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@univ-cotedazur.fr
Laboratoire J. A. Dieudonné
Université Côte d'Azur, CNRS
06108, Nice, France

Abstract

Finite state space hidden Markov models are flexible tools to model phenomena with complex time dependencies: any process distribution can be approximated by a hidden Markov model with enough hidden states. We consider the problem of estimating an unknown process distribution using nonparametric hidden Markov models in the *misspecified setting*, that is when the data-generating process may not be 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, nonparametric statistics, maximum likelihood estimator, model selection, oracle inequality, hidden Markov model

Contents

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 (2018) and least squares estimators in de Castro et al. (2016); Lehéricy (2018). 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, or equivalently a finite moment. 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 issue. 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:

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

where the supremum is taken over all emission densities γ in the models and for some constant $C_{\mathbf{Q}} > 0$. For instance, this assumption holds when all densities are upper bounded by $e^{C_{\mathbf{Q}} \log n}$. A key remark at this point is the dependency of the exponent with n: we allow the models to depend on the sample size. Typically, taking a larger sample makes it possible to consider larger models.

To stabilize the log-likelihood, we modify the models in the following way. First, only keep HMMs whose transition matrix have entries that are neither too small nor too large: when the HMM has K hidden states, the entries of the transition matrix should belong to the interval $[K/(C_\gamma \log n), KC_\gamma \log n]$ for some constant $C_\gamma > 0$. 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 6) is the following oracle inequality: there exist 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 $1 - e^{-t} - n^{-2}$:

$$\mathbf{K}(\hat{\theta}_n) \leqslant (1+\eta) \inf_{\dim(S) \leqslant n} \left\{ \inf_{\theta \in S} \mathbf{K}(\theta) + 2\mathrm{pen}_n(S) \right\} + \frac{A}{\eta} t \frac{(\log n)^{10}}{n},$$

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 10. 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 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. 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 π and $\mathbf{Q}$ are the initial distribution and transition matrix of a Markov chain on $[K]$ and for all $x\in[K]$, γ_x is a convex combination of M elements of $\{f_\xi\}_{\xi\in\Xi}$.

$\mathbf{L}^2$ 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 π and $\mathbf{Q}$ are the initial distribution and transition matrix of a Markov chain on $[K]$ and for all $x\in[K]$, γ_x is a probability density such that $\gamma_x = g\vee 0$ for a function $g\in E_M$.

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

$$(K,\hat\pi_{K,M,n},\hat{\mathbf{Q}}_{K,M,n},\hat\gamma_{K,M,n})\in\operatorname*{argmax}_{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M,n}}\frac{1}{n}l_n(K,\pi,\mathbf{Q},\gamma).$$

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:

$$(\hat K_n,\hat M_n)\in\operatorname*{argmax}_{K\in\mathcal{K}_n,M\in\mathcal{M}_n}\left(\frac{1}{n}l_n(K,\hat\pi_{K,M,n},\hat{\mathbf{Q}}_{K,M,n},\hat\gamma_{K,M,n})-\operatorname{pen}_n(K,M)\right)$$

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 $(Y_t)_{t\geqslant1}$ is stationary, so that one can extend it into a process $(Y_t)_{t\in\mathbb{Z}}$.

[A⋆tail] There exists $\delta>0$ such that

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

This assumption ensures that the true log-density rarely takes extreme values (see Lemma 14).

[A⋆forget] There exist 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 $y^i_{i-(k\vee k')}\in\mathcal{Y}^{(k\vee k')+1}$,

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

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

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

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⋆mix] There exist two constants $c_* > 0$ and $n_* \in \mathbb{N}^*$ such that

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

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

Even if [A⋆forget] is analog to a ψ-mixing condition (see Bradley (2005) for a survey on mixing conditions) and is proved using the same tool [A⋆mix] 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⋆forget] but not necessarily [A⋆mix]. Conversely, there exist processes satisfying [A⋆mix] but not [A⋆forget].

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⋆forget] and [A⋆mix] 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⋆forget], see for instance Douc et al. (2004), proof of Lemma 2.

For [A⋆mix], the Doeblin condition implies that for all distributions 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$\star$mix] from this inequality. ■

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 $(S_{K,M,n})_{K\in\mathbb{N}^*,M\in\mathcal{M}}$ such that for each K and M, 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.

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

[Aergodic] There exists $C_{\mathbf{Q}} \geqslant 1$ such that for all $(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n$,

$$\begin{aligned} &\forall x,x' \in [K], \qquad (C_{\mathbf{Q}}\log n)^{-1} \leqslant K\mathbf{Q}(x,x') \leqslant C_{\mathbf{Q}}\log n \\ \text{and} \qquad &\forall x \in [K], \qquad (C_{\mathbf{Q}}\log n)^{-1} \leqslant K\pi(x) \leqslant C_{\mathbf{Q}}\log n. \end{aligned}$$

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

$$b_\gamma(y) = \log\left(K^{-1}\sum_x \gamma_x(y)\right).$$

When $(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n$, assumption [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 $(C_{\mathbf{Q}}\log n)^{-1}K^{-1}$, 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 $(C_{\mathbf{Q}}\log n)^{-1}K^{-1}\sum_x \gamma_x(Y_t)$. For the same reason, it is upper bounded by $C_{\mathbf{Q}}(\log n)K^{-1}\sum_x \gamma_x(Y_t)$. Thus, it is enough to bound b_γ to control $p_{(K,\pi,\mathbf{Q},\gamma)}$ without having to handle the dependency in past observations.

The following assumption ensures that the log-likelihood rarely takes extreme values.

[Atail] There exists $C_\gamma \geqslant 1$ such that

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

In practice, it is enough to check the upper deviations, as shown in the following lemma.

Lemma 2 *Assume that there exists $C \geqslant 1$ such that*

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

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 $C_\gamma = C \vee a$.

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 10. This is in agreement with the fact that this component is asymptotically never observed as soon as $a > 1$.

The following assumption means that as far as the bracketing entropy is concerned, the set of emission densities of the model $S_{K,M,n}$ behaves like a parametric model with dimension m_M.

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

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

where d_∞ is the supremum norm distance 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. The following assumption ensures that the models do not grow absurdly fast.

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

$$\sup_{K,M \text{ s.t. } K \leqslant n \text{ and } m_M \leqslant n} \log C_{\text{aux}}(M, K, 3C_\gamma (\log n)^2, n) \leqslant n^\zeta.$$

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. Baring models $\mathbf{S}_n^{(\gamma)}$ that grow so fast with respect to n that [Aentropy] becomes essentially meaningless, [Agrowth] is usually immediately checked once [Aentropy] is established. An example of this approach is given in Section 3.2 for mixtures of exponential power distributions.

2.5 Limit and properties of the normalized log-likelihood

In this section, we focus on the convergence of the normalized log-likelihood.

Lemma 3 (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).

Lemma 4 (Leroux (1992)) *Let K be a positive integer, γ a vector of K probability densities, $\mathbf{Q}$ 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 $\pi(x) > 0$ and $\mathbb{E}^*|\log \gamma_x(Y_1)| < +\infty$ for all $x \in [K]$.*

Then there exists a finite quantity $l(K, \mathbf{Q}, \gamma)$ which does not depend on π such that

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

In particular, $l(K, \mathbf{Q}, \gamma) = \lim_n \mathbb{E}[\frac{1}{n} l_n(K, \pi, \mathbf{Q}, \gamma)]$.

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 $\mathbf{K}(K, \mathbf{Q}, \gamma) \geqslant 0$ since it is the limit of a sequence of Kullback-Leibler divergences: under the assumptions of Lemma 4,

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

where $\mathbb{P}^*_{Y_1^n}$ (respectively $\mathbb{P}_{Y_1^n|(K,\pi,\mathbf{Q},\gamma)}$) is the distribution of Y_1^n under $\mathbb{P}^*$ (respectively $\mathbb{P}_{(K,\pi,\mathbf{Q},\gamma)}$). We will see in the proofs that 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. In a way, it is a prediction error under the parameter $(K, \mathbf{Q}, \gamma)$.

In the particular case where the true distribution of $(Y_t)_t$ comes from 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 5 (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 $\mathbb{E}^*[(\log \gamma_x^*(Y_1))^+] < \infty$ for all $x \in [K^*]$ (so that $l^* < \infty$).*

Then for all $K \leqslant K^*$, for all transition matrices $\mathbf{Q}$ of size K and for all K-uples of probability densities γ, $\mathbf{K}(K, \mathbf{Q}, \gamma) = 0$ 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 6 *Assume [A⋆forget], [A⋆mix], [A⋆tail], [Aergodic], [Atail], [Aentropy] and [Agrowth] hold.*

Let $(w_M)_{M\in\mathcal{M}}$ be a nonnegative sequence such that $\sum_{M\in\mathcal{M}} e^{-w_M} \leqslant e-1$. For all K and M, let

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

$$(\hat{K}, \hat{M}) \in \underset{\substack{K\leqslant n \\ M \text{ s.t. } m_M\leqslant n}}{\operatorname{argmax}} \left(\frac{1}{n} l_n(K, \hat{\pi}_{K,M,n}, \hat{\mathbf{Q}}_{K,M,n}, \hat{\gamma}_{K,M,n}) - pen_n(K,M) \right)$$

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 exist constants A and C_{pen} depending only on $C_{\mathbf{Q}}$, C_γ, n_* and c_* and a constant n_0 depending only on $C_{\mathbf{Q}}$, C_γ, n_*, ζ, n_{growth}, C_*, ρ_*, δ and M_δ such that for all $n \geqslant n_0$, $t \geqslant 1$ and $\eta \leqslant 1$, with probability at least $1 - e^{-t} - 2n^{-2}$,

$$\mathbf{K}(\hat{K}, \hat{\mathbf{Q}}, \hat{\gamma}) \leqslant (1+\eta) \inf_{\substack{K\leqslant n \\ M \text{ s.t. } m_M\leqslant n}} \left\{ \inf_{(K,\pi,\mathbf{Q},\gamma)\in S_{K,M,n}} \mathbf{K}(K,\mathbf{Q},\gamma) + 2pen_n(K,M) \right\} + \frac{A}{\eta} t \frac{(\log n)^{10}}{n}$$

as soon as

$$pen_n(K,M) \geqslant \frac{C_{pen}}{\eta} \frac{(\log n)^{10}}{n} \Big\{ w_M + (\log n)^4 (m_M K + K^2 - 1) \times \left((\log n)^3 \log\log n + \log C_{aux}(M, K, 3C_\gamma (\log n)^2, n) \right) \Big\}.$$

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 may 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 $\log n$ come from:

- The limitation of the dependency to the $\log n$ most recent observations,
- The dependency of the bounds $C_{\mathbf{Q}} \log n$ and $C_\gamma \log n$ on n in assumptions [Aergodic] and [Atail],
- Truncating the emission log-densities (possible thanks to assumptions [Atail] and [A$\star$tail]),
- The use of 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 6 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. Without loss of generality, we identify the true hidden state space with $[K^*]$. 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.* For all $x \in [K^*]$, $\log(\gamma_x^* G_\lambda)$ is locally β-Hölder with $\beta > 0$, i.e. there exist a polynomial L and a constant $R > 0$ such that if r is the largest integer smaller than β, one has for all $x \in [K^*]$,

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

(C2) *Moments.* There exists $\epsilon > 0$ such that for all $x \in [K^*]$,

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

(C3) *Tail.* There exist positive constants c and τ such that for all $x \in [K^*]$,

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

(C4) *Monotonicity.* For all $x \in [K^*]$, $(\gamma_x^* G_\lambda)$ is positive and there exists $y_m < y_M$ such that for all $x \in [K^*]$, $(\gamma_x^* G_\lambda)$ 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 true 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 may 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

- [Aergodic] holds with $C_\sigma = 1$,
- For all $x \in [K]$, there exist $(s_{x,1}, \dots, s_{x,M}) \in [\frac{1}{n}, n]^M$, $(\mu_{x,1}, \dots, \mu_{x,M}) \in [-n, n]^M$ 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 7 (Checking the assumptions) *Assume* $\inf \mathbf{Q}^* > 0$*, then:*

- *[A⋆forget] and [A⋆mix] hold.*
- *Assume* ***(C3)****, then [A⋆tail] holds.*
- *[Atail] holds for all* $n \geqslant 3$ *by taking* $C_\gamma = 10$*.*
- *[Aentropy] and [Agrowth] hold for any* $\zeta > 0$ *by taking* $m_M = 2M$ *and* $C_{aux}(M, K, D, n) = 4pn^3$*, for instance* $\zeta = 2$ *and* $n_{growth} = 4p$*.*

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)**.

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

Remark 8 *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 9 (Approximation rates) *Assume **(C1)-(C4)** hold. Then there exists sequences of mixtures $(g_{M,x})_M$ for each $x \in [K^*]$ such that for M large enough and all $n \geqslant M$, $(n^{-2} + (1-n^{-2})g_{M,x})_{x\in[K^*]} \in S^{(\gamma)}_{K^*,M,n}$ and*

$$\max_{x\in[K^*]} KL(\gamma^*_x \| g_{M,x}) = O(M^{-2\beta}(\log M)^{2\beta\frac{p}{\tau}}).$$

Proof Proof in Section A.1.2. ■

Corollary 10 (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 $n \geqslant M$,*

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

Hence, using Theorem 6 with $pen_n(K, M) = (KM + K^2)(\log n)^{18}/n$, there exists a constant C such that almost surely, there exists a (random) n_0 such that

$$\begin{aligned}\forall n \geqslant n_0, \quad \mathbf{K}(\hat{K}_n, \hat{\mathbf{Q}}_n, \hat{\gamma}_n) &\leqslant Cn^{\frac{-2\beta}{2\beta+1}}(\log n)^{18+\frac{p}{\tau}-\frac{16+\frac{p}{\tau}}{2\beta+1}} \\ &\leqslant Cn^{\frac{-2\beta}{2\beta+1}}(\log n)^{18+\frac{p}{\tau}}.\end{aligned}$$

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, it is still unknown whether increasing the number of states improves the minimax rates of convergence. 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. (2020) 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 6)

5.1 Overview of the proof

By definition of $(\hat{K}, \hat{\pi}, \hat{\mathbf{Q}}, \hat{\gamma})$, one has for all $K \leqslant n$, 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). For all $i \in \mathbb{Z}$, $k \in \mathbb{N}^*$, let

$$L_{i,k}^* = \log p^*(Y_i | Y_{i-k}^{i-1}), \tag{3}$$

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 distributions 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 $p_{(K,\mathbf{Q},\gamma)}(\cdot|X_{i-k}\sim\mu)$ is the density of a HMM with parameters $(K,\mathbf{Q},\gamma)$ starting at time $i-k$ with the distribution μ. When μ is the stationary distribution of the Markov chain under the parameter $(K,\mathbf{Q},\gamma)$, we write $L_{i,k}(K,\mathbf{Q},\gamma)$. The following remark will be useful in our proofs: since

$$\begin{aligned}p_{(K,\pi,\mathbf{Q},\gamma)}(X_k=x|Y_1^{k-1})&=\frac{\sum_{x'\in[K]}p_{(K,\pi,\mathbf{Q},\gamma)}(X_{k-1}=x'|Y_1^{k-2})\mathbf{Q}(x',x)\gamma_{x'}(Y_{k-1})}{\sum_{x'\in[K]}p_{(K,\pi,\mathbf{Q},\gamma)}(X_{k-1}=x'|Y_1^{k-2})\gamma_{x'}(Y_{k-1})}\\&\in[(C_{\mathbf{Q}}\log n)^{-1}K^{-1},C_{\mathbf{Q}}(\log n)K^{-1}]\end{aligned}$$

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

$$|L_{i,k,\mu}(K,\mathbf{Q},\gamma)-b_\gamma(Y_i)|\leqslant\log(C_{\mathbf{Q}}\log n).\tag{4}$$

Assume from now on that $n\geqslant\exp(C_{\mathbf{Q}})$. For all $k,k'\in\mathbb{N}^*$, for all μ, μ' probability distributions and for all $(K,\pi,\mathbf{Q},\gamma),(K',\pi',\mathbf{Q}',\gamma')\in\mathbf{S}_n$,

$$\begin{cases}|L_{i,k,\mu}(K,\mathbf{Q},\gamma)-L_{i,k',\mu'}(K',\mathbf{Q}',\gamma')|\leqslant 4\log\log n+|b_\gamma(Y_i)|+|b_{\gamma'}(Y_i)|,\\|L_{i,k,\mu}(K,\mathbf{Q},\gamma)-L^*_{i,k'}|\leqslant 2\log\log n+|b_\gamma(Y_i)|+|L^*_{i,k'}|.\end{cases}\tag{5}$$

Let $k\geqslant 1$ and $D>0$. Approximate $\nu(K,\pi,\mathbf{Q},\gamma)$ by the deviation

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

where

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

for a fixed measure μ_t, for instance the uniform measure on $[K]$. Note that $\|t^{(D)}_{(K,\mathbf{Q},\gamma)}\|_\infty\leqslant 2(D+\log\log n)$ by equation (5).

Considering these functions $t^{(D)}_{(K,\mathbf{Q},\gamma)}$ has two advantages. The first one is to limit the time dependency on the past to only k observations, 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 Bernstein-like concentration inequalities apply. The error of this approximation is given by the following lemma.

Lemma 11 *Assume [Atail], [Aergodic], [A⋆tail] and [A⋆forget] hold. Then there exists n_0 depending on $C_{\mathbf{Q}}$, C_*,ρ_*,M_δ and δ such that for all $n\geqslant n_0$, for all $u\geqslant 1$, with probability greater than $1-2ne^{-u}$, for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$,*

$$\left|\nu(K,\pi,\mathbf{Q},\gamma)-\bar\nu_k(t^{(C_\gamma(\log n)u)}_{(K,\mathbf{Q},\gamma)})\right|\leqslant 10C_\gamma(\log n)ue^{-u}+\frac{2}{n\rho(1-\rho)^2}+\frac{4\rho^{k-1}}{1-\rho}$$

where $\rho = 1 - (C_{\mathbf{Q}} \log n)^{-2}$. *In particular, if* $k \geqslant C_{\mathbf{Q}}^2 (\log n)^3$ *and* $n \geqslant n_0 \vee \sqrt{30 C_\gamma}$, *for all* $D \geqslant 3C_\gamma (\log n)^2$, *with probability greater than* $1 - 2n^{-2}$,

$$\left| \nu(K, \pi, \mathbf{Q}, \gamma) - \bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)}) \right| \leqslant 13 C_{\mathbf{Q}}^4 \frac{(\log n)^4}{n}. \tag{6}$$

Proof Proof in Section 5.2.1. ■

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

Theorem 12 *Assume [Aergodic], [Aentropy] and [A⋆mix]. Also assume that* $D \geqslant \log n$, *that* $k \geqslant n_* + 1$ *and that there exists* n_1 *such that for all* $n \geqslant n_1$, *for all* $K \leqslant n$ *and* M *such that* $m_M \leqslant n$,

$$14\pi \left(m_M K + K^2 - 1\right) e^{-4D} (\log n)^2 (k + \log C_{aux}(M, K, D, n)) \leqslant n. \tag{7}$$

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 exist 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* $m_M \leqslant n$,

$$pen_n(K, M) \geqslant \frac{C_{pen}}{n} k^2 \left(\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{k} \right) \times \left(w_M + (m_M K + K^2 - 1) D (\log n)^2 (D + k \log\log n + \log C_{aux}) \right). \tag{8}$$

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

$$|\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - pen_n(K, M) \leqslant \epsilon \mathbb{E}[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})^2] + A k^2 \left(\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{k} \right) \frac{s}{n}. \tag{9}$$

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 13 *Assume [Atail], [Aergodic], [A⋆tail] and [A⋆forget] hold. There exists a constant* n_0 *depending on* M_δ, δ, ρ_*, C_* *and* $C_{\mathbf{Q}}$ *such that for all* $n \geqslant n_0$, $k \geqslant C_{\mathbf{Q}} (\log n)^3$, $D > 0$ *and* $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$,

$$\frac{1}{44 C_\gamma^2 (\log n)^4} \mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] \leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) + \frac{22}{n}.$$

Proof Proof in Section 5.2.2. ■

Take $u = 3\log n$ in order to have $ne^{-u} \leqslant n^{-2}$ in Lemma 11. Note that $u \geqslant 1$ for all $n \geqslant e$. Based on Lemma 11 and 13, also take

$$\begin{cases} D = C_\gamma(\log n)u = 3C_\gamma(\log n)^2, \\ k = C_{\mathbf{Q}}^2(\log n)^3. \end{cases}$$

In the following, we assume $n \geqslant e \vee \exp([(n_* + 1)/C_{\mathbf{Q}}^2]^{1/3})$, so that $k \geqslant n_* + 1$ and $D \geqslant \log n$. Let $\eta \leqslant 1$. 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 13, take

$$\frac{1}{\epsilon} = \frac{1}{\eta}44C_\gamma^2(\log n)^4.$$

When assumption [Agrowth] holds and $m_M \leqslant n$ and $K \leqslant n$, equation (7) is implied by

$$28\pi n^2(\log n)^2 e^{-12C_\gamma(\log n)^2}(C_{\mathbf{Q}}^2(\log n)^3\log\log n + n^\zeta) \leqslant n$$

for all $n \geqslant n_{\text{growth}}$, which is true for $n \geqslant n_1$ for a constant n_1 depending only on n_{growth}, $C_{\mathbf{Q}}$ and ζ.

Moreover, there exists a constant C_ϵ depending only on $C_{\mathbf{Q}}$ and C_γ such that for all n,

$$\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{k} \leqslant \frac{C_\epsilon}{\eta}(\log n)^4.$$

Thus, there exists an integer n_0'' depending on $C_{\mathbf{Q}}$ and C_γ (for instance $\exp(3C_\gamma/C_{\mathbf{Q}}^2)$) such that for all $n \geqslant n_0''$ equation (8) is implied by

$$\begin{aligned} \text{pen}_n(K,M) \geqslant \frac{C_{\text{pen}}}{n}C_{\mathbf{Q}}^4(\log n)^6\frac{C_\epsilon}{\eta}(\log n)^4 \\ \times \Bigg[w_M + 6C_\gamma(\log n)^4(m_M K + K^2 - 1) \\ \times \left(C_{\mathbf{Q}}^2(\log n)^3\log\log n + \log C_{\text{aux}}(M,K,3C_\gamma(\log n)^2,n)\right)\Bigg], \end{aligned}$$

so if in addition n is larger than the thresholds of Theorem 12 and Lemma 13, equation (9) and Lemma 13 imply for all $s > 0$, with probability at least $1 - e^{-s}$, for all $K \leqslant n$ and M such that $m_M \leqslant n$ and all $(K,\pi,\mathbf{Q},\gamma) \in S_{K,M,n}$,

$$\begin{aligned} |\bar\nu_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - \text{pen}_n(K,M) \\ &\leqslant \eta\mathbf{K}(K,\mathbf{Q},\gamma) + \frac{22\eta}{n} + AC_{\mathbf{Q}}^4(\log n)^6\frac{C_\epsilon}{\eta}(\log n)^4\frac{s}{n} \\ &\leqslant \eta\mathbf{K}(K,\mathbf{Q},\gamma) + 23AC_{\mathbf{Q}}^4(\log n)^6\frac{C_\epsilon}{\eta}(\log n)^4\frac{s}{n} \end{aligned} \tag{10}$$

since we may assume $A \geqslant 1$ without loss of generality. Therefore, putting together equations (6) and (10) shows

$$\begin{aligned}|\nu(K,\pi,\mathbf{Q},\gamma)|-\mathrm{pen}_n(K,M) & \\ &\leqslant \eta\mathbf{K}(K,\mathbf{Q},\gamma) + \frac{23AC_{\mathbf{Q}}^4C_\epsilon}{\eta}s\frac{(\log n)^{10}}{n} + 13C_{\mathbf{Q}}^4\frac{(\log n)^4}{n} \\ &\leqslant \eta\mathbf{K}(K,\mathbf{Q},\gamma) + \frac{36AC_{\mathbf{Q}}^4C_\epsilon}{\eta}s\frac{(\log n)^{10}}{n}\end{aligned}$$

which is equation (2) with the appropriate residual terms for Theorem 6.

5.2 Proofs

Let us first state two lemmas that will be of use in subsequent proofs.

Lemma 14 *Assume [Atail] and [A⋆tail]. Then there exists a constant n_0 depending on δ and M_δ such that for all $n \geqslant n_0$, for all i,k, and for all $u \geqslant 1$,*

$$\mathbb{P}^*[|L_{i,k}^*| \geqslant C_\gamma(\log n)u] \leqslant e^{-u}$$

where $L_{i,k}^* = \log p^*(Y_i|Y_{i-k}^{i-1})$ as defined in (3), and writing $D = C_\gamma(\log n)u$,

$$\mathbb{E}\left[\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|\,\mathbf{1}_{\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|\geqslant D}\right] \vee \mathbb{E}\left[|L_{i,k}^*|\mathbf{1}_{|L_{i,k}^*|\geqslant D}\right] \leqslant 2De^{-u},$$

$$\mathbb{E}\left[\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|^2\,\mathbf{1}_{\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|\geqslant D}\right] \vee \mathbb{E}\left[|L_{i,k}^*|^2\mathbf{1}_{|L_{i,k}^*|\geqslant D}\right] \leqslant 5D^2e^{-u}.$$

Proof Let $i \in \mathbb{Z}$, $k \in \mathbb{N}$ and $v > 0$. By [A⋆tail] and Markov's inequality,

$$\begin{aligned}\mathbb{P}^*\left[L_{i,k}^* \geqslant v\right] &= \mathbb{P}^*\left[p^*(Y_i|Y_{i-k}^{i-1}) \geqslant e^v\right] \\ &\leqslant e^{-\delta v}\mathbb{E}^*\left[(p^*(Y_i|Y_{i-k}^{i-1}))^\delta\right] \\ &\leqslant e^{\log M_\delta - \delta v}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{P}^*\left[L_{i,k}^* \leqslant -v\right] &= \mathbb{P}^*\left[p^*(Y_i|Y_{i-k}^{i-1}) \leqslant e^{-v}\right] \\ &= \mathbb{E}^*\left[\int \mathbf{1}_{p^*(y|Y_{i-k}^{i-1})\leqslant e^{-v}}p^*(y|Y_{i-k}^{i-1})\lambda(dy)\right] \\ &\leqslant e^{-v}.\end{aligned}$$

Thus, there exists $B^* \geqslant 1$ such that if $u \geqslant 1$, $\mathbb{P}^*[|L_{i,k}^*| \geqslant B^*u] \leqslant e^{-u}$. Therefore, for all $n \geqslant \exp(B^*)$, the first equation holds, and under [Atail], the variables $|L_{i,k}^*|$ and $\sup_{\gamma\in\mathbf{S}_n^{(\gamma)}}|b_\gamma(Y_1)|$ are dominated by $C_\gamma(\log n)(W \vee 1)$ where W is an exponential random variable with parameter 1. To conclude, note that for all $u > 0$,

$$\begin{aligned}\mathbb{E}^*[W\mathbf{1}_{W\geqslant u}] &\leqslant (1+u)e^{-u} \leqslant 2ue^{-u}, \\ \mathbb{E}^*[W^2\mathbf{1}_{W\geqslant u}] &\leqslant (u^2+2u+2)e^{-u} \leqslant 5u^2e^{-u}.\end{aligned}$$

■

Lemma 15 *Assume [Aergodic] and [A⋆forget].*

1. *Let $\rho = 1 - (C_{\mathbf{Q}} \log n)^{-2}$. 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, 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. *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 and k,*

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

3. *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 $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$. Assume [A⋆mix], [Atail] and [A⋆tail], then they are also ergodic, integrable and*

$$l(K,\mathbf{Q},\gamma) = \mathbb{E}^*[L_{1,\infty}(K,\mathbf{Q},\gamma)] \quad \textit{and} \quad l^* = \mathbb{E}^*[L^*_{1,\infty}].$$

Proof The first point is a result from Douc et al. (2004).

The second point follows directly from [A⋆forget].

The third point follows from the ergodicity of $(Y_t)_{t\geqslant 1}$ under [A⋆mix], from the integrability of $L_{i,\infty}$ and $L^*_{i,\infty}$ under [Atail] and [A⋆tail] by Lemma 14 and from Lemmas 3 and 4 for the definition of l and l^*. ■

5.2.1 Proof of Lemma 11

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

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

Thus, by Lemma 15,

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

as soon as $\rho_* \leqslant \rho$ and $C_* \leqslant 1/(1-\rho)$, which holds when $C_{\mathbf{Q}} \log n \geqslant (C_* \vee (1-\rho_*)^{-1})^{1/2}$, in particular when $\log n \geqslant (C_* \vee (1-\rho_*)^{-1})^{1/2}$.

Let $u \geqslant 1$ and $D = C_\gamma (\log n) u$ and assume that $n \geqslant n_0$ from Lemma 14. Then

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

We restrict ourselves to the event $\bigcap_{i=1}^n\{|L^*_{i,k}| \vee (\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_i)|) \leqslant D\}$, which occurs with probability greater than $1-2ne^{-u}$ using assumption [Atail] and Lemma 14. On this event,

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

Moreover,

$$\begin{aligned}|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)]| \\ = \mathbb{E}^*[|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)|\mathbf{1}_{|L^*_{0,k}|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|)>D}].\end{aligned}$$

Equation (5) ensures that $|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)| \leqslant |L^*_{0,k}| + \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| + 2\log\log n$ when $n \geqslant \exp(C_{\mathbf{Q}})$, so that

$$\begin{aligned}&|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)]| \\ &\leqslant \mathbb{E}^*\left[|L^*_{0,k}|\left(\mathbf{1}_{|L^*_{0,k}|>D} + \mathbf{1}_{|L^*_{0,k}|\leqslant D< \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|}\right)\right] \\ &\quad + \mathbb{E}^*\left[\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\left(\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>D} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\leqslant D<|L^*_{0,k}|}\right)\right] \\ &\quad + 2(\log\log n)\mathbb{E}^*\left[\left(\mathbf{1}_{|L^*_{0,k}|>D} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>D}\right)\right].\end{aligned}$$

Thus, by Lemma 14,

$$\begin{aligned}&|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)]| \\ &\qquad\leqslant 2De^{-u} + D\mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| > D\right) + 2De^{-u} + D\mathbb{P}^*(|L^*_{0,k}| > D) \\ &\qquad\quad + 2(\log\log n)\left(\mathbb{P}^*\left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| > D\right) + \mathbb{P}^*(|L^*_{0,k}| > D)\right) \\ &\qquad\leqslant 6De^{-u} + 4(\log\log n)e^{-u}.\end{aligned}$$

Finally, using $\log\log n \leqslant \log n \leqslant C_\gamma \log nu$ when $u \geqslant 1$ concludes the proof of the first equation.

For the second equation, take $u = 3\log n$. Since $\rho^k \leqslant n^{-1}$ when $k \geqslant C_{\mathbf{Q}}^2(\log n)^3$ and $u \longmapsto ue^{-u}$ is nonincreasing on $[1, +\infty)$, for all $D \geqslant 3C_\gamma(\log n)^2$,

$$\begin{aligned}\left|\nu(K,\pi,\mathbf{Q},\gamma) - \bar\nu_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})\right| &\\ &\leqslant 10C_\gamma(\log n)(3\log n)e^{-3\log n} + \frac{2(C_{\mathbf{Q}}\log n)^4}{n\rho} + \frac{4(C_{\mathbf{Q}}\log n)^2}{n\rho}\\ &\leqslant 30C_\gamma\frac{(\log n)^2}{n^3} + \frac{12(C_{\mathbf{Q}}\log n)^4}{n} \leqslant \frac{13(C_{\mathbf{Q}}\log n)^4}{n}\end{aligned}$$

for $n \geqslant \sqrt{30C_\gamma}$ (using $\rho \leqslant 1/2$ for the second line).

5.2.2 Proof of Lemma 13

Lemma 16 *Assume [Atail], [Aergodic] and [A⋆tail] hold. 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 $n \geqslant e^4 \vee \exp(C_{\mathbf{Q}})$,

$$\frac{1}{44C_\gamma^2(\log n)^4}\mathbf{V}(K,\mathbf{Q},\gamma) \leqslant \mathbf{K}(K,\mathbf{Q},\gamma) + \frac{11}{n}.$$

Proof We need the following lemma:

Lemma 17 (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:

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

Let $n \in \mathbb{N}^*$ and $D' = C_\gamma(\log n)^2$. Take $P = \mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}}$ and $Q = \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)}$, so that $\mathbb{E}_P(\log\frac{p}{q})^2 = \mathbf{V}(K,\mathbf{Q},\gamma)$. Using equation (5) for $n \geqslant \exp(C_{\mathbf{Q}})$,

$$\begin{aligned}\left(\log\frac{p}{q}\right)^2 &\leqslant \left(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| + |L^*_{0,\infty}| + 2\log\log n\right)^2\\ &\leqslant 3\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|^2 + 3|L^*_{0,\infty}|^2 + 12(\log\log n)^2\end{aligned}$$

Let $\lambda > 0$ be such that $2D' = \log \frac{1}{\lambda} - 2\log\log n$. Note that $\lambda \leqslant e^{-4}$ when $n \geqslant e^4$. By equation (5),

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

hence

$$\begin{aligned}&8\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)^2 \mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right)\right] \\ &\leqslant 24\mathbb{E}^*\left[|L^*_{0,\infty}|^2\left(\mathbf{1}_{|L^*_{0,\infty}|>D'} + \mathbf{1}_{|L^*_{0,\infty}|\leqslant D' < \sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|}\right)\right] \\ &\quad + 24\mathbb{E}^*\left[\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|^2\left(\mathbf{1}_{\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|>D'} + \mathbf{1}_{\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|\leqslant D' < |L^*_{0,\infty}|}\right)\right] \\ &\quad + 96(\log\log n)^2\mathbb{E}^*\left[\mathbf{1}_{|L^*_{0,\infty}|>D'} + \mathbf{1}_{\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)|>D'}\right],\end{aligned}$$

and by Lemma 14 (for n large enough)

$$\begin{aligned}8\mathbb{E}_P\left[\left(\log\frac{p}{q}\right)^2 \mathbf{1}\left(\frac{p}{q} \geqslant \frac{1}{\lambda}\right)\right] &\leqslant 24\left(\frac{5D'^2}{n} + \frac{D'^2}{n}\right) + 24\left(\frac{5D'^2}{n} + \frac{D'^2}{n}\right) \\ &\quad + 96\frac{2(\log\log n)^2}{n} \\ &\leqslant \frac{480}{n}D'^2\end{aligned}$$

using $\log\log n \leqslant D'$ for $n \geqslant e$. Therefore, by Lemma 17,

$$\begin{aligned}\mathbf{V}(K,\mathbf{Q},\gamma) &\leqslant \mathbb{E}^*_{Y_{-\infty}^{-1}}\left[H(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}}, \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})^2\right](12 + 2(2D' + 2\log\log n)^2) \\ &\quad + \frac{480}{n}D'^2 \\ &\leqslant \mathbb{E}^*_{Y_{-\infty}^{-1}}\left[KL(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}} \| \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})\right](12 + 32D'^2) + \frac{480}{n}D'^2\end{aligned}$$

using that the Kullback Leibler divergence is lower bounded by the Hellinger distance. Finally, since $\mathbb{E}^*_{Y_{-\infty}^{-1}}\left[KL(\mathbb{P}^*_{Y_0|Y_{-\infty}^{-1}} \| \mathbb{P}_{Y_0|Y_{-\infty}^{-1},(K,\mathbf{Q},\gamma)})\right] = \mathbf{K}(K,\mathbf{Q},\gamma)$,

$$\mathbf{V}(K,\mathbf{Q},\gamma) \leqslant 44D'^2\mathbf{K}(K,\mathbf{Q},\gamma) + \frac{480}{n}D'^2.$$

■

Next, let $D > 0$ and let us bound the difference between $\mathbf{V}(K, \mathbf{Q}, \gamma)$ and $\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2]$. Taking $t_{(K,\mathbf{Q},\gamma)} : Y^0_{-k} \longmapsto L^*_{0,k} - L_{0,k,x}(K, \mathbf{Q}, \gamma)$, 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,

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

by Lemma 15 and equation (5), provided $\rho_* \leqslant \rho$ and $C_* \leqslant 1/(1-\rho)$ (which is ensured by $\log n \geqslant (C_* \vee (1-\rho_*)^{-1})^{1/2}$). Note that the condition $k \geqslant C^2_{\mathbf{Q}}(\log n)^3$ ensures that $\rho^k \leqslant n^{-1}$, and that $\rho \leqslant 1/2$ when $n \geqslant e^4$. The expectation can be upper bounded using Lemma 14 with $u = 1$:

$$\begin{aligned}
|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] - \mathbf{V}(K, \mathbf{Q}, \gamma)| &\leqslant \frac{2}{n\rho(1-\rho)}(8C_\gamma \log n + 4\log\log n) \\
&\leqslant \frac{48C^2_{\mathbf{Q}}C_\gamma}{n}(\log n)^3.
\end{aligned}$$

Therefore, under the assumptions of Lemma 16, if $D \geqslant C_\gamma(\log n)^2$,

$$\begin{aligned}
\frac{\mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2]}{44C^2_\gamma(\log n)^4} &\leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) + \frac{11}{n} + \frac{1}{44C^2_\gamma(\log n)^4}\frac{48C^2_{\mathbf{Q}}C_\gamma}{n}(\log n)^3 \\
&\leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) + \frac{11}{n} + \frac{48C^2_{\mathbf{Q}}}{44n\log n} \\
&\leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) + \frac{22}{n}
\end{aligned}$$

for n larger than a constant that only depends on $C_{\mathbf{Q}}$, which concludes the proof.

Acknowledgements

I am grateful to Élisabeth Gassiat for her precious advice and insightful discussions. I would also like to thank the anonymous referee for his patience and very helpful review.

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

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

Randal Douc, Éric 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 François Roueff. Posterior consistency for partially observed Markov models. *Stochastic Processes and their Applications*, 130(2):733–759, 2020.

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. *The Journal of Machine Learning Research*, 19(1):1432–1477, 2018.

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. *arXiv preprint arXiv:1612.04932*, 2016.

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

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

Élodie 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 7 (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 $\gamma \in S_{K,M,n}^{(\gamma)}$,

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

where we recall that the maximum is taken over $\mu \in [-n, n]$ and $s \in [\frac{1}{n}, n]$.

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

$$\begin{aligned} b_\gamma(y) &\leqslant 0 \vee \big(\log n + \log(1+y^2) + \log \pi\big) \\ &\leqslant \log n + 1 + 2\log n + \log \pi \quad \text{since } n \geqslant 1 \\ &\leqslant 3\log n + \log(\pi e) \leqslant 5 \log n \end{aligned}$$

as soon as $n \geqslant 3$. Otherwise, one can take $y \geqslant n$ and then

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

by writing $Y = y - n$ and using that $\log(a+b) \leqslant \log a + \log(1+b) \leqslant \log a + 1 + \log b$ when $a, b \geqslant 1$. Thus, writing $Y' = Y/n$,

$$\begin{aligned} b_\gamma(y) &\leqslant 3\log n + \log(2e\pi) + 0 \vee (-(Y')^p + \log(1 + 2n^2 (Y')^2)) \\ &\leqslant \begin{cases} 3\log n + \log(2e\pi) + \log(1+2n^2) & \text{if } Y' \leqslant 1 \\ 3\log n + \log(2e\pi) + 0 \vee (-(Y')^p + 1 + \log(2n^2(Y')^2)) & \text{otherwise} \end{cases} \\ &\leqslant 5\log n + \log(4e^2\pi) + 0 \vee (-(Y')^p + 2\log(Y')) \\ &\leqslant 5\log n + \log(4e^2\pi), \end{aligned}$$

so that $b_\gamma(y) \leqslant 10 \log n$ as soons as $n \geqslant 3$.

Checking [Aentropy] and [Agrowth] 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 $([n,n] \times [\frac{1}{n}, n])^M$ provides a bracket covering of $\{\gamma_x\}_{\gamma \in \mathbf{S}_n^{(\gamma)}, x \in [K]}$ of size $L_p \epsilon$. Since there exists a bracket covering of size ϵ of $[n,n] \times [\frac{1}{n}, n]$ for the supremum norm with less than $(\frac{4n}{\epsilon} \vee 1)^2$ brackets, one gets [Aentropy] by taking $C_{\text{aux}}(M, K, D, n) = 4L_p n$ and $m_M = 2M$.

Let us now check that this constant L_p exists.

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

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

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

as soon as $p \geqslant 2$. Thus, one can take $L_p = pn^2$ and $C_{\text{aux}}(M, K, D, n) = 4pn^3$. With this C_{aux}, checking [Agrowth] is straightforward for all $\zeta > 0$: with $n \geqslant 4p$, it is ensured by $\log n^4 \leqslant n^\zeta$, which is always true for $\zeta = 2$ for instance.

A.1.2 Proof of Lemma 9 (approximation rates)

Let $F(y) = e^{-c|y|^\tau}$. Lemma 4 of Kruijer et al. (2010) ensures that there exist $c' > 0$ and $H \geqslant 6\beta + 4p$ such that for all $x \in [K^*]$ and $s > 0$, there exists a mixture $g_{s,x}$ with $O(s^{-1}|\log s|^{p/\tau})$ components, each with density $\frac{1}{s}\psi(\frac{\cdot - \mu}{s})$ with respect to the Lebesgue measure for some $\mu \in \{y \,|\, F(y) \geqslant c's^H\}$, such that $g_{s,x}$ approximates the emission density γ_x^*:

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

For M large enough, the condition $M \geqslant O(s^{-1}|\log s|^{p/\tau})$ (number of components) is ensured by $s^{-1} \leqslant cM(\log M)^{-p/\tau}$ for some small enough constant $c > 0$. Take this s in the following and $g_{M,x} = g_{s,x}$.

Let us now check that $(n^{-2} + (1 - n^{-2})g_{M,x})_{x\in[K^*]} \in S^{(\gamma)}_{K^*,M,n}$. When $M \leqslant n$ is large enough that $s \leqslant 1$, this s is indeed in $[\frac{1}{n}, n]$. When $|\mu| \geqslant s^{-1}$, $F(\mu) \leqslant \exp(-cs^{-\tau}) = o(c's^H)$. Thus, for s small enough (i.e. for M large enough), all translation parameters μ belong to $[-s^{-1}, s^{-1}]$, which is indeed in $[-n, n]$ when $M \leqslant n$.

A.1.3 Proof of Corollary 10 (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 15 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 18 *Assume that [A⋆tail], [A⋆forget], [Atail] and [Aergodic] hold. Then there exists a constant n_1 depending on C_γ, $C_{\mathbf{Q}}$, δ and M_δ such that for all $n \geqslant n_1$, for all $(K, \mathbf{Q}, \gamma) \in \mathbf{S}_n$,*

$$\mathbf{H}^2(K, \mathbf{Q}, \gamma) \leqslant \mathbf{K}(K, \mathbf{Q}, \gamma) \leqslant 7C_\gamma(\log n)^2\left(\mathbf{H}^2(K, \mathbf{Q}, \gamma) + \frac{2}{n}\right).$$

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 by equation (4), $\log\frac{p}{q} \leqslant |b_\gamma| + |L^*_{1,\infty}| + \log(C_{\mathbf{Q}}\log n)$ where $L^*_{1,\infty}$ is as in Lemma 15 and $\mathbf{1}\left\{\log\frac{p}{q} \geqslant v\right\} \leqslant \mathbf{1}\{|b_\gamma| \geqslant \frac{1}{2}(v - \log(C_{\mathbf{Q}}\log n))\} \vee \mathbf{1}\{|L^*_{1,\infty}| \geqslant \frac{1}{2}(v - \log(C_{\mathbf{Q}}\log n))\}$. There exists n_1 depending only on C_γ and $C_{\mathbf{Q}}$ such that for all $n \geqslant n_1$, $\log(C_{\mathbf{Q}}\log n) \leqslant C_\gamma(\log n)^2$. Assume $n \geqslant n_1$ and take $v = 3C_\gamma(\log n)^2$, then $\frac{1}{2}(v - \log(C_{\mathbf{Q}}\log n)) \geqslant (C_\gamma \log n)^2$ and $1 + 2v \leqslant 7C_\gamma(\log n)^2$, so that

$$\begin{aligned}\mathbf{K}(K,\mathbf{Q},\gamma) \leqslant\ & 7C_\gamma(\log n)^2\mathbf{H}^2(K,\mathbf{Q},\gamma)\\ &+2C_\gamma(\log n)^2\left\{\mathbb{P}^*(|b_\gamma| \geqslant C_\gamma(\log n)^2) + \mathbb{P}^*(|L^*_{1,\infty}| \geqslant C_\gamma(\log n)^2)\right\}\\ &+2\mathbb{E}^*[(|L^*_{1,\infty}| + |b_\gamma|)\\ &\qquad\times(\mathbf{1}\{|L^*_{1,\infty}| \geqslant C_\gamma(\log n)^2\} \vee \mathbf{1}\{|b_\gamma| \geqslant C_\gamma(\log n)^2\})].\end{aligned}$$

By Lemma 14, which also holds for $L^*_{1,\infty}$ using the uniform convergence of Lemma 15, $\mathbb{P}^*(|L^*_{1,\infty}| \geqslant C_\gamma(\log n)^2) \leqslant \exp(-\log n) \leqslant n^{-1}$ for $n \geqslant n_0$ where n_0 is defined in Lemma 14 (and depends on δ and M_δ). Likewise, by [Atail], $\mathbb{P}^*(|b_\gamma| \geqslant C_\gamma(\log n)^2) \leqslant n^{-1}$.

The last expectation of the above equation can be written as

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

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

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

by Lemma 14 for the first term and [Atail] for the second one. Likewise,

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

so that finally

$$\mathbf{K}(K,\mathbf{Q},\gamma) \leqslant 7C_\gamma(\log n)^2\mathbf{H}^2(K,\mathbf{Q},\gamma) + 14C_\gamma\frac{(\log n)^2}{n},$$

which concludes the proof. ∎

Let $M \in \mathbb{N}^*$. Let $g_{M,x}$ be the approximating densities given by Lemma 9 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 19 *Let $\sigma^* > 0$ be such that $\sigma^* \leqslant K^*\mathbf{Q}^*(x,x') \leqslant (\sigma^*)^{-1}$ for all $x, x' \in [K^*]$. Then*

$$\mathbf{H}^2(K^*, \mathbf{Q}^*, (\gamma_{M,x})_x) \leqslant \left(2 + \frac{32(K^*)^3}{(\sigma^*)^{11}}\right)\sum_{x\in[K^*]} h^2(\gamma^*_x, \gamma_{M,x})$$

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 $[\frac{\sigma^*}{K^*}, \frac{1}{K^*\sigma^*}]$ by assumption on $\mathbf{Q}^*$,

$$\sum_x (\sqrt{p_x} - \sqrt{p_x^*})^2 \in \left[\frac{K^*\sigma^*}{4}, \frac{K^*}{4\sigma^*}\right] \sum_x (p_x - p_x^*)^2.$$

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 and that their reasoning works with $\rho_\star = 1 - \min \mathbf{Q}_\star / \max \mathbf{Q}_\star$.

$$\sum_x |p_x - p_x^*| \leqslant \frac{4K^*}{(\sigma^*)^3} \sum_{i=0}^{+\infty} (1 - (\sigma^*)^2)^i \frac{\max_x |\gamma_x^*(Y_{-i}) - \gamma_{M,x}(Y_{-i})|}{\sum_x \gamma_x^*(Y_{-i}) \vee \sum_x \gamma_{M,x}(Y_{-i})}.$$

Therefore, using Cauchy-Schwarz's inequality:

$$\begin{aligned}
\sum_x (p_x - p_x^*)^2 &\leqslant \left(\sum_x |p_x - p_x^*|\right)^2 \\
&\leqslant \frac{16(K^*)^2}{(\sigma^*)^8} \sum_{i=0}^{+\infty} (1 - (\sigma^*)^2)^i \left(\frac{\max_x |\gamma_x^*(Y_{-i}) - \gamma_{M,x}(Y_{-i})|}{\sum_x \gamma_x^*(Y_{-i}) \vee \sum_x \gamma_{M,x}(Y_{-i})}\right)^2.
\end{aligned}$$

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

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

so that

$$\begin{aligned}\mathbb{E}^*\left[\sum_x(\sqrt{p_x^*}-\sqrt{p_x})^2\right] &\leqslant \frac{K^*}{4\sigma^*}\mathbb{E}^*\left[\sum_x(p_x-p_x^*)^2\right]\\ &\leqslant \frac{16(K^*)^3}{(\sigma^*)^{11}}\sum_x h^2(\gamma_x^*,\gamma_{M,x}),\end{aligned}$$

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 $c\geqslant 0$, for all x,

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

Therefore,

$$\begin{aligned}\mathbf{K}(K^*,\mathbf{Q}^*,(\gamma_{M,x})_x) \leqslant{}& 14C_\gamma\frac{(\log n)^2}{n}\\ &+7C_\gamma(\log n)^2\left(2+\frac{32(K^*)^3}{(\sigma^*)^{11}}\right)\sum_{x\in[K^*]}\left(\frac{4}{n^2}+2KL(\gamma_x^*,g_{M,x})\right).\end{aligned}$$

Thus, there exists a constant C such that for all $n\geqslant 3$,

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

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

The choice of penalty verifies the lower bound of Theorem 6. Thus, the oracle inequality of Theorem 6 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 $K^*\leqslant M_n\leqslant n/2$ for all n:

$$\begin{aligned}\mathbf{K}(\hat{K},\hat{\mathbf{Q}},\hat{\gamma}) &\leqslant 2\mathbf{K}(K^*,\mathbf{Q}^*,(\gamma_{M_n,x})_x)+4\mathrm{pen}_n(K^*,M_n)+A\frac{(\log n)^{10}}{n}\\ &\leqslant 2C(\log n)^2\left(\frac{1}{n}+M_n^{-2\beta}(\log n)^{2\beta\frac{p}{\tau}}\right)\\ &\quad+4K^*\frac{(\log n)^{18}}{n}M_n+2A\frac{(\log n)^{10}}{n}.\end{aligned}$$

Taking $M_n\sim n^{\frac{1}{2\beta+1}}(\log n)^{\frac{2\beta p/\tau-16}{2\beta+1}}$ leads to the desired rate.

Appendix B. Proof of the control of $\bar{\nu}_k$ (Theorem 12)

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

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

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.

Without loss of generality, we assume $n \geqslant \exp(C_{\mathbf{Q}})$ and $D \geqslant \log n$ so that $\|t^{(D)}_{(K,\mathbf{Q},\gamma)}\|_\infty \leqslant 4D$ for all $(K,\pi,\mathbf{Q},\gamma) \in S_{K,M,n}$ by equation (5) and n larger than the constant n_0 from Lemma 14.

Changes of notations. In the rest of this section, we omit the dependency of $W_{K,M}$, $x_{K,M}$ and $S_{K,M}$ on n in the notations. We also introduce the notation $\theta \in \mathbf{S}_n$ instead of $(K,\pi,\mathbf{Q},\gamma) \in \mathbf{S}_n$ to make the notation shorter. Given $\theta \in \mathbf{S}_n$, we write π_θ, $\mathbf{Q}_\theta$ and γ_θ its components. To avoid multiple subscripts, we write $\gamma_\theta(y|x)$ instead of $\gamma_{\theta,x}(y)$.

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 $d_k^2(t_1,t_2) = \mathbb{E}^*[(t_1-t_2)^2(Z_0)]$. For any semi-distance d, write $N(A,d,\epsilon) = e^{H(A,d,\epsilon)}$ 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 20 *Assume [A⋆mix] 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 $\mathcal{Y}^{k+1}$ and 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 $n \in \mathbb{N}$:

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

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

Lemma 21 (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 exist 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

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

Assumption [A⋆mix] 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 21, one gets that there exist two positive constants C_1 and C_2 depending on c_* and n_* such that for all functions 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 $V = \mathbb{E}^* t^2(Z_0)$:

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

using [A⋆mix]. 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 $(A_i)_{1\leqslant i\leqslant k}$,

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

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 22 *Assume [A⋆mix] 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 exist 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 sets A such that $\mathbb{P}^*(A) > 0$:

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

where

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

By taking $\mathcal{T} = \{t_\theta^{(D)} \,|\theta \in \mathbf{B}_\sigma\}$ and $b = 4D$, one gets the following lemma from Lemma 4.23 and Lemma 2.4 of Massart (2007):

Lemma 23 *Assume that there exist a function φ and constants C and $\sigma_{K,M}$ such that $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing and*

$$\forall \sigma \geqslant \sigma_{K,M} \qquad E \leqslant C\varphi(\sigma)\sqrt{n}. \tag{11}$$

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

$$W_{K,M} := \sup_{\theta \in S_{K,M}} \left| \frac{|\bar{\nu}_k(t_\theta^{(D)})|}{\mathbb{E}^*[t_\theta^{(D)}(Z_0)^2] + x_{K,M}^2} \right| \leqslant 4C^*(n_* + k + 1)\left[C\frac{\varphi(x_{K,M})}{x_{K,M}^2\sqrt{n}} + \sqrt{\frac{z}{x_{K,M}^2 n}} + 4D\frac{z(\log n)^2}{x_{K,M}^2 n} \right]. \tag{12}$$

The two remaining steps are the control of the bracketing entropy which will lead to equation (11) (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

In this section, we show that for all $k \geqslant 2$ and $\epsilon > 0$,

$$H(\epsilon) \leqslant 2(m_M K + K^2 - 1)\log \max\left(\frac{95De^{2D}\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{k+3/2}\sqrt{kKC_{\text{aux}}{}'}}{\epsilon}, 14\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{k+1/2}\sqrt{kKC_{\text{aux}}{}'} \right)$$

where $C_{\text{aux}}{}' = (C_{\text{aux}}e^D) \vee (K-1)$.

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

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

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

$$|L^*_{k,k}| \vee \sup_{\theta' \in \mathbf{S}_n} |b_{\theta'}(y_k)| \leqslant D \tag{13}$$

since if this is not the case, then $t^{(D)}_\theta(y_k) = t^{(D)}_{\theta'}(y_k) = 0$. *This means that only the* y_k *satisfying equation* (13) *are relevant for the construction of the brackets.*

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

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

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

$$(C_\mathbf{Q}(\log n)e^D)^{-1} \leqslant \sum_{x\in[K]} g_{\theta,x} \leqslant C_\mathbf{Q}(\log n)e^D. \tag{14}$$

Let $[a, b]$ be a bracket of size ϵ for $\mathcal{G}$ with the distance $d_\mathcal{G}$ such that

$$(2C_\mathbf{Q}(\log n)e^D)^{-1} \leqslant \sum_x a_x \leqslant \sum_x b_x \leqslant 2C_\mathbf{Q}(\log n)e^D. \tag{15}$$

Then

$$\begin{aligned}
\left(\log\sum_x a_x - \log\sum_x b_x\right)^2 &\leqslant 2\log\left(2C_\mathbf{Q}(\log n)e^D\right)\left|\log\sum_x a_x - \log\sum_x b_x\right| \\
&\leqslant 8D \times 2C_\mathbf{Q}(\log n)e^D\sum_x |a_x - b_x|
\end{aligned}$$

when $n \geqslant e^2$ using that $|\log a - \log b| \leqslant |a - b|/(a \wedge b)$. Therefore,

$$\begin{aligned}
&d_k\left(\log\sum_x a_x, \log\sum_x b_x\right)^2 \\
&= \mathbb{E}^*_{Y_0^{k-1}}\left[\int\left(\log\sum_x a_x - \log\sum_x b_x\right)^2 (Y_0^{k-1}, y_k)p^*(Y_k = y_k|Y_0^{k-1})\lambda(dy_k)\right] \\
&\leqslant 16DC_\mathbf{Q}(\log n)e^D\mathbb{E}^*_{Y_0^{k-1}}\left[\int\sum_x |a_x - b_x|(Y_0^{k-1}, y_k)\exp(L^*_{k,k})\lambda(dy_k)\right] \\
&\leqslant 16DC_\mathbf{Q}(\log n)e^{2D}d_\mathcal{G}(a,b),
\end{aligned}$$

so that

$$N(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, \epsilon) \leqslant \bar{N}\left(\mathcal{G}, d_\mathcal{G}, \left(\frac{\epsilon}{16DC_\mathbf{Q}(\log n)e^{2D}}\right)^2\right) \tag{16}$$

where $\bar{N}$ is the minimal cardinality of a bracket covering of $\mathcal{G}$ such that all brackets $[a, b]$ satisfy equation (15).

B.2.2 Decomposition into simple sets

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

Lemma 25 *Assume $k \geqslant 2$ and let $\epsilon \in \left(0, \frac{70}{168}\right)$. Then*

$$\begin{aligned}\bar{N}\left(\mathcal{G}, d_\mathcal{G}, \epsilon\right) \leqslant N&\left(\{\pi_\theta\}_{\theta\in S_{K,M}}, d_\infty, \frac{\epsilon}{70k\left(\sqrt{2}C_\mathbf{Q}\log n\right)^{2k+1}K}\right)\\ &\times N\left(\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}, d_\infty, \frac{\epsilon}{70k\left(\sqrt{2}C_\mathbf{Q}\log n\right)^{2k+1}K}\right)\\ &\times N\left(\{\gamma_\theta\}_{\theta\in S_{K,M}}, d_\infty, \frac{\epsilon\, e^{-D}}{70k\left(\sqrt{2}C_\mathbf{Q}\log n\right)^{2k+1}K}\right)\end{aligned}$$

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, we assume $(C_\mathbf{Q}\log n)^{-1}K^{-1} \leqslant a(x) \leqslant b(x) \leqslant C_\mathbf{Q}(\log n)K^{-1}$ and $(C_\mathbf{Q}\log n)^{-1}K^{-1} \leqslant p(x, x') \leqslant q(x, x') \leqslant C_\mathbf{Q}(\log n)K^{-1}$ 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. We also assume that the brackets aren't empty: 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, for all $y \in \mathcal{Y}$,

$$Ke^{-D}(1-\epsilon) \leqslant \sum_x u(y|x) \leqslant \sum_x v(y|x) \leqslant K(e^D + \epsilon e^{-D}). \tag{17}$$

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

$$\beta_{i|k}(x_i) = \sum_{x_{i+1}^k \in [K]^{k-i}} \mathbf{Q}_\theta(x_i, x_{i+1})\gamma_\theta(y_{i+1}|x_{i+1}) \dots \mathbf{Q}_\theta(x_{k-1}, x_k)\gamma_\theta(y_k|x_k),$$

for $0 \leqslant i \leqslant k-1$ and $\beta_{k|k}(x) = 1$ for all $x \in [K]$,

$$\mu_{0|k}^\theta(x) = \frac{\pi_\theta(x)\beta_{0|k}(x)\gamma_\theta(y_0|x)}{\sum_{x' \in [K]} \pi_\theta(x')\beta_{0|k}(x')\gamma_\theta(y_0|x')}$$
$$\text{and} \quad F_{i|k}^\theta(x_{i-1}, x_i) = \frac{\beta_{i|k}(x_i)\mathbf{Q}_\theta(x_{i-1}, x_i)\gamma_\theta(y_i|x_i)}{\sum_{x \in [K]} \beta_{i|k}(x)\mathbf{Q}_\theta(x_{i-1}, x)\gamma_\theta(y_i|x)}.$$

To clarify the role of these quantities, observe that

$$\begin{aligned}
&\beta_{i|k}(x_i) = p_\theta(Y_{i+1}^k | X_i = x_i),\\
&\mu_{0|k}^\theta(x) = \mathbb{P}_\theta(X_0 = x | Y_0^k),\\
&F_{i|k}^\theta(x_{i-1}, x_i) = \mathbb{P}_\theta(X_i = x_i | Y_i^k, X_{i-1} = x_{i-1}),
\end{aligned}$$

so that

$$\left(\mu_{0|k}^\theta F_{1|k}^\theta \dots F_{k|k}^\theta\right)_x = \mathbb{P}_\theta(X_k = x | Y_0^k).$$

Now, let

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

for $0 \leqslant i \leqslant k-1$ and $\alpha_{k|k}(x) = \delta_{k|k}(x) = 1$ for all $x \in [K]$,

$$\begin{cases}
\nu(x) = \dfrac{a(x)\alpha_{0|k}(x)u(y_0|x)}{\sum_{x' \in [K]} b(x')\delta_{0|k}(x')v(y_0|x')}\\
\\
\omega(x) = \dfrac{b(x)\delta_{0|k}(x)v(y_0|x)}{\sum_{x' \in [K]} a(x')\alpha_{0|k}(x')u(y_0|x')}
\end{cases},$$

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 $\{\mu_{0|k}^\theta\}_{\theta \in S_{K,M}}$ and $\{F_{i|k}^\theta\}_{\theta \in S_{K,M}}$ 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 26 *Assume $\epsilon\leqslant\frac{1}{2}$, then*

$$\sup_{0\leqslant i\leqslant k}\frac{\sum_{x\in[K]}|\alpha_{i|k}(x)u(y_i|x)-\delta_{i|k}(x)v(y_i|x)|}{\sum_{x\in[K]}\alpha_{i|k}(x)u(y_i|x)}\leqslant 4\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1}K\epsilon.$$

Proof Using minimalist notations,

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

Then, note that for all $j\in\{i+1,\dots,k\}$,

$$\begin{aligned}
&\sum_{x_i^k\in[K]^{k-i+1}}u_ip_i^{i+1}\dots p_{j-2}^{j-1}u_{j-1}|p_{j-1}^j-q_{j-1}^j|v_jq_j^{j+1}\dots q_{k-1}^kv_k\\
&\qquad\leqslant\epsilon(C_{\mathbf{Q}}(\log n)K^{-1})^{k-j}\sum_{x_i^{j-1}\in[K]^{j-i}}u_ip_i^{i+1}\dots p_{j-2}^{j-1}u_{j-1}\\
&\qquad\qquad\times\sum_{x_j\in[K]}(u_j+\epsilon e^{-D})\cdots\sum_{x_k\in[K]}(u_k+\epsilon e^{-D})
\end{aligned}$$

and for all $j\in\{i,\dots,k\}$ (with a special case for $j=i$),

$$\begin{aligned}
\sum_{x\in[K]}\alpha_{i|k}(x)u(y_i|x)&=\sum_{x_i^k\in[K]^{k-i+1}}u_ip_i^{i+1}\dots p_{j-2}^{j-1}u_{j-1}p_{j-1}^ju_jp_j^{j+1}\dots p_{k-1}^ku_k\\
&\geqslant(C_{\mathbf{Q}}(\log n)K)^{-(k-j+1)}\sum_{x_i^{j-1}\in[K]^{j-i}}u_ip_i^{i+1}\dots p_{j-2}^{j-1}u_{j-1}\sum_{x_j\in[K]}u_j\cdots\sum_{x_k\in[K]}u_k.
\end{aligned}$$

so that

$$\begin{aligned}
\frac{\sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1}\dots u_{j-1}|p_{j-1}^j - q_{j-1}^j|v_j\dots q_{k-1}^k v_k}{\sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1}\dots u_{j-1}p_{j-1}^j u_j\dots p_{k-1}^k u_k} &\\
&\leqslant \epsilon K (C_{\mathbf{Q}}\log n)^{2(k-j)+1}\prod_{\ell=j}^{k}\frac{Ke^{-D}\epsilon + \sum_{x_\ell} u_\ell}{\sum_{x_\ell} u_\ell}\\
&\leqslant \epsilon K (C_{\mathbf{Q}}\log n)^{2(k-j)+1}\prod_{\ell=j}^{k}\left(1+\frac{Ke^{-D}\epsilon}{Ke^{-D}(1-\epsilon)}\right)\\
&\leqslant \epsilon K \frac{(C_{\mathbf{Q}}\log n)^{2(k-j)+1}}{(1-\epsilon)^{k-j+1}}.
\end{aligned}$$

Likewise, for all $j\in\{i,\dots,k\}$,

$$\frac{\sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1}\dots p_{j-1}^j|u_j - v_j|q_j^{j+1}\dots q_{k-1}^k v_k}{\sum_{x_i^k\in[K]^{k-i+1}} u_i p_i^{i+1}\dots u_{j-1}p_{j-1}^j u_j\dots p_{k-1}^k u_k} \leqslant \epsilon K\frac{(C_{\mathbf{Q}}\log n)^{2(k-j)+1}}{(1-\epsilon)^{k-j+1}}.$$

Therefore, when $\epsilon\leqslant 1/2$,

$$\begin{aligned}
\frac{\sum_{x\in[K]}|\alpha_{i|k}(x)u(y_i|x)-\delta_{i|k}(x)v(y_i|x)|}{\sum_{x\in[K]}\alpha_{i|k}(x)u(y_i|x)} &\leqslant 2\frac{\epsilon K}{C_{\mathbf{Q}}\log n}\sum_{j=i}^{k}\left(2(C_{\mathbf{Q}}\log n)^2\right)^{k-j+1}\\
&\leqslant 4\epsilon K C_{\mathbf{Q}}(\log n)\frac{\left(2(C_{\mathbf{Q}}\log n)^2\right)^{k-i}-1}{2(C_{\mathbf{Q}}\log n)^2-1}\\
&\leqslant 4\epsilon K\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2(k-i)+1}
\end{aligned}$$

since $n\geqslant e^2$, which gives the desired result. ■

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

$$\|\nu-\omega\|_1 \leqslant 6\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon$$

and

$$\sup_{0\leqslant i\leqslant k}\sup_{x\in[K]}\|f_{i|k}(x,\cdot)-g_{i|k}(x,\cdot)\|_1 \leqslant 6\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon. \tag{18}$$

Proof With minimalist notations,

$$\begin{aligned}
\sum|\nu-\omega| &= \sum\left|\frac{a\alpha u}{\sum b\delta v}-\frac{b\delta v}{\sum a\alpha u}\right|\\
&\leqslant \frac{\sum|a\alpha u - b\delta v|}{\sum b\delta v}+\sum|b\delta v|\left|\frac{1}{\sum a\alpha u}-\frac{1}{\sum b\delta v}\right|\\
&\leqslant \frac{\sum|a\alpha u - b\delta v|}{\sum b\delta v}+\frac{\sum|a\alpha u - b\delta v|}{\sum a\alpha u}\\
&\leqslant 2C_{\mathbf{Q}}(\log n)K\frac{\sum|a\alpha u - b\delta v|}{\sum\alpha u}
\end{aligned}$$

using $(C_{\mathbf{Q}}(\log n)K)^{-1} \leqslant a \leqslant b \leqslant C_{\mathbf{Q}}(\log n)K^{-1}$, $0 \leqslant \alpha \leqslant \delta$ and $0 \leqslant u \leqslant v$. Thus,

$$\begin{aligned}\sum |\nu - \omega| &\leqslant 2C_{\mathbf{Q}}(\log n)K\left(\frac{\sum b|\alpha u - \delta v|}{\sum \alpha u} + \frac{\sum |a-b|\alpha u}{\sum \alpha u}\right)\\
&\leqslant 2C_{\mathbf{Q}}(\log n)K\left(C_{\mathbf{Q}}(\log n)K^{-1}\frac{\sum |\alpha u - \delta v|}{\sum \alpha u} + \epsilon\right)\\
&\leqslant 2C_{\mathbf{Q}}(\log n)K\left(C_{\mathbf{Q}}(\log n)4\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1}\epsilon + \epsilon\right) \text{ (Lemma 26)}\\
&\leqslant 6\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon.\end{aligned}$$

The control of $\sum_{x'\in[K]} |g_{i|k} - f_{i|k}|(x,x')$ is the same after replacing a and b by p and q. ■

Write $\eta = 6\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon$. Equation (18) implies that as soon as $\eta < 1$, 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 27 such that $f'_{i|k}/(1-\eta)$ and $g'_{i|k}/(1+\eta)$ are transition matrices.

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 28 *Assume $\epsilon \leqslant \frac{1}{12k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K}$. Then for all y_0^k,*

$$\sum_{x\in[K]} |(\nu f'_{1|k}\dots f'_{k|k})_x - (\omega g'_{1|k}\dots g'_{k|k})_x| \leqslant 7k\eta = 42k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon$$

and

$$\sum_{x\in[K]} |(\nu f'_{1|k}\dots f'_{k|k}p)_x - (\omega g'_{1|k}\dots g'_{k|k}q)_x| \leqslant 64k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon.$$

Proof First,

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

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

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

By Lemma 27, $\|\omega\|_1 \leqslant 1+\eta$ (since the bracket $[\nu,\omega]$ contains a probability distribution $\mu^{\theta}_{0|k}$ for some $\theta \in S_{K,M}$) and $\sup_{1\leqslant i\leqslant k}\sup_{x\in[K]}\|f'_{i|k}(x,\cdot)-g'_{i|k}(x,\cdot)\|_1 \leqslant 3\eta$, so that

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

For all $x\in[0,\frac{1}{2}]$, $3(1+x)(e^x-1)\leqslant 6x$. Since $k\eta\leqslant\frac{1}{2}$ by the assumption on ϵ,

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

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 $[p,q]$ is a non-empty bracket of $\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}$, $\sum_x q_{x',x}\leqslant 1+K\epsilon$ for all x' and since $\nu f'_{1|k}\dots f'_{k|k}$ is the lower bound of a non empty bracket of $\{p_{X_k|Y_1^k,\theta}\}_{\theta\in S_{K,M}}$, $\sum_{x'}(\nu f'_{1|k}\dots f'_{k|k})_{x'}\leqslant 1$. Hence,

$$\begin{aligned}&\sum_{x\in[K]}|(\nu f'_{1|k}\dots f'_{k|k}p)_x-(\omega g'_{1|k}\dots g'_{k|k}q)_x|\\ &\leqslant(1+K\epsilon)\sum_{x'}|(\nu f'_{1|k}\dots f'_{k|k})_{x'}-(\omega g'_{1|k}\dots g'_{k|k})_{x'}|+K\epsilon\sum_{x'}(\nu f'_{1|k}\dots f'_{k|k})_{x'}\\ &\leqslant(1+K\epsilon)42\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon+K\epsilon\qquad\text{(by the first part of the lemma)}\\ &\leqslant 64k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+3}K\epsilon\end{aligned}$$

since $\epsilon\leqslant\frac{1}{2K}$ under the assumption of the lemma and $k\wedge(\sqrt{2}C_{\mathbf{Q}}\log n)\geqslant 1$. ■

Lemma 29 *Assume* $\epsilon\leqslant\frac{1}{12k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1}K}$. *Then*

$$d_{\mathcal{G}}(A,B)\leqslant 70k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1}K\epsilon.$$

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},

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

Since the brackets are not empty, for all $x \in [K]$, $\int v(y|x)\lambda(dy) \leqslant 1 + \epsilon e^{-D}$ 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, Lemma 28 entails

$$\begin{aligned}
d_{\mathcal{G}}(A,B) &\leqslant \epsilon e^{-D} \sum_x (\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x\\
&\qquad + (1+\epsilon e^{-D}) \sum_x |(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x - (\omega g'_{1|k-1}\dots g'_{k-1|k-1}q)_x|\\
&\leqslant \epsilon e^{-D} + (1+\epsilon e^{-D}) 64(k-1)\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2(k-1)+3} K\epsilon\\
&\leqslant 70k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K\epsilon
\end{aligned}$$

since $1 + \epsilon e^{-D} \leqslant 13/12$ under the assumption of the lemma. ∎

Assume $k \geqslant 2$ and let $\eta' := 42(k-1)\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K\epsilon$. Lemma 28 implies $\sum_x |(\nu f'_{1|k-1}\dots f'_{k-1|k-1})_x - (\omega g'_{1|k-1}\dots g'_{k-1|k-1})_x| \leqslant \eta'$. Since the bracket $[\nu f'_{1|k-1}\dots f'_{k-1|k-1}, \omega g'_{1|k-1}\dots g'_{k-1|k-1}]$ is not empty, it contains a probability measure. Thus, using $(C_{\mathbf{Q}}\log n)^{-1}K^{-1} \leqslant p \leqslant q \leqslant C_{\mathbf{Q}}(\log n)K^{-1}$, for all $x \in [K]$,

$$\begin{aligned}
(C_{\mathbf{Q}}\log n)^{-1}K^{-1}(1-\eta') &\leqslant (\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x\\
&\leqslant (\omega g'_{1|k-1}\dots g'_{k-1|k-1}q)_x \leqslant C_{\mathbf{Q}}(\log n)K^{-1}(1+\eta').
\end{aligned}$$

Therefore, by equation (17),

$$\begin{aligned}
(C_{\mathbf{Q}}\log n)^{-1}K^{-1}(1-\eta')e^{-D}K(1-\epsilon) &\leqslant \sum_{x\in[K]} A_x\\
&\leqslant \sum_{x\in[K]} B_x \leqslant C_{\mathbf{Q}}(\log n)K^{-1}(1+\eta')K(e^D + \epsilon e^{-D}).
\end{aligned}$$

The inequality $(2C_{\mathbf{Q}}(\log n)e^D)^{-1} \leqslant \sum_{x\in[K]} A_x \leqslant \sum_{x\in[K]} B_x \leqslant 2C_{\mathbf{Q}}(\log n)e^D$ required in the definition of $\bar{N}$ follows as soon as $(1-\eta')(1-\epsilon) \geqslant 1/2$ and $(1+\eta')(1+\epsilon e^{-2D}) \leqslant 2$, for instance when $(1-\eta')^2 \geqslant 1/2$ since $\eta' \geqslant \epsilon$ and $D \geqslant 0$, which holds when $\eta' \leqslant 1/4$, in other words when

$$\epsilon \leqslant \frac{1}{168(k-1)\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K}.$$

Thus, taking $\epsilon' = 70k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K\epsilon$ ensures that if $\epsilon' \leqslant \frac{70}{168}$, then $d_{\mathcal{G}}(A,B) \leqslant \epsilon'$. Lemma 25 follows.

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

Lemma 30 *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 25 ensures that for all $\epsilon \leqslant \frac{70}{168}$,

$$\log \bar{N}\left(\mathcal{G}, d_{\mathcal{G}}, \epsilon\right) \leqslant (m_M K + K^2 - 1)\log\max\left(\frac{70k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K{C_{\text{aux}}}'}{\epsilon}, 1\right),$$

so that using Equation (16) and letting $H(u) = H(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, u)$, one gets for all $\epsilon \leqslant 16DC_{\mathbf{Q}}(\log n)e^{2D}\sqrt{70/168}$ and in particular for all $\epsilon \leqslant 7D(\sqrt{2}C_{\mathbf{Q}}\log n)e^{2D}$:

$$\begin{aligned} H(\epsilon) &\leqslant (m_M K + K^2 - 1)\log\max\left(\frac{(16DC_{\mathbf{Q}}(\log n)e^{2D})^2 70k\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{2k+1} K{C_{\text{aux}}}'}{\epsilon^2}, 1\right) \\ &\leqslant 2(m_M K + K^2 - 1)\log\max\left(\frac{95De^{2D}\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{k+3/2}\sqrt{kK{C_{\text{aux}}}'}}{\epsilon}, 1\right). \end{aligned}$$

Thus, for all $\epsilon > 0$,

$$H(\epsilon) \leqslant 2(m_M K + K^2 - 1)\log\max\left(\frac{95De^{2D}\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{k+3/2}\sqrt{kK{C_{\text{aux}}}'}}{\epsilon}, 14\left(\sqrt{2}C_{\mathbf{Q}}\log n\right)^{k+1/2}\sqrt{kK{C_{\text{aux}}}'}\right).$$

B.3 Choice of parameters

The goal of this section is to find a function φ and a constant C for which equation (11) holds, and to choose the weights $x_{K,M}$ of Lemma 23.

Lemma 31 *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

$$\varphi(u) = u\sqrt{2\pi(m_M K + K^2 - 1)}\Bigg(1+ \left\{ \log \max\left(\frac{95De^{2D}\left(\sqrt{2}C_{\mathbf{Q}} \log n\right)^{k+3/2}\sqrt{kKC_{\text{aux}}{}'}}{u}, 14\left(\sqrt{2}C_{\mathbf{Q}} \log n\right)^{k+1/2}\sqrt{kKC_{\text{aux}}{}'}\right)\right\}^{1/2}\Bigg).$$

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 $(1 + 2\sqrt{D}\log n)\varphi(x) = \sqrt{n}x^2$ with unknown x, when a solution exists. By the definition of E in Lemma 22,

$$\begin{aligned} \forall \sigma \geqslant \sigma_{K,M}, \quad E &\leqslant \sqrt{n}\varphi(\sigma) + 4D(\log n)^2 \frac{\varphi(\sigma)^2}{\sigma^2} \\ &\leqslant \left(1 + \frac{4D(\log n)^2}{1 + 2\sqrt{D}\log n}\right)\varphi(\sigma)\sqrt{n} \\ &\leqslant \left(1 + 2\sqrt{D}\log n\right)\varphi(\sigma)\sqrt{n}. \end{aligned}$$

Using equation (12), for all $z > 0$ and $x_{K,M} \geqslant \sigma_{K,M}$, with probability larger than $1 - e^{-z}$,

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

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 $2\theta + 4D(\log n)^2\theta^2 \leqslant \frac{\epsilon}{4C^*(n_*+k+1)}$. Then

$$W_{K,M} \leqslant 4C^*(n_* + k + 1)\left[\theta + \theta + 4D(\log n)^2\theta^2\right] \leqslant \epsilon$$

and

$$\begin{aligned} W_{K,M}x_{K,M}^2 &\leqslant 4C^*(n_* + k + 1)\left[\sigma_{K,M}x_{K,M} + \sqrt{\frac{z}{n}}x_{K,M} + 4D(\log n)^2\frac{z}{n}\right] \\ &\leqslant 4C^*(n_* + k + 1)\left[\theta x_{K,M}^2 + 4D(\log n)^2\frac{z}{n}\right] \\ &\leqslant 8C^*(n_* + k + 1)\left[\frac{1}{\theta}\sigma_{K,M}^2 + \left(4D(\log n)^2 + \frac{1}{\theta}\right)\frac{t}{n}\right]. \end{aligned}$$

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

$$\mathrm{pen}_n(K, M) \geqslant 8C^*(n_* + k + 1)\left[\frac{1}{\theta}\sigma_{K,M}^2 + \left(4D(\log n)^2 + \frac{1}{\theta}\right)\frac{w_M + K}{n}\right],$$

it holds

$$W_{K,M}x_{K,M}^2 - \mathrm{pen}_n(K, M) \leqslant 8C^*(n_* + k + 1)\left(4D(\log n)^2 + \frac{1}{\theta}\right)\frac{s}{n}.$$

A θ that satisfies $2\theta + 4D(\log n)^2\theta^2 = \frac{\epsilon}{4C^*(n_*+k+1)}$ is

$$\theta = \frac{1}{4D(\log n)^2}\left(\sqrt{1 + \frac{\epsilon D(\log n)^2}{C^*(n_* + k + 1)}} - 1\right).$$

Let us take this θ. Since $\frac{1}{\sqrt{1+x}-1} \leqslant \max(1, \frac{3}{x})$ for all $x > 0$,

$$\frac{1}{\theta} \leqslant 12C^*(n_* + k + 1)\max\left(\frac{D(\log n)^2}{3C^*(n_* + k + 1)}, \frac{1}{\epsilon}\right).$$

Therefore,

$$\begin{aligned} &W_{K,M}x_{K,M}^2 - \mathrm{pen}_n(K, M) \\ &\quad\leqslant 96(C^*)^2(n_* + k + 1)^2\left(\frac{D(\log n)^2}{3C^*(n_* + k + 1)} + \frac{1}{\epsilon} \vee \frac{D(\log n)^2}{3C^*(n_* + k + 1)}\right)\frac{s}{n} \\ &\quad\leqslant 192(C^*)^2(n_* + k + 1)^2\left(\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{3C^*(n_* + k + 1)}\right)\frac{s}{n} \end{aligned}$$

as soon as

$$\mathrm{pen}_n(K, M) \geqslant 96(C^*)^2(n_* + k + 1)^2\left(\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{3C^*(n_* + k + 1)}\right)\left(\sigma_{K,M}^2 + 2\frac{w_M + K}{n}\right).$$

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

Lemma 32 *Let A, B, C and E be functions $\mathbb{N} \longrightarrow [1,\infty)$, and $\varphi_n : x \longmapsto xA(n)(1+\sqrt{\log\max(\frac{B(n)}{x}, C(n))})$. Let σ_n be the only solution of the equation $\frac{\varphi_n(x)}{x^2\sqrt{n}} = \frac{1}{E(n)}$ 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

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

In our case,

$$\begin{cases} A(n) = \sqrt{2\pi(m_M K + K^2 - 1)}, \\ B(n) = 95De^{2D}(\sqrt{2}C_{\mathbf{Q}}\log n)^{k+3/2}\sqrt{kKC_{\text{aux}}{}'}, \\ C(n) = 14(\sqrt{2}C_{\mathbf{Q}}\log n)^{k+1/2}\sqrt{kKC_{\text{aux}}{}'}, \\ E(n) = 1 + 2\sqrt{D}\log n \leqslant 3\sqrt{D}\log n. \end{cases}$$

Hence

$$\begin{aligned} f(n) &\leqslant 18\pi\,(m_M K + K^2 - 1)D(\log n)^2\left(\frac{14}{95De^{2D}}\right)^2\left(1+\sqrt{\log B(n)+\log n}\right)^2 \\ &\leqslant \frac{4}{5}\pi\,(m_M K + K^2 - 1)(\log n)^2\frac{e^{-4D}}{D}\Bigg(1 + \log n + \log 95 + \log D + 2D + \\ &\qquad\qquad \left(k+\frac{3}{2}\right)\log(\sqrt{2}C_{\mathbf{Q}}\log n) + \frac{1}{2}\log(kKC_{\text{aux}}{}')\Bigg) \\ &\leqslant \frac{4}{5}\pi\,(m_M K + K^2 - 1)(\log n)^2\frac{e^{-4D}}{D}\left(15D + 2k\log\log n + \frac{1}{2}\log C_{\text{aux}}\right) \end{aligned}$$

when $\log n \geqslant \sqrt{2}C_{\mathbf{Q}} \geqslant 1$ by using that $1 \leqslant k, K \leqslant n$, $\log x \leqslant x$ for all $x \geqslant 0$, $D \geqslant \log n$ by assumption and $\log C_{\text{aux}}{}' \leqslant \log C_{\text{aux}} + D + \log K$. Thus,

$$f(n) \leqslant \tilde{f}_{K,M}(n) := 14\pi\,(m_M K + K^2 - 1)e^{-4D}(\log n)^2(k + \log C_{\text{aux}}).$$

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

$$\begin{aligned} \sigma_{K,M}^2 &\leqslant \frac{36\pi\,(m_M K + K^2 - 1)D(\log n)^2}{n}(1 + \log n + \log B) \\ &\leqslant \frac{36\pi\,(m_M K + K^2 - 1)D(\log n)^2}{n}\left(15D + 2k\log\log n + \frac{1}{2}\log C_{\text{aux}}\right). \end{aligned}$$

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

$$\begin{aligned} \text{pen}_n(K,M) \geqslant \frac{C_{\text{pen}}}{n}(n_* + k + 1)^2\left(\frac{1}{\epsilon} \vee \frac{D(\log n)^2}{3C^*(n_* + k + 1)}\right)\Big(w_M + \\ (m_M K + K^2 - 1)D(\log n)^2(D + k\log\log n + \log C_{\text{aux}})\Big). \end{aligned}$$