Estimation adaptative pour les modèles de Markov cachés non paramétriques

Luc Lehéricy

NNT : 2018SACLS550

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Laboratoire d'accueil : Laboratoire de mathématiques d'Orsay, UMR 8628 CNRS

Spécialité de doctorat : Mathématiques appliquées

Luc LEHÉRICY

Estimation adaptative pour les modèles de Markov cachés non paramétriques

Date de soutenance : 14 décembre 2018

Après avis des rapporteurs :
M. Yannick BARAUD (Université de Nice Sophia Antipolis)
M. Richard NICKL (University of Cambridge)

Jury de soutenance :

M. Yannick BARAUD	(Université de Nice Sophia Antipolis)	Rapporteur
M. Randal DOUC	(Telecom SudParis)	Examinateur
Mme Élisabeth GASSIAT	(Université Paris-Sud)	Directrice de thèse
M. Hajo HOLZMANN	(Philipps-Universität Marburg)	Examinateur
Mme Catherine MATIAS	(CNRS, Sorbonne Université)	Présidente
M. Richard NICKL	(University of Cambridge)	Rapporteur

REMERCIEMENTS

Je tiens tout d'abord à remercier ma directrice de thèse Élisabeth Gassiat pour le riche sujet de thèse qu'elle m'a proposé, ainsi que pour ses conseils avisés et sa disponibilité malgré de nombreuses contraintes. Sa bienveillance et la liberté qu'elle m'a laissée au cours de ma thèse furent une expérience précieuse pour laquelle je lui suis profondément reconnaissant.

I am very grateful to Yannick Baraud and Richard Nickl for reviewing my thesis in great detail. I also thank Randal Douc, Hajo Holzmann and Catherine Matias deeply for accepting to be part of my jury.

Ces trois ans au laboratoire de mathématiques d'Orsay furent une expérience inoubliable, en grande partie grâce aux si sympathiques collègues, permanents ou non, avec qui j'ai eu l'occasion d'interagir. Je tiens à remercier mes camarades doctorants, en particulier Augustin, Jeanne et Martin pour de nombreuses discussions enrichissantes, ainsi qu'Amine, Armand, Benjamin, Claire, Élodie, Gabriel, Guillaume, Hugo, Joseph, Margaux, Mor, Pierre, Romain, Solène, Thomas et Valérie. Je remercie également Claire Lacour, Sylvain Le Corff et Yohann de Castro pour leurs conseils et leur enthousiasme contagieux. Je suis reconnaissant à tous ceux qui m'ont aidé avec le sourire dans les nombreuses démarches administratives qui rythment la vie d'un doctorant : merci à Catherine Ardin, Corentin Guéneron, Valérie Lavigne, Marie-Christine Myoupo et Florence Rey, ainsi qu'à Stéphane Nonnenmacher et Frédéric Paulin.

Merci également à mes amis qui m'ont aidé à respirer quand la pression se faisait trop sentir. Merci à Alexis, François, Ismail, Julien, Marc, Paul, Sarah, Wei. Merci aussi à tous les membres d'Achor pour l'ambiance chaleureuse qui sait si bien égayer les mercredis soirs.

Et enfin, je remercie ma famille, mes parents, mes grand-parents et mon frère, sans qui rien de tout cela n'aurait été possible.

CONTENTS

NOTATIONS

Sets

- $\mathbb{N} := \{0, 1, 2, \dots\}$ is the set of nonnegative integers. $\mathbb{N}^* := \{1, 2, \dots\}$ is the set of positive integers, $\mathbb{Z} := \{\dots, -1, 0, 1, \dots\}$ the set of all integers.
- $\mathbb{R}$ is the set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers.
- When A is a finite set, we write $|A|$ or $\#A$ its cardinality.
- For all $K \in \mathbb{N}^*$, we write $[K] := \{1, \dots, K\}$.
- For all $K \in \mathbb{N}^*$, $\mathfrak{S}(K)$ is the set of permutations of $[K]$.
- $\mathrm{Span}(A)$ is the linear space spanned by the family A.
- When $(A, \mathcal{A}, \mu)$ is a measured space, we write $\mathbf{L}^2(A, \mathcal{A}, \mu)$ the Hilbert space of measurable and square integrable functions on A with respect to the measure μ. When the σ-field $\mathcal{A}$ is not ambiguous, we simply write $\mathbf{L}^2(A, \mu)$, and when the measure is not ambiguous, we write $\mathbf{L}^2(A)$.
- If E_1 and E_2 are two sets of functions, we denote $E_1 \otimes E_2$ their tensor product, that is the linear space spanned by the tensor products of their elements: $E_1 \otimes E_2 = \mathrm{Span}(f_1 \otimes f_2 : f_1 \in E_1, f_2 \in E_2)$ (see below for the definition of the tensor product of two functions).

Functions

- log is the natural logarithm.
- $\lfloor \cdot \rfloor$ is the floor function: for all $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer that is smaller or equal to x.
- $a \wedge b$ is the minimum of a and b, $a \vee b$ the maximum.
- For $x \in \mathbb{R}$, we write $x^+ = x \vee 0$;

- When f and g are two functions defined on F and G respectively with values in $\mathbb{R}$, we write $f \otimes g$ the function defined on $F \times G$ by $(f \otimes g)(x, y) = f(x)g(y)$.
- $\|\cdot\|_F$ is the Frobenius norm. We implicitely extend the definition of the Frobenius norm to tensors with more than two dimensions.
- When $A \in \mathbb{R}^{n \times p}$ is a $n \times p$ matrix, we write $\sigma_1(A) \geqslant \cdots \geqslant \sigma_{n \wedge p}(A)$ its singular values.
- Let $(u_n)_{n \in \mathbb{N}^*}$ and $(v_n)_{n \in \mathbb{N}^*}$ be two real-valued sequences such that v_n is nonnegative for n large enough. We write:
  - $u_n = o(v_n)$ when there exists a nonnegative sequence $w_n \longrightarrow 0$ such that $|u_n| \leqslant w_n v_n$ for n large enough,
  - $u_n = O(v_n)$ when there exists a constant $C > 0$ such that $|u_n| \leqslant C v_n$ for n large enough,
  - $u_n \asymp v_n$ when u_n is nonnegative for n large enough and $u_n = O(v_n)$ and $v_n = O(u_n)$.

Miscellaneous

- For $K \in \mathbb{N}^*$, Id_K is the identity matrix of size K.
- Wherever it is defined, Leb denotes the Lebesgue measure.
- For all $(Y_i)_{i \in \mathbb{Z}}$ and $(a, b) \in \mathbb{Z}^2$ with $a \leqslant b$, we write Y_a^b the uple $(Y_a, \ldots, Y_b)$. When $a > b$, the notation Y_a^b corresponds to an empty uple.
- When appropriate, we write p_X the density of the distribution of the random variable X with respect to the (previously chosen) dominating measure. Likewise, we write $p_{X|Z}$ the density of the distribution of X conditionally to Z. When the distribution depends on a parameter θ, we write it as a superscript as in $p^\theta_{X|Z}$.

CHAPTER

1

INTRODUCTION

1.1 Modèles de Markov cachés

Dans ma thèse, je me suis intéressé aux modèles de Markov cachés à espace d'états fini. Ces modèles ont été formalisés par Baum and Petrie (1966), qui ont également étudié leurs propriétés asymtotiques. Ces modèles et leurs généralisations ont depuis donné lieu à une abondante littérature, tant théorique qu'appliquée. Une revue exhaustive de toute la littérature à leur sujet dépasserait le cadre de cette thèse ; nous présenterons les résultats les plus significatifs pour les thèmes abordés ici. Le lecteur intéressé pourra par exemple consulter le livre de Cappé et al. (2005).

1.1.1 Description du modèle

Les modèles de Markov cachés (en anglais *hidden Markov model*, que nous abrégerons en HMM) sont un cas particulier de modèle à variable latente, qui peut se définir ainsi.

Définition 1.1 (Modèle à variable latente)**.** *Un modèle à variable latente est un processus bivarié* $(X_t, Y_t)_{t\in\mathbb{N}^*}$ *où seules les* observations $(Y_t)_{t\in\mathbb{N}^*}$ *sont accessibles. Les variables* $(X_t)_{t\in\mathbb{N}^*}$*, appelées* états cachés*, ne sont pas observées.*

Dans cette thèse, nous étudions essentiellement des modèles où les variables latentes prennent un nombre fini de valeurs. L'intérêt de tels modèles apparaît pleinement lorsque connaître les états cachés permet d'écrire la loi des observations $(Y_t)_t$ de manière simple. C'est le cas des modèles de mélange et des modèles de Markov cachés, que l'on définit comme suit.

Définition 1.2 (Modèle de mélange fini)**.** *Soient* $\mathcal{X}$ *un espace fini et* $(\mathcal{Y}, \mathcal{B})$ *un espace mesurable. Soit* μ *une mesure de probabilité sur* $\mathcal{X}$ *et* $\nu = (\nu_x)_{x\in\mathcal{X}}$ *un vecteur de mesures de probabilité sur* $\mathcal{Y}$*.*

On dit que le processus $(Y_t)_{t\in\mathbb{N}^*}$ *est généré par un modèle de mélange de paramètres* $(\mathcal{X}, \mu, \nu)$ *s'il existe des variables aléatoires* $(X_t)_{t\in\mathbb{N}^*}$ *telles que*

(i) les variables $(X_t)_{t\in\mathbb{N}^*}$ *sont indépendantes et identiquement distribuées (i.i.d.) de loi* μ *;*

(ii) conditionnellement à $(X_t)_{t\in\mathbb{N}^*}$, les variables $(Y_t)_{t\in\mathbb{N}^*}$ sont indépendantes ;

(iii) pour tout $t \in \mathbb{N}^*$, conditionnellement à $\{X_t = x\}$, Y_t suit la loi ν_x.

Remarque. *Dans ce modèle, les variables Y_t sont i.i.d. de loi $\sum_{x\in\mathcal{X}} \mu_x\nu_x$.*

Les modèles de mélange apparaissent naturellement lorsque les observations proviennent de plusieurs groupes ayant chacun des caractéristiques propres. La variable latente identifie alors le groupe de l'individu. Ce cadre est très utilisé pour la classification de données.

Une autre application des modèles de mélange est de pouvoir approcher des lois complexes à l'aide d'un petit nombre de composantes élémentaires. C'est le cas des mélanges de lois normales, voir par exemple les travaux de Kruijer et al. (2010) et Maugis-Rabusseau and Michel (2013).

Les modèles de mélange ne permettent cependant pas de rendre compte d'une dépendance entre les observations. La manière la plus simple de le faire est de supposer que les états cachés, au lieu d'être i.i.d., sont une chaîne de Markov : c'est le principe des modèles de Markov cachés.

Définition 1.3 (Modèle de Markov caché)**.** *Soit $\mathcal{X}$ un espace fini et $(\mathcal{Y}, \mathcal{B})$ un espace mesuré. Soit π une mesure de probabilité sur $\mathcal{X}$, Q un noyau de transition de $\mathcal{X}$ dans $\mathcal{X}$ et $\nu = (\nu_x)_{x\in\mathcal{X}}$ un vecteur de mesures de probabilité sur $\mathcal{Y}$.*

On dit que le processus $(Y_t)_{t\in\mathbb{N}^*}$ est généré par un modèle de Markov caché de paramètres $(\mathcal{X}, \pi, Q, \nu)$ s'il existe des variables aléatoires $(X_t)_{t\in\mathbb{N}^*}$ telles que

(i) le processus $(X_t)_{t\in\mathbb{N}^*}$ est une chaîne de Markov à valeurs dans $\mathcal{X}$ de loi initiale π et de noyau de transition Q ;

(ii) conditionnellement à $(X_t)_{t\in\mathbb{N}^*}$, les variables $(Y_t)_{t\in\mathbb{N}^*}$ sont indépendantes ;

(iii) pour tout $t \in \mathbb{N}^*$, conditionnellement à $\{X_t = x\}$, Y_t suit la loi ν_x.

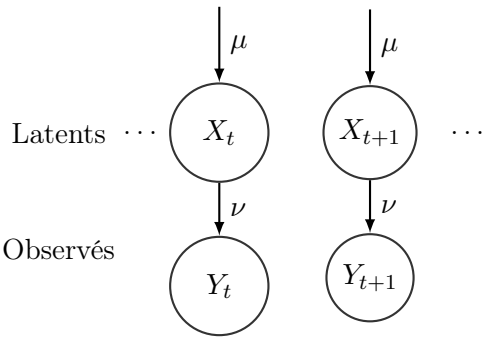
$\mathcal{X}$ est appelé l'espace des états *du modèle de Markov caché (*state space *en anglais) et $\mathcal{Y}$* l'espace des observations *(*observation space*). Enfin, les lois de probabilité $(\nu_x)_{x\in\mathcal{X}}$ sont appelées* lois d'émission.

Une remarque importante à ce stade est que contrairement au modèle de mélange, il n'est pas possible d'écrire la loi du processus $(Y_t)_{t\in\mathbb{N}^*}$ de manière simple sans faire intervenir les variables $(X_t)_{t\in\mathbb{N}^*}$. De la même manière, bien que le processus bivarié $(X_t, Y_t)_{t\in\mathbb{N}^*}$ soit une chaîne de Markov, ce n'est pas le cas des observations $(Y_t)_{t\in\mathbb{N}^*}$. Le fait de ne pas observer les états cachés donne ainsi une grande souplesse aux modèles de Markov cachés en leur permettant de rendre compte de dépendances complexes.

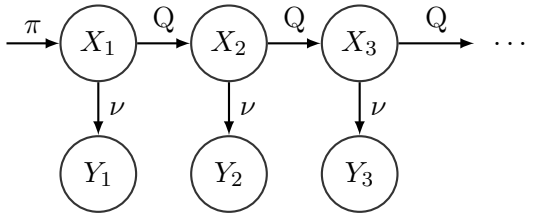
La proposition suivante illustre comment calculer la loi des observations d'un modèle de Markov caché.

Proposition 1.1. *Soit $(Y_t)_{t\in\mathbb{N}^*}$ un processus à valeurs dans $\mathcal{Y}$ généré par un modèle de Markov caché de paramètres $(\mathcal{X}, \pi, Q, \nu)$. Soit λ une mesure sur $\mathcal{Y}$. Supposons que les lois d'émission $(\nu_x)_{x\in\mathcal{X}}$ sont absolument continues par rapport à λ, de densités $(\gamma_x)_{x\in\mathcal{X}}$. Alors pour tout $n \in \mathbb{N}^*$, la loi de $(Y_1, \dots, Y_n)$ admet une densité par rapport à $\lambda^{\otimes n}$ qui s'écrit*

$$p_{Y_1^n}^{(\mathcal{X},\pi,Q,\gamma)}(y_1^n) = \sum_{x_1^n\in\mathcal{X}^n} \pi_{x_1} Q_{x_1,x_2} \dots Q_{x_{n-1},x_n} \prod_{i=1}^{n} \gamma_{x_i}(y_i).$$



(a) Modèle de mélange de paramètres $(\mathcal{X}, \mu, \nu)$.



(b) Modèle de Markov caché de paramètres $(\mathcal{X}, \pi, Q, \nu)$.

Figure 1.1: Représentation sous forme de graphe orienté acyclique des modèles de mélange et de Markov caché. L'étiquette des arêtes correspond au noyau de transition.

Preuve. On utilise pour cela la structure de dépendance du modèle. En conditionnant par les états cachés :

$$\begin{aligned} p_{Y_1^n}^{(\mathcal{X},\pi,Q,\nu)}(y_1^n) &= \sum_{x_1^n \in \mathcal{X}^n} p_{X_1^n}^{(\mathcal{X},\pi,Q,\nu)}(x_1^n)\, p_{Y_1^n|X_1^n}^{(\mathcal{X},\pi,Q,\nu)}(y_1^n|x_1^n) \\ &= \sum_{x_1^n \in \mathcal{X}^n} p_{X_1^n}^{(\mathcal{X},\pi,Q,\nu)}(x_1^n) \prod_{i=1}^{n} p_{Y_i|X_1^n}^{(\mathcal{X},\pi,Q,\nu)}(y_i|x_1^n) \\ &= \sum_{x_1^n \in \mathcal{X}^n} p_{X_1^n}^{(\mathcal{X},\pi,Q,\nu)}(x_1^n) \prod_{i=1}^{n} \gamma_{x_i}(y_i). \end{aligned}$$

Ensuite, on utilise que $(X_t)_{t\in\mathbb{N}^*}$ est une chaîne de Markov de paramètres (π, Q) pour développer la densité de X_1^n. □

Définition 1.4 (Ordre d'un modèle de Markov caché)**.** *On dit que le processus $(Y_t)_{t\in\mathbb{N}^*}$ est généré par un modèle de Markov caché (à espace d'états fini) s'il existe des paramètres $(\mathcal{X}, \pi, Q, \nu)$ avec $\mathcal{X}$ fini tel que ce processus est généré par un modèle de Markov caché de paramètres $(\mathcal{X}, \pi, Q, \nu)$. On appelle* ordre *du processus $(Y_t)_{t\in\mathbb{N}^*}$ le plus petit cardinal possible de $\mathcal{X}$ pour l'ensemble des paramètres $(\mathcal{X}, \pi, Q, \nu)$ qui génèrent ce processus.*

L'ordre ainsi défini existe et est unique. Par contre, il n'existe pas toujours une seule paramétrisation qui le réalise.

Remarque. *Dans la littérature, la notion d'ordre peut également désigner l'ordre de dépendance de la chaîne de Markov, c'est-à-dire le nombre de variables passées dont dépend la variable présente. Cet ordre vaut 1 dans les chaînes de Markov usuelles : la variable actuelle ne dépend que de la précédente. Dans toute la suite, nous supposerons l'ordre de dépendance égal à 1, et le terme « ordre » désignera toujours le plus petit cardinal possible de l'espace des états.*

1.1.2 Unicité des paramètres

Une question naturelle à ce stade est : étant donné un processus observé $(Y_t)_{t\in\mathbb{N}^*}$, est-il possible de définir « le » HMM qui le génère ? Autrement dit, sous quelle condition peut-on garantir

l'unicité, en un certain sens, des paramètres qui génèrent ce processus ? Si oui, on dit que le modèle est *identifiable*.

Tout d'abord, notons que la loi des observations est invariante par permutation des états cachés, autrement dit par une permutation de l'espace des états. En notant $\sigma : \mathcal{X} \to \mathcal{X}'$ cette transformation, cela correspond à remplacer les paramètres $(\mathcal{X}, \pi, Q, \nu)$ du HMM par les paramètres $(\mathcal{X}', \pi', Q', \nu')$ définis par

$$\forall (x, x') \in \mathcal{X}^2, \quad \begin{cases} \pi'_{\sigma(x)} = \pi_x, \\ Q'_{\sigma(x),\sigma(x')} = Q_{x,x'}, \\ \nu'_{\sigma(x)} = \nu_x. \end{cases}$$

Cette transformation revient à changer « l'étiquette » des états cachés (d'où l'appellation *relabelling* en anglais), mais ne change essentiellement rien au HMM. Lorsque l'espace des états $\mathcal{X}$ est de cardinal K, cela signifie qu'on peut toujours l'identifier à $[K] := \{1, \dots, K\}$. Cela signifie aussi que tous les résultats d'identifiabilité seront à permutation des états cachés près, du moins en l'absence de condition permettant de définir une permutation canonique. Dans la suite, lorsque $|\mathcal{X}| = K$, on écrira (K, π, Q, ν) au lieu de $(\mathcal{X}, \pi, Q, \nu)$.

Des travaux récents de Gassiat et al. (2015) et Alexandrovich et al. (2016) montrent qu'il suffit de connaître la loi d'un certain nombre d'observations consécutives pour retrouver les paramètres à permutation des états cachés près.

Dans les théorèmes qui suivent, on note $\mathbb{P}^{(K,\pi,Q,\nu)}_{Y_a^b}$ la loi du vecteur $Y_a^b := (Y_a, \dots, Y_b)$ lorsque le processus $(Y_t)_{t\in\mathbb{N}^*}$ est généré par un modèle de Markov de paramètres (K, π, Q, ν).

Le premier théorème est adapté de Gassiat et al. (2015) et montre que lorsque les lois d'émissions sont libres, les paramètres du modèle sont caractérisés par la loi de trois observations consécutives. Noter que cet article suppose le nombre d'états connus, ce qui n'est en fait pas nécessaire : sous les hypothèses du théorème, l'ordre du modèle est identifiable à partir de deux observations consécutives. Cette propriété sera développée et utilisée dans le Chapitre 2 portant sur l'estimation de l'ordre.

Théorème 1.2 (Identifiabilité, lois d'émission libres)**.** *Supposons que $\mathcal{Y}$ est un espace polonais (c'est-à-dire métrique, séparable et complet).*

Soient $K \in \mathbb{N}^*$ un entier naturel, $\pi = (\pi_x)_{x\in[K]}$ une loi de probabilité sur $[K]$, Q une matrice de transition de taille K et $\nu = (\nu_x)_{x\in[K]}$ un vecteur de mesures de probabilité.

Supposons que Q est inversible, que $\pi_x > 0$ pour tout $x \in [K]$ et que la famille $(\nu_x)_{x\in[K]}$ est libre. Alors les paramètres K, π, Q et ν sont identifiables à partir de la loi de (Y_1, Y_2, Y_3), autrement dit : pour tout entier $K' \in \mathbb{N}^*$, pour toute loi de probabilité π' sur $[K']$, pour toute matrice de transition Q' de taille K' et pour tout vecteur de mesures de probabilité ν' de taille K', on a

$$\left(\mathbb{P}^{(K,\pi,Q,\nu)}_{Y_1^3} = \mathbb{P}^{(K',\pi',Q',\nu')}_{Y_1^3} \text{ et } K' \leqslant K\right) \iff (K, \pi, Q, \nu) = (K', \pi', Q', \nu') \text{ à permutation des états cachés près.}$$

Quelques remarques sur ce théorème :

- Cela assure en particulier que K est bien l'ordre du processus généré par ce HMM, et qu'il n'existe qu'une seule paramétrisation possible, à permutation des états cachés près, qui réalise l'ordre.
- Aucune condition n'est requise sur les paramètres alternatifs (K', π', Q', ν'), à part $K' \leqslant K$. En particulier, la matrice Q' peut ne pas être inversible.

- La condition $K' \leqslant K$ est essentielle : il y a une infinité de manières d'ajouter un état à un modèle de Markov caché sans changer sa loi. Une façon de le faire est de s'assurer que la probabilité de transition vers ce nouvel état est nulle.

Le théorème suivant a été démontré par Alexandrovich et al. (2016). Il permet de remplacer l'hypothèse de liberté des lois d'émission par l'hypothèse qu'elles sont seulement deux à deux distinctes, au prix de devoir prendre plus d'observations en compte.

Théorème 1.3 (Identifiabilité, lois d'émission distinctes)**.** *Supposons que $\mathcal{Y}$ est inclus dans un espace euclidien (c'est-à-dire dans $\mathbb{R}^q$ pour un certain entier q).*

Soient $K \in \mathbb{N}^*$ un entier naturel, $\pi = (\pi_x)_{x\in[K]}$ une loi de probabilité sur $[K]$, Q une matrice de transition de taille K et $\nu = (\nu_x)_{x\in[K]}$ un vecteur de mesures de probabilité.

Supposons que Q est inversible et ergodique (c'est-à-dire irréductible et apériodique) et que les lois d'émission $(\nu_x)_{x\in[K]}$ sont deux à deux distinctes. Alors les paramètres K, π, Q et ν sont identifiables à partir de la loi de $(Y_1, \dots, Y_{(2K+1)(K^2-2K+2)+1})$, autrement dit : pour tout entier $K' \in \mathbb{N}^*$, pour toute loi de probabilité π' sur $[K']$, pour toute matrice de transition Q' de taille K' et pour tout vecteur de mesures de probabilité ν' de taille K', on a

$$\left(\mathbb{P}^{(K,\pi,Q,\nu)}_{Y_1^{(2K+1)(K^2-2K+2)+1}} = \mathbb{P}^{(K',\pi',Q',\nu')}_{Y_1^{(2K+1)(K^2-2K+2)+1}} \text{ et } K' \leqslant K\right)$$
$$\iff \quad (K,\pi,Q,\nu) = (K',\pi',Q',\nu') \text{ à permutation des états cachés près.}$$

Quelques remarques sur ce théorème :

- Le nombre d'observations nécessaire dépend de l'ordre du processus généré par ce HMM.
- En général, la condition d'ergodicité de Q est peu contraignante : on la suppose par ailleurs pour s'assurer que le processus généré par ce HMM est lui-même ergodique.

Lorsque le processus observé $(Y_t)_{t\in\mathbb{N}^*}$ est généré par un modèle de Markov caché dont les paramètres vérifient l'un de ces deux théorèmes, on a donc unicité de ces paramètres. Cela permet d'identifier le processus observé et le HMM ayant ces paramètres. On parle alors d'*ordre du HMM* (ou alternativement de son *nombre d'états cachés* ou *nombre d'états*) et de *paramètres du HMM* (au lieu de « paramètres de l'unique HMM générant ce processus et vérifiant les conditions d'identifiabilité »).

1.2 Cadre d'étude

1.2.1 Problématique

L'objectif de cette thèse peut se résumer ainsi :

> **Problématique :** Prouver qu'on peut estimer, dans un cadre général et de façon optimale et robuste, tous les paramètres d'un modèle de Markov caché.

Développons les différents aspects de cette problématique.

Cadre général. L'objectif est ici de faire le moins d'hypothèses préalables sur les paramètres du HMM. Cela implique notamment :

- Pas de borne a priori sur l'ordre du modèle. Autrement dit, on ne suppose pas avoir accès à un entier K_0 connu tel que l'ordre K du HMM vérifie $K \leqslant K_0$. Cette hypothèse est requise dans la plupart des travaux concernant l'estimation de l'ordre.

- Pas de restriction paramétrique sur les lois d'émission. Cela signifie qu'il n'existe pas d'application connue à l'avance de $\mathbb{R}^d$ (pour un certain entier d) dans l'ensemble des lois de probabilité sur $\mathcal{Y}$ dont l'image contienne les lois d'émissions du HMM. Par exemple, cela exclut le cas où on sait les lois d'émission gaussiennes, car celles-ci sont alors caractérisées par leur moyenne et leur variance. Si les modèles paramétriques sont en général plus simples à manipuler, il peut être difficile d'en trouver un suffisamment flexible. Les modèles non paramétriques permettent de s'affranchir de cette difficulté.

  Les résultats sur les modèles de Markov cachés non paramétriques sont récents : citons par exemple l'article de de Castro et al. (2016), qui ont pour la première fois montré qu'il était possible d'estimer les lois d'émission d'un HMM de manière minimax en non paramétrique pour un estimateur des moindres carrés.

Optimalité. Dans toute la suite, on suppose que les lois d'émission admettent une densité, qu'on appelle *densité d'émission*, par rapport à une mesure λ connue. On note $\gamma = (\gamma_x)_{x\in\mathcal{X}}$ le vecteur des densités d'émission. Notre but est de montrer que les estimateurs de γ convergent à la « meilleure » vitesse possible. La théorie minimax fournit une définition naturelle de cette « meilleure » vitesse.

Le risque minimax se définit ainsi. Soit $\{\mathbb{P}_f : f \in \mathcal{F}\}$ un ensemble de lois de probabilité associées à un processus aléatoire $(Z_t)_{t\in\mathbb{N}^*}$, tel que $\mathbb{P}_f$ est la loi de $(Z_t)_{t\in\mathbb{N}^*}$ sous le paramètre f. Soit R un risque sur $\mathcal{F}$, c'est-à-dire une fonction $\mathcal{F}\times\mathcal{F} \longrightarrow \mathbb{R}_+$. R peut être vue comme une mesure de la distance entre deux paramètres. Soit $\mathcal{G} \subset \mathcal{F}$ un sous-ensemble des paramètres, et $\mathcal{E}_n$ l'ensemble des estimateurs fondés sur les n premières observations, c'est-à-dire l'ensemble des fonctions à valeurs dans $\mathcal{F}$ mesurables pour la tribu engendrée par $(Z_1, \dots, Z_n)$.

En notant $\mathbb{E}_f$ l'espérance sous le paramètre f, le risque d'un estimateur $\hat{f}_n \in \mathcal{E}_n$ sur la classe $\mathcal{G}$ est défini comme le pire risque sur la classe $\mathcal{G}$, autrement dit

$$\bar{R}_n(\hat{f}_n) := \sup_{f\in\mathcal{G}} \mathbb{E}_f\left[R(\hat{f}_n, f)\right].$$

Le risque minimax pour le risque R, la classe $\mathcal{G}$ et n observations est alors le plus petit risque sur $\mathcal{G}$ pour tous les estimateurs possibles :

$$\bar{R}_n^{\text{minimax}} := \inf_{\hat{f}_n\in\mathcal{E}_n} \sup_{f\in\mathcal{G}} \mathbb{E}_f\left[R(\hat{f}_n, f)\right].$$

En particulier, tout estimateur a nécessairement un risque supérieur sur au moins un élément de $\mathcal{G}$, quelle que soit la manière dont cet estimateur est construit. Pour qu'un estimateur soit « bon » , il faut qu'il ait un petit risque uniformément sur la classe $\mathcal{G}$. On dit qu'une suite d'estimateurs $(\hat{f}_n)_{n\in\mathbb{N}^*}$ est *minimax* ou *converge à vitesse minimax* sur la classe $\mathcal{G}$ si

$$\bar{R}_n(\hat{f}_n) \asymp \bar{R}_n^{\text{minimax}}.$$

Si un estimateur est minimax sur la classe $\mathcal{G}$, il est donc impossible de faire mieux, sauf à constante multiplicative près.

Dans la suite, nous nous intéresserons surtout à l'estimation de densités dans $\mathbf{L}^2(\mathcal{Y}, \lambda)$ avec pour risque la norme $\mathbf{L}^2$, c'est-à-dire $R(f, f') = \|f - f'\|_2^2$. Pour l'estimation de la densité de variables i.i.d. à valeurs dans $\mathcal{Y} \subset \mathbb{R}^d$, un certain nombre de vitesses minimax sont connues, voir par exemple l'article de Goldenshluger and Lepski (2014) et les références qu'il contient. Un cas particulier emblématique est celui des classes de Hölder lorsque $\mathcal{Y} = [0, 1]$.

Soient s et L deux réels strictement positifs, et soit j le plus grand entier strictement plus petit que s. Alors en notant $\mathcal{C}^j([0, 1])$ la classe des fonctions de $[0, 1]$ dans $\mathbb{R}$ j fois dérivables de

dérivée j-ième continue, la classe de Hölder $\mathcal{H}(s, L)$ est définie par

$$\mathcal{H}(s,L) := \Big\{ f \in \mathcal{C}^j([0,1]) \text{ t.q. } \forall 0 \leqslant i \leqslant j, \|f^{(i)}\|_\infty \leqslant L \\ \text{et } \forall x, y \in [0,1], |f^{(j)}(x) - f^{(j)}(y)| \leqslant L|x-y|^{s-j} \Big\}.$$

Le paramètre s est appelé *régularité* de ces fonctions.

Soit $\mathcal{D}([0,1])$ l'ensemble des densités de probabilité sur $[0,1]$ par rapport à la mesure de Lebesgue. Pour tout $f \in \mathcal{D}([0,1])$, on se donne une loi de probabilité $\mathbb{P}_f$ et des variables aléatoires $Z_1, \ldots, Z_n$ telles que les $(Z_t)_{1 \leqslant t \leqslant n}$ sont i.i.d. de loi de densité f par rapport à la mesure de Lebesgue sous $\mathbb{P}_f$. Alors on a

$$\inf_{\hat{f}_n \in \mathcal{E}_n} \sup_{f \in \mathcal{D}([0,1]) \cap \mathcal{H}(s,L)} \mathbb{E}_f \left[\|\hat{f}_n - f\|_2^2 \right] \asymp n^{\frac{-2s}{2s+1}}.$$

Autrement dit, le risque minimax sur la classe de Hölder $\mathcal{H}(s, L)$ sur un compact de $\mathbb{R}$ pour la norme 2 est de l'ordre de $n^{-2s/(2s+1)}$. C'est ce risque qu'on cherchera à atteindre *à un facteur logarithmique près*, c'est-à-dire qu'on cherchera à avoir un risque dominé par $n^{-2s/(2s+1)}(\log n)^a$ pour une constante $a > 0$.

Noter que ce risque minimax reste pertinent pour les modèles de Markov cachés : les variables i.i.d. sont un cas particulier de HMM (à un seul état caché).

Adaptativité. Une nuance importante à garder en tête dans la définition précédente est que si un estimateur est minimax, il l'est pour une classe de paramètres bien précise (par exemple la classe de Hölder $\mathcal{H}(s, L)$). Cela ne garantit pas qu'il soit minimax pour une autre classe de paramètres.

Certains estimateurs sont minimax, à un facteur logarithmique près, pour plusieurs classes de paramètres. On peut en distinguer deux types :

- Ceux qui utilisent la régularité dans leur construction, autrement dit qui savent à l'avance dans quelle classe de paramètres se trouve le paramètre à estimer. Ces estimateurs sont dits *minimax non adaptatifs* puisqu'ils ne peuvent pas s'adapter à la régularité du paramètre.

- Ceux qui ne nécessitent pas de connaître la régularité. Ces estimateurs sont dits *minimax et adaptatifs*. Comme la régularité du paramètre cible n'est en général pas accessible, c'est ce type d'estimateurs qu'on cherche à construire.

Tous les paramètres. Le premier paramètre à estimer est l'ordre du HMM. Cette étape est cruciale pour au moins deux raisons.

La première est que c'est un prérequis à la convergence des paramètres : si le nombre d'états estimé est strictement plus grand que l'ordre, il n'y a plus unicité des paramètres qui génèrent le processus observé ; si le nombre d'états est strictement plus petit que l'ordre, il n'est plus possible de retrouver la loi du processus.

La seconde est que dans certaines applications, l'ordre est en lui-même un paramètre d'intérêt. Les modèles de Markov cachés sont très appréciés pour leur capacité à expliquer un phénomène comme un mélange de différents comportements spécifiques qui se succèdent au cours du temps. Par exemple, un oiseau marin se déplace différemment selon qu'il vole, qu'il pêche ou qu'il dorme. Son déplacement peut donc se modéliser comme un HMM dont l'état caché sera son type de comportement à l'instant considéré. Estimer le nombre de types de comportement permet alors de mieux comprendre le mode de vie de l'oiseau et d'interpréter et expliquer ces comportements.

Les autres paramètres à estimer sont la loi initiale de la chaîne cachée, sa matrice de transition et les lois d'émission.

Il est impossible de retrouver exactement la loi initiale : en n'ayant accès qu'à une seule réalisation du processus des observations $(Y_t)_{t\in\mathbb{N}^*}$, le mieux qu'on puisse espérer retrouver est de quel état est partie la chaîne cachée. En général, même cette information ne peut être retrouvée exactement. Pour contourner ce problème, on peut se fixer une loi initiale (par exemple la loi uniforme ou un Dirac en un des états) ou considérer que la loi initiale est une loi invariante de la chaîne de Markov.

Robustesse. Dans la pratique, il arrive que les modèles de Markov cachés ne puissent pas rendre compte de toute la complexité du processus. Il est donc important d'étudier comment les estimateurs issus de ces modèles se comportent quand on considère des processus qui ne sont pas générés par un HMM ou lorsqu'on enrichit les modèles pour rendre compte d'autres aspects de la loi des observations.

La première approche part du principe qu'aucun processus réel ne peut être parfaitement décrit par un HMM. On dit alors que le modèle est *mal spécifié*. Même s'il n'est pas possible de retrouver exactement la loi du processus, un estimateur robuste est capable de trouver celle qui l'approche au mieux dans le modèle. C'est le cadre étudié dans le Chapitre 4.

La deuxième approche consiste à étudier des extensions des HMMs. Dans le Chapitre 5, nous étudions une généralisation non homogène des HMMs, c'est-à-dire dont la loi dépend du temps : les modèles de Markov cachés avec tendances. Ce sont des processus $(X_t, Z_t, Y_t)_{t\in\mathbb{N}^*}$ à valeurs dans $\mathcal{X}\times\mathbb{R}^d\times\mathbb{R}^d$ tels que seul le processus $(Y_t)_{t\in\mathbb{N}^*}$ est observé et tels qu'il existe un vecteur de fonctions $(T_x)_{x\in\mathcal{X}}$ de $\mathbb{N}^*$ dans $\mathbb{R}^d$ tel que $(X_t, Z_t)_{t\in\mathbb{N}^*}$ est un modèle de Markov caché et

$$Y_t = T_{X_t}(t) + Z_t.$$

Chaque « état » dérive donc avec sa propre tendance. Cela ajoute un paramètre au modèle, qu'il faut également estimer : le vecteur des tendances $(T_x)_{x\in\mathcal{X}}$.

1.2.2 Sélection de modèle

Quelle que soit la méthode considérée, il n'est en général pas possible de construire directement un estimateur adaptatif dans le cadre non paramétrique. La sélection de modèle permet de résoudre ce problème. Elle consiste à se donner une famille de modèles paramétriques–pour lesquels il est possible de calculer un estimateur–puis de sélectionner le « meilleur » modèle, en un sens à préciser.

Pour donner une idée de son fonctionnement, considérons le problème de l'estimation d'une densité de probabilité f dans $\mathbf{L}^2(\mathcal{Y},\lambda)$, avec comme risque la norme 2. Soit $(\varphi_a)_{a\in\mathbb{N}^*}$ une base orthonormée de $\mathbf{L}^2(\mathcal{Y},\mu)$, et donnons-nous comme modèles les espaces $\mathfrak{P}_M$ engendrés par les M premiers vecteurs de cette base pour tout $M\in\mathbb{N}^*$. Soient $(Z_t)_{t\in\mathbb{N}^*}$ des variables i.i.d. de loi de densité f par rapport à λ. Un estimateur naturel dans le modèle $\mathfrak{P}_M$ est donné par

$$\hat{f}_{M,n} = \sum_{a=1}^{M}\left(\frac{1}{n}\sum_{t=1}^{n}\varphi_a(Z_t)\right)\varphi_a.$$

Comme $\frac{1}{n}\sum_{t=1}^{n}\varphi_a(Z_t)$ converge vers $\langle f,\varphi_a\rangle$ par la loi des grands nombres et comme $(\varphi_a)_{a\in\mathbb{N}^*}$ est une base orthonormée de $\mathbf{L}^2(\mathcal{Y},\mu)$, cela assure que $\hat{f}_{M,n}$ converge vers la projection de f sur $\mathfrak{P}_M$. Notons f_M cette projection.

L'erreur totale de l'estimateur $\hat{f}_{M,n}$ est donnée par

$$\begin{aligned}\|\hat{f}_{M,n} - f\|_2^2 &= \|\hat{f}_{M,n} - f_M + f_M - f\|_2^2 \\ &= \underbrace{\|\hat{f}_{M,n} - f_M\|_2^2}_{\text{variance}} + \underbrace{\|f_M - f\|_2^2}_{\text{biais}}\end{aligned}$$

d'après le théorème de Pythagore.

Nous voyons apparaître deux termes au comportement opposé. Le premier est un terme de *variance*, ou *erreur d'estimation*, qui croît quand la dimension du modèle croît : plus un modèle est grand, plus il est difficile de trouver le meilleur paramètre à l'intérieur, et plus le bruit se fait sentir. Le second est un terme de *biais*, ou *erreur d'approximation*, qui décroît quand la dimension du modèle croît. Ici, c'est la distance entre la vraie densité et le modèle : plus le modèle grandit, plus cette distance est petite.

La clé de la sélection de modèle est de choisir un modèle $\hat{M}_n$ telle que l'erreur totale de l'estimateur correspondant, c'est-à-dire la somme de ces deux termes, soit la plus faible possible. On parle de *compromis biais-variance*. Nous utiliserons la pénalisation pour effectuer ce compromis.

Pénalisation et inégalité oracle. Deux des estimateurs étudiés dans cette thèse peuvent se résumer comme suit. Soit $(Y_1, \dots, Y_n)$ un n-échantillon. Considérons un espace $\mathcal{F}$ et une famille de modèles $(\mathfrak{P}_M)_M$ telle que $\mathfrak{P}_M \subset \mathcal{F}$ pour tout M. Enfin, soient $R_n : \mathcal{F} \longrightarrow \mathbb{R}_+$ et $R : \mathcal{F} \longrightarrow \mathbb{R}_+$ deux fonctions de risque. La première fonction de risque, le *risque empirique*, est celle qu'on peut calculer à partir des observations. La seconde, le *risque théorique*, nécessite de connaître la loi du processus et ne peut donc être utilisée pour l'estimation. On construit l'estimateur avec le risque empirique et on veut que son risque théorique soit petit.

Considérons une famille d'estimateurs $(\hat{f}_M)_M$ définie par : pour tout M,

$$\hat{f}_M \in \operatorname*{arg\,min}_{f \in \mathfrak{P}_M} R_n(f).$$

La sélection de modèle par pénalisation consiste à sélectionner le modèle $\hat{M}$ qui minimise le risque pénalisé

$$M \longmapsto R_n(\hat{f}_M) + \operatorname{pen}(M)$$

pour une pénalité $M \longmapsto \operatorname{pen}(M)$ bien choisie. L'idée est de pénaliser les gros modèles qui ont tendance à surapprendre.

Le modèle $\hat{M}$ choisi vérifiera alors par définition : pour tout M et tout $f \in \mathfrak{P}_M$

$$\begin{aligned}R_n(\hat{f}_{\hat{M}}) + \operatorname{pen}(\hat{M}) &\leqslant R_n(\hat{f}_M) + \operatorname{pen}(M) \\ &\leqslant R_n(f) + \operatorname{pen}(M),\end{aligned}$$

donc pour tout M et tout $f \in \mathfrak{P}_M$

$$R(\hat{f}_{\hat{M}}) + (R_n - R)(\hat{f}_{\hat{M}}) + \operatorname{pen}(\hat{M}) \leqslant R(f) + (R_n - R)(f) + \operatorname{pen}(M).$$

Le « bon » choix de pénalité est celui pour lequel $\operatorname{pen}(M) \geqslant |(R_n - R)(f)|$ pour tout $f \in \mathfrak{P}_M$ et pour tout M, ce qui permet d'obtenir l'inégalité oracle

$$R(\hat{f}_{\hat{M}}) \leqslant \inf_M \left\{ \inf_{f \in \mathfrak{P}_M} R(f) + 2\operatorname{pen}(M) \right\}.$$

On voit à nouveau apparaître un compromis biais-variance : le terme de biais est $\inf_{f \in \mathfrak{P}_M} R(f)$ et le terme de variance est contrôlé par la pénalité. En pratique, le contrôle de $(R_n - R)$ est

plus astucieux, et les inégalités que nous obtenons sont de la forme : lorsque la pénalité est plus grande qu'une borne explicite, on a avec grande probabilité

$$R(\hat{f}_{\hat{M}}) \leqslant C \inf_{M} \left\{ \inf_{f \in \mathfrak{P}_M} R(f) + 2\,\mathrm{pen}_n(M) \right\} + r_n$$

où

- $C \geqslant 1$ est une constante, idéalement proche de 1 ;
- r_n est un terme résiduel déterministe qui tend vers 0 quand n tend vers l'infini.

Les inégalités oracle garantissent donc que l'estimateur final $\hat{f}_{\hat{M}}$ réalise un compromis biais-variance. Lorsque la pénalité est du même ordre que la variance, cela assure qu'il fait aussi bien que le meilleur estimateur de la famille à constante et à terme résiduel près.

Noter que le terme de droite de l'inégalité oracle est déterministe, et peut être calculé grâce à la théorie de l'approximation pour certaines familles de modèles $(\mathfrak{P}_M)_M$. C'est ce que nous utiliserons pour montrer que les estimateurs sont minimax et adaptatifs.

1.2.3 Estimateurs

Dans cette section, nous donnons un aperçu des trois types d'estimateurs étudiés dans cette thèse : l'estimateur du maximum de vraisemblance, l'estimateur des moindres carrés et l'estimateur spectral.

Maximum de vraisemblance

Un estimateur du maximum de vraisemblance (abrégé EMV, ou MLE pour *Maximum Likelihood Estimator* en anglais) est un paramètre qui maximise la vraisemblance des observations, c'est-à-dire la fonction

$$(K, \pi, Q, \gamma) \longmapsto p_{Y_1^n}^{(K,\pi,Q,\gamma)}(y_1^n),$$

ou de manière équivalente la log-vraisemblance

$$l_n : (K, \pi, Q, \gamma) \longmapsto \log p_{Y_1^n}^{(K,\pi,Q,\gamma)}(y_1^n),$$

où y_1^n est la valeur observée du vecteur de variables aléatoire Y_1^n et $p_{Y_1^n}^{(K,\pi,Q,\gamma)}$ est sa densité par rapport à la mesure dominante. Noter qu'on a ici remplacé les lois d'émissions $(\nu_x)_{x\in[K]}$ par leurs densités $\gamma = (\gamma_x)_{x\in[K]}$ dans la notation des paramètres.

Notre estimateur est construit en deux étapes. La première consiste à trouver un maximiseur de la vraisemblance dans une famille de modèles paramétriques $(S_{K,M})_{(K,M)\in\mathbb{N}^*\times\mathcal{M}}$ indexés par un indice M et par un nombre d'états K. Par modèle paramétrique, on entend ici un ensemble de paramètres de HMM en bijection avec un sous-ensemble de $\mathbb{R}^d$ pour un certain entier d. On impose de plus que tous les paramètres du modèle $S_{K,M}$ correspondent à un HMM à K états. On calcule donc pour tout K et M un maximiseur

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

puis on sélectionne le modèle ayant la plus grande vraisemblance pénalisée.

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

L'estimateur final est

$$(\hat{K}_n, \hat{\pi}_{\hat{K}_n,\hat{M}_n,n}, \hat{Q}_{\hat{K}_n,\hat{M}_n,n}, \hat{\gamma}_{\hat{K}_n,\hat{M}_n,n}).$$

Les estimateurs du maximum de vraisemblance sont très utilisés en pratique grâce à l'algorithme de Baum-Welch, combinaison de l'algorithme EM (espérance-maximisation) de Dempster et al. (1977), particulièrement adapté aux modèles de mélange, et de l'algorithme forward-backward, qui permet de calculer la vraisemblance des observations ainsi que la loi *a posteriori* des états cachés pour les HMMs. Il est en général facile à implémenter, s'applique à un grand nombre de modèles et fournit une suite de paramètres qui converge toujours vers un maximiseur local de la vraisemblance. Son principal inconvénient est qu'il s'agit d'une procédure d'optimisation locale : la suite obtenue converge en général lentement et vers un paramètre sous-optimal. Cela impose de relancer plusieurs fois l'algorithme pour accroître les chances de tomber sur un bon maximiseur, ce qui augmente d'autant la durée de calcul sans garantie d'avoir trouvé le maximum global.

Un avantage théorique de l'estimateur du maximum de vraisemblance est qu'il se fonde sur la loi du processus entier pour retrouver les paramètres. Ceci permet d'utiliser le résultat d'identifiabilité du Théorème 1.3. Rappelons que celui-ci garantit l'identifiabilité sous des conditions plus générales que le Théorème 1.2, au prix de devoir connaître la loi d'un L-uplet d'observations (où L augmente polynômialement avec le nombre d'états cachés) au lieu d'un triplet d'observations. Prendre en compte la totalité du processus permet donc de retrouver les paramètres dans un plus grand nombre de situations.

Moindres carrés

Rappelons que le Théorème 1.2 montre qu'on peut retrouver les paramètres du HMM à partir de la loi d'un triplet d'observations. Suivant ce principe, l'estimateur des moindres carrés cherche à estimer la loi du triplet pour en déduire les paramètres du HMM. Plus précisément, si on note $(K^*, \pi^*, Q^*, \gamma^*)$ les vrais paramètres, on cherche des paramètres (K, π, Q, γ) tels que la distance

$$\left\| p_{Y_1^3}^{(K,\pi,Q,\gamma)} - p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|_{\mathbf{L}^2(\mathcal{Y}^3)}$$

soit la plus petite possible, ou de manière équivalente que

$$\left\| p_{Y_1^3}^{(K,\pi,Q,\gamma)} - p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)} - \left\| p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)}$$

soit la plus petite possible.

Ce critère n'est pas calculable en pratique car il nécessite de connaître les vrais paramètres. On le remplace donc par le critère empirique

$$\gamma_n : g \in \mathbf{L}^2(\mathcal{Y}^3) \longmapsto \|g\|^2_{\mathbf{L}^2(\mathcal{Y}^3)} - \frac{2}{n-2} \sum_{t=1}^{n-2} g(Y_t, Y_{t+1}, Y_{t+2})$$

Noter que pour tout $n \in \mathbb{N}^*$,

$$\begin{aligned} \lim_m \gamma_m(g) = \mathbb{E}\gamma_n(g) &= \|g\|^2_{\mathbf{L}^2(\mathcal{Y}^3)} - 2\langle g, p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \rangle_{\mathbf{L}^2(\mathcal{Y}^3)} \\ &= \left\| g - p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)} - \left\| p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)}, \end{aligned}$$

ce qui justifie son appellation de critère empirique.

De même que pour l'estimateur du maximum de vraisemblance, on procède par pénalisation : étant donné une famille de modèles $(S_{K,M})_{(K,M)\in\mathbb{N}^*\times\mathcal{M}}$, on construit des estimateurs

$$(K, \hat{\pi}_{K,M,n}, \hat{Q}_{K,M,n}, \hat{\gamma}_{K,M,n}) \in \underset{(K,\pi,Q,\gamma)\in S_{K,M}}{\arg\min} \gamma_n\left(p_{Y_1^3}^{(K,\pi,Q,\gamma)}\right),$$

puis on sélectionne

$$(\hat{K}_n, \hat{M}_n) \in \underset{(K,M)\in\mathbb{N}^*\times\mathcal{M}}{\arg\min} \left\{\gamma_n\left(p_{Y_1^3}^{(K,\hat{\pi}_{K,M,n},\hat{Q}_{K,M,n},\hat{\gamma}_{K,M,n})}\right) + \mathrm{pen}_n(K,M)\right\}.$$

Noter que le critère empirique, tout comme la log-vraisemblance pour l'EMV, n'est pas une fonction convexe. On utilise donc une procédure d'optimisation itérative pour trouver un minimum local du critère.

Remarque. *Cet estimateur se généralise au cadre du Théorème 1.3 lorsqu'une borne a priori sur l'ordre est disponible. Il s'agit alors d'estimer la loi d'un L-uplet d'observations, où L est pris suffisamment grand pour que le résultat d'identifiabilité s'applique.*

Spectral

L'algorithme spectral se fonde également sur la loi de trois observations consécutives. Plus précisément, on considère les tenseurs contenant les coordonnées de cette loi sur une base orthonormée $(\varphi_a)_{a\in\mathbb{N}^*}$ de $\mathbf{L}^2(\mathcal{Y})$: pour tout $a, b, c \in \mathbb{N}^*$, on note

$$\begin{aligned}
\mathbf{L}(a) &= \mathbb{E}[\varphi_a(Y_1)] &&= \langle p_{Y_1}^{(K^*,\pi^*,Q^*,\gamma^*)}, \varphi_a\rangle,\\
\mathbf{N}(a,b) &= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)] &&= \langle p_{(Y_1,Y_2)}^{(K^*,\pi^*,Q^*,\gamma^*)}, \varphi_a\otimes\varphi_b\rangle,\\
\mathbf{P}(a,c) &= \mathbb{E}[\varphi_a(Y_1)\varphi_c(Y_3)] &&= \langle p_{(Y_1,Y_3)}^{(K^*,\pi^*,Q^*,\gamma^*)}, \varphi_a\otimes\varphi_c\rangle,\\
\mathbf{M}(a,b,c) &= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)] &&= \langle p_{(Y_1,Y_2,Y_3)}^{(K^*,\pi^*,Q^*,\gamma^*)}, \varphi_a\otimes\varphi_b\otimes\varphi_c\rangle.
\end{aligned}$$

En utilisant l'expression de la densité des observations d'un HMM donné par la Proposition 1.1, il est possible d'écrire ces tenseurs comme produits matriciels de la loi initiale π^*, la matrice de transition Q^* et les coordonnées des densités d'émission sur la base $\mathbf{L}^2$. Une suite d'opérations matricielles simples (décomposition en valeurs singulières, inversions et diagonalisations) permet alors de retrouver les différents paramètres. Une version complète de cet algorithme est donnée dans l'Appendice 3.A du Chapitre 3.

Le premier point fort de cette méthode est son coût algorithmique : pour un grand nombre d'observations (plusieurs centaines de milliers), le temps d'exécution de l'algorithme spectral est plus court de plusieurs ordres de grandeur que celui des algorithmes précédents.

Le second est son indépendance des conditions initiales. Cet algorithme retrouve directement les paramètres du HMM sans passer par une procédure d'optimisation locale. Ces procédures, communes aux estimateurs du maximum de vraisemblance et des moindres carrés, dépendent fortement de l'initialisation choisie. Une mauvaise initialisation peut conduire à un maximum local sous-optimal, or les garanties théoriques ne s'appliquent qu'au maximum global.

Cet algorithme souffre toutefois de deux défauts. Le premier est son instabilité. Comme cette méthode repose sur la diagonalisation et l'inversion de matrices bruitées, elle peut donner des résultats aberrants lorsque le bruit est trop important. Cela se produit notamment lorsqu'on a peu d'observations, lorsque le nombre d'états considérés est grand ou encore lorsque les lois d'émission sont presque liées.

Le second défaut est son manque de généralisabilité. Le coeur de cette méthode est la structure du modèle de Markov caché. Elle ne fonctionne plus dès qu'on ajoute des relations de dépendance au modèle, par exemple entre les observations, contrairement à l'estimateur du maximum de vraisemblance pour lequel l'algorithme de Baum-Welch s'adapte naturellement.

1.3 État de l'art et contributions

Dans cette section, je présente mes contributions ainsi que la bibliographie qui s'y rapporte. J'y suis globalement le même ordre que dans les chapitres qui suivent. La première partie traite de l'estimation de l'ordre du HMM. La deuxième aborde la question de l'estimation adaptative et minimax des autres paramètres, principalement des densités d'émission. Dans la troisième partie, j'étends l'étude au cas des modèles mal spécifiés, c'est-à-dire quand les observations ne sont pas générées par un HMM. Enfin, j'étudie dans une quatrième partie un cas particulier de HMMs non homogènes (c'est-à-dire dont la loi dépend du temps) avec les HMMs avec tendance. Cette dernière partie est issue d'un travail en cours en collaboration avec Augustin Touron, en thèse CIFRE avec EDF, qui travaille sur les générateurs stochastiques de temps sous la direction d'Élisabeth Gassiat et Thi-Thu-Huong Hoang.

L'objectif de cette section est avant tout de présenter le contexte et les apports de cette thèse. Les résultats sont présentés en détail dans les chapitres concernés.

1.3.1 Estimation de l'ordre

Les premiers résultats théoriques sur l'estimation de l'ordre d'un HMM remontent à une vingtaine d'années. Un point commun à toutes les approches est le découpage de cette étude en deux parties : montrer que l'estimateur ne sous-estime pas l'ordre et montrer qu'il ne le sur-estime pas.

L'identifiabilité est l'argument clé pour traiter la sous-estimation : puisque l'ordre est le plus petit nombre d'états requis pour générer les observations, aucun HMM avec moins d'états ne pourra avoir la bonne loi. Le vrai HMM et l'ensemble des HMMs avec moins d'états sont donc écartés d'une distance non nulle qui devient visible quand on a suffisamment d'observations.

La plupart des résultats théoriques avant les travaux de Gassiat et al. (2015) sur l'identifiabilité des modèles de Markov cachés supposent en plus soit que l'espace des observations est fini, soit que les mélanges des lois d'émission sont identifiables. Autrement dit, si on note $(f_\theta)_{\theta\in\Theta}$ la famille de lois dans laquelle on cherche les lois d'émission, on a pour toutes mesures de probabilité G et H à support fini dans Θ

$$\int f_\theta dG(\theta) = \int f_\theta dH(\theta) \quad \Longrightarrow \quad G = H.$$

Cette propriété se vérifie au cas par cas pour des familles spécifiques. Elle est par exemple satisfaite pour les lois de Poisson, Gaussiennes, Gamma, de Paréto... Mais sa rigidité fait qu'elle ne permet pas de traiter la plupart des HMMs paramétriques, *a fortiori* les HMMs non paramétriques. Par exemple, elle n'est plus vérifiée lorsque la famille des lois d'émission est l'ensemble des mélanges de deux Gaussiennes.

Une deuxième hypothèse répandue est une borne *a priori* sur l'ordre. Autrement dit, on suppose connu un entier K_0 tel que le vrai ordre K vérifie $K \leqslant K_0$.

Dans la revue bibliographique qui suit, tous les HMMs sont paramétriques et, sauf mention contraire, leur espace d'observations n'est pas fini (il s'agit en général d'une partie de $\mathbb{R}^d$ pour un certain entier d).

L'approche dominante, inaugurée par Rydén (1995), repose sur la sélection de l'ordre par pénalisation de la vraisemblance, en l'occurrence d'une variante de la vraisemblance obtenue en considérant que les observations forment des blocs indépendants. La taille des blocs en question dépend d'une borne *a priori* sur l'ordre du HMM. Lorsque les mélanges des lois d'émission sont identifiables, l'ordre n'est pas sous-estimé asymptotiquement.

Toujours avec une borne *a priori* sur l'ordre, Gassiat (2002) étudie la vraisemblance obtenue en considérant que les observations sont i.i.d., comme dans un modèle de mélange. Elle montre alors que la vraisemblance pénalisée conduit à un estimateur consistant de l'ordre. Noter que cet article ne suppose pas que les mélanges des lois d'émission sont identifiables, mais définit comme ordre le plus petit nombre de composantes nécessaire pour obtenir la loi marginale d'une observation, ce qui peut différer de notre définition.

En utilisant des vraisemblances issues de la théorie de l'information, Gassiat and Boucheron (2003) construisent deux estimateurs fortement consistants de l'ordre par pénalisation de la vraisemblance. Leur résultat est cependant restreint à un espace d'observations fini, mais n'a besoin ni de l'identifiabilité des paramètres du HMM autres que l'ordre, ni de borne *a priori* sur l'ordre.

Lorsque les lois d'émission sont des Gaussiennes ou des lois de Poisson, Chambaz et al. (2009) montrent que les estimateurs de l'ordre obtenus par pénalisation de la vraie vraisemblance et de la vraisemblance obtenue en considérant que les observations sont i.i.d. sont consistants, sans avoir besoin de borne *a priori*.

Même si la vraisemblance reste extrêmement majoritaire dès qu'on cherche à construire un critère par pénalisation, d'autres critères sont possibles. En supposant que les mélanges de lois d'émission sont identifiables, MacKAY (2002) pénalise ainsi la distance de Kolmogorov-Smirnov sur la loi d'un m-uplet d'observations, où m dépend en général d'une borne *a priori* de l'ordre. L'estimateur obtenu est consistant.

Une autre approche repose sur les tests du rapport de vraisemblance : en notant K l'ordre, il s'agit de tester successivement $K = K_0$ contre $K > K_0$ pour différentes valeurs de K_0. L'ordre est le plus petit K_0 pour lequel l'hypothèse $K = K_0$ n'est pas rejetée. Malheureusement, cette approche échoue en général : Gassiat and Keribin (2000) montrent que la statistique du rapport de vraisemblance n'est pas bornée dans le test $K = 1$ contre $K = 2$ (autrement dit mélange contre HMM à deux états). Citons tout de même Dannemann and Holzmann (2008), qui étudient un test du rapport de vraisemblance $K = 2$ contre $K = 3$ en modifiant la vraisemblance en considérant que les observations sont indépendantes. Leur résultat nécessite que les mélanges des lois d'émission soient identifiables et revient à tester le nombre de composantes dans la loi marginale d'une observation.

Enfin, des approches bayésiennes ont été étudiées plus récemment. Gassiat and Rousseau (2014) montrent que la loi *a posteriori* se concentre autour des paramètres ayant le bon ordre ; van Havre et al. (2016) étudient des HMMs où le nombre d'états est pris volontairement trop grand mais dont la loi *a priori* est telle que la loi *a posteriori* donne un poids arbitrairement petit aux états en trop. Ces deux résultats, de par leur construction, nécessitent une borne *a priori* sur l'ordre.

Dans le Chapitre 2, nous introduisons deux nouveaux estimateurs fortement consistants de l'ordre d'un HMM. Ces estimateurs ont le double avantage de s'appliquer aux HMMs non paramétriques et de ne pas nécessiter de borne *a priori* sur l'ordre. Ils s'appliquent également à des espaces d'observation quelconques.

Le premier estimateur est l'estimateur des moindres carrés pénalisés introduit dans la Section 1.2.3. Soit $(\mathfrak{P}_M)_{M\in\mathbb{N}^*}$ une famille de sous-espaces vectoriels de $\mathbf{L}^2(\mathcal{Y})$ croissante au sens de l'inclusion, d'union dense dans $\mathbf{L}^2(\mathcal{Y})$. Pour simplifier, on suppose que $\dim(\mathfrak{P}_M) = M$ pour tout

M. Pour tout K et M, le modèle $S_{K,M}$ est l'ensemble des paramètres de HMMs stationnaires à K états dont les lois d'émission sont dans $\mathfrak{P}_M$.

On construit d'abord pour tout K et M les estimateurs

$$(K, \hat{\pi}_{K,M,n}, \hat{Q}_{K,M,n}, \hat{\gamma}_{K,M,n}) \in \underset{(K,\pi,Q,\gamma)\in S_{K,M}}{\arg\min} \gamma_n \left(p_{Y_1^3}^{(K,\pi,Q,\gamma)} \right),$$

puis on sélectionne l'estimateur de l'ordre

$$\hat{K}_n \in \underset{K\leqslant n}{\arg\min} \min_{M\leqslant n} \left\{ \gamma_n \left(p_{Y_1^3}^{(K,\hat{\pi}_{K,M,n},\hat{Q}_{K,M,n},\hat{\gamma}_{K,M,n})} \right) + \mathrm{pen}_n(K,M) \right\}.$$

Noter que lorsqu'on dispose de n observations, on ne considère que les HMMs avec moins de n états et des densités d'émission dans des espaces de dimension inférieure à n. Cette borne est peu contraignante : les plus gros modèles considérés ont plus de degrés de liberté qu'il n'y a d'observations. On peut se restreindre à un nombre d'états et à une dimension inférieure à $M_n^{\max} < n$ sans perdre les propriétés théoriques de l'estimateur, tant que cette borne tend vers l'infini avec le nombre d'observations.

Le Corollaire 2.5 assure que cet estimateur est fortement consistant. Ce résultat fait appel à un contrôle fin des déviations du critère empirique sur tous les modèles à la fois. Sa preuve se décompose en deux parties. D'abord, le Théorème 2.3 montre que pour un choix général de pénalité, la probabilité de sous-estimer l'ordre décroît exponentiellement vite. Ensuite, le Théorème 2.4 montre que sous une certaine condition sur la pénalité, la probabilité de sur-estimer l'ordre décroît polynomialement. Cette condition sert à s'assurer que la pénalité croît assez vite pour compenser le sur-apprentissage des modèles, et est typiquement vérifiée pour une pénalité de type BIC alourdie, autrement dit

$$\mathrm{pen}_n(K,M) = u_n(MK + K^2 - 1)\frac{\log n}{n}$$

où u_n tend vers l'infini. Le terme $MK + K^2 - 1$ correspond à la dimension du modèle : K densités d'émissions, chacune dans un espace de dimension M ; $K(K-1)$ coefficients libres dans la matrice de transition et $K-1$ coefficients libres dans la loi initiale de la chaîne cachée.

Cette méthode a également l'avantage d'estimer simultanément les autres paramètres du HMM de manière minimax et adaptative. C'est l'objet du Corollaire 2.11, sur lequel nous reviendrons dans la section suivante.

Le second estimateur est fondé sur la méthode spectrale, plus particulièrement sur la matrice des coordonnées de la loi de deux observations consécutives. Reprenons la famille d'espaces vectoriels $(\mathfrak{P}_M)_{M\in\mathbb{N}^*}$. Il existe une base orthonormée $(\varphi_a)_{a\in\mathbb{N}^*}$ telle que pour tout M, $\mathfrak{P}_M$ est engendré par les M premiers vecteurs de cette base. On définit alors pour tout M

$$\forall (a,b) \in [M], \quad \mathbf{N}_M(a,b) = \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)].$$

Sous les hypothèses d'identifiabilité du Théorème 1.2, pour M assez grand, le rang de cette matrice est égal à l'ordre du HMM. Estimer l'ordre revient donc à estimer le rang de cette matrice. En pratique, on doit se fonder sur la matrice empirique

$$\forall (a,b) \in [M], \quad \hat{\mathbf{N}}_M(a,b) = \frac{1}{n-1}\sum_{t=1}^{n-1} \varphi_a(Y_t)\varphi_b(Y_{t+1}),$$

qui est, elle, presque sûrement de rang M. Pour obtenir le rang, on procède par seuillage des valeurs singulières de cette matrice, c'est-à-dire qu'on définit

$$\hat{K} = \sup\{i : \sigma_i(\hat{\mathbf{N}}_M) > C\sqrt{\log(n)/n}\}$$

pour une constante C assez grande. Le terme en $\sqrt{\log(n)/n}$ apparaît lors du contrôle de l'écart entre $\hat{\mathbf{N}}_M$ et $\mathbf{N}_M$. La probabilité que cet estimateur ne trouve pas le bon ordre décroît polynomialement avec le nombre d'observations, ce qui implique qu'il est fortement consistant.

Cette seconde méthode se distingue nettement des méthodes envisagées jusqu'à présent car elle estime l'ordre de façon directe : il n'est pas nécessaire d'estimer un paramètre pour différents nombres d'états puis de sélectionner le meilleur. Il suffit de calculer une matrice et de regarder ses valeurs singulières. Par conséquent, elle a le double avantage d'être rapide et d'avoir une interprétation visuelle : en général, l'ordre se manifeste sur son spectre par un « coude » qui sépare les valeurs singulières significatives, correspondant aux valeurs singulières non nulles de $\mathbf{N}_M$, et les valeurs singulières non significatives, provenant du bruit. La Figure 2.5 (page 44) illustre ce phénomène. Elle peut également être généralisée en utilisant d'autres méthodes pour estimer l'ordre de cette matrice.

Ces deux méthodes sont étudiées empiriquement sur des données simulées. Nous détaillons dans la Section 2.5 du Chapitre 2 comment les implémenter en pratique et les résultats des simulations, qui confirment leur capacité à retrouver l'ordre du HMM.

1.3.2 Estimation minimax et adaptative des densités d'émission

Dans cette section, je développe le contexte de l'estimation minimax pour les modèles de Markov cachés non paramétriques. J'y distingue notamment les estimateurs globalement minimax, dont les performances se dégradent lorsque les paramètres n'ont pas la même régularité, et les estimateurs minimax état par état. Je décris enfin une nouvelle méthode permettant de construire de manière systématique des estimateurs minimax état par état. C'est à ma connaissance le premier résultat d'adaptation état par état existant.

De nombreux résultats théoriques existent pour les modèles de Markov cachés paramétriques, voir par exemple le livre de Cappé et al. (2005) et les références qu'il contient. Les modèles de Markov cachés non paramétriques n'ont été étudiés que beaucoup plus récemment : les premiers résultats ont été développés en parallèle par de Castro et al. (2016) et Bonhomme et al. (2016b).

Dans toute la revue bibliographique qui suit, sauf précision contraire, les articles supposent que le nombre d'états cachés est connu.

En se fondant sur des travaux d'identifiabilité de Allman et al. (2009), Bonhomme et al. (2016b) introduisent une méthode spectrale permettant de retrouver les paramètres de modèles à variable latente multivariés, dont les HMMs peuvent être vus comme un cas particulier. Leur méthode se fonde sur les mêmes outils que celle introduite par Anandkumar et al. (2012) et Hsu et al. (2012) : exploiter les corrélations entre observations pour retrouver la structure latente. Leur méthode permet d'estimer les coordonnées des densités d'émission sur les M premiers vecteurs d'une base orthonormée $(\varphi_a)_{a\in\mathbb{N}^*}$. Faire tendre M vers l'infini avec le nombre d'observations permet alors de retrouver les densités d'émission. Mieux encore, ils montrent que si la régularité des densités est connue, il est possible de choisir M de sorte que les estimateurs des densités d'émission convergent à la vitesse minimax. Leurs estimateurs ne sont pas adaptatifs.

Toujours lorsque le nombre d'états cachés est connu, De Castro et al. (2017) introduisent une méthode spectrale pour les HMMs non paramétriques, également fondée sur l'approche de Anandkumar et al. (2012). Leur estimateur n'est toutefois pas minimax. Leur résultat principal relie l'erreur sur les estimateurs des paramètres à l'erreur sur les probabilités de filtrage et de lissage, autrement dit la loi des états cachés conditionnellement aux observations, dans un cadre non paramétrique. Ce résultat est crucial dans les applications nécessitant de retrouver les états cachés.

Cette méthode spectrale est reprise par de Castro et al. (2016), qui l'utilisent comme estimateur auxiliaire de la matrice de transition du HMM. Cet article introduit une méthode pour

retrouver les lois d'émission d'un HMM non paramétrique par sélection de modèle en utilisant un critère des moindres carrés pénalisés. L'estimateur qui en découle est minimax et adaptatif, mais nécessite que le nombre d'états cachés soit connu.

Vernet (2015b) étudie les vitesses de convergence a posteriori d'un estimateur bayésien pour les HMMs non paramétriques et montre qu'il se concentre à vitesse minimax, et ce de manière adaptative, lorsque le nombre d'états cachés est connu.

Enfin, Gassiat et al. (2016), étudient le cas particulier des HMMs dont les lois d'émission sont des versions translatées d'une loi inconnue. Autrement dit, les observations s'écrivent $Y_i = m_{X_i} + \epsilon_i$ où la loi de ϵ_i ne dépend pas de l'état caché X_i. Dans ce cadre, elles montrent qu'il est possible de retrouver non seulement les paramètres de translation $(m_x)_{x\in\mathcal{X}}$ mais également le nombre d'états cachés et la loi de ϵ_i dans un cadre non paramétrique.

Dans cette thèse, je propose et étudie plusieurs estimateurs pour les HMMs non paramétriques. L'idée générale de ces deux estimateurs a été introduite dans la Section 1.2.3.

Le premier est un estimateur spectral qui améliore celui de De Castro et al. (2017). Cet estimateur est rendu minimax (mais non adaptatif en l'état) en s'inspirant de l'idée de Bonhomme et al. (2016b) de traiter différemment les différentes observations du triplet dont on estime la loi. De plus, il est stabilisé en considérant plusieurs rotations aléatoires dans l'étape de codiagonalisation au lieu d'une seule. Cet estimateur est étudié dans le Chapitre 3 et détaillé en Section 3.A.

Le second estimateur est un estimateur des moindres carrés. Cet estimateur est fondé sur celui de de Castro et al. (2016) mais ne nécessite pas d'estimateur auxiliaire : il estime simultanément tous les paramètres du modèle, le nombre d'états, la matrice de transition et les densités d'émission, le tout de façon minimax et adaptative. L'estimation de l'ordre est capitale car le nombre d'états cachés n'est pas connu en général. De plus, ne pas recourir à des estimateurs auxiliaires permet d'éviter les cas où ceux-ci échouent à retrouver les paramètres, ce qui peut arriver à l'estimateur auxiliaire de de Castro et al. (2016). Cet estimateur est étudié dans le Chapitre 2, et son contrôle est affiné dans le Chapitre 3.

Étude locale du risque. Une des principales difficultés rencontrées lors de l'étude des estimateurs fondés sur la minimisation d'un risque est de relier l'erreur sur le risque à l'erreur sur les paramètres. Pour l'estimateur des moindres carrés par exemple, il s'agit de relier l'erreur

$$\left\| p_{Y_1^3}^{(K^*,\hat{\pi},\hat{Q},\hat{\gamma})} - p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)}$$

à l'erreur sur les paramètres (π^*, Q^*, γ^*) (en supposant le nombre d'états connus pour simplifier). Je montre dans le Chapitre 2 que cette erreur est équivalente à

$$d_{\mathrm{perm}}((\hat{\pi},\hat{Q},\hat{\gamma}),(\pi^*,Q^*,\gamma^*))^2 = \|\hat{Q}-Q^*\|_F^2 + \|\hat{\pi}-\pi^*\|_2^2 + \sum_{x\in[K^*]} \|\hat{\gamma}_x - \gamma_x^*\|_2^2, \tag{1.1}$$

modulo permutation des états cachés près. Par équivalence, j'entends qu'il existe deux constantes $c > 0$ et $C < \infty$ telles que

$$\begin{aligned} c\, d_{\mathrm{perm}}&((\hat{\pi},\hat{Q},\hat{\gamma}),(\pi^*,Q^*,\gamma^*))^2 \\ &\leqslant \left\| p_{Y_1^3}^{(K^*,\hat{\pi},\hat{Q},\hat{\gamma})} - p_{Y_1^3}^{(K^*,\pi^*,Q^*,\gamma^*)} \right\|^2_{\mathbf{L}^2(\mathcal{Y}^3)} \leqslant C\, d_{\mathrm{perm}}((\hat{\pi},\hat{Q},\hat{\gamma}),(\pi^*,Q^*,\gamma^*))^2. \end{aligned} \tag{1.2}$$

L'analyse théorique montre que le risque, ici l'erreur sur la densité des triplets, se concentre à la bonne vitesse. Cette équivalence est cruciale pour en déduire que les estimateurs des paramètres se concentrent à la vitesse minimax.

Lorsque la matrice de transition et la loi initiale sont fixées, ces inégalités ont été étudiées par de Castro et al. (2016). La majoration est toujours vérifiée. Le principal résultat de leur article est que la minoration est génériquement vérifiée pour tout K^*, et toujours vérifiée pour $K^* = 2$. Plus précisément, ils montrent que pour tout Q^* et γ^* en dehors des zéros d'un polynôme non nul, il existe un voisinage de Q^* tel que l'équation (1.2) soit vraie en remplaçant Q^* et $\hat{Q}$ par un Q dans ce voisinage et π^* et $\hat{\pi}$ par la loi invariante associée. La constante c ne dépend alors que du voisinage et de γ^*.

Nous étendons cette étude au cas où la matrice de transition Q est également inconnue. Notre principal résultat à ce sujet est le Théorème 3.7, qui démontre qu'il est possible de choisir la constante c comme une fonction des paramètres (π^*, Q^*, γ^*) semi-continue inférieurement et génériquement non nulle, et qui fournit une expression explicite de cette constante au voisinage des vrais paramètres.

La clé de la preuve de l'équation (1.2) est l'étude du critère au voisinage des vrais paramètres, plus particulièrement sa Hessienne, la matrice de ses dérivées secondes (lorsqu'elle peut être définie). Comme le vrai paramètre est un minimiseur du critère, cette Hessienne est positive. Lorsqu'elle est de plus non dégénérée, elle définit une norme qui est équivalente au critère au voisinage des vrais paramètres. La seconde partie consiste à s'assurer que cette équivalence reste vraie même loin des vrais paramètres.

C'est le même outil qui intervient dans l'étude de l'estimateur du maximum de vraisemblance pour les modèles paramétriques. Tous les résultats de normalité asymptotiques font appel à la matrice d'information de Fisher, qui est en fait la Hessienne de la vraisemblance : on montre en général que l'estimateur d'un paramètre $\theta^* \in \mathbb{R}^q$ vérifie

$$\frac{1}{\sqrt{n}}(\hat{\theta} - \theta^*) \longrightarrow \mathcal{N}(0, I(\theta^*)^{-1})$$

où $I(\theta^*)$ est une matrice $q \times q$ appelée information de Fisher du modèle en θ^*, supposée définie positive. Si on effectue la transformation $\vartheta \longleftarrow \sqrt{nI(\theta^*)}(\theta - \theta^*)$, qui transforme la métrique définie par l'information de Fisher en la métrique euclidienne usuelle, cette loi devient une Gaussienne multivariée centrée réduite. C'est donc bien la Hessienne du risque qui définit la « bonne » métrique pour étudier le comportement asymptotique des estimateurs.

Le plus gros problème de cette approche apparaît lors du passage au cadre non paramétrique, autrement dit à la dimension infinie. En dimension finie, toutes les normes sont équivalentes. Il suffit donc que la Hessienne soit non dégénérée pour que la norme qu'elle définit soit équivalente à n'importe quelle norme usuelle. Malheureusement, en dimension infinie, il n'y a plus aucune garantie que la norme issue de la Hessienne et les normes usuelles sont équivalentes. On se retrouve donc avec une métrique impossible à exploiter puisqu'elle dépend de manière non explicite du vrai paramètre, supposé inconnu !

Ce problème peut être résolu pour l'estimateur des moindres carrés en utilisant explicitement la structure d'espace de Hilbert de $\mathbf{L}^2(\mathcal{Y})$. Cela consiste à séparer la forme quadratique en deux parties : la première traite la projection des densités d'émission sur l'espace engendré par les vraies densités d'émission (qui est de dimension finie), et la seconde traite leur partie orthogonale aux vraies densités d'émission. La première partie se traite comme dans le cas paramétrique, et la seconde ne pose pas problème.

Malheureusement, une telle approche n'est pas possible pour l'estimateur du maximum de vraisemblance. Nous montrerons néanmoins qu'il est possible de démontrer des résultats de convergence minimax et adaptative sur l'estimateur de maximum de vraisemblance, même dans un cadre non paramétrique, dans le Chapitre 4.

Estimateurs globalement minimax et estimateurs minimax état par état. Dans tous les résultats précédents, le terme « minimax » se rapporte à une erreur globale qui prend en

compte tous les paramètres. Il s'agit en général de l'erreur donnée par la distance d_{perm} définie dans l'Équation (1.1), qui est elle-même équivalente à l'erreur

$$\|\hat{Q} - Q^*\|_F^2 \vee \|\hat{\pi} - \pi^*\|_2^2 \vee \max\left(\|\hat{\gamma}_x - \gamma_x^*\|_2^2 : x \in [K^*]\right).$$

Autrement dit, les résultats précédents étudient la convergence minimax pour la plus grande erreur sur les paramètres. C'est également le cas des estimateurs spectraux : même s'ils ne reposent pas sur la minimisation d'un risque, les opérations matricielles sur lesquelles ils sont fondés font intervenir tous les paramètres à la fois, ce qui répercute l'erreur sur chaque paramètre sur les autres paramètres.

Détaillons la vitesse de chacune de ces erreurs dans le cas particulier des densités Hölder :

- la vitesse minimax d'estimation de Q^* et de π^* est au mieux $n^{-1/2}$: c'est la vitesse paramétrique ;
- si les densités d'émission γ_x^* sont Hölder sur $[0,1]^d$ de régularités respectives s_x, la vitesse minimax d'estimation de γ_x^* est au mieux $n^{-s_x/(2s_x+d)}$.

Ainsi, la vitesse minimax de convergence de l'erreur globale est au mieux $n^{-\inf_x s_x/(2\inf s_x+d)}$, ce qui est la plus lente des vitesses individuelles. Comme il n'est pas possible de faire mieux que cette vitesse, cela implique qu'il suffit qu'une seule densité soit peu régulière pour dégrader la vitesse d'estimation de tous les autres paramètres.

Dans le Chapitre 3, j'introduis la notion d'estimateur minimax *état par état*. Un tel estimateur doit être capable d'estimer chaque paramètre avec la vitesse minimax qui lui est propre. Ces estimateurs sont donc capables de traiter des paramètres avec des régularités très différentes sans dégrader leur vitesse de convergence.

La méthode utilisée se fonde sur la sélection de modèle, et s'inspire de la méthode de Goldenshluger et Lepski. Elle permet, à partir d'une famille d'estimateurs auxiliaires, de construire des estimateurs minimax état par état de manière adaptative. Pour cela, on calcule un proxy du biais de chaque estimateur des densités d'émission directement à partir de la famille d'estimateurs. Ce critère est construit séparément pour chaque état caché, puis pénalisé, ce qui permet de sélectionner un modèle différent selon l'état caché considéré. Cette méthode est détaillée dans la Section 3.2.2, dans laquelle nous montrons également que les estimateurs obtenus vérifient une inégalité oracle (Théorème 3.2), ce qui assure qu'ils sont minimax état par état et adaptatifs lorsque la famille de départ est fondée sur les estimateurs spectraux (Corollaire 3.4) ou des moindres carrés (Corollaire 3.9).

Cette méthode est totalement indépendante de la manière dont sont construits les estimateurs auxiliaires : n'importe quelle famille pouvant donner un estimateur globalement minimax (même non adaptatif) peut être employée pour fournir des estimateurs minimax état par état et adaptatifs.

Un autre avantage de cette méthode est son coût algorithmique. Cette méthode est rapide, quelques secondes sur un ordinateur de bureau récent lorsque la famille compte quelques centaines d'estimateurs. De plus, ce coût ne dépend que du nombre d'estimateurs dans la famille, pas du nombre d'observations disponibles. Il est donc possible de l'appliquer avec de très gros jeux de données pour construire un estimateur qui reste optimal même en présence de différences de régularité, et ce pour un temps supplémentaire négligeable par rapport à la méthode globalement minimax sur laquelle elle se greffe.

Le principal paramètre de cette méthode, et le seul pour lequel une calibration automatique n'est pas disponible, est la pénalité. Cette pénalité doit être une borne de la variance des estimateurs. Un choix heuristique type BIC est possible. Nous montrons dans la Section 3.3 que ce choix est justifié pour les deux familles d'estimateurs considérées.

Enfin, nous examinons les performances de cette méthode et de quelques variantes sur des données simulées et sur des données réelles à la fin du Chapitre 3.

1.3.3 Modèles mal spécifiés

Aucun processus réel ne peut être parfaitement décrit par un HMM ou de manière générale par un modèle statistique. Lorsque la loi du processus observé n'appartient pas au modèle, on dit que ce modèle est *mal spécifié*.

L'étude des modèles mal spécifiés permet de s'assurer que les estimateurs sont robustes à une erreur de modélisation. Il s'agit en général de montrer, sous des hypothèses générales, que la loi estimée converge vers la meilleure approximation de la vraie loi dans le modèle. Si un certain nombre de travaux ont été effectués lorsque les observations sont indépendantes, le cas des observations dépendantes est moins bien connu.

Tous les résultats de la revue bibliographique suivante portent sur l'estimateur du maximum de vraisemblance dans le cadre paramétrique. Ils se donnent un espace métrique compact $\Theta \subset \mathbb{R}^p$ pour un certain entier p et une application de Θ dans un ensemble de paramètres de HMMs (la *paramétrisation*). L'objectif est de trouver le point de Θ qui donne la loi la plus proche de celle du processus observé.

Mevel and Finesso (2004) étudient les HMMs avec un nombre fini et connu d'états cachés à valeurs dans $\mathbb{R}^d$. Ils montrent que la distance entre l'EMV et l'ensemble des maximiseurs de la log-vraisemblance limite tend presque sûrement vers zéro, ce qui tient lieu de consistance. Lorsque cet ensemble de maximiseurs est un ensemble de points isolés, l'estimateur est également asymptotiquement normal. Ce résultat s'étend au cas où la matrice de Fisher n'est pas inversible, modulo une perte de contrôle sur les directions données par le noyau de cette matrice.

Le résultat de Douc and Moulines (2012) porte sur un cadre plus général mais ne montre que la consistance de l'EMV. Leur résultat s'applique à des HMMs à espaces d'états et d'observations quelconques (qui peuvent n'être ni finis ni compacts). Ils montrent que la distance entre l'EMV et l'ensemble des maximiseurs de la log-vraisemblance limite tend presque sûrement vers zéro.

Enfin, Pouzo et al. (2016) étudient une généralisation des HMMs où chaque état caché dépend de l'état caché et de l'observation de l'instant précédent, et où chaque observation dépend de l'observation de l'instant précédent et de l'état caché actuel. L'espace des états cachés est supposé fini et de taille connue. Contrairement aux résultats précédents qui considéraient un processus observé quelconque, ils supposent que le processus observé est généré par un tel modèle. L'aspect mal spécifié vient du fait que les paramètres associés à la vraie loi n'appartiennent pas à l'ensemble des paramètres considérés par l'estimateur. Ils montrent que la distance entre l'EMV et l'ensemble des maximiseurs de la log-vraisemblance limite tend presque sûrement vers zéro. De plus, ils fournissent un développement asymptotique de la log-vraisemblance duquel ils déduisent la normalité asymptotique de l'EMV.

Tous ces résultats ont en commun qu'ils s'intéressent à la paramétrisation : ils cherchent à montrer que l'estimateur $\hat{\theta}$ converge vers le meilleur élément de Θ. Quand le modèle est mal spécifié, cette paramétrisation n'apporte plus d'information sur la structure du processus. Tout ce qu'on peut espérer est que la loi estimée soit proche de la vraie loi. C'est cela que j'étudie.

La vraisemblance fournit une mesure naturelle de l'écart entre deux lois : la divergence de Kullback-Leibler. Supposons que le processus observé est ergodique et stationnaire de loi $\mathbb{P}^*$ associé à une log-vraisemblance ℓ_n^*. Notons $\mathbb{P}^{(K,Q,\gamma)}$ la loi d'un HMM stationnaire de paramètres (K, Q, γ) et $\ell_n(K, Q, \gamma)$ la log-vraisemblance associée. Alors sous des hypothèses générales, il

existe une fonction $\mathbf{K}$ telle que

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

Cette fonction est positive et vérifie $\mathbf{K}(K,Q,\gamma) = 0$ si et seulement si le processus observé est généré par un HMM stationnaire de paramètres (K,Q,γ). Elle peut abusivement s'écrire sous la forme d'une divergence de Kullback-Leibler entre les lois d'une observation conditionnée à tout son passé :

$$\mathbf{K}(K,Q,\gamma) = \mathbb{E}^*_{Y_{-\infty}^0}\left[KL(\mathbb{P}^*_{Y_1|Y_{-\infty}^0}\|\mathbb{P}^{(K,Q,\gamma)}_{Y_1|Y_{-\infty}^0})\right].$$

C'est donc une mesure de la similarité entre la loi du HMM et la vraie loi du processus. Les articles de la bibliographie traitent cette quantité comme un intermédiaire pour contrôler ce qui se passe dans Θ. Cette approche requiert des hypothèses restrictives sur la paramétrisation, dont la principale est que l'espace Θ est compact. D'autres hypothèses sont que les minimiseurs de $\mathbf{K}$ sont isolés, voire qu'il n'en existe qu'un et que celui-ci soit dans l'intérieur de Θ.

Dans le Chapitre 4, j'introduis une méthode de sélection de modèle par vraisemblance pénalisée. D'une part, la sélection de modèle permet de considérer des modèles non paramétriques, capables d'approcher plus de lois. D'autre part, elle permet de se passer de l'hypothèse de compacité. Cette hypothèse est en général utilisée pour montrer que la différence des log-vraisemblances converge vers le critère $\mathbf{K}$ uniformément sur tout le modèle. Ici, je considère des modèles compacts que j'autorise à croître avec le nombre d'observations, de sorte que leur union est dense dans l'espace des paramètres. La clé est de choisir judicieusement la vitesse de croissance des modèles pour s'assurer que la convergence uniforme de la log-vraisemblance a toujours lieu et pour en contrôler la vitesse.

Le principal résultat du chapitre est l'inégalité oracle du Théorème 4.8. Cette inégalité assure que la loi estimée approche au mieux la vraie loi au sens du critère $\mathbf{K}$. Noter qu'on autorise le nombre d'états cachés à être arbitrairement grand, ce qui permet d'approcher des dépendances complexes.

Les hypothèses sur le processus observé sont qu'il est exponentiellement mélangeant et qu'il oublie ses conditions initiales exponentiellement vite. Ces hypothèses sont classiques pour les HMMs bien spécifiés, où elles découlent de l'ergodicité géométrique de la chaîne de Markov cachée. Je suppose également que la densité du processus observé a des queues sous-polynomiales. Je ne suppose rien sur la structure du processus ; en particulier, il n'est pas forcément généré par un HMM.

La première hypothèse sur les modèles est que toutes les matrices de transition sont minorées par une constante strictement positive, ce qui assure l'ergodicité géométrique de la chaîne de Markov cachée. Cette constante peut tendre vers zéro quand le nombre d'observations tend vers l'infini. La deuxième est une hypothèse sur les queues des densités d'émissions. Les deux dernières hypothèses contrôlent la complexité des modèles au sens de l'entropie à crochets et sont habituelles en concentration de la mesure.

Ces hypothèses sont par exemple vérifiées lorsque les modèles sont des HMMs dont les densités d'émission sont des mélanges finis de Gaussiennes. Lorsque le processus est généré par un HMM à valeurs dans $\mathbb{R}$, la loi associée à l'EMV construit à partir de ces modèles converge à vitesse minimax et adaptative vers la vraie loi. C'est à notre connaissance le seul résultat théorique sur l'estimateur du maximum de vraisemblance pour les HMMs non paramétriques, avec celui d'Alexandrovich et al. (2016) qui montre que l'EMV est consistant lorsque le modèle est bien spécifié, que le nombre d'états cachés est connu et que les lois d'émission sont des mélanges (non paramétriques) de lois paramétriques.

1.3.4 Modèles non homogènes

Dans les modèles de Markov cachés et la plupart de leurs généralisations, le processus $(X_t, Y_t)_{t\geqslant 1}$ est une chaîne de Markov. On dit que le modèle est *non homogène* lorsque cette chaîne est non homogène, c'est-à-dire que la loi de (X_t, Y_t) conditionnellement à (X_{t-1}, Y_{t-1}) dépend de t. Ainsi, dans un HMM non homogène, la matrice de transition et les lois d'émission peuvent dépendre du temps.

L'étude de ces modèles se fonde sur la constatation que beaucoup de processus réels ne peuvent pas être supposés stationnaires. Sur le plan théorique, l'étude de généralisations non homogènes des HMMs est un sujet très récent. Dans cette section, nous présentons les résultats existants puis détaillons notre contribution à ce domaine. Tous les articles cités ci-après considèrent des modèles paramétriques.

Diehn et al. (2018) étudient le cas où un phénomène transitoire affecte la loi des observations. Leur modèle est un processus trivarié $(X_t, Y_t, Z_t)_{t\geqslant 1}$ où seul le processus $(Z_t)_{t\geqslant 1}$ est observé tel que $(X_t, Y_t)_{t\geqslant 1}$ est un HMM homogène et $(X_t, Z_t)_{t\geqslant 1}$ est un HMM non homogène. L'hypothèse centrale est que la distance entre Z_t et Y_t tend vers zéro suffisamment vite lorsque t tend vers l'infini. Ainsi, le processus $(Z_t)_{t\geqslant 1}$ est le processus $(Y_t)_{t\geqslant 1}$ perturbé par un bruit non homogène qui s'atténue rapidement. Les auteurs introduisent deux estimateurs, l'un étant l'estimateur du maximum de vraisemblance usuel, l'autre un estimateur du maximum de la quasi-log-vraisemblance : en notant $p^{\theta}_{Y_1^n}$ la densité du vecteur de variables aléatoires Y_1^n (et de même pour Z_1^n), l'estimateur du maximum de vraisemblance est

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^{\theta}_{Z_1^n}(Z_1^n)$$

et l'estimateur du maximum de la quasi-log-vraisemblance est

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^{\theta}_{Y_1^n}(Z_1^n).$$

Autrement dit, ce dernier estimateur est obtenu en faisant comme si les observations n'étaient pas perturbées. En pratique, la quasi-vraisemblance est plus intéressante car elle ne nécessite pas de connaître la loi de la perturbation non homogène.

Le résultat théorique principal de l'article est que ces deux estimateurs sont consistants. La clé de la preuve est de montrer que les propriétés asymptotiques de $(Y_t)_{t\geqslant 1}$ se transfèrent au processus $(Z_t)_{t\geqslant 1}$, notamment l'ergodicité du processus et la convergence de la log-vraisemblance, ce qui permet d'adapter des preuves de consistance pour les HMMs homogènes.

Touron (2018) introduit des modèles de Markov cachés saisonniers, où la matrice de transition et les lois d'émission dépendent du temps de façon périodique. L'auteur montre que ce modèle est identifiable sous des hypothèses générales et que l'estimateur du maximum de vraisemblance est consistant. La preuve de la consistance repose sur une réécriture du processus sous forme de HMM homogène, et la preuve de l'identifiabilité se fonde sur une méthode spectrale.

Par souci d'exhaustivité, citons les modèles dynamiques non homogènes à régimes markoviens, étudiés par exemple par Ailliot and Pene (2015) et Pouzo et al. (2016). Ces modèles sont une généralisation des HMMs dans laquelle l'état caché X_t dépend à la fois de l'état caché précédent X_{t-1} et de l'observation précédente Y_{t-1}, et l'observation Y_t dépend à la fois de l'état caché au même instant X_t et de l'observation précédente. L'appellation « non homogène » vient du fait que le noyau de transition de la chaîne cachée $(X_t)_{t\geqslant 1}$ dépend des observations, ce qui en fait une chaîne de Markov non homogène. Toutefois, ce modèle est bien homogène au sens défini précédemment, c'est-à-dire que le processus joint $(X_t, Y_t)_{t\geqslant 1}$ est une chaîne de Markov homogène.

Le Chapitre 5 contient un travail en collaboration avec Augustin Touron. Nous y introduisons une nouvelle généralisation non homogène : les HMMs avec tendances. Cela permet de traiter des phénomènes non périodiques et non transitoires, tels que le réchauffement climatique si on s'intéresse à la température. Ces modèles sont des processus trivariés $(X_t, Y_t, Z_t)_{t\geqslant 1}$ à valeurs dans $\mathcal{X} \times \mathbb{R}^d \times \mathbb{R}^d$ pour un certain entier d où seul le processus $(Y_t)_{t\geqslant 1}$ est observé. On suppose de plus que $(X_t, Z_t)_{t\geqslant 1}$ est un HMM homogène et qu'il existe un vecteur de fonctions $(T_x)_{x\in\mathcal{X}}$ de $\mathbb{N}^*$ dans $\mathbb{R}^d$, les *tendances*, tel que

$$Y_t = T_{X_t}(t) + Z_t.$$

Notre approche permet de traiter des tendances polynomiales, qui peuvent donc diverger. Nous montrons que l'estimateur du maximum de vraisemblance retrouve tous les paramètres du HMM homogène $(X_t, Z_t)_{t\geqslant 1}$ ainsi que les tendances au sens de la convergence en norme infinie. L'analyse se décompose en deux étapes principales.

La première consiste à contrôler la partie divergente des tendances. Nous montrons qu'à partir d'un certain rang, chaque tendance estimée se trouve dans un tube de taille bornée autour d'une des vraies tendances. En regroupant les tendances selon le tube auquel elles appartiennent, il est donc possible de définir des « blocs » de tendances. Ces blocs $(B_t)_{t\geqslant 1}$ permettent de construires des observations détendancées

$$Z'_t = Z_t + \left[T^*_{X_t}(t) - T^\theta_{B_t}(t)\right],$$

où le terme entre crochets, la différence entre la vraie tendance de l'état ayant généré cette observation $T^*_{X_t})(t)$ et la tendance du paramètre θ associée au bloc de cet état, est borné.

La deuxième étape consiste à montrer que pour l'estimateur du maximum de vraisemblance, cette différence de tendances varie lentement au cours du temps. Intuitivement, cela signifie que le processus $(B_t, Z'_t)_{t\geqslant 1}$ est presque stationnaire, ce qui permet d'adapter les résultats existants de consistance pour les HMMs stationnaires. On en déduit à la fois l'identifiabilité du modèle et la consistance de l'estimateur du maximum de vraisemblance.

1.4 Summary

The contributions of my PhD are summarized in the following table. Each line corresponds to one of the topics I studied, and each column to one estimator.

	Least squares estimator	Spectral estimator	Maximum likelihood estimator
Order estimation	✔	✔	(1)
Minimax adaptive parameter estimation	✔	✔	(2)
State-by-state adaptivity		✔	
Misspecified models	(3)		✔
HMMs with trends			✔

The checked cells correpond to proved results. The three numbered cells have not been fully solved but deserve some remarks.

Cell (1) corresponds to the order estimation with the maximum likelihood estimator. The tools used for studying the least squares estimator can be adapted straighforwardly to prove that the MLE does not underestimate the order asymptotically. However, proving that the MLE does not overestimate the order in a general setting is much more challenging and is still an open question as far as we know.

Cell (2) corresponds to showing that the nonparametric MLE recovers the parameters at minimax and adaptive rates. To prove such a result, one would need to relate the error on the likelihood, which is what the proofs control, to the error on the parameters. Such a link is possible using the Fisher information matrix in a parametric setting, however it cannot be generalized to the nonparametric setting. This issue is what prevents us from checking this cell.

Cell (3) corresponds to guarantees on the least squares estimator for misspecified HMMs. While I did not write results about this case, all techniques used for the MLE can be adapted directly to the least squares estimator. Thus, one can expect a similar result to hold, that is that the $\mathbf{L}^2$ error on the density of three consecutive observations given by the estimator will converge to the lowest possible error within the model. Note that it estimates the distribution of a 3-uple, or more generally of a L-uple for some fixed integer L, so that it may not be able to recover long-term dependencies.

Finally, let us comment on the row labelled "state-by-state adaptivity". I introduce a new method to adapt to the regularity of each emission densities instead of adapting to the worst regularity. This method takes a family of auxiliary estimators as input, but it does not depend on how the auxiliary estimators are computed. Thus, any of the three methods studied here–or even other methods–can be used. On a theoretical side, the only requirement is that the auxiliary estimators satisfy an appropriate variance bound. I have shown that such a bound holds for the least squares and spectral estimators, and would hold for the MLE as well if the issue of cell (2) was solved.

CHAPTER

2

ORDER ESTIMATION AND GLOBALLY MINIMAX ADAPTIVE DENSITY ESTIMATION

This chapter is based on the accepted paper Lehéricy (2019), to appear in Bernoulli.

We consider the problem of estimating the number of hidden states (the order*) of a nonparametric hidden Markov model (HMM). We propose two different methods and prove their almost sure consistency without any prior assumption, be it on the order or on the emission distributions. This is the first time a consistency result is proved in such a general setting without using restrictive assumptions such as a priori upper bounds on the order or parametric restrictions on the emission distributions. Our main method relies on the minimization of a penalized least squares criterion. In addition to the consistency of the order estimation, we also prove that this method yields rate minimax adaptive estimators of the parameters of the HMM - up to a logarithmic factor. Our second method relies on estimating the rank of a matrix obtained from the distribution of two consecutive observations. Finally, numerical experiments are used to compare both methods and study their ability to select the right order in several situations.*

2.1 Introduction

2.1.1 Context and motivation

Hidden Markov models (HMM in short) are powerful tools to study time-evolving processes on heterogeneous populations. Nonparametric HMMs–that is, hidden Markov models where the parameters are not restricted to a finite-dimensional space–have proved useful in a wide range of applications, see for instance Couvreur and Couvreur (2000) for voice activity detection, Lambert et al. (2003) for climate state identification, Lefèvre (2003) for automatic speech recognition,

Shang and Chan (2009) for facial expression recognition, Volant et al. (2014) for methylation comparison of proteins, Yau et al. (2011) for copy number variants identification in DNA analysis.

In practice, the hidden states often have an interpretation in the modelling of the phenomenon. It is thus important to be able to infer the right order in addition to the parameters when dealing with hidden Markov models. However, this task is notoriously difficult: Gassiat and Keribin (2000) show that the likelihood ratio statistic is unbounded even in the simple case where one wants to test if a HMM has 1 or 2 hidden states. As far as we know, no consistency result has been proved about order selection for nonparametric HMMs. Even for parametric HMMs, no estimator has been proved to be consistent in a general setting without assuming that an *a priori* upper bound on the order is known beforehand.

Not only is the order estimation useful in order to interpret the model, it is also necessary to ensure stability. This is because over estimating the order causes a loss of identifiability: there are several ways to add one state to a HMM without changing anything to its distribution. The spectral estimators (Anandkumar et al. (2012); De Castro et al. (2017)) are especially sensitive to this problem, as shown by Lehéricy (2015) and Figure 2.6: as soon as the HMM becomes close to a HMM with fewer hidden states, the estimators give absurd results. Thus, estimating the right order is crucial for such methods to be effective.

Formally, a hidden Markov model is a markovian process $(X_t, Y_t)_{t\geqslant 1}$ taking value in $\mathcal{X}\times\mathcal{Y}$. $(X_t)_{t\geqslant 1}$ is a Markov chain and the observations Y_t depend only on the associated X_t (i.e. the $(Y_t)_{t\geqslant 1}$ are independent conditionally on $(X_t)_{t\geqslant 1}$). The states $(X_t)_{t\geqslant 1}$ are assumed to be hidden, so that one has only access to the observations $(Y_t)_{t\geqslant 1}$. When the number of hidden states $|\mathcal{X}|$ (which we call the *order* of the HMM) is finite, the model is completely defined by its order, the initial distribution and the transition matrix of the hidden Markov chain, and the possible distributions of an observation Y_t conditionally to the values of its hidden state X_t, which we call the *emission distributions*. The goal of the estimation procedures is to recover these parameters by using only the observations $(Y_t)_{t\geqslant 1}$.

Up to now, most theoretical results on hidden Markov models dealt with the parametric frame, that is with a finite number of parameters. However, it is not always possible to restrict the model to such a convenient finite-dimensional space. Theoretical results in the nonparametric framework were only developed recently and do not address the order estimation problem. de Castro et al. (2016) propose an adaptive quasi-rate minimax least squares method. De Castro et al. (2017) and Bonhomme et al. (2016b) study spectral methods. The latter is also proved to reach the minimax convergence rate but is not adaptive: it requires the regularity of the emission distributions to be known. All these methods require the order of the HMM to be known.

Our work is novel on three points. First, it deals with the nonparametric setting: we need no parametric or regularity assumption on the emission densities. Note that all our results also apply to parametric settings or even to finite observation spaces, since these are just special cases of nonparametric estimation. Secondly, we do not require any *a priori* upper bound on the order, an assumption that is often made in earlier works, both frequentist and bayesian. Finally, our least squares method yields estimators of all model parameters at the same time, without requiring any prior information. Oracle inequalities show that these estimators are rate minimax adaptive up to a logarithmic factor.

2.1.2 Related works

The first step to obtain theoretical results was to understand when hidden Markov models are identifiable. This challenging issue was only solved a few years ago, see Gassiat et al. (2015) (following Allman et al. (2009) and Hsu et al. (2012)) and with weaker assumptions Alexandrovich et al. (2016). Both proved that under generic assumptions, the parameters of

the HMM can be recovered from the distribution of a finite number of consecutive observations, thus paving the way for guarantees on parameter estimation.

HMM inference is generally decomposed in two parts. The first one is the estimation of the order, and the second one is the estimation of the parameters once the order is known.

From a theoretical point of view, the order estimation problem remains widely open in the HMM framework. One can distinguish two kinds of results. The first kind does not need an *a priori* upper bound on the order, but is only applicable to restrictive cases. For instance, using tools from coding theory, Gassiat and Boucheron (2003) introduced a penalized maximum likelihood order estimator for which they prove strong consistency without *a priori* upper bound on the order of the HMM. Nevertheless, their result is restricted to a finite observation space and they have to use heavy penalties that grow as a power of the order. For the special case of Gaussian or Poisson emission distributions, Chambaz et al. (2009) showed that the penalized maximum likelihood estimator is strongly consistent without any *a priori* upper bound on the order. The second kind of results is more general but requires an *a priori* upper bound of the order just to get weak consistency of order estimators, for penalized likelihood criterion (Gassiat (2002)) as well as Bayesian approaches (Gassiat and Rousseau (2014); van Havre et al. (2016)).

On a practical side, several order estimation methods using penalized likelihood criterion have been studied numerically, see for instance Volant et al. (2014) when emission distributions are a mixture of parametric densities or Celeux and Durand (2008) for parametric HMMs. The latter also introduced cross-validation procedures that aimed for circumventing the lack of independance of the observations. In the case of nonparametric HMMs, Langrock et al. (2015) studied a method using P-splines with a custom penalization.

Then comes the question of estimating the parameters of the HMM once its order is known. In the parametric setting, the asymptotic behaviour of the maximum likelihood estimator is rather well understood (see for instance Bickel et al. (1998) or Douc et al. (2004) using techniques from Le Gland and Mevel (2000)), but so far the question of its nonasymptotic behaviour remains open. Hsu et al. (2012) and Anandkumar et al. (2012) proposed a spectral method for parametric HMMs based on joint diagonalization of a set of matrices and controlled its nonasymptotic error. Bonhomme et al. (2016b) and De Castro et al. (2017) extended this method to the nonparametric setting, and de Castro et al. (2016) used the latter to obtain an estimator of the transition matrix of the hidden chain for a quasi-rate minimax adaptive least squares estimator of the emission densities. Our least squares estimation method is a generalization of their procedure that is able to deal with all parameters at once and does not require auxiliary estimators.

2.1.3 Contribution

The aim of this chapter is twofold. Firstly, we introduce two estimators of the order for nonparametric HMMs and show that both converge almost surely to the right order under minimal assumptions. Secondly, we numerically assess their ability to select the right order and compare their efficiency.

Our first and main method is the penalized least squares estimator. This method is based on estimating the projection of the emission distributions onto a family of nested parametric subspaces. Our results hold for any Hilbert space, including parametric sets of emission densities and finite observation spaces. Then, for each subspace and for each possible value K of the order, we look for the HMM with K hidden states and with emission distributions in the chosen subspace that matches the observations "best"–where "best" means minimizing the empirical equivalent of an $\mathbf{L}^2$ distance. This step provides an empirical distance between the observations and the model, which is then penalized in order to counterbalance the overfitting phenomenon that occurs when considering large models. Our first main result is that for a suitable choice of

the penalty, choosing the model (i.e. the order and the subspace) which minimizes this penalized distance leads to a strongly consistent estimator of the order, see Corollary 2.5.

In addition, this method also provides estimators of the other parameters of the HMM for free, by taking the parameters of the HMM corresponding to the selected model. We prove an oracle inequality on the $\mathbf{L}^2$ risk of these estimators, which shows that they achieve the minimax adaptive rate of convergence, up to a logarithmic term, see Theorem 2.10 and Corollary 2.11.

Our second estimator comes from spectral methods. Just like for our least squares procedure, we consider a nested family of parametric subspaces of a Hilbert space. Let us choose one of them, and denote by $(\varphi_a)_a$ an orthonormal basis of this subspace. Then, consider the matrix $\mathbf{N}$ defined by

$$\mathbf{N}(a,b) := \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)].$$

This matrix contains the coordinates of the density of (Y_1, Y_2) in the orthonormal basis $(\varphi_a \otimes \varphi_b)_{a,b}$. It is proved in Section 2.4 that the rank of $\mathbf{N}$ is exactly equal to the order of the HMM as soon as the subspace is large enough. Therefore, finding its rank means finding the number of hidden states. However, in practice, one only has access to an empirical version of this matrix. The difficulty comes from the fact that this noisy version will almost surely have full rank. Thus, the key point is to recover the order of the true matrix given its empirical (full rank) counterpart. We achieve this by thresholding the spectrum of the empirical matrix. Notice that other methods exist to estimate the rank of a matrix based on a noisy observation, see for instance Kleibergen and Paap (2006) and references therein. Unfortunately, most can not be applied directly to our setting since they require an invertibility condition on the covariance matrix of the matrix entries. The CRT statistics from Robin and Smith (2000) is a notable exception, however their test of rank also requires the tuning of an unknown parameter in order to be weakly consistent.

Then, we run an implementation of these two methods and compare their efficiency on simulated data. The difficulty at this stage comes from the fact that both method involve an unknown tuning parameter. This is a common issue that appears in every model selection method in one form or another, and many heuristics have been proposed to circumvent this difficulty.

For the least squares estimator, we compare two methods which have been both proved to be theoretically valid in simple cases and empirically validated in a large variety of situations: the slope heuristics (see for instance Baudry et al. (2012) and references therein) and the dimension jump heuristics (introduced and proved to lead to an optimal penalization in the gaussian model selection framework by Birgé and Massart (2007)). Both behave well with our estimator and lead to a satisfying calibration of the penalty.

For the spectral estimator, we introduce a custom heuristics based on the fact that the smallest singular values of the empirical version of the matrix $\mathbf{N}$ decrease in a simple manner. It is thus possible to calibrate an entirely data-driven threshold to distinguish "significant" singular values–that is, the ones corresponding to non-zero singular values of the real $\mathbf{N}$–from noise.

The numerical validation shows that our least squares method performs well in almost any situation. It is able to select the right order accurately with notably fewer observations than the spectral estimator, and is easier to calibrate. On the other hand, the spectral method is very fast, which allows to take more observations into account. This allows to obtain satisfying estimators in a short amount of time.

Regarding the inference of the other parameters, our least squares estimator offers several advantages when compared to previous methods. First, it does not need a preliminary estimation of the transition matrix or of the order, unlike de Castro et al. (2016) who used the transition matrix given by spectral estimators. Nevertheless, our method still reaches the adaptive minimax

convergence rate for the estimation of the emission densities, up to a logarithmic factor. This is especially useful to avoid the cases where their auxiliary estimator fails. For instance, the spectral method that de Castro et al. (2016) used is unreliable when the order is over estimated or when the states are almost linearly dependent, see for instance Lehéricy (2015) or Figure 2.6. Then, our least squares method is robust to an overestimation of the order, both theoretically and numerically, thanks to the iterative initialization procedure that we introduce. This initialization method consists in using estimators from smaller models as initial point for the minimization algorithm in order to avoid getting stuck in suboptimal local extrema. We believe it can be of practical interest since it produces robust estimators and can also be used in other settings, for instance as initialization for expectation maximization algorithm for maximum likelihood estimators.

2.1.4 Outline of the chapter

This chapter is organized as follows.

Section 2.2 is devoted to the notations, the model and the assumptions.

Our main procedure, the penalized least squares method, is introduced in Section 2.3. We first state an identifiability proposition which we use to prove strong consistency of the estimator of the order. This is done in two steps. Firstly, we control the probability to underestimate the order. This is done thanks to Proposition 2.1, and gives an exponential bound on the probability of error, see Theorem 2.3. Secondly, we control the probability to overestimate the order, see Theorem 2.4. For this, we introduce a general condition on the penalty, which we use to prove polynomial decrease rate, and illustrate how to easily satisfy this condition. Finally, we state oracle inequalities on the estimators of the density of L consecutive observations and on the parameters of the hidden Markov model under a generic assumption, see Theorem 2.10 and Corollary 2.11, which shows that they reach the minimax convergence rate up to a logarithmic factor.

In Section 2.4, we introduce the spectral algorithm and propose a strongly consistent estimator of the order. This is done by thresholding the spectrum of the empirical version of the matrix $\mathbf{N}$, which describes the projection of the distribution of two observations on an orthonormal basis, see Theorem 2.13.

In Section 2.5, we propose practical algorithms to apply both methods and compare them. Firstly, we set the parameters on which we will test both procedures. Secondly, we compare their results and discuss their performance. Lastly, we introduce and discuss the heuristics we used to practically implement both methods.

Our main technical result, Lemma 2.16, can be found at the beginning of Section 2.6. It is used extensively for both the consistency of the estimator of the order and the oracle inequalities on the HMM parameters. The rest of this section is dedicated to the proofs of the results.

Appendix 2.A contains the spectral algorithm from De Castro et al. (2017) and de Castro et al. (2016) that we use in our simulations. Appendix 2.B gathers the proofs of Section 2.3.4, which deals with the oracle inequalities for the least squares method. Finally, Appendix 2.C contains the proof of Lemma 2.16, and Appendix 2.D contains miscellaneous lemmas and proofs.

2.2 Definitions and assumptions

We will use the following notations throughout the chapter.

- $\Delta_K = \{\pi \in [0,1]^K | \sum_{k=1}^K \pi_k = 1\}$ is the simplex in dimension K. It will be seen as the set of probability measures on a finite set of size K.

- $\mathcal{Q}_K \subset \mathbb{R}^{K\times K}$ is the set of irreducible transition matrices of size K.
- The notation $C \equiv C(a, b, \dots)$ for a constant C will mean that the value of C depends on the specified parameters a, b, ... For several constants depending on the same parameters, we will write $(C, D) \equiv (C, D)(a, b, \dots)$.

In the following, L is a positive integer which will denote the number of consecutive observations used for the estimation procedure.

2.2.1 Hidden Markov models

Let $(X_j)_{j\geqslant 1}$ be a Markov chain with finite state space $\mathcal{X}$ of size K^* with transition matrix $\mathbf{Q}^*$ and initial distribution π^*. Without loss of generality, we can set $\mathcal{X} = [K^*]$.

Let $(Y_j)_{j\geqslant 1}$ be random variables on a measured space $(\mathcal{Y}, \mu)$ with μ σ-finite such that conditionally on $(X_j)_{j\geqslant 1}$ the Y_j's are independent with a distribution depending only on X_j. Let ν_k^* be the distribution of Y_j conditionally to $\{X_j = k\}$. Assume that ν_k^* has density $f_k^* \in \mathbf{L}^2(\mathcal{Y}, \mu)$ with respect to μ. We call $(\nu_k^*)_{k\in\mathcal{X}}$ the *emission distributions* and $\mathbf{f}^* = (f_1^*, \dots, f_{K^*}^*)$ the *emission densities*.

Then $(X_j, Y_j)_{j\geqslant 1}$ is a hidden Markov model with parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, K^*)$. The hidden chain $(X_j)_{j\geqslant 1}$ is assumed to be unknown, so that the estimator only has access to the observations $(Y_j)_{j\geqslant 1}$.

For $K \in \mathbb{N}^*$, $\pi \in \mathbb{R}^K$, $\mathbf{Q} \in \mathbb{R}^{K\times K}$ and $\mathbf{f} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$, let

$$g^{\pi,\mathbf{Q},\mathbf{f},K} = \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1}, k_i) \bigotimes_{i=1}^{L} f_{k_i}.$$

When π is a probability distribution on $[K]$, $\mathbf{Q}$ a $K \times K$ transition matrix and $\mathbf{f}$ a K-uple of probability densities, $g^{\pi,\mathbf{Q},\mathbf{f},K}$ is the density of the first L observations of a HMM with parameters $(\pi, \mathbf{Q}, \mathbf{f}, K)$.

For the sake of readability, we will drop the dependence in K in the following and write $g^{\pi,\mathbf{Q},\mathbf{f}}$ instead of $g^{\pi,\mathbf{Q},\mathbf{f},K}$. Moreover, if $\mathbf{Q}$ is irreducible with stationary distribution π, we simply write $g^{\mathbf{Q},\mathbf{f}}$, and we write the true density $g^* := g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}$.

2.2.2 Assumptions

Let $\mathcal{F}$ be a subset of $\mathbf{L}^2(\mathcal{Y}, \mu)$ and $(\mathfrak{P}_M)_{M\in\mathcal{M}\subset\mathbb{N}}$ be a sequence of nested subspaces of $\mathbf{L}^2(\mathcal{Y}, \mu)$ such that $\mathfrak{P}_M$ has dimension M for all $M \in \mathcal{M}$ and their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mu)$. $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ will be the subspaces on which the projections of the emission densities will be estimated.

We will need the following assumptions.

[HX] $(X_k)_{k\geqslant 1}$ is a stationary ergodic Markov chain with parameters $(\pi^*, \mathbf{Q}^*)$;

[HidA] $\mathbf{Q}^*$ is invertible, $L \geqslant 3$ and the family $\mathbf{f}^*$ is linearly independent;

[HidB] $\mathbf{Q}^*$ is invertible, $L \geqslant (2K^*+1)((K^*)^2-2K^*+2)+1$ and the emission densities $(f_k^*)_{k\in\mathcal{X}}$ are all distinct;

[HF] $\mathbf{f}^* \in \mathcal{F}^{K^*}$, $\mathcal{F}$ is closed under projection on $\mathfrak{P}_M$ for all M and

$$\forall f \in \mathcal{F}, \quad \begin{cases} \|f\|_\infty \leqslant C_{\mathcal{F},\infty} \\ \|f\|_2 \leqslant C_{\mathcal{F},2} \end{cases}$$

with $C_{\mathcal{F},\infty}$ and $C_{\mathcal{F},2}$ larger than 1.

The ergodicity assumption in **[HX]** is completely standard in order to obtain convergence results. In this case, the initial distribution is forgotten exponentially fast, so that the HMM will essentially behave like a stationary process. In order to simplify the proofs, we assume the Markov chain to be stationary. One can check that our results are essentially the same when the initial distribution is not the stationary one.

[HidA] appears in spectral methods, with the hypothesis that $\pi^* > 0$ elementwise, see for instance Hsu et al. (2012). **[HidA]** and **[HidB]** also appear in identifiability issues, possibly combined with the stationarity hypothesis, see Alexandrovich et al. (2016) and Gassiat et al. (2015). Note that the condition on L in **[HidB]** only involves the real order K^*.

Even though **[HidB]** appears less restrictive than **[HidA]** about the emission densities, it is delicate to use here. The problem lies in the condition on the number of consecutive observations L. For **[HidB]**, one has to take L larger than an increasing function of the order, so it requires to have an *a priori* upper bound on the order to choose L. This is less interesting than **[HidA]**, which can work without prior bound since it only requires $L = 3$ for any value of the order.

2.3 Least squares estimation

In this section, we introduce our penalized least squares estimator and study its asymptotic properties.

2.3.1 Approximation spaces and estimators

We want to estimate the density of L consecutive observations g^* by minimizing the quadratic loss $t \mapsto \|t - g^*\|_2^2 - \|g^*\|_2^2$. We thus take the corresponding empirical loss

$$\gamma_n(t) = \|t\|_2^2 - \frac{2}{n}\sum_{s=1}^{n} t(Z_s)$$

where $Z_s = (Y_s, \dots, Y_{s+L-1})$ for an observation sequence $(Y_t)_{1 \leqslant t \leqslant n+L-1}$ of length $n + L - 1$ coming from a single HMM $(X_t, Y_t)_{t \geqslant 1}$.

Define for all $K \in \mathbb{N}^*$, $M \in \mathcal{M}$:

$$S_{K,M} := \{g^{\mathbf{Q},\mathbf{f}}, \mathbf{Q} \in \mathcal{Q}_K, \mathbf{f} \in (\mathcal{F} \cap \mathfrak{P}_M)^K\}$$
$$S_K := \{g^{\mathbf{Q},\mathbf{f}}, \mathbf{Q} \in \mathcal{Q}_K, \mathbf{f} \in \mathcal{F}^K\}$$

where $\mathcal{F}$ and $(\mathfrak{P}_M)_{M \in \mathcal{M}}$ are defined in Section 2.2.2. In the following, we will always implicitly consider $M \in \mathcal{M}$.

For all K and M, we define the corresponding estimators

$$\hat{g}_{K,M} = g^{\hat{\mathbf{Q}}_{K,M}, \hat{\mathbf{f}}_{K,M}} \in \operatorname*{arg\,min}_{t \in S_{K,M}} \gamma_n(t)$$

where we dropped the dependency in n for ease of notation. Then, we select the parameters using the penalized empirical loss:

$$(\hat{K}_{\text{l.s.}}, \hat{M}) \in \operatorname*{arg\,min}_{K \leqslant n,\, M \leqslant n} \{\gamma_n(\hat{g}_{K,M}) + \text{pen}(n, M, K)\}$$

which leads to the estimators

$$\hat{g} := \hat{g}_{\hat{K}_{\text{l.s.}}, \hat{M}}$$
$$\hat{\mathbf{Q}} := \hat{\mathbf{Q}}_{\hat{K}_{\text{l.s.}}, \hat{M}}$$
$$\hat{\mathbf{f}} := \hat{\mathbf{f}}_{\hat{K}_{\text{l.s.}}, \hat{M}}$$

2.3.2 Underestimation of the order

Note that the distribution of the HMM remains unchanged under permutation of the hidden states. We will therefore use a pseudo-distance d_{perm} that is invariant by permutation on the set of parameters.

We define it as follows. Let $K \geqslant 1$, $\pi_1, \pi_2 \in \Delta_K$, $\mathbf{Q}_1$ and $\mathbf{Q}_2$ transition matrices of size K, $\mathbf{f}_1, \mathbf{f}_2 \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$. Let $\mathfrak{S}(\mathcal{X})$ be the set of permutations of $\mathcal{X}$. For all $\tau \in \mathfrak{S}(\mathcal{X})$, define the swapped parameters $\tau\pi_1$, $\tau\mathbf{Q}_1$ and $\tau\mathbf{f}_1$ by

$$\begin{aligned}(\tau\pi_1)(k) &:= \pi_1(\tau(k)) \\ (\tau\mathbf{Q}_1)(k,l) &:= \mathbf{Q}_1(\tau(k), \tau(l)) \\ (\tau\mathbf{f}_1)_k &:= f_{1,\tau(k)}\end{aligned}$$

and finally

$$d_{\text{perm}}((\pi_1, \mathbf{Q}_1, \mathbf{f}_1), (\pi_2, \mathbf{Q}_2, \mathbf{f}_2)) := \inf_{\tau \in \mathfrak{S}(\mathcal{X})} \left(\|\tau\pi_1 - \pi_2\|_2^2 + \|\tau\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2 + \sum_{k=1}^{K} \|(\tau\mathbf{f}_1)_k - f_{2,k}\|_2^2 \right)^{1/2}.$$

The following properties will be of use to prove the consistency of the order estimator, but we think it can also be of independent interest to better understand the identifiability of the model. The first one is a generalization of previous identifiability results from Alexandrovich et al. (2016); Gassiat et al. (2015); de Castro et al. (2016).

Proposition 2.1. *Let $K \geqslant 1$, $\pi \in \Delta_K$ such that $\pi_k > 0$ for all $k \in \mathcal{X}$, $\mathbf{Q}$ transition matrix of size K and $\mathbf{f} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$ such that **[HidA]** or **[HidB]** hold for the order K. Then, for all $K' \geqslant 1$, for all $\pi' \in \Delta_{K'}$, for all transition matrix $\mathbf{Q}'$ of size K' and all $\mathbf{f}' \in (\mathbf{L}^2(\mathcal{Y}, \mu))^{K'}$, the following holds:*

$$\left(g^{\pi,\mathbf{Q},\mathbf{f}} = g^{\pi',\mathbf{Q}',\mathbf{f}'} \text{ and } K' \leqslant K\right) \Rightarrow \left(K = K' \text{ and } d_{perm}((\pi, \mathbf{Q}, \mathbf{f}), (\pi', \mathbf{Q}', \mathbf{f}')) = 0\right).$$

Remark. *This property does not require two assumptions that appear in Alexandrovich et al. (2016) and Gassiat et al. (2015): that $\mathbf{f}$ is a family of probability densities and that the Markov chain is stationary.*

In particular, the fact that $\mathbf{f}$ may not be a family of probability densities is crucial in the proof of Corollary 2.2, which is necessary to prove the strong consistency of the estimator of the order.

Proof. Assume **[HidA]**. The spectral algorithm from De Castro et al. (2017) applied on the linear space spanned by both sets of densities allows to retrieve the order from two consecutive observations and the parameters from three consecutive observations. Their proof works when the emission densities are not probability densities and when the chain is not stationary.

Assume **[HidB]**. A careful reading of the proofs of Alexandrovich et al. (2016) shows that their result can be extended to general observation spaces and do not require the measures to be probabilities. □

The second property is the following corollary, which states that the $\mathbf{L}^2$ distance between the actual model and the models where the order is underestimated is positive. It is worth noting that we do not need $\mathcal{F}$ to be compact.

Corollary 2.2. *Assume **[HX]**, (**[HidA]** or **[HidB]**) and **[HF]** hold. Then, for all $K < K^*$:*

$$d_K := \inf_{t \in S_K} \|t - g^*\|_2 > 0$$

Proof. Proof in Section 2.6.2. □

Our first theorem shows that the probability to underestimate the order decreases exponentially with the number of observations. This comes from Corollary 2.2: since the empirical criterion converges to the $\mathbf{L}^2$ distance (plus some constant that does not depend on the model), the penalized error will eventually become larger for orders under K^* than for orders over K^*, which means that we won't underestimate the real order. The exponential decrease rate brings to mind the one studied in Gassiat and Boucheron (2003): in both cases, the exponents involve the distance between the actual model and models with underestimated orders, as can be seen in our proof.

Theorem 2.3. *Assume **[HX]**, (**[HidA]** or **[HidB]**) and **[HF]** hold. There exists positive constants $\rho \equiv \rho(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ and $\beta \equiv \beta(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, (d_K)_{K<K^*}, L)$ such that the following holds.*

Assume that

$$\forall n,\ \forall M,\ \forall K, \quad pen(n, M, K) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n},$$

and

$$\forall M,\ \forall K, \quad pen(n, M, K) \underset{n\to\infty}{\longrightarrow} 0$$

then there exists n_0 such that for all $n \geqslant n_0$,

$$\mathbb{P}(\hat{K}_{l.s.} < K^*) \leqslant e^{-\beta n}.$$

Proof. Proof in Section 2.6.3. □

2.3.3 Overestimation of the order and consistency

Our second theorem controls the probability to overestimate the order. It consists in overpenalizing large models so that the estimated order remains small.

We will need the following technical condition on the penalty:

Condition (**[Hpen]**(α, ρ))**.** *The penalty function pen satisfies*

$$\exists n_1,\ \forall n \geqslant n_1,\ \forall M \leqslant n,\ \forall K \leqslant n \text{ s.t. } K > K^*,$$
$$pen(n, M, K) - pen(n, M, K^*) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n} + \alpha\frac{\log(n)}{n},$$

We can now state the theorem and its corollary proving the strong consistency of our estimator of the order. Note that it does not require any identifiability assumption.

Theorem 2.4. *Assume **[HX]** and **[HF]** hold. There exists positive constants $(\rho, \beta) \equiv (\rho, \beta)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ such that the following holds.*

*Assume **[Hpen]**(α, ρ) holds for some $\alpha \geqslant 0$, then there exists n_0 such that for all $n \geqslant n_0$,*

$$\mathbb{P}(\hat{K}_{l.s.} > K^*) \leqslant n^{-\beta\alpha}.$$

Proof. Proof in Section 2.6.3. □

Corollary 2.5. *Assume **[HX]**, **[HF]** and (**[HidA]** or **[HidB]**) hold. There exists positive constants $(\rho, \beta) \equiv (\rho, \beta)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ such that the following holds.*

Assume that the penalty function satisfies

$$\begin{cases} \forall n,\ \forall M \leqslant n,\ \forall K \leqslant n, \quad pen(n,M,K) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n} \\ \forall M,\ \forall K, \quad pen(n,M,K) \underset{n\to+\infty}{\longrightarrow} 0 \end{cases}$$

*and **[Hpen]**$(\alpha/\beta, \rho)$ holds for some $\alpha > 1$, then*

$$\mathbb{P}(\hat{K}_{l.s.} \neq K^*) = O(n^{-\alpha}).$$

In particular, $\hat{K}_{l.s.} \longrightarrow K^*$ almost surely.

Let us comment on the condition **[Hpen]** when using a penalty of the form $\text{pen}(n, M, K) = C(MK + K^2 - 1)\log(n)/n$ where C may depend on n.

- If one has an *a priori* bound on the order, i.e. if $K^* \leqslant K_0$ for some known K_0, then direct computations show that for all α, ρ, there exists $C \geqslant 0$ depending on K_0 (for instance, $C = 2\rho(1 + K_0^2 \vee \frac{\alpha}{\rho})$ works) such that **[Hpen]**(α, ρ) holds for all $K^* \leqslant K_0$ (instead of $K \leqslant n$). This means that if one has an *a priori* bound K_0 on the order, then by taking a constant C large enough and $\hat{K}_{\text{l.s.}} \leqslant K_0$, the estimator $\hat{K}_{\text{l.s.}} > K^*$ is almost surely consistent.

- If one does not have an *a priori* bound on K^*, taking a constant C does not allow to get **[Hpen]**(α, ρ) for all possible K^*, which means we can't apply Corollary 2.5. However, by taking C as a sequence indexed by n that tends to infinity, we get that for all K^* and α, ρ, **[Hpen]**(α, ρ) holds. This implies consistency with polynomial decrease of the probability of error, at the cost of overpenalizing.

  Overpenalizing is actually necessary if one wants to satisfy **[Hpen]** for all K^*. This is stated in the following proposition:

Proposition 2.6. *Let $\rho > 0$ and pen be a positive penalty such that for all K^*, **[Hpen]**$(0, \rho)$ holds, then there exists a sequence $(u_n)_{n\geqslant 1} \longrightarrow \infty$ such that for all $n \geqslant 1$, $M \leqslant n$ and $K \leqslant n$, $pen(n, M, K) \geqslant u_n(MK + K^2 - 1)\log(n)/n$.*

Proof. Proof in Appendix 2.D.1. □

2.3.4 Oracle inequalities

Our first result for this section is an oracle inequality on the density of L consecutive observations for the least squares estimator.

Theorem 2.7. *Assume **[HX]** and **[HF]** hold. Then there exists positive constants $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ such that if the penalty satisfies*

$$\forall n,\ \forall M \leqslant n,\ \forall K \leqslant n, \quad pen(n,M,K) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n}$$

then for all $n \geqslant n_0$, for all $x > 0$, it holds with probability larger than $1 - e^{-x}$ that

$$\|\hat{g} - g^*\|_2^2 \leqslant 3 \inf_{K\leqslant n,\, M\leqslant n} \left\{\|g^*_{K,M} - g^*\|_2^2 + 2pen(n,M,K)\right\} + 4A\frac{x}{n}.$$

Proof. Proof in Section 2.B.1. □

Remark. *The constant 3 before the infimum can be replaced by any constant $\kappa > 1$, at the cost of changing the constants n_0, ρ and A.*

We would like to deduce an oracle inequality on the parameters of the HMM from this result. Using Cauchy-Schwarz inequality, it is easy to upper bound the error on the density g^* by the error on the parameters: for all probability distributions π_1 and π_2 on $[K]$, for all transition matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ of size K and for all $\mathbf{f}_1, \mathbf{f}_2 \in \mathcal{F}^K$,

$$\|g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2}\|_2 \leqslant C^L_{\mathcal{F},2}\sqrt{LK}\, d_{\mathrm{perm}}((\pi_1,\mathbf{Q}_1,\mathbf{f}_1),(\pi_2,\mathbf{Q}_2,\mathbf{f}_2)) \tag{2.1}$$

as soon as **[HF]** holds. The proof of this equation is detailed in Section 2.B.2.

Thus, all we need to deduce an oracle inequality on the parameters is to lower bound the error on g^* by the error on the parameters. Let $\mathfrak{C} \subset \mathbb{R}^{K^*} \times \mathbb{R}^{K^*\times K^*} \times \mathbb{R}^{K^*\times K^*}$ be the set of parameters (p, q, A) such that

$$\begin{cases} \forall i \in \mathcal{X}, & \sum_{j\in\mathcal{X}} q(i,j) = 0 \\ \forall j \in \mathcal{X}, & \sum_{i\in\mathcal{X}} A(i,j) = 0 \end{cases} \tag{2.2}$$

Note that $\mathfrak{C}$ can be identified with the set

$$\begin{aligned} \mathfrak{C}_{\mathrm{red}} :=& \{((p_i)_{i\geqslant 2}, (q(i,j))_{i,j\geqslant 2}, (A(i,j))_{i\geqslant 2,j}) \mid (p,q,A) \in \mathfrak{C}\} \\ =& \mathbb{R}^{K^*-1} \times \mathbb{R}^{K^*\times(K^*-1)} \times \mathbb{R}^{K^*\times(K^*-1)} \end{aligned}$$

These assumptions are natural since they are necessary (but not sufficient) to ensure that if $(p, q, A) \in \mathfrak{C}$ and π is a probability distribution, $\mathbf{Q}$ a transition matrix and $\mathbf{f}$ a vector of probability densities, then $\pi + p$ is also a probability distribution, $\mathbf{Q} + q$ a transition matrix and $\mathbf{f} + A\mathbf{f}$ a vector of probability densities.

The first step in order to get a lower bound along the same lines as equation (2.1) is to control the behaviour of the difference near the true parameters, which comes down to proving that the quadratic form M derived from the second-order expansion of

$$\mathfrak{N} : (p,q,A) \in \mathbb{R}^{K^*} \times \mathbb{R}^{K^*\times K^*} \times \mathbb{R}^{K^*\times K^*} \longmapsto \|g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}}\|_2^2$$

is positive definite on $\mathfrak{C}$ for $(\pi, \mathbf{Q}, \mathbf{f}) = (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. One can write the coefficients of the matrix of this quadratic form as polynomials in the coefficients of π, $\mathbf{Q}$ and of the Gram matrix $G(\mathbf{f}) := (\langle f_i, f_j\rangle)_{i,j\in\mathcal{X}}$. However, this matrix may not be invertible: one has to consider its restriction to the space $\mathfrak{C}$, which is equivalent to considering the quadratic form $M_{\mathfrak{C}}$ defined on $\mathfrak{C}_{\mathrm{red}}$ by the second-order expansion of $x \in \mathfrak{C}_{\mathrm{red}} \longmapsto \mathfrak{N}(I_{\mathfrak{C}}(x))$ where $I_{\mathfrak{C}}$ is the natural linear injection from $\mathfrak{C}_{\mathrm{red}}$ to $\mathfrak{C}$ (note that $I_{\mathfrak{C}}$ is bijective and bicontinous under **[HF]**). Since the quadratic form $M_{\mathfrak{C}}$ is always nonnegative, we only need its determinant to be non zero in order for the quadratic form M to be positive definite on $\mathfrak{C}$.

Thus, let H be determinant of the matrix of this quadratic form. H is also a polynomial in the coefficients of π, $\mathbf{Q}$ and $G(\mathbf{f})$. The following lemma shows that there exists some parameters $\pi, \mathbf{Q}$ and $\mathbf{f}$ satisfying the conditions for which H is not zero.

Lemma 2.8. *There exists some parameters $(\pi, \mathbf{Q}, \mathbf{f})$ satisfying the conditions **[HX]** and **[HidA]** such that $H(\pi, \mathbf{Q}, G(\mathbf{f})) \neq 0$.*

Proof. Proof in Section 2.B.3. □

What should be retained from this lemma is that H is a polynomial which is not identically zero on the set of parameters satisfying the identifiability conditions. This means that one can generically assume it to be different from zero, which corresponds to the assumption

[Hdet] $H(\pi^*, \mathbf{Q}^*, G(\mathbf{f}^*)) \neq 0$.

Since we assumed π^* to be the stationary distribution of $\mathbf{Q}^*$, its coefficients–and by extension H–can be expressed as a rational function of the coefficients of $\mathbf{Q}^*$. Taking H_1 as the numerator of the rational function deduced from H, one gets another polynomial in the coefficients of $\mathbf{Q}^*$ and $G(\mathbf{f}^*)$ which is also non-zero. Thus, the following assumption–which we will need to lower bound the error on the density g^* by the error on the parameters–is generically satisfied.

[HdetStat] $H_1(\mathbf{Q}^*, G(\mathbf{f}^*)) \neq 0$.

Note that **[Hdet]** and **[HdetStat]** are equivalent under the assumption **[HX]**.

Theorem 2.9. *Assume **[HidA]** and **[Hdet]** hold. Then there exists a positive constant $c(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ such that for all $\pi \in \Delta_{K^*}$, for all transition matrix $\mathbf{Q}$ of size K^* and for all $\mathbf{h} \in \mathcal{F}^{K^*}$ such that $\int h_i d\mu = 1$ for all $i \in [K^*]$,*

$$\|g^{\pi,\mathbf{Q},\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \geqslant c(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)\, d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*))^2.$$

Proof. Proof in Section 2.B.4. □

The following theorem is a direct consequence of the above results. It provides an oracle inequality on the parameters conditionally to the fact that the order has been correctly estimated.

Theorem 2.10. *Assume **[HX]**, **[HidA]**, **[HF]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$.*

Then there exists positive constants $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ such that if the penalty satisfies

$$\forall n,\ \forall M \leqslant n,\ \forall K \leqslant n, \quad pen(n, M, K) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n}$$

then for all $n \geqslant n_0$, for all $x > 0$, conditionally to $\{\hat{K}_{l.s.} = K^*\}$, with probability larger than $1 - e^{-x}$:

$$d_{perm}((\hat{\pi}, \hat{\mathbf{Q}}, \hat{\mathbf{f}}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)) \leqslant \frac{4C_{\mathcal{F},2}^L\sqrt{LK^*}}{c(\mathbf{Q}^*, \mathbf{f}^*)} \times \left[\inf_{M \leqslant n}\left\{\sum_{k=1}^{K^*} \|f_{M,k}^* - f_k^*\|_2^2 + pen(n, M, K^*)\right\} + A\frac{x}{n}\right],$$

where $f_{M,k}^*$ is the projection of f_k^* on $\mathfrak{P}_M$.

It is now possible to get the convergence rate of the estimators of the parameters. In order to take the event where $\hat{K}_{\text{l.s.}} \neq K^*$ into account, we agree that the distance between the parameters of two HMMs with different orders is bounded by some constant C_{err}. Note that C_{err} could even be taken as a power of n without changing anything to our result.

Corollary 2.11. *Assume **[HX]**, **[HidA]**, **[HF]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$, and that the penalty satisfies*

$$\forall n,\ \forall M \leqslant n,\ \forall K \leqslant n, \quad pen(n, M, K) = (MK + K^2 - 1)\frac{\log(n)^2}{n}$$

Then there exists a positive constant $A \equiv A(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ *such that for all* $\beta > 1$*, there exists a positive constant* $n_0 \equiv n_0(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L, \beta)$ *such that for all* $n \geqslant n_0$ *and for all* $C_{err} > 0$*,*

$$\mathbb{E}\left[\mathbf{1}_{\hat{K}\neq K^*} C_{err} + \mathbf{1}_{\hat{K}=K^*} d_{perm}((\hat{\pi}, \hat{\mathbf{Q}}, \hat{\mathbf{f}}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*))\right] \leqslant \frac{4C_{\mathcal{F},2}^L \sqrt{LK^*}}{c(\mathbf{Q}^*, \mathbf{f}^*)} \times \inf_{M \leqslant n} \left\{ \sum_{k=1}^{K^*} \|f_{M,k}^* - f_k^*\|_2^2 + pen(n, M, K^*) \right\} + \frac{A}{c(\mathbf{Q}^*, \mathbf{f}^*)\, n} + \frac{C_{err}}{n^\beta},$$

and $\mathbb{P}(\hat{K}_{l.s.} \neq K^*) = O(n^{-\beta})$.

Let us discuss what this corollary implies. The approximation error $\sum_{k=1}^{K^*} \|f_{M,k}^* - f_k^*\|_2^2$ can be bounded in a standard way by $O(M^{-2s/D})$ where $s > 0$ is the regularity of the emission densities, see for instance DeVore and Lorentz (1993). One can obtain a trade-off between approximation error and penalty by choosing $M \approx (n/\log(n)^2)^{D/(2s+D)}$, which leads to the optimal rate of convergence $(n/\log(n)^2)^{-2s/(2s+D)}$, up to a logarithmic factor. This shows that our estimators are adaptive, quasi-rate minimax and converge almost surely to the right number of states, all at the same time.

2.4 Spectral estimation

In this section, we introduce our spectral order estimator. We will assume **[HX]** and **[HidA]** hold.

The idea of this method is to use the matrix containing the coordinates of the density of two consecutive observations in an orthonormal basis. Take $M \in \mathcal{M}$ and let $\Phi_M = (\varphi_1^{(M)}, \ldots, \varphi_M^{(M)})$ be an orthonormal basis of $\mathfrak{P}_M$. For ease of notation, we will drop the dependency in M and write φ_a instead of $\varphi_a^{(M)}$. Let us introduce the matrice $\mathbf{N}_M$ and its empirical estimator, defined by

$$\forall a, b \in [M], \quad \mathbf{N}_M(a,b) := \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)],$$
$$\forall a, b \in [M], \quad \hat{\mathbf{N}}_M(a,b) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1}).$$

$\mathbf{N}_M$ contains the coordinates of the density of (Y_1, Y_2) with respect to $\mu^{\otimes 2}$ on the basis Φ_M. It holds that

$$\mathbf{N}_M = \mathbf{O}_M \mathrm{Diag}(\pi^*) \mathbf{Q}^* \mathbf{O}_M^\top, \tag{2.3}$$

with $\mathbf{O}_M$ the coordinates of the emission densities on the orthonormal basis:

$$\forall m \in [M],\ \forall k \in \mathcal{X}, \quad \mathbf{O}_M(m,k) := \mathbb{E}[\varphi_m(Y_1)|X_1 = k] = \int \varphi_m f_k^* d\mu.$$

When the emission densities are linearly independent, $\mathbf{O}_M$ has full rank for M large enough.

The key remark for our method is that $\mathbf{N}_M$ contains explicit information about the order of the HMM, as stated in the following lemma:

Lemma 2.12. *There exists* $M_0 \equiv M_0(\mathbf{Q}^*, \Phi_M, \mathbf{f}^*)$ *such that for all* $M \geqslant M_0$*,* $\mathbf{N}_M$ *has rank* K^**.*

In the following, we will assume $M \geqslant M_0$ for M_0 given by this lemma.

In practice, one only has access to the matrix $\hat{\mathbf{N}}_M$, which can be seen as a noisy version of $\mathbf{N}_M$. In particular, there is no reason for it to have only K^* nonzero singular values. On

the contrary, the spectrum becomes noisy, and when some singular values of $\mathbf{N}_M$ are too small, they can be masked by this noise. As seen in equation (2.3), this can occur when $\mathbf{Q}^*$ or $\mathbf{O}_M$ are close to not having full rank, which means for $\mathbf{O}_M$ that the emission densities are almost linearly dependent.

Denote by $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \dots$ the singular values of the matrix A. We can now state the theorem proving the consistency of the spectral order estimator:

Theorem 2.13. *Let* $\hat{K}_{sp.}(C) = \#\{i \mid \sigma_i(\hat{\mathbf{N}}_M) > C\sqrt{\log(n)/n}\}$.

There exists $C_0 \equiv C_0(\mathbf{Q}^*, \Phi_M)$ *and* $n_0 \equiv n_0(\mathbf{Q}^*, \Phi_M, \mathbf{O}_M^*)$ *such that for all* $C \geqslant C_0$ *and* $n \geqslant n_0 C^2(1 + \log(C))$,

$$\mathbb{P}(\hat{K}_{sp.}(C) \neq K^*) \leqslant n^{-2}$$

so that $\hat{K}_{sp.}(C) \longrightarrow K^*$ *almost surely.*

Remark. *It is possible to take* $M \longrightarrow \infty$, n_0 *constant and* C_0 *depending on* M *in an explicit way as long as* M *grows slowly enough, that is* $\eta_2(\Phi_M) \leqslant cst \cdot \sqrt{n/\log(n)}$ *and* $C_0 = cst \cdot \eta_2(\Phi_M)$ *where* $\eta_2(\Phi_M)$ *is defined in Lemma 2.14.*

Proof. The following result from appendix E of De Castro et al. (2017) allows to control the difference between the spectra of $\mathbf{N}_M$ and $\hat{\mathbf{N}}_M$.

Lemma 2.14. *There exists some constant* $\mathcal{C}_*$ *depending only on* $\mathbf{Q}^*$ *such that for any positive* u, M *and* n,

$$\mathbb{P}\left[\|\mathbf{N}_M - \hat{\mathbf{N}}_M\|_F \geqslant \frac{\eta_2(\Phi_M)\mathcal{C}_*}{\sqrt{n}}(1+u)\right] \leqslant e^{-u^2}$$

where

$$\eta_2^2(\Phi_M) = \sup_{y,y' \in \mathcal{Y}^2} \sum_{a,b=1}^{M} (\varphi_a(y_1)\varphi_b(y_2) - \varphi_a(y_1')\varphi_b(y_2'))^2.$$

In particular, taking $u = \sqrt{2\log(n)}$ and assuming $u > 1$ and $n \geqslant 2$, one has with probability $1 - n^{-2}$ that

$$\sigma_1(\mathbf{N}_M - \hat{\mathbf{N}}_M) \leqslant C\sqrt{\frac{\log(n)}{n}}$$

for all $C \geqslant C_0 := 2\sqrt{2}\eta_2(\Phi_M)\mathcal{C}_*$, using that for any matrix A, one has $\sigma_1(A) \leqslant \|A\|_F$.

Let $C \geqslant C_0$. We will need Weyl's inequality (a proof may be found in Stewart and Sun (1990) for instance):

Lemma 2.15 (Weyl's inequality)**.** *Let* A, B *be* $p \times q$ *matrices with* $p \geqslant q$, *then for all* $i = 1, \dots, q$,

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

Using this inequality, one gets that with probability at least $1 - n^{-2}$, for all $1 \leqslant i \leqslant K^*$, $\sigma_i(\hat{\mathbf{N}}_M) > \sigma_{K^*}(\mathbf{N}_M) - C\sqrt{\log(n)/n}$ and for all $i > K^*$, $\sigma_i(\hat{\mathbf{N}}_M) < C\sqrt{\log(n)/n}$.

In particular, if $2C\sqrt{\log(n)/n} < \sigma_{K^*}(\mathbf{N}_M)$, then with probability at least $1 - n^{-2}$, the order is exactly the number of singular values of $\hat{\mathbf{N}}_M$ which are larger than $C\sqrt{\log(n)/n}$. Finally, observe that under the condition $n \geqslant n_0 C^2(1 + \log(C))$,

$$\begin{aligned} C\sqrt{\frac{\log(n)}{n}} &\leqslant \sqrt{\frac{2\log(C) + \log(1 + \log(C))}{n_0(1 + \log(C))}} \\ &\leqslant \sqrt{\frac{3}{n_0}}\sqrt{\frac{\log(C)}{1 + \log(C)}}, \end{aligned}$$

since one can assume without loss of generality that $C_0 \geqslant 1$. By taking $n_0 = 12/\sigma_{K^*}(\mathbf{N}_M)^2$, this concludes the proof. □

2.5 Numerical experiments

In this section, we show the results of our estimators on simulated data. The simulation parameters are introduced in Section 2.5.1. We show the numerical results and discuss their ability to select the right order in practice in Section 2.5.2, and we present the data-driven methods and heuristics we used for the numerical implementation in Section 2.5.3.

2.5.1 Simulation parameters

We will consider $\mathcal{Y} = [0,1]$ with μ being the Lebesgue measure. We will use a trigonometric basis on $\mathbf{L}^2([0,1])$ to generate the approximation spaces $(\mathfrak{P}_M)_M$. More precisely, define

$$\begin{aligned}\varphi_0(t) &= 1\\ \varphi_a(t) &= \sqrt{2}\cos(\pi a t)\end{aligned}$$

for all $t \in [0,1]$ and $a \in \mathbb{N}^*$. We take $\mathfrak{P}_M = \mathrm{Span}(\{\varphi_a \mid 0 \leqslant a < M\})$ the spaces induced by the trigonometric basis.

Remark. *Taking the same vectors in all bases is not mandatory to ensure theoretical consistency, but in practice it allows us to take an additional initial point for the minimization step and improves the stability of the algorithm (see Step 1 below).*

We will assume $\mathbf{f}^*$ to be linearly independent, so that one only needs $L = 3$ observations to recover the parameters of the HMM.

In order to assess the performances of the different procedures, we generate n observations of a HMM of order 3 for several values of n, using the following parameters:

- Emission distributions: Beta distributions with two possible sets of parameters: $[(1.5;5)$, $(7;2)$ and $(6;6)]$ or $[(2;5)$, $(4;2)$ and $(4;4)]$;
- Markov chain parameters:

$$\begin{aligned}\mathbf{Q}^* &= \begin{pmatrix} 0.8 & 0.1 & 0.1\\ 0.2 & 0.7 & 0.1\\ 0.07 & 0.13 & 0.8\end{pmatrix},\\ \pi^* &= \begin{pmatrix}\frac{47}{120} & \frac{11}{40} & \frac{1}{3}\end{pmatrix}\\ &\approx (0.3917 \quad 0.2750 \quad 0.3333).\end{aligned}$$

Finally, we take $M_{\max} = 50$ and $K_{\max} = 5$ the maximum values of M and K for which we will compute the estimators.

The simulation codes are available in MATLAB at `https://www.normalesup.org/~llehericy/HMM_order_simfiles/`.

2.5.2 Numerical results

Figure 2.1 summarizes the results of both procedures. Both select the right order as soon as the number of observations is sufficient.

The spectral method is easily put in pratice and runs extremely fast. It doesn't need a time-consuming contrast minimization step or an initial point. However, the thresholding of the singular values is a delicate issue, and if the order is incorrect, then the theoretical results about the spectral estimators of the parameters don't hold and this method may behave poorly.

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$	$\mathbb{P}(\hat{K}_{\text{sp.}} = K^*)$
999	0.2	0
3 000	1	0
9 999	1	1
19 998	1	1

(a) Beta parameters $(1.5; 5)$, $(7; 2)$ and $(6; 6)$.

n	$\mathbb{P}(\hat{K}_{\text{l.s.}} = K^*)$	$\mathbb{P}(\hat{K}_{\text{sp.}} = K^*)$
7 500	0.3	0
19 998	0.9	0
30 000	1	0
49 998	1	0.1

(b) Beta parameters $(2; 5)$, $(4; 2)$ and $(4; 4)$

Figure 2.1: Probability to select the right order for the two methods ($\hat{K}_{\text{l.s.}}$ for the least squares method and $\hat{K}_{\text{sp.}}$ for the spectral method). 10 simulations have been done for each n. Parameters for spectral selection are $M = 40$, $M_{\text{reg}} = 35$ and $\tau = 1.5$ (see Section 2.5.3 for the definition of these parameters).

The performances of the least squares method are much better (see Figure 2.1 for comparing the order estimators and de Castro et al. (2016) for comparing the emission densities estimators). In addition, the model selection step is easy to handle and gives an estimator of the order that we proved to be consistent, estimators of the HMM parameters that we proved to be quasi-rate minimax and a way to check whether the model fits the data well (see Section 2.5.3), all at the same time. However, the minimization of the (non-convex) empirical contrast is a time-consuming step, especially for large samples and large models.

Choosing the right method is thus a question of computational power and amount of available data. For small datasets where one wants to get accurate results, the least squares method is best. Conversely, on large datasets and large models, the spectral method is a good choice in order to obtain many estimators in a reasonable amount of time.

2.5.3 Practical implementation

Least squares method

The first issue that one encounters when trying to minimize the least squares criterion γ_n is that it is not convex. Several algorithms have been proposed to overcome this difficulty. We chose to use CMA-ES (for Covariance Matrix Adaptation Estimation Strategy, see Hansen (2006)) in order to find a minimizer. This estimator is easy to use and works well in many situations, but–like all approximate minimization algorithms–it requires a good initial point since it might otherwise remain stuck in local minima.

One part of our method consists in using previous estimates as initial points for further steps to counter this phenomenon, since it is likely that this way the estimators stay near the real minimizer. Our practical algorithm is the following:

1. Minimize γ_n on each model, for $M \leqslant M_{\max}$ and $K \leqslant K_{\max}$. We take several initial points for model (K, M) according to the following cases:

   - $K = 1$. Use a HMM with a single state and a uniform emission distribution.
   - $K > 1$. Take the estimator from model $(K - 1, M)$. For each hidden state of the corresponding HMM, use the model where this state is duplicated. More precisely, the Markov chain $\tilde{X}$ where state I is duplicated is obtained by replacing the state I

from chain X with two states I_1 and I_2 such that for each state $S \neq I_1, I_2$,

$$\begin{aligned}\mathbb{P}(\tilde{X}_{t+1} = I_1 \mid \tilde{X}_t = S) &= \frac{1}{2}\mathbb{P}(X_{t+1} = I \mid X_t = S) \\ &= \mathbb{P}(\tilde{X}_{t+1} = I_2 \mid \tilde{X}_t = S) \\ \mathbb{P}(\tilde{X}_{t+1} = S \mid \tilde{X}_t = I_1) &= \mathbb{P}(X_{t+1} = S \mid X_t = I) \\ &= \mathbb{P}(\tilde{X}_{t+1} = S \mid \tilde{X}_t = I_2)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(\tilde{X}_{t+1} = I_2 \mid \tilde{X}_t = I_1) &= \frac{1}{2}\mathbb{P}(X_{t+1} = I \mid X_t = I) \\ &= \mathbb{P}(\tilde{X}_{t+1} = I_1 \mid \tilde{X}_t = I_2)\end{aligned}$$

- $M > 1$. Use estimator from model $(K, M-1)$ with the M-th coordinate of each emission density set to zero. This is only interesting if all $\mathfrak{P}_M$ are spanned by the first M vectors of a given orthonormal basis, like for trigonometric spaces.

Then, after minimization from each one of these initial points, take the estimator that minimizes γ_n.

2. Tune the parameter ρ of the penalty with the slope heuristics or the dimension jump method (see below) and select $\hat{M}$ and $\hat{K}$.

3. Return the estimator for $M = \hat{M}$ and $K = \hat{K}$.

This iterative initialization procedure relies on the heuristics that when the order is underestimated, then several states are "merged" together. Duplicating a merged state will allow to separate them effectively while still taking advantage of the computations done up to now. It is meant to avoid having to recalculate all states at the same time (which could get us stuck in sub-optimal local minima) when the best solution is likely to be a small modification of the previous estimator. In addition, when the order is overestimated, it allows to make sure the empirical criterion is indeed decreasing with the dimension of the model by giving an estimator that performs at least as well as those from smaller models. This makes our method robust to an overestimation of the order.

The last practical issue is a very common one in the model selection setting: the constant ρ of the penalty is unknown and has to be estimated before one can select the right model. Several data-driven estimators have been proposed to circumvent this difficulty, for instance dimension jump heuristics, slope heuristics, bootstrap or cross validation. We focus on the first two, which have several advantages in our setting. First, they are easy to use, are proved to be theoretically valid in many settings and work well in a wide range of applications (see for instance Baudry et al. (2012) and references therein). Secondly, they take advantage of the structure of our problem and both give a qualitative way to check whether the choice of penalty is valid or not, and by extension whether the model is misspecified or not.

Dimension jump heuristics In this paragraph, we study the selected parameters

$$\rho \longmapsto (\hat{M}(\rho), \hat{K}(\rho)) \in \arg\min\{\gamma_n(\hat{g}_{K,M}) + \rho \operatorname{pen}_{\text{shape}}(n, M, K)\}$$

and the selected complexity

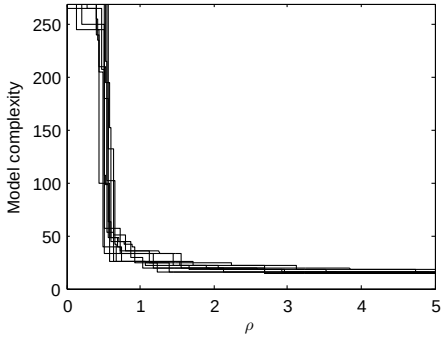
$$\rho \longmapsto \operatorname{Comp}(\rho) = \hat{M}(\rho)\hat{K}(\rho) + \hat{K}(\rho))[\hat{K}(\rho)) - 1]$$

with $\mathrm{pen}_{\mathrm{shape}}(n, M, K) = (MK + K^2 - 1)\log(n)/n$.

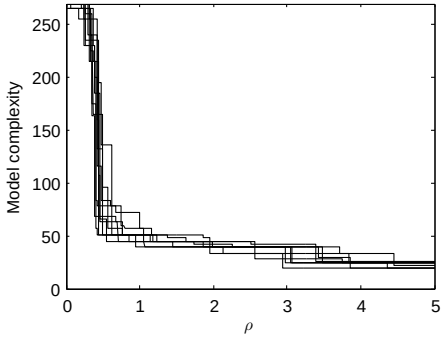
Assume that there exists κ such that $\kappa\,\mathrm{pen}_{\mathrm{shape}}$ is a *minimal penalty*, that is a penalty such that as n tends to infinity, for all $\rho > \kappa$, the size of the model chosen for penalty $\rho\,\mathrm{pen}_{\mathrm{shape}}$ remains small in some sense and for all $\rho < \kappa$, the size of the model becomes huge. Then, for n large enough, this will appear on the graph of the selected model complexity as a "dimension jump": around some constant ρ_{jump}, the complexity will abruptly drop from large models to small models. This is clearly the case in Figure 2.2. Figure 2.3 shows the behaviour of $\hat{M}$ and $\hat{K}$ with ρ. A dimension jump also occurs with these functions. It is most visible for $\hat{M}$.

Finally, once the dimension jump location ρ_{jump} has been estimated, we take $\hat{\rho} = 2\rho_{\mathrm{jump}}$ to select the final parameters.

It is worth noting that this jump method also gives a qualitative way to check whether the choice of parameters is sensible: if no clear jump can be identified, then either one didn't consider enough models to make the jump clear, or the penalty isn't the right one, or the model cannot approximate the data distribution well.

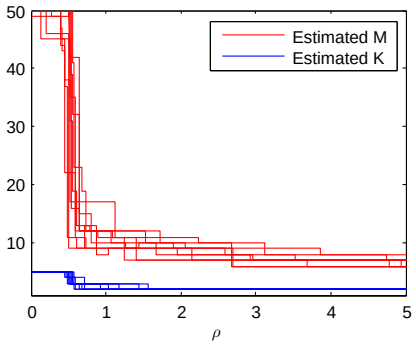


(a) $n = 7,500$.

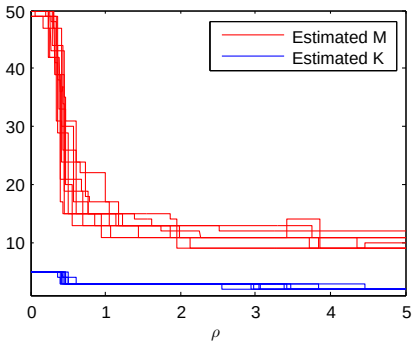


(b) $n = 49,998$.

Figure 2.2: Graph of $\rho \mapsto \mathrm{Comp}(\rho)$ for 10 sets of n consecutive observations. Here, the parameters of the Beta distribution are $(2;5)$, $(4;2)$ and $(4;4)$.



(a) $n = 7,500$.



(b) $n = 49,998$.

Figure 2.3: Graph of $\rho \mapsto \hat{M}(\rho)$ and $\rho \mapsto \hat{K}(\rho)$ for 10 sets of n consecutive observations. Here, the parameters of the Beta distribution are $(2;5)$, $(4;2)$ and $(4;4)$.

Slope heuristics This heuristics relies on the fact that when $\mathrm{pen}_{\mathrm{shape}}$ is a minimal penalty, then the empirical contrast function is expected to behave like $\rho_{\mathrm{min}}\mathrm{pen}_{\mathrm{shape}}$ for large models and for some constant ρ_{min}. This gives both a way to calibrate the constant of the penalty and to

check if the chosen penalty has the right shape (see Baudry et al. (2012)). The final penalty is then taken as $2\hat{\rho}_{\min}\text{pen}_{\text{shape}}$.

Figure 2.4 shows the graph of the empirical contrast depending on $\text{pen}_{\text{shape}}$. The slope heuristics works well in this situation, suggesting that our penalty has the right shape.

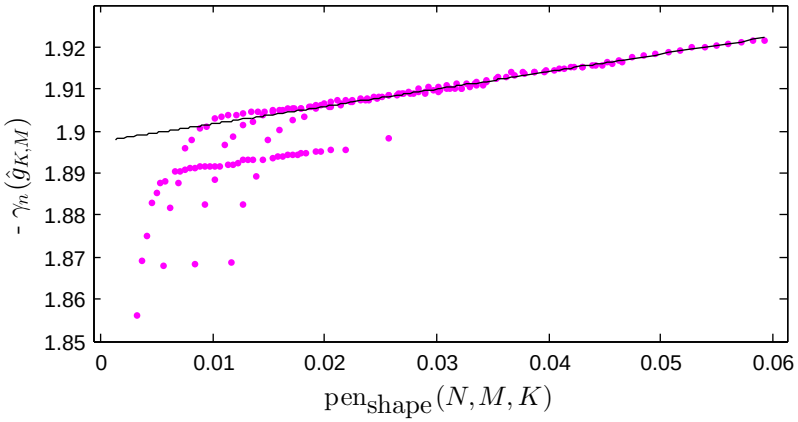


Figure 2.4: Empirical constrast and calibrated penalty for $n = 49,998$. Here, the parameters of the Beta distribution are $(2;5)$, $(4;2)$ and $(4;4)$.

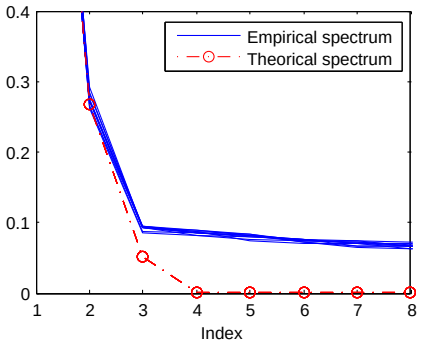
Spectral method

The idea of the spectral order estimation is to recover the rank of the matrix $\mathbf{N}_M$. However, this is not always possible: if one singular value of $\mathbf{N}_M$ is smaller than the noise (which is the case when $\mathbf{O}_M$ is close from not being invertible, i.e. when the emission densities are close from being linearly dependent, and when there are only few observations), then this method will not be able to "see" the corresponding hidden state.

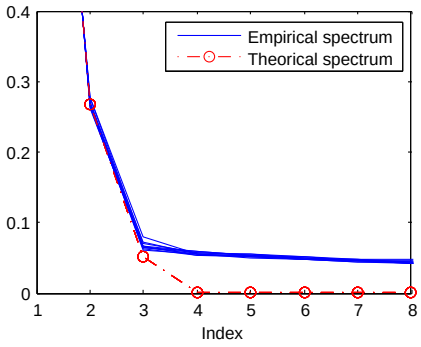
Figure 2.5–and in particular Figure 2.5a–illustrates this problem: the third singular value is smaller than several noisy singular values, which means it won't be possible to recover it. Even if one knows the right order, the fact that the singular value is smaller than the noise can make it impossible for spectral methods to recover the true parameters. Figure 2.6 shows the result when trying to estimate the densities in the situation of Figure 2.5a: when the singular value is drowned by the noise, the output of the spectral estimator is aberrant. Notice that it is not a fatality: in the same situation, the least squares method manages to give sensible estimators of the emission densities. This is an intrinsic limitation of the spectral method.

Therefore, what we need is a way to threshold the parameters in order to distinguish noise from significant singular values. The estimator $\hat{K}_{\text{sp.}}(C)$ is one way to achieve this, but the calibration of C is a tricky problem, since the right choice of C depends on the parameters of the HMM. We will use a different method, which relies on the same idea: identifying the noisy singular values which stand out from the others and saying they correspond to nonzero singular values of $\mathbf{N}_M$. Our heuristics relies on the fact that when one sorts the singular values in decreasing order, then the smallest ones approximately follow an affine relation with respect to their index. This tendency is shown in Figure 2.7.

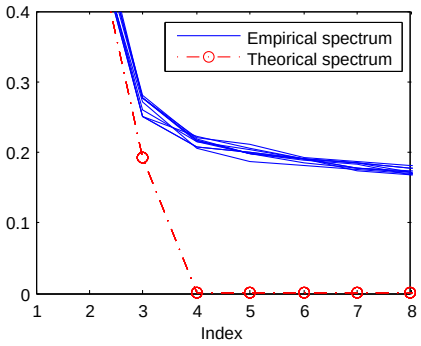
We proceed as follows. Let M and M_{reg} be two positive integers such that $M_{\text{reg}} \leqslant M \leqslant M_{\max}$. We estimate the affine dependance of the singular values of $\hat{\mathbf{N}}_M$ with respect to their index with a linear regression using its M_{reg} smallest singular values. Then, we set a thresholding parameter $\tau > 1$. We say a singular value is *significant* if it is above τ times the value that the regression predicts for it. Lastly, we take $\hat{K}_{\text{sp.}}$ as the number of consecutive significant singular values



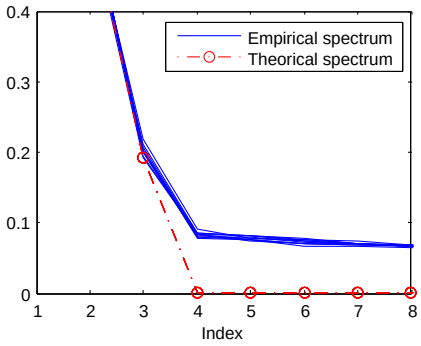
(a) $n = 19,998$, Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$.



(b) $n = 49,998$, Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$.



(c) $n = 3,000$, Beta parameters $(1.5;5)$, $(7;2)$ and $(6;6)$.



(d) $n = 19,998$, Beta parameters $(1.5;5)$, $(7;2)$ and $(6;6)$.

Figure 2.5: Spectrum of the empirical matrix $\hat{\mathbf{N}}_M$ and the theoretical matrix $\mathbf{N}_M$ for $M = 40$ and 10 simulations. The first singular values are too large to appear here.

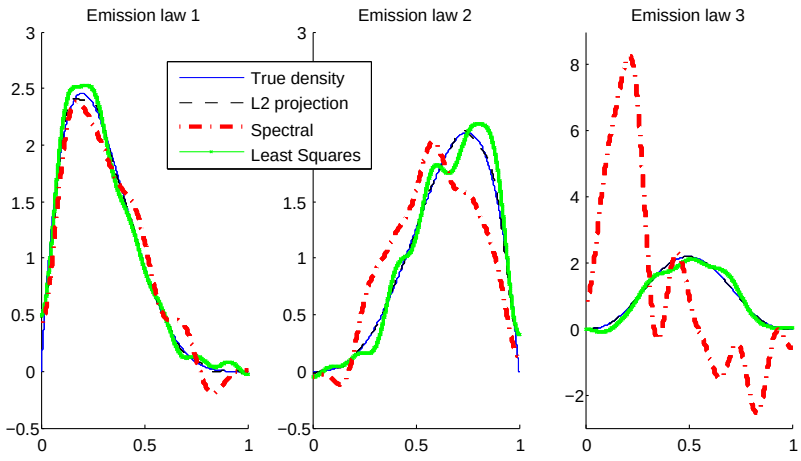


Figure 2.6: Estimators of the emission densities for $n = 19,998$ and Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$. We took $K = \hat{K}_{\text{l.s.}} = 3$ and $M = \hat{M} = 13$. The bad behaviour of the spectral algorithm when the emission densities are poorly separated is clearly visible on the third emission distribution.

starting from the largest one. This heuristics seems to work as soon as τ is large enough, e.g.

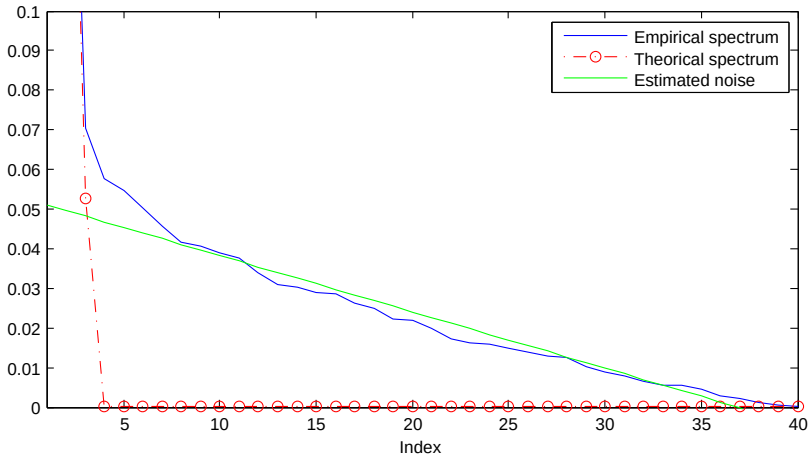


Figure 2.7: Spectrum of $\mathbf{N}_M$ for $M = 40$ and $n = 49,998$ for Beta parameters $(2;5)$, $(4;2)$ and $(4;4)$. The regression (green line) has been performed on the 35 smallest singular values. The two largest singular values are too large to appear here.

$\tau = 1.5$.

2.6 Proofs

2.6.1 Main technical result

The following lemma is the main technical result of this chapter. It is the key for both the strong consistency and the oracle inequalities. It allows to control the difference between the empirical criterion γ_n and the theoretical $\mathbf{L}^2$ loss for all models at the same time.

Define $\nu : t \longmapsto \frac{1}{n}\sum_{s=1}^{n} t(Z_s) - \int t g^*$, so that

$$\forall t \in \mathbf{L}^2(\mathcal{Y}^L, \mu^{\otimes L}), \quad \gamma_n(t) + \|g^*\|_2^2 = \|t - g^*\|_2^2 - 2\nu(t) \tag{2.4}$$

Let

$$\begin{aligned} s = (s_{K,M})_{K,\,M} \in \mathbf{S} := \prod_{K\in\mathbb{N}^*,\, M\in\mathcal{M}} \left(\bigcup_K S_K\right) & \\ \longmapsto (Z_{K,M}(s))_{K,M} & \\ := \left(\sup_{t\in S_{K,M}} \left[\frac{|\nu(t - s_{K,M})|}{\|t - s_{K,M}\|_2^2 + x_{K,M}^2}\right]\right)_{K,M} & \end{aligned} \tag{2.5}$$

Remark. *It is not necessary to assume that $s_{K,M} \in S_{K,M}$. In particular, one can take $s_{K,M} = g^*$ for all K, M. In that case, we will simply write $Z_{K,M}(g^*)$.*

Lemma 2.16. *Assume **[HX]** and **[HF]** hold. Then there exists a sequence $(x_{K,M})_{K,M} \equiv (x_{K,M})_{K,M}(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ and positive constants $(n_0, \rho, A) \equiv (n_0, \rho, A)(C_{\mathcal{F},2}, C_{\mathcal{F},\infty}, \mathbf{Q}^*, L)$ such that if the penalty $\widetilde{pen}$ satisfies*

$$\forall n,\ \forall M \leqslant n,\ \forall K \leqslant n \quad \widetilde{pen}(n, M, K) \geqslant \rho(MK + K^2 - 1)\frac{\log(n)}{n}$$

then for all $s \in \mathbf{S}$, $n \geqslant n_0$ *and* $x > 0$, *one has with probability larger than* $1 - e^{-x}$*:*

$$\begin{cases} \sup_{K' \leqslant n,\, M' \leqslant n} Z_{K',M'}(s) \leqslant \dfrac{1}{4} \\ \sup_{K' \leqslant n,\, M' \leqslant n} \left(2 Z_{K',M'}(s) x^2_{K',M'} - \widetilde{pen}(n, M', K')\right) \leqslant A \dfrac{x}{n} \end{cases}$$

Remark. *One can replace the constant* $1/4$ *in the first upper bound by any* $\epsilon > 0$*, at the cost of changing the constants* n_0*,* ρ *and* A*.*

The structure of the proof follows the usual method to control empirical processes, see for instance Massart (2007), Chapter 6, adapted to the HMM structure by de Castro et al. (2016). The novelty and main difficulty of the proof comes from the generalization to both nonparametric densities and an unknown number of states: we had to introduce a much finer control of the constants and of the bracketing entropy of the models in order to take the dependency in the order of the HMM into account.

The details of the proof can be found in appendix 2.C.

2.6.2 Identifiability proofs

Proof of Corollary 2.2

Denote by Proj_A the orthogonal projection on a linear space A.

Since the union of $(\mathfrak{P}_M)_{M \in \mathcal{M}}$ is dense in $\mathcal{F}$, we can take M such that **[HidA]** or **[HidB]** holds for $\mathbf{f}^*_M = (f^*_{M,k})_{k \in \mathcal{X}} := (\mathrm{Proj}_{\mathfrak{P}_M} f^*_k)_{k \in \mathcal{X}}$.

We will need the following lemma.

Lemma 2.17.

$$\forall \pi \in \mathbb{R}^K,\ \forall \mathbf{Q} \in \mathbb{R}^{K \times K},\ \forall \mathbf{f} \in \mathcal{F}^K,\ \forall M, \qquad Proj_{\mathfrak{P}_M^{\otimes L}} \left(g^{\pi, \mathbf{Q}, \mathbf{f}} \right) = g^{\pi, \mathbf{Q}, Proj_{\mathfrak{P}_M}(\mathbf{f})}$$

Proof. By linearity of the projection operator, it is enough to prove that for all $(t_1, \ldots, t_L) \in (\mathbf{L}^2(\mathcal{Y}, \mu))^L$,

$$\mathrm{Proj}_{\mathfrak{P}_M^{\otimes L}}(t_1 \otimes \cdots \otimes t_L) = \mathrm{Proj}_{\mathfrak{P}_M}(t_1) \otimes \cdots \otimes \mathrm{Proj}_{\mathfrak{P}_M}(t_L)$$

which is easy to check. □

We will make a proof by contradiction. Assume that $\inf_{t \in S_K} \|t - g^*\|_2 = 0$ for some $K < K^*$. Then there exists a sequence $(g_n)_{n \geqslant 1} = (g^{\pi_n, \mathbf{Q}_n, \mathbf{f}_n})_{n \geqslant 1}$ such that $g_n \longrightarrow g^*$ in $\mathbf{L}^2(\mathcal{Y}^L, \mu^{\otimes L})$, with $\pi_n \in \Delta_K$, $\mathbf{Q}_n$ a transition matrix of size K and $\mathbf{f}_n \in \mathcal{F}^K$.

The orthogonal projection on $\mathfrak{P}_M^{\otimes L}$ is continuous, so by using Lemma 2.17, one gets that

$$g^{\pi_n, \mathbf{Q}_n, \mathrm{Proj}_{\mathfrak{P}_M}(\mathbf{f}_n)} \longrightarrow g^{\pi^*, \mathbf{Q}^*, \mathbf{f}^*_M}$$

Then, using the compacity of Δ_K and of the set of transition matrices of size K and the relative compacity of $(\mathcal{F} \cap \mathfrak{P}_M)^K$ (which is a bounded subset of a finite dimension linear space), one gets (up to extraction of a subsequence) that there exists $\pi_\infty \in \Delta_K$, $\mathbf{Q}_\infty$ a transition matrix of size K and $\mathbf{f}_\infty \in (\mathfrak{P}_M)^K$ such that $\pi_n \longrightarrow \pi_\infty$, $\mathbf{Q}_n \longrightarrow \mathbf{Q}_\infty$ and $\mathrm{Proj}_{\mathfrak{P}_M}(\mathbf{f}_n) \longrightarrow \mathbf{f}_\infty$.

Finally, using the continuity of the function $(\pi, \mathbf{Q}, \mathbf{f}) \longmapsto g^{\pi, \mathbf{Q}, \mathbf{f}}$ and the unicity of the limit, one gets

$$g^{\pi_\infty, \mathbf{Q}_\infty, \mathbf{f}_\infty} = g^{\pi^*, \mathbf{Q}^*, \mathbf{f}^*_M}.$$

Then Proposition 2.1 contradicts the assumption $K < K^*$, which is enough to conclude.

2.6.3 Consistency proofs

The definition of $\hat{K}_{\text{l.s.}}$ is equivalent to the following one:

$$\hat{K}_{\text{l.s.}} \in \arg\min_{K \leqslant n} \{\gamma_n(\hat{g}_{K,\hat{M}_K}) + \text{pen}(n, \hat{M}_K, K)\}$$

where

$$\hat{M}_K \in \arg\min_{M \leqslant n} \{\gamma_n(\hat{g}_{K,M}) + \text{pen}(n, M, K)\}$$

Choosing K rather than K^* means that K is better than K^*, i.e.

$$\{\hat{K}_{\text{l.s.}} = K\} \subset \Big\{0 \geqslant \inf_{M \leqslant n} \{\gamma_n(\hat{g}_{K,M}) + \text{pen}(n, M, K)\} - \inf_{M \leqslant n} \{\gamma_n(\hat{g}_{K^*,M}) + \text{pen}(n, M, K^*)\}\Big\}.$$

Let

$$\begin{aligned} D_{n,K} &:= \inf_{M \leqslant n} \{\gamma_n(\hat{g}_{K,M}) + \text{pen}(n, M, K)\} \\ &\qquad - \inf_{M \leqslant n} \{\gamma_n(\hat{g}_{K^*,M}) + \text{pen}(n, M, K^*)\} \\ &= \gamma_n(\hat{g}_{K,\hat{M}_K}) + \text{pen}(n, \hat{M}_K, K) \\ &\qquad - \inf_{M \leqslant n} \{\inf_{t \in S_{K^*,M}} \gamma_n(t) + \text{pen}(n, M, K^*)\}. \end{aligned}$$

Then

$$\{\hat{K}_{\text{l.s.}} = K\} \subset \{D_{n,K} \leqslant 0\}.$$

We will thus control the probability of the latter event for all $K < K^*$ in the first case and $K > K^*$ in the second case.

Proof of Theorem 2.3 Let $M_0 \in \mathcal{M}$. We will choose a suitable value for this integer later in the proof. Assume $n \geqslant M_0$. Then by definition of $D_{n,K}$ and of ν (equation (2.4)),

$$\begin{aligned} D_{n,K} &\geqslant \gamma_n(\hat{g}_{K,\hat{M}_K}) + \text{pen}(n, \hat{M}_K, K) - \gamma_n(g^*_{K^*,M_0}) - \text{pen}(n, M_0, K^*)) \\ &\geqslant \|g^* - \hat{g}_{K,\hat{M}_K}\|_2^2 - \|g^* - g^*_{K^*,M_0}\|_2^2 - 2\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,M_0}) \\ &\qquad + \text{pen}(n, \hat{M}_K, K) - \text{pen}(n, M_0, K^*). \end{aligned}$$

Using the definition of $Z_{K,M}$ (equation (2.5)), one gets that

$$\begin{aligned} |\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,M_0})| &\leqslant |\nu(\hat{g}_{K,\hat{M}_K} - g^*)| + |\nu(g^* - g^*_{K^*,M_0})| \\ &\leqslant Z_{K,\hat{M}_K}(g^*)(\|g^* - \hat{g}_{K,\hat{M}_K}\|_2^2 + x^2_{K,\hat{M}_K}) \\ &\qquad + Z_{K^*,M_0}(g^*)(\|g^* - g^*_{K^*,M_0}\|_2^2 + x^2_{K^*,M_0}). \end{aligned}$$

Let n_0, ρ and A be as in Lemma 2.16. We can assume that $n_0 \geqslant K^*$ so that $K^* \leqslant n$. Let us introduce the function $\widetilde{\text{pen}}(n, M, K) = \rho(MK + K^2 - 1)\frac{\log(n)}{n}$. Let $n \geqslant n_0$ and $x > 0$ and assume we are in the event of probability $1 - e^{-x}$ of Lemma 2.16. Then, for all $K \leqslant n$:

$$\begin{aligned} |\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,M_0})| &\leqslant \frac{1}{4}\|g^* - \hat{g}_{K,\hat{M}_K}\|_2^2 + \frac{1}{2}A\frac{x}{n} + \frac{1}{2}\widetilde{\text{pen}}(n, \hat{M}_K, K) \\ &\qquad + \frac{1}{4}\|g^* - g^*_{K^*,M_0}\|_2^2 + \frac{1}{2}A\frac{x}{n} + \frac{1}{2}\widetilde{\text{pen}}(n, M_0, K^*) \end{aligned}$$

and

$$D_{n,K} \geqslant \frac{1}{2}\|g^* - \hat{g}_{K,\hat{M}_K}\|_2^2 - \frac{3}{2}\|g^* - g^*_{K^*,M_0}\|_2^2 - 2A\frac{x}{n} + \mathrm{pen}(n,\hat{M}_K,K) \\ - \mathrm{pen}(n,M_0,K^*) - \widetilde{\mathrm{pen}}(n,\hat{M}_K,K) - \widetilde{\mathrm{pen}}(n,M_0,K^*).$$

We assumed pen $\geqslant \widetilde{\mathrm{pen}}$, so that

$$D_{n,K} \geqslant \frac{1}{2}\|g^* - \hat{g}_{K,\hat{M}_K}\|_2^2 - \frac{3}{2}\|g^* - g^*_{K^*,M_0}\|_2^2 - 2A\frac{x}{n} - 2\mathrm{pen}(n,M_0,K^*)$$

Corollary 2.2 ensures that

$$d := \inf_{K<K^*}\inf_{t\in S_K}\|t - g^*\|_2 > 0,$$

so that for all $K < K^*$,

$$D_{n,K} \geqslant \frac{d^2}{2} - \frac{3}{2}\|g^* - g^*_{K^*,M_0}\|_2^2 - 2A\frac{x}{n} - 2\mathrm{pen}(n,M_0,K^*).$$

By density of $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ in $\mathcal{F}$, one gets that

$$\inf_M \|g^*_{K^*,M} - g^*\|_2 = 0$$

so that there exists M_0 such that $\|g^* - g^*_{K^*,M_0}\|_2^2 \leqslant d^2/6$. If we choose this M_0, we get that

$$D_{n,K} \geqslant \frac{d^2}{4} - 2A\frac{x}{n} - 2\mathrm{pen}(n,M_0,K^*).$$

Which implies that $D_{n,K} > 0$ as soon as $2Ax/n < d^2/4 - 2\mathrm{pen}(n,M_0,K^*)$, i.e.

$$x < \left(\frac{d^2}{8} - \mathrm{pen}(n,M_0,K^*)\right)\frac{n}{A}.$$

To conclude, note that there exists $\tilde{n_0} \geqslant \max(n_0,M_0)$ such that for all $n \geqslant \tilde{n_0}$, $\mathrm{pen}(n,M_0,K^*) \leqslant \frac{d^2}{16}$. Then, letting $\beta = \frac{d^2}{16A}$, one has for all $n \geqslant \tilde{n_0}$, with probability $1 - e^{-\beta n}$, for all $K < K^*$, $D_{n,K} > 0$, which implies that $\hat{K}_{\text{l.s.}} \neq K$.

Proof of Theorem 2.4 For all $K \geqslant K^*$,

$$D_{n,K} \geqslant \gamma_n(\hat{g}_{K,\hat{M}_K}) + \mathrm{pen}(n,\hat{M}_K,K) - \gamma_n(g^*_{K^*,\hat{M}_K}) - \mathrm{pen}(n,\hat{M}_K,K^*)$$

and

$$\gamma_n(\hat{g}_{K,\hat{M}_K}) - \gamma_n(g^*_{K^*,\hat{M}_K}) = \|\hat{g}_{K,\hat{M}_K} - g^*\|_2^2 - \|g^*_{K^*,\hat{M}_K} - g^*\|_2^2 \\ - 2\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K}).$$

Note that $g^*_{K^*,\hat{M}_K} = g^*_{K,\hat{M}_K}$ is the orthogonal projection of g^* on $\mathfrak{P}^{\otimes L}_{\hat{M}_K}$ and $\hat{g}_{K,\hat{M}_K} \in S_{K,\hat{M}_K} \subset \mathfrak{P}^{\otimes L}_{\hat{M}_K}$, so that, using the Pythagorean Theorem,

$$\|\hat{g}_{K,\hat{M}_K} - g^*\|_2^2 - \|g^*_{K^*,\hat{M}_K} - g^*\|_2^2 = \|\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K}\|_2^2.$$

Let n_0, ρ and A be as in Lemma 2.16. We can assume that $n_0 \geqslant K^*$ so that $K^* \leqslant n$. Let us introduce the function $\widetilde{\mathrm{pen}}(n,M,K) = \rho(MK + K^2 - 1)\frac{\log(n)}{n}$. Let $n \geqslant n_0$ and $x > 0$ and

assume we are in the event of probability $1-e^{-x}$ of Lemma 2.16. Then, for all $K \leqslant n$ such that $K \geqslant K^*$:

$$\begin{aligned}
|\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K^*,\hat{M}_K})| &= |\nu(\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K})| \\
&\leqslant Z_{K,\hat{M}_K}((g^*_{K',M'})_{K',M'})\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 \\
&\qquad + Z_{K,\hat{M}_K}((g^*_{K',M'})_{K',M'})x^2_{K,\hat{M}_K} \\
&\leqslant \frac{1}{4}\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 + \frac{1}{2}A\frac{x}{n} + \frac{1}{2}\widetilde{\mathrm{pen}}(n,\hat{M}_K,K),
\end{aligned}$$

which implies

$$\begin{aligned}
\gamma_n(\hat{g}_{K,\hat{M}_K}) - \gamma_n(g^*_{K^*,\hat{M}_K}) &\geqslant \frac{1}{2}\|\hat{g}_{K,\hat{M}_K} - g^*_{K,\hat{M}_K}\|_2^2 - A\frac{x}{n} - \widetilde{\mathrm{pen}}(n,\hat{M}_K,K) \\
&\geqslant -A\frac{x}{n} - \widetilde{\mathrm{pen}}(n,\hat{M}_K,K)
\end{aligned}$$

so that for all $K \leqslant n$ such that $K \geqslant K^*$:

$$D_{n,K} \geqslant \mathrm{pen}(n,\hat{M}_K,K) - \mathrm{pen}(n,\hat{M}_K,K^*) - \widetilde{\mathrm{pen}}(n,\hat{M}_K,K) - A\frac{x}{n}.$$

Now, assume that **[Hpen]**(α,ρ) holds for some $\alpha > 0$ and the above constant ρ. Then there exists n_1 such that for all $n \geqslant n_1$ and for all $K \leqslant n$ such that $K \geqslant K^*$,

$$D_{n,K} \geqslant \alpha\frac{\log(n)}{n} - A\frac{x}{n},$$

which is strictly positive as soon as $x < \alpha\log(n)/A$. Thus, letting $\beta = 1/(2A)$, one has for all $n \geqslant \max(n_0,n_1,K^*)$, with probability $1-n^{-\beta\alpha}$, for all $K \leqslant n$ such that $K > K^*$, $D_{n,K} > 0$, which implies that $\hat{K}_{\mathrm{l.s.}} \neq K$. This concludes the proof.

Acknowledgments

We would like to thank Elisabeth Gassiat for her precious advice and Yohann de Castro for his codes which were at the root of our numerical experiments.

APPENDICES

2.A Spectral algorithm

Algorithm 1: Spectral estimation of HMM parameters (de Castro et al. (2016), De Castro et al. (2017))

Data: An observed chain $(Y_1, \dots, Y_n)$ and an order K.
Result: Spectral estimators $\hat{\pi}$, $\hat{\mathbf{Q}}$ and the estimators $(\hat{f}_{M,k})_{k\in\mathcal{X}}$ of $(f_k^*)_{k\in\mathcal{X}}$ in $\mathfrak{P}_M$ (equipped with an orthonormal basis $\Phi_M = (\varphi_1, \dots, \varphi_M)$).

[Step 1] Consider the following empirical estimators: for any a, b, c in $\{1, \dots, M\}$,

- $\hat{\mathbf{L}}_M(a) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_1^{(s)})$,
- $\hat{\mathbf{M}}_M(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_1^{(s)})\varphi_b(Y_2^{(s)})\varphi_c(Y_3^{(s)})$,
- $\hat{\mathbf{N}}_M(a,b) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_1^{(s)})\varphi_b(Y_2^{(s)})$,
- $\hat{\mathbf{P}}_M(a,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_1^{(s)})\varphi_c(Y_3^{(s)})$.

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

[Step 3] Form the matrices $\hat{\mathbf{B}}(b) := (\hat{\mathbf{U}}^\top \hat{\mathbf{P}}_M \hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^\top \hat{\mathbf{M}}_M(.\,, b, .\,)\hat{\mathbf{U}}$ for all $b \in \{1, \dots, M\}$.

[Step 4] Set Θ a $(K \times K)$ uniformly drawn random unitary matrix and form the matrices $\hat{\mathbf{C}}(k) := \sum_{b=1}^{M} (\hat{\mathbf{U}}\Theta)(b,k)\hat{\mathbf{B}}(b)$ for all $k \in \{1, \dots, K\}$.

[Step 5] Compute $\hat{\mathbf{R}}$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{\mathbf{C}}(1)$: $\hat{\mathbf{R}}^{-1}\hat{\mathbf{C}}(1)\hat{\mathbf{R}} = \mathrm{Diag}(\hat{\Lambda}(1,1), \dots, \hat{\Lambda}(1,K))$.

[Step 6] Set $\hat{\Lambda}(k,k') := (\hat{\mathbf{R}}^{-1}\hat{\mathbf{C}}(k)\hat{\mathbf{R}})(k',k')$ for all $k, k' \in \mathcal{X}$ and $\hat{\mathbf{O}}_M := \hat{\mathbf{U}}\Theta\hat{\Lambda}$.

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

[Step 8] Set $\tilde{\pi} := \Pi_{\Delta_K}\Big(\big(\hat{\mathbf{U}}^\top \hat{\mathbf{O}}_M\big)^{-1}\hat{\mathbf{U}}^\top \hat{\mathbf{L}}_M\Big)$ where Π_{Δ_K} denotes the projection onto the simplex in dimension K.

[Step 9] Consider the transition matrix estimator:

$$\hat{\mathbf{Q}} := \Pi_{\mathrm{TM}}\Big(\big(\hat{\mathbf{U}}^\top \hat{\mathbf{O}}_M \mathrm{Diag}\tilde{\pi}\big)^{-1}\hat{\mathbf{U}}^\top \hat{\mathbf{N}}_M \hat{\mathbf{U}}\big(\hat{\mathbf{O}}_M^\top \hat{\mathbf{U}}\big)^{-1}\Big),$$

where Π_{TM} denotes the projection onto the convex set of transition matrices, and define $\hat{\pi}$ as the stationary distribution of $\hat{\mathbf{Q}}$.

2.B Proofs of the oracle inequalities

2.B.1 Proof of Theorem 2.7

Let $K \leqslant n$ and $M \leqslant n$. Then

$$\begin{aligned}\gamma_n(\hat{g}) + \mathrm{pen}(n, \hat{M}, \hat{K}_{\mathrm{l.s.}}) &\leqslant \gamma_n(\hat{g}_{K,M}) + \mathrm{pen}(n, M, K) \\ &\leqslant \gamma_n(g^*_{K,M}) + \mathrm{pen}(n, M, K)\end{aligned}$$

where the first inequality comes from the definition of $(\hat{K}_{\mathrm{l.s.}}, \hat{M})$ and the second from the definition of $\hat{g}_{K,M}$. Therefore,

$$\gamma_n(\hat{g}) - \gamma_n(g^*_{K,M}) \leqslant \mathrm{pen}(n, M, K) - \mathrm{pen}(n, \hat{M}, \hat{K}_{\mathrm{l.s.}}).$$

By definition of ν (equation 2.4),

$$\gamma_n(t_1) - \gamma_n(t_2) = \|t_1 - g^*\|_2^2 - \|t_2 - g^*\|_2^2 - 2\nu(t_1 - t_2)$$

so that

$$\|\hat{g} - g^*\|_2^2 \leqslant \|g^*_{K,M} - g^*\|_2^2 + \mathrm{pen}(n, M, K) - \mathrm{pen}(n, \hat{M}, \hat{K}_{\mathrm{l.s.}}) + 2\nu(\hat{g}_{\hat{M},\hat{K}_{\mathrm{l.s.}}} - g^*_{K,M})$$

Now we want to control the ν term. By linearity,

$$\nu(\hat{g}_{\hat{K}_{\mathrm{l.s.}},\hat{M}} - g^*_{K,M}) = \nu(\hat{g}_{\hat{K}_{\mathrm{l.s.}},\hat{M}} - g^*) + \nu(g^* - g^*_{K,M})$$

Using the definition of $Z_{K,M}$ (equation 2.5), we get that

$$\begin{cases} |\nu(\hat{g}_{\hat{K}_{\mathrm{l.s.}},\hat{M}} - g^*)| \leqslant Z_{\hat{K}_{\mathrm{l.s.}},\hat{M}}(g^*)(\|\hat{g}_{\hat{K}_{\mathrm{l.s.}},\hat{M}} - g^*\|_2^2 + x^2_{\hat{K}_{\mathrm{l.s.}},\hat{M}}) \\ |\nu(g^*_{K,M} - g^*)| \leqslant Z_{K,M}(g^*)(\|g^*_{K,M} - g^*\|_2^2 + x^2_{K,M}) \end{cases}$$

so that, using Lemma 2.16, for all $n \geqslant n_0$ and $x > 0$, with probability larger than $1 - e^{-x}$, for all $M \leqslant n$ and $K \leqslant n$,

$$|\nu(\hat{g}_{\hat{K}_{\mathrm{l.s.}},\hat{M}} - g^*_{K,M})| \leqslant \frac{1}{4}\|\hat{g} - g^*\|_2^2 + \frac{1}{4}\|g^*_{K,M} - g^*\|_2^2 + A\frac{x}{n} + \frac{1}{2}\mathrm{pen}(n, \hat{M}, \hat{K}_{\mathrm{l.s.}}) + \frac{1}{2}\mathrm{pen}(n, M, K)$$

so that

$$\begin{aligned}\|\hat{g} - g^*\|_2^2 \leqslant& \|g^*_{K,M} - g^*\|_2^2 + 2\mathrm{pen}(n, M, K) \\ &+ \frac{1}{2}\|\hat{g} - g^*\|_2^2 + \frac{1}{2}\|g^*_{K,M} - g^*\|_2^2 + 2A\frac{x}{n},\end{aligned}$$

which means that

$$\frac{1}{2}\|\hat{g} - g^*\|_2^2 \leqslant \frac{3}{2}\|g^*_{K,M} - g^*\|_2^2 + 2\mathrm{pen}(n, M, K) + 2A\frac{x}{n}$$

and finally

$$\|\hat{g} - g^*\|_2^2 \leqslant 3 \inf_{K \leqslant n,\, M \leqslant n} \{\|g^*_{K,M} - g^*\|_2^2 + 2\mathrm{pen}(n, M, K)\} + 4A\frac{x}{n}$$

which is the expected inequality.

2.B.2 Proof of Equation 2.1

First, decompose the difference in three terms.

$$\|g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2}\|_2 \leqslant \|g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_1,\mathbf{f}_1}\|_2 + \|g^{\pi_2,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_1}\|_2 + \|g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2}\|_2$$

Then we can control each term separately. Let $(\varphi_m)_{m\in\mathbb{N}^*}$ be an orthonormal basis of $\cup_M \mathfrak{P}_M$.

$$\begin{aligned}
\|g^{\pi_1,\mathbf{Q},\mathbf{f}} - g^{\pi_2,\mathbf{Q},\mathbf{f}}\|_2^2 &= \left\| \sum_{\mathbf{k}\in\mathcal{X}^L} (\pi_1-\pi_2)_{k_1} \mathbf{Q}_{k_1,k_2}\dots\mathbf{Q}_{k_{L-1},k_L} \bigotimes_{i=1}^{L} f_{k_i} \right\|_2^2 \\
&= \sum_{\mathbf{m}\in(\mathbb{N}^*)^L} \left(\sum_{\mathbf{k}\in\mathcal{X}^L} (\pi_1-\pi_2)_{k_1} \mathbf{Q}_{k_1,k_2}\dots\mathbf{Q}_{k_{L-1},k_L} \prod_{i=1}^{L} \langle f_{k_i}, \varphi_{m_i}\rangle \right)^2 \\
&\leqslant \sum_{\mathbf{k}\in\mathcal{X}^L} (\pi_1-\pi_2)_{k_1}^2 \mathbf{Q}_{k_1,k_2}\dots\mathbf{Q}_{k_{L-1},k_L} \\
&\quad \times \sum_{\mathbf{k}'\in\mathcal{X}^L} \mathbf{Q}_{k'_1,k'_2}\dots\mathbf{Q}_{k'_{L-1},k'_L} \prod_{i=1}^{L} \sum_{m_i\in\mathbb{N}^*} \langle f_{k'_i}, \varphi_{m_i}\rangle^2
\end{aligned}$$

using Cauchy-Schwarz inequality. Then, since $\sum_{m_i\in\mathbb{N}^*} \langle f_{k'_i}, \varphi_{m_i}\rangle^2 = \|f_{k'_i}\|_2^2 \leqslant C_{\mathcal{F},2}^2$ by **[HF]** and $\mathbf{Q}$ is a transition matrix, we get that

$$\|g^{\pi_1,\mathbf{Q},\mathbf{f}} - g^{\pi_2,\mathbf{Q},\mathbf{f}}\|_2^2 \leqslant K C_{\mathcal{F},2}^{2L} \|\pi_1 - \pi_2\|^2$$

A similar decomposition leads to

$$\|g^{\pi,\mathbf{Q}_1,\mathbf{f}} - g^{\pi,\mathbf{Q}_2,\mathbf{f}}\|_2^2 \leqslant (L-1) K C_{\mathcal{F},2}^{2L} \|\mathbf{Q}_1 - \mathbf{Q}_2\|_F^2$$

and

$$\|g^{\pi,\mathbf{Q},\mathbf{f}_1} - g^{\pi,\mathbf{Q},\mathbf{f}_2}\|_2^2 \leqslant LK C_{\mathcal{F},2}^{2(L-1)} \sum_{k\in\mathcal{X}} \|(\mathbf{f}_1)_k - (\mathbf{f}_2)_k\|_2^2$$

These inequalities remain true if the states of the second set of parameters are swapped. Then, we use that $C_{\mathcal{F},2} \geqslant 1$ by **[HF]** to conclude.

2.B.3 Proof of Lemma 2.8

In the following, we will identify the quadratic form M derived from the second order expansion of $x \mapsto \mathfrak{N}(x)$ and its matrix. Likewise, we will identify the quadratic form $M_{\mathfrak{C}}$ derived from the second order expansion of $x \mapsto \mathfrak{N}(I_{\mathfrak{C}}(x))$ with its matrix. Without loss of generality, one can assume $L = 3$.

Choice of parameters and expression of M. Let $\pi \in \Delta_{K^*}$ be the uniform distribution on $\mathcal{X}$, $\mathbf{Q} = \mathrm{Id}_{K^*}$ and $\mathbf{f}$ such that $\langle f_i, f_j\rangle = F\mathbf{1}_{i=j}$ for some constant $F > 0$. For instance, the f_i's are F times the indicating functions of distinct measurable sets with same measure $\frac{1}{F}$ for μ. In that case, $(f_i/\sqrt{F})_i$ is an orthonormal basis, and the quantity $g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}}$ can be broken down into three order one terms in p, q and A:

- the term in p: $\sum_i p_i f_i \otimes f_i \otimes f_i$;

- the term in q: $\sum_{i,k} q(i,k) f_i \otimes (f_i + f_k) \otimes f_k$;
- the term in A: $\sum_i ((Af)_i \otimes f_i \otimes f_i + f_i \otimes (Af)_i \otimes f_i + f_i \otimes f_i \otimes (Af)_i)$.

Now we can make the list of all second-order terms in the expansion of the quantity $\|g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+A\mathbf{f}} - g^{\pi,\mathbf{Q},\mathbf{f}}\|_2^2$:

- p and p: $F^3 \sum_i p_i^2$;
- p and q: $2F^3 \sum_i p_i q(i,i)$;
- p and A: $3F^3 \sum_i p_i A_{i,i}$;
- q and q: $2F^3 \sum_{i,k} q(i,k)^2 + 2F^3 \sum_i q(i,i)^2$;
- q and A: $F^3 \sum_{i,k} q(i,k) A_{k,i} + F^3 \sum_{i,k} q(i,k) A_{i,k} + 4F^3 \sum_i q(i,i) A_{i,i}$;
- A and A: $6F^3 \sum_i A_{i,i}^2 + 3F^3 \sum_{i,k} A_{i,k}^2$.

We can now write the matrix M. In order to clarify the structure of this matrix, let us swap the components of the parameters (p, q, A) and consider the new parameters $(A_{\text{diag}}, A_{\text{else}}, p, q_{\text{diag}}, q_{\text{else}})$, where A_{diag} (resp. q_{diag}) is a vector of size K^* containing the diagonal coefficients of A (resp. q) and A_{else} (resp. q_{else}) contains its other coefficients. Then the matrix is:

$$M_{\text{swapped}} = F^3 \left(\begin{array}{cc|c|cc} 9\mathrm{Id}_{K^*} & 0 & 3\mathrm{Id}_{K^*} & 6\mathrm{Id}_{K^*} & 0 \\ 0 & 3\mathrm{Id}_{K^*(K^*-1)} & 0 & 0 & X \\ \hline 3\mathrm{Id}_{K^*} & 0 & \mathrm{Id}_{K^*} & 2\mathrm{Id}_{K^*} & 0 \\ \hline 6\mathrm{Id}_{K^*} & 0 & 2\mathrm{Id}_{K^*} & 4\mathrm{Id}_{K^*} & 0 \\ 0 & X & 0 & 0 & 2\mathrm{Id}_{K^*(K^*-1)} \end{array} \right),$$

where $X[(A_{i,j})_{i \neq j}] = (A_{i,j} + A_{j,i})_{i \neq j}$.

Kernel of M. Substracting the first block of lines to the third and fourth blocks of lines and then the first block of columns to the third and fourth blocks of columns does not change the rank and leads to the matrix

$$F^3 \left(\begin{array}{cc|c|cc} 9\mathrm{Id}_K & 0 & 0 & 0 & 0 \\ 0 & 3\mathrm{Id}_{K(K-1)} & 0 & 0 & X \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & X & 0 & 0 & 2\mathrm{Id}_{K(K-1)} \end{array} \right)$$

Thus $\dim(\mathrm{Ker}(M)) \geqslant 2K$, where $\mathrm{Ker}(M)$ is the kernel of M and dim denotes the dimension. If one takes away the lines and columns corresponding to p and q_{diag}, one gets the matrix

$$F^3 \left(\begin{array}{ccc} 9\mathrm{Id}_{K^*} & 0 & 0 \\ 0 & 3\mathrm{Id}_{K^*(K^*-1)} & X \\ 0 & X & 2\mathrm{Id}_{K^*(K^*-1)} \end{array} \right).$$

This matrix is invertible. Therefore, $\dim(\mathrm{Ker}(M)) = 2K$. Now, for all $i \in [K^*]$, let e_i^1 and e_i^2 be the vectors defined as

$$\begin{cases} (e_i^1)_{p_k} = 0 & \text{for all } k \\ (e_i^1)_{A_{k,l}} = 0 & \text{for all } (k,l) \neq (i,i) \\ (e_i^1)_{q(k,l)} = 0 & \text{for all } (k,l) \neq (i,i) \\ (e_i^1)_{A_{i,i}} = 2 & \\ (e_i^1)_{q(i,i)} = -3 & \end{cases}$$

and

$$\begin{cases} (e_i^2)_{p_k} = 0 & \text{for all } k \neq i \\ (e_i^2)_{A_{k,l}} = 0 & \text{for all } (k,l) \neq (i,i) \\ (e_i^2)_{q(k,l)} = 0 & \text{for all } (k,l) \\ (e_i^2)_{A_{i,i}} = 1 \\ (e_i^2)_{p_i} = -3 \end{cases}$$

One can easily check that these vectors are linearly independant and are all in $\mathrm{Ker}(M)$. Thus, they are a basis of the kernel of M: $\mathrm{Ker}(M) = \mathrm{Span}(\{e_i^1, e_i^2 \,|\, i \in [K]\})$.

Nondegeneracy of M restricted on $\mathfrak{C}$. Since M is symmetric, and thus diagonalisable in an orthonormal basis,

$$M = P_{\mathrm{Ker}(M)^\perp}^\top M_{\mathrm{Ker}(M)^\perp} P_{\mathrm{Ker}(M)^\perp} \tag{2.6}$$

where $P_{\mathrm{Ker}(M)^\perp}$ is the orthogonal projection on the space of vectors orthogonal to $\mathrm{Ker}(M)$ and $M_{\mathrm{Ker}(M)^\perp}$ is a symmetric positive definite matrix, whose smallest eigenvalue will be written c_0 in the following. The last step to conclude will require the two following lemmas:

Lemma 2.18. *$Ker(M) \cap \mathfrak{C} = \{0\}$.*

Proof. Let $x \in \mathrm{Ker}(M) \cap \mathfrak{C}$, then $x = \sum_i (\lambda_i e_i^1 + \mu_i e_i^2)$ because $(e_i^1, e_i^2)_i$ is a basis of $\mathrm{Ker}(M)$. Since $x \in \mathfrak{C}$, one gets $\lambda_i = 0$ for all i because of the conditions on q. Then, the conditions on A imply $\mu_i = 0$ for all i, so that $x = 0$. □

Lemma 2.19. *There exists a constant $\kappa > 0$ such that for all $x \in \mathfrak{C}$,*

$$\|P_{Ker(M)^\perp} x\|_F^2 \geqslant \kappa \|x\|_F^2. \tag{2.7}$$

Proof. $P_{\mathrm{Ker}(M)^\perp}$ is continuous. By compacity, the quantity

$$\kappa := \inf\{\|P_{\mathrm{Ker}(M)^\perp} x\|_F^2 \,|\, x \in \mathfrak{C}, \|x\|_F^2 = 1\}$$

is reached for some $x_0 \in \mathfrak{C} \setminus \{0\}$. If $\kappa = 0$, then $x_0 \in \mathrm{Ker}(M)$, but this is impossible because of Lemma 2.18. Therefore $\kappa > 0$. □

Finally, for all $x \in \mathfrak{C}$,

$$\begin{aligned} x^\top M x =& x^\top P_{\mathrm{Ker}(M)^\perp}^\top M_{\mathrm{Ker}(M)^\perp} P_{\mathrm{Ker}(M)^\perp} x \\ =& (P_{\mathrm{Ker}(M)^\perp} x)^\top M_{\mathrm{Ker}(M)^\perp} (P_{\mathrm{Ker}(M)^\perp} x) \\ \geqslant& c_0 \|P_{\mathrm{Ker}(M)^\perp} x\|_F^2 \\ \geqslant& c_0 \kappa \|x\|_F^2. \end{aligned}$$

Therefore, the quadratic form with matrix M is nondegenerate on $\mathfrak{C}$, which shows that H is non-zero for these $(\pi, \mathbf{Q}, \mathbf{f})$. To conclude, observe that H is continuous and that our choice of parameters can be approximated by parameters satisfying **[HX]** and **[HidA]**.

2.B.4 Proof of Theorem 2.9

First, assume that the quadratic form obtained from the second order expansion of

$$\mathfrak{D} : (p, q, \mathbf{h}) \in \{(p, q, A\mathbf{f}) \,|\, (p, q, A) \in \mathfrak{C}\} \longmapsto \|g^{\pi+p,\mathbf{Q}+q,\mathbf{f}+\mathbf{h}} - g^{\pi,\mathbf{Q},\mathbf{f}}\|_2^2$$

is nondegenerate. Then, a careful reading of the proof of Theorem 8 of de Castro et al. (2016) shows that their result can be adapted to our setting and leads to the desired minoration.

Thus, what we need to show is that **[Hdet]** (which implied the nondegeneracy of the quadratic form from $\mathfrak{N}$) implies the nondegeneracy of the quadratic form from $\mathfrak{D}$ (the trick being that $\mathfrak{D}$ takes $\mathbf{h} = A\mathbf{f}$ as parameter while $\mathfrak{N}$ takes A). Assume **[Hdet]**, then there exists $c_0 > 0$ such that

$$\mathrm{Quad}_{\mathfrak{N}}(p, q, A) \geqslant c_0(\|p\|_2^2 + \|\mathbf{Q}\|_F^2 + \|A\|_F^2)$$

with the notation $\mathrm{Quad}_{\mathfrak{N}}$ referring to the quadratic form in the second order expansion of $\mathfrak{N}$.

Since **[HidA]** holds, $\mathbf{f}$ is linearly independent, so that the application $J : A \in \mathbb{R}^{K^* \times K^*} \longmapsto A\mathbf{f} \in \mathrm{Span}(\mathbf{f})^{K^*}$ is invertible. Thus, $\mathbf{h} \longmapsto \sum_{i\in\mathcal{X}} \|h_i\|_2^2$ and $\mathbf{h} \longmapsto \|J^{-1}(\mathbf{h})\|_F^2$ are two norms on the same finite-dimensional linear space $\mathrm{Span}(\mathbf{f})^{K^*}$, so that they are equivalent. In particular, there exists a constant $c_1 \leqslant 1$ such that $\|J^{-1}(\mathbf{h})\|_F^2 \geqslant c_1 \sum_{i\in\mathcal{X}} \|h_i\|_2^2$. Therefore,

$$\begin{aligned}\mathrm{Quad}_{\mathfrak{D}}(p, q, \mathbf{h}) &= \mathrm{Quad}_{\mathfrak{N}}(p, q, J^{-1}(\mathbf{h})) \\ &\geqslant c_0 c_1 \left(\|p\|_2^2 + \|\mathbf{Q}\|_F^2 + \sum_{i\in\mathcal{X}} \|h_i\|_2^2 \right),\end{aligned}$$

which is what we wanted to prove.

2.C Proof of the control of $Z_{K,M}$

This section contains the proof of Lemma 2.16.

2.C.1 Concentration inequality on $Z_{K,M}$

Define for all $\sigma > 0$ the sets

$$B_\sigma = \{t \in S_{K,M},\ C_{\mathcal{F},\infty}^{L/2} \|t - s_{K,M}\|_2 \leqslant \sigma\}$$

Let d_{g^*} be the semi-distance defined by $d_{g^*}^2(t_1, t_2) = \mathbb{E}[(t_1 - t_2)^2(Z_1)] = \int g^*(t_1 - t_2)^2 \mathrm{d}\mu^{\otimes L}$, and d_2 the distance induced by the norm on $\mathbf{L}^2(\mathcal{Y}^L, \mu^{\otimes L})$.

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

The following lemma is a Bernstein-like inequality that follows from Paulin (2013), Theorem 2.4:

Lemma 2.20. *Let t be a real valued and measurable bounded function on $\mathcal{Y}^L$. Let $V = \mathbb{E}[t^2(Z_1)]$. There exists a positive constant c^* depending only on $\mathbf{Q}^*$ and L such that for all $0 \leqslant \lambda \leqslant 1/(2\sqrt{2}c^*\|t\|_\infty)$ and for all $n \in \mathbb{N}$:*

$$\log \mathbb{E} \exp\left[\lambda \sum_{s=1}^{n} (t(Z_s) - \mathbb{E}t(Z_s))\right] \leqslant \frac{2nc^*V\lambda^2}{1 - 2\sqrt{2}c^*\|t\|_\infty\lambda}$$

The following lemma is an extension of Theorem 6.8 from Massart (2007) and allows to obtain a concentration inequality on the supremum on all functions of a class when one can control its bracketing entropy.

Lemma 2.21. *Let Ξ be some measurable space, $(\xi_i)_{1\leqslant i\leqslant n}$ a sequence of random variables on Ξ, $\mathcal{T}$ some countable class of real valued and measurable functions on Ξ. Assume that there exists some positive numbers a and b such that for all $t\in\mathcal{T}$, $\|t\|_\infty\leqslant b$ and $\sup_i \mathbb{E}[t^2(\xi_i)]\leqslant a^2$.*

Assume also that there exists some constant $C_\xi\geqslant 1/4$ such that for all $0\leqslant\lambda\leqslant 1/(2\sqrt{2}C_\xi b)$ and for all $t\in\mathcal{T}$:

$$\log\mathbb{E}\exp\left[\lambda\sum_{s=1}^{n}(t(\xi_s)-\mathbb{E}t(\xi_s))\right]\leqslant\frac{2nC_\xi a^2\lambda^2}{1-2\sqrt{2}C_\xi b\lambda} \tag{2.8}$$

Assume furthermore that for any positive number δ, there exists some finite set $\mathcal{B}_\delta$ of brackets covering $\mathcal{T}$ such that for any bracket $[t_1,t_2]\in\mathcal{B}_\delta$, $\|t_1-t_2\|_\infty\leqslant b$ and $\sup_i\mathbb{E}[(t_1-t_2)^2(\xi_i)]\leqslant\delta^2$. Let $e^{H(\delta)}$ denote the minimal cardinality of such a covering. Then, there exists a numerical constant $\kappa>0$ such that for any measurable set A such that $\mathbb{P}(A)>0$,

$$\mathbb{E}^A\left(\sup_{t\in\mathcal{T}}\sum_{s=1}^{n}(t(\xi_s)-\mathbb{E}t(\xi_s))\right)\leqslant\kappa C_\xi\left[E+a\sqrt{n\log\left(\frac{1}{\mathbb{P}(A)}\right)}+b\log\left(\frac{1}{\mathbb{P}(A)}\right)\right]$$

where

$$E=\sqrt{n}\int_0^a\sqrt{H(u)\wedge n}du+(b+a)H(a)$$

and for any measurable random variable W, $\mathbb{E}^A[W]=\mathbb{E}[W\mathbf{1}_A]/\mathbb{P}(A)$.

Remark. *The assumption $C_\xi\geqslant 1/4$ is only used to factorise the upper bound by C_ξ. Without it, the upper bound would be*

$$\kappa'\left[E+a\sqrt{C_\xi n\log\left(\frac{1}{\mathbb{P}(A)}\right)}+bC_\xi\log\left(\frac{1}{\mathbb{P}(A)}\right)\right]$$

In practice, this assumption doesn't cost anything: if equation (2.8) holds for some constant C_ξ, then it holds for any constant $C'\geqslant C_\xi$.

We will apply this lemma to $\Xi=\mathcal{Y}^L$ and $\xi_i=Z_i$. Using Lemma 2.20, equation (2.8) holds with $C_\xi=\max(c^*,1/4)$.

Take $\mathcal{T}=(B_\sigma-s_{K,M})\bigcup(-B_\sigma+s_{K,M})$, so that $\sup_{t\in\mathcal{T}}\nu(t)=\sup_{t\in B_\sigma}|\nu(t-s_{K,M})|$. Then, one can check using Lemma 2.28 that the assumptions $\|t\|_\infty\leqslant b$ and $\sup_i\mathbb{E}[t^2(\xi_i)]\leqslant a^2$ hold with

- $b=2C_{\mathcal{F},\infty}^L$
- $a=2\min(\sigma,C_{\mathcal{F},\infty}^{L/2}C_{\mathcal{F},2}^L)$

and $H(u)\leqslant\log(2)+H(u,B_\sigma-s_{K,M},d_{g^*})$.

We can do without assuming $\mathcal{T}$ to be countable. Indeed, ν is continuous on $\mathcal{T}$ equipped with the infinity norm. This entails that the supremum of ν over $\mathcal{T}$ is equal to the supremum of ν over any dense subset of $(\mathcal{T},\|\cdot\|_\infty)$. Since $\mathcal{T}\subset(\mathfrak{P}_M)^{\otimes L}$, which is a finite dimensional metric linear space for the infinity norm, it is separable. Therefore, without loss of generality, we can get rid of the countability assumption on $\mathcal{T}$.

Rewriting these results, we get the following lemma:

Lemma 2.22. *There exists a constant C^* depending only on $\mathbf{Q}^*$ and L such that for all $\sigma > 0$, for all measurable A such that $\mathbb{P}(A) > 0$:*

$$\mathbb{E}^A\left(\sup_{t\in B_\sigma}|\nu(t-s_{K,M})|\right) \leqslant C^*\left[\frac{E}{n}+\sigma\sqrt{\frac{1}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)}+\frac{2C^L_{\mathcal{F},\infty}}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)\right]$$

where

$$\begin{aligned}E =&\sqrt{n}\int_0^\sigma\sqrt{H(u,B_\sigma-s_{K,M},d_{g^*})\wedge n}du+\log(2)\sigma\sqrt{n}\\&+2(C^L_{\mathcal{F},\infty}+C^{L/2}_{\mathcal{F},\infty}C^L_{\mathcal{F},2})H(\sigma,B_\sigma-s_{K,M},d_{g^*})\end{aligned}$$

The core of the proof consists in controlling the bracketing entropy in order to find a "good" function φ and constants C and $\sigma_{K,M}$ depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$ and L such that $x\mapsto\frac{\varphi(x)}{x}$ is nonincreasing and

$$\forall\sigma\geqslant\sigma_{K,M}\qquad E\leqslant C\varphi(\sigma)\sqrt{n}. \tag{2.9}$$

For ease of notation, we did not write the dependency of C and φ on K and M.

Let us see how to conclude with such an inequality. We shall use the following result (lemma 4.23 from Massart (2007)).

Lemma 2.23. *Let S be some countable set, $u\in S$ and $a:S\mapsto\mathbb{R}_+$ such that $a(u)=\inf_{t\in S}a(t)$. Let Z be some process indexed by S and assume that $\sup_{t\in\mathbf{B}(\lambda)}Z(t)-Z(u)$ has finite expectation for any positive number $\lambda\geqslant 0$, where*

$$\mathbf{B}(\lambda)=\{t\in S, a(t)\leqslant\lambda\}$$

Then, for any function ϕ on $\mathbb{R}_+$ such that $x\mapsto\phi(x)/x$ is nonincreasing on $\mathbb{R}_+$ and satisfies for some $\lambda_*\geqslant 0$ to

$$\forall\lambda\geqslant\lambda_*\geqslant 0,\qquad \mathbb{E}\left[\sup_{t\in\mathbf{B}(\lambda)}Z(t)-Z(u)\right]\leqslant\phi(\lambda)$$

one has for any $x\geqslant\lambda_*$:

$$\mathbb{E}\left[\sup_{t\in S}\left(\frac{Z(t)-Z(u)}{a(t)^2+x^2}\right)\right]\leqslant 4\frac{\phi(x)}{x^2}$$

In our case,

$$\begin{cases}S=S_{K,M}-s_{K,M}\\u=s_{K,M}\\a(t)=C^{L/2}_{\mathcal{F},\infty}\|t-s_{K,M}\|_2\\Z(t)=|\nu(t-s_{K,M})|\\\lambda_*=\sigma_{K,M}\\\phi(x)=C^*\left[C\frac{\varphi(x)}{\sqrt{n}}+x\sqrt{\frac{1}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)}+\frac{2C^L_{\mathcal{F},\infty}}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)\right]\end{cases}$$

With this choice of S, a and Z, this proposition holds even if S is not countable for the same reason as in Lemma 2.22.

It follows that for all $x\geqslant\sigma_{K,M}$:

$$\mathbb{E}^A\left[\sup_{t\in S_{K,M}}\frac{|\nu(t-s_{K,M})|}{C^L_{\mathcal{F},\infty}\|t-s_{K,M}\|_2^2+x^2}\right]\leqslant 4\frac{\phi(x)}{x^2},$$

so that if $x_{K,M} \geqslant \frac{\sigma_{K,M}}{C_{\mathcal{F},\infty}^{L/2}}$:

$$\mathbb{E}^A[Z_{K,M}(s)] \leqslant 4\frac{\phi(x_{K,M}C_{\mathcal{F},\infty}^{L/2})}{x_{K,M}^2}$$

and then

$$\begin{aligned}\mathbb{E}^A[Z_{K,M}(s)] \leqslant & 4\frac{C^*}{x_{K,M}^2}\left[C\frac{\varphi(x_{K,M}C_{\mathcal{F},\infty}^{L/2})}{\sqrt{n}} + x_{K,M}C_{\mathcal{F},\infty}^{L/2}\sqrt{\frac{1}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)}\right.\\ & \left. + \frac{2C_{\mathcal{F},\infty}^{L}}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)\right]\\ =: & \psi\left(\log\left(\frac{1}{\mathbb{P}(A)}\right)\right)\end{aligned}$$

Note that the function ψ is nondecreasing. On the event $A = \{Z_{K,M}(s) \geqslant \psi(x)\}$,

$$\psi(x) \leqslant \mathbb{E}^A[Z_{K,M}(s)] \leqslant \psi\left(\log\left(\frac{1}{\mathbb{P}(A)}\right)\right)$$

so that $x \leqslant \log\left(\frac{1}{\mathbb{P}(A)}\right)$ and finally $\mathbb{P}(A) \leqslant e^{-x}$.

It follows that with probability $1 - e^{-z_{K,M}-z}$:

$$Z_{K,M}(s) \leqslant 4C^*\left[C\frac{\varphi(x_{K,M}C_{\mathcal{F},\infty}^{L/2})}{x_{K,M}^2\sqrt{n}} + C_{\mathcal{F},\infty}^{L/2}\sqrt{\frac{z_{K,M}+z}{x_{K,M}^2 n}} + 2C_{\mathcal{F},\infty}^{L}\frac{z_{K,M}+z}{x_{K,M}^2 n}\right] \tag{2.10}$$

and the last step of the proof will be to choose the right $x_{K,M}$ and $z_{K,M}$ (see Section 2.C.3).

2.C.2 Control of the bracketing entropy

The goal of this section is to prove equation (2.9), that is to find φ, C and $\sigma_{K,M}$ such that

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

The bracketing entropy is invariant under translation and increasing with respect to the inclusion relation, so

$$H(u, B_\sigma - s_{K,M}, d_{g^*}) = H(u, B_\sigma, d_{g^*}) \leqslant H(u, S_{K,M}, d_{g^*})$$

Using Lemma 2.28, we get that for all $t \in \mathbf{L}^2(\mathcal{Y}^L, \mathbb{R})$, $\int t^2 g^* \mathrm{d}\mu \leqslant C_{\mathcal{F},\infty}^L \|t\|_2^2$. Therefore, a bracket of size $u/C_{\mathcal{F},\infty}^{L/2}$ for d_2 is also a bracket of size u for d_{g^*}, which implies that

$$H(u, B_\sigma - s_{K,M}, d_{g^*}) \leqslant H\left(\frac{u}{C_{\mathcal{F},\infty}^{L/2}}, S_{K,M}, d_2\right) \tag{2.11}$$

Let us now rewrite the definition of $S_{K,M}$:

$$\begin{aligned}S_{K,M} &= \left\{\sum_{\mathbf{k}\in\{1,\dots,K\}^L} \pi_{k_1}\prod_{i=2}^{L}\mathbf{Q}_{k_{i-1},k_i}\bigotimes_{i=1}^{L} f_{k_i},\ \mathbf{Q}\in\mathcal{Q}_K,\ \pi\mathbf{Q}=\pi,\ \mathbf{f}\in(\mathcal{F}\cap\mathfrak{P}_M)^K\right\}\\ &\subset \left\{\sum_{\mathbf{k}\in\{1,\dots,K\}^L}\mu_{\mathbf{k}}\phi_{\mathbf{k}},\ \mu\in\mathcal{U},\ \phi\in\Phi\right\}\end{aligned}$$

where

$$\begin{cases} \mathcal{U} = \left\{ (\pi_{k_1} \prod_{i=2}^{L} \mathbf{Q}_{k_{i-1},k_i})_{k_1,\dots,k_L},\ \mathbf{Q} \text{ transition matrix } K \times K,\ \pi \geqslant 0,\ \pi \in \mathbb{S}_{K-1} \right\} \\ \Phi = \left\{ (\bigotimes_{i=1}^{L} f_{k_i})_{k_1,\dots,k_L},\ \mathbf{f} \in (\mathcal{F} \cap \mathfrak{P}_M)^K \right\} \end{cases}$$

$\mathcal{U}$ is equipped with the distance $d_2(a,b) = (\sum_{\mathbf{k}} (b^i_{\mathbf{k}} - a^i_{\mathbf{k}})^2)^{1/2}$. A bracket for $\mathcal{U}$ will be a set $[a,b] = \{c \,|\, \forall \mathbf{k} \in \{1,\dots,K\}^L,\ a_{\mathbf{k}} \leqslant c_{\mathbf{k}} \leqslant b_{\mathbf{k}}\}$.

Φ is equipped with the distance $d_{\infty,2}(u,v) = \max_{\mathbf{k}} \|v^i_{\mathbf{k}} - u^i_{\mathbf{k}}\|_2$. A bracket Φ will be a set $[u,v] = \{t \,|\, \forall \mathbf{k} \in \{1,\dots,K\}^L,\ u_{\mathbf{k}}(\cdot) \leqslant t_{\mathbf{k}}(\cdot) \leqslant v_{\mathbf{k}}(\cdot)\}$.

Controlling the bracketing entropy on each of these sets will allow to control the bracketing entropy of $S_{K,M}$. Let us start with them:

Lemma 2.24. *There exists a bracket covering $\{[a^i,b^i]\}_{1\leqslant i\leqslant N_{\mathcal{U}}(\epsilon)}$ of size ϵ of $\mathcal{U}$ for the distance d_2 with cardinality*

$$N_{\mathcal{U}}(\epsilon) \leqslant \max\left(\frac{2LK^{L/2}}{\epsilon}, 1\right)^{K^2-1} \tag{2.12}$$

such that for all i and $\mathbf{k}$, $0 \leqslant a^i_{\mathbf{k}} \leqslant 1$.

Lemma 2.25. *There exists a bracket covering $\{[u^i,v^i]\}_{1\leqslant i\leqslant N_\Phi(\epsilon)}$ of size ϵ of Φ for the distance $d_{\infty,2}$ of cardinality*

$$N_\Phi(\epsilon) \leqslant \max\left(\frac{L(4C_{\mathcal{F},2}M)^L}{\epsilon}, 1\right)^{MK} \tag{2.13}$$

such that

$$\max_i \max_{\mathbf{k}} \|v^i_{\mathbf{k}}\|_2^2 \leqslant (8M^2C^2_{\mathcal{F},2})^L.$$

Let us take such bracketings and consider the following set of brackets:

$$\left\{ \left[\sum_{\mathbf{k}} \mathcal{A}^{i,j}_{\mathbf{k}}, \sum_{\mathbf{k}} \mathcal{B}^{i,j}_{\mathbf{k}} \right] \right\}_{1\leqslant i\leqslant N_{\mathcal{U}}(\epsilon), 1\leqslant j\leqslant N_\Phi(\epsilon)}$$

where

$$\forall y \in \mathcal{Y}^L, \quad \begin{cases} \mathcal{A}^{i,j}_{\mathbf{k}}(y) = \min\{a^i_{\mathbf{k}} u^j_{\mathbf{k}}(y), b^i_{\mathbf{k}} v^j_{\mathbf{k}(y)}\} \\ \mathcal{B}^{i,j}_{\mathbf{k}}(y) = \max\{a^i_{\mathbf{k}} u^j_{\mathbf{k}}(y), b^i_{\mathbf{k}} v^j_{\mathbf{k}(y)}\} \end{cases}.$$

This set covers $S_{K,M}$: for all $\mu \in \mathcal{U}$, $\phi \in \Phi$, there exists $i \in \{1,\dots,N_{\mathcal{U}}(\epsilon)\}$ and $j \in \{1,\dots,N_\Phi(\epsilon)\}$ such that $\mu \in [a^i,b^i]$ and $\phi \in [u^j,v^j]$, and then by construction $\sum_{\mathbf{k}} \mu_{\mathbf{k}}\phi_{\mathbf{k}} \in [\sum_{\mathbf{k}} \mathcal{A}^{i,j}_{\mathbf{k}}, \sum_{\mathbf{k}} \mathcal{B}^{i,j}_{\mathbf{k}}]$.

Let us now bound the size of these brackets. Let $[a,b] \in \{[a^i,b^i]\}_{1\leqslant i\leqslant N_{\mathcal{U}}(\epsilon)}$ and $[u,v] \in \{[u^i,v^i]\}_{1\leqslant i\leqslant N_\Phi(\epsilon)}$, then if one denotes by $[\mathcal{A},\mathcal{B}]$ the corresponding bracket, there exists $(\sigma_{\mathbf{k}})_{\mathbf{k}} \in$

$\{-1,1\}^{K^L}$ such that:

$$\begin{aligned}
\left\|\sum_{\mathbf{k}}\mathcal{A}_{\mathbf{k}}-\sum_{\mathbf{k}}\mathcal{B}_{\mathbf{k}}\right\|_2^2 &= \left\|\sum_{\mathbf{k}}\sigma_{\mathbf{k}}(b_{\mathbf{k}}v_{\mathbf{k}}-a_{\mathbf{k}}u_{\mathbf{k}})\right\|_2^2 \\
&\leqslant \left\|\sum_{\mathbf{k}}|b_{\mathbf{k}}v_{\mathbf{k}}-a_{\mathbf{k}}u_{\mathbf{k}}|\right\|_2^2 \\
&\leqslant K^L\sum_{\mathbf{k}}\|a_{\mathbf{k}}u_{\mathbf{k}}-b_{\mathbf{k}}v_{\mathbf{k}}\|_2^2 \\
&= K^L\sum_{\mathbf{k}}\|(a_{\mathbf{k}}-b_{\mathbf{k}})v_{\mathbf{k}}+a_{\mathbf{k}}(u_{\mathbf{k}}-v_{\mathbf{k}})\|_2^2 \\
&\leqslant 2K^L\left(\sum_{\mathbf{k}}\|(a_{\mathbf{k}}-b_{\mathbf{k}})v_{\mathbf{k}}\|_2^2+\sum_{\mathbf{k}}\|a_{\mathbf{k}}(u_{\mathbf{k}}-v_{\mathbf{k}})\|_2^2\right) \\
&= 2K^L\left(\sum_{\mathbf{k}}(a_{\mathbf{k}}-b_{\mathbf{k}})^2\|v_{\mathbf{k}}\|_2^2+\sum_{\mathbf{k}}a_{\mathbf{k}}^2\|u_{\mathbf{k}}-v_{\mathbf{k}}\|_2^2\right).
\end{aligned}$$

Then, by definition of the brackets, $\|u_{\mathbf{k}}-v_{\mathbf{k}}\|_2^2\leqslant\epsilon^2$ and $\sum_{\mathbf{k}}(a_{\mathbf{k}}-b_{\mathbf{k}})^2\leqslant\epsilon^2$. In addition, we assumed $|a_{\mathbf{k}}|\leqslant 1$ and $\|v_{\mathbf{k}}\|_2^2\leqslant(8M^2C_{\mathcal{F},2}^2)^L$ for all $\mathbf{k}$, so that

$$\|\sum_{\mathbf{k}}\mathcal{A}_{\mathbf{k}}-\sum_{\mathbf{k}}\mathcal{B}_{\mathbf{k}}\|_2^2\leqslant 2K^L\epsilon^2((8M^2C_{\mathcal{F},2}^2)^L+K^L),$$

which implies

$$\begin{aligned}
N(\epsilon,S_{K,M},d_2)\leqslant N_{\mathcal{U}}&\left(\frac{\epsilon}{\sqrt{2K^L(K^L+(8M^2C_{\mathcal{F},2}^2)^L)}}\right) \\
&\times N_{\Phi}\left(\frac{\epsilon}{\sqrt{2K^L(K^L+(8M^2C_{\mathcal{F},2}^2)^L)}}\right), \quad (2.14)
\end{aligned}$$

and finally by combining 2.11, 2.12, 2.13 and 2.14:

$$\begin{aligned}
H(u, B_\sigma - s_{K,M}, d_{g^*}) &\leqslant (K^2-1)\log\max\left(\frac{C_{\mathcal{F},\infty}^{L/2}\sqrt{2K^L(K^L+(8M^2C_{\mathcal{F},2}^2)^L)}2LK^{L/2}}{u},1\right)\\
&\quad + MK\log\max\left(\frac{C_{\mathcal{F},\infty}^{L/2}\sqrt{2K^L(K^L+(8M^2C_{\mathcal{F},2}^2)^L)}L(4C_{\mathcal{F},2}M)^L}{u},1\right)\\
&\leqslant (MK+K^2-1)\log\max\left(\frac{C_{\mathcal{F},\infty}^{L/2}\sqrt{2(K^L+(8M^2C_{\mathcal{F},2}^2)^L)}LK^L(4C_{\mathcal{F},2}M)^L}{u},1\right)\\
&\leqslant (MK+K^2-1)\log\max\left(\frac{C_{\mathcal{F},\infty}^{L/2}\sqrt{2(n^L+(8C_{\mathcal{F},2}^2)^Ln^{2L})}nn^L(4C_{\mathcal{F},2})^Ln^L}{u},1\right)\\
&\leqslant (MK+K^2-1)\log\max\left(\frac{C_{\mathcal{F},\infty}^{L/2}2(8C_{\mathcal{F},2}^2)^{L/2}n^Lnn^L(4C_{\mathcal{F},2})^Ln^L}{u},1\right)\\
&\leqslant (MK+K^2-1)\log\max\left(\frac{2(16C_{\mathcal{F},\infty}^{1/2}C_{\mathcal{F},2}^2)^Ln^{3L+1}}{u},1\right)\\
&\leqslant (MK+K^2-1)\log\max\left(\frac{n^{6L}}{u},1\right)
\end{aligned}$$

for n large enough ($n \geqslant n_0 := 16C_{\mathcal{F},\infty}^{1/2}C_{\mathcal{F},2}^2$) because we assumed $M \leqslant n$, $K \leqslant n$, $L \leqslant n$, $C_{\mathcal{F},2} \geqslant 1$ and $C_{\mathcal{F},\infty} \geqslant 1$. Thus, Lemma 2.29 implies that if we write $C_0 = \sqrt{\pi}$ and

$$\varphi(\sigma) = C_0\sigma\sqrt{MK+K^2-1}\left(1+\sqrt{\log\left(\max\left\{\frac{n^{6L}}{\sigma},1\right\}\right)}\right)$$

then for all $n \geqslant n_0$ and $\sigma > 0$,

$$\begin{cases}
\sigma^2 H(\sigma, S_{K,M}, d_2) \leqslant \varphi(\sigma)^2\\
\displaystyle\int_0^\sigma \sqrt{H(u, S_{K,M}, d_2)}du \leqslant \varphi(\sigma)\\
\log(2)\sigma \leqslant \varphi(\sigma)
\end{cases}$$

Let us now check that this function φ satisfies equation (2.9). First, note that $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing, so that $x \mapsto \frac{\varphi(x)}{x^2}$ is also nonincreasing. Thus, we may define $\sigma_{K,M}$ as the unique solution of the equation $\varphi(x) = \sqrt{n}x^2$, and then for all $\sigma \geqslant \sigma_{K,M}$:

$$H(\sigma, B_\sigma - s_{K,M}, d_{g^*}) \leqslant \frac{\varphi(\sigma)^2}{\sigma^2} \leqslant \frac{\varphi(\sigma)}{\sigma}\sigma\sqrt{n} = \varphi(\sigma)\sqrt{n}$$

Equation (2.9) follows immediately with $C = 2(1 + C_{\mathcal{F},\infty}^L + C_{\mathcal{F},\infty}^{L/2}C_{\mathcal{F},2}^L)$.

Proof of Lemma 2.24 Let $\epsilon \in (0,2)$.

We start with the family $\{[k/n, (k+1)/n],\ k \in \{0,\dots,n-1\}\}$ with n an integer between $1/\epsilon$ and $2/\epsilon$, which gives a bracket covering of size ϵ of $[0,1]$ with cardinality smaller than $2/\epsilon$. These brackets will be used to control each free component of $\mathbf{Q}$ and π, that is K^2-1 components.

More precisely, we define the following bracket set:

$$\Big\{[A,B] \mid A_{\mathbf{k}} = \frac{1}{n^L} p_{k_1} \prod_{i=2}^{L} a_{k_{i-1},k_i},\ B_{\mathbf{k}} = \frac{1}{n^L}(p_{k_1}+1)\prod_{i=2}^{L}(a_{k_{i-1},k_i}+1),$$
$$p \in \{0,\dots,n-1\}^{K-1},\ \sum_{k=1}^{K-1} p_k < n,\ p_K = n - \sum_{k=1}^{K-1}(p_k+1),$$
$$a \in \{0,\dots,n-1\}^{K\times(K-1)},\ \forall i \in \{1,\dots,K\},\ \sum_{k=1}^{K-1} a_{i,k} < n \text{ and}$$
$$a_{i,K} = n - \sum_{k=1}^{K-1}(a_{i,k}+1)\Big\}.$$

This set covers $\mathcal{U}$ and its cardinality is smaller than $\left(\frac{2}{\epsilon}\right)^{K^2-1}$. To get the bracket's size, note that

$$\begin{aligned}
&\sum_{\mathbf{k}\in\{1,\dots,K\}^L} \left(\frac{1}{n^L} p_{k_1}\prod_{i=2}^{L} a_{k_{i-1},k_i} - \frac{1}{n^L}(p_{k_1}+1)\prod_{i=2}^{L}(a_{k_{i-1},k_i}+1)\right)^2 \\
&\qquad = \frac{1}{n^{2L}} \sum_{\mathbf{k}\in\{1,\dots,K\}^L} \left(\prod_{i=2}^{L} a_{k_{i-1},k_i} + \sum_{j=2}^{L} p_{k_1} \prod_{i\neq j, i\geqslant 2} a_{k_{i-1},k_i}\right)^2 \\
&\qquad \leqslant \frac{L^2 n^{2L-2} K^L}{n^{2L}} \\
&\qquad \leqslant L^2 K^L \epsilon^2,
\end{aligned}$$

and in the end

$$N(u,\mathcal{U},d_2) \leqslant \max\left(\frac{LK^{L/2}}{u}, 1\right)^{K^2-1}.$$

Proof of Lemma 2.25 All $f \in \mathcal{F}\cap\mathfrak{P}_M$ can be written as $\sum_{m=1}^{M}\lambda_m\varphi_m$ where $(\varphi_m)_{m\in\{1,\dots,M\}}$ is an orthonormal basis of $\mathfrak{P}_M$. Then, assumption **[HF]** implies that $|\lambda_m| \leqslant C_{\mathcal{F},2}$ for all $m \in \{1,\dots,M\}$.

We will therefore start from a bracket covering of the euclidian ball of radius $C_{\mathcal{F},2}$ of $\mathbb{R}^M$, from which we will construct a covering of $\mathcal{F}\cap\mathfrak{P}_M$ and of Φ.

Lemma 2.26. *Let $\epsilon \in (0,4)$. There exists a bracket covering $\{[a^i,b^i]\}_{1\leqslant i\leqslant N_M}$ of size ϵ of the euclidian ball of radius $C_{\mathcal{F},2}$ of $\mathbb{R}^M$ with cardinality*

$$N_M \leqslant \max\left(\frac{4C_{\mathcal{F},2}\sqrt{M}}{\epsilon}, 1\right)^M$$

such that for all $m \in \{1,\dots,M\}$, $i \in \{1,\dots,N_M\}$, $-C_{\mathcal{F},2} \leqslant a_m^i \leqslant b_m^i \leqslant C_{\mathcal{F},2}$.

Proof. We start with a bracket covering of size $\epsilon/\sqrt{M}$ of the infinity ball of radius $C_{\mathcal{F},2}$ of $\mathbb{R}^M$. This can be done by a regular partition with $\max(\lceil 2C_{\mathcal{F},2}/\epsilon\rceil, 1)$ pieces along each coordinate. One can easily check that such a covering is also a covering of size ϵ of the euclidian ball of radius $C_{\mathcal{F},2}$ of $\mathbb{R}^M$. To conclude, it is enough to notice that $\lceil x\rceil \leqslant 2x$ as soon as $x > 1/2$, and that $2C_{\mathcal{F},2}/\epsilon > 1/2$ because $C_{\mathcal{F},2} \geqslant 1$ and $\epsilon < 4$. □

Let $\{[a^i, b^i]\}_{1\leqslant i\leqslant N_M}$ be such a covering. For all $m \in \{1,\dots,M\}$, $i \in \{1,\dots,N_M\}$ and $y \in \mathcal{Y}$, let

$$u_m^i(y) = \begin{cases} a_m^i & \text{if } \varphi_m(y) \leqslant 0 \\ b_m^i & \text{otherwise} \end{cases}$$
$$v_m^i(y) = a_m^i + b_m^i - u_m^i(y)$$

and for all $i \in \{1,\dots,N_M\}$ and $y \in \mathcal{Y}$,

$$\begin{cases} U_1^i(y) = \sum_{m=1}^M u_m^i(y)\varphi_m(y) \\ U_2^i(y) = \sum_{m=1}^M v_m^i(y)\varphi_m(y) \end{cases}$$

and finally for all $\mathbf{i} = (i_1,\dots,i_K) \in \{1,\dots,N_M\}^K$ and $\mathbf{k} = (k_1,\dots,k_L) \in \{1,\dots,K\}^L$:

$$\begin{cases} (V^{\mathbf{i}})_{\mathbf{k}} = \min\left\{ \bigotimes_{\beta=1}^L U_{\sigma_\beta}^{i_{k_\beta}};\quad \sigma \in \{1,2\}^L \right\} \\ (W^{\mathbf{i}})_{\mathbf{k}} = \max\left\{ \bigotimes_{\beta=1}^L U_{\sigma_\beta}^{i_{k_\beta}};\quad \sigma \in \{1,2\}^L \right\} \end{cases}.$$

It is enough to show that $\{[V^{\mathbf{i}}, W^{\mathbf{i}}],\ \mathbf{i} \in \{1,\dots,N_M\}^K\}$ is a bracket covering of size $L(4C_{\mathcal{F},2}M)^{L-1}\sqrt{M}\epsilon$ of Φ that satisfies

$$\max_i \max_{\mathbf{k}} \int (W_{\mathbf{k}}^i)^2 \mathrm{d}\mu^{\otimes L} \leqslant (8M^2 C_{\mathcal{F},2}^2)^L.$$

Applying the Cauchy-Schwarz inequality, one gets that for all $i \in \{1,\dots,N_M\}$,

$$\begin{aligned} \|U_2^i - U_1^i\|_2^2 &= \| \sum_{m=1}^M |b_m^i - a_m^i|.|\varphi_m| \|_2^2 \\ &\leqslant M\|b^i - a^i\|_2^2 \\ &\leqslant M\epsilon^2. \end{aligned}$$

Moreover, for all $i \in \{1,\dots,N_M\}$ and $\sigma \in \{1,2\}$,

$$\begin{aligned} \|U_\sigma^i\|_2^2 &\leqslant \| \sum_{m_i=1}^M |b_m^i + a_m^i|.|\varphi_m| \|_2^2 \\ &\leqslant 2M(\|a^i\|_2^2 + \|b^i\|_2^2) \\ &\leqslant 4M^2 C_{\mathcal{F},2}^2. \end{aligned}$$

We then use that for all $\mathbf{i} \in \{1,\dots,N_M\}^K$ and $\mathbf{k} \in \{1,\dots,K\}^L$,

$$\begin{aligned} |W_{\mathbf{k}}^{\mathbf{i}} - V_{\mathbf{k}}^{\mathbf{i}}|(y) &\leqslant \sum_{\gamma=1}^L \left| U_2^{i_{k_\gamma}} - U_1^{i_{k_\gamma}} \right|(y_\gamma) \max_{\mathbf{j}\in\{1,2\}^L} \prod_{\beta\neq\gamma,\beta=1}^L \left| U_{j_\beta}^{i_{k_\beta}} \right|(y_\beta) \\ &\leqslant \sum_{\gamma=1}^L \left| U_2^{i_{k_\gamma}} - U_1^{i_{k_\gamma}} \right|(y_\gamma) \prod_{\beta\neq\gamma,\beta=1}^L \left(\left| U_1^{i_{k_\beta}} \right| + \left| U_2^{i_{k_\beta}} \right| \right)(y_\beta) \end{aligned}$$

so that

$$\begin{aligned}\|W_{\mathbf{k}}^{\mathbf{i}} - V_{\mathbf{k}}^{\mathbf{i}}\|_2^2 \leqslant & L\sum_{\gamma=1}^{L}\left\|U_2^{i_{k_\gamma}} - U_1^{i_{k_\gamma}}\right\|_2^2 \prod_{\beta\neq\gamma,\beta=1}^{L} 2\left(\left\|U_1^{i_{k_\beta}}\right\|_2^2 + \left\|U_2^{i_{k_\beta}}\right\|_2^2\right)\\ \leqslant & L2^{L-1}\sum_{\gamma=1}^{L} M\epsilon^2 \prod_{\beta\neq\gamma,\beta=1}^{L}(2\times 4M^2C_{\mathcal{F},2}^2)\\ = & L^2(16M^2C_{\mathcal{F},2}^2)^{L-1}M\epsilon^2\\ = & (L(4C_{\mathcal{F},2}M)^{L-1}\sqrt{M}\epsilon)^2\end{aligned}$$

and finally $d_{\infty,2}(W^{\mathbf{i}}, V^{\mathbf{i}}) \leqslant L(4C_{\mathcal{F},2}M)^{L-1}\sqrt{M}\epsilon$ for all $\mathbf{i} \in \{1,\dots,N_M\}^K$.

The last part of the lemma is proved by noting that for all $\mathbf{i}$ and $\mathbf{k}$,

$$\begin{aligned}(W^{\mathbf{i}})_{\mathbf{k}}^2 &= \max\{\bigotimes_{\beta=1}^{L}(U_{\sigma_\beta}^{i_{k_\beta}})^2; \quad \sigma\in\{1,2\}^L\}\\ &\leqslant \sum_{\sigma\in\{1,2\}^L}\bigotimes_{\beta=1}^{L}(U_{\sigma_\beta}^{i_{k_\beta}})^2,\end{aligned}$$

so that

$$\begin{aligned}\int (W^{\mathbf{i}})_{\mathbf{k}}^2 \mathrm{d}\mu^{\otimes L} &\leqslant \sum_{\sigma\in\{1,2\}^L}\prod_{\beta=1}^{L}\|U_{\sigma_\beta}^{i_{k_\beta}}\|_2^2\\ &\leqslant \sum_{\sigma\in\{1,2\}^L}(4M^2C_{\mathcal{F},2}^2)^L\\ &\leqslant (8M^2C_{\mathcal{F},2}^2)^L.\end{aligned}$$

2.C.3 Choice of parameters

Let us come back to equation (2.10). Since $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing, one has $\frac{\varphi(x_{K,M}C_{\mathcal{F},\infty}^{L/2})}{x_{K,M}\sqrt{n}} \leqslant \sigma_{K,M}C_{\mathcal{F},\infty}^{L/2}$ as soon as $x_{K,M} \geqslant \frac{\sigma_{K,M}}{C_{\mathcal{F},\infty}^{L/2}}$, so with probability $1-e^{-z_{K,M}-z}$:

$$Z_{K,M}(s) \leqslant 4C^*\left[CC_{\mathcal{F},\infty}^{L/2}\frac{\sigma_{K,M}}{x_{K,M}} + C_{\mathcal{F},\infty}^{L/2}\sqrt{\frac{z_{K,M}+z}{x_{K,M}^2 n}} + 2C_{\mathcal{F},\infty}^{L}\frac{z_{K,M}+z}{x_{K,M}^2 n}\right].$$

Let $C' = C^*\max(C,1)C_{\mathcal{F},\infty}^L$. One gets

$$Z_{K,M}(s) \leqslant 4C'\left[\frac{\sigma_{K,M}}{x_{K,M}} + \sqrt{\frac{z_{K,M}+z}{x_{K,M}^2 n}} + \frac{z_{K,M}+z}{x_{K,M}^2 n}\right].$$

Let $x_{K,M} = \theta^{-1}\sqrt{\sigma_{K,M}^2 + \frac{z_{K,M}+z}{n}}$ with θ such that $2\theta+\theta^2 \leqslant 1/(16C')$. Then, with probability $1-e^{-z_{K,M}-z}$:

$$Z_{K,M}(s) \leqslant 4C'(\theta+\theta+\theta^2) \leqslant \frac{1}{4}$$

Now choose $z_{K,M} = M+K$, it follows that $\sum_{K\in\mathbb{N}^*,M\in\mathcal{M}} e^{-z_{K,M}} \leqslant (e-1)^{-2} \leqslant 1$ and the first point of the lemma is proved.

Moreover, one has with probability $1-e^{-z}$, for all K, M:

$$\begin{aligned} Z_{K,M}(s)x_{K,M}^2 &\leqslant 4C'\left[\sigma_{K,M}x_{K,M} + x_{K,M}\sqrt{\frac{z_{K,M}+z}{n}} + \frac{z_{K,M}+z}{n}\right] \\ &\leqslant 4C'\left[2\theta x_{K,M}^2 + \frac{z_{K,M}+z}{n}\right] \\ &= 4C'\left[2\theta^{-1}\sigma_{K,M}^2 + (2\theta^{-1}+1)\frac{M+K}{n} + (2\theta^{-1}+1)\frac{z}{n}\right] \end{aligned}$$

Let $A = 4C'(2\theta^{-1}+1)$. We get that with probability $1-e^{-z}$, for all K, M:

$$Z_{K,M}(s)x_{K,M}^2 \leqslant A\left[\sigma_{K,M}^2 + \frac{M+K}{n} + \frac{z}{n}\right]$$

Therefore the lemma holds as soon as

$$\forall K \leqslant n,\ \forall M \leqslant n, \quad \widetilde{\mathrm{pen}}(n,M,K) \geqslant A\left[\sigma_{K,M}^2 + \frac{M+K}{n}\right] \tag{2.15}$$

Lemma 2.27. *There exists constants C_1 and n_1 such that for all $n \geqslant n_1$:*

$$\sigma_{K,M} \leqslant C_1\sqrt{\frac{MK+K^2-1}{n}}(1+\sqrt{\log(n)})$$

Proof. Let $x(C) = C\sqrt{\frac{MK+K^2-1}{n}}(1+\sqrt{\log(n)})$.

$\sigma_{K,M}$ is defined by the equation $\frac{\varphi(x)}{x^2\sqrt{n}} = 1$. The function $x \mapsto \frac{\varphi(x)}{x^2}$ is nondecreasing, so it is enough to show that $\frac{\varphi(x(C))}{x(C)^2\sqrt{n}} \leqslant 1$ for some constant C that we can assume to be greater than 1.

It is easy to check that there exists a constant n_1 such that for all $n \geqslant n_1$, $\frac{\varphi(n^{6L})}{(n^{6L})^2\sqrt{n}} \leqslant 1$, so that $\sigma_{K,M} \leqslant n^{6L}$, which makes it possible to assume $x(C) \leqslant n^{6L}$. Then

$$\begin{aligned} \frac{\varphi(x(C))}{x(C)^2\sqrt{n}} &= \frac{C_0}{C}\frac{1+\sqrt{\log\left(\frac{n^{6L+1/2}}{C\sqrt{(MK+K^2-1)}\left(1+\sqrt{\log(n)}\right)}\right)}}{1+\sqrt{\log(n)}} \\ &\leqslant \frac{C_0}{C}\frac{1+\sqrt{\log(n^{7L})}}{1+\sqrt{\log(n)}} \\ &= \frac{C_0}{C}\frac{1+\sqrt{7L}\sqrt{\log(n)}}{1+\sqrt{\log(n)}} \end{aligned}$$

and by taking $C_1 = \max(C_0\sqrt{7L}, 1)$, one gets that

$$\frac{\varphi(x(C_1))}{x(C_1)^2\sqrt{n}} \leqslant 1$$

which means that $\sigma_{K,M} \leqslant x(C_1)$. □

The condition of equation (2.15) becomes

$$\widetilde{\mathrm{pen}}(n,M,K) \geqslant A\left[\frac{C_1^2(MK+K^2-1)(1+\sqrt{\log(n)})^2 + M + K}{n}\right]$$

which is implied by

$$\widetilde{\mathrm{pen}}(n,M,K) \geqslant \rho(MK+K^2-1)\frac{\log(n)}{n}$$

for some constant ρ depending only on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$, $\mathbf{Q}^*$ and L. This concludes the proof.

2.D Miscellaneous

2.D.1 Proof of Proposition 2.6

Let $m \geqslant 3$. Note $r = \frac{m}{m-1}$ and $K_0 = (m-1)^m$. One can check that $K = K_0 r^m \geqslant 2K_0$ and $K_0 r^k \in \mathbb{N}^*$ for all $k \in \{0, \dots, m\}$.

Denote by $n(K)$ the integer n_1 in the hypothesis **[Hpen]**$(0, \rho)$ corresponding to $K^* = K$. Then for all $n \geqslant \sup_{k\in\{0,\dots,m-1\}} n(K_0 r^k)$, for all M and for all $k \in \{1, \dots, m\}$,

$$\mathrm{pen}(n, M, K_0 r^k) \;-\; \mathrm{pen}(n, M, K_0 r^{k-1}) \;\geqslant\; \rho(MK_0 r^k \;+\; K_0^2 (r^2)^k \;-\; 1)\frac{\log(n)}{n}.$$

Taking the sum over $k \in \{1, \dots, m\}$, one gets that

$$\begin{aligned}\mathrm{pen}(n, M, K) &\geqslant \rho\left(M\frac{r}{r-1}(K - K_0) + \frac{r^2}{r^2-1}(K^2 - K_0^2) - m\right)\frac{\log(n)}{n} \\ &\geqslant \rho\left(M\frac{r}{r-1}(K - K_0) + \frac{r^2}{r^2-1}(K^2 - 2K_0^2)\right)\frac{\log(n)}{n}\end{aligned}$$

since $m \leqslant K_0^2 = (m-1)^{2m}$. Using that $K \geqslant 2K_0$,

$$\mathrm{pen}(n, M, K) \geqslant \frac{\rho}{2}\left(\frac{r}{r-1}MK + \frac{r^2}{r^2-1}K^2\right)\frac{\log(n)}{n}.$$

Let $v_m = \frac{\rho}{2}\min\left(\frac{r}{r-1}, \frac{r^2}{r^2-1}\right)$. One gets

$$\mathrm{pen}(n, M, K) \geqslant v_m(MK + K^2)\frac{\log(n)}{n}.$$

Therefore, there exists a non-decreasing sequence $(u_n)_{n\geqslant 1}$ such that

$$\begin{cases}\forall n,\ \forall M \leqslant n,\ \forall K \leqslant n,\ \mathrm{pen}(n, M, K) \geqslant u_n(MK + K^2 - 1)\frac{\log(n)}{n} \\ \forall m,\ u_{\max(m, \sup_{k\in\{0,\dots,m-1\}} n(K_0 r^k))} \geqslant v_m\end{cases}.$$

and since $v_m \longrightarrow \infty$, we get that $u_n \longrightarrow \infty$, which concludes the proof.

We could for instance take

$$u_n = \max\left(0,\ \sup\left\{v_i \;\middle|\; i \leqslant n \text{ s.t. } \sup_{k\in\{0,\dots,i-1\}} n(i^k(i-1)^{i-k}) \leqslant n\right\}\right).$$

2.D.2 Auxiliary lemmas

Lemma 2.28.

$$\forall t \in \bigcup_K S_K, \quad \begin{cases}\|t\|_\infty \leqslant C^L_{\mathcal{F},\infty} \\ \|t\|_2 \leqslant C^L_{\mathcal{F},2} \\ \mathbb{E}[t^2] \leqslant C^L_{\mathcal{F},\infty}\|t\|_2^2\end{cases}$$

Proof. t can be written as $t = g^{\pi,\mathbf{Q},\mathbf{f}}$ with π a probability K-uple, $\mathbf{Q}$ a transition matrix of size K and $\mathbf{f} \in \mathcal{F}^K$ for some $K \geqslant 1$.

The first point follows from

$$\begin{aligned}
\|t\|_\infty =& \left\| \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \bigotimes_{i=1}^{L} f_{k_i} \right\|_\infty \\
\leqslant & \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \left\| \bigotimes_{i=1}^{L} f_{k_i} \right\|_\infty \\
\leqslant & \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \prod_{i=1}^{L} \|f_{k_i}\|_\infty \\
\leqslant & C_{\mathcal{F},\infty}^{L} \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \\
= & C_{\mathcal{F},\infty}^{L}
\end{aligned}$$

For the second point, we use the Cauchy-Schwarz inequality:

$$\begin{aligned}
\|t\|_2^2 =& \int \left(\sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \prod_{i=1}^{L} f_{k_i}(y_i) \right)^2 \mathrm{d}\mu(y_1)\dots\mathrm{d}\mu(y_L) \\
=& \int \left(\sum_{k_1,\dots,k_L=1}^{K} \sqrt{\pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i)} \right. \\
& \left. \left(\sqrt{\pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i)} \prod_{i=1}^{L} f_{k_i}(y_i) \right) \right)^2 \mathrm{d}\mu(y_1)\dots\mathrm{d}\mu(y_L) \\
\leqslant & \int \left(\sum_{k'_1,\dots,k'_L=1}^{K} \pi(k'_1) \prod_{i=2}^{L} \mathbf{Q}(k'_{i-1},k'_i) \right) \\
& \left(\sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \prod_{i=1}^{L} f_{k_i}^2(y_i) \right) \mathrm{d}\mu(y_1)\dots\mathrm{d}\mu(y_L) \\
=& \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \int \prod_{i=1}^{L} f_{k_i}^2(y_i) \mathrm{d}\mu(y_1)\dots\mathrm{d}\mu(y_L) \\
=& \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) \prod_{i=1}^{L} \|f_{k_i}\|_2^2 \\
\leqslant & \sum_{k_1,\dots,k_L=1}^{K} \pi(k_1) \prod_{i=2}^{L} \mathbf{Q}(k_{i-1},k_i) C_{\mathcal{F},2}^{2L} \\
=& C_{\mathcal{F},2}^{2L}
\end{aligned}$$

The last point comes from

$$\mathbb{E}[t^2] = \int g^* t^2 \mathrm{d}\mu^{\otimes L}$$

$$
\begin{aligned}
&\leqslant \int \|g^*\|_\infty t^2 \mathrm{d}\mu^{\otimes L} \\
&\leqslant C_{\mathcal{F},\infty}^L \int t^2 \mathrm{d}\mu^{\otimes L} \quad \text{par le premier point} \\
&= C_{\mathcal{F},\infty}^L \|t\|_2^2
\end{aligned}
$$

□

Lemma 2.29. *Let $A, B \in \mathbb{R}_+^*$. Let $H : x \in \mathbb{R}_+^* \mapsto A \log \max(\frac{B}{x}, 1)$, and $\varphi(x) : x \in \mathbb{R}_+^* \mapsto x\sqrt{\pi A}(1 + \sqrt{\log \max(\frac{B}{x}, 1)})$. Then:*

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

Proof. The first point is straightforward.

For the second point, we have two cases.

Case 1: $x \leqslant B$. Then $H(x) = \log(\frac{B}{x})$. Therefore, we can use that $\int_0^\sigma \sqrt{\log(\frac{B}{x})} dx \leqslant \sigma(\sqrt{\pi} + \sqrt{\log(\frac{B}{\sigma})})$, which is enough to conclude.

Case 2: $x \geqslant B$. Then $H(x) = 0$ and $\varphi(x) = x\sqrt{\pi A} \geqslant B\sqrt{\pi A} = \varphi(B)$. Thus,

$$
\begin{aligned}
\int_0^x \sqrt{H(u)} du &= \int_0^B \sqrt{H(u)} du \\
&\leqslant \varphi(B) \\
&\leqslant \varphi(x)
\end{aligned}
$$

□

CHAPTER

3

STATE-BY-STATE MINIMAX ADAPTIVE DENSITY ESTIMATION

This chapter is based on the article Lehéricy (2018a).

In this chapter, we introduce a new estimator for the emission densities of a nonparametric hidden Markov model. It is adaptive and minimax with respect to each state's regularity–as opposed to globally minimax estimators, which adapt to the worst regularity among the emission densities. Our method is based on Goldenshluger and Lepski's methodology. It is computationally efficient and only requires a family of preliminary estimators, without any restriction on the type of estimators considered. We present two such estimators that allow to reach minimax rates up to a logarithmic term: a spectral estimator and a least squares estimator. We show how to calibrate it in practice and assess its performance on simulations and on real data.

3.1 Introduction

Finite state space hidden Markov models, or HMMs in short, are powerful tools for studying discrete time series and have been used in a variety of applications such as economics, signal processing and image analysis, genomics, ecology, speech recognition and ecology among others. The core idea is that the behaviour of the observations depends on a hidden variable that evolves like a Markov chain.

Formally, a hidden Markov model is a process $(X_j, Y_j)_{j\geqslant 1}$ in which $(X_j)_j$ is a Markov chain on $\mathcal{X}$, the Y_i's are independent conditionally on $(X_j)_j$ and the conditional distribution of Y_i given $(X_j)_j$ depends only on X_i. The parameters of the HMM are the parameters of the Markov chain, that is its initial distribution and transition matrix, and the parameters of the observations, that is the *emission distributions* $(\nu_k^*)_{k\in\mathcal{X}}$ where ν_k^* is the distribution of Y_j conditionally to $X_j = k$. Only the observations $(Y_j)_j$ are available.

In this chapter, we focus on estimating the emission distributions in a nonparametric setting. More specifically, assume that the emission distributions have a density with respect to some

known dominating measure μ, and write f_k^* their densities–which we call the *emission densities*. The goal of this chapter is to estimate all f_k^*'s with their minimax rate of convergence when the emission densities are not restricted to belong to a set of densities described by finitely many parameters.

3.1.1 Nonparametric state-by-state adaptivity

Theoretical results in the nonparametric setting have only been developed recently. De Castro et al. (2017) and Bonhomme et al. (2016b) introduce spectral methods, and the latter is proved to be minimax but not adaptive–which means one needs to know the regularity of the densities beforehand to reach the minimax rate of convergence. de Castro et al. (2016) introduce a least squares estimator which is shown to be minimax adaptive up to a logarithmic term. However, all these papers have a common drawback: they study the emission densities as a whole and can not handle them separately. This comes from their error criterion, which is the supremum of the errors on all densities: what they actually prove is that $\max_{k\in\mathcal{X}} \|\hat{f}_k - f_k^*\|_2$ converges with minimax rate when $(\hat{f}_k)_k$ are their density estimators. In general, the regularity of each emission density could be different, leading to different rates of convergence. This means that having just one emission density that is very hard to estimate is enough to deteriorate the rate of convergence of all emission densities.

In this chapter, we construct an estimator that is adaptive and estimates each emission density with its own minimax rate of convergence. We call this property state-by-state adaptivity. Our method does so by handling each emission density individually in a way that is theoretically justified–reaching minimax and adaptive rates of convergence with respect to the regularity of the emission densities–and computationally efficient thanks to its low computational and sample complexity.

Our approach for estimating the densities nonparametrically is model selection. The core idea is to approximate the target density using a family of parametric models that is dense within the nonparametric class of densities. For a square integrable density f^*, we consider its projection f_M^* on a finite-dimensional space $\mathfrak{P}_M$ (the parametric model), where M is a model index. This projection introduces an error, the *bias*, which is the distance $\|f^* - f_M^*\|_2$ between the target quantity and the model. The larger the model, the smaller the bias. On the other hand, larger models will make the estimation harder, resulting in a larger *variance* $\|\hat{f}_M - f_M^*\|_2^2$. The key step of model selection is to select a model with a small total error–or alternatively, a good *bias-variance tradeoff*.

In many situations, it is possible to reach the minimax rate of convergence with a good bias-variance tradeoff. Previous estimators of the emission densities of a HMM perform such a tradeoff based on an error that takes the transition matrix and all emission densities into account. Such an error leads to a rate of convergence that corresponds to the slowest minimax rate among the different parameters. In contrast, our method performs a bias-variance tradeoff for each emission density using an error term that depends only on the density in question, which makes it possible to reach the minimax rates for each density.

3.1.2 Plug-in procedure

The method we propose is based on the method developed in the seminal papers of Goldenshluger and Lepski (2011, 2014) for density estimation, extended by Goldenshluger and Lepski (2013) to the white noise and regression models. It takes a family of estimators as input and chooses the estimator that performs a good bias-variance tradeoff separately for each hidden state. We recommend the article of Lacour et al. (2017) for an insightful presentation of this method in the case of conditional density estimation.

Our method and assumptions are detailed in Section 3.2. Let us give a quick overview of the method. Assume the densities belong to a Hilbert space $\mathcal{H}$. Given a family of subsets of finite-dimensional subspaces of $\mathcal{H}$ (the models) indexed by M and estimators $\hat{f}_k^{(M)}$ of the emission densities for each hidden state k and each model M, one computes a substitute for the bias of the estimators by

$$A_k(M) = \sup_{M'} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M \wedge M')} \right\|_2 - \sigma(M') \right\}.$$

for some penalty σ. Then, for each state k, one selects the estimator $\hat{M}_k$ from the model M minimizing the quantity $A_k(M) + 2\sigma(M)$. The penalty σ can also be interpreted as a variance bound, so that this penalization procedure can be seen as performing a bias-variance tradeoff. The novelty of this method is that it selects a different $\hat{M}_k$, that is a different model, for each hidden state: this is where the state-by-state adaptivity comes from. Also note that contrary to Goldenshluger and Lepski (2013), we do not make any assumption on how the estimators are computed, provided a variance bound holds.

The main theoretical result is an oracle inequality on the selected estimators $\hat{f}_k^{(\hat{M}_k)}$, see Theorem 3.2. As a consequence, we are able to get a rate of convergence that is different for each state. These rates of convergence will even be adaptive minimax up to a logarithmic factor when the method is applied to our two families of estimators: spectral estimators and least squares estimators. To the best of our knowledge, this is the first state-by-state adaptive algorithm for hidden Markov models.

Note that finding the right penalty term σ is essential in order to obtain minimax rates of convergence. This requires a fine theoretical control of the variance of the auxiliary estimators, in the form of assumption **[H**(ϵ)**]** (see Section 3.2.1). To the best of our knowledge, there is no suitable result in the literature. This is the second theoretical contribution of this chapter: we control two families of estimators in a way that makes it possible to reach adaptive minimax rate with our state-by-state selection method, up to a logarithmic term.

On the practical side, we run this method and several variants on data simulated from a HMM with three hidden states and one irregular density, as illustrated in Section 3.4. The simulations confirm that it converges with a different rate for each emission density, and that the irregular density does not alter the rate of convergence of the other ones, which is exactly what we wanted to achieve.

Better still, the added computation time is negligible compared to the computation time of the estimators: even for the spectral estimators of Section 3.3.2 (which can be computed much faster than the least squares estimators and the maximum likelihood estimators using EM), computing the estimators on 200 models for 50,000 observations (the lower bound of our sample sizes) of a 3-states HMM requires a few minutes, compared to a couple of seconds for the state-by-state selection step. The difference becomes even larger for more observations, since the complexity of the state-by-state selection step is independent of the sample size: for instance, computing the spectral estimators on 300 models for 2,200,000 observations requires a bit less than two hours, and a bit more than ten hours for 10,000,000 observations, compared to less than ten seconds for the selection step in both cases. We refer to Section 3.4.6 for a more detailled discussion about the algorithmic complexity of the algorithms.

3.1.3 Families of estimators

We use two methods to construct families of estimators and apply the selection algorithm. The motivation and key result of this part of the chapter is to control the variances of the estimators by the right penalty σ. This part is crucial if one wants to get adaptive minimax rates, and has not been adressed in previous papers. For both methods, we develop new theoretical results

that allow to obtain a penalty σ that leads to adaptive minimax rates of convergence up to a logarithmic term. We present the algorithms and theoretical guarantees in Section 3.3.

The first method is a spectral method and is detailed in Section 3.3.2. Several spectral algorithms were developed, see for instance Anandkumar et al. (2012) and Hsu et al. (2012) in the parametric setting, and Bonhomme et al. (2016b) and De Castro et al. (2017) in a nonparametric framework. The main advantages of spectral methods are their computational efficiency and the fact that they do not resort to optimization procedure such as the EM and more generally nonconvex optimization algorithm, thus avoiding the well-documented issue of getting stuck into local sub-optimal minima.

Our spectral algorithm is based on the one studied in De Castro et al. (2017). However, their estimator cannot reach the minimax rate of convergence: the variance bound $\sigma(M)$ deduced from their results is proportional to M^3, while reaching the minimax rate requires $\sigma(M)$ to be proportional to M. To solve this issue, we introduce a modified version of their algorithm and show that it has the right variance bound, so that it is able to reach the adaptive minimax rate after our state-by-state selection procedure, up to a logarithmic term. Our algorithm also has an improved complexity: it is at most quasi-linear in the number of observations and in the model dimension, instead of cubic in the model dimension for the original algorithm.

The second method is a least squares method and is detailed in Section 3.3.3. Nonparametric least squares methods were first introduced by de Castro et al. (2016) to estimate the emission densities and extended by Lehéricy (2019) to estimate all parameters at once. They rely on estimating the density of three consecutive observations of the HMM using a least squares criterion. Since the model is identifiable from the distribution of three consecutive observations when the emission distributions are linearly independent, it is possible to recover the parameters from this density. In practice, these methods are more accurate than the spectral methods and are more stable when the models are close to not satisfying the identifiability condition, see for instance de Castro et al. (2016) for the accuracy and Lehéricy (2019) for the stability. However, since they rely on the minimization of a nonconvex criterion, the computation times of the corresponding algorithms are often longer than the ones from spectral methods.

A key step in proving theoretical guarantees for least squares methods is to relate the error on the density of three consecutive observations to the error on the HMM parameters in order to obtain an oracle inequality on the parameters from the oracle inequality on the density of three observations. More precisely, the difficult part is to lower bound the error on the density by the error on the parameters. Let us write g and g' the densities of the first three observations of a HMM with parameters θ and θ' respectively (these parameters actually correspond to the transition matrix and the emission densities of the HMM). Then one would like to get

$$\|g - g'\|_2 \geqslant \mathcal{C}(\theta)\, d(\theta, \theta')$$

where d is the natural $\mathbf{L}^2$ distance on the parameters and $\mathcal{C}(\theta)$ is a positive constant which does not depend on θ'. Such inequalities are then used to lower bound the variance of the estimator of the density of three observations g^* by the variance of the parameter estimators: let g be the projection of g^* and g' be the estimator of g^* on the current approximation space (with index M). Denote θ^*_M and $\hat{\theta}_M$ the corresponding parameters and assume that the error $\|g - g'\|_2$ is bounded by some constant $\sigma'(M)$, then the result will be that

$$d(\hat{\theta}_M, \theta^*_M) \leqslant \frac{\sigma'(M)}{\mathcal{C}(\theta^*_M)}.$$

Such a result is crucial to control the variance of the estimators by a penalty term σ, which is the result we need for the state-by-state selection method. In the case where only the emission densities vary, de Castro et al. (2016) proved that such an inequality always holds for HMMs

with 2 hidden states using brute-force computations, but it is still unknown whether it is always true for larger number of states. When the number of states is larger than 2, they show that this inequality holds under a generic assumption. Lehéricy (2019) extended this result to the case where all parameters may vary. However, the constants deduced from both articles are not explicit, and their regularity (when seen as a function of θ) is unknown, which makes it impossible to use in our setting: one needs the constants $\mathcal{C}(\theta^*_M)$ to be lower bounded by the same positive constant, which requires some sort of regularity on the function $\theta \longmapsto \mathcal{C}(\theta)$ in the neighborhood of the true parameters.

To solve this problem, we develop a finer control of the behaviour of the difference $\|g - g'\|_2$, which is summarized in Theorem 3.7. We show that it is possible to assume $\mathcal{C}$ to be lower semicontinuous and positive without any additional assumption. In addition, we give an explicit formula for the constant when θ' and θ are close, which gives an explicit bound for the asymptotical rate of convergence. This result allows us to control the variance of the least squares estimators by a penalty σ which ensures that the state-by-state method reaches the adaptive minimax rate up to a logarithmic term.

3.1.4 Numerical validation and application to real data sets

Section 3.4 shows how to apply the state-by-state selection method in practice and shows its performance on simulated data and a comparison with a method based on cross validation that does note estimate state by state.

Note that the theoretical results give a penalty term σ known only up to a multiplicative constant which is unknown in practice. This problem, the *penalty calibration* issue, is usual in model selection methods. It can be solved using algorithms such as the dimension jump heuristics, see for instance Birgé and Massart (2007), who introduce this heuristics and prove that it leads to an optimal penalization in the special case of Gaussian model selection framework. This method has been shown to behave well in practice in a variety of domains, see for instance Baudry et al. (2012). We describe the method and show how to use this heuristics to calibrate the penalties in Section 3.4.2.

We propose and compare several variants of our algorithm. Section 3.4.2 shows some variants in the calibration of the penalties and Section 3.4.3 shows other ways to select the final estimator. We discuss the result of the simulations and the convergence of the selected estimators in Section 3.4.4.

In Section 3.4.5, we compare our method with a non state-by-state adaptive method based on cross validation. Finally, we discuss the complexities of the auxiliary estimation methods and of our selection procedures in Section 3.4.6.

In Section 3.5, we apply our algorithm to two sets of GPS tracks. The first data set contains trajectories of artisanal fishers from Madagascar, recorded using a regular sampling with 30 seconds time steps. The second data set contains GPS positions of Peruvian seabird, recorded with 1 second time steps. We convert these tracks into the average velocity during each time step and apply our method using spectral estimators as input. The observed behaviour confirms the ability of our method to adapt to the different regularities by selecting different dimensions for each emission density.

Section 3.6 contains a conclusion and perspectives for this work.

Finally, Appendix A contains the details of our spectral algorithm and Appendix 3.B is dedicated to the proofs.

3.1.5 Notations

Throughout the chapter, for all vectors of functions $\mathbf{f} = (f_1, \dots, f_K) \in \mathbf{L}^2(\mathcal{Y}, \mu)^K$, we write $G(\mathbf{f})$ the Gram matrix of $\mathbf{f}$, defined by $G(\mathbf{f})_{i,j} = \langle f_i, f_j \rangle$ for all $i, j \in [K]$.

3.2 The state-by-state selection procedure

In this section, we introduce the framework and our state-by-state selection method.

In Section 3.2.1, we introduce the notations and assumptions. In Section 3.2.2, we present our selection method and prove that it satisfies an oracle inequality.

3.2.1 Framework and assumptions

Let $(X_j)_{j \geqslant 1}$ be a Markov chain with finite state space $\mathcal{X}$ of size K. Let $\mathbf{Q}^*$ be its transition matrix and π^* be its initial distribution. Let $(Y_j)_{j \geqslant 1}$ be random variables on a measured space $(\mathcal{Y}, \mu)$ with μ σ-finite such that conditionally on $(X_j)_{j \geqslant 1}$ the Y_j's are independent with a distribution depending only on X_j. Let ν_k^* be the distribution of Y_j conditionally to $\{X_j = k\}$. Assume that ν_k^* has density f_k^* with respect to μ. We call $(\nu_k^*)_{k \in \mathcal{X}}$ the *emission distributions* and $\mathbf{f}^* = (f_k^*)_{k \in \mathcal{X}}$ the *emission densities*. Then $(X_j, Y_j)_{j \geqslant 1}$ is a hidden Markov model with parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. The hidden chain $(X_j)_{j \geqslant 1}$ is assumed to be unobserved, so that the estimators are based only on the observations $(Y_j)_{j \geqslant 1}$.

Let $(\mathfrak{P}_M)_{M \in \mathbb{N}}$ be a nested family of finite-dimensional subspaces such that their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mu)$. The spaces $(\mathfrak{P}_M)_{M \in \mathbb{N}}$ are our models; in the following we abusively call M the model instead of $\mathfrak{P}_M$. For each index $M \in \mathbb{N}$, we write $\mathbf{f}^{*,(M)} = (f_k^{*,(M)})_{k \in \mathcal{X}}$ the projection of $\mathbf{f}^*$ on $(\mathfrak{P}_M)^K$. It is the best approximation of the true densities within the model M.

In order to estimate the emission densities, we do not need to use every models. Typically there is no point in taking models with more dimensions than the sample size, since they will likely be overfitting. Let $\mathcal{M}_n \subset \mathbb{N}$ be the set of indices which will be used for the estimation from n observations. For each $M \in \mathcal{M}_n$, we assume we are given an estimator $\hat{\mathbf{f}}_n^{(M)} = (\hat{f}_{n,k}^{(M)})_{k \in \mathcal{X}} \in (\mathfrak{P}_M)^K$. We will need to assume that for all models, the variance–that is the distance between $\hat{\mathbf{f}}_n^{(M)}$ and $\mathbf{f}^{*,(M)}$–is small with high probability. In the following, we drop the dependency in n and simply write $\mathcal{M}$ and $\hat{\mathbf{f}}^{(M)}$.

The following result is what one usually obtains in model selection. It bounds the distance between the estimators $\hat{\mathbf{f}}^{(M)}$ and the projections $\mathbf{f}^{*,(M)}$ by some penalty function σ. Thus, $\sigma/2$ can be seen as a bound of the variance term.

[H(ϵ)] With probability $1 - \epsilon$,

$$\forall M \in \mathcal{M}, \quad \inf_{\tau_{n,M} \in \mathfrak{S}(K)} \max_{k \in \mathcal{X}} \left\| \hat{f}_k^{(M)} - f_{\tau_{n,M}(k)}^{*,(M)} \right\|_2 \leqslant \frac{\sigma(M, \epsilon, n)}{2}$$

where the upper bound $\sigma : (M, \epsilon, n) \in \mathcal{M} \times [0,1] \times \mathbb{N}^* \longmapsto \sigma(M, \epsilon, n) \in \mathbb{R}_+$ is nondecreasing in M. We show in Sections 3.3.2 and 3.3.3 how to obtain such a result for a spectral method and for a least squares method (using an algorithm from Lehéricy (2019)). In the following, we omit the parameters ϵ and n in the notations and only write $\sigma(M)$.

What is important for the selection step is that the permutation $\tau_{n,M}$ does not depend on the model M: one needs all estimators $(\hat{f}_k^{(M)})_{M \in \mathcal{M}}$ to correspond to the same emission density, namely $f_{\tau_n(k)}^*$ when $\tau_{n,M} = \tau_n$ is the same for all models M. This can be done in the following

way: let $M_0 \in \mathcal{M}$ and let

$$\hat{\tau}^{(M)} \in \underset{\tau \in \mathfrak{S}(K)}{\arg\min} \left\{ \max_{k \in \mathcal{X}} \left\| \hat{f}^{(M)}_{\tau(k)} - \hat{f}^{(M_0)}_{k} \right\|_2 \right\}$$

for all $M \in \mathcal{M}$. Then, consider the estimators obtained by swapping the hidden states by these permutations. In other words, for all $k \in \mathcal{X}$, consider

$$\hat{f}^{(M)}_{k,\text{new}} = \hat{f}^{(M)}_{\hat{\tau}^{(M)}(k)}.$$

Now, assume that the error on the estimators is small enough. More precisely, write $B_{M,M_0} = \max_{k \in \mathcal{X}} \left\| f^{*,(M)}_k - f^{*,(M_0)}_k \right\|_2$ the distance between the projections of $\mathbf{f}^*$ on the models M and M_0 and assume that $2\left[\sigma(M)/2 + \sigma(M_0)/2 + B_{M,M_0}\right]$ (that is twice the upper bound of the distance between two estimated emission densities corresponding to the same hidden states in models M and M_0) is smaller than $m(\mathbf{f}^*, M_0) := \min_{k' \neq k} \left\| f^{*,(M_0)}_k - f^{*,(M_0)}_{k'} \right\|_2$, which is the smallest distance between two different densities of the vector $\mathbf{f}^{*,(M_0)}$.

Then **[H(ϵ)]** ensures that with probability as least $1 - \epsilon$, for all k, there exists a single component of $\hat{\mathbf{f}}^{(M)}$ that is closer than $\sigma(M)/2 + \sigma(M_0)/2$ of $f^{*,(M_0)}_k$, and this component will be $\hat{f}^{(M)}_{\hat{\tau}^{(M)}(k)}$ by definition. This is summarized in the following lemma.

Lemma 3.1. *Assume **[H(ϵ)]** holds. Then with probability $1 - \epsilon$, there exists a permutation $\tau_n \in \mathfrak{S}(K)$ such that for all $k \in \mathcal{X}$ and for all $M \in \mathcal{M}$ such that*

$$\sigma(M) + \sigma(M_0) + 2B_{M,M_0} < m(\mathbf{f}^*, M_0),$$

one has

$$\max_{k \in \mathcal{X}} \left\| \hat{f}^{(M)}_{k,new} - f^{*,(M)}_{\tau_n(k)} \right\|_2 \leqslant \frac{\sigma(M)}{2}. \tag{3.1}$$

Proof. Proof in Section 3.B.1. □

Thus, this property holds asymptotically as soon as $\inf \mathcal{M}$ tends to infinity and $\sup_{M \in \mathcal{M}} \sigma(M)$ tends to zero.

3.2.2 Estimator and oracle inequality

Let us now introduce our selection procedure. This method and the following theorem are based on the approach of Goldenshluger and Lepski (2011), but do not require any assumption on the structure of the estimators, provided a variance bound such as Equation (3.1) holds.

For each $k \in \mathcal{X}$ and $M \in \mathcal{M}$, let

$$A_k(M) = \sup_{M' \in \mathcal{M}} \left\{ \left\| \hat{f}^{(M')}_k - \hat{f}^{(M \wedge M')}_k \right\|_2 - \sigma(M') \right\}.$$

$A_k(M)$ serves as a replacement for the bias of the estimator $\hat{f}^{(M)}_k$, as can be seen in Equation (3.2). This comes from the fact that for large M', the quantity $\|\hat{f}^{(M')}_k - \hat{f}^{(M)}_k\|_2$ is upper bounded by the variances $\|\hat{f}^{(M')}_k - f^{*,(M')}_k\|_2$ and $\|\hat{f}^{(M)}_k - f^{*,(M)}_k\|_2$ (which are bounded by $\sigma(M')/2$) plus the bias $\|f^{*,(M)}_k - f^*_k\|_2$. Thus, only the bias term remains after substracting the variance bound $\sigma(M')$.

Then, for all $k \in \mathcal{X}$, select a model through the bias-variance tradeoff

$$\hat{M}_k \in \underset{M \in \mathcal{M}}{\arg\min} \{A_k(M) + 2\sigma(M)\}$$

and finally take

$$\hat{f}_k = \hat{f}_k^{(\hat{M}_k)}.$$

The following theorem shows an oracle inequality on this estimator.

Theorem 3.2. *Let $\epsilon \geqslant 0$ and assume equation (3.1) holds for all $k \in \mathcal{X}$ with probability $1-\epsilon$. Then with probability $1-\epsilon$,*

$$\forall k \in \mathcal{X}, \quad \|\hat{f}_k - f^*_{\tau_n(k)}\|_2 \leqslant 4 \inf_{M \in \mathcal{M}} \left\{ \|f^{*,(M)}_{\tau_n(k)} - f^*_{\tau_n(k)}\|_2 + \sigma(M,\epsilon) \right\}.$$

Proof. We restrict ourselves to the event of probability at least $1-\epsilon$ where equation (3.1) holds for all $k \in \mathcal{X}$.

The first step consists in decomposing the total error: for all $M \in \mathcal{M}$ and $k \in \mathcal{X}$,

$$\begin{aligned}\left\| \hat{f}_k^{(\hat{M}_k)} - f^*_{\tau_n(k)} \right\|_2 \leqslant & \left\| \hat{f}_k^{(\hat{M}_k)} - \hat{f}_k^{(\hat{M}_k \wedge M)} \right\|_2 + \left\| \hat{f}_k^{(\hat{M}_k \wedge M)} - \hat{f}_k^{(M)} \right\|_2 \\ & + \left\| \hat{f}_k^{(M)} - f^{*,(M)}_{\tau_n(k)} \right\|_2 + \left\| f^{*,(M)}_{\tau_n(k)} - f^*_{\tau_n(k)} \right\|_2.\end{aligned}$$

From now on, we will omit the subscripts k and $\tau_n(k)$. Using equation (3.1) and the definition of $A(M)$ and $\hat{M}$, one gets

$$\begin{aligned}\left\| \hat{f}^{(\hat{M})} - f^* \right\|_2 \leqslant & (A(M) + \sigma(\hat{M})) + (A(\hat{M}) + \sigma(M)) \\ & + \sigma(M) + \left\| f^{*,(M)} - f^* \right\|_2 \\ \leqslant & 2A(M) + 4\sigma(M) + \left\| f^{*,(M)} - f^* \right\|_2.\end{aligned}$$

Then, notice that $A(M)$ can be bounded by

$$\begin{aligned}A(M) \leqslant & \sup_{M'} \left\{ \left\| \hat{f}^{(M')} - f^{*,(M')} \right\|_2 + \left\| \hat{f}^{(M \wedge M')} - f^{*,(M \wedge M')} \right\|_2 - \sigma(M') \right\} \\ & + \sup_{M'} \left\| f^{*,(M')} - f^{*,(M \wedge M')} \right\|_2.\end{aligned}$$

Since σ is nondecreasing, $\sigma(M \wedge M') \leqslant \sigma(M')$, so that the first term is upper bounded by zero thanks to equation (3.1). The second term can be controlled since the orthogonal projection is a contraction. This leads to

$$A(M) \leqslant \left\| f^* - f^{*,(M)} \right\|_2, \tag{3.2}$$

which is enough to conclude. □

Remark. *The oracle inequality also holds when taking*

$$A_k(M) = \sup_{M' \geqslant M} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M)} \right\|_2 - \sigma(M') \right\}_+.$$

Remark. *Note that the selected $\hat{M}_k$ implicitely depends on the probability of error ϵ through the penalty σ.*

In the asymptotic setting, we take ϵ as a function of n, so that $\hat{M}_k$ is a function of n only. This will be used to get rid of ϵ when proving that the estimators reach the minimax rates of convergence.

3.3 Plug-in estimators and theoretical guarantees

In this section, we introduce two methods to construct families of estimators of the emission densities. We show that they satisfy assumption **[H(ϵ)]** for a given variance bound σ.

In Section 3.3.1, we introduce the assumptions we will need for both methods. Section 3.3.2 is dedicated to the spectral estimator and Section 3.3.3 to the least squares estimator.

3.3.1 Framework and assumptions

Recall that we approximate $\mathbf{L}^2(\mathcal{Y}, \mu)$ by a nested family of finite-dimensional subspaces $(\mathfrak{P}_M)_{M\in\mathcal{M}}$ such that their union is dense in $\mathbf{L}^2(\mathcal{Y}, \mu)$ and write $f_k^{*,(M)}$ the orthogonal projection of f_k^* on $\mathfrak{P}_M$ for all $k \in \mathcal{X}$ and $M \in \mathcal{M}$. We assume that $\mathcal{M} \subset \mathbb{N}$ and that the space $\mathfrak{P}_M$ has dimension M. A typical way to construct such spaces is to take $\mathfrak{P}_M$ spanned by the first M vectors of an orthonormal basis.

Both methods will construct an estimator of the emission densities for each model of this family. These estimators will then be plugged in the state-by-state selection method of Section 3.2.2, which will select one model for each state of the HMM.

We will need the following assumptions.

[HX] $(X_j)_{j\geqslant 1}$ is a stationary ergodic Markov chain with parameters $(\pi^*, \mathbf{Q}^*)$;

[Hid] $\mathbf{Q}^*$ is invertible and the family $\mathbf{f}^*$ is linearly independent.

The ergodicity assumption in **[HX]** is standard in order to obtain convergence results. In this case, the initial distribution is forgotten exponentially fast, so that the HMM will essentially behave like a stationary process after a short period of time. For the sake of simplicity, we assume the Markov chain to be stationary.

[Hid] appears in identifiability results, see for instance Gassiat et al. (2015) and Theorem 3.5. It is sufficient to ensure identifiability of the HMM from the law of three consecutive observations. Note that it is in general not possible to recover the law of a HMM from two observations (see for instance Appendix G of Anandkumar et al. (2012)), so that three is actually the minimum to obtain general identifiability.

3.3.2 The spectral method

Algorithm 2 is a variant of the spectral algorithm introduced in De Castro et al. (2017). Unlike the original one, it is able to reach the minimax rate of convergence thanks to two improvements. The first one consists in decomposing the joint density on different models, hence the use of two dimensions m and M. The second one consists in trying several randomized joint diagonalizations instead of just one, and selecting the best one, hence the parameter r. These additional parameters do not actually add much to the complexity of the algorithm: in theory, the choice $m, r \approx \log(n)$ is fine (see Corollary 3.4), and in practice, any large enough constant works, see Section 3.4 for more details.

For all $M \in \mathcal{M}$, let $(\varphi_1^M, \dots, \varphi_M^M)$ be an orthonormal basis of $\mathfrak{P}_M$. Let

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

The following theorem follows the proof of Theorem 3.1 of De Castro et al. (2017), with modifications that allow to control the error of the spectral estimators in expectation and are essential to obtain the right rates of convergence in Corollary 3.4.

Algorithm 2: Spectral estimation of the emission densities of a HMM (short version)

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, two dimensions $m \leqslant M$, an orthonormal basis $(\varphi_1, \dots, \varphi_M)$ and number of retries r.

Result: Spectral estimators $(\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$.

[Step 1] Consider the following empirical estimators: for any $a, c \in [m]$ and $b \in [M]$,

- $\hat{\mathbf{M}}_{m,M,m}(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$
- $\hat{\mathbf{P}}_{m,m}(a,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_c(Y_{s+2})$.

[Step 2] Let $\hat{\mathbf{U}}_m$ be the $m \times K$ matrix of orthonormal left singular vectors of $\hat{\mathbf{P}}_{m,m}$ corresponding to its top K singular values. $\hat{\mathbf{U}}_m$ can be seen as a projection. Denote by $\mathbf{P}'$ and $\mathbf{M}'(\cdot, b, \cdot)$ the projected tensors, defined by $\mathbf{P}' = \hat{\mathbf{U}}_m^\top \hat{\mathbf{P}}_{m,m}\hat{\mathbf{U}}_m$ and likewise for $\mathbf{M}'$.

[Step 3] Form the matrices $\mathbf{B}(b) := (\mathbf{P}')^{-1}\mathbf{M}'$ for all $b \in [M]$.

[Step 4] Construct a matrix $\hat{\mathbf{O}}$ by taking the best approximate simultaneous diagonalization of all $\mathbf{B}(b)$ among r attempts: for all $b \in [M]$, $\mathbf{B}(b) \approx \mathbf{R}\mathrm{diag}[\hat{\mathbf{O}}(b,\cdot)]\mathbf{R}^{-1}$ for some matrix $\mathbf{R}$ (see details in Algorithm 4, in Appendix 3.A).

[Step 5] Define the emission densities estimators $\hat{\mathbf{f}}^{(M,r)} := (\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$ by: for all $k \in \mathcal{X}$, $\hat{f}_k^{(M,r)} := \sum_{b=1}^{M} \hat{\mathbf{O}}(b,k)\varphi_b$.

Theorem 3.3. *Assume **[HX]** and **[Hid]** hold. Then there exists a constant M_0 depending on $\mathbf{f}^*$ and constants C_σ and n_1 depending on $\mathbf{f}^*$ and $\mathbf{Q}^*$ such that for all $\epsilon \in (0,1)$, for all $m, M \in \mathcal{M}$ such that $M \geqslant m \geqslant M_0$ and for all $n \geqslant n_1\eta_3^2(m,M)(-\log\epsilon)^2$, with probability greater than $1-6\epsilon$,*

$$\inf_{\tau\in\mathfrak{S}(K)} \max_{k\in\mathcal{X}} \|\hat{f}_k^{(M,\lceil t\rceil)} - f_{\tau(k)}^{*,(M)}\|_2^2 \leqslant C_\sigma \eta_3^2(m,M)\frac{(-\log\,\epsilon)^2}{n}$$

Proof. Proof in Section 3.B.2. □

Note that the constants n_1 and C_σ depend on $\mathbf{Q}^*$ and $\mathbf{f}^*$. This dependency will not affect the rates of convergence of the estimators (with respect to the sample size n), but it can change the constants of the bounds and the minimum sample size needed to reach the asymptotic regime.

Let us now apply the state-by-state selection method to these estimators. The following corollary shows that it is possible to reach the minimax rate of convergence up to a logarithmic term separately for each state under standard assumptions. Note that we need to bound the resulting estimators by some power of n, but this assumption is not very restrictive since α can be arbitrarily large.

Corollary 3.4. *Assume **[HX]** and **[Hid]** hold. Also assume that $\eta_3^2(m,M) \leqslant C_\eta m^2 M$ for a constant $C_\eta > 0$ and that for all $k \in \mathcal{X}$, there exists s_k such that $\|f_k^{*,(M)} - f_k^*\|_2 = O(M^{-s_k})$. Then there exists a constant C_σ depending on $\mathbf{f}^*$ and $\mathbf{Q}^*$ such that the following holds.*

Let $\alpha > 0$ and $C \geqslant 2(1+2\alpha)\sqrt{C_\eta C_\sigma}$. Let $\hat{\mathbf{f}}^{sbs}$ be the estimators selected from the family $(\hat{\mathbf{f}}^{(M,\lceil(1+2\alpha)\log(n)\rceil)})_{M\leqslant M_{\max}(n)}$ with $M_{\max}(n) = n/\log(n)^5$, $m_M = \log(n)$ and $\sigma(M) = C\sqrt{\frac{M\log(n)^4}{n}}$ for all M. Then there exists a sequence of random permutations $(\tau_n)_{n\geqslant 1}$ such

that

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\|(-n^\alpha) \vee (\hat{f}^{sbs}_{\tau_n(k)} \wedge n^\alpha) - f^*_k\right\|_2^2\right] = O\left(\left(\frac{n}{\log(n)^4}\right)^{\frac{-2s_k}{2s_k+1}}\right).$$

The novelty of this result is that each emission density is estimated with its own rate of convergence: the rate $\frac{-s_k}{2s_k+1}$ is different for each emission density, even though the original spectral estimators did not handle them separately. This is due to our state-by-state selection method.

Moreover, it is able to reach the minimax rate for each density in an adaptive way. For instance, in the case of a β-Hölder density on $\mathcal{Y} = [0,1]^D$ (equipped with a trigonometric basis), one can easily check the control of η_3, and the control $\|f^{*,(M)}_k - f^*_k\|_2 = O(M^{-\beta/D})$ follows from standard approximation results, see for instance DeVore and Lorentz (1993). Thus, our estimators converge with the rate $(n/\log(n)^4)^{-2\beta/(2\beta+D)}$ to this density: this is the minimax rate up to a logarithmic factor.

Remark. *By aligning the estimators like in Section 3.2.1, one can replace the sequence of permutations in Corollary 3.4 by a single permutation, in other words there exists a random permutation τ which does not depend on n such that*

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\|(-n^\alpha) \vee (\hat{f}^{sbs}_{\tau(k)} \wedge n^\alpha) - f^*_k\right\|_2^2\right] = O\left(\left(\frac{n}{\log(n)^4}\right)^{\frac{-2s_k}{2s_k+1}}\right).$$

This means that the sequence $(\hat{f}^{sbs}_k)_{n\geqslant 1}$ is an adaptive rate-minimax estimator of f^*_k–or more precisely of one of the emission densities $(f^*_{k'})_{k'\in\mathcal{X}}$, but since the distribution of the HMM is invariant under relabelling of the hidden states, one can assume the limit to be f^*_k without loss of generality–up to a logarithmic term.

At this point, it is important to note that the choice of the constant $C \geqslant 2(1+2\alpha)\sqrt{C_\eta C_\sigma}$ depends on the hidden parameters of the HMM and as such is unknown. This penalty calibration problem is very common in the model selection framework and can be solved in practice using methods such as the slope heuristics or the dimension jump method which have been proved to be theoretically valid in several cases, see for instance Baudry et al. (2012) and references therein. We use the dimension jump method and explain its principle and implementation in Section 3.4.2.

Proof. Using Theorem 3.3, one gets that for all n and for all $M \in \mathcal{M}$ such that $n \geqslant n_1\eta_3^2(m_M, M)(1+2\alpha)^2\log(n)^2$, with probability $1 - 6n^{-1-2\alpha}$,

$$\begin{aligned}\inf_{\tau\in\mathfrak{S}(K)} \max_{k\in\mathcal{X}} \|\hat{f}^{(M,\lceil t\rceil)}_k - f^{*,(M)}_{\tau(k)}\|_2^2 \leqslant & C_\sigma \eta_3^2(m_M, M)\frac{(1+2\alpha)^2\log(n)^2}{n} \\ \leqslant & (1+\alpha)^2 C_\sigma C_\eta M \frac{\log(n)^4}{n} \\ \leqslant & \frac{\sigma(M)^2}{4}\end{aligned}$$

where $\sigma(M) = C\sqrt{\frac{M\log(n)^4}{n}}$ with C such that $C^2 \geqslant 4(1+2\alpha)^2 C_\sigma C_\eta$.

The condition on M becomes

$$n \geqslant n_1 \log(n)^4 M (1+2\alpha)^2$$

and is asymptotically true for all $M \leqslant M_{\max}(n)$ as soon as $M_{\max}(n) = o(n/\log(n)^4)$.

Thus, **[H**$(6n^{-(1+2\alpha)}$**)]** is true for the family $(\hat{\mathbf{f}}^{(M,\lceil(1+2\alpha)\log(n)\rceil)})_{M\leqslant M_{\max}(n)}$. Note that the assumption $M_{\max}(n) = o(n/\log(n)^4)$ also implies that there exists M_1 such that for n large enough, Lemma 3.1 holds for all $M \geqslant M_1$, so that Theorem 3.2 implies that for n large enough, there exists a permutation τ_n such that with probability $1 - 6n^{-(1+2\alpha)}$, for all $k \in \mathcal{X}$,

$$\begin{aligned}\|\hat{f}^{\mathrm{sbs}}_{\tau_n(k)} - f^*_k\|_2 &\leqslant 4 \inf_{M_1\leqslant M\leqslant M_{\max}} \{\|f^{*,(M)}_k - f^*_k\|_2 + \sigma(M)\} \\ &= O\left(\inf_{M_1\leqslant M\leqslant M_{\max}} \left\{M^{-s_k} + \sqrt{\frac{M\log(n)^4}{n}}\right\}\right) \\ &= O\left(\left(\frac{n}{\log(n)^4}\right)^{-s_k/(1+2s_k)}\right),\end{aligned}$$

where the tradeoff is reached for $M = (\frac{n}{\log(n)^4})^{1/(1+2s_k)}$, which is in $[M_1, M_{\max}(n)]$ for n large enough.

Finally, write A the event of probability smaller than $6n^{-(1+2\alpha)}$ where **[H**$(6n^{-(1+\alpha)}$**)]** doesn't hold, then for n large enough and for all $k \in \mathcal{X}$,

$$\begin{aligned}\mathbb{E}\left[\left\|(-n^\alpha) \vee (\hat{f}^{\mathrm{sbs}}_{\tau_n(k)} \wedge n^\alpha) - f^*_k\right\|_2^2\right] &\leqslant \mathbb{E}\left[\mathbf{1}_A \left\|\hat{f}^{\mathrm{sbs}}_{\tau_n(k)} - f^*_k\right\|_2^2\right] + \mathbb{E}\left[\mathbf{1}_{A^c}(n^{2\alpha} + \|f^*_k\|_2^2)\right] \\ &= O\left(\left(\frac{n}{\log(n)^4}\right)^{-2s_k/(1+2s_k)}\right) + O\left(\frac{n^{2\alpha} + \|f^*_k\|_2^2}{n^{1+2\alpha}}\right) \\ &= O\left(\left(\frac{n}{\log(n)^4}\right)^{-2s_k/(1+2s_k)}\right).\end{aligned}$$

□

3.3.3 The penalized least squares method

Let $\mathcal{F}$ be a subset of $\mathbf{L}^2(\mathcal{Y}, \mu)$. We will need the following assumption on $\mathcal{F}$ in order to control the deviations of the estimators:

[HF] $\mathbf{f}^* \in \mathcal{F}^{K^*}$, $\mathcal{F}$ is closed under projection on $\mathfrak{P}_M$ for all $M \in \mathcal{M}$ and

$$\forall f \in \mathcal{F}, \quad \begin{cases}\|f\|_\infty \leqslant C_{\mathcal{F},\infty} \\ \|f\|_2 \leqslant C_{\mathcal{F},2}\end{cases}$$

with $C_{\mathcal{F},\infty}$ and $C_{\mathcal{F},2}$ larger than 1.

A simple way to construct such a set $\mathcal{F}$ when μ is a finite measure is to take the sets $(\mathfrak{P}_M)_M$ spanned by the first M vectors of an orthonormal basis $(\varphi_i)_{i\geqslant 0}$ whose first vector φ_0 is proportional to $\mathbf{1}$. Then any set $\mathcal{F}$ of densities such that $\int f d\mu = 1$, $\sum_i \langle f, \varphi_i\rangle^2 \leqslant C_{\mathcal{F},2}$ and $\sum_i |\langle f, \varphi_i\rangle| \|\varphi_i\|_\infty \leqslant C_{\mathcal{F},\infty}$ for given constants $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ and for all $f \in \mathcal{F}$ satisfies **[HF]**.

When $\mathbf{Q} \in \mathbb{R}^{K\times K}$, $\pi \in \mathbb{R}^K$ and $\mathbf{f} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$, let

$$g^{\pi,\mathbf{Q},\mathbf{f}}(y_1, y_2, y_3) = \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)\mathbf{Q}(k_1, k_2)\mathbf{Q}(k_2, k_3) f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3).$$

When π is a probability distribution, $\mathbf{Q}$ a transition matrix and $\mathbf{f}$ a K-tuple of probability densities, then $g^{\pi,\mathbf{Q},\mathbf{f}}$ is the density of the first three observations of a HMM with parameters $(\pi, \mathbf{Q}, \mathbf{f})$. The motivation behind estimating $g^{\pi,\mathbf{Q},\mathbf{f}}$ is that it allows to recover the true parameters under the identifiability assumption **[Hid]**, as shown in the following theorem.

Let $\mathcal{Q}$ be the set of transition matrices on $\mathcal{X}$ and Δ the set of probability distributions on $\mathcal{X}$. For a permutation $\tau \in \mathfrak{S}(K)$, write $\mathbb{P}_\tau$ its matrix (that is the matrix defined by $\mathbb{P}_\tau(i,j) = \mathbf{1}_{\{j=\tau(i)\}}$). Finally, define the distance on the HMM parameters

$$d_{\mathrm{perm}}((\pi_1, \mathbf{Q}_1, \mathbf{f}_1), (\pi_2, \mathbf{Q}_2, \mathbf{f}_2))^2 = \inf_{\tau \in \mathfrak{S}(K)} \left\{ \|\pi_1 - \mathbb{P}_\tau \pi_2\|_2^2 + \|\mathbf{Q}_1 - \mathbb{P}_\tau \mathbf{Q}_2 \mathbb{P}_\tau^\top\|_F^2 + \sum_{k \in \mathcal{X}} \|f_{1,k} - f_{2,\tau(k)}\|_2^2 \right\}.$$

This distance is invariant under permutation of the hidden states. This corresponds to the fact that a HMM is only identifiable up to relabelling of its hidden states.

Theorem 3.5 (Identifiability)**.** *Let $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \in \Delta \times \mathcal{Q} \times (\mathbf{L}^2(\mathcal{Y}, \mu))^K$ such that $\pi_x^* > 0$ for all $x \in \mathcal{X}$ and **[Hid]** holds. Then for all $(\pi, \mathbf{Q}, \mathbf{f}) \in \Delta \times \mathcal{Q} \times (\mathbf{L}^2(\mathcal{Y}, \mu))^K$,*

$$(g^{\pi,\mathbf{Q},\mathbf{f}} = g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}) \;\Rightarrow\; d_{perm}((\pi, \mathbf{Q}, \mathbf{f}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)) = 0.$$

Proof. The spectral algorithm of De Castro et al. (2017) applied on the finite dimensional space spanned by the components of $\mathbf{f}$ and $\mathbf{f}^*$ allows to recover all the parameters even when the emission densities are not probability densities and when the Markov chain is not stationary. □

Define the empirical contrast

$$\gamma_n(t) = \|t\|_2^2 - \frac{2}{n} \sum_{j=1}^{n} t(Z_j)$$

where $Z_j := (Y_j, Y_{j+1}, Y_{j+2})$ and $(Y_j)_{1 \leqslant j \leqslant n+2}$ are the observations. It is a biased estimator of the $\mathbf{L}^2$ loss: for all $t \in (\mathbf{L}^2(\mathcal{Y}, \mu))^3$,

$$\mathbb{E}[\gamma_n(t)] = \|t - g^*\|_2^2 - \|g^*\|_2^2$$

where $g^* = g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}$. Since the bias does not depend on the function t, one can hope that the minimizers of γ_n are close to minimizers of $\|t - g^*\|_2$. We will show that this is indeed the case.

The least squares estimators of all HMM parameters are defined for each model $\mathfrak{P}_M$ by

$$(\hat{\pi}^{(M)}, \hat{\mathbf{Q}}^{(M)}, \hat{\mathbf{f}}^{(M)}) \in \operatorname*{arg\,min}_{\pi \in \Delta,\, \mathbf{Q} \in \mathcal{Q},\, \mathbf{f} \in (\mathfrak{P}_M \cap \mathcal{F})^K} \gamma_n(g^{\pi,\mathbf{Q},\mathbf{f}}).$$

The procedure is summarized in Algorithm 3. Note that with the notations of the algorithm,

$$\gamma_n(g^{\pi,\mathbf{Q},\mathbf{O}^\top \Phi}) = \|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} - \hat{\mathbf{M}}_M\|_F^2 - \|\hat{\mathbf{M}}_M\|_F^2.$$

Then, the proof of the oracle inequality of Lehéricy (2019) allows to get the following result.

Theorem 3.6. *Assume **[HF]**, **[HX]** and **[Hid]** hold.*

Then there exists constants C and n_0 depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$ and $\mathbf{Q}^*$ such that for all $n \geqslant n_0$, for all $t > 0$, with probability greater than $1 - e^{-t}$, one has for all $M \in \mathcal{M}$ such that $M \leqslant n$:

$$\|\hat{g}^{\hat{\pi}^{(M)},\hat{\mathbf{Q}}^{(M)},\hat{\mathbf{f}}^{(M)}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^{*,(M)}}\|_2^2 \leqslant C\Big(\frac{t}{n} + M\frac{\log(n)}{n}\Big).$$

Algorithm 3: Least squares estimation of the emission densities of a HMM

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, a dimension M and an orthonormal basis $\Phi = (\varphi_1, \dots, \varphi_M)$.

Result: Least squares estimators $\hat{\pi}^{(M)}$, $\hat{\mathbf{Q}}^{(M)}$ and $(\hat{f}_k^{(M)})_{k\in\mathcal{X}}$.

[Step 1] Compute the tensor $\hat{\mathbf{M}}_M$ defined by $\hat{\mathbf{M}}_M(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$ for all $a, b, c \in [M]$.

[Step 2] Find a minimizer $(\hat{\pi}^{(M)}, \hat{\mathbf{Q}}^{(M)}, \hat{\mathbf{O}})$ of $(\pi, \mathbf{Q}, \mathbf{O}) \longmapsto \|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} - \hat{\mathbf{M}}_M\|_F^2$ where

- $\pi \in \mathbb{R}^K$ is a probability distribution on $\mathcal{X}$, i.e. $\sum_{k\in\mathcal{X}} \pi_k = 1$;
- $\mathbf{Q} \in \mathbb{R}^{K\times K}$ is a transition matrix on $\mathcal{X}$, i.e. $\sum_{k'\in\mathcal{X}} Q(k,k') = 1$ for all $k \in \mathcal{X}$;
- $\mathbf{O}$ is a $M \times K$ matrix such that for all $k \in \mathcal{X}$, $\sum_{b=1}^{M} \mathbf{O}(b,k)\varphi_b \in \mathcal{F}$;
- $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})} \in \mathbb{R}^{M\times M\times M}$ is defined by $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{O})}(\cdot, b, \cdot) = \mathbf{O}\mathrm{diag}[\pi]\mathbf{Q}\mathrm{diag}[\mathbf{O}(b,\cdot)]\mathbf{Q}\mathbf{O}^\top$ for all $b \in [M]$.

[Step 3] Consider the emission densities estimators $\hat{\mathbf{f}}^{(M)} := (\hat{f}_k^{(M)})_{k\in\mathcal{X}}$ defined by for all $k \in \mathcal{X}$, $\hat{f}_k^{(M)} := \sum_{b=1}^{M} \hat{\mathbf{O}}(b,k)\varphi_b$.

In order to deduce a control of the error on the parameters–and in particular on the emission densities–from the previous result, we will need to assume that the quadratic form derived from the second-order expansion of $(\pi, \mathbf{Q}, \mathbf{f}) \in \Delta \times \mathcal{Q} \times \mathcal{F}^K \longmapsto \|g^{\pi,\mathbf{Q},\mathbf{f}} - g^*\|_2^2$ around $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ is nondegenerate.

It is still unknown whether this nondegeneracy property is true for all parameters $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ such that **[Hid]** and **[HX]** hold. de Castro et al. (2016) prove it for $K = 2$ hidden states when only the emission densities are allowed to vary by using brute-force computations. To do so, they introduce an (explicit) polynomial in the coefficients of π^*, $\mathbf{Q}^*$ and of the Gram matrix of $\mathbf{f}^*$ and prove that its value is nonzero if and only if the quadratic form is nondegenerate for the corresponding parameters. The difficult part of the proof is to show that this polynomial is always nonzero.

For the expression of this polynomial–which we will write H–in our setting, we refer to Section 3.B.3. Note that Lehéricy (2019) proves that this polynomial H is non identically zero: it is shown that there exists parameters $(\pi, \mathbf{Q}, \mathbf{f})$ satisfying **[HX]** and **[Hid]** such that $H(\pi, \mathbf{Q}, \mathbf{f}) \neq 0$, which means that the following assumption is generically satisfied:

[Hdet] $H(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \neq 0$.

The following result allows to lower bound the $\mathbf{L}^2$ error on the density of three consecutive observations by the error on the parameters of the HMM using this condition. It is an improvement of Theorem 6 of de Castro et al. (2016) and Theorem 9 of Lehéricy (2019). The main difference is that the constant $c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F})$ does not depend on the $\mathbf{f}$ around which the parameters are taken. This is crucial to obtain Corollary 3.8, from which we will deduce **[H0]**. Note that we do not need $\mathbf{f}$ to be in a compact neighborhood of $\mathbf{f}^*$. Another improvement is that the constant in the minoration only depends on the true parameters and on the set $\mathcal{F}$.

Theorem 3.7. *1. Assume that **[HF]** holds and that for all $f \in \mathcal{F}$, $\int f d\mu = 1$.*

*Then there exist a lower semicontinuous function $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \longmapsto c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F})$ that is positive when **[Hid]** and **[Hdet]** hold and a neighborhood $\mathcal{V}$ of $\mathbf{f}^*$ in $\mathcal{F}^K$ depending only*

on π^*, $\mathbf{Q}^*$, $\mathbf{f}^*$ and $\mathcal{F}$ such that for all $\mathbf{f} \in \mathcal{V}$ and for all $\pi \in \Delta$, $\mathbf{Q} \in \mathcal{Q}$ and $\mathbf{h} \in \mathcal{F}^K$,

$$\|g^{\pi,\mathbf{Q},\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2 \geqslant c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F}) d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}))^2.$$

2. *There exists a continuous function $\epsilon : (\pi^*, \mathbf{Q}^*, \mathbf{f}^*) \mapsto \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ that is positive when **[Hid]** and **[Hdet]** hold and such that for all $\pi \in \Delta$, $\mathbf{Q} \in \mathcal{Q}$ and $\mathbf{h} \in (\mathbf{L}^2(\mathcal{Y}, \mu))^K$ a K-tuple of probability densities such that $d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*)) \leqslant \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$, one has*

$$\|g^{\pi,\mathbf{Q},\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \geqslant c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) d_{perm}((\pi, \mathbf{Q}, \mathbf{h}), (\pi^*, \mathbf{Q}^*, \mathbf{f}^*))^2.$$

where

$$c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*) = \frac{(\inf_{k \in \mathcal{X}} \pi^*(k))\, \sigma_K(\mathbf{Q}^*)^4 \sigma_K(G(\mathbf{f}^*))^2}{4} \wedge \frac{H(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)}{2(1 \wedge K\|G(\mathbf{f}^*)\|_\infty)(3K^3(1 \vee \|G(\mathbf{f}^*)\|_\infty^4))^{K^2 - K/2}}.$$

Proof. Proof in Section 3.B.4. □

Corollary 3.8. *Assume **[HX]**, **[HF]**, **[Hid]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$.*

Then there exists a constant n_0 depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$ and $\mathbf{Q}^*$ and constants M_0 and C' depending on $\mathcal{F}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ such that for all $n \geqslant n_0$ and $t > 0$, with probability greater than $1 - e^{-t}$, one has for all $M \in \mathcal{M}$ such that $M_0 \leqslant M \leqslant n$:

$$\inf_{\tau \in \mathfrak{S}(K)} \max_{k \in \mathcal{X}} \|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\|_2^2 \leqslant C' \left(M \frac{\log(n)}{n} + \frac{t}{n} \right).$$

Remark. *Using the second point of Theorem 3.7, one can alternatively take n_0 and M_0 depending on $\mathcal{F}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$, and C' depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ only. For instance, one can take $C' = C/c_0(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$ with the notations of Theorems 3.6 and 3.7.*

In particular, this means that the asymptotic variance bound of the least squares estimators (and therefore the rate of convergence of the estimators selected by our state-by-state selection method) does not depend on the set $\mathcal{F}$, but only on the HMM parameters and on the bounds $C_{\mathcal{F},2}$ and $C_{\mathcal{F},\infty}$ on the square and supremum norms of the emission densities. Note that this universality result is essentially an asymptotic one since it requires n_0 to depend on $\mathcal{F}$ in a non-explicit way.

Proof. Let $\mathcal{V}$ be the neighborhood given by Theorem 3.7, then there exists M_0 such that for all $M \geqslant M_0$, $\mathbf{f}^{*,(M)} \in \mathcal{V}$. Then Theorem 3.6 and Theorem 3.7 applied to $\pi = \hat{\pi}^{(M)}$, $\mathbf{Q} = \hat{\mathbf{Q}}^{(M)}$, $\mathbf{h} = \hat{\mathbf{f}}^{(M)}$ and $\mathbf{f} = \mathbf{f}^{*,(M)}$ for all M allow to conclude. □

We may now state the following result which shows that the state-by-state selection method applied to these estimators reaches the minimax rate of convergence (up to a logarithmic factor) in an adaptive manner under generic assumptions. Its proof is the same as the one of Corollary 3.4.

Corollary 3.9. *Assume **[HX]**, **[HF]**, **[Hid]** and **[Hdet]** hold. Also assume that for all $f \in \mathcal{F}$, $\int f d\mu = 1$ and that for all k, there exists s_k such that $\|f_k^{*,(M)} - f_k^*\|_2 = O(M^{-s_k})$. Then there exists a constant C_σ depending on $C_{\mathcal{F},2}$, $C_{\mathcal{F},\infty}$, $\mathbf{Q}^*$ and $\mathbf{f}^*$ such that the following holds.*

Let $C \geqslant C_\sigma$ and let $\hat{\mathbf{f}}^{sbs}$ be the estimators selected from the family $(\hat{\mathbf{f}}^{(M)})_{M \leqslant n}$ with $\sigma(M) = C\sqrt{\frac{M\log(n)}{n}}$ for all M, aligned like in Remark 3.3.2. Then there exists a random permutation τ which does not depend on n such that

$$\forall k \in \mathcal{X}, \qquad \mathbb{E}\left[\left\|\hat{f}^{sbs}_{\tau(k)} - f^*_k\right\|_2\right] = O\left(\left(\frac{n}{\log(n)}\right)^{\frac{-s_k}{2s_k+1}}\right).$$

3.4 Numerical experiments

This section is dedicated to the discussion of the practical implementation of our method. We run the spectral estimators on simulated data for different number of observations and study the rate of convergence of the selected estimators for several variants of our method. Finally, we discuss the algorithmic complexity of the different estimators and selection methods.

In Section 3.4.1, we introduce the parameters with which we generate the observations. In Section 3.4.2, we discuss how to calibrate the constant of the penalty in practice. In Section 3.4.3, we introduce two other ways to select the final estimators, the POS and MAX variants. Section 3.4.4 contains the results of the simulations for each variant and calibration method. In Section 3.4.5, we present a cross validation procedure and compare its results with the one obtained using our method. Finally, we discuss the algorithmic complexity of the different algorithms and estimators in Section 3.4.6.

3.4.1 Setting and parameters

We take $\mathcal{Y} = [0,1]$ equipped with the Lebesgue measure. We choose the approximation spaces spanned by a trigonometric basis: $\mathfrak{P}_M := \mathrm{Span}(\varphi_1, \ldots, \varphi_M)$ with

$$\begin{cases} \varphi_1(x) & = 1 \\ \varphi_{2m}(x) & = \sqrt{2}\cos(2\pi m x) \\ \varphi_{2m+1}(x) & = \sqrt{2}\sin(2\pi m x) \end{cases}$$

for all $x \in [0,1]$ and $m \in \mathbb{N}^*$. We will consider a hidden Markov model with $K = 3$ hidden states and the following parameters:

- Transition matrix

$$\mathbf{Q}^* = \begin{pmatrix} 0.7 & 0.1 & 0.2 \\ 0.08 & 0.8 & 0.12 \\ 0.15 & 0.15 & 0.7 \end{pmatrix};$$

- Emission densities (see Figure 3.1)
  - Uniform distribution on $[0;1]$;
  - Symmetrized Beta distribution, that is a mixture with the same weight of $\frac{2}{3}X$ and $1 - \frac{1}{3}X'$ with X, X' iid following a Beta distribution with parameters $(3, 1.6)$;
  - Beta distribution with parameters $(3, 7)$.

We generate n observations and run the spectral algorithm in order to obtain estimators for the models $\mathfrak{P}_M$ with `M_min` $\leqslant M \leqslant$ `M_max`, $m = 20$ and $r = \lceil 2\log(n) + 2\log(M)\rceil$, where `M_min` $= 3$ and `M_max` $= 300$. Finally, we use the state-by-state selection method to choose the final estimator for each emission density. The main reason for using spectral estimators instead

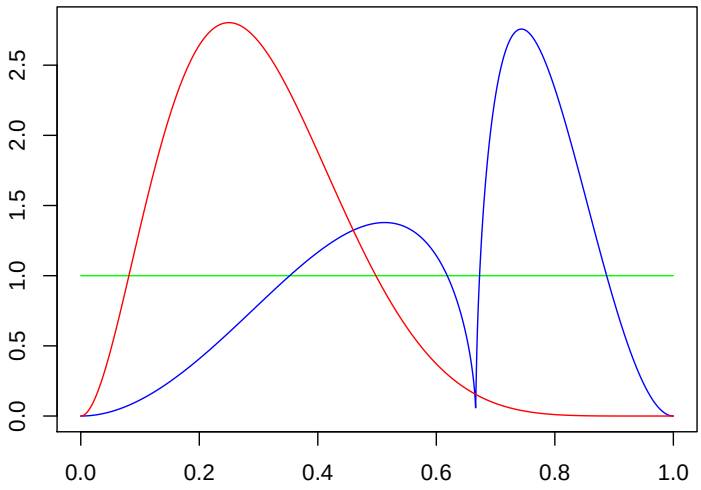


Figure 3.1: Emission densities. In all following figures, the uniform distribution corresponds to the green lines, the Beta distribution to the red lines and the symmetrized Beta distribution to the blue lines.

of maximum likelihood estimation or least squares estimation is its computational speed: it is much faster for large n than the least squares algorithm or the EM algorithm, which makes studying asymptotic behaviours possible.

We made 300 simulations, 20 per value of n, with n taking values in $\{5 \times 10^4, 7 \times 10^4, 1 \times 10^5, 1.5 \times 10^5, 2.2 \times 10^5, 3.5 \times 10^5, 5 \times 10^5, 7 \times 10^5, 1 \times 10^6, 1.5 \times 10^6, 2.2 \times 10^6, 3.5 \times 10^6, 5 \times 10^6, 7 \times 10^6, 1 \times 10^7\}$.

3.4.2 Penalty calibration

It is important to note that when considering spectral and least squares methods, the penalty σ in the state-by-state selection procedure depends on the hidden parameters of the HMM and as such is unknown in practice. This penalty calibration problem is well known and several procedures exist that allow to solve it, for instance the slope heuristics and the dimension jump method (see for instance Baudry et al. (2012) and references therein). In the following, we will use the dimension jump method to calibrate the penalty in the state-by-state selection procedure.

Consider a penalty shape $\mathrm{pen}_{\mathrm{shape}}$ and define $\hat{M}_k(\rho)$ the model selected for the hidden state k by the state-by-state selection estimator using the penalty $\rho\,\mathrm{pen}_{\mathrm{shape}}$:

$$\hat{M}_k(\rho) \in \operatorname*{arg\,min}_{M \in \mathcal{M}} \{A_k(M) + 2\rho\,\mathrm{pen}_{\mathrm{shape}}(M)\}.$$

where

$$A_k(M) = \sup_{M' \in \mathcal{M}} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M \wedge M')} \right\|_2 - \rho\,\mathrm{pen}_{\mathrm{shape}}(M') \right\}.$$

The dimension jump method relies on the heuristics that there exists a constant C such that $C\,\mathrm{pen}_{\mathrm{shape}}$ is a *minimal penalty*. This means that for all $\rho < C$, the selected models $\hat{M}_k(\rho)$ will be very large, while for $\rho > C$, the models will remain small. This translates into a sharp jump located around a value $\rho_{\mathrm{jump},k} = C$ in the plot of $\rho \longmapsto \hat{M}_k(\rho)$. The final step consists in taking

twice this value to calibrate the constant of the penalty, thus selecting the model $\hat{M}(2\rho_{\text{jump},k})$. In practice, we take $\rho_{\text{jump},k}$ as the position of the largest jump of the function $\rho \longmapsto \hat{M}_k(\rho)$.

Figure 3.2 shows the resulting dimension jumps for $n = 220,000$ observations. Each curve corresponds to one of the $\hat{M}_k(\rho)$ and has a clear dimension jump, which confirms the relevance of the heuristics. Several methods may be used to calibrate the constant of the penalty:

eachjump. Calibrate the constant independently for each state. This method has the advantage of being easy to calibrate since there is usually a single sharp jump in each state's complexity. However, our theoretical results do not suggest that the penalty constant is different for each state;

jumpmax. Calibrate the constant for all states together using only the latest jump. This consists in taking the maximum of the $\rho_{\text{jump},k}$ to select the final models. Since the penalty is known up to a multiplicative constant and taking a constant larger than needed does not affect the rates of convergence–contrary to smaller constants–this is the "safe" option;

jumpmean. Calibrate the constant for all states together using the mean of the positions of the different jumps.

We try and compare these calibration methods in Section 3.4.4.

3.4.3 Alternative selection procedures

Variant POS.

As mentionned in Section 3.2.2, it is also possible to select the estimators using the criterion

$$A_k(M) = \sup_{M' \geqslant M} \left\{ \left\| \hat{f}_k^{(M')} - \hat{f}_k^{(M)} \right\|_2 - \sigma(M') \right\}_+$$

followed by

$$\hat{M}_k \in \arg\min_{M \in \mathcal{M}} \{A_k(M) + 2\sigma(M)\}.$$

This positivity condition was in the original Goldenshluger-Lepski method. The theoretical guarantees remain the same as the previous method and both behave almost identically in practice, as shown in Section 3.4.4.

Variant MAX.

In the context of kernel density estimation, Lacour et al. (2017) show that the Goldenshluger-Lepski method still works when the biais estimate $A_k(M)$ of the model M is replaced by the distance between the estimator of the model M and the estimator with the smallest bandwidth (the analog of the largest model in our setting). They also prove an oracle inequality for this method after adding a corrective term to the penalty.

The following variant is based on the same idea. It consists in selecting the model

$$\hat{M}_k \in \arg\min_{M \in \mathcal{M}} \{\|\hat{f}_k^{(M_{\max})} - \hat{f}_k^{(M)}\|_2 + \sigma(M)\}$$

for each $k \in \mathcal{X}$ and takes

$$\hat{f}_k = \hat{f}_k^{(\hat{M}_k)},$$

where σ is the same penalty as the one in the usual state-by-state selection method.

An advantage of this algorithm is its lower complexity, since it requires $O(M_{\max})$ computations of $\mathbf{L}^2$ norms instead of $O(M_{\max}^2)$. We do not study this method theoretically in our setting.

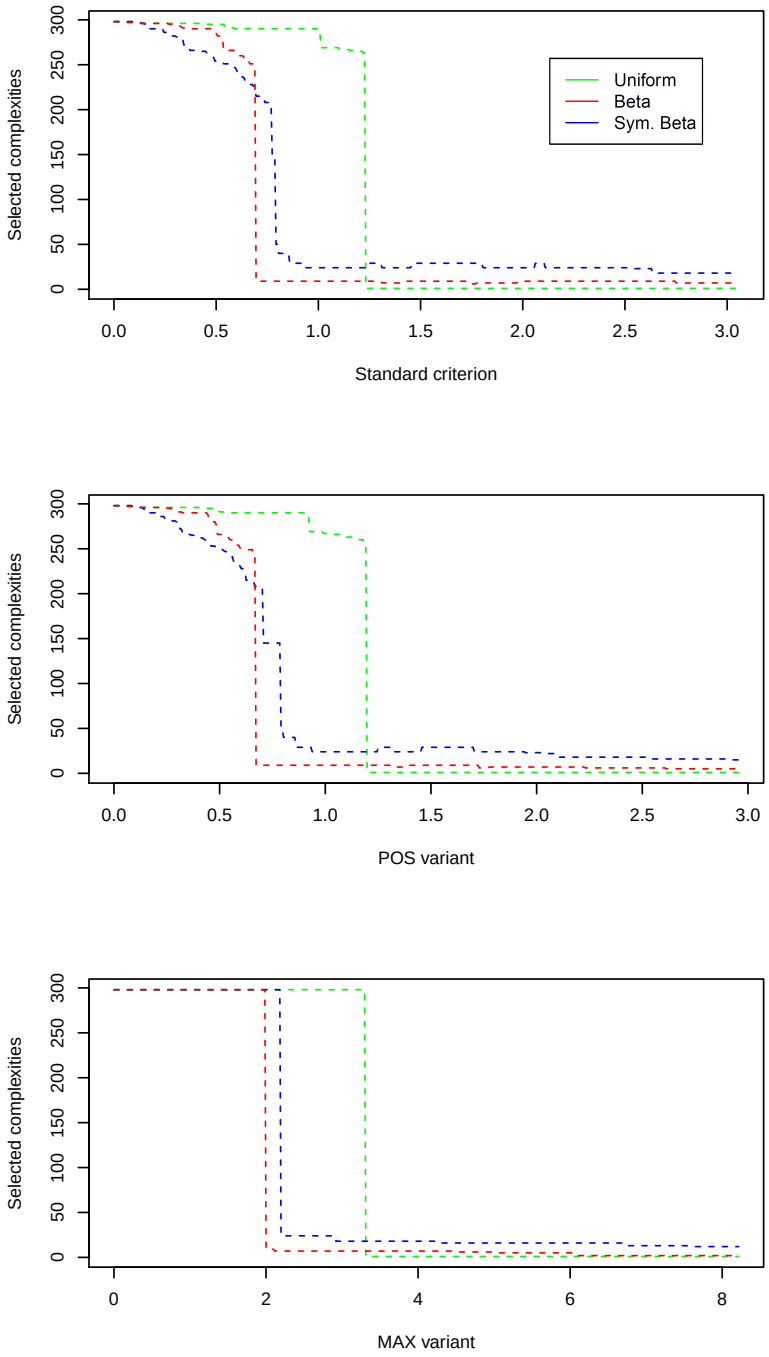


Figure 3.2: Selected complexities with respect to the penalty constant ρ for the same simulation of $n = 500,000$ observations. The colored dashed lines correspond to the single-state complexities $M_k(\rho)$.

However, the simulations (and in particular Figure 3.4) show that it behaves similarly to the standard state-by-state selection method in the asymptotic regime and even has a smaller error for small number of observations. In addition, the dimension jumps are much sharper for this method than for the usual state-by-state selection method (see Figure 3.2), which makes the

calibration heuristics easier to use.

3.4.4 Results

Figure 3.3 shows the evolution of the error $\|\hat{f}_k - f_k^*\|_2$ for each state k with respect to the number of observations n, for all penalty calibration methods and all variants of the model selection procedure. Figure 3.4 compares the evolution of the median error for the different calibration methods and for the different selection variants, and Figure 3.5 compares two estimators with the oracle estimators.

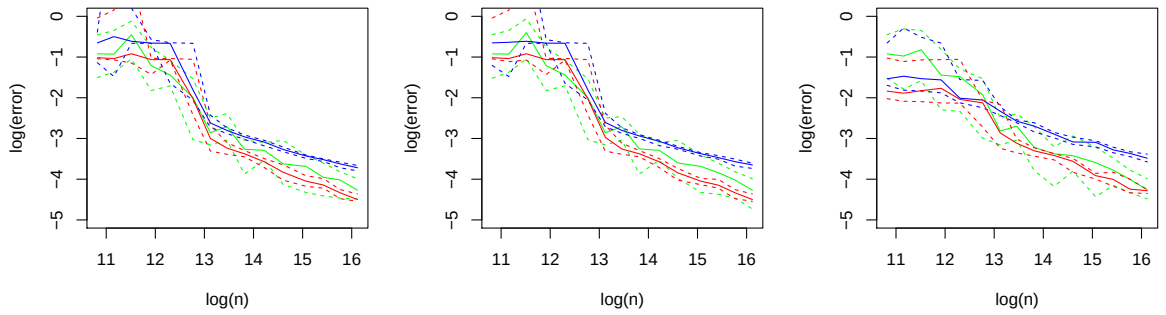
When the number of observations n is large enough, the logarithm of the error decreases linearly with respect to $\log(n)$. This corresponds to the asymptotic convergence regime: the error is expected to decrease as a power of the number of observations n when n tends to infinity. The corresponding slopes are listed in Table 3.1.

Estimator	Convergence rate exponents		
	Uniform	Sym. Beta	Beta
Eachjump	-0.500 ± 0.046	-0.347 ± 0.007	-0.470 ± 0.015
Eachjump POS	-0.503 ± 0.047	-0.327 ± 0.008	-0.469 ± 0.015
Eachjump MAX	-0.480 ± 0.052	-0.335 ± 0.009	-0.449 ± 0.015
Jumpmean	-0.532 ± 0.048	-0.349 ± 0.006	-0.471 ± 0.017
Jumpmean POS	-0.540 ± 0.048	-0.350 ± 0.006	-0.456 ± 0.016
Jumpmean MAX	-0.493 ± 0.049	-0.374 ± 0.009	-0.437 ± 0.015
Jumpmax	-0.500 ± 0.046	-0.349 ± 0.006	-0.464 ± 0.016
Jumpmax POS	-0.492 ± 0.046	-0.358 ± 0.006	-0.442 ± 0.015
Jumpmax MAX	-0.480 ± 0.052	-0.404 ± 0.009	-0.466 ± 0.015
Cross Validation	-0.434 ± 0.007	-0.263 ± 0.011	-0.377 ± 0.008
Oracle	-0.517 ± 0.048	-0.360 ± 0.006	-0.459 ± 0.017
Hidden states known	-0.526 ± 0.031	-0.293 ± 0.005	-0.428 ± 0.007
Minimax (Hölder)	-0.5	$-3/11 \approx -0.273$	$-3/7 \approx -0.429$

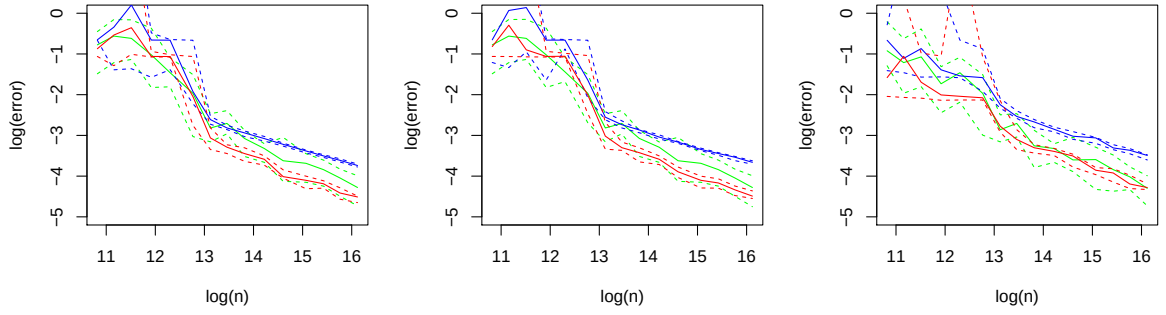
Table 3.1: Exponents of the rates of convergence for the different algorithms. The rates are obtained from a linear regression with the relation $\log(\|\hat{f}_k - f_k^*\|_2) \sim \log(n)$ for the estimators $\hat{f}_k$ computed with $n \geqslant 700,000$ observations ($n \geqslant 1,000,000$ for the cross validation estimators from Section 3.4.5). The smaller the exponent, the faster the estimators converge. The line "hidden states known" is obtained by density estimation when the hidden states are observed.

For each state, the confidence intervals of the rates of all estimators–including the oracle estimators–have a common intersection (except for the symmetrized Beta distribution in the jumpmax MAX variant, whose estimators seem to converge faster than the others). This tends to confirm that the calibration and selection variants are asymptotically equivalent. This phenomenon is also visible in Figures 3.3 and 3.4: in the asymptotic regime, the errors decrease in a similar way for all methods.

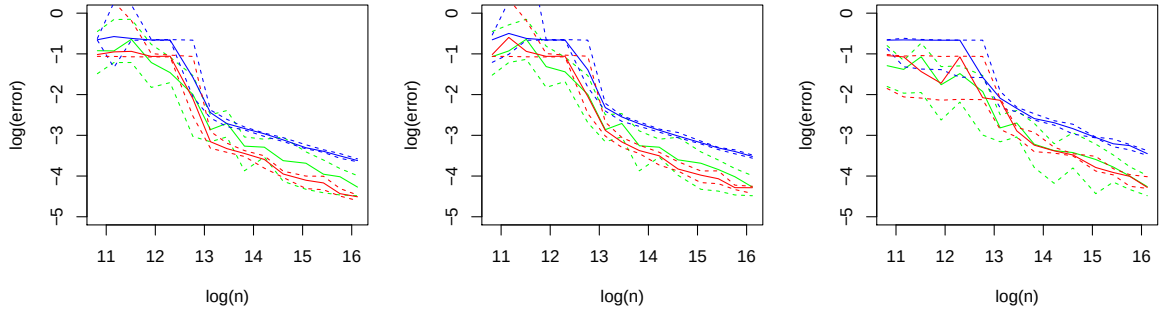
Furthermore, the rates of convergence are clearly distinct. The uniform distribution is estimated with a rate of convergence of approximately $n^{-1/2}$, which is also the best possible rate (it corresponds to a parametric estimation rate). In comparison, the rate of convergence for the



(a) eachjump (b) eachjump POS (c) eachjump MAX



(d) jumpmean (e) jumpmean POS (f) jumpmean MAX

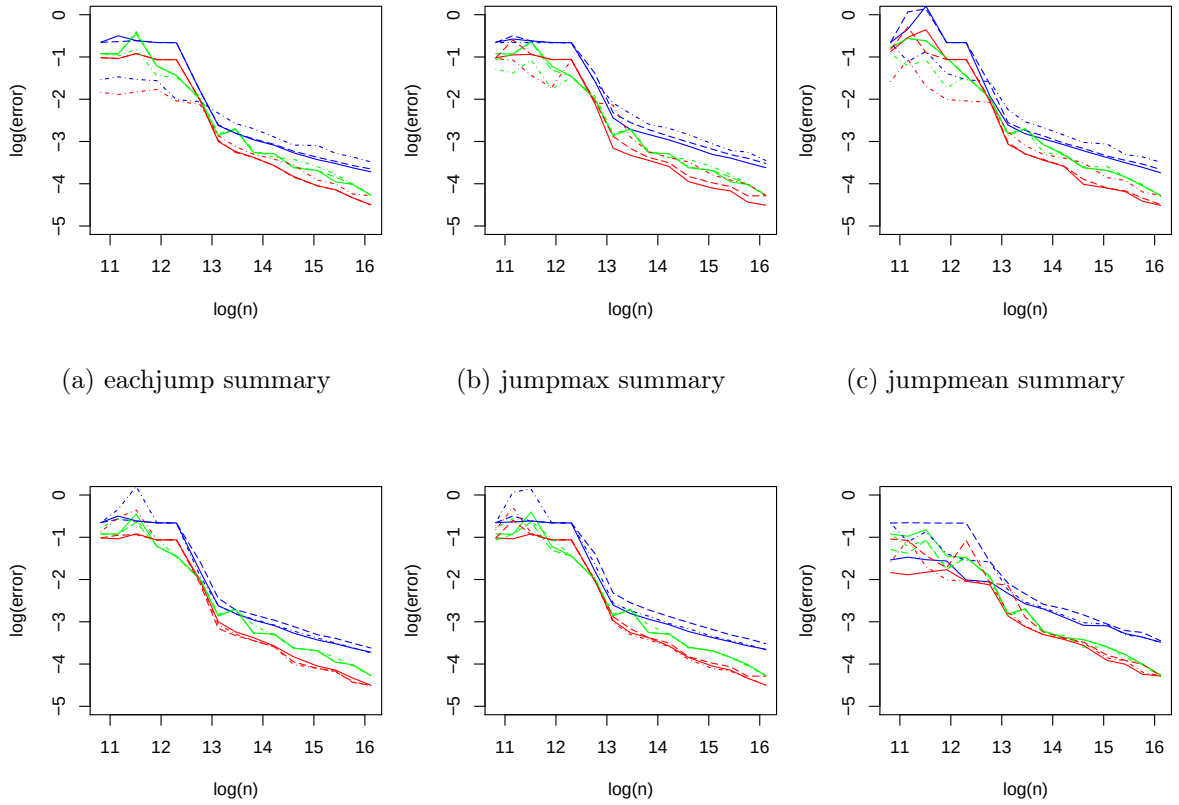


(g) jumpmax (h) jumpmax POS (i) jumpmax MAX

Figure 3.3: Logarithm of the $\mathbf{L}^2$ error on each emission densities depending on the logarithm of the number of observations for each of the selection and calibration methods. Each color corresponds to one emission density. The full lines are the medians of the 20 observations and the dashed ones are the 25 and 75 percentiles.

symmetrized Beta distribution is much slower (around $n^{-0.36}$). This shows that the algorithm effectively adapts to the regularity of each state and that one irregular emission density does not deteriorate the rates of convergence of the other densities.

Note that the above rates are in accordance with the minimax rates as far as the Hölder regularity is concerned. The minimax Hölder rate for the symmetrized Beta (which is 0.6-



(a) eachjump summary (b) jumpmax summary (c) jumpmean summary

(d) none summary (e) POS summary (f) MAX summary

Figure 3.4: Superposition of the median lines of Figure 3.3 by selection method and by calibration variant. Each color corresponds to one emission density. In Subfigures (a)-(c), the full lines correspond to the basic selection method, the dashed ones to the POS method and the dotted ones to the MAX method. In Subfigures (d)-(f), the full lines correspond to the eachjump method, the dashed ones to the jumpmax method and the dotted ones to the jumpmean method.

Hölder) is $n^{-3/11}$, or approximately $n^{-0.27}$, which means our estimator converges faster than the minimax rate would suggest. The minimax Hölder rate for the Beta distribution (which is 3-Hölder) is $n^{-3/7}$, or approximately $n^{-0.43}$, which is around the observed value.

3.4.5 Comparison with cross validation

In this section, we use a cross validation procedure based on our spectral estimators to check whether our method actually improves estimation accuracy.œ

When estimating a density by taking an estimator within some class (the model), two sources of error appear: the bias, that is the (deterministic) distance between the true density and the model, and the variance, that is the (random) error of the estimation within the model. Small models will have a large bias but a small variance, while large models will have a small bias and a large variance. The core issue of model selection is to select a model that minimizes the total error, that is large enough to accurately describe the true densities and small enough to prevent overfitting: in other words, perform a bias-variance tradeoff.

Cross validation seeks to achieve such a tradeoff by computing an estimate of the total

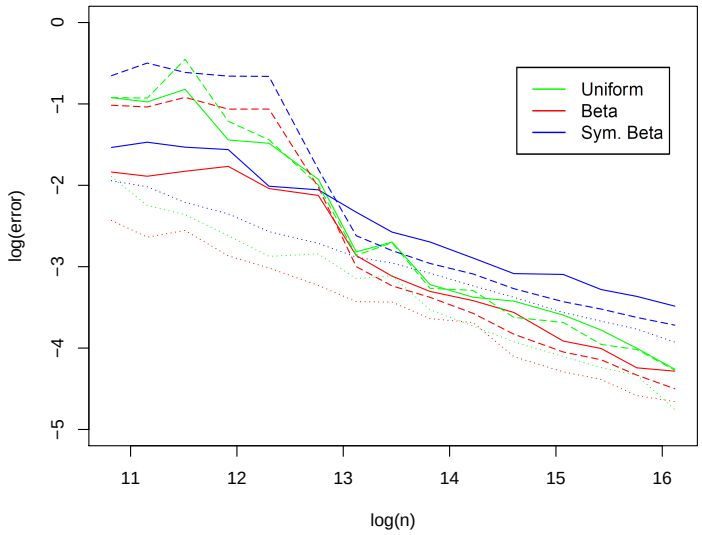


Figure 3.5: Comparison of the errors for the eachjump MAX method (full lines), for the eachjump method (dashed lines) and for the oracle estimators (dotted lines). For each k, the oracle estimator is defined as $\hat{f}_k^{(M_k^{\text{oracle}})}$ where M_k^{oracle} minimizes $M \longmapsto \|\hat{f}_k^{(M)} - f_k^*\|_2$. The oracle corresponds to the best estimator one could possibly select among the preliminary estimator if the true densities were known.

error. This is done by splitting the sample into two sets, the training sample being used for the calibration of the estimator and the validation sample for measuring the error. Taking the mean of these errors for different splits between training and validation samples provides an estimator of the total error. This method has become popular for its simplicity of use. We refer to the survey of Arlot and Celisse (2010) for an overview on this method and its guarantees.

Risk

We use the least squares criterion of Algorithm 3 to quantify the error of the estimators. Since the guarantees on spectral estimators rely on the $\mathbf{L}^2$ norm, a least squares criterion is more natural than the likelihood. In addition, the spectral estimators might take negative values depending on the orthonormal basis, which is not a problem as far as $\mathbf{L}^2$ error is concerned but can be an issue for the likelihood.

Let us first recall this criterion. Given an orthonormal basis $(\varphi_i)_{i\in\mathbb{N}}$ of $\mathbf{L}^2(\mathcal{Y},\mu)$, define the coordinate tensor of the empirical distribution of the triplet (Y_1,Y_2,Y_3) on this basis by

$$\hat{\mathbf{M}}(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}) \quad \text{for all } a,b,c \in \mathbb{N}.$$

Given a transition matrix $\mathbf{Q}$ of size K, a stationary distribution π of $\mathbf{Q}$ and a vector of densities $\mathbf{f} = (f_1,\dots,f_K)$, define the coordinate matrix $\mathbf{O}$ of $\mathbf{f}$ by $\mathbf{O}(b,k) = \langle \varphi_b, f_k\rangle$. Let $\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})}$ be the coordinate tensor of the distribution of (Y_1,Y_2,Y_3) under the parameters $(\pi,\mathbf{Q},\mathbf{f})$, that is

$$\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})}(\cdot,b,\cdot) = \mathbf{O}\text{diag}[\pi]\mathbf{Q}\text{diag}[\mathbf{O}(b,\cdot)]\mathbf{Q}\mathbf{O}^\top \quad \text{for all } b \in \mathbb{N}.$$

The empirical least squares criterion is $\|\mathbf{M}_{(\pi,\mathbf{Q},\mathbf{f})} - \hat{\mathbf{M}}\|_F^2$. It corresponds to the $\mathbf{L}^2$ error

between the empirical distribution of three consecutive observations and the theoretical distribution under the parameters $(\pi, \mathbf{Q}, \mathbf{f})$.

Implementation

We use 10-fold cross validation, that is we split the sequence into 10 segments of same size $I_1, \dots, I_{10}$. In order to avoid interferences between samples, we prune the ends of each segment, so that the observations in each segment can be considered independent. In practice, we take a gap of 30 observations between two segments.

We ran 150 simulations, 10 per value of n, with the same parameters as in Section 3.4.1. Each simulation is as follows.

For each segment I_j, we run the spectral algorithm on all models $\mathfrak{P}_M$ for $\texttt{M\_min} \leqslant M \leqslant \texttt{M\_max}$ using only the observations from the other segments. The transition matrix is estimated using an additional step of the spectral method which is adapted from Steps 8 and 9 of Algorithm 1 of De Castro et al. (2017). Then, we compute the least squares criterion for the estimated parameters using the segment I_j as observed sample. Finally, for each M, we average this error on all segments I_j, which gives the least squares cross validation error $E_{\mathrm{VC}}(M)$.

This cross validation criterion is used to select one model $\hat{M}_{\mathrm{VC}} \in \arg\min_M E_{\mathrm{VC}}(M)$, from which we construct the final estimators of the emission densities $\hat{f}_k = \hat{f}_k^{(\hat{M}_{\mathrm{VC}})}$ for all k. Note that the selected model is the same for all emission densities.

Results

Figure 3.6 compares the selected model dimensions for each n using our state-by-state selection method and using the cross validation method. When the number of observations n becomes larger than 10^6, the cross validation tends to always pick the largest model, which means that it does not prevent overfitting as well as our method.

The $\mathbf{L}^2$ errors on the emission densities are shown in Figure 3.7. It appears that the cross validation has a lower error for small n ($n \leqslant 350,000$) than our method. However, for larger values of n, the errors becomes larger than the ones of our method (see Figure 3.5) by up to one order of magnitude, and only start decreasing once the selected model is set to the maximum dimension.

Finally, the estimated rates of convergence are shown in Table 3.1. Our state-by-state method outperforms the cross validation method for all emission densities. The cross validation estimators only reach the minimax rate of convergence for the less regular density: the symmetrized Beta, and even then they converge slower than the state-by-state estimator. All other emission densities are estimated slower than their minimax rate.

3.4.6 Algorithmic complexity

In the following, we treat K as a constant as far as the algorithmic complexity is concerned. The different complexities are summarized in Table 3.2.

Spectral algorithm (see Section 3.3.2)

We consider the algorithmic complexity of estimating the emission densities for all models M such that $M_{\min} \leqslant M \leqslant M_{\max}$ with n observations and auxiliary parameters r and m depending on n and M (upper bounded by $m_{\max}$ and $r_{\max}$).

Step 1 can be computed for all models with $O(nM_{\max}m_{\max}^2)$ operations. It is the only step whose complexity depends on n. Steps 2 and 3 require $O(m^3 M)$ operations for each model

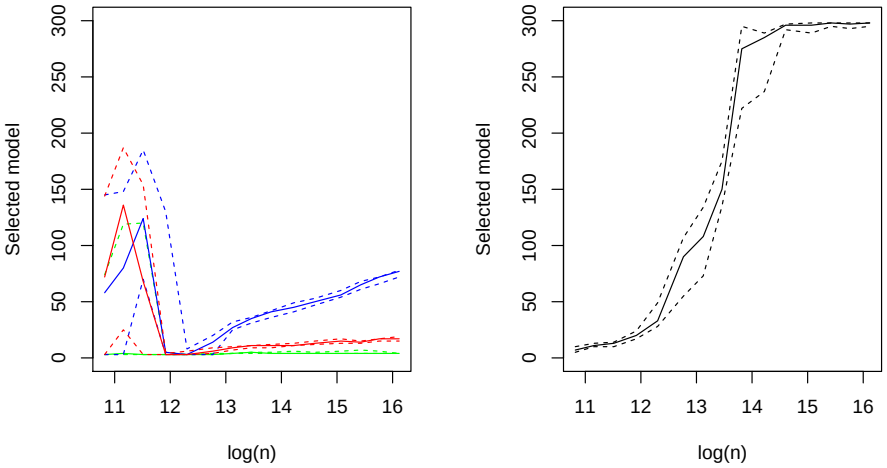


Figure 3.6: Selected model dimensions for each n using our state-by-state selection method (left) and 10-fold cross validation (right). The full lines are the median model dimensions and the dashed lines are the 25 and 75 percentiles.

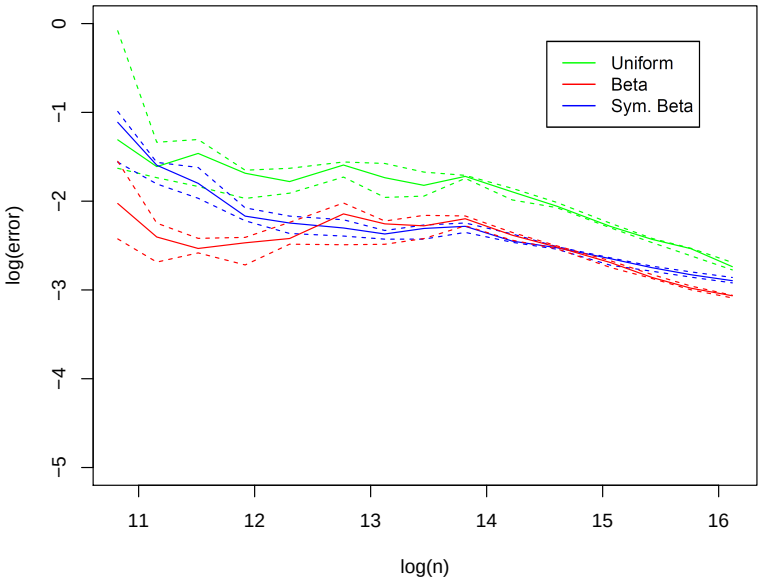


Figure 3.7: Error of the cross validation estimators for each n using 10-fold cross validation. The full lines are the median errors for each density and the dashed lines are the 25 and 75 percentiles.

and Steps 4 to 7 require $O(Mr)$ operations for each model, for a total of $O(nM_{\max}m_{\max}^2 + M_{\max}^2 m_{\max}^3 + M_{\max}^2 r_{\max})$ operations.

In practice, one takes $m \propto \log(n)$, $r \propto \log(n) + \log(M)$ and $M_{\max} \leqslant n$, so that the total complexity of the spectral algorithm is $O(n \log(n)^2 M_{\max})$.

In comparison, the complexity of the spectral algorithm of De Castro et al. (2017) is $O(nM_{\max}^3)$ because of Step 1. This becomes much larger than our complexity when $M_{\max}$ grows as a power of n (which is necessary in order to reach minimax rates).

Least squares algorithm (see Section 3.3.3)

We consider the algorithmic complexity of estimating the emission densities for all models M such that $M_{\min} \leqslant M \leqslant M_{\max}$ with n observations.

Step 1 is similar to the one of the spectral algorithm, but with $O(nM_{\max}^3)$ operations. The complexity of Step 2 is more difficult to evaluate. Since the criterion is nonconvex, finding the minimizer requires to run an approximate minimization algorithm whose complexity C_n will depend on the desired precision–which will in turn depend on the number of observations n–and on the initial points. As discussed in Lehéricy (2019), this is usually the longest step when computing least squares estimators. Thus, the total complexity of the least squares algorithm is $O(nM_{\max}^3 + C_n)$.

Note that despite the worse sample complexity, the least squares algorithm is tractable and can greatly improve the estimation for small sample size. As shown in Section 3.4.4, the spectral algorithm is unstable for small samples, which makes the state-by-state selection procedure return abnormal results. This can be explained by the matrix inversions of the spectral method, which sometimes lead to nearly singular matrices when the noise is too large. On the other hand, the least squares method does not involve any matrix inversion, and often gives better results than the spectral estimators, as shown in de Castro et al. (2016), thus making it a relevant choice for small to medium data sets.

Selection method and POS variant (see Sections 3.2.2 and 3.4.3)

We consider the algorithmic complexity of selecting estimators from a family of estimators $(\hat{\mathbf{f}}^{(M)})_{M_{\min} \leqslant M \leqslant M_{\max}}$. The selection algorithms can be decomposed in two parts.

- Compute the distances $\|\hat{f}_k^{(M)} - \hat{f}_k^{(M')}\|_2$ for all M, M' and k. This has complexity $O(M_{\max}^3)$: it requires to compute the $\mathbf{L}^2$ distance of at most $M_{\max}^2$ couples of functions in a Hilbert space of dimension $M_{\max}$.
- Compute $\hat{\rho}_k$ defined as the abscissa of the largest jump of the function $\rho \longmapsto \hat{M}_k(\rho)$ for all k, where $\hat{M}_k$ is defined as in Section 3.4.2. Note that computing $\hat{M}_k(\rho)$ requires $O(M_{\max}^2)$ operations. An approximate value of $\hat{\rho}_k$ can be computed in $O(\log(\hat{\rho}_k) M_{\max}^2)$ operations, which is usually $O(M_{\max}^2)$.

Once the $\hat{\rho}_k$ are known, it is possible to calibrate the penalty in constant time for the three calibrations methods (eachjump, jumpmax and jumpmean) and to select the final models in $O(M_{\max}^2)$ operations.

Thus, the total complexity of the selection algorithm and of its POS variant is $O(M_{\max}^3)$.

Selection method, MAX variant (see Section 3.4.3)

In the MAX variant, the first step of the standard selection procedure is replaced by computing the distances $\|\hat{f}_k^{(M_{\max})} - \hat{f}_k^{(M)}\|_2$ for all M. This has complexity $O(M_{\max}^2)$. The complexity of the other steps remains unchanged.

	Algorithm	Complexity
Preliminary estimators	Spectral method	$O(n\log(n)^2 M_{\max})$
	Spectral method (De Castro et al. (2017))	$O(nM_{\max}^3)$
	Least squares method	$O(nM_{\max}^3 + C_n)$
Selection step	Standard and POS variant	$O(M_{\max}^3)$
	MAX variant	$O(M_{\max}^2)$

Table 3.2: Complexities of the different algorithms. n is the number of observations, $M_{\max}$ is the largest model dimension considered.

Thus, the total complexity of the MAX variant of the selection algorithm is $O(M_{\max}^2)$.

3.5 Application to real data

In this section, we present the results of our method on two sets of trajectories. Trajectories are a typical example of dependent data that shows several behaviours depending on the activity of the entity being tracked, which makes hidden Markov models a popular modelling choice. For instance, the movement of a fisher is not the same depending on whether he's travelling to the next fishing zone or actually fishing.

The first data set follows artisanal fishers in Madagascar. The second one contains seabird movements. Studying the movements of fishers and seabirds has many applications, for instance understanding the fishing habits of the tracked entity, controlling the fishing pressure on local ecosystems and monitoring the dynamics of coastal ecosystems, see for instance Boyd et al. (2014); Vermard et al. (2010) and references therein.

3.5.1 Artisanal fishery

We use GPS tracks of artisanal fishers with a regular sampling period of 30 seconds. These tracks were produced by Faustinato Behivoke (Institut Halieutiques et des Sciences Marines, Université de Toliara, Madagascar) and Marc Léopold (IRD), who recorded artisanal fishers from Ankilibe, in Madagascar. Their fishing method is a seine netting.

From this data, we compute the velocity of the fisher during each time step. In order to estimate densities on $[0,1]$, we divide this velocity by an upper bound of the maximum observed velocity. We consider the observation space $\mathcal{Y} = [0,1]$ endowed with the dominating measure $\delta_0 + \text{Leb}$, where δ_0 is the dirac measure in zero and Leb is the Lebesgue measure on $[0,1]$. As a proof of concept, we use the orthonormal basis consisting of the trigonometric basis on $[0,1]$ and the indicator function of $\{0\}$, that is the family $(\varphi_m)_{m\in\mathbb{N}}$ defined on $[0,1]$ by

$$\text{if } x = 0, \begin{cases} \varphi_0(x) = 1 \\ \varphi_m(x) = 0 \quad \text{for all } m \in \mathbb{N}^* \end{cases}$$

$$\text{if } x \neq 0, \begin{cases} \varphi_0(x) = 0 \\ \varphi_1(x) = 1 \\ \varphi_{2m}(x) = \sqrt{2}\cos(2\pi m x) \quad \text{for all } m \in \mathbb{N}^* \\ \varphi_{2m+1}(x) = \sqrt{2}\sin(2\pi m x) \quad \text{for all } m \in \mathbb{N}^* \end{cases}$$

The number of hidden states is chosen using the spectral thresholding method of Lehéricy (2019). This methods consists is based on the fact that the rank of the spectral tensor $\mathbb{E}\hat{\mathbf{N}}_{m,m}$

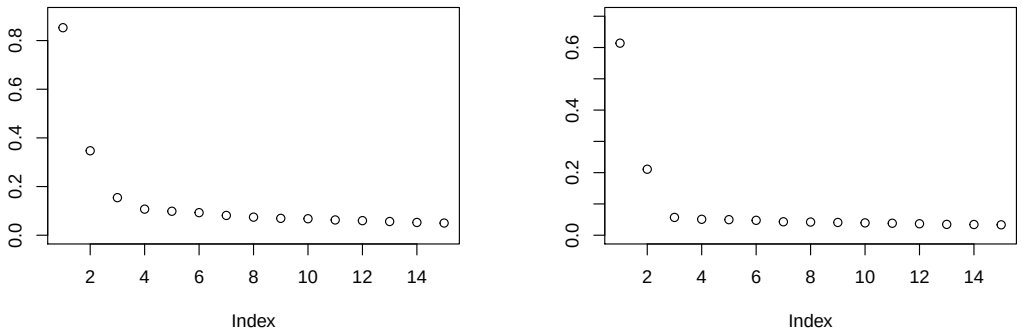


Figure 3.8: First 15 eigenvalues of the spectrum of the empirical tensor $\hat{\mathbf{N}}_{50,50}$ (see Algorithm 4 in Appendix 3.A). Left: fisher 1, right: fisher 2.

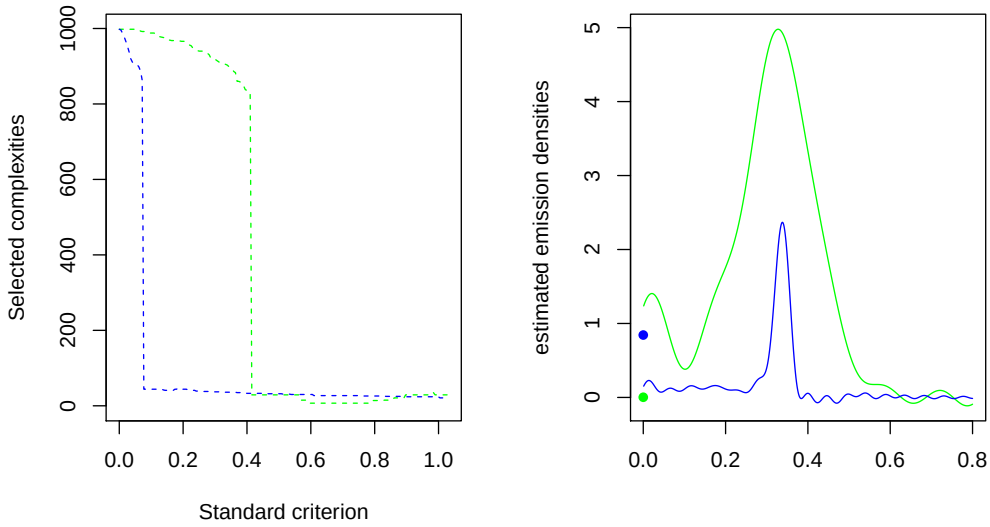


Figure 3.9: Selected complexities and estimated densities on artisanal fishery data (fisher 1, $n = 17,300$). Green = state 1, blue = state 2. The dirac component is shown as a dot at $y = 0$. The selected dimensions are $(14, 41)$.

(with the notations of Algorithm 4 in Appendix 3.A) is the number of hidden states. This is visible in the spectrum of $\hat{\mathbf{N}}_{m,m}$ by an elbow, as shown in Figure 3.8. Based on these spectra, we use two hidden states.

The results using `M_max` $= 1000$ are shown in Figures 3.9 and 3.10. We took the normalizing velocity large enough that all observed normalized velocities belong to $[0, 0.8]$, hence the plot betweeen 0 and 0.8 for the densities.

In both cases, the selected model complexities differ greatly depending on the state. This comes from the fact that in both cases, one of the density is spiked, thus requiring more vectors of the orthonormal basis to be approximated. This illustrates that our method is able to estimate the smoother densities with fewer vectors of the basis, thus preventing overfitting.

As a side note, we needed considerably less observations than in the simulations: around 10,000, compared to 500,000 in the simulations. This can be explained by the fact that each state is very stable, with an estimated probability of leaving the states below 0.02–compared to 0.3 in the simulations. This is encouraging, as hidden states in real data are expected to be rather

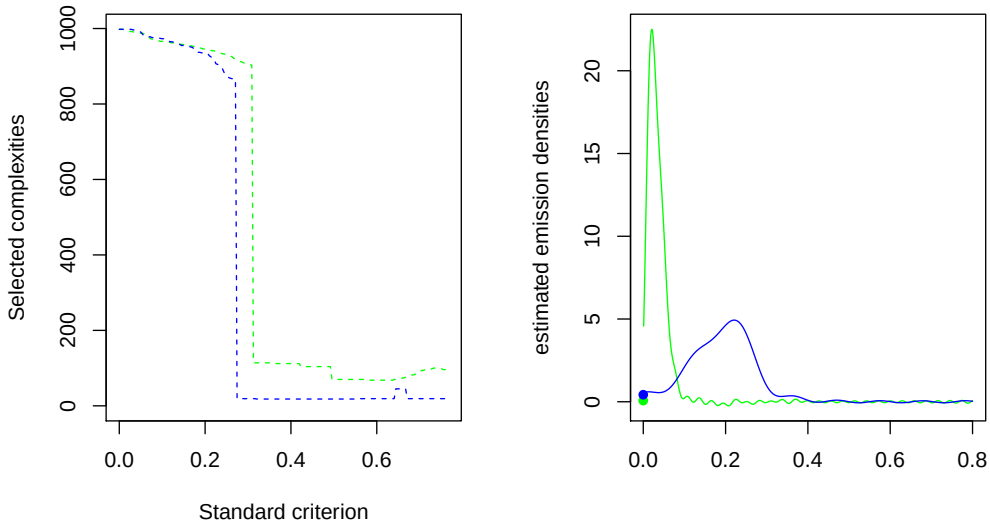


Figure 3.10: Selected complexities and estimated densities on artisanal fishery data (fisher 2, $n = 11,600$). Green = state 1, blue = state 2. The dirac component is shown as a dot at $y = 0$. The selected dimensions are $(68, 18)$.

stable, especially when the sampling frequency is high, as long as the conditional independance of the observations can be assumed to hold.

3.5.2 Seabird foraging

In this section, we consider the seabird data from Bertrand et al. (2015) and we focus on the tracks named cormorant d in this paper.

We apply the same transformation as in the previous section to obtain normalized velocities in $[0, 0.8]$ (after removal of anomalous velocities exceeding 150 m/s) and run the spectral algorithm with the trigonometric basis on $[0, 1]$ plus the indicator of $\{0\}$. The spectral thresholding gives a number of hidden states equal to two; we set it to three to account for more complex behaviours of the seabirds. The results are shown in Figure 3.11.

Note that the use of the trigonometric basis allows the estimated densities to take negative values. This is not a problem as far as minimax rates of convergence (in $\mathbf{L}^2$ norm) are concerned, however this can become an issue if one wants to use these densities in a forward-backward algorithm in order to get an estimator of the hidden states. One way to circumvent this problem is to use simplex projection to compute an approximation of the projection of these estimated density on the simplex of all probability densities. Note that since this is an $\mathbf{L}^2$ projection on a convex set which contains the true densities, the projected densities have an even smaller error, thus keeping the minimax rate of convergence of the original estimators. The resulting densities are shown in Figure 3.12

The number of observations in this setting is even smaller than for the fishery's data set: our algorithm was able to recover three emission densities from less than 3,000 observations, despite the states being less stable than in the fishery data set: the diagonal terms of the estimated transition matrix using the EM algorithm are $(0.83, 0.93, 0.98)$. In addition, the result of our method is consistent with other estimation methods, as shown in Figure 3.12: estimating the parameters with the EM algorithm using piecewise constant densities leads to a very similar result.

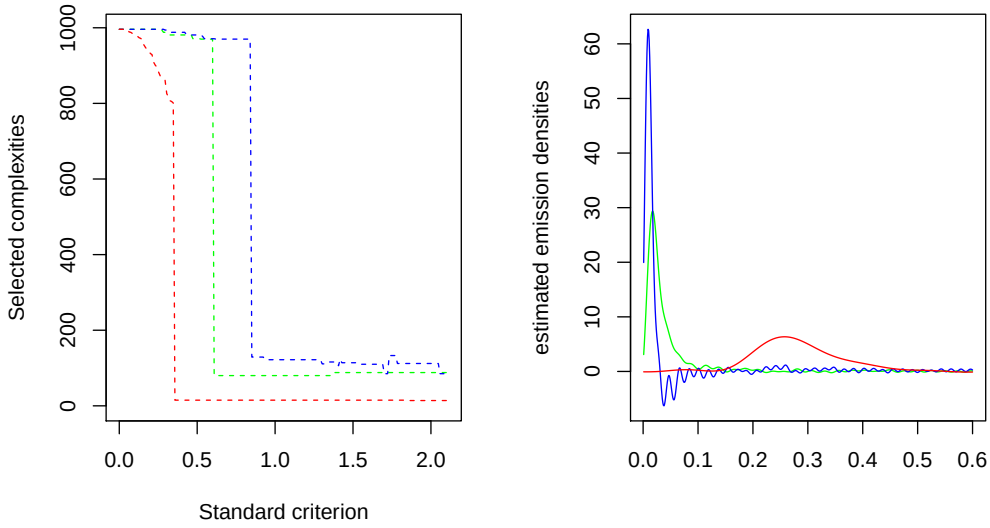


Figure 3.11: Selected complexities and estimated densities for Cormorant d's trajectory ($n = 2,891$). Green = state 1, blue = state 2, red = state 3. The selected dimensions are $(80, 110, 15)$.

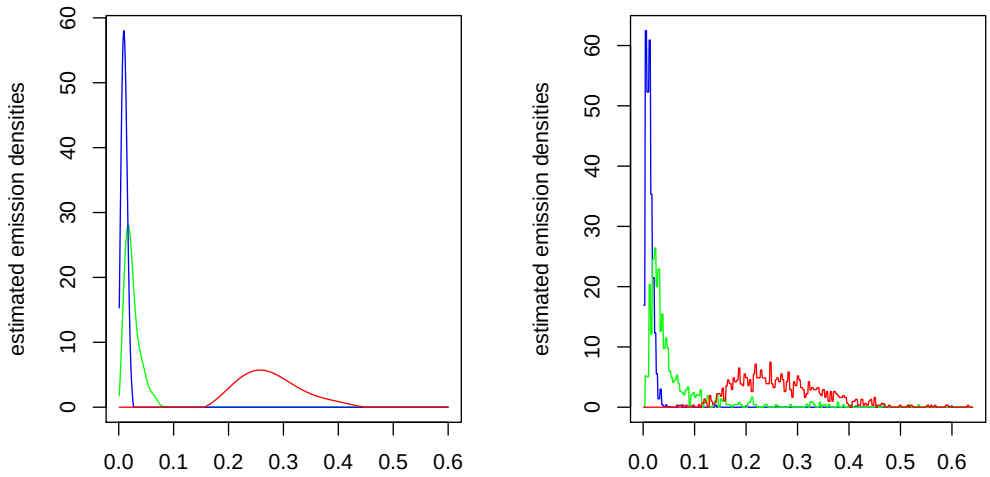


Figure 3.12: Projection of the estimated densities of Figure 3.11 for Cormorant d on the set of probability densities (left) and comparison with an estimation with histogram densities on a regular partition of size 300 using the EM algorithm (right).

3.6 Conclusion and perspectives

We propose a state-by-state selection method to infer the emission densities of a HMM. Using a family of estimators, our method selects one estimator for each hidden state in a way that is adaptive with respect to this state's regularity. This method does not depend on the type of preliminary estimator, as long as a suitable variance bound is available. As such, it may be seen as a plug-in that takes a family of estimators and the corresponding variance bound and outputs the selected estimator. Note that its complexity does not depend on the number of observations

used to compute the estimators, which makes it applicable to arbitrarily large data sets.

To apply this method, we present two families of estimators: a least squares estimator and a spectral estimator. For both, we prove a bound on their variance and show that this bound allows to recover the minimax rate of convergence separately on each hidden state, up to a logarithmic factor. The variance bounds are similar to a BIC penalty, with an additional logarithmic factor for the spectral estimators.

We carry out a numerical study of the method and some variants on simulated data. We use the spectral estimators, which are both fast and don't suffer from initialization issues, unlike the least squares and maximum likelihood estimators. The simulations show that our selection method is very fast compared to the computation of the estimators and that indeed, the final estimators reach the minimax rate of convergence on each state.

Then, we compare our method with a cross validation estimator based on a least square risk. This estimator only reaches the minimax rate corresponding to the worst regularity among the emission densities and fails to select models with small dimensions. It is still noteworthy that the cross validation returns relevant results for small sample sizes, whereas our method requires the sample size to be large enough to work properly. An interesting problem would be to investigate whether cross validation or other methods can be combined with our state-by-state selection method to give an algorithm that is at the same time fast, stable for small sample sizes and optimal in the asymptotic setting.

Finally, we apply our algorithm to real trajectory data sets. On this data, our method proves that it is able to match the regularity of the underlying emission densities. In addition, it is able to produce sensible results with far fewer observations than in our simulation study.

Our state-by-state selection method can be easily applied to multiview mixture models (also named mixture models with repeated measurement, see for instance Bonhomme et al. (2016a) and Gassiat et al. (2018)). Let us first describe the model. A multiview mixture model consists of two random variables, a hidden state U and an observation vector $\mathbf{Y} := (Y_i)_{i\in[m]}$ such that conditionally to U, the components Y_i of $\mathbf{Y}$ are independent with a distribution depending only on U and i. Let us assume that U takes its values in a finite set $\mathcal{X}$ of size K and that the Y_i have some density $f^*_{u,i}$ conditionally to $U = u$ with respect to a dominating measure. A question of interest is to estimate the densities $f^*_{u,i}$ from a sequence of observed $(\mathbf{Y}_n)_{n\geqslant 1}$.

Our state-by-state selection method can be applied directly to such a model as long as estimators with a proper variance bound are available (see assumption $\mathbf{[H}(\epsilon)\mathbf{]}$ in Section 3.2.1). Indeed, we never use the dependency structure of the model. Regarding the development of preliminary estimators, multiview mixture models appear closely related to hidden Markov models: Anandkumar et al. (2012) and Bonhomme et al. (2016b) develop spectral methods that work for both multiview mixtures and HMMs at the same time using the same theoretical arguments. Thus, it seems clear that variance bounds such as the ones we developed can also be written for multiview mixture models.

Acknowledgments

I am grateful to Elisabeth Gassiat and Claire Lacour for their precious advice. I thank Augustin Touron for providing me with a R implementation of the spectral algorithm. I would also like to thank Marie-Pierre Etienne and of course Faustinato Behivoke (Institut Halieutiques et des Sciences Marines, Université de Toliara, Madagascar), Marc Léopold (IRD) and Sophie Bertrand (IRD) for letting me work on their data sets.

APPENDICES

3.A Spectral algorithm, full version

Algorithm 4: Spectral estimation of the emission densities of a HMM (full version)

Data: A sequence of observations $(Y_1, \dots, Y_{n+2})$, two dimensions $m \leqslant M$, an orthonormal basis $(\varphi_1, \dots, \varphi_M)$ and number of retries r.

Result: Spectral estimators $(\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$, $\hat{\mathbf{Q}}$ and $\hat{\pi}$.

[Step 1] Consider the following empirical estimators: for any $a, c \in [m]$ and $b \in [M]$,

- $\hat{\mathbf{L}}_m(a) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)$
- $\hat{\mathbf{M}}_{m,M,m}(a,b,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$
- $\hat{\mathbf{N}}_{m,M}(a,b) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})$
- $\hat{\mathbf{P}}_{m,m}(a,c) := \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_c(Y_{s+2})$.

[Step 2] Let $\hat{\mathbf{U}}_m$ be the $m \times K$ matrix of orthonormal left singular vectors and $\hat{\mathbf{V}}_M$ be the $M \times K$ matrix of orthonormal right singular vectors of $\hat{\mathbf{N}}_{m,M}$ corresponding to its top K singular values.

[Step 3] Form the matrices for all $b \in [M]$, $\hat{\mathbf{B}}(b) := (\hat{\mathbf{U}}_m^\top \hat{\mathbf{P}}_{m,m} \hat{\mathbf{U}}_m)^{-1} \hat{\mathbf{U}}_m^\top \hat{\mathbf{M}}_{m,M,m}(.,b,.)\hat{\mathbf{U}}_m$.

[Step 4] Set $(\Theta_i)_{1\leqslant i\leqslant r}$ r iid $(K \times K)$ unitary matrix uniformly drawn. Form the matrices for all $k \in \mathcal{X}$ and $i \in [r]$, $\hat{\mathbf{C}}_i(k) := \sum_{b=1}^{M} (\hat{\mathbf{V}}_M \Theta_i)(b,k)\hat{\mathbf{B}}(b)$.

[Step 5] Compute $\hat{\mathbf{R}}_i$ a $(K \times K)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{\mathbf{C}}_i(1)$: $\hat{\mathbf{R}}_i^{-1}\hat{\mathbf{C}}_i(1)\hat{\mathbf{R}}_i = \mathrm{diag}(\hat{\Lambda}_i(1,1), \dots, \hat{\Lambda}_i(1,K))$.

[Step 6] Set for all $k, k' \in \mathcal{X}$, $\hat{\Lambda}_i(k,k') := (\hat{\mathbf{R}}_i^{-1}\hat{\mathbf{C}}_i(k)\hat{\mathbf{R}}_i)(k',k')$. Choose i_0 maximizing $\min_k \min_{k_1\neq k_2} |\hat{\Lambda}_i(k,k_1) - \hat{\Lambda}_i(k,k_2)|$ and set $\hat{\mathbf{O}} := \hat{\mathbf{V}}_M \Theta_{i_0} \hat{\Lambda}_{i_0}$.

[Step 7] Consider the emission densities estimators $\hat{\mathbf{f}}^{(M,r)} := (\hat{f}_k^{(M,r)})_{k\in\mathcal{X}}$ defined by for all $k \in \mathcal{X}$, $\hat{f}_k^{(M,r)} := \sum_{b=1}^{M} \hat{\mathbf{O}}(b,k)\varphi_b$.

[Step 8] Let $\hat{\mathbf{O}}_m$ be the $m \times K$ matrix containing the first m rows of $\hat{\mathbf{O}}$. Set $\hat{\pi} = \Pi_\Delta\left((\hat{\mathbf{U}}_m^\top \hat{\mathbf{O}}_m)^{-1}\hat{\mathbf{U}}_m^\top \hat{\mathbf{L}}_m\right)$ where Π_Δ is the $\mathbf{L}^2$ projection onto the probability simplex.

[Step 9] Let $\hat{\mathbf{Q}}$ be the transition matrix defined by $\hat{\mathbf{Q}} = \Pi_{\mathrm{TM}}\left((\hat{\mathbf{U}}_m^\top \hat{\mathbf{O}}_m \mathrm{diag}[\hat{\pi}])^{-1}\hat{\mathbf{U}}_m^\top \hat{\mathbf{N}}_{m,M}\hat{\mathbf{V}}_M(\hat{\mathbf{O}}^\top \hat{\mathbf{V}}_M)^{-1}\right)$ where Π_{TM} is the projection onto the set of transition matrices. This projection is obtained by projecting each line of the matrix onto the probability simplex.

3.B Proofs

3.B.1 Proof of Lemma 3.1

Let $\tau_{n,M}$ be the permutation that minimizes $\tau \longmapsto \max_{k\in\mathcal{X}} \left\|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\right\|_2$. **[H($\epsilon$)]** means that with probability $1-\epsilon$, one has $\max_{k\in\mathcal{X}} \left\|\hat{f}_k^{(M)} - f_{\tau(k)}^{*,(M)}\right\|_2 \leqslant \frac{\sigma(M)}{2}$.

Let $M \in \mathcal{M}$. Let us show that $\left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k')} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2 > \left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k)} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2$ for all $k, k' \in \mathcal{X}$ such that $k' \neq k$. If this holds, then the definition of $\hat{\tau}^{(M)}$ implies that $\hat{\tau}^{(M)} = \tau_{n,M}^{-1} \circ \tau_{n,M_0}$. Thus, one has $\max_{k\in\mathcal{X}} \left\|\hat{f}_{k,\text{new}}^{(M)} - f_{\tau_{n,M_0}(k)}^{*,(M)}\right\|_2 \leqslant \frac{\sigma(M)}{2}$, which is exactly Equation (3.1) with $\tau_n = \tau_{n,M_0}$.

Applying the triangular inequality leads to

$$\begin{aligned}\left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k)} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2 \leqslant& \left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k)} - f_k^{*,(M)}\right\|_2 + \left\|f_k^{*,(M)} - f_k^{*,(M_0)}\right\|_2 + \left\|f_k^{*,(M_0)} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2 \\ \leqslant& \frac{\sigma(M)}{2} + B_{M,M_0} + \frac{\sigma(M_0)}{2}\end{aligned}$$

and

$$\begin{aligned}\left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k')} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2 \geqslant& \left\|f_{k'}^{*,(M_0)} - f_k^{*,(M_0)}\right\|_2 - \left\|\hat{f}^{(M)}_{\tau_{n,M}^{-1}(k')} - f_{k'}^{*,(M)}\right\|_2 \\ &- \left\|f_{k'}^{*,(M)} - f_{k'}^{*,(M_0)}\right\|_2 - \left\|f_k^{*,(M_0)} - \hat{f}^{(M_0)}_{\tau_{n,M_0}^{-1}(k)}\right\|_2 \\ \geqslant& m(\mathbf{f}^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2}.\end{aligned}$$

Thus, the result holds as soon as $m(\mathbf{f}^*, M_0) - \frac{\sigma(M)}{2} - B_{M,M_0} - \frac{\sigma(M_0)}{2} > \frac{\sigma(M)}{2} + B_{M,M_0} + \frac{\sigma(M_0)}{2}$, which is the condition of Lemma 3.1.

3.B.2 Proof of Theorem 3.3

The structure of the proof is the same as the one of Theorem 3.1 of De Castro et al. (2017).

The first difference lies in the fact that we consider different models for each component of the tensors $\hat{\mathbf{N}}_{m,M}$ and $\hat{\mathbf{M}}_{m,M,m}$ in Step 1. As a consequence, we use the left *and* right singular vectors of $\hat{\mathbf{N}}_{m,M}$ instead of just the right singular vectors of $\hat{\mathbf{P}}_{m,m}$. A careful reading shows that their proof can be adapted straightforwardly to this situation.

The second difference consists in generating several independant random unitary matrices in Step 4 and keeping the one that separates the eigenvalues of all $\hat{\mathbf{C}}_i(k)$ best. This allows to replace Lemma F.6 of De Castro et al. (2017) by the following one, based on the independence of the unitary matrices:

Lemma 3.10. *For all $x > 0$ and $r \in \mathbb{N}^*$,*

$$\mathbb{P}\left[\forall k, k_1 \neq k_2, |\hat{\Lambda}_{i_0}(k,k_1) - \hat{\Lambda}_{i_0}(k,k_2)| \geqslant \frac{2e^{-x/r}(1-\epsilon^2_{\mathbf{N}_{m,M}})^{1/2}}{\sqrt{e}K^{5/2}(K-1)}\gamma(\mathbf{O}_M)\right] \geqslant 1 - e^{-x}$$

and

$$\mathbb{P}\left[\|\hat{\Lambda}_{i_0}\|_\infty \geqslant \frac{1+\sqrt{2}\sqrt{x+\log(K^2 r)}}{\sqrt{K}}\|\mathbf{O}_M\|_{2,\infty}\right] \leqslant e^{-x},$$

The notations $\epsilon_{\mathbf{N}_{m,M}}$ (or $\epsilon_{\mathbf{P}_M}$ in the original proof), $\gamma(\mathbf{O}_M)$ et $\|\mathbf{O}_M\|_{2,\infty}$ are introduced in De Castro et al. (2017).

Using this lemma, their proof leads to our result by taking $r = x = t$.

3.B.3 Definition of the polynomial H

Definition

We parameterize the application

$$(\pi, \mathbf{Q}, \mathbf{f}) \in \Delta \times \mathcal{Q} \times \mathrm{Span}(\mathbf{f}^*)^K \longmapsto \|g^{\pi,\mathbf{Q},\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \tag{3.3}$$

in the following way. For $p \in \mathbb{R}^{K-1}$, $q \in \mathbb{R}^{K\times(K-1)}$ and $A \in \mathbb{R}^{K\times(K-1)}$, define the extensions

- $\bar{p} \in \mathbb{R}^K$ defined by $\bar{p}(k) = p(k)$ for all $k \in [K-1]$ and $\bar{p}(K) = -\sum_{k\in[K-1]} p(k)$;
- $\bar{q} \in \mathbb{R}^{K\times K}$ by $\bar{q}(k, K) = -\sum_{k'\in[K-1]} q(k, k')$;
- $\bar{A} \in \mathbb{R}^{K\times K}$ by $\bar{A}(k, K) = -\sum_{k'\in[K-1]} A(k, k')$.

$\bar{p}$ corresponds to $\pi - \pi^*$, $\bar{q}$ to $\mathbf{Q} - \mathbf{Q}^*$ and A to the components of $\mathbf{f} - \mathbf{f}^*$ on $\mathbf{f}^*$ (which is a basis as soon as **[Hid]** holds). The condition on the last component of $\bar{p}$ and of each line of $\bar{q}$ and $\bar{A}$ follows from the fact that $\bar{p}$ corresponds to the difference of two probability vectors, $\bar{q}$ corresponds to the difference of two transition matrices and $\bar{A}$ correspond to the difference of two vectors of probability densities on a basis of probability densities.

Then, consider the quadratic form derived from the Taylor expansion of

$$(p, q, A) \in \mathbb{R}^{K-1} \times \mathbb{R}^{K\times(K-1)} \times \mathbb{R}^{(K-1)\times K} \longmapsto \|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2.$$

Let M be the matrix associated to this quadratic form. We define H as the determinant of M. Direct computations show that H is a polynomial in the coefficients of π^*, $\mathbf{Q}^*$ and $G(\mathbf{f}^*)$.

Link between H and the quadratic form from Equation (3.3)

The goal of this section is to show how H can be used to lower bound the quadratic form from Equation (3.3) by a positive constant times the distance between $(\pi, \mathbf{Q}, \mathbf{f})$ and $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$. We will not need the assumptions **[Hid]**, **[HF]** or **[Hdet]** unless specified otherwise.

Let us start by the relation between the norms of (p, q, A) and $(\bar{p}, \bar{q}, \bar{A})$.

Lemma 3.11. *For all $(p, q, A) \in \mathbb{R}^{K-1} \times \mathbb{R}^{K\times(K-1)} \times \mathbb{R}^{(K-1)\times K}$,*

$$\begin{aligned}
\|p\|_2^2 \leqslant \|\bar{p}\|_2^2 \leqslant K\|p\|_2^2,\\
\|q\|_F^2 \leqslant \|\bar{q}\|_F^2 \leqslant K\|q\|_F^2,\\
\|A\|_F^2 \leqslant \|\bar{A}\|_F^2 \leqslant K\|A\|_F^2.
\end{aligned}$$

Proof. $\|p\|_2^2 \leqslant \|\bar{p}\|_2^2$ is immediate. Then,

$$\begin{aligned}
\|\bar{p}\|_2^2 =& \|p\|_2^2 + \left(\sum_{k\in[K-1]} p(k)\right)^2\\
\leqslant& \|p\|_2^2 + (K-1)\sum_{k\in[K-1]} p(k)^2\\
=& K\|p\|_2^2.
\end{aligned}$$

The proof is the same for q and A. □

The next lemma will be used to link the norms of A and $A\mathbf{f}$.

Lemma 3.12. *For all $\bar{A} \in \mathbb{R}^{K\times K}$ and $\mathbf{f}^* \in (\mathbf{L}^2(\mathcal{Y},\mu))^K$,*

$$\sigma_K(G(\mathbf{f}^*))\|\bar{A}\|_F^2 \leqslant \sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_2^2 \leqslant K\|G(\mathbf{f}^*)\|_\infty \|\bar{A}\|_F^2$$

Proof. For the first inequality, we use that for all $k \in \mathcal{X}$,

$$\begin{aligned}\|(\bar{A}\mathbf{f}^*)_k\|_2^2 =& \bar{A}(k,\cdot)G(\mathbf{f}^*)\bar{A}(k,\cdot)^\top \\ \geqslant& \sigma_K(G(\mathbf{f}^*))\|\bar{A}(k,\cdot)\|_2^2\end{aligned}$$

and the inequality follows by summing over k.

For the second inequality,

$$\begin{aligned}\sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f}^*)_k\|_2^2 &= \sum_{k\in[K]} \int (\bar{A}\mathbf{f}^*)_k(x)^2 \mu(dx) \\ &= \sum_{k\in[K]} \int \left(\sum_{j\in[K]} \bar{A}(k,j) f_j^*(x)\right)^2 \mu(dx) \\ &\leqslant \sum_{k\in[K]} \int K \sum_{j\in[K]} \bar{A}(k,j)^2 (f_j^*)^2(x)\mu(dx) \\ &\leqslant K \left(\sum_{k,j\in[K]} \bar{A}(k,j)^2\right) \sup_{j\in\mathcal{X}} \int (f_j^*)^2(x)\mu(dx) \\ &= K\|\bar{A}\|_F^2 \|G(\mathbf{f}^*)\|_\infty.\end{aligned}$$

□

Finally, we will use the following result (shown in Section 2.B.2 or Chapter 2) in order to upper bound the spectrum of the matrix M.

Lemma 3.13. *For all $\pi_1, \pi_2 \in \Delta$, for all $\mathbf{Q}_1, \mathbf{Q}_2 \in \mathcal{Q}$ and for all $\mathbf{f}_1, \mathbf{f}_2 \in (\mathbf{L}^2(\mathcal{Y},\mu))^K$,*

$$\|g^{\pi_1,\mathbf{Q}_1,\mathbf{f}_1} - g^{\pi_2,\mathbf{Q}_2,\mathbf{f}_2}\|_2 \leqslant \sqrt{3K(\|G(\mathbf{f}_1)\|_\infty^3 \vee \|G(\mathbf{f}_2)\|_\infty^3)}\, d_{perm}((\pi_1,\mathbf{Q}_1,\mathbf{f}_1),(\pi_2,\mathbf{Q}_2,\mathbf{f}_2))$$

Together, these results imply that for all (p,q,A),

$$\begin{aligned}\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 & \\ \leqslant& 3K(\|G(\mathbf{f}^*+\bar{A}\mathbf{f}^*)\|_\infty^3 \vee \|G(\mathbf{f}^*)\|_\infty^3)(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}} \|(\bar{A}\mathbf{f})_k\|_F^2) \\ \leqslant& 3K\|G(\mathbf{f}^*)\|_\infty^3 (1+K^2\|A\|_F^2)^3 (K\|p\|_2^2 + K\|q\|_F^2 + K^2\|G(\mathbf{f}^*)\|_\infty \|A\|_F^2)\end{aligned}$$

so that $\sigma_1(M) \leqslant \sqrt{3K^3}(1 \vee \|G(\mathbf{f})\|_\infty^2)$. Since $H = \prod_{i=1}^{(K-1)(2K+1)} \sigma_i(M)$, one has

$$\sigma_{(K-1)(2K+1)}(M) \geqslant \frac{H}{(3K^3(1\vee\|G(\mathbf{f})\|_\infty^4))^{K^2-K/2}}.$$

Now, assume that **[Hid]** holds, so that $\sigma_K(G(\mathbf{f}^*)) > 0$, then

$$\begin{aligned}\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 &\geqslant \sigma_{(K-1)(2K+1)}(M)(\|p\|_2^2 + \|q\|_F^2 + \|A\|_F^2) \\ &\quad + o(\|p\|_2^2 + \|q\|_F^2 + \|A\|_F^2) \\ &\geqslant \frac{\sigma_{(K-1)(2K+1)}(M)}{1 \wedge K\|G(\mathbf{f}^*)\|_\infty}\left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}}\|(\bar{A}\mathbf{f}^*)_k\|_F^2\right) \\ &\quad + o\left(\frac{1}{1 \wedge \sigma_K(G(\mathbf{f}^*))}\left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}}\|(\bar{A}\mathbf{f}^*)_k\|_F^2\right)\right)\end{aligned}$$

and finally

$$\begin{aligned}&\|g^{\pi^*+\bar{p},\mathbf{Q}^*+\bar{q},\mathbf{f}^*+\bar{A}\mathbf{f}^*} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}^*}\|_2^2 \\ &\qquad \geqslant c_2(\pi^*,\mathbf{Q}^*,\mathbf{f}^*)\left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}}\|(\bar{A}\mathbf{f}^*)_k\|_F^2\right) \\ &\qquad\qquad + o\left(\|\bar{p}\|_2^2 + \|\bar{q}\|_F^2 + \sum_{k\in\mathcal{X}}\|(\bar{A}\mathbf{f}^*)_k\|_F^2\right) \quad (3.4)\end{aligned}$$

where

$$c_2(\pi^*,\mathbf{Q}^*,\mathbf{f}^*) = \frac{H}{(1 \wedge K\|G(\mathbf{f}^*)\|_\infty)(3K^3(1 \vee \|G(\mathbf{f}^*)\|_\infty^4))^{K^2-K/2}}$$

is positive as soon as **[Hid]** and **[Hdet]** hold.

3.B.4 Proof of Theorem 3.7

Let

$$N_{\mathbf{f}}(p,q,\mathbf{h}) = \|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{h}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2$$

and

$$\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2 = d_{\mathrm{perm}}((\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{h}),(\pi^*,\mathbf{Q}^*,\mathbf{f}))^2.$$

We want to show that there exists a constant $c^* > 0$ such that there exists a neighborhood $\mathcal{V}$ of $\mathbf{f}^*$ such that if one writes

$$c_{\mathbf{f}} := \inf_{p\in(\Delta-\Delta),\, q\in(\mathcal{Q}-\mathcal{Q}),\, \mathbf{h}\in(\mathcal{F}-\mathcal{F})^K} \frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2}$$

then $\inf_{\mathbf{f}\in\mathcal{V}} c_{\mathbf{f}} \geqslant c^*$.

The proof follows the structure of the proof of Theorem 6 of de Castro et al. (2016). It consists of three steps: the first one controls the component of $\mathbf{h}$ that is orthogonal to $\mathbf{f}$. This makes it possible to restrict $\mathbf{h}$ to the finite-dimensional space spanned by $\mathbf{f}$ in the two other parts. The second step controls the case when $\mathbf{h}$ is small, so that the behaviour of $N_{\mathbf{f}}$ is given by its quadratic form, and the last step controls the case where $\mathbf{h}$ is far from zero.

The orthogonal part

Let $\mathbf{u}$ be the orthogonal projection of $\mathbf{h}$ on $\mathrm{Span}(\mathbf{f})$. Then

$$N_{\mathbf{f}}(p,q,\mathbf{h}) = N_{\mathbf{f}}(p,q,\mathbf{u}) + M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{h}-\mathbf{u})$$

where

$$
\begin{aligned}
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) = \sum_{i_1,j_1,k_1} \sum_{i_2,j_2,k_2} & (\pi^* + p)(i_1)(\mathbf{Q}^* + q)(i_1,j_1)(\mathbf{Q}^* + q)(j_1,k_1) \\
& (\pi^* + p)(i_2)(\mathbf{Q}^* + q)(i_2,j_2)(\mathbf{Q}^* + q)(j_2,k_2) \\
& \Big(\langle a_{i_1}, a_{i_2} \rangle \langle (f+u)_{j_1}, (f+u)_{j_2} \rangle \langle (f+u)_{k_1}, (f+u)_{k_2} \rangle \\
& + \langle (f+u)_{i_1}, (f+u)_{i_2} \rangle \langle a_{j_1}, a_{j_2} \rangle \langle (f+u)_{k_1}, (f+u)_{k_2} \rangle \\
& + \langle (f+u)_{i_1}, (f+u)_{i_2} \rangle \langle (f+u)_{j_1}, (f+u)_{j_2} \rangle \langle a_{k_1}, a_{k_2} \rangle \\
& \quad + \langle a_{i_1}, a_{i_2} \rangle \langle a_{j_1}, a_{j_2} \rangle \langle (f+u)_{k_1}, (f+u)_{k_2} \rangle \\
& \quad + \langle a_{i_1}, a_{i_2} \rangle \langle (f+u)_{j_1}, (f+u)_{j_2} \rangle \langle a_{k_1}, a_{k_2} \rangle \\
& \quad + \langle (f+u)_{i_1}, (f+u)_{i_2} \rangle \langle a_{j_1}, a_{j_2} \rangle \langle a_{k_1}, a_{k_2} \rangle \\
& \qquad + \langle a_{i_1}, a_{i_2} \rangle \langle a_{j_1}, a_{j_2} \rangle \langle a_{k_1}, a_{k_2} \rangle \Big).
\end{aligned}
$$

Let us write Π' the matrix whose diagonal terms are the elements of $\pi^* + p$ and $\mathbf{Q}'$ the matrix $\mathbf{Q}^* + q$, then $M_{\mathbf{f}}$ can be written as

$$
\begin{aligned}
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) = \sum_{i,j} \Big(& ((\Pi'\mathbf{Q}')^\top G(\mathbf{a}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{f}+\mathbf{u})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \quad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \\
& \qquad + ((\Pi'\mathbf{Q}')^\top G(\mathbf{a}) \Pi'\mathbf{Q}')_{i,j} G(\mathbf{a})_{i,j} (\mathbf{Q}'^\top G(\mathbf{a})\mathbf{Q}')_{i,j} \Big).
\end{aligned}
$$

By the Schur product theorem, these terms are nonnegative since they correspond to Hadamard products of three Gram matrices which are nonnegative. Thus, one can lower bound $M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a})$ by the second term of the sum, which leads to

$$
M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) \geqslant \sum_{i,j=1}^{K} ((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u}) \Pi'\mathbf{Q}')_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j} \langle a_i, a_j \rangle
$$

Assume **[Hid]** holds for the parameters $(\pi^* + p, \mathbf{Q}^* + q, \mathbf{f} + \mathbf{u})$, then the matrices $(\Pi'\mathbf{Q}')^\top G(\mathbf{f} + \mathbf{u})\Pi'\mathbf{Q}'$ and $\mathbf{Q}'^\top G(\mathbf{f} + \mathbf{u})\mathbf{Q}'$ are positive symmetric with respective lowest eigenvalue lower bounded by $(\inf_k(\pi_k^* + p_k)\sigma_K(\mathbf{Q}^* + q))^2 \sigma_K(G(\mathbf{f}+\mathbf{u}))$ and $\sigma_K(\mathbf{Q}^* + q)^2 \sigma_K(G(\mathbf{f}+\mathbf{u}))$. Therefore, their Hadamard product is positive, and one has

$$
(((\Pi'\mathbf{Q}')^\top G(\mathbf{f}+\mathbf{u}) \Pi'\mathbf{Q}')_{i,j} (\mathbf{Q}'^\top G(\mathbf{f}+\mathbf{u})\mathbf{Q}')_{i,j})_{i,j} = (D\mathbf{U})^\top (D\mathbf{U})
$$

with $\mathbf{U}$ an orthogonal matrix and D a diagonal matrix with positive diagonal coefficients. Moreover, the Schur product theorem implies that $\sigma_K(D)^2 \geqslant (\inf_k(\pi_k^* + p_k))^2 \sigma_K(\mathbf{Q}^* + q)^4 \sigma_K(G(\mathbf{f} +$

$\mathbf{u}))^2$. Then

$$\begin{aligned}M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a}) \geqslant& \sum_{i,j=1}^{K}((D\mathbf{U})^\top(D\mathbf{U}))_{i,j}\langle a_i,a_j\rangle\\ =&\sum_{j=1}^{K}\|D\mathbf{U}\mathbf{a}\|_2^2\\ \geqslant&\sigma_K(D)^2\|\mathbf{U}\mathbf{a}\|_2^2\\ \geqslant&(\inf_k(\pi_k^*+p_k))^2\sigma_K(\mathbf{Q}^*+q)^4\sigma_K(G(\mathbf{f}+\mathbf{u}))^2\|\mathbf{a}\|_2^2.\end{aligned}$$

Finally, let $c_1(\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u})=(\inf_k(\pi_k^*+p_k))^2\sigma_K(\mathbf{Q}^*+q)^4\sigma_K(G(\mathbf{f}+\mathbf{u}))^2$. The application $(p,\pi^*,q,\mathbf{Q}^*,\mathbf{u},\mathbf{f})\mapsto c_1(\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u})$ is continuous and nonnegative, it is positive when **[Hid]** holds for the parameters $(\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u})$, and one has

$$M_{\mathbf{f}}(p,q,\mathbf{u},\mathbf{a})\geqslant c_1(\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u})\|\mathbf{a}\|_2^2.$$

We will now control the term $N_{\mathbf{f}}(p,q,\mathbf{u})$. Two cases appear: when $(\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u})$ is close to $(\pi^*,\mathbf{Q}^*,\mathbf{f}^*)$ in some sense and when it is not. The first case will be solved using the nondegeneracy of the quadratic form ensured by **[Hdet]**. The second case will be solved using the identifiability of the HMM.

In the neighborhood of $\mathbf{f}^*$.

The Taylor expansion of

$$(p,q,\mathbf{u})\in(\Delta-\Delta)\times(\mathcal{Q}-\mathcal{Q})\times((\mathcal{F}-\mathcal{F})\cap\mathrm{Span}(\mathbf{f}))^K\mapsto N_{\mathbf{f}}(p,q,\mathbf{u})$$

around $(0,0,0)$ leads to a nonnegative quadratic form and no linear part. **[Hdet]**, **[Hid]** and equation (3.4) ensure that this form is positive for $\mathbf{f}=\mathbf{f}^*$. Let $c_2(\mathbf{Q}^*,\pi^*,\mathbf{f})$ be as defined in Section 3.B.3, then $\mathbf{f}\mapsto c_2(\mathbf{Q}^*,\pi^*,\mathbf{f})$ is continuous and it is positive in the neighborhood of $\mathbf{f}^*$. Moreover, there exists a positive constant η depending on $\|G(\mathbf{f})\|_\infty$ such that for all $(p,q,\mathbf{u})$ such that $\|(p,q,\mathbf{u})\|_{\mathbf{f}}\leqslant 1$, one has

$$N_{\mathbf{f}}(p,q,\mathbf{u})\geqslant c_2(\mathbf{Q}^*,\pi^*,\mathbf{f})\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2-\eta\|(p,q,\mathbf{u})\|_{\mathbf{f}}^3.$$

For instance, $\eta=4000K^6\|G(\mathbf{f})\|_\infty^3$ works: the terms of order 2 or more in the Taylor expansion of $N_{\mathbf{f}}$ are the scalar product of sums of terms of the form $\sum_{i,j,k\in\mathcal{X}}\pi^*(i)\mathbf{Q}^*(i,j)\mathbf{Q}^*(j,k)f_i\otimes f_j\otimes f_k$ where zero to three of the f may be replaced by u, zero to two of the $\mathbf{Q}^*$ by q and π^* may be replaced by p and at least one of them is replaced. There are 63 possibilities, which leads to a sum of $(63K^3)^2$ terms, each of which can be bounded by $\|G(\mathbf{f})\|_\infty^3(\max\{p(i),q(i,j),\|u_i\|_2\mid i,j\in\mathcal{X}\})^r$ where r is the number of replaced terms. By taking the right permutation of states, the max can be bounded by $\|(p,q,\mathbf{u})\|_{\mathbf{f}}$, hence the result.

Then, using $\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2=\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2$ leads to

$$\begin{aligned}\frac{N_{\mathbf{f}}(p,q,\mathbf{h})}{\|(p,q,\mathbf{h})\|_{\mathbf{f}}^2}\geqslant&c_1(\mathbf{Q}^*+q,\pi^*+p,\mathbf{f}+\mathbf{u})\frac{\|\mathbf{a}\|_2^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2}\\ &+c_2(\mathbf{Q}^*,\pi^*,\mathbf{f})\frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2}-\eta\frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{(\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2)^{1/2}}\\ \geqslant&c_1(\mathbf{Q}^*+q,\pi^*+p,\mathbf{f}+\mathbf{u})\frac{\|\mathbf{a}\|_2^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2}\\ &+c_2(\mathbf{Q}^*,\pi^*,\mathbf{f})\frac{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2+\|\mathbf{a}\|_2^2}-\eta\sqrt{\|(p,q,\mathbf{u})\|_{\mathbf{f}}^2}\end{aligned}$$

Let $c_0 = \min(c_1/2, c_2)/2$, then c_0 is continuous and there exists a continuous function $(\pi^*, \mathbf{Q}^*, \mathbf{f}) \mapsto \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f})$ which is positive as soon as **[Hid]** and **[Hdet]** hold for $(\pi^*, \mathbf{Q}^*, \mathbf{f})$ and such that

$$\|(p, q, \mathbf{u})\|_{\mathbf{f}} \leqslant \epsilon(\pi^*, \mathbf{Q}^*, \mathbf{f}) \;\Rightarrow\; \frac{N_{\mathbf{f}}(p, q, \mathbf{h})}{\|(p, q, \mathbf{h})\|_{\mathbf{f}}^2} \geqslant c_0(\mathbf{Q}^*, \pi^*, \mathbf{f}).$$

Thus, there exists positive constants ϵ_0 and c_{near} depending on $\mathbf{Q}^*$, π^* and $\mathbf{f}^*$ such that

$$\forall (p, q, \mathbf{h}, \mathbf{f}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K$$
$$\text{s.t. } \|(p, q, \mathbf{u})\|_{\mathbf{f}} \leqslant \epsilon_0 \text{ and } \sum_{k \in \mathcal{X}} \|f_k - f_k^*\|_2^2 \leqslant \epsilon_0^2, \qquad \frac{N_{\mathbf{f}}(p, q, \mathbf{h})}{\|(p, q, \mathbf{h})\|_{\mathbf{f}}^2} \geqslant c_{\text{near}}.$$

Far from $\mathbf{f}^*$.

Lemma 3.14. *The application*

$$(p, q, \mathbf{u}, \mathbf{f}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K \longmapsto N_{\mathbf{f}}(p, q, \mathbf{u})$$

restricted to the set of $(p, q, \mathbf{u}, \mathbf{f})$ such that $\mathbf{u} \in Span(\mathbf{f})^K$ is uniformly continuous for the norm $\|\cdot\|_{tot}$ defined by

$$\|(p, q, \mathbf{u}, \mathbf{f})\|_{tot}^2 := \|p\|_2^2 + \|q\|_F^2 + \sum_{k \in \mathcal{X}} \left(\|u_k\|_2^2 + \|f_k\|_2^2\right).$$

Thus, by compactness of $(\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(\mathbf{f}))^K$, the application

$$c_{\text{far}} : \mathbf{f} \longmapsto \inf_{(p,q,\mathbf{u}) \in (\Delta-\Delta)\times(\mathcal{Q}-\mathcal{Q})\times((\mathcal{F}-\mathcal{F})\cap \text{Span}(\mathbf{f}))^K \text{ s.t. } \|(p,q,\mathbf{u})\|_{\mathbf{f}} > \epsilon_0} N_{\mathbf{f}}(p, q, \mathbf{u})$$

is continuous. Let us now prove that $c_{\text{far}}(\mathbf{f}^*) > 0$.

Let $(p_n, q_n, \mathbf{u}_n)_n \in ((\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times ((\mathcal{F} - \mathcal{F}) \cap \text{Span}(\mathbf{f}^*))^K)^{\mathbb{N}}$ be a sequence such that $\|(p_n, q_n, \mathbf{u}_n)\|_{\mathbf{f}^*} > \epsilon_0$ for all n and

$$c_{\text{far}}(\mathbf{f}^*) = \lim_n N_{\mathbf{f}^*}(p_n, q_n, \mathbf{u}_n).$$

By compactness, this sequences converges towards a limit $(p, q, \mathbf{u})$ up to taking a subsequence. Necessarily $\|(p, q, \mathbf{u})\|_{\mathbf{f}^*} \geqslant \epsilon_0$. Since **[Hid]** holds, Theorem 3.5 shows that $N_{\mathbf{f}^*}(p, q, \mathbf{u}) > 0$, which implies $c_{\text{far}}(\mathbf{f}^*) > 0$ by continuity of $N_{\mathbf{f}^*}$. Note that $c_{\text{far}}(\mathbf{f}^*)$ may depend on $\mathcal{F}$ in addition to the parameters π^*, $\mathbf{Q}^*$ and $\mathbf{f}^*$.

Thus, by continuity, there exists $\epsilon_1 > 0$ such that for all $\mathbf{f} \in \mathcal{F}^K$ such that $\sum_{k \in \mathcal{X}} \|f_k - f_k^*\|_2^2 \leqslant \epsilon_1^2$, $c_{\text{far}}(\mathbf{f}) \geqslant c_{\text{far}}(\mathbf{f}^*)/2$.

Finally, **[HF]** implies that there exists a constant $\mathcal{C}$ depending only on $C_{\mathcal{F},2}$ such that $\|(p, q, \mathbf{u})\|_{\mathbf{f}}^2 \leqslant \|(p, q, \mathbf{h})\|_{\mathbf{f}}^2 \leqslant \mathcal{C}$ for all $(p, q, \mathbf{h}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K$. Therefore,

$$\forall (p, q, \mathbf{h}, \mathbf{f}) \in (\Delta - \Delta) \times (\mathcal{Q} - \mathcal{Q}) \times (\mathcal{F} - \mathcal{F})^K \times \mathcal{F}^K$$
$$\text{s.t. } \|(p, q, \mathbf{u})\|_{\mathbf{f}} \geqslant \epsilon_0 \text{ and } \sum_{k \in \mathcal{X}} \|f_k - f_k^*\|_2^2 \leqslant \epsilon_1^2, \qquad \frac{N_{\mathbf{f}}(p, q, \mathbf{h})}{\|(p, q, \mathbf{h})\|_{\mathbf{f}}^2} \geqslant \frac{N_{\mathbf{f}}(p, q, \mathbf{u})}{\mathcal{C}}$$
$$\geqslant \frac{c_{\text{far}}(\mathbf{f}^*)}{2\mathcal{C}}.$$

The theorem follows by taking $c^*(\pi^*, \mathbf{Q}^*, \mathbf{f}^*, \mathcal{F}) = \min\left(\frac{c_{\text{far}}(\mathbf{f}^*)}{2\mathcal{C}}, c_{\text{near}}\right)$ and the neighborhood containing all $\mathbf{f} \in \mathcal{F}^K$ such that $\sum_{k \in \mathcal{X}} \|f_k - f_k^*\|_2^2 \leqslant \min(\epsilon_0, \epsilon_1)^2$. Moreover, $(\pi, \mathbf{Q}, \mathbf{f}) \longmapsto c^*(\pi, \mathbf{Q}, \mathbf{f}, \mathcal{F})$ is lower bounded by this value in a neighborhood of $(\pi^*, \mathbf{Q}^*, \mathbf{f}^*)$, so that it can be assumed to be lower semicontinuous.

Note that the dependency of c^* on $\mathcal{F}$ appears during this last step and is made non explicit because of the compactness assumption.

Proof of Lemma 3.14

$$\begin{aligned}
&\Big|N_{\mathbf{f}}(p,q,\mathbf{u}) - N_{\mathbf{f}'}(p',q',\mathbf{u}')\Big| \\
&\quad = \Big| \|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}}\|_2^2 - \|g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}\|_2^2 \Big| \\
&\quad \leqslant 2\|g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'}\|_2^2 + 2\|g^{\pi^*,\mathbf{Q}^*,\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}\|_2^2 \\
&\qquad + 2\left|\left\langle g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}, g^{\pi^*+p,\mathbf{Q}^*+q,\mathbf{f}+\mathbf{u}} - g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} \right\rangle\right| \\
&\qquad + 2\left|\left\langle g^{\pi^*+p',\mathbf{Q}^*+q',\mathbf{f}'+\mathbf{u}'} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'}, g^{\pi^*,\mathbf{Q}^*,\mathbf{f}} - g^{\pi^*,\mathbf{Q}^*,\mathbf{f}'} \right\rangle\right|
\end{aligned}$$

Then, using the fact that $\|g^{\pi,\mathbf{Q},\mathbf{f}} - g^{\pi',\mathbf{Q}',\mathbf{f}'}\|_2 \leqslant \sqrt{3K}C^3_{\mathcal{F},2}\|(\pi-\pi',\mathbf{Q}-\mathbf{Q}',\mathbf{f}-\mathbf{f}',0)\|_{\mathrm{tot}}$ (see Lemma 3.13), that $\|g^{\pi,\mathbf{Q},\mathbf{f}}\|_2 \leqslant C^3_{\mathcal{F},2}$ (see for instance Lemma 29 of Lehéricy (2019)) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
\Big|N_{\mathbf{f}}(p,q,\mathbf{u}) - N_{\mathbf{f}'}(p',q',\mathbf{u}')\Big| \leqslant\, & 6KC^6_{\mathcal{F},2}\|(p-p',q-q',\mathbf{f}+\mathbf{u}-\mathbf{f}'-\mathbf{u}',0)\|^2_{\mathrm{tot}} \\
& + 6KC^6_{\mathcal{F},2}\|(0,0,0,\mathbf{f}-\mathbf{f}')\|^2_{\mathrm{tot}} \\
& + 4\sqrt{3K}C^6_{\mathcal{F},2}\|(p-p',q-q',\mathbf{f}+\mathbf{u}-\mathbf{f}'-\mathbf{u}',0)\|_{\mathrm{tot}} \\
& + 4\sqrt{3K}C^6_{\mathcal{F},2}\|(0,0,0,\mathbf{f}-\mathbf{f}')\|^2_{\mathrm{tot}} \\
\leqslant\, & 24KC^6_{\mathcal{F},2}\Big(\|(p-p',q-q',\mathbf{u}-\mathbf{u}',\mathbf{f}-\mathbf{f}')\|^2_{\mathrm{tot}} \\
& + \|(p-p',q-q',\mathbf{u}-\mathbf{u}',\mathbf{f}-\mathbf{f}')\|_{\mathrm{tot}}\Big),
\end{aligned}$$

which proves the uniform continuity of the application.

CHAPTER

4

MISSPECIFIED MODELS

This chapter is based on the submitted paper Lehéricy (2018b).

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

4.1 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 chapter 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 chapter, 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. (2016a); Lehéricy (2018a) and least squares estimators in de Castro et al. (2016); Lehéricy (2018a). 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 chapter is an oracle inequality that holds as soon as the models have controlled tails. This bound is optimal when the true distribution is a HMM taking values in $\mathbb{R}$. Let us give some details about this result.

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

To approximate a large class of true distributions, we consider nonparametric HMMs, where the parameters are not described by a finite dimensional space. For instance, one may consider HMMs with arbitrary number of states and arbitrary emission distributions. Computing a maximizer of the likelihood directly in a nonparametric model may be hard or result in overfitting. The model selection approach offers a way to circumvent this problem. It consists in considering a countable family of parametric sets $(S_M)_{M\in\mathcal{M}}$–the *models*–and selecting one of them. The larger the union of all models, the more distributions are approximated. Several criteria can be used to select the model, such as bootstrap, cross validation (see for instance Arlot and Celisse (2010)) or penalization (see for instance Massart (2007)). We use a penalized criterion, which

consists in maximizing the function

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

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

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

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

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

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

A simplified version of our main result (Theorem 4.8) is the following oracle inequality: for all $\alpha \geqslant 1$, there exists constants A and n_0 such that if the penalty is large enough, the penalized maximum likelihood estimator $\hat{\theta}_n$ satisfies for all $t \geqslant 1$, $\eta \in (0,1)$ and $n \geqslant n_0$, with probability larger than $1 - e^{-t} - n^{-\alpha}$:

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

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

In the case where the true distribution is a HMM, it is possible to quantify the approximation error $\inf_{\theta \in S} \mathbf{K}(\theta)$. Using the results of Kruijer et al. (2010), we show that the above oracle inequality is optimal in the minimax sense–up to logarithmic factors–for real-valued HMMs, see Corollary 4.11. This is done by taking HMMs whose emission densities are mixtures of exponential power distributions–which include Gaussian mixtures as a special case.

The chapter is organized as follows. We detail the framework of the chapter in Section 4.2. In particular, Section 4.2.3 describes the assumptions on the true distribution, Section 4.2.4 presents the assumptions on the model and Section 4.2.5 introduces the Kullback Leibler criterion used in the oracle inequality. Our main results are stated in Section 4.3. Section 4.3.1 contains the oracle inequality and Section 4.3.2 shows how it can be used to show minimax adaptivity for real-valued HMMs. Section 4.4 lists some perspectives for this work.

One may wish to relax our assumptions depending on the setting. For instance, one could want to change the dependency of the functions $B(n)$ and $\sigma_-(n)$ on n, change the tail conditions

or the rate of forgetting. We give an overview of the key steps of the proof of our oracle inequality in Section 4.5 to make it easier to adapt our result.

Some proofs are postponed the Appendices. Appendix 4.A contains the proof of the minimax adaptivity result and Appendix 4.B contains the proof of the main technical lemma of Section 4.5.

4.2 Notations and assumptions

4.2.1 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.

4.2.2 The model selection estimator

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

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

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

In both cases, we took a lower bound on the coefficients of the transition matrix $\mathbf{Q}$ that tends to zero when the number of observations grows. This allows to estimate parameters for

which some coefficients of the transition matrix are small or zero. We prove the choice $(\log n)^{-1}$ to be a good choice in general in Theorem 4.8.

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

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

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

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

for some penalty pen_n to be chosen.

In the following, we use the following notations.

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

4.2.3 Assumptions on the true distribution

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

Forgetting and mixing

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

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

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

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

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

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

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

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

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

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

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

Lemma 4.1. *Assume that $(Y_t)_t$ is generated by a HMM with a compact metric state space $\mathcal{X}$ (not necessarily finite) endowed with a Borel probability measure μ. Write $\mathcal{Q}^*$ its transition kernel and assume that $\mathcal{Q}^*$ admits a density with respect to μ that is uniformly lower bounded and upper bounded by positive and finite constants σ_-^* and σ_+^*. Write $(\gamma_x^*)_{x\in\mathcal{X}}$ its emission densities and assume that they satisfy $\int \gamma_x^*(y)\mu(dx) \in (0,+\infty)$ for all $y \in \mathcal{Y}$.*

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

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

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

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

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

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

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

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

Extreme values of the true density

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

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

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

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

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

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

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

Lemma 4.2. *Assume that there exists $\delta > 0$ such that*

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

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

4.2.4 Model assumptions

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

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

Transition kernel

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

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

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

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

Tail of the emission densities

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

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

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

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

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

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

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

Lemma 4.3. *Assume that one of the two following assumption holds:*

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

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

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

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

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

Changing the densities as in the lemma amounts to adding a mixture component (with weight n^{-a} and distribution λ) to the emission densities to make sure that they are uniformly lower bounded. We shall see in the following that if $a \geqslant 1$, then this additional component changes nothing to the approximation properties of the models, see the proof of Corollary 4.11. This is in agreement with the fact that this component is asymptotically never observed as soon as $a > 1$.

Complexity of the approximation spaces

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

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

$$N\left(\left\{ y \mapsto \gamma_x(y) \mathbf{1}_{\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(y)| \leqslant D} \right\}_{\gamma \in S_{K,M,n}^{(\gamma)}, x \in [K]}, d_\infty, \delta \right) \leqslant \max\left(\frac{C_{\text{aux}}}{\delta}, 1 \right)^{m_M}, \quad (4.1)$$

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

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

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

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

A typical way to check **[Aentropy]** is to use a parametrization of the emission densities, for instance a lipschitz application $[-1,1]^{m_M} \longrightarrow S^{(\gamma)}_{K,M,n}$. This reduces the construction of a bracket covering on $S^{(\gamma)}_{K,M,n}$ to the construction of a bracket covering of the unit ball of $\mathbb{R}^{m_M}$. In this case, C_{aux} depends on the lipschitz constant of the parametrization. An example of this approach is given in Section 4.3.2 for mixtures of exponential power distributions.

4.2.5 Limit and properties of the log-likelihood

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

Convergence of the log-likelihood

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

Lemma 4.4 (Barron (1985))**.** *Assume that the process $(Y_t)_{t\geqslant 1}$ is ergodic, then there exists a quantity $l^* > -\infty$ such that*

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

and

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

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

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

Lemma 4.5 (Leroux (1992))**.** *Let K be a positive integer, γ a vector of K nonnegative and measurable functions, $\mathbf{Q}$ a transition matrix of size K and π a probability measure on $[K]$.*

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

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

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

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

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

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

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

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

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

Note that when γ is a vector of probability densities, $\mathbf{K}(K, \mathbf{Q}, \gamma) \geqslant 0$ since it is the limit of a sequence of Kullback-Leibler divergences: under the assumptions of Lemma 4.5, if $\inf_{x \in [K]} \pi(x) > 0$,

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

Approximation of the limit

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

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

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

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

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

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

Lemma 4.6.

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

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

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

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

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

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

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

- *if **[Atail]** holds, then for all $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$, $l(K, \mathbf{Q}, \gamma)$ exists, is finite and*

$$l(K, \mathbf{Q}, \gamma) = \mathbb{E}^*[L_{1,\infty}(K, \mathbf{Q}, \gamma)];$$

- *if **[A⋆tail]** holds, then l^* exists and is finite and*

$$l^* = \mathbb{E}^*[L^*_{1,\infty}].$$

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

The third point follows from the ergodicity of $(Y_t)_{t\geqslant 1}$ under **[A⋆mixing]**, from the integrability of $L_{i,\infty}$ and $L^*_{i,\infty}$ under **[Atail]** and **[A⋆tail]** and from Lemmas 4.4 and 4.5. □

Note that under the assumptions of point 3 of Lemma 4.6, one has $\mathbf{K}(K, \mathbf{Q}, \gamma) = \mathbb{E}[L^*_{1,\infty} - L_{1,\infty}(K, \mathbf{Q}, \gamma)] \in [0, +\infty)$ for all $(K, \pi, \mathbf{Q}, \gamma) \in \mathbf{S}_n$ (recall that γ is a vector of probability densities in this case), or with some notation abuses:

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

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

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

Lemma 4.7 (Alexandrovich et al. (2016), Theorem 5)**.** *Assume $(Y_t)_t$ is generated by a finite state space HMM with parameters $(K^*, \pi^*, \mathbf{Q}^*, \gamma^*)$. Assume $\mathbf{Q}^*$ is invertible and ergodic, that the emission densities $(\gamma^*_x)_{x\in[K^*]}$ are all distinct and that $\mathbb{E}^*\left[(\log \gamma^*_x(Y_1))^+\right] < \infty$ for all $x \in [K^*]$ (so that $l^* < \infty$).*

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

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

4.3 Main results

4.3.1 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 4.8. *Assume **[A⋆forgetting]**, **[A⋆mixing]**, **[A⋆tail]**, **[Aergodic]**, **[Atail]**, **[Aentropy]** and **[Agrowth]** hold.*

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

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

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

and let

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

be the nonparametric maximum likelihood estimator.

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

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

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

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

as soon as

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

The proof of this theorem is presented in Section 4.5. Its structure and main steps are detailed in Section 4.5.1, and the proof of these steps are gathered in Section 4.5.2.

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

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

- The limitation of the dependency to the $\log n$ most recent observations, which induces a factor $(\log n)^2$;
- The dependency of $\sigma_-(n)$ and $B(n)$ on n, each of them at the root of a factor $(\log n)^2$;
- Truncating the emission densities (possible thanks to assumptions **[Atail]** and **[A⋆tail]**), which induces a factor $(\log n)^{2q}$;

- The use of a Bernstein inequality for exponentially α-mixing processes, which introduces a factor $(\log n)^2$ compared to a Bernstein inequality for independent variables. However, together with the previous point (the truncation of the emission densities), the two points only induce a factor $(\log n)^{1+q}$.

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

4.3.2 Minimax adaptive estimation using location-scale mixtures

In this section, we show that the oracle inequality of Theorem 4.8 allows to construct an estimator that is adaptive and minimax up to logarithmic factors when the observations are generated by a finite state space hidden Markov model. To do so, we consider models whose emission densities are finite mixtures of exponential power distributions, and use an approximation result by Kruijer et al. (2010).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lemma 4.9 (Checking the assumptions)**.** *Assume* $\inf \mathbf{Q}^* > 0$*, then:*

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

Proof. The first point follows from Lemma 4.1.

The second point follows from the fact that the densities γ_x^* are uniformly bounded under **(C3)** and by taking δ large enough in Lemma 4.2.

[Aergodic] holds by definition of the models.

See Section 4.A.1 for the proof of the last two points. □

Remark. *One can also take* $(s_{x,1}, \dots, s_{x,M}) \in [\frac{1}{n}; n]^M$*, in which case Lemma 4.9 holds by taking* $B(n) = 6 \log n$ *and* $C_{aux} = 2pn^4$*.*

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

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

Lemma 4.10 (Approximation rates)**.** *Assume* ***(C1)-(C4)*** *hold. Then there exists a sequence of mixtures* $(g_{M,x})_M$ *such that* $n^{-2} + (1 - n^{-2}) g_{M,x} \in S^{(\gamma)}_{K^*,M,n}$ *for all* $n \geqslant M$ *and*

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

Proof. Proof in Section 4.A.1. □

Corollary 4.11 (Minimax adaptive estimation rates)**.** *Assume* ***(C1)-(C4)*** *hold. Also assume that* $\inf \mathbf{Q}^* > 0$*. Then there exists a constant* $C > 0$ *such that for all* $M \geqslant 3$ *and* $n \geqslant M$*,*

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

Hence, using Theorem 4.8 with $pen_n(K, M) = (KM+K^2)(\log n)^{15}/n$*, there exists a constant* C *such that almost surely, there exists a (random)* n_0 *such that*

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

Proof. Proof in Section 4.A.1. □

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

4.4 Perspectives

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

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

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

4.5 Proof of the oracle inequality (Theorem 4.8)

4.5.1 Overview of the proof

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

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

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

Let

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

then

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

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

$$|\nu(K, \pi, \mathbf{Q}, \gamma)| - \mathrm{pen}_n(K, M) \leqslant \eta\mathbf{K}(K, \mathbf{Q}, \gamma) + R_n \tag{4.2}$$

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 (4.2). The following remark will be useful in our proofs: since

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

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

$$\begin{aligned}L_{i,k,\mu}(K, \mathbf{Q}, \gamma) \in & \left[\log \sigma_-(n) + \log \sum_{x \in [K]} \gamma_x(Y_i); \log \sum_{x \in [K]} \gamma_x(Y_i)\right] \\ =& [\log \sigma_-(n) + b_\gamma(Y_i); b_\gamma(Y_i)]\end{aligned} \tag{4.3}$$

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

$$\begin{cases} |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L_{i,k',\mu'}(K', \mathbf{Q}', \gamma')| \leqslant \log \frac{1}{\sigma_-(n)} + |b_\gamma(Y_i)| + |b_{\gamma'}(Y_i)| \\ |L_{i,k,\mu}(K, \mathbf{Q}, \gamma) - L_{i,k'}^*| \leqslant \log \frac{1}{\sigma_-(n)} + |b_\gamma(Y_i)| + |L_{i,k'}^*| \end{cases}. \tag{4.4}$$

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

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

where $D>0$ and

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

for a fixed $x\in[K]$. Note that $\|t^{(D)}_{(K,\mathbf{Q},\gamma)}\|_\infty \leqslant 2D + \log\frac{1}{\sigma_-(n)}$ thanks to equation (4.3).

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

Lemma 4.12. *Assume **[Atail]**, **[Aergodic]**, **[A⋆tail]** and **[A⋆forgetting]** hold. Also assume $B(n)\geqslant B^*$ and $\sigma_-(n)\leqslant\frac{1-\rho_*}{2-\rho_*}\wedge\frac{1}{1+C_*}$. Then for all $u\geqslant 1$, with probability greater than $1-2ne^{-u}$, for all $(K,\pi,\mathbf{Q},\gamma)\in\mathbf{S}_n$,*

$$\left|\nu(K,\pi,\mathbf{Q},\gamma) - \bar{\nu}_k(t^{(B(n)u^q)}_{(K,\mathbf{Q},\gamma)})\right| \leqslant \left(6B(n)u^q + \log\frac{1}{\sigma_-(n)}\right)e^{-u} + \frac{2}{n\rho(1-\rho)^2} + \frac{4\rho^{k-1}}{1-\rho} \quad (4.5)$$

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

Proof. Proof in Section 4.5.2. □

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

Theorem 4.13. *Assume **[Aergodic]**, **[Aentropy]** and **[A⋆mixing]**. Also assume that there exists n_1 such that for all $n\geqslant n_1$, for all $K\leqslant n$ and M such that $m_M\leqslant n$,*

$$6900\pi\,(m_M K + K^2 - 1)ke^{-4D}(\log n)^3\,(k+\log C_{aux}) \leqslant n. \quad (4.6)$$

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

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

$$pen_n(K,M) \geqslant \frac{C_{pen}}{n}(n_*+k+1)^2\Bigg[\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2(m_M K+K^2-1) \times\left(\frac{1}{\epsilon}\vee\frac{\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_*+k+1}\right)\left(\log n + k\log\frac{2}{\sigma_-(n)} + D + \log C_{aux}\right) + \left(\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2 + \left(\frac{1}{\epsilon}\vee\frac{\left(D+\log\frac{1}{\sigma_-(n)}\right)(\log n)^2}{n_*+k+1}\right)\right)w_M\Bigg]. \quad (4.7)$$

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

$$|\bar{\nu}_k(t^{(D)}_{(K,\mathbf{Q},\gamma)})| - pen_n(K, M) \leqslant \epsilon \mathbb{E}[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^0_{-k})^2] + A(n_* + k + 1)^2 \left(\left(D + \log \frac{1}{\sigma_-(n)} \right) (\log n)^2 + \frac{1}{\epsilon} \vee \frac{\left(D + \log \frac{1}{\sigma_-(n)} \right) (\log n)^2}{n_* + k + 1} \right) \frac{s}{n}. \quad (4.8)$$

Proof. Proof in Section 4.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 4.14. *Assume **[Atail]**, **[Aergodic]**, **[A⋆tail]** and **[A⋆forgetting]** hold. Also assume that $B(n) \geqslant B^*$ and $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*} \wedge e^{-2}$. Then for all k such that*

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

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

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

Proof. Proof in Section 4.5.2. □

Now that the main lemmas have been stated, let us show how the assumptions of Theorem 4.8 leads to the desired oracle inequality.

Let C_σ and C_B be two positive constants and let

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

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

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

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

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

In order to get $\epsilon \mathbb{E}^*[t^{(D)}_{(K,\mathbf{Q},\gamma)}(Y^i_{i-k})^2] \leqslant \eta \mathbf{K}(K, \mathbf{Q}, \gamma) + \frac{22\eta}{n}$ using Lemma 4.14, one needs

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

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

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

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

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

and therefore

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

Thus, there exists an integer n_0'' depending on C_σ, C_B and α such that for all $n \geqslant n_0'' \vee \exp(C_\sigma((1+C_*) \vee \frac{2-\rho_*}{1-\rho_*} \vee e^2)) \vee \exp(\frac{B^*}{C_B}) \vee \exp\sqrt{\frac{C_\sigma}{2}(n_*+1)}$ (so that $k = \frac{2}{C_\sigma}(\log n)^2 \geqslant n_* + 1$, $B(n) \geqslant B^*$ and $\sigma_-(n) \leqslant \frac{1-\rho_*}{2-\rho_*} \wedge \frac{1}{1+C_*} \wedge e^{-2}$), equation (4.7) is implied by

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

such that equation (4.8) (combined with Lemma 4.14) implies

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

such that equation (4.5) implies

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

and such that when **[Agrowth]** holds and when $m_M \leqslant n$ and $K \leqslant n$, equation (4.6) is implied by

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

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

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

i.e.

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

which holds for all $n \geqslant n_0''$ (up to modification of n_0'') when $C_B \geqslant 1 + \frac{\zeta}{4}$. Putting these equations together proves Theorem 4.8.

4.5.2 Proofs

Upper bounds for the moments of the tails

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

For $u \geqslant 0$, let

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

Lemma 4.15. *For all $u \geqslant 1$,*

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

Proof. One has

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

Likewise,

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

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

Proof of Lemma 4.12 (approximating the likelihood)

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

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

one gets using Lemma 4.6 and **[A⋆forgetting]** that

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

as soon as

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

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

Then, note that

$$\begin{aligned}\bar{\nu}_k(t_{(K,\mathbf{Q},\gamma)}) - \bar{\nu}_k(t_{(K,\mathbf{Q},\gamma)}^{(B(n)u^q)}) = \frac{1}{n}\sum_{i=1}^{n} &t_{(K,\mathbf{Q},\gamma)}(Y_{i-k}^i)\mathbf{1}_{|L_{i,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_i)|)>B(n)u^q}\\ &- \mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)\mathbf{1}_{|L_{0,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|)>B(n)u^q}].\end{aligned}$$

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

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

Moreover,

$$|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t_{(K,\mathbf{Q},\gamma)}^{(B(n)u^q)}(Y_{-k}^0)]| = \mathbb{E}^*[|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)|\mathbf{1}_{|L_{0,k}^*|\vee(\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|)>B(n)u^q}].$$

Equation (4.3) ensures that $|t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0)| \leqslant |L_{0,k}^*| + \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)| + \log\frac{1}{\sigma_-(n)}$, so that

$$\begin{aligned}&|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t_{(K,\mathbf{Q},\gamma)}^{(B(n)u^q)}(Y_{-k}^0)]|\\ &\leqslant \mathbb{E}^*\left[|L_{0,k}^*|\left(\mathbf{1}_{|L_{0,k}^*|>B(n)u^q} + \mathbf{1}_{|L_{0,k}^*|\leqslant B(n)u^q < \sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|}\right)\right]\\ &\quad + \mathbb{E}^*\left[\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\left(\mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)u^q} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|\leqslant B(n)u^q<|L_{0,k}^*|}\right)\right]\\ &\quad + \mathbb{E}^*\left[\left(\log\frac{1}{\sigma_-(n)}\right)\left(\mathbf{1}_{|L_{0,k}^*|>B(n)u^q} + \mathbf{1}_{\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}|b_{\gamma'}(Y_0)|>B(n)u^q}\right)\right]\end{aligned}$$

[Atail] and **[A⋆tail]** imply that $\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} \left|b_{\gamma'}(Y_0)\right| / B(n)$ and $|L_{0,k}^*|/B^*$ can be upper bounded by the random variable W defined in Section 4.5.2, which means that for all $u \geqslant 1$,

$$\begin{aligned}
|\mathbb{E}^*[t_{(K,\mathbf{Q},\gamma)}(Y_{-k}^0) - t_{(K,\mathbf{Q},\gamma)}^{(B(n)u^q)}(Y_{-k}^0)]| & \\
&\leqslant B^* E_q(u) + B(n)u^q \mathbb{P}^* \left(\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| > B(n)u^q \right) \\
&\quad + B(n)E_q(u) + B(n)u^q \mathbb{P}^*(|L_{0,k}^*| > B(n)u^q) \\
&\quad + \left(\log \frac{1}{\sigma_-(n)} \right) \left(\mathbb{P}^* \left(\sup_{\gamma' \in \mathbf{S}_n^{(\gamma)}} |b_{\gamma'}(Y_0)| > B(n)u^q \right) \right. \\
&\qquad \left. + \mathbb{P}^*(|L_{0,k}^*| > B(n)u^q) \right) \\
&\leqslant 6B(n)u^q e^{-u} + 2 \left(\log \frac{1}{\sigma_-(n)} \right) e^{-u}
\end{aligned}$$

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

Proof of Lemma 4.14 (controlling the variance residual)

Lemma 4.16. *Assume **[Atail]**, **[Aergodic]** and **[A⋆tail]** hold. Assume $\sigma_-(n) \leqslant e^{-2}$ and let*

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

Then for all $v \geqslant 1$,

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

Proof. We need the following lemma :

Lemma 4.17 (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:

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

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 (4.4), one gets

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

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

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

so that

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

[Atail] and **[A⋆tail]** imply that $\sup_{\gamma'\in\mathbf{S}_n^{(\gamma)}}\left|b_{\gamma'}(Y_0)\right|/B(n)$ and $|L^*_{0,\infty}|/B^*$ can be upper bounded by the random variable W defined in Section 4.5.2, which means that for all $v \geqslant 1$,

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

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

Therefore, for all $v \geqslant 1$ such that the real number λ defined by $2B(n)v^q = \log\frac{1}{\lambda} - \log\frac{1}{\sigma_-(n)}$

satisfies $\lambda \leqslant e^{-4}$ (i.e. $2B(n)v^q \geqslant 4 - \log \frac{1}{\sigma_-(n)}$, which holds as soon as $v \geqslant 1$ and $\sigma_-(n) \leqslant e^{-1}$),

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

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

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

one gets

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

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

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

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

Then,

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

using Lemma 4.6, equation (4.4), Lemma 4.15, $B(n) \geqslant B^*$ and the condition on $\sigma_-(n)$ (which

implies $\rho \geqslant \rho_*$ and $\frac{1}{1-\rho} \geqslant C_*$). Therefore, under the assumptions of Lemma 4.16, one has

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

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

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

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

$$k \geqslant \frac{1}{\sigma_-(n)}\left(\log n + 2\log\frac{1}{\sigma_-(n)}\right) \tag{4.9}$$

using that $\log\log x \leqslant (\log x)/e$ for all $x > 1$ and that $e(1-1/e) \geqslant 1$. Therefore, for all k satisfying equation (4.9), for all $D > 0$ and $v \geqslant \log n$,

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

which concludes the proof.

Acknowledgements

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

APPENDICES

4.A Proofs for the minimax adaptive estimation

4.A.1 Proofs for the mixture framework

Proof of Lemma 4.9 (checking the assumptions)

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

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

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

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

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

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

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

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

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

as soons as $n \geqslant 10$.

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

Let us now check that this constant L_p exists.

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

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

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

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

Proof of Lemma 4.10 (approximation rates)

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

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

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

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

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

Proof of Corollary 4.11 (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 4.6 for the definition of these conditional distributions).

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

Lemma 4.18. *Assume that **[A⋆tail]**, **[A⋆forgetting]**, **[Atail]** and **[Aergodic]** hold with $B(n) = C_B \log n$ and $\sigma_-(n) = C_\sigma(\log n)^{-1}$.*

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

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

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

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

Take $p = p^*_{Y_1|Y^0_{-\infty}}$ and $q = p_{Y_1|Y^0_{-\infty},(K,\mathbf{Q},\gamma)}$. Then $\log\frac{p}{q} \leqslant |b_\gamma| + |L^*_{1,\infty}| + \log\frac{1}{\sigma_-(n)}$ and $\mathbf{1}\left\{\log\frac{p}{q} \geqslant v\right\} \leqslant \mathbf{1}\{|b_\gamma| \geqslant \frac{1}{2}(v - \log\frac{1}{\sigma_-(n)})\} \vee \mathbf{1}\{|L^*_{1,\infty}| \geqslant \frac{1}{2}(v - \log\frac{1}{\sigma_-(n)})\}$. Taking $v = 2C_B(\log n)^2$, one gets that there exists n_1 depending only on C_B and C_σ such that for all $n \geqslant n_1$, $v + \log\sigma_-(n) \geqslant (C_B\log n)^2$ and $1 + 2v \leqslant 5C_B(\log n)^2$, so that

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

Note that **[A⋆tail]** also holds for $L^*_{1,\infty}$ using the uniform convergence of Lemma 4.6. This implies that $\mathbb{P}^*(|L^*_{1,\infty}| \geqslant C_B(\log n)^2) \leqslant \exp(-\log n) \leqslant n^{-1}$ since $C_B(\log n) \geqslant B^*$ for $n \geqslant \exp(\frac{B^*}{C_B})$. Likewise, **[Atail]** implies that $\mathbb{P}^*(|b_\gamma| \geqslant C_B(\log n)^2) \leqslant n^{-1}$.

The last expectation of the above equation can be written as

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

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

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

using $C_B \log n \geqslant B^*$ and Lemma 4.15 for the first term and **[Atail]** for the second one. Likewise,

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

so that finally

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

which concludes the proof. □

Let $M \in \mathbb{N}^*$. Let $g_{M,x}$ be the approximating densities given by Lemma 4.10 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 4.19. *Assume $\sigma_-(n) \leqslant \inf \mathbf{Q}^*$, then*

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

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

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

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

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

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

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

Therefore, using the Cauchy-Schwarz inequality:

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

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

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

so that

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

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 $c\geqslant 0$, one has 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,

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

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

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

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

The choice of penalty verifies the lower bound of Theorem 4.8. Thus, the oracle inequality of Theorem 4.8 with $\eta=1$, $\alpha=2$ and $t=2\log n$ entails that for n large enough and for any sequence $(M_n)_n$ such that $M_n\leqslant n/2$ for all n:

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

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

4.B Proof of the control of $\bar{\nu}_k$ (Theorem 4.13)

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

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

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

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

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

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

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

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

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

The concentration results are stated in Section 4.B.1. The control of the bracketing entropy is done in Section 4.B.2. Finally, the choice of $x_{K,M,n}$ and the synthesis of the proof are done in Section 4.B.3.

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

4.B.1 Concentration inequality

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

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

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

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

Theorem 4.20. *Assume **[A⋆mixing]** holds. Then there exists a constant C_{mix} depending on c_* and n_* such that the following holds.*

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

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

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

Lemma 4.21 (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 exists positive constants C_1 and C_2 depending on c such that for all $n \geqslant 2$ and all $\lambda \in (0, \frac{1}{C_1 M(\log n)^2})$,

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

where v is defined by

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

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

Consider the processes $(Z_{i(n_*+k+1)+j})_i$ with α-mixing coefficients $\alpha_{Z,j}(n)$. By construction, they satisfy $\alpha_{Z,j}(n) \leqslant e^{-c_* n_* n}$ for all $n \geqslant 1$ and $j \in \{1, \ldots, n_* + k + 1\}$. Apply Lemma 4.21, one gets that there exists two positive constants C_1 and C_2 depending on c_* and n_* such that for all function t, all $\lambda \in (0, \frac{1}{C_1 M(\log n)^2})$ and all $n \in \mathbb{N}$:

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

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

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

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

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

so that

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

which concludes the proof. □

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

Lemma 4.22. *Assume **[A⋆mixing]** holds. Then there exists a constant $C^* \geqslant 1$ depending on n_* and c_* such that the following holds.*

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

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

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

where

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

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

Lemma 4.23. *Assume that there exists a function φ and constants C and $\sigma_{K,M}$ such that $x \mapsto \frac{\varphi(x)}{x}$ is nonincreasing and*

$$\forall \sigma \geqslant \sigma_{K,M} \qquad E \leqslant C\varphi(\sigma)\sqrt{n}. \tag{4.10}$$

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

$$W_{K,M} := \sup_{\theta\in S_{K,M}}\left|\frac{|\bar{\nu}_k(t_\theta^{(D)})|}{\mathbb{E}^*[t_\theta^{(D)}(Z_0)^2] + x_{K,M}^2}\right| \leqslant 4C^*(n_*+k+1)\left[C\frac{\varphi(x_{K,M})}{x_{K,M}^2\sqrt{n}} + \sqrt{\frac{z}{x_{K,M}^2 n}} + \left(2D + \log\frac{1}{\sigma_-}\right)\frac{z(\log n)^2}{x_{K,M}^2 n}\right]. \tag{4.11}$$

The two remaining steps are the control of the bracketing entropy which will lead to equation (4.10) (see Section 4.B.2) and the choice of the parameters $x_{K,M}$ and z (see Section 4.B.3).

4.B.2 Control of the bracketing entropy

Reduction of the set

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

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

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

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

Remark. *In the rest of Section 4.B.2, we always assume that*

$$|L_{k,k}^*| \vee \sup_{\theta' \in \mathbf{S}_n} |b_{\theta'}(y_k)| \leqslant D \tag{4.12}$$

since if this is not the case, then $t_\theta^{(D)}(y_k) = t_{\theta'}^{(D)}(y_k) = 0$. This means that only the y_k satisfying equation (4.12) are relevant for the construction of the brackets.

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

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

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

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

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

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

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

Therefore,

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

so that

$$N(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, \epsilon) \leqslant \bar{N}\left(\mathcal{G}, d_{\mathcal{G}}, \left(\frac{\sigma_-\epsilon}{4(D + \log\frac{1}{\sigma_-})e^{2D}}\right)^2\right) \tag{4.13}$$

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

Decomposition into simple sets

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

Lemma 4.24. *Let* $\epsilon \in \left(0, \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}\right)$. *Then*

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

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, one can assume $\sigma_- \leqslant a(x) \leqslant b(x) \leqslant 1$ and $\sigma_- \leqslant p(x,x') \leqslant q(x,x') \leqslant 1$ for all $x, x' \in [K]$ since all elements of $\{\pi_\theta\}_{\theta\in S_{K,M}}$ and $\{\mathbf{Q}_\theta\}_{\theta\in S_{K,M}}$ satisfy these inequalities. One can also assume that there exists $\theta \in S_{K,M}$ such that $\pi_\theta \in [a,b]$, $\mathbf{Q}_\theta \in [p,q]$ and $\gamma_\theta \in [u,v]$. Under this assumption, all brackets that we construct are non empty and for all $y \in \mathcal{Y}$, $e^{-D}(1 - K\epsilon) \leqslant \sum_x u(y|x) \leqslant \sum_x v(y|x) \leqslant e^D + K\epsilon e^{-D}$.

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

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

where

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

To clarify the role of these quantities, observe that

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

so that

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

Now, let

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

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

and

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

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

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

Lemma 4.25. *Assume $\epsilon \leqslant \frac{1}{2K}$, then*

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

and

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

Proof. Using minimalist notations, one has

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

Then, note that for all j,

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

and

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

so that

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

and likewise

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

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

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

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

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

□

Lemma 4.26. *Assume $\epsilon \leqslant \frac{1}{2K}$, then*

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

and

$$\sup_{1\leqslant i\leqslant k}\sup_{x\in[K]}\|f_{i|k}(x,\cdot)-g_{i|k}(x,\cdot)\|_1 \leqslant 5\left(\frac{2}{\sigma_-}\right)^{k-i+2}\epsilon \leqslant 5\left(\frac{2}{\sigma_-}\right)^{k+1}\epsilon. \tag{4.14}$$

Proof. With minimalist notations, one has

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

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

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

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

□

Define $\eta = 5(\frac{2}{\sigma_-})^{k+1}\epsilon$. Equation (4.14) implies that as soon as $\eta \leqslant 1 - K\sigma_-$ (and in particular $\eta \leqslant 1/2$ since we assume $K \leqslant \frac{1}{2\sigma_-}$), it is possible to enlarge the bracket $[f_{i|k}, g_{i|k}]$ into a bracket $[f'_{i|k}, g'_{i|k}]$ of size smaller than 3η for the norm of Lemma 4.26 such that $f'_{i|k}/(1-\eta)$ and $g'_{i|k}/(1+\eta)$ are transition matrices.

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

Let

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

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

Lemma 4.27. *Assume $\epsilon \leqslant \frac{1}{2K} \wedge \frac{1}{10k}(\frac{\sigma_-}{2})^{k+1}$. Then for all y_0^k,*

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

and

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

Proof. Note that

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

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

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

so that using Lemma 4.26:

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

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

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

For the second part, note that

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

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

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

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

Lemma 4.28. *Assume* $\epsilon \leqslant \frac{1}{2K} \wedge \frac{1}{10k}(\frac{\sigma_-}{2})^k$. *Then*

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

Proof. By definition,

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

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

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

Since we assumed the brackets to be non empty, one has $\int v(y|x)\lambda(dy) \leqslant 1 + \|v-u\|_\infty = 1 + \epsilon e^{-D}$ and $\sum_x(\nu f'_{1|k-1}\dots f'_{k-1|k-1}p)_x \leqslant 1$ (it is the lower bound of a non empty bracket of $\{p_{X_k|Y_0^{k-1},\theta} \,|\, \theta \in S_{K,M}\}$). Therefore, one gets with Lemma 4.27 that

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

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

Lemma 4.27 implies that $\sup_x |(\nu f'_{1|k}\dots f'_{k|k}p)_x - (\omega g'_{1|k}\dots g'_{k|k}q)_x| \leqslant \eta' := 53(k-1)(\frac{2}{\sigma_-})^k\epsilon$. Therefore, since the bracket $[A,B]$ is not empty, one gets by using the assumption on u and v that

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

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

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

which is implied by $\epsilon \leqslant \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}$ since $K \leqslant \frac{1}{\sigma_-}$. This concludes the proof of Lemma 4.24.

Control of the bracketing entropy of the simple sets and synthesis

Lemma 4.29. *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 4.24 ensures that for all $\epsilon \leqslant \frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}$,

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

so that using Equation (4.13) and letting $H(u) = H(\{t_\theta^{(D)} \mid \theta \in \mathbf{B}_\sigma\}, d_k, u)$, one has for all $\epsilon \leqslant \frac{4(D+\log\frac{1}{\sigma_-})e^{2D}}{\sigma_-}\sqrt{\frac{1}{106k}\left(\frac{\sigma_-}{2}\right)^{k+1}}$,

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

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

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

4.B.3 Choice of parameters

The goal of this section is to find a function φ and a constant C for which equation (4.10) holds, and to choose the weights $x_{K,M}$ of Lemma 4.23.

Lemma 4.30. *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

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

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 $(1 + \sqrt{2D + \log\frac{1}{\sigma_-}}\log n)\varphi(x) = \sqrt{n}x^2$ with unknown x, when a solution exists. By definition of E, one has

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

We define $D' := (2D + \log\frac{1}{\sigma_-})(\log n)^2$ in order to lighten the notations. Using equation (4.11), one gets that for all $z > 0$ and $x_{K,M} \geqslant \sigma_{K,M}$, with probability larger than $1 - e^{-z}$,

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

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

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

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

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

and

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

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

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

one has

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

A possible choice of θ is

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

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

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

Therefore,

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

as soon as

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

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

Lemma 4.31. *Let A, B, C and E be functions $\mathbb{N} \longrightarrow [1, \infty)$, and $\varphi : x \longmapsto xA(1 + \sqrt{\log\max(\frac{B}{x}, C)})$. Let σ be the only solution of the equation $\frac{\varphi(x)}{x^2\sqrt{n}} = \frac{1}{E}$ with unknown $x \in \mathbb{R}^*_+$. Let*

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

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

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

In our case,

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

Hence

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

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

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

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

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

so that

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

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

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

CHAPTER

5

HMM WITH TRENDS

This chapter is based on a joint work with Augustin Touron (EDF R&D and Laboratoire de Mathématiques d'Orsay).

5.1 Introduction

Most existing results on hidden Markov model rely heavily on the homogeneity of the process. In practice, some processes cannot be assumed to be stationary. In hidden Markov models and most of their generalizations, the joint process $(X_t, Y_t)_{t\geqslant 1}$ is a Markov chain. We say that the process is *inhomogeneous* when this chain is inhomogeneous, that is when the distribution of (X_t, Y_t) conditionally to (X_{t-1}, Y_{t-1}) depends on t. In inhomogeneous HMM, the transition matrix and emission densities may vary over time.

This chapter is motivated by the study of meteorological data recordings spanning over several decades, in particular temperature. In this setting, it is necessary to include trends to account for the global warming, yet there is no theoretical guarantee that the corresponding maximum likelihood estimator is consistent.

Some inhomogeneous generalizations of HMMs have been studied recently. All of them deal with parametric models.

Diehn et al. (2018) focus on the case where a rapidly fading phenomenon affects the distribution of the observations. Their model is a trivariate process $(X_t, Y_t, Z_t)_{t\geqslant 1}$ where only $(Z_t)_{t\geqslant 1}$ is observed, such that $(X_t, Y_t)_{t\geqslant 1}$ is an homogeneous HMM and $(X_t, Z_t)_{t\geqslant 1}$ is an inhomogeneous HMM. Their key assumption is that the distance between Z_t and Y_t tends to zero fast enough when t tends to infinity. In this sense, the process $(Z_t)_{t\geqslant 1}$ can be seen as a perturbation of the process $(Y_t)_{t\geqslant 1}$ by a rapidly fading inhomogeneous noise.

The authors introduce two estimators. The first one is the usual maximum likelihood estimator. The second one is a so-called quasi-maximum likelihood estimator. Let us write $p^{\theta}_{Y_1^n}$ (resp. $p^{\theta}_{Z_1^n}$) the density of the vector of random variables Y_1^n (resp. Z_1^n) under the parameter θ. The maximum likelihood estimator is

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^{\theta}_{Z_1^n}(Z_1^n)$$

and the quasi-maximum likelihood estimator is

$$\arg\max_{\theta\in\Theta} \frac{1}{n} \log p^{\theta}_{Y_1^n}(Z_1^n).$$

Thus, the second estimator is obtained by doing as if the perturbation wasn't here. The main theoretical result of their article is that both estimators are consistent. The core idea is that the asymptotic properties of $(Y_t)_{t\geqslant 1}$ can be transferred to the process $(Z_t)_{t\geqslant 1}$, in particular the ergodicity and the convergence of the log-likelihood. This makes it possible to adapt existing proofs for homogeneous HMMs.

In practice, the quasi-maximum likelihood estimator does not need to know the structure of the inhomogeneous noise, which makes it easier to use. Their result can also be seen as a proof that the maximum likelihood estimator is robust to a temporary perturbation of the data, in which case the natural estimator is the quasi-maximum likehood one.

A second generalization has been developed by Touron (2018) to handle periodic phenomenon. The author introduces periodic hidden Markov models, where the transition matrix and the emission densities vary periodically over time. He shows that such models are identifiable under general assumptions and that the maximum likelihood estimator is consistent. The consistency proof relies on a transformation of the process into an homogeneous HMM, and the identifiability proof makes use of a spectral method.

For completeness, note that consistency results for non-homogeneous Markov-switching models have also been obtained by Pouzo et al. (2016) and Ailliot and Pene (2015) for instance. These models are a generalization of HMMs where the hidden state X_t depends both of the previous hidden state X_{t-1} and on previous observations, let's say Y_{t-1} for an order one model, and where the observation Y_t depends both on the corresponding hidden state X_t and on previous observations. Such models are called "non-homogeneous" because the transition kernel of the hidden Markov chain depends on time through previous observations. However, this model is actually homogeneous in the previous sense, since the joint process $(X_t, Y_t)_{t\geqslant 1}$ is an homogeneous Markov chain.

In this chapter, we introduce a new inhomogeneous generalization: HMMs with trends. These models allow to deal with non periodic and non vanishing inhomogeneities. A HMM with trends is a trivariate process $(X_t, Y_t, Z_t)_{t\geqslant 1}$ taking values in $\mathcal{X} \times \mathbb{R}^d \times \mathbb{R}^d$ for some integer d (which we assume to be 1 for the sake of simplicity) where only the process $(Y_t)_{t\geqslant 1}$ is observed. In addition, we assume that $(X_t, Z_t)_{t\geqslant 1}$ is an homogeneous HMM and that there exists a vector of functions $(T_x)_{x\in\mathcal{X}}$ from $\mathbb{N}^*$ to $\mathbb{R}^d$, the *trends*, such that

$$Y_t = T_{X_t}(t) + Z_t.$$

We consider polynomial trends. As a consequence, they may diverge. We show that the maximum likelihood estimator recovers all parameters of the homogeneous HMM $(X_t, Z_t)_{t\geqslant 1}$ as well as the trends with respect to the supremum norm.

The first step of the proof is to reduce the set of possible maximizers of the log-likelihood. Sections 5.3 and 5.4 show that the maximum likelihood estimator may be θ only if

- all true trends are near of at least one of the trends of θ (Section 5.4.1),
- all trends of θ are near of at least one of the true trends (Section 5.4.2).

Here, "near" means bounded distance with respect to the supremum norm on $[0, n]$ where n is the number of available observations. The bound on the distance must not depend on n.

These two properties allow to define an equivalence relationship between the trends of possible MLEs and the true trends. In Section 5.5, we define "blocks" as the equivalence classes for this relationship. The second step of the proof is to show that these blocks can be assumed to be observed: since the true trends are either at constant distance or move apart as time passes, the blocks eventually become visible.

Once the blocks $(B_t)_{t\geqslant 1}$ are observed, it is possible to de-trend the observations by substracting a representative of the block for the parameter θ:

$$Z'_t = Z_t + \left[T^*_{X_t}(t) - T^\theta_{B_t}(t)\right].$$

The observations defined in this way are not perfectly de-trended, that is the process $(Z'_t, B_t)_{t\geqslant 1}$ is not homogeneous. However, its distribution varies slowly with time, so that it is nearly homogeneous. Section 5.6 focuses on the approximation of this process by homogeneous processes. This approach allows to compute the limit of the log-likelihood as an *integrated log-likelihood*. The identifiability of the model and the consistency of the maximum likelihood estimator follow from the properties of this limit.

5.2 Model and assumptions

Let $\mathcal{X}^*$ be a finite ordered set. Let $\gamma^* = (\gamma^*_{x^*})_{x^*\in\mathcal{X}^*}$ be a vector of probability densities on $\mathbb{R}$ with respect to the Lebesgue measure. Let $(X_t)_{t\geqslant 1}$ be a Markov chain on $\mathcal{X}^*$ with transition matrix Q^* and initial distribution π^*. For all $x^* \in \mathcal{X}^*$, let $(Z_t^{x^*})_{t\geqslant 1}$ be a sequence of i.i.d. random variables in $\mathbb{R}$ such that these sequences are mutually independent and independent on $(X_t)_{t\geqslant 1}$ and such that for all $x^* \in \mathcal{X}^*$, $Z_1^{x^*}$ has density $\gamma^*_{x^*}$ with respect to the Lebesgue measure. Let $Z_t^{\max} = \max_{x^*\in\mathcal{X}^*} Z_t^{x^*}$ and $Z_t = Z_t^{X_t}$. Finally, let $T^* = (T^*_{x^*})_{x^*\in\mathcal{X}^*}$ be a family of functions $\mathbb{N}^* \longrightarrow \mathbb{R}$ and let $Y_t = Z_t + T^*_{X_t}(t)$ for all $t \geqslant 1$. The $(T^*_{x^*})_{x^*\in\mathcal{X}^*}$ are called *trends*.

Then $(X_t, Z_t)_{t\geqslant 1}$ is a homogeneous hidden Markov model with parameters $(\mathcal{X}^*, \pi^*, Q^*, \gamma^*)$ and $(X_t, Y_t)_{t\geqslant 1}$ is a hidden Markov model with trends with parameters $(\mathcal{X}^*, \pi^*, Q^*, \gamma^*, T^*)$.

Remark. *The random variables $Z_t^{\max}$ are i.i.d. and independent of $(X_t)_{t\geqslant 1}$. They allow to bound Z_t uniformly for all possible values of X_t.*

Consider a sample $(Y_1, \dots, Y_n)$ generated by a hidden Markov model with trends with parameters $(\mathcal{X}^*, \pi^*, Q^*, \gamma^*, T^*)$, which we call the *true parameters*. The goal is to recover these parameters. In the following, we write $\mathbb{P}^*$ the distribution under the true parameters and $\mathbb{E}^*$ the corresponding expectation.

Let Θ be a set such that each $\theta \in \Theta$ is associated to a finite state space $\mathcal{X}^\theta$ and to parameters π^θ, Q^θ, $\gamma^\theta = (\gamma^\theta_x)_{x\in\mathcal{X}^\theta}$ and $T^\theta = (T^\theta_x)_{x\in\mathcal{X}^\theta}$. Without loss of generality, we assume the set $\mathcal{X}^\theta$ to be an ordered set. Write $\mathbb{P}^\theta$ the distribution under the parameter θ. We assume that the true parameters correspond to some $\theta^* \in \Theta$. We will use the following assumptions.

(Adegree) *Bounded degree.* There exists $d \in \mathbb{N}^*$ such that for all $\theta \in \Theta$ and $x \in \mathcal{X}^\theta$, T^θ_x is a polynomial with degree smaller bounded by d.

(Aorder) *Bounded order.* There exists $K \in \mathbb{N}^*$ such that $|\mathcal{X}^\theta| \leqslant K$ for all $\theta \in \Theta$.

(Aenv) *Envelope function.* There exists a nonincreasing function $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that $g \underset{+\infty}{\longrightarrow} 0$ and

$$\forall\theta \in \Theta \quad \forall x \in \mathcal{X}^\theta \quad \forall z \in \mathbb{R} \quad \gamma^\theta_x(z) \leqslant g(|z|).$$

(Amin) *Lower bound function.* There exists a nonincreasing function $m : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\forall \theta \in \Theta \quad \forall x \in \mathcal{X}^\theta \quad \forall z \in \mathbb{R} \quad \gamma_x^\theta(z) \geqslant m(|z|) > 0.$$

(Aerg) *Doeblin condition, geometric ergodicity of the Markov chain.* There exists a constant $\sigma_- > 0$ such that for all $\theta \in \Theta$ and for all $x, x' \in \mathcal{X}^\theta$, $Q^\theta(x, x') \geqslant \sigma_-$.

(Aint) *Integrability of the lower bound function.* The function m defined in **(Amin)** satisfies

$$\forall M > 0, \quad \mathbb{E}^* |\log m(M + |Z_t^{\max}|)| < \infty.$$

(Astable) *Stability under removal of a hidden state.* For all $\theta \in \Theta$ and $x \in \mathcal{X}^\theta$, there exists $\theta' \in \Theta$ such that

$$\forall A \in \sigma(Y_t \,|\, t \geqslant 1), \quad \mathbb{P}^\theta(A \,|\, \forall t \geqslant 1, X_t \neq x) = \mathbb{P}^{\theta'}(A).$$

In other words, there exists a parameter θ' under which the observed process has the same distribution than under the parameter θ without its hidden state x.

(Astable) is typically satisfied when Θ is the set of parameters of all HMMs with no more than K states (resp. exactly K states) and emission densities in a set Γ. Given $\theta \in \Theta$ and $x \in \mathcal{X}^\theta$, it is indeed possible to construct a HMM with $|\mathcal{X}^\theta| - 1$ states that has the same distribution than the HMM with parameter θ conditionned on never visiting x by changing only the transition matrix. Duplicating one of the remaining states results in a HMM with the same number of states as the original HMM.

(Areg) *Regularity of the emission densities.* There exists a modulus of continuity ω (that is a nondecreasing function $\mathbb{R}_+ \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ that is continuous at 0 and such that $\omega(0) = 0$) and a nondecreasing function $L : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\forall z, \eta \in \mathbb{R}, \quad \forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^\theta, \quad \left| \log \frac{\gamma_x^\theta(z+\eta)}{\gamma_x^\theta(z)} \right| \leqslant L(|z|)\omega(|\eta|)$$

and such that

$$\forall M > 0, \quad \mathbb{E}^*[L(M + |Z_1^{\max}|)] < \infty.$$

For all $K' \in \mathbb{N}^*$, let $\Theta_{K'} := \{\theta \in \Theta \text{ s.t. } |\mathcal{X}^\theta| = K'\}$. On this set, we identify $\mathcal{X}^\theta$ and $[K']$.

(Aparam) *Parametric model for the homogeneous part.* For all $K' \in \mathbb{N}$, there exists a compact metric space $\Xi_{K'}$ and mappings $\theta \in \Theta_{K'} \longmapsto \xi(\theta) \in \Xi_{K'}$, $\xi \in \Xi_{K'} \longmapsto Q_\xi \in \mathcal{Q}_{K'}$ and $\xi \in \Xi_{K'} \longmapsto \gamma_\xi \in \mathbf{L}^1(\mathbb{R})^{K'}$ such that:

- for all $\theta \in \Theta_{K'}$,
$$\begin{cases} Q^\theta = Q_{\xi(\theta)}, \\ \gamma^\theta = \gamma_{\xi(\theta)}. \end{cases}$$
In other words, there exists a "component" of the hyperparameter θ living in a compact space that describes the emission densities and the transition matrix.
- for all $x, x' \in [K']$ and for all $z \in \mathbb{R}$, the mappings $\xi \in \Xi_{K'} \longmapsto \gamma_{\xi,x}(z)$ and $\xi \in \Xi_{K'} \longmapsto Q_\xi(x, x')$ are continuous

A natural way of defining a model is taking $\theta = (\xi, \eta)$ where ξ parametrizes the homogeneous parameters (the transition matrix and the emission densities) and where η parametrizes the trends. In this case, **(Aparam)** amounts to assuming that ξ lives in a compact space and that the parametrization $\xi \longmapsto (Q_\xi, \gamma_\xi)$ is continuous. Note that we do not assume anything on the trend parameter η.

(Aid) *Identifiability.* Q^* is invertible and the couples $(\gamma^*_{x^*}(\cdot - \Delta(x^*)), \mathbf{b}^*(x^*))_{x^* \in \mathcal{X}^*}$ are pairwise distinct, where the functions $\mathbf{b}^*$ and Δ are defined in Section 5.5.

(Acentering) *Centering of the emission densities.* 0 is a median of the emission densities, that is:

$$\forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^\theta, \quad \int_{z \leqslant 0} \gamma^\theta_x(z)dz = \frac{1}{2}.$$

(Arate) *Relative entropy rate.* There exists a finite $l(\theta^*)$ such that

$$\frac{1}{n} l_n(\theta^*) \underset{n\to\infty}{\longrightarrow} l(\theta^*).$$

Lemma 5.15 shows that **(Arate)** is a consequence of **(Aenv)**, **(Amin)**, **(Aerg)** and **(Adegree)**.

5.3 Compactness results

In this section, we introduce two subsets $\mathcal{T}(\alpha, n, D)$ and $\mathcal{U}(\beta, n, B)$ of the set of parameters Θ. The first one is the set of parameters which do not approximate one of the true trends. The second one is the set of parameters for which one of the parameter trends is far from all the true trends. We prove some topological properties regarding these subsets. They will later be useful to prove that the MLE recovers the true trends. We shall use Assumptions **(Adegree)** and **(Aorder)**.

5.3.1 The set $\mathcal{T}(\alpha, n, D)$

For all $n \in \mathbb{N}^*$, $\alpha \in (0, 1)$ and $D > 0$, let

$$\mathcal{T}(\alpha, n, D) := \left\{ \theta \in \Theta \text{ s.t. } \frac{\#\left\{ t \in \{1, \dots, n\} \text{ s.t. } \sup_{x^* \in \mathcal{X}^*} \inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(t) - T^\theta_x(t)| \geqslant D \right\}}{n} \geqslant \alpha \right\}$$

be the set of parameters θ for which one of the true trends differ by at least D from all the trends associated with θ on a ratio larger than α of the n first time steps. The set of the corresponding times is denoted by $I_{n,D}(\theta)$:

$$I_{n,D}(\theta) := \left\{ t \in \{1, \dots, n\} \text{ s.t. } \sup_{x^* \in \mathcal{X}^*} \inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(t) - T^\theta_x(t)| \geqslant D \right\}.$$

Note that

$$\mathcal{T}(\alpha, n, D) = \left\{ \theta \in \Theta \text{ s.t. } \frac{\# I_{n,D}(\theta)}{n} \geqslant \alpha \right\}.$$

We also introduce the set of times

$$\tilde{I}_{n,D,x^*}(\theta) := \left\{ t \in \{1, \dots, n\} \text{ s.t. } \inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(t) - T^\theta_x(t)| \geqslant D \right\},$$

for which the true trend of the state x^* is far from all parameter trends. Note that $I_{n,D}(\theta) = \bigcup_{x^* \in \mathcal{X}^*} \tilde{I}_{n,D,x^*}(\theta)$. For $\theta \in \Theta$, $t \geqslant 1$ and $x^* \in \mathcal{X}^*$, let

$$x(\theta, x^*, t) := \arg\min_{x \in \mathcal{X}^\theta} |T^*_{x^*}(t) - T^\theta_x(t)|$$

be the nearest state of $\mathcal{X}^\theta$ from the true trend $T^*_{x^*}$ at time t.

Now let $\alpha \in (0,1)$. We shall prove that for n large enough, for any $D > 0$, $\theta \in \Theta \setminus \mathcal{T}(\alpha, n, D)$ and $x^* \in \mathcal{X}^*$, there exists $x \in \mathcal{X}^\theta$ and a constant $M^{\mathcal{T}}(\alpha, D) > 0$ such that

$$\|T^\theta_x - T^*_{x^*}\|_{\infty,[0,n]} \leqslant M^{\mathcal{T}}(\alpha, D).$$

First, let us rescale the trends on the interval $[0,1]$:

Definition 5.1. *For all $\theta \in \Theta$, $n \in \mathbb{N}^*$ and $x \in \mathcal{X}^\theta$, let*

$$S^{\theta,n}_x : u \in [0,1] \longmapsto T^\theta_x(nu)$$

the trend T^θ_x rescaled from $[0,n]$ to $[0,1]$. Likewise, we define $S^{*,n}_{x^*}$ the true rescaled trend corresponding to the state x^*.

Notice that under Assumption **(Adegree)**, the rescaled trends are polynomials with degree at most d. Let $\theta \in \Theta \setminus \mathcal{T}(\alpha, n, D)$. By definition of $\mathcal{T}(\alpha, n, D)$,

$$\frac{\#\{t \in \{1, \dots, n\} \text{ s.t. } \forall x^* \in \mathcal{X}^*,\ |T^*_{x^*}(t) - T^\theta_{x(\theta,x^*,t)}(t)| \leqslant D\}}{n} \geqslant 1 - \alpha.$$

Let $x^* \in \mathcal{X}^*$. Using **(Aorder)** and the pigeonhole principle, one can define $x_S(x^*, \theta, n)$ as the smallest $x \in \mathcal{X}^\theta$ such that

$$\frac{\#\{t \in \{1, \dots, n\} \text{ s.t. } x(\theta, x^*, t) = x_S(x^*, \theta, n) \text{ and } |T^*_{x^*}(t) - T^\theta_{x(\theta,x^*,t)}(t)| \leqslant D\}}{n} \geqslant \frac{1-\alpha}{K}.$$

Theorem 5.1. *Under Assumptions **(Aorder)** and **(Adegree)**, for all $x^* \in \mathcal{X}^*$,*

$$\mathcal{S}_{x^*} := \bigcup_{n \geqslant \frac{4K(d+1)}{1-\alpha}} \bigcup_{\theta \notin \mathcal{T}(\alpha,n,D)} \{S^{\theta,n}_{x_S(x^*,\theta,n)} - S^{*,n}_{x^*}\}$$

is relatively compact in the set of continuous functions $\left(\mathcal{C}^0([0,1]), \|\cdot\|_\infty\right)$.

Since $(\mathcal{S}_{x^*}, \|\cdot\|_\infty)$ is a metric space, it is enough to prove that its closure is sequentially compact. Hence let us prove that every sequence of elements of $\mathcal{S}_{x^*}$ admits a subsequence that converges in $\left(\mathcal{C}^0([0,1]), \|\cdot\|_\infty\right)$.

Let us first give another representation of the elements of $\mathcal{S}_{x^*}$. This is the object of the two following lemmas.

Lemma 5.2. *For any $S \in \mathcal{S}_{x^*}$, there exists $(u^S_1, \dots, u^S_{d+1}) \in [0,1]^{d+1}$ such that*

$$\begin{cases} \forall i \neq j, \quad |u^S_i - u^S_j| \geqslant \frac{1-\alpha}{4K(d+1)}, \\ \forall i, \quad |S(u^S_i)| \leqslant D. \end{cases}$$

Proof. For $S \in \mathcal{S}_{x^*}$, let $n(S)$ be an integer $n \geqslant \frac{4K(d+1)}{1-\alpha}$ such that $S \in \bigcup_{\theta \notin \mathcal{T}(\alpha,n,D)} \{S^{\theta,n}_{x_S(x^*,\theta,n)} - S^{*,n}_{x^*}\}$. Let us define $(u_1^S, \dots, u_{d+1}^S)$ iteratively. Let

$$\mathcal{A} := \left\{ t \in \{1, \dots, n\} \text{ s.t. } \left| S\left(\frac{t}{n(S)}\right) \right| \leqslant D \right\}$$

be the set of times such that the trend associated to S differs from $T^*_{x^*}$ by no more than D. By definition of $\mathcal{S}_{x^*}$ and $x_S(x^*, \theta, n)$, we have $\#\mathcal{A} \geqslant \frac{1-\alpha}{K} n$.

Let $\mathcal{A}_0 = \mathcal{A}$ and, for all $i \geqslant 1$,

- $t_i^S \in \mathcal{A}_{i-1}$;
- $\mathcal{A}_i = \mathcal{A}_{i-1} \setminus \bar{B}\left(t_i^S, \frac{1-\alpha}{4K(d+1)} n\right)$.

The closed ball $\bar{B}\left(t_i^S, \frac{1-\alpha}{4K(d+1)} n\right)$ contains at most $1 + \left\lfloor 2\frac{1-\alpha}{4K(d+1)} n \right\rfloor \leqslant 3\frac{1-\alpha}{4K(d+1)} n$ elements, as $\frac{1-\alpha}{4K(d+1)} n \geqslant 1$. Thus, for all $i \geqslant 0$,

$$\#\mathcal{A}_i \geqslant n \frac{1-\alpha}{K} \left(1 - \frac{3}{4(d+1)} i\right).$$

In particular, $\mathcal{A}_i \neq \emptyset$ for all $i \in \{0, \dots, d+1\}$, which makes it possible to define $(t_i^S)_{1 \leqslant i \leqslant d+1}$. It is easy to see that taking $u_i^S = \frac{t_i^S}{n(S)}$ for all $i \in \{1, \dots, d+1\}$ satisfies the desired properties. □

The next lemma states that this representation is continuous.

Lemma 5.3. *For all $\epsilon > 0$, the map*

$$(u_1, \dots, u_{d+1}, s_1, \dots, s_{d+1}) \in \left\{ [0,1]^{d+1} \text{ s.t. } \inf_{i \neq j} |u_i - u_j| \geqslant \epsilon \right\} \times \mathbb{R}^{d+1} \longmapsto P_{u,s} \in \left(\mathcal{C}^0([0,1]), \|.\|_\infty\right)$$

is continuous, where $P_{u,s}$ is the only polynomial with degree at most d such that $P(u_i) = s_i$ for all $i \in \{1, \dots, d+1\}$.

Proof. This is a straightforward consequence of the Lagrange form of the interpolation polynomial. □

Let $(\hat{S}_m)_{m \geqslant 1}$ be a sequence in $\mathcal{S}_{x^*}$. It can be represented by the vectors

$$\left(\left(u_i^{\hat{S}_m} \right)_{1 \leqslant i \leqslant d+1}, \left(\hat{S}_m(u_i^{\hat{S}_m}) \right)_{1 \leqslant i \leqslant d+1} \right)_{m \geqslant 1} =: (\hat{v}_m)_{m \geqslant 1}.$$

This sequence takes its values in

$$\left\{ u \in [0,1]^{d+1} \text{ s.t. } \inf_{i \neq j} |u_i - u_j| \geqslant \frac{1-\alpha}{4K(d+1)} \right\} \times [-D, D]^{d+1},$$

which is compact, so that $(\hat{v}_m)_{m \geqslant 1}$ admits a convergent subsequence and, using Lemma 5.3, the corresponding subsequence of $(\hat{S}_m)_{m \geqslant 1}$ converges, which ends the proof of Theorem 5.1.

Theorem 5.1 and Ascoli-Arzelà's theorem imply the following corollary:

Corollary 5.4. *Assume* ***(Adegree)*** *and* ***(Aorder)****. Then the sets* $\mathcal{S}_{x^*}$ *(*$x^* \in \mathcal{X}^*$*) are uniformly equicontinuous and uniformly bounded: there exists a constant* $M^{\mathcal{T}}(\alpha, D) < +\infty$ *and a continuity modulus* ω *(i.e. a function* $\omega : \mathbb{R}_+ \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$ *continuous at* 0 *such that* $\omega(0) = 0$*) such that*

$$\begin{cases} \forall x^* \in \mathcal{X}^* \quad \forall S \in \mathcal{S}_{x^*} \quad \forall u \in [0,1] \qquad |S(u)| \leqslant M^{\mathcal{T}}(\alpha, D), \\ \forall x^* \in \mathcal{X}^* \quad \forall S \in \mathcal{S}_{x^*} \quad \forall u, v \in [0,1] \qquad |S(u) - S(v)| \leqslant \omega(|u - v|). \end{cases}$$

As a consequence, for all $\alpha \in (0,1)$, $n \geqslant \frac{4K(d+1)}{1-\alpha}$, $D > 0$, $\theta \notin \mathcal{T}(\alpha, n, D)$ and $x^* \in \mathcal{X}^*$,

$$\|T^{\theta}_{x_S(x^*,\theta,n)} - T^*_{x^*}\|_{\infty,[0,n]} \leqslant M^{\mathcal{T}}(\alpha, D). \tag{5.1}$$

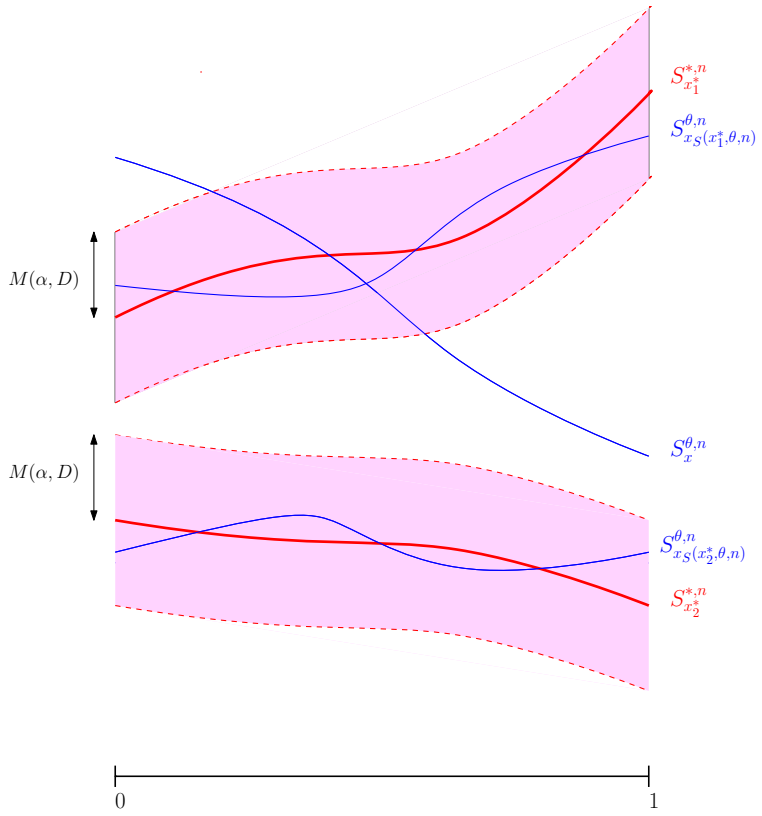


Figure 5.1: Rescaled trends of a parameter not in $\mathcal{T}(\alpha, n, D)$. Every true trend is at bounded distance of at least one parameter trend. However, some parameter trends may be far from all true trends.

This result will be useful to show the consistency of the MLE, as we will later show that for n large enough, $\hat{\theta}_n \notin \mathcal{T}(\alpha, n, D)$.

5.3.2 The set $\mathcal{U}(\beta, n, B)$

For $n \geqslant 1$, $\beta \in (0,1)$ and $B > 0$, let

$$\mathcal{U}(\beta, n, B) := \left\{ \theta \in \Theta \text{ s.t. } \exists x \in \mathcal{X}^\theta : \frac{\#\left\{ t \in \{1, \dots, n\} \text{ s.t. } \inf_{x^* \in \mathcal{X}^*} |T^*_{x^*}(t) - T^{\theta}_{x}(t)| \geqslant B \right\}}{n} \geqslant \beta \right\}$$

be the set of parameters θ for which one of the parameter trends differ from all the true trends by at least B, on a proportion at least β of the first n time steps. Let $x_U(\theta)$ the index of this "isolated" trend, or the smallest index if there are several, and

$$I^U_{n,B}(\theta) := \left\{ t \in \{1,\dots,n\} \text{ s.t. } \inf_{x^* \in \mathcal{X}^*} |T^*_{x^*}(t) - T^\theta_{x_U(\theta)}(t)| \geqslant B \right\}$$

the corresponding times. These definitions imply that if $\theta \in \mathcal{U}(\beta, n, B)$, then $\#I^U_{n,B}(\theta) \geqslant \beta n$.

For $n \geqslant 1$, $\beta \in (0,1)$, $B > 0$ and $\theta \notin \mathcal{U}(\beta, n, B)$, one has

$$\forall x \in \mathcal{X}^\theta \quad \frac{\#\{t \in \{1,\dots,n\} \text{ t.q. } \inf_{x^* \in \mathcal{X}^*} |T^*_{x^*}(t) - T^\theta_x(t)| \leqslant B\}}{n} \geqslant 1 - \beta.$$

For $x^* \in \mathcal{X}^*$ and $x \in \mathcal{X}^\theta$, let

$$L_{n,B}(\theta, x^*, x) := \{t \in \{1,\dots,n\} \text{ s.t. } |T^*_{x^*}(t) - T^\theta_x(t)| \leqslant B\}.$$

As a consequence of **(Aorder)** and the pigeonhole principle, for all $\theta \notin \mathcal{U}(\beta, n, B)$, there exists $x^*_L(\theta, x, n) \in \mathcal{X}^*$ such that

$$\frac{\#L_{n,B}(\theta, x^*_L(\theta, x, n), x)}{n} \geqslant \frac{1-\beta}{K}.$$

To ensure uniqueness, we take $x^*_L(\theta, x, n)$ as the smallest such index when several indices are possible. We get the following theorem, whose proof is identical to the one of Theorem 5.1.

Theorem 5.5. *Under Assumptions **(Aorder)** and **(Adegree)**, the set*

$$\mathcal{S}' := \bigcup_{n \geqslant \frac{4K(d+1)}{1-\beta}} \bigcup_{\theta \notin \mathcal{U}(\beta,n,B)} \bigcup_{x \in \mathcal{X}^\theta} \{S^{\theta,n}_x - S^{*,n}_{x^*_L(\theta,x,n)}\}$$

is relatively compact in $\left(\mathcal{C}^0([0,1]), \|\cdot\|_\infty\right)$. It follows that it is uniformly bounded and uniformly equicontinuous.

Therefore, for all $\beta \in (0,1)$ and $B > 0$, there exists a constant $M^{\mathcal{U}}(\beta, B)$ such that for all $n \geqslant \frac{4K(d+1)}{1-\beta}$, $\theta \notin \mathcal{U}(\beta, n, B)$ and $x \in \mathcal{X}^\theta$,

$$\|T^\theta_x - T^x_{x^*_L(\theta,x,n)}\|_{\infty,[0,n]} \leqslant M^{\mathcal{U}}(\beta, B)$$

5.4 Localization of the MLE

5.4.1 The MLE is not in $\mathcal{T}(\alpha, n, D)$

The key idea of this section is that if one of the true trends is far from all parameter trends, then the observations coming from this true trend will significantly reduce the likelihood.

Let $\alpha \in (0,1)$, $n \geqslant 1$, $D > 0$ and $\theta \in \mathcal{T}(\alpha, n, D)$ as defined in Section 5.3. Recall that for $x \in \mathcal{X}^\theta$, we defined

$$\tilde{I}_{n,D,x^*}(\theta) := \left\{ t \in \{1,\dots,n\} \text{ s.t. } \inf_{x \in \mathcal{X}^\theta} |T^*_{x^*}(t) - T^\theta_x(t)| \geqslant D \right\}.$$

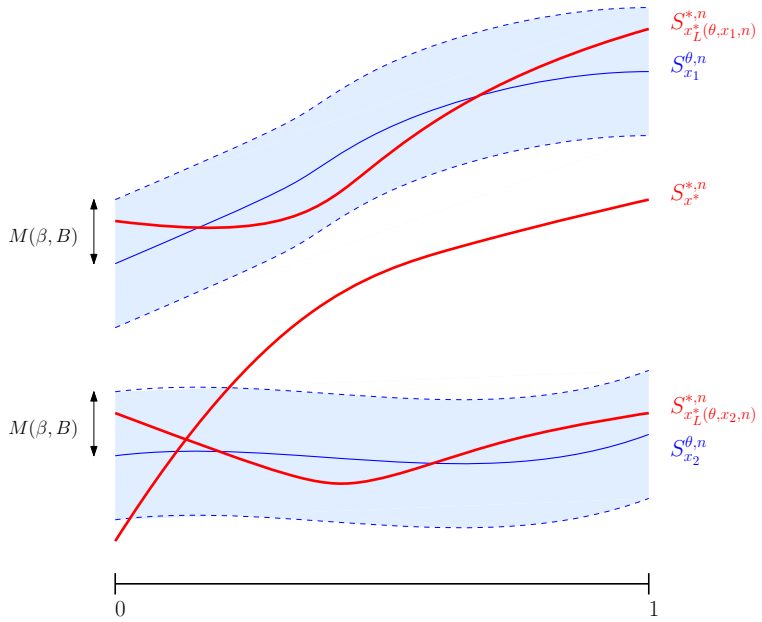


Figure 5.2: Rescaled trends of a parameter not in $\mathcal{U}(\beta,n,B)$. Every parameter trend is at bounded distance of at least one true trend. However, some true trends may be far from all parameter trends.

By the pigeonhole principle and by definition of $\mathcal{T}(\alpha,n,D)$, there exists $x^*_p(\theta)$ such that

$$\frac{\#\tilde{I}_{n,D,x^*_p(\theta)}(\theta)}{n} \geqslant \frac{\alpha}{K^*}.$$

Then,

$$\begin{aligned}\frac{1}{n}l_n(\theta) &= \frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1}) \\ &\leqslant \log g(0) + \frac{1}{n}\sum_{t\in\tilde{I}_{n,D,x^*_p(\theta)}(\theta)} \mathbf{1}_{X_t=x^*_p(\theta)}\log\frac{g(\{D-|Z_t^{\max}|\}_+)}{g(0)}. \qquad (5.2)\end{aligned}$$

We used the fact that under Assumption **(Aenv)**,

$$\begin{aligned}p^\theta(Y_t|Y_1^{t-1}) &= \sum_{x\in\mathcal{X}^\theta} p^\theta(X_t=x|Y_1^{t-1})\gamma^\theta_x(Y_t-T^\theta_x(t)) \\ &\leqslant \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*}\sup_{x\in\mathcal{X}^\theta}\gamma^\theta_x(Z_t+T^*_{x^*}(t)-T^\theta_x(t)) \\ &\leqslant \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*}\sup_{x\in\mathcal{X}^\theta} g(\{|T^*_{x^*}(t)-T^\theta_x(t)|-|Z_t^{\max}|\}_+) \\ &\leqslant \sum_{x^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x^*} g(\inf_{x\in\mathcal{X}^\theta}\{|T^*_{x^*}(t)-T^\theta_x(t)|-|Z_t^{\max}|\}_+).\end{aligned}$$

We will need the following assumption:

(Asegments) There exists $A\in\mathbb{N}^*$ such that for all $\theta\in\Theta$, $n\in\mathbb{N}^*$, $D>0$ and for all $x^*\in\mathcal{X}^*$, $\tilde{I}_{n,D,x^*}(\theta)$ has at most A connected components.

Lemma 5.6. *Under Assumption* ***(Adegree)*** *and* ***(Aorder)****, Assumption* ***(Asegments)*** *is satisfied.*

Proof. The functions $(t \longmapsto T^*_{x^*}(t) - T^\theta_x(t))_{x\in\mathcal{X}^\theta}$ are polynomials whose degree is at most d by **(Adegree)**. Their derivatives vanish at most $d-1$ times, and the set of times t where they are larger than D in absolute value is a union of segments containing either, a zero of their derivative, $+\infty$ or $-\infty$. Hence there are at most $d+1$ such segments. Thus, $\tilde{I}_{n,D,x^*}(\theta)$ is an intersection of at most K sets (using **(Aorder)**), each of them having at most $d+1$ connected components. Therefore, one may take $A = (d+1)^K$. □

Definition 5.2. *For all $n \in \mathbb{N}^*$, $D > 0$, $x^* \in \mathcal{X}^*$ and $\theta \in \Theta$, we denote by $J_{n,D,x^*}(\theta)$ the largest connected component of $\tilde{I}_{n,D,x^*}(\theta)$. In case of tie, we choose the first one for the usual order in $\mathbb{R}$.*

Under Assumption **(Asegments)**, by the pigeonhole principle,

$$\frac{\#J_{n,D,x^*_p(\theta)}(\theta)}{n} \geqslant \frac{\alpha}{AK^*},$$

and using equation (5.2):

$$\frac{1}{n} l_n(\theta) \leqslant \log g(0) + \frac{1}{n} \sum_{t\in J_{n,D,x^*_p(\theta)}(\theta)} \mathbf{1}_{X_t = x^*_p(\theta)} \log \frac{g(\{D - |Z_t^{\max}|\}_+)}{g(0)}. \tag{5.3}$$

Lemma 5.7. *Let $\delta > 0$ and assume* ***(Aerg)****. Then, almost surely,*

$$\liminf_{n\to\infty} \inf_{\substack{S\subset\{1,\dots,n\}\\ S \text{ segment}\\ \#S\geqslant\delta n}} \inf_{x^*\in\mathcal{X}^*} \frac{1}{n} \sum_{t\in S} \mathbf{1}_{X_t = x^*} \geqslant \frac{\delta\sigma_-}{4}.$$

Proof. The idea is to split $\{1,\dots,n\}$ into segments of size $\frac{\delta}{2}n$ and to control the infimum of the empirical mean over each segment. Each segment of size larger than δn contains at least one of those segments. The proof is detailled in Section 5.9.1. □

Applying Lemma 5.7 to $S = J_{n,D,x^*_p(\theta)}(\theta)$, one gets that almost surely,

$$\liminf_{n\to\infty} \inf_{\theta\in\mathcal{T}(\alpha,n,D)} \frac{1}{n} \sum_{t\in J_{n,D,x^*_p(\theta)}(\theta)} \mathbf{1}_{X_t = x^*_p(\theta)} \geqslant \frac{\alpha\sigma_-}{4AK^*}. \tag{5.4}$$

Lemma 5.8. *Let $\delta \in (0,1)$, $(U_t)_{t\geqslant 1}$ a sequence of i.i.d. non-positive integrable random variables and $(\delta_n)_{n\geqslant 1}$ a non-decreasing sequence of $[0,1]$-valued random variables such that $\liminf_{n\to\infty}\delta_n \geqslant \delta$ a.s. For all $\beta \in [0,1]$, let us denote by $q_U(\beta)$ the β-quantile of U_1, i.e.*

$$q_U(\beta) = \inf\{u \text{ t.q. } \mathbb{P}(U_1 \leqslant u) \geqslant \beta\}.$$

Then, almost surely,

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ \#S\geqslant\delta_n n}} \frac{1}{n}\sum_{t\in S} U_t \leqslant \mathbb{E}[U_1 \mathbf{1}_{U_1 > q_U(1-\delta)}].$$

Equivalently, if $(V_t)_{t\geqslant 1}$ is a sequence of non-negative i.i.d. integrable random variables and $(\delta_n)_{n\geqslant 1}$ a non-increasing sequence of $[0,1]$-valued random variables such that $\limsup_{n\to\infty}\delta_n \leqslant \delta$ a.s., one has almost surely

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ \#S\leqslant\delta_n n}} \frac{1}{n}\sum_{t\in S} V_t \leqslant \mathbb{E}[V_1 \mathbf{1}_{V_1 \geqslant q_V(1-\delta)}].$$

Remark. *The supremum is taken over all subsets S, not only segments.*

Proof. Proof in Section 5.9.1. □

For all $t \geqslant 1$ and $D > 0$, let $U_t^D = \log \frac{g(\{D-|Z_t^{\max}|\}_+)}{g(0)}$. U_t^D is non-positive by definition. Then, taking

$$\begin{cases} \delta = \frac{\alpha\sigma_-}{4AK^*}, \\ \delta_n = \inf\limits_{m \geqslant n} \inf\limits_{\theta \in \mathcal{T}(\alpha,m,D)} \dfrac{1}{m} \displaystyle\sum_{t \in J_{m,D,x_p^*(\theta)}(\theta)} \mathbf{1}_{X_t = x_p^*(\theta)}, \end{cases}$$

one has $\liminf_{n\to\infty} \delta_n \geqslant \delta$ by equation (5.4). Therefore, Lemma 5.8 combined with equation (5.3) implies that almost surely,

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{T}(\alpha,n,D)} \frac{1}{n} l_n(\theta) \leqslant \log g(0) + \mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geqslant q_{U^D}(1-\frac{\alpha\sigma_-}{4AK^*})}].$$

Note that $U_t^D = f^D(Z_t^{\max})$ where $f^D : z \in \mathbb{R}_+ \longmapsto \log \frac{g(\{D-z\}_+)}{g(0)}$. f^D is non-decreasing, so for all x and q, $f^D(x) > f^D(q)$ implies $x > q$. Hence

$$\mathbb{P}\left(f^D(|Z_1^{\max}|) > f^D(q_{|Z^{\max}|}(1-\delta))\right) \leqslant \mathbb{P}\left(|Z_1^{\max}| > q_{|Z^{\max}|}(1-\delta)\right) \leqslant \delta.$$

In other words, $\mathbb{P}(U_1^D \leqslant f^D(q_{|Z^{\max}|}(1-\delta))) \geqslant 1-\delta$, hence $q_{U^D}(1-\delta) \leqslant f^D(q_{|Z^{\max}|}(1-\delta))$ by definition of quantiles. Thus, for all $z \geqslant 0$,

$$\mathbf{1}_{z \geqslant q_{|Z^{\max}|}(1-\delta)} \leqslant \mathbf{1}_{f^D(z) \geqslant f^D(q_{|Z^{\max}|}(1-\delta))} \leqslant \mathbf{1}_{f^D(z) \geqslant q_{U^D}(1-\delta)}$$

since f^D is non-decreasing. Therefore,

$$\mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geqslant q_{U^D}(1-\frac{\alpha\sigma_-}{4AK^*})}] \leqslant \mathbb{E}[U_1^D \mathbf{1}_{|Z_1^{\max}| \geqslant q_{|Z^{\max}|}(1-\frac{\alpha\sigma_-}{4AK^*})}].$$

Then, for all $\delta > 0$, the monotone convergence theorem applied to the right-hand side entails

$$\mathbb{E}[U_1^D \mathbf{1}_{U_1^D \geqslant q_{U^D}(1-\delta)}] \underset{D\to+\infty}{\longrightarrow} -\infty.$$

Thus, under Assumption **(Adegree)**, **(Aenv)** and **(Arate)**, there exists $D(\alpha) < \infty$ such that

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{T}(\alpha,n,D(\alpha))} \frac{1}{n} l_n(\theta) \leqslant l(\theta^*) - 1,$$

so that almost surely, for n large enough,

$$\hat{\theta}_n \notin \mathcal{T}(\alpha, n, D(\alpha)). \tag{5.5}$$

Considering the definition of the set $\mathcal{T}(\alpha, n, D)$, this means that for n large enough, every true trend is approximated by a trend of the MLE.

5.4.2 The MLE is not in $\mathcal{U}(\beta, n, B)$

We use the notations of Section 5.3, in particular the sets $\mathcal{T}(\alpha, n, D)$, $\mathcal{U}(\beta, n, B)$ and $I^U_{n,b}(\theta)$ and the quantities $x_U(\theta)$ and $M^{\mathcal{T}}(\alpha, D)$.

Let $\alpha, \beta \in (0,1)$, $n \geqslant \frac{4K(d+1)}{1-\alpha}$, $D > 0$ and $B > M^{\mathcal{T}}(\alpha, D)$, and $\theta \in \mathcal{U}(\beta, n, B) \setminus \mathcal{T}(\alpha, n, D)$. Since $\theta \in \mathcal{U}(\beta, n, B)$, one of its trends is far from all true trends. The key of the proof is to show that removing this trend increases the likelihood of the observations. An interpretation is that the likelihood "expects" some observations to come from this trend because of **(Aerg)**, but that it never observes it.

Let θ^U be the parameter such that

$$\begin{cases} \mathcal{X}^{\theta^U} = \mathcal{X}^\theta \setminus \{x_U(\theta)\}, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad \pi^{\theta^U}(x) = \frac{\pi^\theta(x)}{1-\pi^\theta(x_U(\theta))}, \\ \forall x, x' \in \mathcal{X}^{\theta^U}, \quad Q^{\theta^U}(x, x') = \frac{Q^\theta(x,x')}{1-Q^\theta(x,x_U(\theta))}, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad \gamma_x^{\theta^U} = \gamma_x^\theta, \\ \forall x \in \mathcal{X}^{\theta^U}, \quad T_x^{\theta^U} = T_x^\theta. \end{cases}$$

Note that for all $x, x' \in \mathcal{X}^{\theta^U}$,

$$\begin{aligned} \pi^{\theta^U}(x) &= \mathbb{P}^\theta(X_1 = x \,|\, X_1 \neq x_U(\theta)) \\ &= \mathbb{P}^\theta(X_1 = x \,|\, \forall t \geqslant 1, X_t \neq x_U(\theta)) \end{aligned}$$

and

$$\begin{aligned} \forall s \geqslant 1, \quad Q^{\theta^U}(x, x') &= \mathbb{P}^\theta(X_{s+1} = x' \,|\, X_s = x, X_{s+1} \neq x_U(\theta)) \\ &= \mathbb{P}^\theta(X_{s+1} = x' \,|\, X_s = x, \forall t \geqslant 1, X_t \neq x_U(\theta)), \end{aligned}$$

so that

$$\forall \in \sigma(Y_t \,|\, t \geqslant 1), \quad \mathbb{P}^{\theta^U}(A) = \mathbb{P}^\theta(A \,|\, \forall t \geqslant 1, X_t \neq x_U(\theta)),$$

so that **(Astable)** ensures that θ^U actually corresponds to an element of Θ. θ^U corresponds to the HMM with parameter θ whose state $x_U(\theta)$ has been removed. Then

$$\frac{1}{n} l_n(\theta^U) = \frac{1}{n} \sum_{t=1}^{n} \log p^\theta(Y_t \,|\, Y_1^{t-1}, x_U(\theta) \notin X_1^t),$$

with the abuse of notation

$$x \in X_1^t \quad \iff \quad \exists s \in \{1, \dots, t\} \quad X_s = x.$$

Under Assumption **(Aerg)**, one has

$$\begin{aligned} p^\theta(Y_t \,|\, Y_1^{t-1}) &= p^\theta(Y_t \,|\, X_t = x_U(\theta)) p^\theta(X_t = x_U(\theta) \,|\, Y_1^{t-1}) \\ &\quad + p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1}) p^\theta(X_t \neq x_U(\theta) \,|\, Y_1^{t-1}) \\ &\leqslant (1 - \sigma_-) p^\theta(Y_t \,|\, X_t = x_U(\theta)) + (1 - \sigma_-) p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1}), \end{aligned}$$

hence

$$\frac{1}{n}l_n(\theta) - \underbrace{\frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1})}_{(i)} \leqslant \log(1-\sigma_-) + \underbrace{\frac{1}{n}\sum_{t=1}^{n}\log\left(1+\frac{p^\theta(Y_t \,|\, X_t = x_U(\theta))}{p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1})}\right)}_{(ii)}.$$

The next steps are:

- Prove that (i) is close to $\frac{1}{n}l_n(\theta^U)$ for large enough n.
- Prove that (ii) goes to zero uniformly in $\theta \in \mathcal{U}(\beta, n, B) \setminus \mathcal{T}(\alpha, n, D)$.

First step: controlling (i) We shall prove that for an adequate choice of β and B, one has almost surely

$$\limsup_{n\to\infty} \sup_{\theta\in\mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)} \left|(i) - \frac{1}{n}l_n(\theta^U)\right| \leqslant \frac{-\log(1-\sigma_-)}{3}.$$

The following forgetting property allows to control what happens in the distant past.

Lemma 5.9. *Assume **(Aerg)**. Then for all $t \geqslant 1$, $\theta \in \Theta$, for any probability measures μ and ν on $\mathcal{X}^\theta$, for all $x \in \mathcal{X}^\theta$ and Y_0^t,*

$$|\log p^\theta(Y_t \,|\, Y_0^{t-1}, X_t \neq x, X_0 \sim \mu) - \log p^\theta(Y_t \,|\, Y_0^{t-1}, X_t \neq x, X_0 \sim \nu)| \leqslant C\rho^t$$

where $\rho = 1 - \frac{\sigma_-}{1-\sigma_-}$ and $C = \frac{2}{\rho(1-\rho)^3}$.

Proof. Proof in Section 5.9.2. □

Let $a \in \mathbb{N}^*$. It follows from Lemma 5.9 that for all t, almost surely,

$$\left|\log p^\theta(Y_t \,|\, Y_1^{t-1}, X_t \neq x_U(\theta)) - \log p^\theta(Y_t \,|\, Y_1^{t-1}, X_t \neq x_U(\theta), x_U(\theta) \notin X_1^{t-a})\right| \leqslant C\rho^a. \tag{5.6}$$

It remains to add X_{t-a+1}^{t-1} to the conditioning. This is the goal of the following lemma.

Lemma 5.10. *Assume **(Aerg)**, **(Aorder)**, **(Aenv)** and **(Amin)**. Then for all $a \in \mathbb{N}^*$,*

$$\begin{aligned}\sup_{\theta\in\mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)} &\left|\frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1}, X_t \neq x_U(\theta), x_U(\theta) \notin X_1^{t-a}) \right.\\ &\left. - \frac{1}{n}\sum_{t=1}^{n}\log p^\theta(Y_t|Y_1^{t-1}, x_U(\theta) \notin X_1^t)\right| \\ &\leqslant \frac{2aK^2}{\sigma_-^3}\left((1-\beta) + \frac{1}{n}\sum_{i=1}^{n-1}\left(1 \wedge \frac{g(\{B - |Z_i^{\max}|\}_+)}{m(|Z_i^{\max}| + M^{\mathcal{T}}(\alpha, D))}\right)\right) \\ &=: \frac{2aK^2}{\sigma_-^3}\left((1-\beta) + \frac{1}{n}\sum_{i=1}^{n-1}h_U(B, Z_i^{\max})\right).\end{aligned}$$

Proof. We show that when **(Aerg)** and **(Aorder)** hold, one has for all $\theta \in \mathcal{U}(\beta, n, B) \setminus \mathcal{T}(\alpha, n, D)$, $t \leqslant n$, $a \in \mathbb{N}^*$ and $y_1^t \in \mathcal{Y}^t$,

$$\begin{aligned}
&\left| \log p^\theta(y_t|y_1^{t-1}, X_t \neq x_U(\theta), x_U(\theta) \notin X_1^{t-a}) - \log p^\theta(y_t|y_1^{t-1}, x_U(\theta) \notin X_1^t) \right| \\
&\qquad \leqslant \frac{2}{\sigma_-} p^\theta \left(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} | y_1^{t-1}, X_t \neq x_U(\theta), x_U(\theta) \notin X_1^{t-a} \right) \qquad (5.7) \\
&\qquad \leqslant \frac{2K^2}{\sigma_-^3} \sum_{i=(t-a+1)\vee 1}^{t-1} \frac{\gamma^\theta_{x_U(\theta)}(y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(y_i - T^\theta_x(i))}. \qquad (5.8)
\end{aligned}$$

Then, we show that under **(Aenv)** and **(Amin)**, one has for all $\theta \in \mathcal{U}(\beta, n, B) \setminus \mathcal{T}(\alpha, n, D)$ and $i \leqslant n$

$$\frac{\gamma^\theta_{x_U(\theta)}(Y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} \leqslant \begin{cases} 1 \wedge \dfrac{g(\{B - |Z_i^{\max}|\}_+)}{m(|Z_i^{\max}| + M^{\mathcal{T}}(\alpha, D))} =: h_U(B, Z_i^{\max}) & \text{if } i \in I^U_{n,B}(\theta), \\ 1 & \text{if } i \notin I^U_{n,B}(\theta). \end{cases} \qquad (5.9)$$

so that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \frac{\gamma^\theta_{x_U(\theta)}(Y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x \in \mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} &\leqslant \frac{1}{n} \sum_{i \notin I^U_{n,B}(\theta)} 1 + \frac{1}{n} \sum_{i \in I^U_{n,B}(\theta)} h_U(B, Z_i^{\max}) \\
&\leqslant (1 - \beta) + \frac{1}{n} \sum_{i=1}^n h_U(B, Z_i^{\max})
\end{aligned}$$

using that $\# I^U_{n,B}(\theta) \geqslant \beta n$ for all $\theta \in \mathcal{U}(\beta, n, B)$ and $h_U(b, z) \geqslant 0$ for all $b, z \geqslant 0$. The lemma follows by summing equation (5.8) over t. The details of the proof can be found in Section 5.9.2. □

Thus, equation (5.6) and Lemma 5.10 imply

$$\limsup_{n \to \infty} \sup_{\theta \in \mathcal{U}(\beta,n,B) \setminus \mathcal{T}(\alpha,n,D)} \left| (i) - \frac{1}{n} l_n(\theta^U) \right| \leqslant C\rho^a + \frac{2aK^2}{\sigma_-^3}(1 - \beta) + \frac{2aK^2}{\sigma_-^3} \mathbb{E}^* \left[h_U(B, Z_i^{\max}) \right].$$

Now choose a large enough so that

$$C\rho^a \leqslant \frac{-\log(1 - \sigma_-)}{9},$$

then β such that

$$\frac{2aK^2}{\sigma_-^3}(1 - \beta) \leqslant \frac{-\log(1 - \sigma_-)}{9}.$$

Finally, note that $0 \leqslant h_U(b, z) \leqslant 1$ for all $b, z \geqslant 0$ and that $h_U(b, z) \longrightarrow 0$ when $b \longrightarrow \infty$ for all z, so that by the dominated convergence theorem, there exists B such that

$$\frac{2aK^2}{\sigma_-^3} \mathbb{E}^* \left[h_U(B, Z_i^{\max}) \right] \leqslant \frac{-\log(1 - \sigma_-)}{9},$$

which ensures that for all $\alpha \in (0, 1)$ and $D > 0$, there exists $\beta \in (0, 1)$ and $B > 0$ such that

$$\limsup_{n \to \infty} \sup_{\theta \in \mathcal{U}(\beta,n,B) \setminus \mathcal{T}(\alpha,n,D)} \left| (i) - \frac{1}{n} l_n(\theta^U) \right| \leqslant \frac{-\log(1 - \sigma_-)}{3}.$$

This concludes the proof of the first step.

Second step: controlling (ii)

Lemma 5.11. *Assume* ***(Aerg)***, ***(Aenv)*** *and* ***(Amin)****. Then*

$$\sup_{\theta\in\mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)}(ii)\leqslant\frac{1}{n}\sum_{t=1}^{n}\log\left(1+\frac{g(\{B-|Z_t^{\max}|\}_+)}{\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))}\right)$$
$$+\frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_{t\notin I^U_{n,B}(\theta)}\log\left(\frac{g(0)+\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))}{g(\{B-|Z_t^{\max}|\}_+)+\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))}\right) \tag{5.10}$$
$$=:\frac{1}{n}\sum_{t=1}^{n}h'_U(B,Z_t^{\max})+\frac{1}{n}\sum_{t\notin I^U_{n,B}(\theta)}V_t^B.$$

Proof. We show that

$$\frac{p^\theta(Y_t\,|\,X_t=x_U(\theta))}{p^\theta(Y_t\,|\,X_t\neq x_U(\theta),Y_1^{t-1})}\leqslant\begin{cases}\dfrac{g(\{B-|Z_t^{\max}|\}_+)}{\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))} & \text{if } t\in I^U_{n,B}(\theta),\\ \dfrac{g(0)}{\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))} & \text{if } t\notin I^U_{n,B}(\theta).\end{cases}$$

The lemma follows by summing over t. The details of the proof can be found in Section 5.9.2. □

Note that under Assumption **(Aint)**,

$$\mathbb{E}^*[-\log m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))]<\infty.$$

Hence,

$$\begin{aligned}\mathbb{E}^*|h'_U(0,Z_t^{\max})|&\leqslant\mathbb{E}^*\left[\left\{\log\left(\sigma_-m\left(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D)\right)+g(0)\right)\right\}_+\right]\\&\quad+\mathbb{E}^*[-\log(\sigma_-m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D)))]\\&\leqslant\log 2+|\log g(0)|+\mathbb{E}^*[-\log m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))]-\log\sigma_-\\&<\infty.\end{aligned}$$

Thus, since $b\longmapsto h'_U(b,z)$ is nonincreasing and converges to zero when $b\longrightarrow\infty$ for all z, the dominated convergence theorem together with the law of large numbers imply that there exists B such that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}h'_U(B,Z_t^{\max})\leqslant\frac{-\log(1-\sigma_-)}{6}.$$

Then, apply Lemma 5.8 to the i.i.d. non-negative random variables $(V_t^B)_{t\geqslant1}$ using the fact that $\#I^U_{n,B}(\theta)\geqslant\beta n$, which yields

$$\limsup_{n\to\infty}\sup_{\theta\in\mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)}\frac{1}{n}\sum_{t\notin I^U_{n,B}(\theta)}V_t^B\leqslant\mathbb{E}^*[V_1^B\mathbf{1}_{V_1^B\geqslant q_{V^B}(\beta)}].$$

Note that

$$\mathbb{E}^*V_1^B\leqslant\log((1+\sigma_-)g(0))-\log\sigma_-+\mathbb{E}^*[-\log m(|Z_t^{\max}|+M^{\mathcal{T}}(\alpha,D))],$$

which is finite thanks to **(Aint)**. Thus,

$$\mathbb{E}^*[V_1^B \mathbf{1}_{V_1^B \geqslant q_{V^B}(\beta)}] \underset{\beta \to 1}{\longrightarrow} 0,$$

so that there exists β such that

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)} \frac{1}{n} \sum_{t \notin I^U_{n,B}(\theta)} V_t^B \leqslant \frac{-\log(1-\sigma_-)}{6}.$$

Hence, we proved that there exists $\beta(\alpha, D) \in (0,1)$ and $B(\alpha, D) > 0$ such that

$$\limsup_{n\to\infty} \sup_{\theta \in \mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)} (ii) \leqslant \frac{-\log(1-\sigma_-)}{3},$$

which ends the second step.

Putting together the results of the two steps, one gets that for all $\alpha \in (0,1)$ and $D > 0$, there exists $\beta \in (0,1)$ and $B > 0$ such that almost surely,

$$\limsup_{n\to\infty} \left(\sup_{\theta \in \mathcal{U}(\beta,n,B)\setminus\mathcal{T}(\alpha,n,D)} \frac{1}{n} l_n(\theta) - \sup_{\theta\in\Theta} \frac{1}{n} l_n(\theta) \right) \leqslant \frac{\log(1-\sigma_-)}{3} < 0,$$

so that for n large enough, $\hat{\theta}_n \notin (\mathcal{U}(\beta,n,B) \setminus \mathcal{T}(\alpha,n,D))$.

Together, the results of Sections 5.4.1 and 5.4.2 imply that for all $\alpha \in (0,1)$, there exists $D, B > 0$ and $\beta \in (0,1)$ such that almost surely, for n large enough,

$$\hat{\theta}_n \in \Theta \setminus (\mathcal{T}(\alpha,n,D) \cup \mathcal{U}(\beta,n,B)).$$

5.5 Block approximation

Let us begin with a few definitions that will also be used in the following sections.

Definition 5.3. *Let us define*

$$\mathcal{B}^* = \mathcal{X}^*/\mathcal{R}^*$$

where the equivalence relation $\mathcal{R}^*$ is defined on $\mathcal{X}^*$ by $x\mathcal{R}^*x'$ if and only if $T^*_x - T^*_{x'}$ is constant. $\mathcal{B}^*$ is the set of blocks of true trends. For $b \in \mathcal{B}^*$, we shall use the notation abuse T^*_b to indicate $T^*_{x^*}$ where x^* is the element of b associated with the smallest trend in the class.

Let us denote by $\mathbf{b}^* : \mathcal{X}^* \longrightarrow \mathcal{B}^*$ the quotient mapping, let

$$B_t = \mathbf{b}^*(X_t)$$

be the block from which the observation Y_t is generated,

$$\Delta : x^* \in \mathcal{X}^* \longmapsto T^*_{x^*}(1) - T^*_{\mathbf{b}^*(x^*)}(1)$$

the function which maps the index of a trend to the difference between the corresponding trend and the reference trend of its block, and

$$Z'_t = Y_t - T^*_{B_t}(t)$$

the difference between an observation and the reference trend of its block. Note that $Z'_t = Z_t + \Delta(X_t)$. We finally denote by

$$E : t \in \mathbb{N}^* \longmapsto \inf_{b \neq b' \in \mathcal{B}^*} |T^*_b(t) - T^*_{b'}(t)|$$

the minimum difference between two distinct blocks of true trends at time t.

The following assumption is implied by Assumption **(Adegree)**.

(Adiv) The sequence $(E(t))_{t\geqslant 1}$ diverges to $+\infty$.

Definition 5.4. *Let $\alpha, \beta \in (0,1)$ and $D, B > 0$. The set $\Theta_n^{OK}(\alpha,\beta,D,B)$ is defined by*

$$\Theta_n^{OK}(\alpha,\beta,D,B) := \Theta \setminus (\mathcal{U}(\beta,n,B) \cup \mathcal{T}(\alpha,n,D)).$$

Let $M(\alpha,\beta,D,B)$ be the smallest $M > 0$ such that for all $n \in \mathbb{N}^*$ and $\theta \in \Theta_n^{OK}(\alpha,\beta,D,B)$,

$$\begin{cases} \forall x \in \mathcal{X}^\theta, \quad \|S_x^{\theta,n} - S^{*,n}_{x_L^*(\theta,x,n)}\|_\infty \leqslant M - \|\Delta\|_\infty & (5.11) \\ \forall x^* \in \mathcal{X}^*, \quad \exists x \in \mathcal{X}^\theta, \quad \|S_x^{\theta,n} - S^{*,n}_{x^*}\|_\infty \leqslant M - \|\Delta\|_\infty, & (5.12) \end{cases}$$

where $x_L^*(\theta,x,n)$ is defined in Section 5.3.2.

By Theorems 5.1 and 5.5, $M(\alpha,\beta,D,B)$ exists and is finite for all (α,β,D,B). For the sake of simplicity, we omit the dependency of M and Θ_n^{OK} in (α,β,D,B) in the notations. Finally, let

$$n_1 := \inf\{n \in \mathbb{N}^* \,|\, \forall t \geqslant n, E(t) > 4M\}.$$

*Note that n_1 is finite under **(Adiv)**.*

Under Assumption **(Adiv)**, equations (5.11) and (5.12) imply that for all $n \geqslant n_1$, $\theta \in \Theta_n^{\mathrm{OK}}$ and $x, x' \in \mathcal{X}^\theta$,

$$\|S_x^{\theta,n} - S_{x'}^{\theta,n}\|_\infty \leqslant 2M \iff \mathbf{b}^*(x_L^*(\theta,x,n)) = \mathbf{b}^*(x_L^*(\theta,x',n)). \tag{5.13}$$

From now on, we consider $n \geqslant n_1$ and $\theta \in \Theta_n^{\mathrm{OK}}$. Let us consider the quotient space

$$\mathcal{B}^{\theta,n} = \mathcal{X}^\theta / \mathcal{R}_M$$

where the equivalence relation $\mathcal{R}_M$ is defined by $x\mathcal{R}_M x'$ if and only if $\|S_x^{\theta,n} - S_{x'}^{\theta,n}\|_\infty \leqslant 2M$ (equation 5.13 shows that this is indeed an equivalence relation). $\mathcal{B}^{\theta,n}$ is the set of trend blocks associated with θ.

Equation (5.13) proves that there is an injection

$$b \in \mathcal{B}^{\theta,n} \longmapsto \mathbf{b}^*(x_L^*(\theta,x_b,n)) \in \mathcal{B}^*$$

where x_b is a representative of b (it does not matter which one). This mapping is also surjective: equations (5.12) and(5.11) imply that for all $x^* \in \mathcal{X}^*$, there exists $x \in \mathcal{X}^\theta$ such that

$$\|S_{x^*}^{*,n} - S^{*,n}_{x_L^*(\theta,x,n)}\|_\infty \leqslant 2(M - \|\Delta\|_\infty),$$

so that under Assumption **(Adiv)**, for $n \geqslant n_1$, $\mathbf{b}^*(x^*) = \mathbf{b}^*(x_L^*(\theta,x,n))$. Thus, this mapping is a bijection. In other words, $\mathcal{B}^{\theta,n}$ can be identified to $\mathcal{B}^*$, for all $n \geqslant n_1$ and $\theta \in \Theta_n^{\mathrm{OK}}$. This is what we are doing in the following.

Definition 5.5. *For all $\theta \in \Theta_n^{OK}$, we denote by $\mathbf{b}^\theta : \mathcal{X}^\theta \longrightarrow \mathcal{B}^*$ the function that maps a state to its equivalence class.*

Note that by equation (5.11), for all $n \geqslant n_1$, $\theta \in \Theta_n^{\mathrm{OK}}$ and $x \in \mathcal{X}^\theta$,

$$\sup_{t\in\{1,\dots,n\}} |T^*_{\mathbf{b}^\theta(x)}(t) - T_x^\theta(t)| \leqslant M. \tag{5.14}$$

5.5.1 Idea

The purpose of this section is to prove the following approximations.

$$\begin{aligned}\log p^{\theta}_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) &\approx \log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\\ &\approx \log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1})\\ &= \log p^{\theta}_{Z'_t,B_t|(Z')_1^{t-1},B_1^{t-1}}(Z'_t,B_t|(Z')_1^{t-1},B_1^{t-1}).\end{aligned}$$

The goal is to show that one may assume that the blocks are observed without changing anything to the asymptotic properties of the MLE. We shall see in Section 5.6 that the resulting process $(Z'_t,B_t)_{t\geqslant 1}$ is almost homogeneous: its distribution evolves slowly enough that it behaves locally like a homogeneous HMM.

5.5.2 Step 1: introduction of the trend block B_t in the log-likelihood

Assumptions used: **(Aerg)**, **(Aenv)**, **(Amin)**, **(Aint)** and **(Adiv)**.

The first step consists in showing that the following quantity tends to 0 uniformly in θ.

$$\begin{aligned}&\log p^{\theta}_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\\ &\qquad = \log\left(\frac{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t=x_t|Y_1^{t-1})\gamma_{x_t}(Y_t-T^\theta_{x_t}(t))}{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t=x_t|Y_1^{t-1})\gamma_{x_t}(Y_t-T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)=B_t}}\right).\end{aligned}$$

Note that we can rewrite this as

$$\begin{aligned}&\log p^{\theta}_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\\ &\qquad = \log(p^\theta_{B_t|Y_1^t}(B_t|Y_1^t)^{-1})\\ &\qquad = \log\left(\left\{1-p^\theta_{B_t|Y_1^t}\left(\{b\in\mathcal{B}^*\text{ s.t. }b\neq B_t\}|Y_1^t\right)\right\}^{-1}\right).\end{aligned}\tag{5.15}$$

Intuitively, when t is large, since the trends get further from one another, the probability to get the wrong block converges to zero.

Lemma 5.12. *Assume **(Aenv)**, **(Amin)** and **(Aerg)**. Then for all $t\in\mathbb{N}^*$,*

$$\begin{aligned}&\sup_{\theta\in\Theta_n^{OK}}\left|\log p^{\theta}_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\\ &\qquad\leqslant \log\left(\left\{1-\frac{g(\{E(t)-M-|Z_t^{\max}|-\|\Delta\|_\infty\}_+)}{\sigma_- m(M+|Z_t^{\max}|+\|\Delta\|_\infty)}\right\}_+^{-1}\wedge\frac{g(0)}{\sigma_- m(M+|Z_t^{\max}|+\|\Delta\|_\infty)}\right)\\ &\qquad=: h(E(t),Z_t^{\max})\end{aligned}$$

using the convention $\{z\}_+^{-1}=+\infty$ if $z\leqslant 0$.

The first part of the infimum can be understood thanks to equation (5.15). The second one ensures that the upper bound is integrable.

Proof. Proof in Section 5.9.3. □

After summing, one gets

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\leqslant\frac{1}{n}\sum_{t=1}^{n}h(E(t),Z_t^{\max}).$$

The function $e\in\mathbb{R}_+\longmapsto h(e,z)$ is non-negative, non-increasing for all $z\in\mathbb{R}$ and tends to 0 as e tends to $+\infty$. Moreover, under Assumption **(Aint)**, $h(0,Z_1^{\max})$ is integrable by definition of h. Thus, under Assumption **(Aint)**, the law of large numbers and **(Adiv)** imply that for all $E>0$

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\leqslant\mathbb{E}^*[h(E,Z_1^{\max})].$$

Hence the dominated convergence theorem ensures that $\mathbb{E}^*[h(E,Z_1^{\max})]\longrightarrow 0$ as $E\longrightarrow+\infty$. Thus, almost surely,

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right|\underset{n\to\infty}{\longrightarrow}0. \tag{5.16}$$

5.5.3 Step 2: conditioning on the blocks B_1^{t-1}

Assumptions used: **(Aenv)**, **(Amin)**, **(Aerg)**, **(Aorder)** and **(Adiv)**.

The following lemma is a consequence of the lower bound on the transition matrices, see for instance Lemma 1 and Corollary 1 of Douc et al. (2004).

Lemma 5.13 (Exponential forgetting)**.** *Under Assumption **(Aerg)**, there exists $C>0$ such that for all $n\in\mathbb{N}^*$, $y_1^n\in\mathbb{R}^n$, $\theta\in\Theta$ and for all probability measures π,π' on $\mathcal{X}^\theta$:*

$$\sum_{x\in\mathcal{X}^\theta}|p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)-p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi')|\leqslant C\rho^n$$

with $\rho=1-\frac{\sigma_-}{1-\sigma_-}\in(0,1)$.

Besides, under **(Aerg)**, for all $\theta\in\Theta$, $x\in\mathcal{X}^\theta$, $y_1^{n-1}\in\mathbb{R}^n$ and for all probability measure π on $\mathcal{X}^\theta$:

$$p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)\geqslant\sigma_-.$$

Hence, using the inequality $|\log x-\log y|\leqslant\frac{|x-y|}{x\wedge y}$ for all $x,y>0$: for all $n\in\mathbb{N}^*$, $\theta\in\Theta$, $y_1^n\in\mathbb{R}^n$, $b\in\mathcal{B}^*$ and for all probability measures π,π' on $\mathcal{X}^\theta$:

$$\begin{aligned}
&|\log p^{\theta}_{Y_n,B_n|Y_1^{n-1}}(y_n,b|y_1^{n-1},X_0\sim\pi)-\log p^{\theta}_{Y_n,B_n|Y_1^{n-1}}(y_n,b|y_1^{n-1},X_0\sim\pi')|\\
&\leqslant\frac{\displaystyle\sum_{x\in\mathcal{X}^\theta}|p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi)-p^{\theta}_{X_n|Y_1^{n-1}}(x|y_1^{n-1},X_0\sim\pi')|p^{\theta}_{Y_n,B_n|X_n}(y_n,b|x)}{\sigma_-\displaystyle\sum_{x\in\mathcal{X}^\theta}p^{\theta}_{Y_n,B_n|X_n}(y_n,b|x)}\\
&\leqslant\frac{C}{\sigma_-}\rho^n.
\end{aligned}$$

Changing the constant C if necessary, we have for all $a\in\mathbb{N}^*$:

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})-\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(Y_t,B_t|Y_1^{t-1},B_1^{t-a})\right|\leqslant C\rho^a. \tag{5.17}$$

It remains to condition on B_{t-a+1}^{t-1}.

Lemma 5.14. *Assume* ***(Aerg)***, ***(Aorder)***, ***(Aenv)*** *and* ***(Amin)****. Then for all* $a \in \mathbb{N}^*$,

$$\sup_{\theta\in\Theta_n^{OK}} \left| \frac{1}{n}\sum_{t=1}^{n} \log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(Y_t,B_t|Y_1^{t-1},B_1^{t-a}) - \frac{1}{n}\sum_{t=1}^{n} \log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1}) \right|$$
$$\leqslant \frac{2aK^2}{\sigma_-^3}\frac{1}{n}\sum_{i=1}^{n-1}\left(1 \wedge \frac{g(\{E(i)-M-|Z_i^{\max}|-\|\Delta\|_\infty\}_+)}{m(M+|Z_i^{\max}|+\|\Delta\|_\infty)}\right)$$
$$=: \frac{2aK^2}{\sigma_-^3}\frac{1}{n}\sum_{i=1}^{n-1} h'(E(i),Z_i^{\max}).$$

Proof. We show that when **(Aerg)** and **(Aorder)** hold, one has for all $\theta \in \Theta_n^{\mathrm{OK}}$, $t \leqslant n$, $a \in \mathbb{N}^*$, $y_1^t \in \mathcal{Y}^t$ and $b_1^t \in (\mathcal{B}^*)^t$,

$$\left| \log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a}) - \log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1}) \right|$$
$$\leqslant \frac{2}{\sigma_-} p^{\theta}_{B^{t-1}_{(t-a+1)\vee 1}|Y_1^{t-1},B_1^{t-a}}\left((\mathcal{B}^*)^{(a\wedge t)-1}\setminus\{b^{t-1}_{(t-a+1)\vee 1}\}|y_1^{t-1},b_1^{t-a}\right) \tag{5.18}$$
$$\leqslant \frac{2K^2}{\sigma_-^3}\sum_{i=(t-a+1)\vee 1}^{t-1} \frac{\sup_{x_i\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_i)\neq b_i} \gamma^\theta_{x_i}(y_i - T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(y_i-T^\theta_x(i))}. \tag{5.19}$$

Then, we show that under **(Aenv)** and **(Amin)**, one has for all $\theta \in \Theta_n^{\mathrm{OK}}$ and $i \leqslant n$

$$\frac{\sup_{x_i\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_i)\neq B_i} \gamma^\theta_{x_i}(Y_i - T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_x(Y_i-T^\theta_x(i))} \leqslant 1 \wedge \frac{g(\{E(i)-M-|Z_i^{\max}|-\|\Delta\|_\infty\}_+)}{m(M+|Z_i^{\max}|+\|\Delta\|_\infty)} \tag{5.20}$$
$$=: h'(E(i),Z_i^{\max}),$$

and the lemma follows by summing over t and i. The details of the proof can be found in Section 5.9.3. □

Therefore, one has almost surely

$$\limsup_{n\to+\infty}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1}) - \frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1})\right|$$
$$\leqslant \limsup_{n\to+\infty}\left(C\rho^a + \frac{2aK^2}{\sigma_-^3}\frac{1}{n}\sum_{i=1}^{n-1}h'(E(i),Z_i^{\max})\right)$$
$$\leqslant C\rho^a + \frac{2aK^2}{\sigma_-^3}\mathbb{E}^*[h'(E,Z_1^{\max})]$$
$$\leqslant C'\left(-\mathbb{E}^*[h'(E,Z_1^{\max})]\log\mathbb{E}^*[h'(E,Z_1^{\max})]\right)$$

for some explicit constant C' and for all E sufficiently large to have $\mathbb{E}^*[h'(E,Z_1^{\max})] < 1/2$ using Assumption **(Adiv)**, the law of large numbers, the fact that the mapping $e \longmapsto h'(e,z)$ is non-negative, bounded by 0 and 1 and non-increasing for all z, and by taking $a = \lceil \frac{\log\mathbb{E}^*[h'(E,Z_1^{\max})]}{\log\rho}\rceil$. Since the function $e \longmapsto h'(e,z)$ tends to 0 as e tends to $+\infty$ for all z, the dominated convergence theorem ensures that almost surely,

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1}) - \frac{1}{n}\sum_{t=1}^{n}\log p^{\theta}_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(Y_t,B_t|Y_1^{t-1},B_1^{t-1})\right| \underset{n\to\infty}{\longrightarrow} 0.$$

Combining the above equation with equation (5.16) yields that almost surely,

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-\frac{1}{n}\log p^{\theta}_{(Y,B)_1^n}((Y,B)_1^n)\right|\underset{n\to+\infty}{\longrightarrow}0.$$

Note that $Y_t=Z'_t+T^*_{B_t}(t)$ by definition, so that the previous equation is equivalent to

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-\frac{1}{n}\log p^{\theta}_{(Z',B)_1^n}((Z',B)_1^n)\right|\underset{n\to+\infty}{\longrightarrow}0.$$

5.5.4 Application: existence and finiteness of the relative entropy rate

Since $\theta^*\in\Theta_n^{\mathrm{OK}}$ for all n, one has

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta^*)-\frac{1}{n}\log p^{*}_{(Z',B)_1^n}((Z',B)_1^n)\right|\underset{n\to+\infty}{\longrightarrow}0.$$

Under θ^*, the process $(X_t,(Z'_t,B_t))_{t\geqslant 1}$ is a stationary and ergodic HMM with emission densities $((z',b)\longmapsto\gamma^*_{x^*}(z'-\Delta(x^*))\otimes\mathbf{1}_{\mathbf{b}^*(x^*)}(b))_{x^*\in\mathcal{X}^*}$ with respect to the measure $\mathrm{Leb}\otimes\mu_{\mathcal{B}^*}$, where $\mu_{\mathcal{B}^*}$ is the counting measure on $\mathcal{B}^*$.

Since it is stationary and ergodic, Barron (1985) shows that there exists $l(\theta^*)>-\infty$ such that

$$\frac{1}{n}l_n^{(Z',B)}(\theta^*)\longrightarrow l(\theta^*).$$

Then, all emission densities are upper bounded by $g(0)$ under **(Aenv)**, so that the positive part of their logarithm is integrable. Therefore, Leroux (1992) implies that $l(\theta^*)<+\infty$.

To sum up:

Lemma 5.15. *Assume **(Aenv)**, **(Amin)**, **(Aerg)** and **(Adegree)**. Then there exists a finite $l(\theta^*)$ such that*

$$\frac{1}{n}l_n(\theta^*)\underset{n\to\infty}{\longrightarrow}l(\theta^*),$$

*or in other word **(Arate)** holds.*

5.6 Integrated log-likelihood

In this section, we use the fact that the observed process $(Y_t)_{t\geqslant 1}$ may be replaced by the process $(Z'_t,B_t)_{t\geqslant 1}$. While this process is not homogeneous, its distribution varies slowly over time, thanks to Theorems 5.1 and 5.5. We take advantage of this property to show the uniform convergence of the log-likelihood by approximating $(Z'_t,B_t)_{t\geqslant 1}$ by an homogenized process. The limit can be written as an integral of limits of log-likelihoods of homogeneous HMMs, hence the name *integrated log-likelihood*.

5.6.1 Convergence of the likelihood to the integrated log-likelihood

Assumptions used: **(Aerg)**, **(Aenv)**, **(Amin)**, **(Aint)**, **(Adegree)**, **(Aorder)**, **(Aparam)** and **(Areg)**.

The normalized log-likelihood associated with the HMM $(Z'_t, B_t)_{t\geqslant 1}$ introduced in Section 5.5 can be written as

$$\begin{aligned}\frac{1}{n}l_n^{(Z',B)}(\theta) &= \frac{1}{n}\log \sum_{x_1^n \text{ s.t. } \forall t,\, \mathbf{b}^\theta(x_t)=B_t} \pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n)\prod_{t=1}^n \gamma^\theta_{x_t}(Z'_t + T^*_{B_t}(t) - T^\theta_{x_t}(t)) \\ &= \frac{1}{n}\log \sum_{x_1^n \text{ s.t. } \forall t,\, \mathbf{b}^\theta(x_t)=B_t} \pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n)\prod_{t=1}^n \gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\frac{t}{n}\right)\right),\end{aligned} \tag{5.21}$$

with $D^{\theta,n}_x = S^{\theta,n}_x - S^{*,n}_{\mathbf{b}^\theta(x)}$. Theorem 5.5 ensures that there exists an integer $n_2 \geqslant n_1$ such that the set

$$\mathcal{D} := \bigcup_{n\geqslant n_2}\left\{D^{\theta,n}_x \,|\, \theta\in\Theta_n^{\mathrm{OK}}, x\in\mathcal{X}^\theta\right\} \tag{5.22}$$

is uniformly equicontinuous and uniformly bounded by M. Hence there exists a continuity modulus ν such that for all $\delta>0$,

$$|s-t|\leqslant\delta \;\Rightarrow\; \sup_{n\geqslant n_2}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\sup_{x\in\mathcal{X}^\theta}|D^{\theta,n}_x(s) - D^{\theta,n}_x(t)| \leqslant \nu(\delta). \tag{5.23}$$

Definition 5.6 (Log-likelihood of the homogenized process)**.** *For $\eta>0$ and $\theta\in\Theta_n^{OK}$, let*

$$\begin{aligned}\frac{1}{n}l_n^{(Z',B)}[\eta](\theta) := \frac{1}{n}\log \sum_{x_1^n \;t.q.\; \forall t,\, \mathbf{b}^\theta(x_t)=B_t} &\pi^\theta(x_1)Q^\theta(x_1,x_2)\dots Q^\theta(x_{n-1},x_n) \\ &\times\prod_{t=1}^n \gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\eta\left\lfloor\frac{t}{\eta n}\right\rfloor\right)\right),\end{aligned}$$

be the normalized log-likelihood of the process where each trend is made constant over segments of length η.

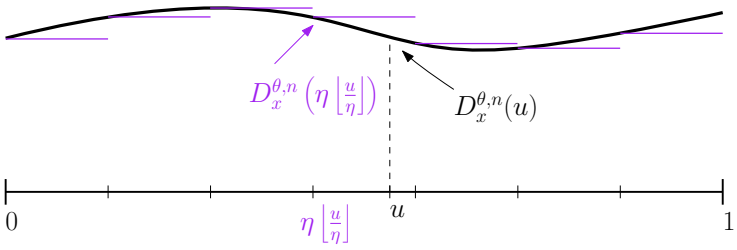


Figure 5.3: Construction of the trends of the homogenized process

$\frac{1}{n}l_n^{(Z',B)}[\eta](\theta)$ is an approximation of the log-likelihood of equation (5.21). Under Assumption **(Areg)**, equation (5.22) ensures that for all $\delta>0$, $n\geqslant n_2$, $\theta\in\Theta_n^{\mathrm{OK}}$, $x\in\mathcal{X}^\theta$ and $t\in\{1,\dots,n\}$,

$$\gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\frac{t}{n}\right)\right) \in \left[e^{-L(|Z'_t|+M)\omega(\nu(\delta))}, e^{L(|Z'_t|+M)\omega(\nu(\delta))}\right]\gamma^\theta_{x_t}\left(Z'_t - D^{\theta,n}_{x_t}\left(\delta\left\lfloor\frac{t}{\delta n}\right\rfloor\right)\right),$$

hence, for all $\delta>0$ and $n\geqslant n_2$,

$$\begin{aligned}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n^{(Z',B)}(\theta) - \frac{1}{n}l_n^{(Z',B)}[\delta](\theta)\right| &\leqslant \omega(\nu(\delta))\times\frac{1}{n}\sum_{t=1}^n L(|Z'_t|+M) \\ &\leqslant \omega(\nu(\delta))\times\frac{1}{n}\sum_{t=1}^n L(\|\Delta\|_\infty + M + |Z^{\max}_t|).\end{aligned} \tag{5.24}$$

Remark. *Under **(Areg)**, the law of large numbers entails that almost surely, for all $N \geqslant 1$,*

$$\limsup_{n\to+\infty} \sup_{\theta\in\Theta_n^{OK}} \left| \frac{1}{n} l_n^{(Z',B)}(\theta) - \frac{1}{n} l_n^{(Z',B)} \left[\frac{1}{N} \right] (\theta) \right| \leqslant \omega\left(\nu\left(\frac{1}{N}\right)\right) \mathbb{E}^*[L(\|\Delta\|_\infty + M + |Z_1^{\max}|)].$$

Definition 5.7. *For all $K' \in \mathbb{N}^*$, for all K'-uple $\gamma = (\gamma_x)_{x\in[K']}$ of nonnegative measurable functions and for all $\mathbf{D} = (D_x)_{x\in[K']} \in \mathbb{R}^{K'}$, let*

$$\tau(\gamma, \mathbf{D}) := (z' \longmapsto \gamma_x(z' - D_x))_{x\in[K']}$$

the vector of functions γ translated by the vector $\mathbf{D}$.

For a probability measure π on $[K']$, a $K' \times K'$ transition matrix Q, a vector γ of K' emission distributions, we denote by $\frac{1}{n} l_n^{stat}(\pi, Q, \gamma, \mathbf{b})\{(z', b)_1^n\}$ (resp. $l^{stat}(Q, \gamma, \mathbf{b})$) the normalized log-likelihood associated with the observations $(z', b)_1^n$ (resp. the limit of the log-likelihood) of the HMM $((Z', B)_t)_{t\geqslant 1}$ with parameters $(\pi, Q, \gamma, \mathbf{b})$, if it exists, that is

$$\frac{1}{n} l_n^{stat}(\pi, Q, \gamma, \mathbf{b})\{(z', b)_1^n\} = \frac{1}{n} \log \sum_{x_1^n \in [K']^n} \pi(x_1) Q(x_1, x_2) \dots Q(x_{n-1}, x_n) \prod_{t=1}^n \gamma_{x_t}(z'_t) \mathbf{1}_{\mathbf{b}(x_t)=b_t}$$

and

$$l^{stat}(Q, \gamma, \mathbf{b}) = \lim_{n\to\infty} \frac{1}{n} l_n^{stat}(\pi, Q, \gamma, \mathbf{b})\{(Z', B)_1^n\}$$

where the limit is taken almost surely under $\mathbb{P}^*$ (we show that it exists in the following Lemma).

Lemma 5.16. *Assume **(Aerg)**, **(Aenv)**, **(Amin)** and **(Aint)**. Let $K' \in \mathbb{N}^*$. Then, for all $\xi \in \Xi_{K'}$, $\mathbf{D} \in [-M, M]^{K'}$ and $\mathbf{b} : [K'] \longrightarrow \mathcal{B}^*$, the quantity*

$$l^{stat}(Q_\xi, \tau(\gamma_\xi, \mathbf{D}), \mathbf{b})$$

exists and is finite.

*Let us denote by $Cl(\mathcal{D})$ the closure of $\mathcal{D}$ as defined in equation (5.22) (with respect to the topology of the supremum norm). If we further assume **(Adegree)**, **(Aorder)** and **(Aparam)**, then for all $K' \in \mathbb{N}^*$, the mapping*

$$(\xi, \mathbf{D}, u, \mathbf{b}) \in \Xi_{K'} \times Cl(\mathcal{D})^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'} \longmapsto l^{stat}(Q_\xi, \tau(\gamma_\xi, \mathbf{D}(u)), \mathbf{b}) \tag{5.25}$$

*is continuous (when $Cl(\mathcal{D})$ is equipped with the supremum norm). Its domain is compact because of Assumption **(Aparam)** and Theorem 5.5, so that this mapping is uniformly continuous.*

In addition, for all $N \in \mathbb{N}^*$,

$$\sup_{(\pi,\xi,\mathbf{D},u,\mathbf{b})} \sup_{s\in\{0,\dots,(N-1)n\}} \left| \frac{1}{n} l_n^{stat}(\pi, Q_\xi, \tau(\gamma_\xi, \mathbf{D}(u)), \mathbf{b})\{(Z', B)_{s+1}^{s+n}\} \right.$$
$$\left. - l^{stat}(Q_\xi, \tau(\gamma_\xi, \mathbf{D}(u)), \mathbf{b}) \right| \underset{n\to\infty}{\longrightarrow} 0$$

where the supremum is taken for all $(\pi, \xi, \mathbf{D}, u, \mathbf{b}) \in \Delta_{K'} \times \Xi_{K'} \times Cl(\mathcal{D})^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'}$.

Proof. Proof in Section 5.9.4 □

As a consequence, the family of functions

$$\bigcup_{K'\leqslant K} \{u \in [0,1] \longmapsto l^{\text{stat}}(Q_\xi, \tau(\gamma_\xi, \mathbf{D}(u)), \mathbf{b})\}_{\xi\in\Xi_{K'}, \mathbf{D}\in\mathcal{D}^{K'}, \mathbf{b}\in(\mathcal{B}^*)^{K'}}$$

is uniformly equicontinuous, which ensures the following result.

Corollary 5.17 (Riemann approximation of the integral)**.** *The quantity*

$$R_N := \sup_{n\geqslant n_1}\sup_{\theta\in\Theta_n^{OK}}\left|\frac{1}{N}\sum_{i=0}^{N-1} l^{stat}\left(Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}\left(\frac{i}{N}\right)\right),\mathbf{b}^\theta\right)-\int l^{stat}\left(Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}(u)\right),\mathbf{b}^\theta\right)du\right| \tag{5.26}$$

satisfies

$$R_N \underset{N\to+\infty}{\longrightarrow} 0.$$

By the triangle inequality and using Equations (5.26) and (5.24), for all $n\geqslant n_1$ and $N\in\mathbb{N}^*$,

$$\begin{aligned}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}&\left|\frac{1}{n}l_n^{(Z',B)}(\theta)-\int l^{\mathrm{stat}}\left(Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}(u)\right),\mathbf{b}^\theta\right)du\right|\\ &\leqslant \omega\left(\nu\left(\frac{1}{N}\right)\right)\frac{1}{n}\sum_{t=1}^{n}L(\|\Delta\|_\infty+|Z_t^{\max}|)+R_N\\ &\quad+\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n^{(Z',B)}\left[\frac{1}{N}\right](\theta)-\frac{1}{N}\sum_{i=0}^{N-1}l^{\mathrm{stat}}\left(Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}\left(\frac{i}{N}\right)\right),\mathbf{b}^\theta\right)\right|.\end{aligned}$$

For the sake of simplicity, let us assume that $\frac{n}{N}$ is an integer. Note that for all $\theta\in\Theta_n^{\mathrm{OK}}$,

$$\begin{aligned}\frac{1}{n}l_n^{(Z',B)}\left[\frac{1}{N}\right](\theta)&=\frac{1}{n}\sum_{i=0}^{N-1}\log p^{(Q^\theta,\tau(\gamma^\theta,\mathbf{D}^{\theta,n}(\frac{i}{N})),\mathbf{b}^\theta)}\left((Z',B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}}\mid(Z',B)_1^{i\frac{n}{N}}\right)\\ &=\frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{\frac{n}{N}}l_{\frac{n}{N}}^{\mathrm{stat}}\left(\pi_{i\frac{n}{N}}^\theta,Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}\left(\frac{i}{N}\right)\right),\mathbf{b}^\theta\right)\left\{(Z',B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}}\right\},\end{aligned}$$

where $\pi_{i\frac{n}{N}}^\theta$ is defined as the distribution of $X_{1+i\frac{n}{N}}$ conditionally to $(Z',B)_{1+i\frac{n}{N}}^{\frac{n}{N}+i\frac{n}{N}}$ under the parameter θ. Hence, Lemma 5.16 implies that almost surely,

$$\begin{aligned}\limsup_{n\to+\infty}\sup_{\theta\in\Theta_n^{\mathrm{OK}}}&\left|\frac{1}{n}l_n^{(Z',B)}(\theta)-\int l^{\mathrm{stat}}\left(Q^\theta,\tau\left(\gamma^\theta,\mathbf{D}^{\theta,n}(u)\right),\mathbf{b}^\theta\right)du\right|\\ &\leqslant\inf_{N\in\mathbb{N}^*}\left[\omega\left(\nu\left(\frac{1}{N}\right)\right)\mathbb{E}^*[L(\|\Delta\|_\infty+|Z_1^{\max}|)]+R_N\right]\\ &=0.\end{aligned}$$

To sum up: let

$$l^{\mathrm{int}}:(\xi,\mathbf{D},\mathbf{b})\in\Xi\times\mathcal{D}^K\times(\mathcal{B}^*)^K\longmapsto\int l^{\mathrm{stat}}(Q_\xi,\tau(\gamma_\xi,\mathbf{D}(u)),\mathbf{b})du$$

be what we call the *integrated log-likelihood*. Then l^{int} is continuous by uniform continuity of l^{stat} and one has almost surely

$$\sup_{\theta\in\Theta_n^{\mathrm{OK}}}\left|\frac{1}{n}l_n(\theta)-l^{\mathrm{int}}(\xi(\theta),\mathbf{D}^{\theta,n},\mathbf{b}^\theta)\right|\underset{n\to\infty}{\longrightarrow}0. \tag{5.27}$$

5.6.2 Maximizers of the integrated log-likelihood and identifiability

Assumptions used: **(Aenv)**, **(Amin)**, **(Aparam)**, **(Areg)**, **(Aid)** and **(Acentering)**.

In this section, we assume that $K^* = K$ is known. We identify $\mathcal{X}^\theta$, $\mathcal{X}^*$ and $[K]$ for all $\theta \in \Theta_K$.

Lemma 5.18. *Assume **(Aid)**, **(Aparam)**, **(Areg)**, **(Aenv)**, **(Amin)** and **(Acentering)**. Let $(\xi, \mathbf{D}, \mathbf{b}) \in \Xi_K \times \mathcal{D}(\mathcal{Y})^K \times (\mathcal{B}^*)^K$ be a maximizer of l^{int}, then $\mathbf{D}$ is constant and $(\mathbf{D}, Q_\xi, \gamma_\xi, \mathbf{b}) = (\Delta, Q^*, \gamma^*, \mathbf{b}^*)$ up to permutation of the hidden states.*

Remark. *Assumptions **(Areg)**, **(Aenv)** and **(Amin)** can be replaced by*

$$\begin{cases} \forall \theta \in \Theta, \quad \forall x \in \mathcal{X}^*, \quad z \in \mathbb{R} \longmapsto \gamma^\theta(z) \text{ is continuous,} \\ \forall x \in \mathcal{X}^*, \quad \gamma_x^*(z) \underset{|z| \to +\infty}{\longrightarrow} 0 \\ \forall x \in \mathcal{X}^*, \quad \forall z \in \mathbb{R}, \quad \gamma_x^*(z) > 0. \end{cases}$$

Proof. The maximum of l^{int} is reached by $(\xi, \mathbf{D}, \mathbf{b})$ if and only if the integrand is maximal for almost every $u \in [0,1]$, which means under **(Aid)**

$$\Big(Q_\xi, \big(\gamma_\xi(\cdot - D_x(u)) \otimes \mathbf{1}_{\mathbf{b}(x)}\big)_{x \in \mathcal{X}^*}\Big) = \Big(Q^*, \big(\gamma_x^*(\cdot - \Delta(x)) \otimes \mathbf{1}_{\mathbf{b}^*(x)}\big)_{x \in \mathcal{X}^*}\Big)$$

up to permutation of the hidden states for all $u \in [0,1]$.

Let us assume that the permutation is not constant at u. Since there are only a finite number of possible permutations of $[K^*]$, there exists two sequences $(u_i)_{i \geqslant 1}$ and $(v_i)_{i \geqslant 1}$ converging to u, one corresponding to a permutation p and the other to a permutation $p' \neq p$, that is

$$\forall i \geqslant 1, \quad \forall x \in \mathcal{X}^*, \quad \begin{cases} \gamma_{\xi,x}(\cdot - D_x(u_i)) = \gamma_{p(x)}^*(\cdot - \Delta(p(x))) \quad \text{and} \quad \mathbf{b}(x) = \mathbf{b}^*(p(x)) \\ \gamma_{\xi,x}(\cdot - D_x(v_i)) = \gamma_{p'(x)}^*(\cdot - \Delta(p'(x))) \quad \text{and} \quad \mathbf{b}(x) = \mathbf{b}^*(p'(x)) \end{cases}$$

Therefore, by continuity under **(Aparam)**, one has for all $x \in \mathcal{X}^*$

$$(\gamma_{p(x)}^*(\cdot - \Delta(p(x))), \mathbf{b}^*(p(x))) = (\gamma_{p'(x)}^*(\cdot - \Delta(p'(x))), \mathbf{b}^*(p'(x))),$$

so that $p = p'$ according to **(Aid)**, which contradicts the assumption that the permutation is not constant in u. Therefore, the permutation does not depend on u.

One may assume without loss of generality that the permutation is the identity, in other words $Q_\xi = Q^*$, $\mathbf{b} = \mathbf{b}^*$ and

$$\forall u \in [0,1], \quad \forall x \in \mathcal{X}^*, \quad \gamma_{\xi,x}(\cdot - D_x(u)) = \gamma_x^*(\cdot - \Delta(x)).$$

Here, we took u in the whole segment $[0,1]$ instead of a subset with measure 1 because the mapping $u \in [0,1] \longmapsto \gamma_{\xi,x}(\cdot - D_x(u))$ is continuous under **(Areg)**. If D_x is not constant at some $x \in \mathcal{X}^*$, this entails that γ_x^* is invariant by translation, so that it is constant, which contradicts **(Aenv)**. Therefore, $\mathbf{D}$ is constant.

Finally, one has

$$\forall x \in \mathcal{X}^*, \quad \frac{1}{2} = \int_{z \leqslant D_x} \gamma_{\xi,x}(z - D_x) dz = \int_{z \leqslant D_x} \gamma_x^*(z - \Delta(x)) dz$$

using **(Acentering)**, so that D_x is a median of γ_x^*. To conclude, note that under **(Amin)** and **(Acentering)**, $\Delta(x)$ is the only median of γ_x^*. □

5.7 Consistency

In this section, we assume that $K^* = K$ is known. We identify $\mathcal{X}^\theta$, $\mathcal{X}^*$ and $[K]$ for all $\theta \in \Theta_K$.

We showed in Section 5.4 that almost surely, there exists a (random) n_0 such that for all $n \geqslant n_0$, $\hat{\theta}_n \in \Theta_n^{\mathrm{OK}}$. For $n \geqslant n_0$, let

$$\begin{cases} \xi_n := \xi(\hat{\theta}_n), \\ \mathbf{D}_n := D^{\hat{\theta}_n, n}, \\ \mathbf{b}_n := \mathbf{b}^{\hat{\theta}_n}. \end{cases}$$

For all $n \geqslant n_0$, $(\xi_n, \mathbf{D}_n, \mathbf{b}_n) \in \Xi_K \times \mathrm{Cl}(\mathcal{D})^K \times (\mathcal{B}^*)^K$, and this set is compact by **(Aparam)** and Theorem 5.5. Let $(\xi, \mathbf{D}, \mathbf{b})$ be the limit of a convergent subsequence $(\xi_{\varphi(n)}, \mathbf{D}_{\varphi(n)}, \mathbf{b}_{\varphi(n)})_{n \geqslant 1}$, then by continuity of l^{int} and by the uniform convergence of equation (5.27), one has

$$\frac{1}{\varphi(n)} l_{\varphi(n)}(\hat{\theta}_n) \underset{n \to \infty}{\longrightarrow} l^{\mathrm{int}}(\xi, \mathbf{D}, \mathbf{b}) \leqslant l^{\mathrm{int}}(\xi(\theta^*), \Delta, \mathbf{b}^*) = l(\theta^*),$$

by Lemma 5.18, and by definition of the MLE

$$\frac{1}{\varphi(n)} l_{\varphi(n)}(\theta^*) \leqslant \frac{1}{\varphi(n)} l_{\varphi(n)}(\hat{\theta}_n).$$

Hence $l^{\mathrm{int}}(\xi, \mathbf{D}, \mathbf{b}) = l(\theta^*) = l^{\mathrm{int}}(\xi(\theta^*), \Delta, \mathbf{b}^*)$, which means that $(Q_\xi, \gamma_\xi, \mathbf{D}, \mathbf{b}) = (Q^*, \gamma^*, \Delta, \mathbf{b}^*)$ up to permutation of the hidden states by Lemma 5.18. Thus, the MLE sequence $(\xi_n, \mathbf{D}_n, \mathbf{b}_n)_{n \geqslant 1}$ has only one possible limit: the true parameters. Therefore, it converges almost surely to the true parameters up to permutation of the hidden states. More precisely, one has almost surely, after permutation of the hidden states:

$$\begin{cases} Q^{\hat{\theta}_n} \underset{n \to \infty}{\longrightarrow} Q^*, \\ \forall z \in \mathbb{R}, \quad \forall x \in [K], \quad \gamma_x^{\hat{\theta}_n}(z) \underset{n \to \infty}{\longrightarrow} \gamma_x^*(z), \\ \forall x \in [K], \quad \|D_x^{\hat{\theta}_n, n} - \Delta(x)\|_\infty \underset{n \to \infty}{\longrightarrow} 0, \\ \mathbf{b}^{\hat{\theta}_n} = \mathbf{b}^* \quad \text{for } n \text{ large enough}, \end{cases}$$

and the last two points imply

$$\forall x \in [K], \quad \|T_x^{\hat{\theta}_n} - T_x^*\|_{\infty, [0,n]} \underset{n \to \infty}{\longrightarrow} 0.$$

5.8 Simulations

A naive approach to retrieve the blocks of trends when they diverge is using a clustering algorithm in the plane (t, Y_t), as shown in Figure 5.4. If the trends diverge, the clustering algorithm will identify the true blocks of trends. Then we can apply a simple maximum likelihood method separately for each block in order to retrieve the other parameters. It only works if the trends have indeed diverged in the range of observations, that is, the blocks are well separated. Figure 5.4 shows an example with two linear trends, gaussian emission distributions and 2000 observations (left panel). The colors correspond to the true states. If we consider the first 500 observations only (center panel), the data points are not well separated, thus a clustering algorithm will not identify the two states. On the other hand, if we look at the last 500 data points, the trends have diverged, and a clustering algorithm will retrieve the true states (for example DBSCAN with appropriate parameters makes no classification error).

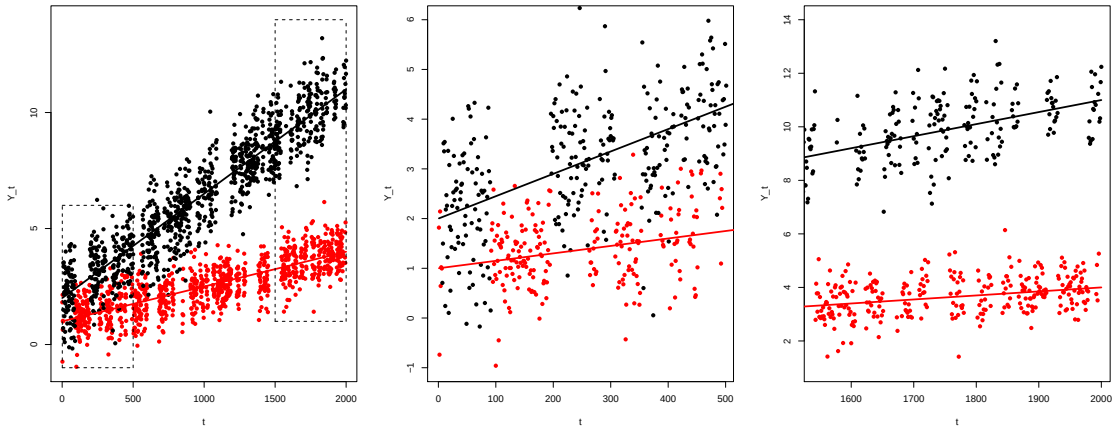


Figure 5.4: Example with linear trends: once the trends have diverged, we can identify them using a clustering algorithm

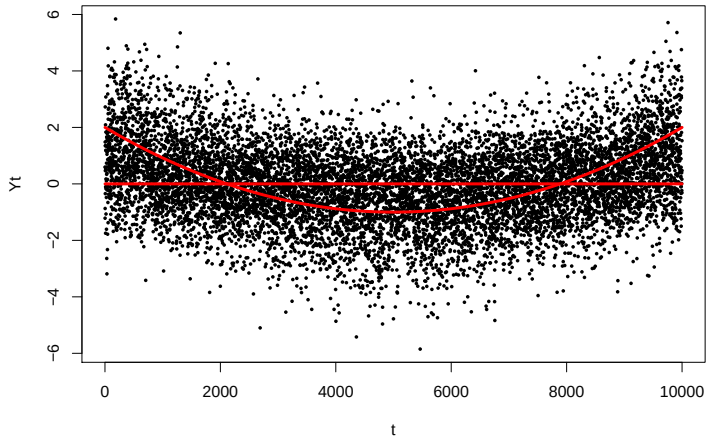


Figure 5.5: Simulated data points

Let us consider another example where the trends have not diverged yet. Let $K = 2$, $T_1(t) = 0$ and $T_2(t) = 3\left(\frac{t-\frac{n}{2}}{\frac{n}{2}}\right)^2 - 1$. The emission distributions of Z_t are normally distributed, with mean 0 and variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$, and $n = 10000$. The transition matrix is given by $Q = \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix}$.

Looking at Figure 5.5, the two states are not separated. They will eventually diverge, but we don't have enough observations to make use of this. Because of this, the naive approach fails. However, using the direct maximum likelihood approach, it is possible to identify the trends even if they have not diverged yet. Here we consider that $K = 2$ is known and that the degrees of the trends are bounded by $d = 4$ (known). The variances are unknown.

Figure 5.6 shows the true and estimated trends, obtained by maximum likelihood inference using the EM algorithm. The estimated transition matrix and variances are

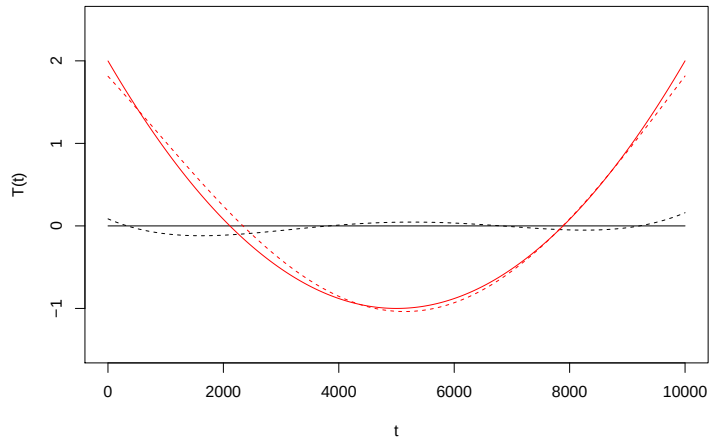


Figure 5.6: True (full lines) and estimated (dashed lines) trends

$$\hat{Q} = \begin{pmatrix} 0.74 & 0.26 \\ 0.22 & 0.78 \end{pmatrix}, \quad (\hat{\sigma}_1^2, \hat{\sigma}_2^2) = (1.13, 2.11).$$

The precision of these estimations can be improved by increasing the number of data points, but the trends and the other parameters are already well estimated. The naive approach could not have achieved this because it does not take full advantage of the temporal structure of the data.

A noteworthy remark at this point is that even if the trends have not diverged–an assumption on which the proofs rely heavily–the MLE is still able to recover the trends and the homogeneous parameters accurately. In the example, an explanation is that the process is slowly varying in the sense of Section 5.6. This is especially relevant for practical applications where one may not have enough time to see the trends diverge.

5.9 Proofs

5.9.1 Proof of the concentration inequalities

Proof of Lemma 5.7

Let us first state a Hoeffding inequality for uniformly ergodic Markov chains, using **(Aerg)** (see e.g. Glynn and Ormoneit (2002)): for all $\epsilon > 0$, $x_1 \in \mathcal{X}^*$ and $n \geqslant \frac{1}{2\epsilon\sigma_-}$,

$$\mathbb{P}\left(\mathbb{P}(X_1 = x^*) - \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{X_t = x^*} \geqslant \epsilon \;\middle|\; X_1 = x_1 \right) \leqslant \exp\left(-\frac{\sigma_-^2 \epsilon^2}{2} n \right).$$

The value of $\mathbb{P}(X_1 = x^*)$ in the inequality is the one corresponding to the stationary distribution, so it is bounded below by σ_- using **(Aerg)**. Thus, for all $\delta > 0$, $\epsilon > 0$, $n \geqslant \frac{1}{\delta\epsilon\sigma_-}$ and $x_1 \in \mathcal{X}^*$,

$$\mathbb{P}\left(\frac{2}{\delta n}\sum_{t=1}^{\delta n/2} \mathbf{1}_{X_t = x^*} \leqslant \sigma_- - \epsilon \;\middle|\; X_1 = x_1 \right) \leqslant \exp\left(-\frac{\sigma_-^2 \epsilon^2 \delta}{4} n \right).$$

Assume $n \geqslant \frac{2}{\delta(\sigma_-)^2}$. Choose $\epsilon = \sigma_-/2$ and let us apply a union bound on a covering $\mathcal{R}$ of $\{1, \dots, n\}$ in at most $2n/\delta$ segments of size $\delta n/2$:

$$\mathbb{P}\left(\inf_{S\in\mathcal{R}} \frac{1}{n}\sum_{t\in S} \mathbf{1}_{X_t=x^*} \leqslant \frac{\delta\sigma_-}{4}\right) \leqslant \frac{2n}{\delta}\exp\left(-\frac{\sigma_-^4\delta}{16}n\right).$$

We conclude using Borel-Cantelli's lemma.

Proof of Lemma 5.8

Without loss of generality, we assume $\delta_n \longrightarrow \delta$ almost surely (this is possible by replacing δ_n by $\delta_n \wedge \delta$ in the first statement and $\delta_n \vee \delta$ in the second).

Let us first show that the two statements are equivalent. Assume the second one to be satisfied. Let $(U_t)_{t\geqslant 1}$ and $(\delta_n)_n$ as in the first statement. Applying the second one to the i.i.d sequence of non-negative integrable random variables $(-U_t)_{t\geqslant 1}$, we get

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ \#S\leqslant\delta_n n}} \frac{1}{n}\sum_{t\in S}(-U_t) &\leqslant \mathbb{E}[(-U_1)\mathbf{1}_{-U_1\geqslant -q_U(\delta)}] \\ &\leqslant -\mathbb{E}[U_1\mathbf{1}_{U_1\leqslant q_U(\delta)}].\end{aligned}$$

Adding $\mathbb{E}[U_1]$ on each side and using the law of large numbers, we obtain

$$\limsup_{n\to\infty} \sup_{\substack{S\subset\{1,\dots,n\}\\ \#S\leqslant\delta_n n}} \frac{1}{n}\sum_{t\notin S} U_t \leqslant \mathbb{E}[U_1\mathbf{1}_{U_1> q_U(\delta)}].$$

Replacing δ_n by $1-\delta_n$, δ by $1-\delta$ and the sets S by their complementary yields the first statement.

Now let us show the second statement (regarding non-negative random variables). For a random vector $(V_1, \dots, V_n)$, we denote by $V_{(1)} \leqslant V_{(2)} \leqslant \dots \leqslant V_{(n)}$ its order statistics. Let $\delta \in (0,1)$ and let $(\delta_n)_n$ be a non-increasing sequence of $[0,1]$-valued random variables whose limit is δ almost surely.

For all $\beta \in (0,1)$, let us denote by $\hat{q}_V(\beta) := V_{(\lfloor \beta n\rfloor)}$ the empirical β-quantile. Then

$$\sup_{\substack{S\subset\{1,\dots,n\}\\ \#S\leqslant\delta_n n}} \frac{1}{n}\sum_{t\in S} V_t = \frac{1}{n}\sum_{t=1}^{n} V_t\mathbf{1}_{V_t\geqslant\hat{q}_V(1-\delta_n)}$$

Let us show that almost surely

$$\left|\frac{1}{n}\sum_{t=1}^{n} V_t\mathbf{1}_{V_t\geqslant\hat{q}_V(1-\delta_n)} - \frac{1}{n}\sum_{t=1}^{n} V_t\mathbf{1}_{V_t\geqslant q_V(1-\delta)}\right| \underset{n\to\infty}{\longrightarrow} 0$$

and the result will follow by the law of large numbers. Thus we have to show that

$$\frac{1}{n}\sum_{t=1}^{n} V_t\left(\mathbf{1}_{q_V(1-\delta)\leqslant V_t\leqslant\hat{q}_V(1-\delta_n)} + \mathbf{1}_{\hat{q}_V(1-\delta_n)\leqslant V_t\leqslant q_V(1-\delta)}\right)$$

goes to 0 almost surely. Using Hoeffding's inequality, we have, for all $\beta \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{\#\{t\in\{1,\dots,n\} \text{ s.t. } V_t\geqslant q_V(\beta)\}}{n} - \mathbb{P}(V_1\geqslant q_V(\beta))\right| \geqslant \sqrt{\frac{\log n}{n}}\right) \leqslant 2n^{-2}.$$

In particular, taking $\beta = 1 - \delta$, Borel-Cantelli's lemma shows that almost surely, for large enough n,

$$\hat{q}_V\left(1-\delta-\sqrt{\frac{\log n}{n}}\right) \leqslant q_V(1-\delta) \leqslant \hat{q}_V\left(1-\delta+\sqrt{\frac{\log n}{n}}\right),$$

so that there are at most $\sqrt{n \log n}$ terms between $q_V(1-\delta)$ and $\hat{q}_V(1-\delta)$. Hence there are at most $\sqrt{n \log n} + (\delta_n - \delta)n$ terms between $q_V(1-\delta)$ and $\hat{q}_V(1-\delta_n)$. As $(\delta_n)_n$ is non-increasing, this yields

$$\hat{q}_V(1-\delta_n) \leqslant \hat{q}_V(1-\delta) \leqslant q_V\left(1-\delta+\sqrt{\frac{\log n}{n}}\right),$$

which ensures that these terms are bounded above by $q_V(1-\delta+\sqrt{\frac{\log n}{n}})$. Thus, almost surely, for n large enough,

$$\frac{1}{n}\sum_{t=1}^{n} V_t\left(\mathbf{1}_{q_V(1-\delta)\leqslant V_t\leqslant \hat{q}_V(1-\delta_n)} + \mathbf{1}_{\hat{q}_V(1-\delta_n)\leqslant V_t\leqslant q_V(1-\delta)}\right) \leqslant \left[\sqrt{\frac{\log n}{n}} + (\delta_n - \delta)\right] q_V(1-\delta/2),$$

which indeed converges to 0.

5.9.2 Proof of the localization of the MLE

Proof of Lemma 5.9

We shall use the inequality

$$|\log C_\mu - \log C_\nu| \leqslant \frac{|C_\mu - C_\nu|}{C_\mu \wedge C_\nu}$$

with $C_\mu = p^\theta(Y_t \mid Y_0^{t-1}, X_t \neq x, X_0 \sim \mu)$ and $C_\nu = p^\theta(Y_t \mid Y_0^{t-1}, X_t \neq x, X_0 \sim \nu)$.

$$\begin{aligned}|C_\mu - C_\nu| &= \left|\frac{p^\theta(Y_t, X_t \neq x \mid Y_0^{t-1}, X_0 \sim \mu)}{p^\theta(X_t \neq x \mid Y_0^{t-1}, X_0 \sim \mu)} - \frac{p^\theta(Y_t, X_t \neq x \mid Y_0^{t-1}, X_0 \sim \nu)}{p^\theta(X_t \neq x \mid Y_0^{t-1}, X_0 \sim \nu)}\right| \\ &=: \left|\frac{B_\mu}{A_\mu} - \frac{B_\nu}{A_\nu}\right|.\end{aligned}$$

One has

$$\begin{aligned}B_\mu &= \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') p^\theta(X_t = x' \mid Y_0^{t-1}, X_0 \sim \mu) \\ &= \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \sum_{x'' \in \mathcal{X}^\theta} Q^\theta_{x''x'} p^\theta(X_{t-1} = x'' \mid Y_0^{t-1}, X_0 \sim \mu),\end{aligned}$$

which yields

$$\sigma_- \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x') \leqslant B_\mu \leqslant (1-\sigma_-) \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')$$

and the same result holds for B_ν. Besides,

$$\begin{aligned} A_\mu &= \sum_{x \neq x'} p^\theta(X_t = x' \mid Y_0^{t-1}, X_0 \sim \mu) \\ &= \sum_{x' \neq x} \sum_{x'' \in \mathcal{X}^\theta} Q^\theta_{x''x'} p^\theta(X_{t-1} = x'' \mid Y_0^{t-1}, X_0 \sim \mu). \end{aligned}$$

Hence,

$$\sigma_- \leqslant A_\mu \leqslant 1 - \sigma_-$$

and the same result holds for A_ν. Then, letting $\phi_\mu(x') = p^\theta(X_{t-1} = x' \mid Y_0^{t-1}, X_0 \sim \mu)$, we get, using the above expressions:

$$\begin{aligned} |A_\mu - A_\nu| &\leqslant (1 - \sigma_-)\|\phi_\mu - \phi_\nu\|_1 \\ |B_\mu - B_\nu| &\leqslant (1 - \sigma_-) \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')\|\phi_\mu - \phi_\nu\|_1 \end{aligned}$$

Thus,

$$\begin{aligned} |C_\mu - C_\nu| &= \left| \frac{B_\mu}{A_\mu} - \frac{B_\nu}{A_\nu} \right| \\ &\leqslant \frac{1}{A_\mu A_\nu} \left(B_\mu |A_\mu - A_\nu| + A_\mu |B_\mu - B_\nu| \right) \\ &\leqslant \frac{2(1 - \sigma_-)^2}{\sigma_-^2} \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')\|\phi_\mu - \phi_\nu\|_1 \end{aligned}$$

Furthermore,

$$\frac{1}{C_\mu \wedge C_\nu} \leqslant \frac{(1 - \sigma_-)}{\sigma_- \sum_{x' \neq x} p^\theta(Y_t \mid X_t = x')}$$

Finally,

$$|\log C_\mu - \log C_\nu| \leqslant \frac{2}{(1 - \rho)^3} \|\phi_\mu - \phi_\nu\|_1$$

It remains to prove that $\|\phi_\mu - \phi_\nu\|_1 \leqslant \rho^{t-1}$, which follows from the geometric ergodicity of the HMM,see for instance Corollary 1 of Douc et al. (2004) or Proposition 2.1 of De Castro et al. (2017).

Proof of Lemma 5.10

Proof of equation (5.7)**.** For all $t \geqslant 1$ and $y_1^t \in \mathbb{R}^t$,

$$\begin{aligned} &p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\ &\quad = p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^t) \times p^\theta(x_U(\theta) \notin X^{t-1}_{(t-a+1)\vee 1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\ &\qquad + p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta), x_U(\theta) \in X^{t-1}_{(t-a+1)\vee 1}) \\ &\qquad\quad \times p^\theta(x_U(\theta) \in X^{t-1}_{(t-a+1)\vee 1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)), \end{aligned}$$

so that

$$\begin{aligned}
&|p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) - p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^t)| \\
&\leqslant p^\theta(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&\quad \times \Big(p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^t) + p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta), x_U(\theta) \in X_{t-a+1}^{t-1}) \Big) \\
&\leqslant 2p^\theta(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \sum_{x \in \mathcal{X}^\theta \setminus \{x_U(\theta)\}} \gamma_x^\theta(y_t - T_x^\theta(t)).
\end{aligned}$$

In addition, under **(Aerg)**,

$$\begin{aligned}
&p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&\qquad = \sum_{x \in \mathcal{X}^\theta \setminus \{x_U(\theta)\}} p^\theta(y_t | X_t = x) p^\theta(X_t = x \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}) \\
&\qquad \geqslant \sigma_- \sum_{x \in \mathcal{X}^\theta \setminus \{x_U(\theta)\}} \gamma_x^\theta(y_t - T_x^\theta(t))
\end{aligned}$$

and the same holds for $p^\theta(y_t \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^t)$, so that using that $|\log x - \log y| \leqslant \frac{|x-y|}{x \wedge y}$ for all $x, y > 0$, we obtain that for all $t \geqslant 1$ and $y_1^t \in \mathbb{R}^t$

$$\begin{aligned}
&|\log p^\theta(Y_t \,|\, X_1^{t-a} \neq x_U(\theta), X_t \neq x_U(\theta), Y_1^{t-1}) - \log p^\theta(Y_t \,|\, X_1^t \neq x_U(\theta), Y_1^{t-1})| \\
&\qquad \leqslant \frac{2}{\sigma_-} p^\theta(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)). \quad (5.28)
\end{aligned}$$

Proof of equation (5.8)**.** By union bound,

$$\begin{aligned}
&p^\theta(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&\leqslant \sum_{i=(t-a+1)\vee 1}^{t-1} p^\theta(X_i = x_U(\theta) \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&= \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} p^\theta(X_i = x_U(\theta) \,|\, y_i, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \\
&\qquad \times p^\theta(X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&\leqslant \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} p^\theta(X_i = x_U(\theta) \,|\, y_i, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \\
&= \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1}, x_{i+1} \in \mathcal{X}^\theta} \frac{p^\theta(X_i = x_U(\theta) \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \gamma_{x_U(\theta)}^\theta(y_i - T_{x_U(\theta)}^\theta(i))}{\sum_{x \in \mathcal{X}^\theta} p^\theta(X_i = x \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \gamma_x^\theta(y_i - T_x^\theta(i))}.
\end{aligned}$$

Using the Markov property and **(Aerg)**, we obtain that for all $x_{i-1}, x_{i+1} \in \mathcal{X}^\theta$,

$$p^\theta(X_i = x \,|\, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}) \in [\sigma_-^2, 1].$$

Hence,

$$\begin{aligned}
p^\theta(x_U(\theta) \in X_{(t-a+1)\vee 1}^{t-1} \,|\, y_1^{t-1}, x_U(\theta) \notin X_1^{t-a}, X_t \neq x_U(\theta)) \\
&\leqslant \sum_{i=(t-a+1)\vee 1}^{t-1} \sum_{x_{i-1},x_{i+1}\in\mathcal{X}^\theta} \frac{\gamma^\theta_{x_U(\theta)}(y_i - T^\theta_{x_U(\theta)}(i))}{\sigma_-^2 \sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(y_i - T^\theta_x(i))} \\
&\leqslant \frac{K^2}{\sigma_-^2} \sum_{i=(t-a+1)\vee 1}^{t-1} \frac{\gamma^\theta_{x_U(\theta)}(y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(y_i - T^\theta_x(i))}
\end{aligned}$$

which concludes the proof.

Proof of equation (5.9)**.** This quantity is always bounded by 1 since all terms are nonnegative. In addition, under Assumptions **(Aenv)** and **(Amin)**, for all $i \in I^U_{n,B}(\theta)$,

$$\begin{aligned}
\frac{\gamma^\theta_{x_U(\theta)}(Y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} &\leqslant \sum_{x\in\mathcal{X}^*} \mathbf{1}_{X_i=x^*} \left(1 \wedge \frac{\gamma^\theta_{x_U(\theta)}(Z_i + T^*_{x^*}(i) - T^\theta_{x_U(\theta)}(i))}{\sup_{x\in\mathcal{X}^\theta} \gamma^\theta_x(Z_i + T^*_{x^*}(i) - T^\theta_x(i))}\right) \\
&\leqslant 1 \wedge \frac{g\left(\left\{\inf_{x^*\in\mathcal{X}^*} |T^*_{x^*}(i) - T^\theta_{x_U(\theta)}(i)| - |Z_i^{\max}|\right\}_+\right)}{m\left(|Z_i^{\max}| + \sup_{x^*\in\mathcal{X}^*} \inf_{x\in\mathcal{X}^\theta} |T^*_{x^*}(i) - T^\theta_x(i)|\right)}.
\end{aligned}$$

Since $\theta \in \mathcal{U}(\beta,n,B) \setminus \mathcal{T}(\alpha,n,D)$, Corollary 5.4 ensures that $\inf_{x\in\mathcal{X}^\theta} |T^*_{x^*}(i) - T^\theta_x(i)| \leqslant M^{\mathcal{T}}(\alpha,D)$ for all $x^* \in \mathcal{X}^*$ and $i \in \{1,\dots,n\}$. Moreover, by definition of $I^U_{n,B}(\theta)$, one has $\inf_{x^*\in\mathcal{X}^*} |T^*_{x^*}(i) - T^\theta_{x_U(\theta)}(i)| \geqslant B$ for all $i \in I^U_{n,B}(\theta)$, so that

$$\frac{\gamma^\theta_{x_U(\theta)}(Y_i - T^\theta_{x_U(\theta)}(i))}{\sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} \leqslant 1 \wedge \frac{g(\{B - |Z_i^{\max}|\}_+)}{m(|Z_i^{\max}| + M^{\mathcal{T}}(\alpha,D))}.$$

for all $i \in I^U_{n,B}(\theta)$, which concludes the proof.

Proof of Lemma 5.11

Under Assumption **(Aenv)**, we have

$$\begin{aligned}
p^\theta(Y_t \,|\, X_t = x_U(\theta)) &= \gamma^\theta_{x_U(\theta)}(Y_t - T^\theta_{x_U(\theta)}(t)) \\
&= \sum_{x^*\in\mathcal{X}^*} \mathbf{1}_{X_t=x^*} \gamma^\theta_{x_U(\theta)} \left(Z_t + T^*_{x^*}(t) - T^\theta_{x_U(\theta)}(t)\right) \\
&\leqslant \sup_{x^*\in\mathcal{X}^*} g\left(\left\{|T^*_{x^*}(t) - T^\theta_{x_U(\theta)}(t)| - |Z_t^{\max}|\right\}_+\right) \\
&\leqslant g\left(\left\{\inf_{x^*\in\mathcal{X}^*} |T^*_{x^*}(t) - T^\theta_{x_U(\theta)}(t)| - |Z_t^{\max}|\right\}_+\right),
\end{aligned}$$

hence, for all θ,

$$p^\theta(Y_t \,|\, X_t = x_U(\theta)) \leqslant \begin{cases} g(\{B - |Z_t^{\max}|\}_+) & \text{if } t \in I_{n,B}^U(\theta), \\ g(0) & \text{otherwise.} \end{cases} \tag{5.29}$$

On the other hand, under Assumptions **(Amin)** and **(Aerg)**,

$$\begin{aligned}
p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1}) &= \frac{p^\theta(Y_t, X_t \neq x_U(\theta) \,|\, Y_1^{t-1})}{p^\theta(X_t \neq x_U(\theta) \,|\, Y_1^{t-1})} \\
&\geqslant \sum_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} p^\theta(Y_t \,|\, X_t = x, Y_1^{t-1}) p^\theta(X_t = x \,|\, Y_1^{t-1}) \\
&\geqslant \sigma_- \sum_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} p^\theta(Y_t \,|\, X_t = x) \\
&= \sigma_- \sum_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} \gamma_x^\theta(Y_t - T_x^\theta(t)) \\
&= \sigma_- \sum_{x^* \in \mathcal{X}^*} \mathbf{1}_{X_t = x^*} \sum_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} \gamma_x^\theta(Z_t + T_{x^*}^*(t) - T_x^\theta(t)) \\
&\geqslant \sigma_- \inf_{x^* \in \mathcal{X}^*} \sup_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} m(|Z_t^{\max}| + |T_{x^*}^*(t) - T_x^\theta(t)|) \\
&\geqslant \sigma_- m\left(|Z_t^{\max}| + \sup_{x^* \in \mathcal{X}^*} \inf_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} |T_{x^*}^*(t) - T_x^\theta(t)| \right).
\end{aligned}$$

Using the fact that $\theta \notin \mathcal{T}(\alpha, n, D)$ and Corollary 5.4, for all $x^* \in \mathcal{X}^*$,

$$\inf_{x \in \mathcal{X}^\theta, x \neq x_U(\theta)} |T_{x^*}^*(t) - T_x^\theta(t)| \leqslant M^{\mathcal{T}}(\alpha, D).$$

For example, we can choose $x = x_S(x^*, \theta, n)$ (see Equation 5.1). We know that $x_U(\theta) \neq x_S(x^*, \theta, n)$ because we chose $B > M^{\mathcal{T}}(\alpha, D)$, so that $|T_{x^*}^*(i) - T_{x_U(\theta)}^\theta(i)| > M^{\mathcal{T}}(\alpha, D)$ for at least one $i \in \{1, \dots, n\}$, and because $|T_{x^*}^*(i) - T_{x_S(x^*,\theta,n)}^\theta(i)| \leqslant M^{\mathcal{T}}(\alpha, D)$ for all $i \in \{1, \dots, n\}$. Therefore, for all $\theta \in \mathcal{U}(\beta, n, B) \setminus \mathcal{T}(\alpha, n, D)$,

$$p^\theta(Y_t \,|\, X_t \neq x_U(\theta), Y_1^{t-1}) \geqslant \sigma_- m(|Z_t^{\max}| + M^{\mathcal{T}}(\alpha, D)),$$

which concludes the proof together with equation (5.29).

5.9.3 Proof of the block approximation lemmas

Proof of Lemma 5.12 (current block)

First note that this quantity is non-negative: the denominator contains less terms, and all of them are non-negative. Hence it is enough to find an upper bound. To this aim we will use Assumptions **(Aenv)**, **(Amin)** and **(Aerg)**:

$$\begin{aligned}
&\left| \log p^\theta_{Y_t | Y_1^{t-1}}(Y_t | Y_1^{t-1}) - \log p^\theta_{Y_t, B_t | Y_1^{t-1}}(Y_t, B_t | Y_1^{t-1}) \right| \\
&\leqslant \log \left(\frac{g(0)}{\sigma_- \sup_{x_t \in \mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t) = B_t} m(|Y_t - T_{x_t}^\theta(t)|)} \right) \\
&\leqslant \log \frac{g(0)}{\sigma_-} + \sum_{x_t^* \in \mathcal{X}^*} \mathbf{1}_{X_t = x_t^*} \left(- \log m \left(\inf_{x_t \in \mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t) = \mathbf{b}^*(x_t^*)} |Z_t + T_{x_t^*}^*(t) - T_{x_t}^\theta(t)| \right) \right).
\end{aligned}$$

When $X_t = x_t^*$,

$$\begin{aligned}\inf_{x_t\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| \\ &\leqslant |Z_t| + \inf_{x_t\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |T^*_{\mathbf{b}^*(x_t^*)}(t) - T^\theta_{x_t}(t)| + \Delta(x_t^*) \\ &\leqslant |\tilde{Z}_t| + M + \|\Delta\|_\infty\end{aligned}$$

using equation (5.14), hence

$$-\log m\left(\inf_{x_t\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t)=\mathbf{b}^*(x_t^*)} |Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|\right) \leqslant -\log m(M + |Z_t^{\max}| + \|\Delta\|_\infty).$$

This yields

$$\left|\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1})\right| \leqslant \log\frac{g(0)}{\sigma_- m(M + |Z_t^{\max}| + \|\Delta\|_\infty)}.$$

Let us show the second bound. We can rewrite it as

$$\begin{aligned}&\log p^\theta_{Y_t|Y_1^{t-1}}(Y_t|Y_1^{t-1}) - \log p^\theta_{Y_t,B_t|Y_1^{t-1}}(Y_t,B_t|Y_1^{t-1}) \\ &= -\log\left(1 - \frac{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)\neq B_t}}{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Y_t - T^\theta_{x_t}(t))}\right) \\ &= -\log\left(1 - \sum_{x_t^*\in\mathcal{X}^*}\mathbf{1}_{X_t=x_t^*}\underbrace{\frac{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t))\mathbf{1}_{\mathbf{b}^\theta(x_t)\neq B_t}}{\displaystyle\sum_{x_t\in\mathcal{X}^\theta} p^\theta(X_t = x_t|Y_1^{t-1})\gamma_{x_t}(Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t))}}_{(*)}\right).\end{aligned}$$

Using **(Aenv)**, **(Amin)** and **(Aerg)**, we get

$$\begin{aligned}(*) &\leqslant \frac{\displaystyle\sup_{x_t\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t)\neq B_t} g(|Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|)}{\sigma_- \displaystyle\sum_{x_t\in\mathcal{X}^\theta} m(|Z_t + T^*_{x_t^*}(t) - T^\theta_{x_t}(t)|)} \\ &\leqslant \frac{\displaystyle\sup_{x_t\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_t)\neq B_t} g(\{|T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| - |Z_t^{\max}|\}_+)}{\sigma_- \displaystyle\sup_{x_t\in\mathcal{X}^\theta} m(|T^*_{x_t^*}(t) - T^\theta_{x_t}(t)| + |Z_t^{\max}|)}.\end{aligned}$$

Let $x \in \mathcal{X}^\theta$ such that $\mathbf{b}^\theta(x) \neq B_t$. Then, when $X_t = x_t^*$,

$$\begin{aligned}|Y_t - T^\theta_x(t)| &= |Z_t + T^*_{x_t^*}(t) - T^\theta_x(t)| \\ &\geqslant |T^*_{B_t}(t) - T^\theta_x(t)| - |Z_t^{\max}| - \Delta(x_t^*) \\ &\geqslant |T^*_{B_t}(t) - T^*_{\mathbf{b}^\theta(x)}(t)| - M - |Z_t^{\max}| - \|\Delta\|_\infty \\ &\geqslant E(t) - M - |Z_t^{\max}| - \|\Delta\|_\infty\end{aligned}$$

and

$$\sup_{x_t\in\mathcal{X}^\theta} m(|T^*_{x_t^*}(t)-T^\theta_{x_t}(t)|+|Z_t^{\max}|) \leqslant m(\inf_{x_t\in\mathcal{X}^\theta}|T^*_{x_t^*}(t)-T^\theta_{x_t}(t)|+|Z_t^{\max}|)$$
$$\leqslant m(M+|Z_t^{\max}|+\|\Delta\|_\infty)$$

using equation (5.14). We finally obtain

$$\textbf{(*)} \leqslant \frac{g(\{E(t)-M-|Z_t^{\max}|-\|\Delta\|_\infty\}_+)}{\sigma_- m(M+|Z_t^{\max}|+\|\Delta\|_\infty)}. \tag{5.30}$$

Proof of Lemma 5.14 (recent blocks)

Proof of equation (5.18)**.** Without loss of generality, one may assume $a \leqslant t$ (otherwise, the proof holds by replacing a by $a\wedge t$).

$$\begin{aligned}p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a}) &= p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1})\\ &\quad\times p^\theta_{B^{t-1}_{t-a+1}|Y_1^{t-1},B_1^{t-a}}(b^{t-1}_{t-a+1}|y_1^{t-1},b_1^{t-a})\\ &\quad+ p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a},(\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\})\\ &\quad\times p^\theta_{B^{t-1}_{t-a+1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\}|y_1^{t-1},b_1^{t-a}),\end{aligned}$$

hence

$$\begin{aligned}&|p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a}) - p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1})|\\ &\quad\leqslant p^\theta_{B^{t-1}_{t-a+1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\}|y_1^{t-1},b_1^{t-a})\Big[p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1})\\ &\qquad\qquad+ p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-a},(\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\})\Big]\\ &\quad\leqslant 2p^\theta_{B^{t-1}_{t-a+1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\}|y_1^{t-1},b_1^{t-a})\sum_{x\in\mathcal{X}^\theta} p^\theta_{Y_t,B_t|X_t}(y_t,b_t|x).\end{aligned}$$

Finally, since under **(Aerg)**

$$\begin{aligned}p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-a}}(y_t,b_t|y_1^{t-1},b_1^{t-a}) &= \sum_{x\in\mathcal{X}} p^\theta_{Y_t,B_t|X_t}(y_t,b_t|x)p^\theta_{X_t|Y_1^{t-1},B_1^{t-a}}(x|y_1^{t-1},b_1^{t-a})\\ &\geqslant \sigma_-\sum_{x\in\mathcal{X}} p^\theta_{Y_t,B_t|X_t}(y_t,b_t|x)\end{aligned}$$

and the same inequality holds for $p^\theta_{Y_t,B_t|Y_1^{t-1},B_1^{t-1}}(y_t,b_t|y_1^{t-1},b_1^{t-1})$, we obtain equation (5.18) using $|\log x-\log y|\leqslant\frac{|x-y|}{x\wedge y}$ for all $x,y>0$.

Proof of equation (5.19)**.** Since

$$(\mathcal{B}^*)^{a-1}\setminus\{b^{t-1}_{t-a+1}\} = \bigcup_{i=t-a+1}^{t-1}(\mathcal{B}^*)^{i-(t-a+1)}\times(\mathcal{B}^*\setminus\{b_i\})\times(\mathcal{B}^*)^{t-1-i},$$

we get using a union bound that

$$\begin{aligned}
&p^{\theta}_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b_{t-a+1}^{t-1}\}|y_1^{t-1},b_1^{t-a})\\
&\leqslant \sum_{i=t-a+1}^{t-1} p^{\theta}_{B_i|Y_1^{t-1},B_1^{t-a}}(\mathcal{B}^*\setminus\{b_i\}|y_1^{t-1},b_1^{t-a})\\
&= \sum_{i=t-a+1}^{t-1}\sum_{x_i\in\mathcal{X}^\theta} p^{\theta}_{B_i|X_i}(\mathcal{B}^*\setminus\{b_i\}|x_i)p^{\theta}_{X_i|Y_1^{t-1},B_1^{t-a}}(x_i|y_1^{t-1},b_1^{t-a})\\
&= \sum_{i=t-a+1}^{t-1}\sum_{x_i\in\mathcal{X}^\theta} \mathbf{1}_{\mathbf{b}^\theta(x_i)\neq b_i}\sum_{x_{i-1},x_{i+1}\in\mathcal{X}^\theta} p^{\theta}_{X_i|Y_i,X_{i-1},X_{i+1}}(x_i|y_i,x_{i-1},x_{i+1})\\
&\qquad\qquad\times p^{\theta}_{X_{i-1},X_{i+1}|Y_1^{t-1},B_1^{t-a}}(x_{i-1},x_{i+1}|y_1^{t-1},b_1^{t-a}).
\end{aligned}$$

Then we use the fact that for all $x_{i-1},x_{i+1}\in\mathcal{X}^\theta$,

$$p^{\theta}_{X_{i-1},X_{i+1}|Y_1^{t-1},B_1^{t-a}}(x_{i-1},x_{i+1}|y_1^{t-1},b_1^{t-a})\leqslant 1.$$

Moreover, using the Markov property and **(Aerg)**, we see that for all $x_{i-1},x,x_{i+1}\in\mathcal{X}^\theta$

$$\begin{aligned}
p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x|x_{i-1},x_{i+1}) &= \frac{p^{\theta}_{X_{i+1}|X_i}(x_{i+1}|x)p^{\theta}_{X_i|X_{i-1}}(x|x_{i-1})}{p^{\theta}_{X_{i+1}|X_{i-1}}(x_{i+1}|x_{i-1})}\\
&\geqslant Q^\theta(x,x_{i+1})Q^\theta(x_{i-1},x)\\
&\geqslant \sigma_-^2,
\end{aligned}$$

hence,

$$\begin{aligned}
&p^{\theta}_{X_i|Y_i,X_{i-1},X_{i+1}}(x_i|y_i,x_{i-1},x_{i+1})\\
&\qquad= \frac{p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta} p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x|x_{i-1},x_{i+1})\gamma^\theta_{x}(y_i-T^\theta_{x}(i))}\\
&\qquad\leqslant \frac{p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sigma_-^2\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_{x}(y_i-T^\theta_{x}(i))}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&p^{\theta}_{B_{t-a+1}^{t-1}|Y_1^{t-1},B_1^{t-a}}((\mathcal{B}^*)^{a-1}\setminus\{b_{t-a+1}^{t-1}\}|y_1^{t-1},b_1^{t-a})\\
&\leqslant \sum_{i=t-a+1}^{t-1}\sum_{x_{i-1},x_{i+1}\in\mathcal{X}^\theta}\frac{\sum_{x_i\in\mathcal{X}^\theta}\mathbf{1}_{\mathbf{b}^\theta(x_i)\neq b_i}p^{\theta}_{X_i|X_{i-1},X_{i+1}}(x_i|x_{i-1},x_{i+1})\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sigma_-^2\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_{x}(y_i-T^\theta_{x}(i))}\\
&\leqslant \frac{K^2}{\sigma_-^2}\sum_{i=t-a+1}^{t-1}\frac{\sup_{x_i\in\mathcal{X}^\theta\text{ t.q. }\mathbf{b}^\theta(x_i)\neq b_i}\gamma^\theta_{x_i}(y_i-T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta}\gamma^\theta_{x}(y_i-T^\theta_{x}(i))}.
\end{aligned}$$

Proof of equation (5.20)**.** Using **(Aenv)** and **(Amin)**,

$$\frac{\sup_{x_i\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_i)\neq B_i} \gamma^\theta_{x_i}(Y_i - T^\theta_{x_i}(i))}{\sum_{x\in\mathcal{X}^\theta} \gamma^\theta_x(Y_i - T^\theta_x(i))} \leqslant \left(1 \wedge \frac{\sup_{x_i\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_i)\neq B_i} g(|Y_i - T^\theta_{x_i}(i)|)}{\sup_{x\in\mathcal{X}^\theta} m(|Y_i - T^\theta_x(i)|)}\right)$$

$$\leqslant \sum_{x_i^*\in\mathcal{X}^*} \mathbf{1}_{X_t=x_i^*} \left(1 \wedge \frac{\sup_{x_i\in\mathcal{X}^\theta \text{ t.q. } \mathbf{b}^\theta(x_i)\neq \mathbf{b}^*(x_i^*)} g(|Z_i + T^*_{x_i^*}(i) - T^\theta_{x_i}(i)|)}{\sup_{x\in\mathcal{X}^\theta} m(|Z_i + T^*_{x_i^*}(i) - T^\theta_x(i)|)}\right)$$

$$\leqslant 1 \wedge \frac{g(\{E(t) - M - |Z_i^{\max}| - \|\Delta\|_\infty\}_+)}{m(M + |Z_i^{\max}| + \|\Delta\|_\infty)}$$

by the same arguments as in the control of **(*)** in equation (5.30).

5.9.4 Uniform convergence of the log-likelihood

The following theorem is a reformulation of Proposition 2 of Douc et al. (2004) by noticing that their proof also works when the space of parameters is not parametric.

Theorem 5.19. *Let $\mathcal{V}$ be a Polish space and write $\mathcal{D}(\mathcal{V})$ the set of nonnegative functions of $\mathcal{V}$. Let $K \in \mathbb{N}^*$. Let $\mathcal{Q}_K$ be the set of transition matrices of size K and Δ_K the set of probability measures on $[K]$. Let $(V_t)_{t\geqslant 1}$ be an ergodic and stationary process taking values in $\mathcal{V}$ with distribution $\mathbb{P}^*$.*

Consider a compact metric space Ω and mappings $\omega \longmapsto Q^\omega \in \mathcal{Q}_K$ and $\omega \longmapsto \gamma^\omega \in \mathcal{D}(\mathcal{V})^K$. Assume that $\omega \longmapsto Q^\omega$ is continuous and for all $v \in \mathcal{V}$, the mapping $\omega \longmapsto \gamma^\omega(v) \in \mathbb{R}_+^K$ is continuous. Finally, assume that there exists a constant $\sigma_- > 0$ such that

$$\inf_{\omega\in\Omega} \inf_{x,x'\in[K']} Q^\omega(x,x') \geqslant \sigma_-,$$

$$\sup_{\omega\in\Omega} \sup_{x\in[K']} \sup_{v\in\mathcal{V}} \gamma^\omega_x(v) < \infty,$$

$$\mathbb{E}^*\left[\sup_{\omega\in\Omega} \Big(\log \sum_{x\in[K']} \gamma^\omega_x(V_1)\Big)_-\right] < \infty.$$

For all $\pi \in \Delta_K$, $\omega \in \Omega$ and $v_1^n \in \mathcal{V}^n$, let

$$\frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{v_1^n\} := \frac{1}{n} \log \sum_{x_1^n\in[K']^n} \pi(x_1) Q^\omega(x_1,x_2)\dots Q^\omega(x_{n-1},x_n) \prod_{t=1}^n \gamma^\omega_{x_t}(v_t)$$

be the log-likelihood corresponding to the HMM with parameters $(\pi, Q^\omega, \gamma^\omega)$ and to the observations v_1^n.

Then for all $\pi \in \Delta_K$ and $\omega \in \Omega$, there exists a finite $l(Q^\omega, \gamma^\omega)$ such that almost surely,

$$\frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_1^n\} \underset{n\to\infty}{\longrightarrow} l(Q^\omega, \gamma^\omega).$$

In addition, for all $N \in \mathbb{N}^*$, the mapping $\omega \longmapsto l(Q^\omega, \gamma^\omega)$ is continuous and

$$\sup_{\omega\in\Omega} \sup_{\pi\in\Delta_{K'}} \left|\frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_1^n\} - l(Q^\omega, \gamma^\omega)\right| \underset{n\to\infty}{\longrightarrow} 0$$

almost surely.

Let us check these assumptions. First, let $\mathcal{V} = \mathbb{R} \times \mathcal{B}^*$, $V_t = (Z'_t, B_t)$, $K = K'$ and

$$\Omega = \left\{ \omega = (\xi, \mathbf{D}, u, \mathbf{b}) \in \Xi_{K'} \times \mathrm{Cl}(\mathcal{D})^{K'} \times [0,1] \times (\mathcal{B}^*)^{K'} \right\}.$$

By assumption **(Aparam)** and by Theorem 5.5, Ω is compact. It is also metrizable. By the uniform continuity of $\mathrm{Cl}(\mathcal{D})$, the applications

$$\begin{cases} \omega \longmapsto Q^\omega := Q_\xi \\ \omega \longmapsto \gamma^\omega(z', b) := (\gamma_{\xi,x}(z' - D_x(u))\mathbf{1}_{\mathbf{b}(x)}(b))_{x \in [K']} \end{cases}$$

are continuous for all $(z', b) \in \mathbb{R} \times \mathcal{B}^*$. The minoration of the transition matrices is ensured by **(Aerg)** and the majoration of the densities by **(Aenv)**. Finally, the integrability condition follows from the fact that for all $\omega \in \Omega$,

$$\sum_{x \in [K']} \gamma^\omega(Z'_1, B_1) \geqslant \inf_{x \in [K']} \gamma_\xi(Z'_1 - \mathbf{D}(u)) \geqslant m(M + |Z_1^{\max}|)$$

by **(Amin)**, and $\mathbb{E}^*[-\log m(M + |Z_1^{\max}|)] < \infty$ by **(Aint)**.

Thus, the previous theorem holds, which shows that the application

$$\omega \longmapsto l(Q^\omega, \gamma^\omega) =: l^{\mathrm{stat}}(Q_\xi, \tau(\gamma_\xi, \mathbf{D}(u)), \mathbf{b})$$

is continuous on Ω. For the uniform convergence, let π_U be the uniform distribution on $[K']$ and let

$$S_{s,n}(\omega) = \frac{1}{n} l_n(\pi_U, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\}$$

for all $s, n \in \mathbb{N}^*$ and $\omega \in \Omega$.

The theorem implies that almost surely,

$$\lim_{n \to \infty} \sup_{\omega \in \Omega} \left| \frac{1}{n} S_{0,n}(\omega) - l(Q^\omega, \gamma^\omega) \right| = 0.$$

Hence, for all $\epsilon > 0$, there exists a (random) $n(\epsilon)$ such that,

$$\forall n \geqslant n(\epsilon), \quad \sup_{\omega \in \Omega} \left| \frac{1}{n} S_{0,n}(\omega) - l(Q^\omega, \gamma^\omega) \right| \leqslant \epsilon. \tag{5.31}$$

The following Lemma is a reformulation of Lemma 2 of Douc et al. (2004) for compact nonparametric parameter spaces.

Lemma 5.20. *Under the same assumptions as the previous theorem, one has for all $v_1^n \in \mathcal{V}^n$*

$$\sup_{\omega \in \Omega} \sup_{\pi \in \Delta_K} |l_n(\pi, Q^\omega, \gamma^\omega)\{v_1^n\} - l_n(\pi_U, Q^\omega, \gamma^\omega)\{v_1^n\}| \leqslant \frac{1}{\sigma_-^2}.$$

Therefore,

$$|n S_{s,n}(\omega) - l_n(\pi_{X_{s+1}|V_1^s, X_1 \sim \pi_U}, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\}| \leqslant \frac{1}{\sigma_-^2}.$$

Note that

$$\begin{aligned} l_n(\pi_{X_{s+1}|V_1^s, X_1 \sim \pi_U}, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\} &= l_{s+n}(\pi_U, Q^\omega, \gamma^\omega)\{V_1^{s+n}\} - l_s(\pi_U, Q^\omega, \gamma^\omega)\{V_1^s\} \\ &= (s+n) S_{0,s+n}(\omega) - s S_{0,s}(\omega), \end{aligned}$$

so that

$$|nS_{s,n}(\omega) - (s+n)S_{0,s+n}(\omega) - sS_{0,s}(\omega)| \leqslant \frac{1}{\sigma_-^2}.$$

Thus, equation (5.31) entails that for all $s \geqslant 1$, $n \geqslant n(\epsilon)$ and $\omega \in \Omega$,

$$|nS_{s,n}(\omega) - nl(Q^\omega, \gamma^\omega)| \leqslant (2s+n)\epsilon + \frac{1}{\sigma_-^2}.$$

Therefore, by Lemma 5.20, one has for all $n \geqslant n(\epsilon)$ and $s \in \{0, \ldots, (N-1)n\}$:

$$\sup_{\omega\in\Omega} \sup_{\pi\in\Delta_{K'}} \left| \frac{1}{n} l_n(\pi, Q^\omega, \gamma^\omega)\{V_{s+1}^{s+n}\} - l(Q^\omega, \gamma^\omega) \right| \leqslant (2N-1)\epsilon + \frac{2}{n\sigma_-^2},$$

which concludes the proof.

BIBLIOGRAPHY

Pierre Ailliot and Françoise Pene. Consistency of the maximum likelihood estimate for non-homogeneous markov–switching models. *ESAIM: Probability and Statistics*, 19:268–292, 2015.

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

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

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.

Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

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.

Sophie Bertrand, Rocío Joo, and Ronan Fablet. Generalized Pareto for pattern-oriented random walk modelling of organisms' movements. *PloS one*, 10(7):e0132231, 2015.

Peter J Bickel, Ya'acov Ritov, Tobias Ryden, et al. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635, 1998.

Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

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, 2016a.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Estimating multivariate latent-structure models. *The Annals of Statistics*, 44(2):540–563, 2016b.

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.

Olivier Cappé, Éric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

Gilles Celeux and Jean-Baptiste Durand. Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564, 2008.

Antoine Chambaz, Aurelien Garivier, and Elisabeth Gassiat. A minimum description length approach to hidden Markov models with Poisson and Gaussian emissions. application to order identification. *Journal of Statistical Planning and Inference*, 139(3):962–977, 2009.

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.

Jörn Dannemann and Hajo Holzmann. Testing for two states in a hidden Markov model. *Canadian Journal of Statistics*, 36(4):505–520, 2008.

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

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

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

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

Manuel Diehn, Axel Munk, and Daniel Rudolf. Maximum likelihood estimation in hidden markov models with inhomogeneous noise. *arXiv preprint arXiv:1804.04034*, 2018.

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

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

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

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

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

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

Elisabeth Gassiat. Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'IHP Probabilités et statistiques*, volume 38, pages 897–906, 2002.

Elisabeth Gassiat and Stéphane Boucheron. Optimal error exponents in hidden Markov models order estimation. *Information Theory, IEEE Transactions on*, 49(4):964–980, 2003.

Elisabeth Gassiat and Christine Keribin. The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM: Probability and Statistics*, 4:25–52, 2000.

Elisabeth Gassiat and Judith Rousseau. About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 2014.

Elisabeth Gassiat, Alice Cleynen, and Stéphane Robin. Finite state space non parametric hidden Markov models are in general identifiable. *Stat. Comp.*, pages 1–11, 2015.

Elisabeth Gassiat, Judith Rousseau, et al. Nonparametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016.

Elisabeth Gassiat, Judith Rousseau, and Elodie Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electronic Journal of Statistics*, 12(1): 703–740, 2018.

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.

Peter W Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics & probability letters*, 56(2):143–146, 2002.

Alexander Goldenshluger and Oleg Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, pages 1608–1632, 2011.

Alexander Goldenshluger and Oleg Lepski. General selection rule from a family of linear estimators. *Theory of Probability & Its Applications*, 57(2):209–226, 2013.

Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

Nikolaus Hansen. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer, 2006.

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.

Frank Kleibergen and Richard Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1):97–126, 2006.

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.

Claire Lacour, Pascal Massart, and Vincent Rivoirard. Estimator selection: a new method with applications to kernel density estimation. *Sankhya A*, 79(2):298–335, 2017.

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.

Roland Langrock, Thomas Kneib, Alexander Sohn, and Stacy L DeRuiter. Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, 71(2):520–528, 2015.

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. *Journal of Machine Learning Research*, 19(39):1–46, 2018a.

Luc Lehéricy. Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. *arXiv preprint arXiv:1807.03997*, 2018b.

Luc Lehéricy. Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, 25(1):464–498, 2019.

Luc Lehéricy. Estimation adaptative non paramétrique pour les modèles à chaîne de Markov cachée. Mémoire de M2, Orsay, 2015.

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

Rachel J MacKAY. Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30(4):573–589, 2002.

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.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings. *arXiv preprint arxiv:1212.2015v2*, 2013.

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

Jean-Marc Robin and Richard J Smith. Tests of rank. *Econometric Theory*, 16(02):151–175, 2000.

Tobias Rydén. Estimating the order of hidden Markov models. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):345–354, 1995.

Lifeng Shang and Kwok-Ping Chan. Nonparametric discriminant HMM and application to facial expression recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2090–2096. IEEE, 2009.

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

GW Stewart and Ji-Guang Sun. Matrix perturbation theory (computer science and scientific computing), 1990.

Augustin Touron. Consistency of the maximum likelihood estimator in seasonal hidden markov models. *arXiv preprint arXiv:1802.08161*, 2018.

Zoé van Havre, Judith Rousseau, Nicole White, and Kerrie Mengersen. Overfitting hidden Markov models with an unknown number of states. *arXiv preprint arXiv:1602.02466*, 2016.

Youen Vermard, Etienne Rivot, Stéphanie Mahévas, Paul Marchal, and Didier Gascuel. Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian hidden Markov models. *Ecological Modelling*, 221(15):1757–1769, 2010.

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

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

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

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

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Estimation adaptative pour les modèles de Markov cachés non paramétriques

Mots Clefs : Statistiques non paramétriques, modèles de Markov cachés, adaptativité minimax, sélection de modèle

Résumé : Dans cette thèse, j'étudie les propriétés théoriques des modèles de Markov cachés non paramétriques. Le choix de modèles non paramétriques permet d'éviter les pertes de performance liées à un mauvais choix de paramétrisation, d'où un récent intérêt dans les applications.
Dans une première partie, je m'intéresse à l'estimation du nombre d'états cachés. J'y introduis deux estimateurs consistants : le premier fondé sur un critère des moindres carrés pénalisés, le second sur une méthode spectrale.
Une fois l'ordre connu, il est possible d'estimer les autres paramètres.
Dans une deuxième partie, je considère deux estimateurs adaptatifs des lois d'émission, c'est-à-dire capables de s'adapter à leur régularité. Contrairement aux méthodes existantes, ces estimateurs s'adaptent à la régularité de chaque loi au lieu de s'adapter seulement à la pire régularité.
Dans une troisième partie, je me place dans le cadre mal spécifié, c'est-à-dire lorsque les observations sont générées par une loi qui peut ne pas être un modèle de Markov caché. J'établis un contrôle de l'erreur de prédiction de l'estimateur du maximum de vraisemblance sous des conditions générales d'oubli et de mélange de la vraie loi.
Enfin, j'introduis une variante non homogène des modèles de Markov cachés : les modèles de Markov cachés avec tendances, et montre la consistance de l'estimateur du maximum de vraisemblance.

Title : Adaptive estimation for nonparametric hidden Markov models

Keys words : Nonparametric statistics, hidden Markov models, minimax adaptive estimation, model selection

Abstract : During my PhD, I have been interested in theoretical properties of nonparametric hidden Markov models. Nonparametric models avoid the loss of performance coming from an inappropriate choice of parametrization, hence a recent interest in applications.
In a first part, I have been interested in estimating the number of hidden states. I introduce two consistent estimators: the first one is based on a penalized least squares criterion, and the second one on a spectral method.
Once the order is known, it is possible to estimate the other parameters.
In a second part, I consider two adaptive estimators of the emission distributions. Adaptivity means that their rate of convergence adapts to the regularity of the target distribution. Contrary to existing methods, these estimators adapt to the regularity of each distribution instead of only the worst regularity.
The third part is focussed on the misspecified setting, that is when the observations may not come from a hidden Markov model. I control the prediction error of the maximum likelihood estimator when the true distribution satisfies general forgetting and mixing assumptions.
Finally, I introduce a nonhomogeneous variant of hidden Markov models : hidden Markov models with trends, and show that the maximum likelihood estimator of such models is consistent.