7k\eta = 35k \left(\frac{2}{\sigma_-} \right)^{k+1} \epsilon$$

and

$$\sum_{x \in [K]} |(\nu f'_{1|k} \dots f'_{k|k} p)_x - (\omega g'_{1|k} \dots g'_{k|k} q)_x| \leqslant 53k \left(\frac{2}{\sigma_-} \right)^{k+1} \epsilon.$$

Proof Note that

$$\sum_{x\in[K]} |(\nu f'_{1|k}\dots f'_{k|k})_x - (\omega g'_{1|k}\dots g'_{k|k})_x| \leqslant \sum_{x\in[K]} |((\nu-\omega) f'_{1|k}\dots f'_{k|k})_x| + \sum_{j=1}^{k}\sum_{x\in[K]} |(\omega g'_{1|k}\dots g'_{j-1|k}(g'_{j|k}-f'_{j|k})f'_{j+1|k}\dots f'_{k|k})_x|.$$

Then, we use that $f'_{i|k}/(1-\eta)$ and $g'_{i|k}/(1+\eta)$ are transition matrices (and thus are 1-Lipschitz linear operators of $\mathbf{L}^1([K])$):

$$\|\nu f'_{1|k}\dots f'_{k|k} - \omega g'_{1|k}\dots g'_{k|k}\|_1 \leqslant \|\omega-\nu\|_1(1-\eta)^k + \sum_{j=1}^{k}\|\omega\|_1(1+\eta)^{j-1}\sup_{1\leqslant i\leqslant k}\sup_{x\in[K]}\|f'_{i|k}(x,\cdot)-g'_{i|k}(x,\cdot)\|_1(1-\eta)^{k-j},$$

so that using Lemma 28:

$$\begin{aligned}\|\nu f'_{1|k}\dots f'_{k|k} - \omega g'_{1|k}\dots g'_{k|k}\|_1 &\leqslant \eta + \|\omega\|_1\sum_{j=1}^{k}(1+\eta)^{j-1}3\eta\\ &\leqslant \eta\left(1+3(1+\eta)\sum_{j=0}^{k-1}(1+\eta)^j\right)\\ &\leqslant \eta\left(1+3(1+\eta)\frac{(1+\eta)^k-1}{\eta}\right)\\ &\leqslant \eta + 3(1+\eta)(e^{k\eta}-1).\end{aligned}$$

One can check that for all $x\in[0,\frac{1}{2}]$, $3(1+x)(e^x-1)\leqslant 6x$. Replacing x by $k\eta$, one gets that for all $\eta\leqslant\frac{1}{2k}$,

$$\|\nu f'_{1|k}\dots f'_{k|k} - \omega g'_{1|k}\dots g'_{k|k}\|_1 \leqslant \eta + 6k\eta \leqslant 7k\eta.$$

For the second part, note that

$$\begin{aligned}&\sum_{x\in[K]} |(\nu f'_{1|k}\dots f'_{k|k}p)_x - (\omega g'_{1|k}\dots g'_{k|k}q)_x|\\ &\quad\leqslant \sum_x\sum_{x'} |(\nu f'_{1|k}\dots f'_{k|k})_{x'}p_{x',x} - (\omega g'_{1|k}\dots g'_{k|k})_{x'}q_{x',x}|\\ &\quad\leqslant \sum_x\sum_{x'} |(\nu f'_{1|k}\dots f'_{k|k})_{x'} - (\omega g'_{1|k}\dots g'_{k|k})_{x'}|q_{x',x}\\ &\qquad+\sum_x\sum_{x'}(\nu f'_{1|k}\dots f'_{k|k})_{x'}|p_{x',x}-q_{x',x}|.\end{aligned}$$

Since the brackets are not empty, one has $\sum_x q_{x',x}\leqslant 1+K\epsilon$ for all x' and $\sum_{x'}(\nu f'_{1|k}\dots f'_{k|k})_{x'}\leqslant 1$ (since $\nu f'_{1|k}\dots f'_{k|k}$ is the lower bound of a non empty bracket of $\{p_{X_k|Y_1^k,\theta}\,|\,\theta\in S_{K,M}\}$), so

that

$$\begin{aligned}
&\sum_{x\in[K]} |(\nu f'_{1|k}\dots f'_{k|k}p)_x - (\omega g'_{1|k}\dots g'_{k|k}q)_x| \\
&\leqslant (1+K\epsilon)\sum_{x'} |(\nu f'_{1|k}\dots f'_{k|k})_{x'} - (\omega g'_{1|k}\dots g'_{k|k})_{x'}| + K\epsilon\sum_{x'}(\nu f'_{1|k}\dots f'_{k|k})_{x'} \\
&\leqslant (1+K\epsilon)35k\left(\frac{2}{\sigma_-}\right)^{k+1}\epsilon + K\epsilon.
\end{aligned}$$

Finally, we use that since $\epsilon \leqslant \frac{1}{2K}$, one has $(1+K\epsilon)35 \leqslant \frac{105}{2}$ and since $K\sigma_- \leqslant 1$, one has $K \leqslant \frac{1}{2}(\frac{2}{\sigma_-})^{k+1}$. ■

Lemma 30 *Assume* $\epsilon \leqslant \frac{1}{2K} \wedge \frac{1}{10k}(\frac{\sigma_-}{2})^k$. *Then*

$$d_{\mathcal{G}}(A,B) \leqslant 106k\left(\frac{2}{\sigma_-}\right)^k \epsilon.$$

Proof By definition,

$$d_{\mathcal{G}}(A,B) = \mathbb{E}^*_{Y_0^{k-1}} \sum_{x\in[K]} \int |A_x(Y_0^k) - B_x(Y_0^k)|\lambda(dY_k).$$

Taking some fixed Y_0^{k-1}, one has

$$\begin{aligned}
&\sum_x \int |A_x(y_k) - B_x(y_k)|\lambda(dy_k) \\
&= \sum_x \int |u(y_k|x)(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x - v(y_k|x)(\omega g'_{1|k-1}\dots g'_{k-1|k-1}q)_x|\lambda(dy_k) \\
&\leqslant \sum_x \int |u(y_k|x) - v(y_k|x)|(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x\lambda(dy_k) \\
&\quad + \sum_x \int v(y_k|x)|(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x - (\omega g'_{1|k-1}\dots g'_{k-1|k-1}q)_x|\lambda(dy_k).
\end{aligned}$$

Since we assumed the brackets to be non empty, one has $\int v(y|x)\lambda(dy) \leqslant 1+\|v-u\|_\infty = 1+\epsilon e^{-D}$ and $\sum_x(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x \leqslant 1$ (it is the lower bound of a non empty bracket of $\{p_{X_k|Y_0^{k-1},\theta}\,|\,\theta\in S_{K,M}\}$). Therefore, one gets with Lemma 29 that

$$\begin{aligned}
d_{\mathcal{G}}(A,B) \leqslant& \epsilon e^{-D}\sum_x(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x \\
&+ (1+\epsilon e^{-D})\sum_x |(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x - (\omega g'_{1|k-1}\dots g'_{k-1|k-1}q)_x| \\
\leqslant& \epsilon e^{-D} + (1+\epsilon e^{-D})53(k-1)\left(\frac{2}{\sigma_-}\right)^k\epsilon.
\end{aligned}$$

Finally, notice that $\epsilon e^{-D} \leqslant 1$ and $1 \leqslant 53(\frac{2}{\sigma_-})^k$ to conclude. ■

Lemma 29 implies that $\sup_x |(\nu f'_{1|k} \dots f'_{k|k} p)_x - (\omega g'_{1|k} \dots g'_{k|k} q)_x| \leqslant \eta' := 53(k-1)(\frac{2}{\sigma_-})^k \epsilon$. Therefore, since the bracket $[A, B]$ is not empty, one gets by using the assumption on u and v that

$$(\sigma_- - \eta')e^{-D}(1 - K\epsilon) \leqslant \sum_{x\in[K]} A_x \leqslant \sum_{x\in[K]} B_x \leqslant (1 + \eta')(e^D + K\epsilon e^{-D}),$$

from which we deduce that the desired inequality $\sigma_- e^{-D}/2 \leqslant \sum_{x\in[K]} A_x \leqslant \sum_{x\in[K]} B_x \leqslant 2e^D$ holds as soon as $\eta' \leqslant \frac{\sigma_-}{4}$ and $\epsilon \leqslant \frac{1}{4K}$, i.e.

$$\epsilon \leqslant \frac{\sigma_-}{4(53(k-1)(\frac{2}{\sigma_-})^k)} \wedge \frac{1}{4K},$$

which is implied by $\epsilon \leqslant \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}$ since $K \leqslant \frac{1}{\sigma_-}$. This concludes the proof of Lemma 26.

B.2.3 Control of the bracketing entropy of the simple sets and synthesis

Lemma 31 *Let $\delta > 0$, then*

$$N\left(\{\pi_\theta\}_{\theta\in S_{K,M}}, d_\infty, \delta\right) \leqslant \max\left(\frac{K-1}{\delta}, 1\right)^{K-1},$$

$$N\left(\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}, d_\infty, \delta\right) \leqslant \max\left(\frac{K-1}{\delta}, 1\right)^{K(K-1)},$$

Let $C_{\text{aux}}{}' = C_{\text{aux}} e^D \vee (K-1)$, then by **[Aentropy]**,

$$N\left(\{\gamma_\theta\}_{\theta\in S_{K,M}}, d_\infty, \delta e^{-D}\right) \leqslant \max\left(\frac{C_{\text{aux}}{}'}{\delta}, 1\right)^{m_M K}.$$

Then, Lemma 26 ensures that for all $\epsilon \leqslant \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}$,

$$\log \bar{N}\left(\mathcal{G}, d_{\mathcal{G}}, \epsilon\right) \leqslant (m_M K + K^2 - 1) \log \max\left(\left(\frac{2}{\sigma_-}\right)^k \frac{106 k e^D C_{\text{aux}}{}'}{\epsilon}, 1\right),$$

so that using Equation (13) and letting $H(u) = H(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, u)$, one has for all $\epsilon \leqslant \frac{4(D+\log\frac{1}{\sigma_-})e^{2D}}{\sigma_-}\sqrt{\frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}}$,

$$\begin{aligned} H(\epsilon) &\leqslant (m_M K + K^2 - 1) \log \max\left(\left(\frac{4(D + \log\frac{1}{\sigma_-})e^{2D}}{\sigma_-}\right)^2 \left(\frac{2}{\sigma_-}\right)^k \frac{106 k e^D C_{\text{aux}}{}'}{\epsilon^2}, 1\right) \\ &\leqslant 2(m_M K + K^2 - 1) \log \max\left(\left(D + \log\frac{1}{\sigma_-}\right)\left(\frac{2}{\sigma_-}\right)^{k/2+1} \frac{21 e^{5D/2}\sqrt{k C_{\text{aux}}{}'}}{\epsilon}, 1\right). \end{aligned}$$

Then, since $2/\sigma_- \geqslant 1$, one gets that for all $\epsilon > 0$,

$$H(\epsilon) \leqslant 2(m_M K + K^2 - 1) \log \max \left(\left(D + \log \frac{1}{\sigma_-} \right) \left(\frac{2}{\sigma_-} \right)^{k+1/2} \frac{21 e^{5D/2} \sqrt{k C_{\text{aux}}{}'}}{\epsilon}, \right.$$
$$\left. 109 \left(\frac{2}{\sigma_-} \right)^{k+1/2} k e^{D/2} \sqrt{C_{\text{aux}}{}'} \right).$$

B.3 Choice of parameters

The goal of this section is to find a function φ and a constant C for which equation (10) holds, and to choose the weights $x_{K,M}$ of Lemma 24.

Lemma 32 *Let $A, B, C \in \mathbb{R}_+^*$, $H : x \in \mathbb{R}_+^* \mapsto A \log \max(\frac{B}{x}, C)$, and $\varphi(x) : x \in \mathbb{R}_+^* \mapsto x\sqrt{\pi A}(1 + \sqrt{\log \max(\frac{B}{x}, C)})$. Then:*

$$\begin{cases} x^2 H(x) \leqslant \varphi(x)^2, \\ \displaystyle\int_0^x \sqrt{H(u)} du \leqslant \varphi(x). \end{cases}$$

Let

$$\varphi(u) = u\sqrt{2\pi(m_M K + K^2 - 1)} \Bigg(1 + \Bigg\{ \log \max \left(\left(D + \log \frac{1}{\sigma_-} \right) \left(\frac{2}{\sigma_-} \right)^{k+1/2} \frac{21 e^{5D/2} \sqrt{k C_{\text{aux}}{}'}}{u}, \right.$$
$$\left. 109 \left(\frac{2}{\sigma_-} \right)^{k+1/2} k e^{D/2} \sqrt{C_{\text{aux}}{}'} \right) \Bigg\}^{1/2} \Bigg).$$

The function $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing, so $x \mapsto \frac{\varphi(x)}{x^2}$ is decreasing and one can define $\sigma_{K,M}$ as the unique solution of the equation $(1 + \sqrt{2D + \log \frac{1}{\sigma_-}} \log n)\varphi(x) = \sqrt{n} x^2$ with unknown x, when a solution exists. By definition of E, one has

$$\begin{aligned} \forall \sigma \geqslant \sigma_{K,M}, \quad E &\leqslant \varphi(\sigma)\sqrt{n} + \left(2D + \log \frac{1}{\sigma_-} \right) (\log n)^2 \frac{\varphi(\sigma)^2}{\sigma^2} \\ &\leqslant \left(1 + \frac{(2D + \log \frac{1}{\sigma_-})(\log n)^2}{1 + \sqrt{2D + \log \frac{1}{\sigma_-}} \log n} \right) \varphi(\sigma)\sqrt{n} \\ &\leqslant \left(1 + \sqrt{2D + \log \frac{1}{\sigma_-}} \log n \right) \varphi(\sigma)\sqrt{n}. \end{aligned}$$

We define $D' := (2D + \log \frac{1}{\sigma_-})(\log n)^2$ in order to lighten the notations. Using equation (11), one gets that for all $z > 0$ and $x_{K,M} \geqslant \sigma_{K,M}$, with probability larger than $1 - e^{-z}$,

$$\begin{aligned} W_{K,M} \leqslant & 4C^*(n_* + k + 1) \left[(1 + \sqrt{D'}) \frac{\varphi(x_{K,M})}{x_{K,M}^2 \sqrt{n}} + \sqrt{\frac{z}{x_{K,M}^2 n}} + D' \frac{z}{x_{K,M}^2 n} \right] \\ \leqslant & 4C^*(n_* + k + 1) \left[\frac{\sigma_{K,M}}{x_{K,M}} + \sqrt{\frac{z}{x_{K,M}^2 n}} + D' \frac{z}{x_{K,M}^2 n} \right]. \end{aligned}$$

Let $\epsilon > 0$, and let us take

$$x_{K,M} = \frac{1}{\theta} \left(\sigma_{K,M} + \sqrt{\frac{z}{n}} \right),$$

where $\theta > 0$ is such that $2\theta + D'\theta^2 \leqslant \frac{\epsilon}{4C^*(n_* + k + 1)}$. Then

$$\begin{aligned} W_{K,M} \leqslant & 4C^*(n_* + k + 1) \left[\theta + \theta + D'\theta^2 \right] \\ \leqslant & \epsilon \end{aligned}$$

and

$$\begin{aligned} W_{K,M} x_{K,M}^2 \leqslant & 4C^*(n_* + k + 1) \left[\sigma_{K,M} x_{K,M} + \sqrt{\frac{z}{n}} x_{K,M} + D' \frac{z}{n} \right] \\ \leqslant & 4C^*(n_* + k + 1) \left[\theta x_{K,M}^2 + D' \frac{z}{n} \right] \\ \leqslant & 8C^*(n_* + k + 1) \left[\frac{1}{\theta} \sigma_{K,M}^2 + (D' + \frac{1}{\theta}) \frac{t}{n} \right]. \end{aligned}$$

Take $z = s + w_M + K$, then since $\sum_M e^{-w_M} \leqslant e - 1$, one gets that with probability larger than $1 - e^{-s}$, for all M, K and for all functions pen such that

$$\mathrm{pen}_n(K, M) \geqslant 8C^*(n_* + k + 1) \left[\frac{1}{\theta} \sigma_{K,M}^2 + (D' + \frac{1}{\theta}) \frac{w_M + K}{n} \right],$$

one has

$$W_{K,M} x_{K,M}^2 - \mathrm{pen}_n(K, M) \leqslant 8C^*(n_* + k + 1)(D' + \frac{1}{\theta}) \frac{s}{n}.$$

A possible choice of θ is

$$\theta = \frac{1}{D'} \left(\sqrt{1 + \frac{\epsilon D'}{4C^*(n_* + k + 1)}} - 1 \right).$$

Using that $\frac{1}{\sqrt{1+x}-1} \leqslant \max(1, \frac{3}{x})$ for all $x > 0$, one gets that there exists θ such that $2\theta + D'\theta^2 \leqslant \frac{\epsilon}{4C^*(n_* + k + 1)}$ and

$$\frac{1}{\theta} \leqslant 3C^*(n_* + k + 1) \max \left(\frac{D'}{12C^*(n_* + k + 1)}, \frac{1}{\epsilon} \right).$$

Therefore,

$$W_{K,M}x_{K,M}^2 - \mathrm{pen}_n(K,M) \leqslant 24(C^*)^2(n_*+k+1)^2\left(D' + \frac{1}{\epsilon} \vee \frac{D'}{12C^*(n_*+k+1)}\right)\frac{s}{n}$$

as soon as

$$\mathrm{pen}_n(K,M) \geqslant 24(C^*)^2(n_*+k+1)^2\Bigg[\left(\frac{1}{\epsilon} \vee \frac{D'}{12C^*(n_*+k+1)}\right)\sigma_{K,M}^2 + \left(D' + \frac{1}{\epsilon} \vee \frac{D'}{12C^*(n_*+k+1)}\right)\frac{w_M+K}{n}\Bigg].$$

The last step of the proof is to find an upper bound of $\sigma_{K,M}$.

Lemma 33 *Let A, B, C and E be functions $\mathbb{N} \longrightarrow [1,\infty)$, and $\varphi : x \longmapsto xA(1+\sqrt{\log\max(\frac{B}{x}, C)})$. Let σ be the only solution of the equation $\frac{\varphi(x)}{x^2\sqrt{n}} = \frac{1}{E}$ with unknown $x \in \mathbb{R}_+^*$. Let*

$$f(n) = \left[\frac{A(n)C(n)E(n)}{B(n)}(1+\sqrt{\log B(n) + \log n})\right]^2.$$

Assume that there exists n_1 such that for all $n \geqslant n_1$, $f(n) \leqslant n$. Then

$$\forall n \geqslant n_1, \qquad \sigma \leqslant \frac{A(n)E(n)}{\sqrt{n}}(1+\sqrt{\log B(n) + \log n}).$$

In our case,

$$\begin{cases} A = \sqrt{2\pi(m_M K + K^2 - 1)}, \\ B = \left(D + \log\dfrac{1}{\sigma_-}\right)\left(\dfrac{2}{\sigma_-}\right)^{k+1/2} 21e^{5D/2}\sqrt{k{C_{\mathrm{aux}}}'}, \\ C = 109\left(\dfrac{2}{\sigma_-}\right)^{k+1/2} ke^{D/2}\sqrt{{C_{\mathrm{aux}}}'}, \\ E = 1+\sqrt{D'} \leqslant 2\sqrt{D'} \end{cases}.$$

Hence

$$f(n) \leqslant 862.2\pi\,(m_M K + K^2 - 1)\frac{k}{D + \log\frac{1}{\sigma_-}}e^{-4D}(\log n)^2 \left(2\log\left(D + \log\frac{1}{\sigma_-}\right) + (2k+1)\log\frac{2}{\sigma_-} + 2\log 21 + 5D + \log k + \log {C_{\mathrm{aux}}}' + 2\log n\right).$$

By using that $1 \leqslant k \leqslant n$, that $\log(D + \log\frac{1}{\sigma_-}) \leqslant D + \log\frac{1}{\sigma_-}$, that $\log {C_{\mathrm{aux}}}' \leqslant \log C_{\mathrm{aux}} + D + \log n$, that $\frac{1}{\sigma_-} \geqslant 2K \geqslant 4$ and by assuming $n \geqslant 3$ and $k \geqslant 2$, one gets:

$$f(n) \leqslant \tilde{f}_{K,M}(n) := 6900\pi\,(m_M K + K^2 - 1)ke^{-4D}(\log n)^3\,(k + \log C_{\mathrm{aux}}).$$

Now, assume that there exists n_1 such that $\tilde{f}_{K,M}(n) \leqslant n$ for all $n \geqslant n_1$, then

$$\forall n \geqslant n_1, \qquad \sigma^2 \leqslant \frac{8\pi\,(m_M K + K^2 - 1)D'}{n}\Bigg(2 + 2\log 21 + 3\log n \\ + 2\log\left(D + \log\frac{1}{\sigma_-}\right) + (2k+1)\log\frac{2}{\sigma_-} + 6D + \log k + \log C_{\mathrm{aux}}\Bigg),$$

so that

$$\forall n \geqslant n_1, \quad \sigma^2 \leqslant \frac{64\pi\,(m_M K + K^2 - 1)D'}{n}\Big(\log n + k\log\frac{2}{\sigma_-} + D + \log C_{\mathrm{aux}}\Big).$$

Therefore, there exists a numerical constant C_{pen} such that the condition on the penalty is implied by

$$\mathrm{pen}_n(K,M) \geqslant \frac{C_{\mathrm{pen}}}{n}(n_* + k + 1)^2\Bigg[D'\left(\frac{1}{\epsilon} \vee \frac{D'}{C^*(n_* + k + 1)}\right)(m_M K + K^2 - 1) \\ \Big(\log n + k\log\frac{2}{\sigma_-} + D + \log C_{\mathrm{aux}}\Big) + \left(D' + \frac{1}{\epsilon} \vee \frac{D'}{C^*(n_* + k + 1)}\right) w_M\Bigg].$$