Estimations pour les modèles de Markov cachés et approximations particulaires : Application à la cartographie et à la localisation simultanées.

Sylvain Le Corff

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2012-ENST-052

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et image »

présentée et soutenue publiquement par

Sylvain LE CORFF

le 28 Septembre 2012

Estimations pour les modèles de Markov cachés et approximations particulaires

Application à la cartographie et à la localisation simultanées

Directeur de thèse : **Eric MOULINES**
Co-encadrement de la thèse : **Gersende FORT**

Jury

M. Gilles PAGÈS, Professeur, LPMA, Université Pierre et Marie Curie	Rapporteur
M. Arnaud DOUCET, Professeur, Department of Statistics, Oxford University	Rapporteur
Mme Elisabeth GASSIAT, Professeur, Laboratoire de Mathématiques, Université Paris-Sud	Examinateur
M. Jean-Michel MARIN, Professeur, Institut de Mathématiques, Université Montpellier II	Examinateur
Mlle Gersende FORT, Directrice de recherche, LTCI, CNRS - TELECOM ParisTech	Examinateur
M. Eric MOULINES, Professeur, LTCI, CNRS - TELECOM ParisTech	Examinateur

T H È S E

TELECOM ParisTech
école de l'Institut Télécom - membre de ParisTech

Table des matières

Chapitre 1

Introduction

1.1 Présentation générale

Un modèle de Markov caché (HMM pour *Hidden Markov Models*) est un processus stochastique à temps discret $\{(X_k, Y_k)\}_{k\geq 0}$ tel que $\{X_k\}_{k\geq 0}$ soit une chaîne de Markov à valeurs dans un espace d'état $\mathbb{X}$ et tel que les variables aléatoires $\{Y_k\}_{k\geq 0}$ soient indépendantes conditionnellement à $\{X_k\}_{k\geq 0}$ et à valeurs dans un espace d'observations $\mathbb{Y}$. De plus, pour tout $m \geq 0$, la loi de Y_m conditionnellement à la suite $\{X_k\}_{k\geq 0}$ ne dépend que de l'état courant X_m. La suite $\{X_k\}_{k\geq 0}$ constitue la *chaîne cachée* (les X_k sont aussi appelées *données manquantes*) et les $\{Y_k\}_{k\geq 0}$ sont les *observations*. Les HMM sont utilisés dans une grande classe de problèmes en finance, en biologie ou en traitement du signal, voir [Cappé *et al.*, 2005] et [Mac Donald et Zucchini, 2009] et les références incluses dans ces livres.

Dans ce document, nous nous intéressons à l'estimation de paramètres dans les chaînes de Markov cachées au sens du maximum de vraisemblance. La consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance ont été prouvées par [Baum et Petrie, 1966] dans le cas où $\mathbb{X}$ et $\mathbb{Y}$ sont finis. La consistance de cet estimateur a ensuite été étendue par [Leroux, 1992] pour des chaînes à espace d'état $\mathbb{X}$ fini et à espace d'observations $\mathbb{Y}$ général. La normalité asymptotique a quant à elle été établie par [Bickel *et al.*, 1998] sous les mêmes hypothèses. Le résultat fourni par [Bickel *et al.*, 1998] a été étendu par [Jensen et Petersen, 1999] pour un espace d'état général, sous des hypothèses d'ergodicité géométrique uniforme. Notons que ces auteurs supposent la consistance de l'estimateur qui a ensuite été établie sous des hypothèses similaires par [Douc *et al.*, 2004b] (des modèles plus généraux que les HMM, à savoir les modèles autorégressifs fonctionnels à régime markovien, sont considérés dans cet article). Plus récemment, la consistance de l'estimateur du maximum de vraisemblance pour des modèles HMM généraux a été établie par [Douc *et al.*, 2011b] en supposant que la chaîne sous-jacente est géométriquement ergodique.

Tous ces résultats sont établis dans le cas de modèles bien spécifiés, c'est-à-dire lorsque les observations sont issues d'un modèle dépendant d'un paramètre $\theta_\star$ inconnu. Une étude du comportement asymptotique de l'estimateur du maximum de vraisemblance pour des modèles mal spécifiés a été proposée par [Douc et Moulines, 2012].

Les résultats présentés précédemment mettent en avant les propriétés asymptotiques de l'estimateur du maximum de vraisemblance. Dans cette thèse, nous nous intéressons à la fois à l'estimation de paramètres en dimension finie et en dimension infinie. Dans le cas de la dimension finie, nous imposons des contraintes sur le calcul de l'estimateur proposé : un premier volet de cette thèse est l'estimation *en ligne* d'un paramètre au sens du maximum de vraisemblance. Le fait d'estimer en ligne signifie que les estimations doivent être produites sans avoir besoin de conserver toutes les observations et sans devoir toutes les utiliser pour produire une seule estimation. Ce genre de procédure est nécessaire dans le cas de grandes bases de données (pour lesquelles un parcours de toutes les observations serait trop coûteux) ou dans des situations de type *temps réel* (pour lesquelles une mise à jour fréquente du paramètre ne permettant pas un accès à toutes les observations est indispensable). Une approche naturelle pour l'estimation en ligne est d'utiliser une méthode de gradient stochastique de type Robbins-Monro, voir [Robbins et Monro, 1951]. De tels algorithmes sont bien connus pour des observations indépendantes, voir par exemple [Benaim, 1999] (voir également [Laruelle, 2011] et [Laruelle et Pagès, 2012] pour l'étude de la convergence d'algorithmes d'approximation stochastique dans un cadre plus général). L'utilisation d'un gradient stochastique pour l'estimation en ligne dans les HMM a été étudiée par [Le Gland et Mevel, 1997] pour une chaîne de Markov cachée à espace d'état fini et espace d'observations général. Ces résultats ont été précisés dans le travail récent de [Tadić, 2010]. L'extension de ces méthodes à des espaces d'état généraux sous des hypothèses d'ergodicité uniforme de la chaîne cachée a été considérée dans [Poyiadjis *et al.*, 2005] et [Doucet *et al.*, 2011]. Dans ce cas, le calcul du score de Fisher (gradient du logarithme de la vraisemblance conditionnelle) n'est plus explicite et requiert une procédure d'approximation qui est effectuée à l'aide de méthodes de Monte Carlo séquentielles (filtres particulaires). La convergence de tels algorithmes reste un problème ouvert.

Une autre approche proposée pour l'estimation de paramètres en ligne dans les HMM et plus généralement pour les modèles à données latentes est d'utiliser une version récursive de l'algorithme Expectation Maximization (EM). Les travaux pionniers dans cette direction ont été menés par [Titterington, 1984] dans le cas où les observations sont indépendantes (estimation de paramètres de mélange par exemple lorsque les données cachées sont indépendantes). Dans cet algorithme, chaque nouvelle observation permet d'obtenir une nouvelle estimation à partir d'un calcul du score. La nor-

malité asymptotique de cet estimateur a été établie par [Titterington, 1984]. Un autre algorithme de type EM en ligne utilisant une étape d'approximation stochastique a été proposé par [Cappé et Moulines, 2009] toujours pour des observations indépendantes et pour une grande famille de vraisemblances (pour les vraisemblances complètes appartenant à la famille exponentielle courbe). La consistance et la normalité asymptotique de cet estimateur récursif ont été établies sous des hypothèses très générales. L'extension de cet algorithme EM en ligne au cadre des HMM a été entreprise par [Cappé, 2011a] (voir aussi [Mongillo et Denève, 2008]) dans le cas d'un espace d'état fini. Il n'existe pas de résultat de convergence pour ces algorithmes, bien que certains résultats intermédiaires soient proposés dans [Cappé, 2011a]. L'algorithme présenté dans [Cappé, 2011a] a été étendu au cas d'espaces d'état plus généraux à l'aide de méthodes de Monte Carlo séquentielles par [Cappé, 2009] et [Del Moral *et al.*, 2010b].

Dans cette thèse, nous proposons une nouvelle méthode d'estimation en ligne pour les HMM basée sur l'algorithme EM appelée *Block Online Expectation Maximization* (BOEM). Cet algorithme est défini pour des HMM à espace d'état et espace d'observations généraux. La consistance de l'algorithme ainsi que des vitesses de convergence en probabilité ont été prouvées. Nous nous sommes également intéressés à des algorithmes dits "moyennisés" permettant d'obtenir de meilleures vitesses de convergence.

Dans le cas d'espaces d'états généraux, l'implémentation numérique de l'algorithme BOEM requiert d'introduire des méthodes de Monte Carlo séquentielles - aussi appelées méthodes particulaires - pour approcher des espérances conditionnelles sous des lois de lissage qui ne peuvent être calculées explicitement. Nous avons donc proposé une approximation Monte Carlo de l'algorithme BOEM appelée Monte Carlo BOEM. Parmi les hypothèses nécessaires à la convergence de l'algorithme Monte Carlo BOEM, un contrôle de la norme L_p de l'erreur d'approximation Monte Carlo explicite en le nombre d'observations T et le nombre de particules N est nécessaire. Par conséquent, un second volet de cette thèse a été consacré à l'obtention de tels contrôles pour plusieurs méthodes de Monte Carlo séquentielles : l'algorithme Forward Filtering Backward Smoothing introduit par [Doucet *et al.*, 2000] et l'algorithme Forward Filtering Backward Simulation proposé par [Godsill *et al.*, 2004]. Des contrôles de la norme L_p de l'erreur d'approximation Monte Carlo de ces algorithmes ont été obtenus par [Del Moral *et al.*, 2010a] et [Douc *et al.*, 2011a]. Néanmoins, les bornes obtenues dans [Douc *et al.*, 2011a] pour le contrôle des lois de lissage ne sont pas explicites en T et ne peuvent pas être utilisées dans notre cadre d'estimation. Les résultats que nous proposons améliorent ceux fournis par [Del Moral *et al.*, 2010a], notamment en ce qui concerne la dépendance en T et N. Nous avons également obtenu des inégalités de déviations exponentielles pour le contrôle des lois de lissage plus précises que celles proposées dans [Douc *et al.*, 2011a].

Dans un troisième volet de cette thèse, nous considérons des applications de l'algorithme Monte Carlo BOEM à des problèmes de cartographie et de localisation simultanées (SLAM). Ces problèmes se posent lorsque un mobile se déplace dans un environnement inconnu. Il s'agit alors de localiser le mobile tout en construisant une carte de son environnement. De nombreuses solutions ont été apportées pour résoudre ce problème suivant les hypothèses effectuées sur le modèle dynamique du mobile et sur son modèle d'observation (voir [Burgard *et al.*, 2005] pour une présentation de ces solutions). Nous choisissons une modélisation du problème similaire à celle adoptée par [Martinez-Cantin, 2008] dans laquelle le SLAM est formulé comme un problème d'estimation de paramètres dans les modèles de Markov cachés. Nous répondons au problème de cartographie par l'estimation d'un paramètre et au problème de localisation à l'aide de méthodes particulaires. A la différence de [Martinez-Cantin, 2008] qui propose de réaliser l'estimation par maximum de vraisemblance récursif, notre approche est basée sur l'algorithme Monte Carlo BOEM. Nous illustrons l'intérêt de nos méthodes d'estimation pour la résolution du problème du SLAM en considérant :

a) un problème de SLAM plan où les observations reçues par le mobile sont des distances et des angles à des obstacles puis

b) un problème de SLAM indoor où les observations reçues sont des mesures de puissances de signaux WiFi.

Enfin, la dernière partie de cette thèse est relative à l'estimation non paramétrique dans les HMM. Le problème considéré a été très peu étudié et nous avons donc choisi de l'aborder dans un cadre précis. Nous supposons que $\{X_k\}_{k\geq 0}$ est une marche aléatoire sur un sous-espace compact de $\mathbb{R}^m$ dont la loi des incréments est connue à un facteur d'échelle $a_\star$ près. Nous supposons également que, pour tout $k \geq 0$, Y_k est une observation dans un bruit additif gaussien de $f_\star(X_k)$, où $f_\star$ est une fonction à valeurs dans $\mathbb{R}^\ell$ que nous cherchons à estimer. La difficulté de ce problème provient du fait que les points en lesquels la fonction $f_\star$ est évaluée ne sont pas observés.

Dans les modèles à erreurs sur les variables (modèles *errors-in-variables*), les $\{X_k\}_{k\geq 0}$ sont stationnaires et observés en présence de bruit. Il existe de nombreuses méthodes pour estimer la fonction $f_\star$ dans ce contexte. Certaines de ces solutions proposent d'approcher la loi conditionnelle de Y_0 sachant X_0 pour produire une estimation de $f_\star$. Pour cela, il est nécessaire d'obtenir une estimation de la loi de X_0, *i.e.* de répondre au problème de *déconvolution* (voir par exemple [Carroll et Hall, 1988] et [Carroll et Stefanski, 1990]). Des estimateurs à noyau de $f_\star$ ont été proposés (voir [Fan et Truong, 1993]) ainsi que des méthodes de sélection de modèles utilisant une estimation par minimisation de contraste pénalisé (voir [Comte et Taupin, 2007]). Toutes ces procédures nécessitent une observation de l'état caché permettant une estimation de la densité de X_0, elles ne peuvent donc être utilisées dans notre

cas où la seule information disponible est contenue dans les observations $\{Y_k\}_{k\geq 0}$.

Lorsque $\{X_k\}_{k\geq 0}$ est une chaîne de Markov, une estimation de la densité de sa loi invariante et une estimation de la densité de son noyau ont été proposées par [Lacour, 2008a] lorsque les observations sont données par

$$Y_k = X_k + \epsilon_k \ ,$$

où les variables $\{\epsilon_k\}_{k\geq 0}$ sont i.i.d. et de loi connue. Cependant, dans le modèle considéré par [Lacour, 2008a], c'est la chaîne cachée qui est observée au travers du processus $\{Y_k\}_{k\geq 0}$, il n'y a pas de transformation de la chaîne par une fonction $f_\star$.

Le premier résultat que nous avons établi dans notre situation est l'identifiabilité du modèle statistique considéré. Nous avons également proposé une estimation de la fonction $f_\star$ et du paramètre $a_\star$ à partir de la log-vraisemblance par paires des observations. Nous avons prouvé la convergence en probabilité de ces estimateurs lorsque le nombre d'observations utilisées tend vers $+\infty$.

Les résultats obtenus durant cette thèse ont fait l'objet d'articles soumis à des revues internationales. Nous avons choisi d'insérer ces articles dans ce document dans les chapitres 6, 7 et 8. Un préambule est associé à chacun de ces articles pour en expliquer le contenu et pour mettre en valeur notre contribution (voir chapitres 2, 3 et 4). Nous donnons enfin au Chapitre 5 des applications de nos travaux au problème de cartographie et de localisation simultanées. Ce chapitre présente des travaux soumis à des revues qui ne sont pas insérés dans ce document par souci de brièveté (nous décrivons ici la modélisation statistique et nous donnons les résultats expérimentaux, la méthodologie étant déjà donnée dans les autre chapitres). Nous proposons ci-dessous un court résumé de chaque article pour mettre en avant notre contribution. Nous donnons également en Appendice A des preuves supplémentaires pour le Chapitre 6.

i) Chapitre 5 : [Le Corff et Fort, 2011a] et [Dumont et Le Corff, 2012b].
Convergence of a Particle-based Approximation of the Block Online Expectation Maximization Algorithm. *Article accepté pour publication dans la revue Transactions on Modeling and Computer Simulation.* Dans cet article, nous proposons d'appliquer les résultats de [Le Corff et Fort, 2011b] sur l'estimation en ligne des chaînes de Markov cachées. Nous utilisons une méthode de Monte Carlo séquentielle compatible avec la nécessité d'une estimation en ligne. L'erreur d'approximation introduite est contrôlée par les résultats donnés dans l'ar-

ticle [Dubarry et Le Corff, 2011]. L'algorithme proposé est appliqué à des problèmes de cartographie et de localisation simultanées.

Simultaneous localization and mapping problem in wireless sensor networks. *Article soumis pour publication dans la revue IEEE Transactions on Signal Processing.* Cet article introduit une modification de l'algorithme BOEM proposé dans [Le Corff et Fort, 2011b] pour résoudre un problème de cartographie et de localisation simultanées dans un contexte différent de celui donné par [Le Corff et Fort, 2011a]. Dans ce cas, le mobile est localisé à partir de cartes de propagation de signaux WiFi. La méthode proposée a fait l'objet d'un brevet.

ii) Chapitre 6 : [Le Corff et Fort, 2011b], préambule : Chapitre 2.

Online Expectation Maximization based algorithms for inference in Hidden Markov Models. *Article soumis pour publication dans la revue Annals of Statistics.* Dans cet article, nous proposons une nouvelle méthode d'estimation en ligne de chaînes de Markov cachées basée sur l'algorithme EM. Cette méthode traite les observations par blocs de tailles croissantes. Pour chacun de ces blocs d'observations, les statistiques de l'EM sont calculées en ligne, la valeur des paramètres étant fixée. Les paramètres sont mis à jour lorsque toutes les observations du bloc ont été traitées. Les statistiques à calculer sur chaque bloc sont des espérances conditionnelles de fonctionnelles additives sous les lois de lissage jointes qui peuvent être soit calculées explicitement (lorsque l'espace d'état est fini ou pour les modèles linéaires gaussiens) soit approchées par des méthodes de Monte Carlo (pour les modèles HMM généraux). Nous prouvons la consistance de l'algorithme, *i.e.*, la convergence de la suite d'estimateurs vers l'ensemble des points stationnaires de la limite de la log-vraisemblance normalisée lorsque le nombre d'observations tend vers $+\infty$. Nous établissons également des vitesses de convergence en probabilité pour nos estimateurs et pour les estimateurs moyennisés. Cette étude démontre l'intérêt des algorithmes moyennisés : leur vitesse de convergence ne dépend pas de la taille des blocs d'observations et décroît comme $T_n^{-1/2}$, où T_n est le nombre total d'observations utilisées pour produire la n-ième estimation.

iii) Chapitre 7 : [Dubarry et Le Corff, 2011], préambule : Chapitre 3.

Nonasymptotic deviation inequalities for smoothed additive functionals in non-linear state-space models. *Article accepté pour publication dans la revue Bernoulli.* Nous étudions l'approximation de l'espérance conditionnelle de fonctionnelles additives sous les lois de lissage jointes par des méthodes de Monte Carlo séquentielles. Nous nous intéressons à deux algorithmes très employés dans la pratique : l'algorithme *Forward Filtering Backward Smoothing* et l'algorithme *Forward Filtering Backward Simulation.* Les algorithmes considérés procèdent en deux étapes. Dans la première étape, un algorithme de filtrage particu-

laire est utilisé pour approcher les lois de filtrage à l'aide de familles de particules pondérées. Dans une seconde étape, les lois de lissage jointes sont approchées soit en modifiant les poids des particules soit en simulant des trajectoires utilisant les particules et les poids produits dans la première étape. Nous donnons dans cet article des bornes non asymptotiques pour le contrôle de la norme L_p de l'erreur Monte Carlo pour les deux méthodes considérées. Nous proposons également des inégalités de déviation exponentielles pour cette même erreur. Nous montrons en particulier que l'erreur d'approximation en temps long est bornée par une fonction de T/N (où T est le nombre d'observations et N le nombre de particules) uniformément en T et en N sous des hypothèses de mélange fort. La méthode de preuve est basée sur une décomposition de l'erreur d'approximation sous la forme d'une somme d'erreurs locales contrôlées à l'aide d'inégalités de déviation pour les martingales bornées couplées à des résultats d'ergodicité uniforme de la chaîne cachée.

iv) Chapitre 8 : [Dumont et Le Corff, 2012a], préambule : Chapitre 4.
Nonparametric estimation using partially observed Markov chains. *Article soumis pour publication dans la revue Annals of Statistics.* Nous introduisons une nouvelle méthode d'estimation non paramétrique pour les chaînes de Markov cachées : nous cherchons à estimer $f_\star$ dans le cas où, pour tout $k \geq 0$, Y_k est une observation bruitée de $f_\star(X_k)$ et où $\{X_k\}_{k\geq 0}$ est une marche aléatoire restreinte à un sous-espace de $\mathbb{R}^m$, la loi des incréments de la marche étant connue à un facteur d'échelle près. La méthode introduite est basée sur un algorithme de type maximum de vraisemblance non paramétrique par paires (*pairwise maximum likelihood*). Le critère à maximiser est la vraisemblance des paires d'observations qui dépend de la fonction $f_\star$. Pour prendre en compte le caractère non paramétrique du problème, cette vraisemblance est pénalisée par une norme fonctionnelle du paramètre à estimer (dans notre cas une norme Sobolev mais d'autres normes fonctionnelles pourraient être considérées). Nous avons d'abord établi l'identifiabilité du modèle en utilisant des outils de géométrie différentielle (formule de Sard, théorème du point fixe). Nous montrons ensuite la convergence de l'estimation de la loi des paires d'observations, ce qui nous permet enfin d'établir la consistance de l'estimateur de $f_\star$.

1.2 Modèles de Markov cachés : notations

Nous présentons dans cette section les notations sur les modèles de Markov cachés qui seront utilisées dans tous les chapitres de ce document.

Θ désigne un ensemble de paramètres dont la structure varie selon les chapitres. Nous précisons dans chaque chapitre la nature de Θ.

$\mathbb{X}$ et $\mathbb{Y}$ sont deux espaces topologiques munis des tribus boréliennes $\mathcal{X}$ et $\mathcal{Y}$. Dans la plupart des chapitres, $\mathbb{X}$ et $\mathbb{Y}$ sont des sous-espaces de $\mathbb{R}^{d_x}$ et de $\mathbb{R}^{d_y}$.

Un noyau de Markov M défini sur $\mathbb{X} \times \mathcal{X}$ est une application à valeurs dans $[0,1]$ vérifiant les deux propriétés suivantes :

– Pour tout $x \in \mathbb{X}$, $A \mapsto M(x,A)$ est une probabilité sur $(\mathbb{X}, \mathcal{X})$.

– Pour tout $A \in \mathcal{X}$, $x \mapsto M(x,A)$ est une application $\mathcal{X}$-mesurable.

Modèles de Markov cachés

On considère une famille de noyaux de Markov $\{M_\theta\}_{\theta\in\Theta}$, $M_\theta : \mathbb{X} \times \mathcal{X} \to [0,1]$, une mesure σ-finie μ sur $(\mathbb{Y}, \mathcal{Y})$ et une famille de densités de probabilité de transition par rapport à μ, $\{g_\theta(x,\cdot)\}_{\theta\in\Theta, x\in\mathbb{X}}$, $g_\theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}_+$.

Pour tout $\theta \in \Theta$, soit K_θ le noyau de Markov défini sur $(\mathbb{X}\times\mathbb{Y})\times(\mathcal{X}\otimes\mathcal{Y})$, pour tout $(x,y) \in \mathbb{X} \times \mathbb{Y}$ et tout $C \in \mathcal{X} \otimes \mathcal{Y}$, par

$$K_\theta\left[(x,y), C\right] \stackrel{\text{def}}{=} \int \mathbf{1}_C(x',y')\, g_\theta(x',y')\, \mu(\mathrm{d}y')\, M_\theta(x, \mathrm{d}x') \; .$$

Pour tout $\theta \in \Theta$ et toute probabilité χ sur $(\mathbb{X}, \mathcal{X})$, soit $\mathbb{P}_\theta^\chi$ la probabilité sur l'espace canonique $((\mathbb{X} \times \mathbb{Y})^{\mathbb{N}}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes\mathbb{N}})$ telle que $\{(X_k, Y_k)\}_{k\geq 0}$ soit une chaîne de Markov de distribution initiale

$$\mathbb{P}_\theta^\chi((X_0,Y_0) \in C) = \int \mathbf{1}_C(x,y)\, g_\theta(x,y)\, \mu(\mathrm{d}y)\, \chi(\mathrm{d}x)$$

et de noyau de Markov K_θ. L'espérance par rapport à $\mathbb{P}_\theta^\chi$ est notée $\mathbb{E}_\theta^\chi$.

On suppose que, pour tout $\theta \in \Theta$ et tout $x \in \mathbb{X}$, $M_\theta(x,\cdot)$ a pour densité $m_\theta(x,\cdot)$ par rapport à une mesure de référence λ sur $(\mathbb{X}, \mathcal{X})$ supposée finie.

Lois de lissage jointes

Pour toute probabilité χ sur $(\mathbb{X}, \mathcal{X})$, tout $\theta \in \Theta$ et tout $s \leq u \leq v \leq t$, on définit la *loi de lissage jointe* $\phi^\chi_{\theta,u:v|s:t}$ par

$$\mathbb{E}_\theta^\chi\left[f(X_{u:v})|Y_{s:t}\right] \stackrel{\text{def}}{=} \int f(x_{u:v})\phi^\chi_{\theta,u:v|s:t}(\mathrm{d}x_{u:v}) \; , \tag{1.1}$$

où, pour tout $u \leq v$, nous utilisons la notation $x_{u:v}$ pour la suite $(x_u, \cdots, x_v)$. $\phi^\chi_{\theta,u:v|s:t}$ est la loi des variables $X_{u:v}$ conditionnellement aux observations $Y_{s:t}$. Dans le cas où $s = 0$, les lois de lissage jointes sont notées

$$\phi^\chi_{\theta,u:v|t} \stackrel{\text{def}}{=} \phi^\chi_{\theta,u:v|0:t} \; .$$

Ainsi l'espérance sous la loi $\phi^{\chi}_{\theta,u:v|t}$ s'exprime par

$$\begin{aligned}\mathbb{E}^{\chi}_{\theta}\left[f(X_{u:v})|Y_{0:t}\right] &= \int f(x_{u:v})\phi^{\chi}_{\theta,u:v|t}(\mathrm{d}x_{u:v}) \\ &= \frac{\int g_\theta(x_0,Y_0)\prod_{i=1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}f(x_{u:v})\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:t})}{\int g_\theta(x_0,Y_0)\prod_{i=1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:t})} .\end{aligned}$$

Il nous sera également utile d'introduire des notations similaires dans le cas où la loi initiale χ est la loi de la variable X_r (et non de la variable X_0). Pour tout $\theta \in \Theta$ et tout $r \leq u \leq v \leq t$, on note alors

$$\begin{aligned}&\mathbb{E}^{\chi,r}_{\theta}\left[f(X_{u:v})|Y_{r:t}\right] \\ &= \frac{\int g_\theta(x_r,Y_r)\prod_{i=r+1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}f(x_{u:v})\chi(\mathrm{d}x_r)\lambda(\mathrm{d}x_{r+1:t})}{\int g_\theta(x_r,Y_r)\prod_{i=r+1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}\chi(\mathrm{d}x_r)\lambda(\mathrm{d}x_{r+1:t})} .\end{aligned} \tag{1.2}$$

Lois de filtrage

Pour toute probabilité χ sur $(\mathbb{X},\mathcal{X})$, tout $\theta \in \Theta$ et tout $t \geq 0$, on introduit la loi de filtrage $\phi^{\chi}_{\theta,t}$ définie par

$$\phi^{\chi}_{\theta,t} \stackrel{\text{def}}{=} \phi^{\chi}_{\theta,t:t|t} .$$

On a alors

$$\begin{aligned}\mathbb{E}^{\chi}_{\theta}\left[f(X_t)|Y_{0:t}\right] &= \int f(x_t)\phi^{\chi}_{\theta,t}(\mathrm{d}x_t) \\ &= \frac{\int g_\theta(x_0,Y_0)\prod_{i=1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}f(x_t)\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:t})}{\int g_\theta(x_0,Y_0)\prod_{i=1}^{t}\{m_\theta(x_{i-1},x_i)g_\theta(x_i,Y_i)\}\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:t})} .\end{aligned}$$

Vraisemblance complète

Nous introduisons également la vraisemblance complète du modèle,

$$p_\theta(x_{0:T},y_{0:T}) \stackrel{\text{def}}{=} g_\theta(x_0,y_0)\prod_{i=0}^{T-1} m_\theta(x_i,x_{i+1})g_\theta(x_{i+1},y_{i+1}) . \tag{1.3}$$

La structure de dépendance des modèles de Markov cachés, définie par (1.3), est résumée par le modèle graphique donnée en Figure 1.1.

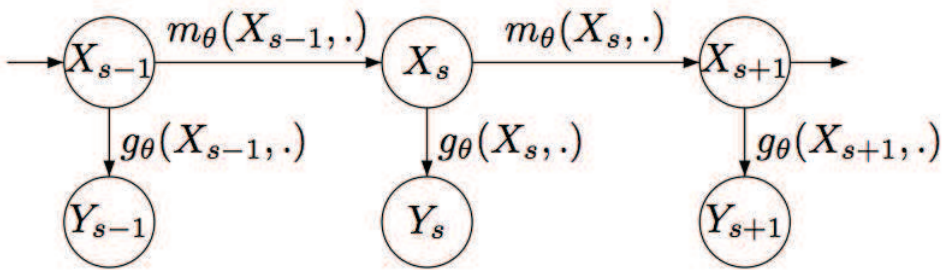


FIGURE 1.1 – Modèle graphique d'un HMM.

Vraisemblance des observations

La log-vraisemblance des observations $y_{0:T}$ est définie, pour tout $T \geq 0$, tout $\theta \in \Theta$ et toute distribution initiale χ sur $(\mathbb{X}, \mathcal{X})$, par

$$\ell^{\chi}_{\theta,T}(y_{0:T}) \stackrel{\text{def}}{=} \log \int p_\theta(x_{0:T}, y_{0:T}) \chi(\mathrm{d}x_0) \lambda(\mathrm{d}x_{1:T}) \; . \tag{1.4}$$

1.3 PRODUCTION SCIENTIFIQUE

Les travaux contenus dans cette thèse ont fait l'objet d'articles soumis à des revues internationales, du dépôt d'un brevet (pour le problème du SLAM indoor) et de communications.

Articles dans des revues internationales

i) Nonasymptotic deviation inequalities for smoothed additive functionals in non-linear state-space models, Dubarry C. and Le Corff S.
Article accepté pour publication dans la revue Bernoulli.

ii) Online Expectation Maximization based algorithms for inference in Hidden Markov Models, Le Corff S. and Fort G.
Article soumis pour publication dans la revue Annals of Statistics.

iii) Convergence of a Particle-based Approximation of the Block Online Expectation Maximization Algorithm, Le Corff S. and Fort G.
Article accepté pour publication dans la revue Transactions on Modeling and Computer Simulation.

iv) Simultaneous localization and mapping problem in wireless sensor networks, Dumont T. and Le Corff S.
Article soumis pour publication dans la revue IEEE Transactions on Signal Processing.

v) Nonparametric estimation using partially observed Markov chains, Dumont T. and Le Corff S.
Article soumis pour publication dans la revue Annals of Statistics.

Actes de conférences internationales

i) New Online EM algorithms for general Hidden Markov models. Application to the SLAM problem, [Le Corff *et al.*, 2012].
10th International Conference on Latent Variable Analysis and Source Separation (LVA/ICA).

ii) Block Online EM for hidden Markov models with general state-space, [Le Corff *et al.*, 2011a].
14th International Conference Applied Stochastic Models and Data Analysis (ASMDA).

iii) Online EM algorithm to solve the SLAM problem, [Le Corff *et al.*, 2011b].
IEEE Workshop on Statistical Signal Processing (SSP).

iv) Fast computation of smoothed additive functionals in general state-space models, [Dubarry et Le Corff, 2011].
IEEE Workshop on Statistical Signal Processing (SSP).

Communications dans des conférences internationales

i) Nonasymptotic upper bounds in a linear version of the FFBS algorithm.
28th European Meeting of Statisticians (EMS), Athènes, 2010.

ii) Nonparametric estimation in hidden Markov models.
8th International Congress on Probability and Statistics (IMS), Istanbul, 2012.

Communications dans des séminaires

i) Convergence properties of particle filters and estimation of smoothed additive functionals in non-linear state-space models.
SMILE : Statistical Machine Learning in Paris, Paris, 2010.

ii) Block online EM for hidden Markov models with application to parameter estimation in general state-spaces.
BigMC : Séminaire Méthodes de Monte Carlo en grande dimension, Paris, 2011.

Codes C et Matlab

Les différents codes écrits en relation avec les algorithmes présentés dans cette thèse (Monte Carlo BOEM, SLAM basé sur des *landmarks*, SLAM indoor) sont disponibles à l'adresse `http://perso.telecom-paristech.fr/~lecorff/software.html`.

Chapitre 2

Estimation en ligne dans les modèles de Markov cachés (préambule)

Dans ce chapitre, nous proposons une nouvelle procédure d'estimation en ligne dans les HMM basée sur l'algorithme Expectation Maximization (EM). La méthode proposée s'applique lorsque la quantité intermédiaire de l'EM est calculable explicitement ou lorsqu'elle est remplacée par une approximation Monte Carlo si l'on sait suffisamment contrôler l'erreur d'approximation. Nous prouvons la consistance de l'estimateur et nous donnons des vitesses de convergence en probabilité. Nous montrons également que la vitesse de convergence optimale est obtenue lorsque l'on utilise un algorithme moyennisé.

L'estimation de paramètres statiques est un problème classique dans les modèles de Markov cachés. Le terme *en ligne* signifie que la procédure d'estimation doit respecter deux critères. D'une part, de nouvelles estimations doivent être produites au fur et à mesure que les observations sont acquises, et ce, de telle sorte qu'il n'est pas nécessaire d'accéder à toutes les observations passées. D'autre part, la capacité de stockage des informations ne doit pas croître avec le temps. Les observations sont donc utilisées puis supprimées puisqu'elles ne seront pas utiles aux estimations futures.

L'intérêt pour ces algorithmes provient d'applications pour lesquelles les données ne peuvent pas être conservées ou qui reposent sur des bases de données trop volumineuses pour être traitées en bloc. Les applications nécessitant un traitement en temps réel sont candidates à l'utilisation de telles techniques en ligne. L'exemple du problème de cartographie et de localisation simultanées est étudié au Chapitre 5.

Dans cette thèse, nous nous intéressons à l'estimation au sens du maximum de vraisemblance et plus particulièrement aux algorithmes de type EM, [Dempster *et al.*, 1977].

2.1 Introduction

Nous commençons par rappeler le fonctionnement de l'algorithme EM et par donner les différentes notations utiles qui s'y rapportent.

Nous considérons une loi χ sur $(\mathbb{X}, \mathcal{X})$ et deux familles de noyaux de Markov dont les densités de probabilité sont données par $\{m_\theta\}_{\theta\in\Theta}$ sur $\mathbb{X}\times\mathbb{X}$ et $\{g_\theta\}_{\theta\in\Theta}$ sur $\mathbb{X}\times\mathbb{Y}$ (voir Section 1.2). Disposant de $T+1$ observations $Y_{0:T}$ et d'une estimation initiale θ_0, l'algorithme EM produit de façon itérative une suite d'estimation $\{\theta_n\}_{n\geq 1}$. Chaque itération de l'EM se déroule en deux étapes. Si θ_n est l'estimation courante du paramètre nous avons alors :

1. **Etape E :** calculer pour tout $\theta \in \Theta$

$$Q(\theta, \theta_n) \stackrel{\text{def}}{=} \mathbb{E}^{\chi}_{\theta_n}\left[\frac{1}{T}\log p_\theta(X_{0:T}, Y_{0:T})\middle| Y_{0:T}\right] ,$$

   où les quantités $\mathbb{E}^{\chi}_{\theta_n}[\cdot|Y_{0:T}]$ et p_θ sont définies par (1.1) et (1.3),

2. **Etape M :** définir θ_{n+1} comme un élément de l'ensemble

$$\text{Argmax}_{\theta\in\Theta} Q(\theta, \theta_n) .$$

Chaque itération vérifie l'inégalité suivante

$$\ell^{\chi}_{\theta,T}(Y_{0:T}) - \ell^{\chi}_{\theta',T}(Y_{0:T}) \geq Q(\theta, \theta') - Q(\theta', \theta') ,$$

où $\ell^{\chi}_{\theta,T}$ est la log-vraisemblance des observations $Y_{0:T}$ définie par (1.4) (voir par exemple [Wu, 1983, Section 1]). Cette inégalité est à la base de l'intérêt pour l'algorithme EM puisqu'elle signifie que chaque itération augmente la valeur de la log-vraisemblance des observations. L'algorithme EM présente cependant deux inconvénients majeurs. D'une part la nécessité d'évaluer la fonction $\theta \mapsto Q(\theta, \theta_n)$ pour toutes les valeurs possibles du paramètre et, d'autre part, la possibilité de maximiser cette quantité sur l'ensemble Θ.

Dans la suite nous nous plaçons dans le cas de modèles de Markov cachés dont la vraisemblance complète appartient à la famille exponentielle courbe. Cette hypothèse nous assure l'existence de fonctions S, ϕ et ψ telles que, pour tout $(x, x') \in \mathbb{X}^2$ et tout $y \in \mathbb{Y}$,

$$m_\theta(x, x')g_\theta(x', y) = \exp\left\{\phi(\theta) + \left\langle S(x, x,', y), \psi(\theta)\right\rangle\right\} . \tag{2.1}$$

Par définition de p_θ, voir (1.3), nous avons pour tout $x_{0:T} \in \mathbb{X}^{T+1}$ et tout $y_{0:T} \in \mathbb{Y}^{T+1}$,

$$\begin{aligned}\frac{1}{T}\log p_\theta(x_{0:T}, y_{0:T}) \\ &= \frac{1}{T}\log g_\theta(x_0, y_0) + \frac{1}{T}\sum_{k=1}^{T} \log\left\{m_\theta(x_{k-1}, x_k) g_\theta(x_k, y_k)\right\} \\ &= \phi(\theta) + \left\langle \frac{1}{T}\sum_{k=1}^{T} S(x_{k-1}, x_k, y_k), \psi(\theta) \right\rangle + \frac{1}{T}\log g_\theta(x_0, y_0) \ .\end{aligned}$$

Dans la suite, la contribution de $g_\theta(x_0, y_0)$ est omise (cette simplification est classique dans l'étude des algorithmes de type EM). L'équation (2.1) permet alors d'écrire, pour tout $\theta \in \Theta$,

$$Q(\theta, \theta_n) = \phi(\theta) + \left\langle \frac{1}{T}\sum_{k=1}^{T} \mathbb{E}_{\theta_n}^{\chi}\left[S(X_{k-1}, X_k, Y_k) | Y_{0:T}\right], \psi(\theta) \right\rangle \ .$$

Nous supposons de plus que ce modèle est tel que, pour tout s dans l'enveloppe convexe de $S(\mathbb{X} \times \mathbb{X} \times \mathbb{Y})$, la fonction $\theta \mapsto \phi(\theta) + \langle s, \psi(\theta)\rangle$ a un maximum unique donné par $\bar{\theta}(s)$. Ces deux hypothèses sont classiques pour l'étude de l'algorithme EM et de ses versions stochastiques (voir par exemple [Delyon *et al.*, 1999], [Fort et Moulines, 2003] et [Cappé, 2011b]). L'algorithme EM devient alors :

1. **Etape E :** calculer

$$\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}_{\theta_n}^{\chi}\left[S(X_{k-1}, X_k, Y_k) | Y_{0:T}\right] \ , \tag{2.2}$$

2. **Etape M :** définir θ_{n+1} par

$$\theta_{n+1} \stackrel{\text{def}}{=} \bar{\theta}\left(\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}_{\theta_n}^{\chi}\left[S(X_{k-1}, X_k, Y_k) | Y_{0:T}\right]\right) \ .$$

L'étape E se résume donc maintenant au seul calcul d'une espérance conditionnelle sous la valeur θ_n du paramètre.

Ce chapitre est organisé de la façon suivante. Nous commençons par présenter notre contribution en Section 2.2 : nous y définissons l'algorithme Block Online Expectation Maximization (BOEM) et nous étudions ses propriétés de convergence. Les résultats présentés dans la Section 2.2 ont fait l'objet de l'article [Le Corff et Fort, 2011b], soumis dans une revue internationale et inséré au Chapitre 6. Cette section détaille également différents algorithmes pour l'estimation en ligne proposés dans la littérature : des

méthodes de type gradient ou basées sur l'algorithme EM sont mises en avant dans le cadre de données manquantes indépendantes ou dans le cadre général des HMM. Le chapitre se termine par l'Annexe 2.3 qui explicite les différentes méthodes permettant les calculs en ligne nécessaires à la mise en place de l'algorithme BOEM.

2.2 Contribution

2.2.1 Algorithmes pour l'estimation en ligne par blocs

Cette section présente notre contribution algorithmique à l'estimation en ligne dans le cadre de modèles de Markov cachés généraux. Nous proposons un nouvel algorithme, le *Block Online EM*, ainsi qu'une version moyennisée. Ces algorithmes nécessitent le calcul d'espérances conditionnelles sous les lois de lissage (voir Section 1.2). Nous proposons un algorithme basé sur une approximation Monte Carlo de ces espérances quand elles ne sont pas calculables explicitement.

L'algorithme proposé, proche de l'EM original, parcourt les observations disponibles en ne mettant à jour l'estimation du paramètre qu'à certains instants prédéfinis de façon déterministe. Si la suite $\{T_k\}_{k\geq 0}$ représente les instants de mises à jour, la valeur du paramètre reste inchangée pendant le parcours des données $Y_{T_k+1:T_{k+1}}$. Nous effectuons, à partir de ces données et de l'estimation courante du paramètre, le calcul de la quantité intermédiaire de l'EM. A la fin de ce bloc de données, le paramètre est alors mis à jour par application de la fonction $\bar{\theta}$.

L'algorithme BOEM

La définition de l'algorithme BOEM nécessite l'introduction de quelques quantités :

i) une suite d'entiers $\{\tau_k\}_{k\geq 1}$ correspondant à la taille des blocs d'observations. Nous posons alors

$$T_0 \stackrel{\text{def}}{=} 0 \quad \text{et, pour tout } n \geq 1 \,, T_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \tau_k \,,$$

ii) une suite $\{\chi_k\}_{k\geq 1}$ de lois sur $(\mathbb{X}, \mathcal{X})$ correspondant aux lois initiales de l'état caché utilisées sur chaque bloc pour calculer la quantité intermédiaire de l'algorithme BOEM.

Pour tout $T > 0$, $\tau > 0$, $\theta \in \Theta$ et toute distribution initiale χ sur $(\mathbb{X}, \mathcal{X})$, nous définissons la quantité $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ par :

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{k=1}^{\tau} \mathbb{E}_\theta^{\chi,T} \left[S(X_{T+k-1}, X_{T+k}, Y_{T+k}) | Y_{T:T+\tau} \right] , \tag{2.3}$$

où $\mathbb{E}_\theta^{\chi,T}[\cdot|Y_{T:T+\tau}]$ est définie par (1.2). $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ correspond à la quantité intermédiaire de l'EM, donnée par (2.2), calculée avec les observations $Y_{T:T+\tau}$, lorsque la distribution de l'état caché à l'instant T est χ et lorsque les lois régissant le HMM sont paramétrées par θ .

L'algorithme BOEM définit alors une suite d'estimations $\{\theta_n\}_{n\geq 0}$ par :

$$\theta_n \stackrel{\text{def}}{=} \bar{\theta} \left[\bar{S}_{\tau_n}^{\chi_{n-1},T_{n-1}}(\theta_{n-1}, \mathbf{Y}) \right] . \tag{2.4}$$

L'algorithme BOEM moyennisé

Nous proposons une version moyennisée de l'algorithme BOEM, qui à chaque fin de bloc, remplace la statistique $\bar{S}_{\tau_n}^{\chi_{n-1},T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ par une moyenne pondérée de toutes les statistiques calculées jusqu'au bloc courant :

$$\Sigma_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{j=1}^{n} \tau_j \, \bar{S}_{\tau_j}^{\chi_{j-1},T_{j-1}}(\theta_{j-1}, \mathbf{Y}) .$$

Nous notons alors $\{\widetilde{\theta}_n\}_{n\geq 0}$ la suite d'estimateurs correspondante :

$$\widetilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta} \left(\Sigma_n \right) .$$

Notons que Σ_n se calcule à partir de $\bar{S}_{\tau_n}^{\chi_{n-1},T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ et de Σ_{n-1} : il n'est pas nécessaire de conserver toutes les statistiques depuis le départ.

L'algorithme Monte Carlo BOEM

Nous considérons enfin un algorithme qui remplace la statistique intermédiaire $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n, \mathbf{Y})$ par une approximation de type Monte Carlo $\widetilde{S}_n$ dans le cas où le calcul exact n'est pas possible.

Il est important à ce stade de voir que la quantité $\bar{S}_{\tau_n}^{\chi_{n-1},T_{n-1}}(\theta_{n-1}, \mathbf{Y})$ peut se calculer en ligne au sein du bloc courant dans le cas d'espaces d'état finis. Dans le cas général, une approximation en ligne peut être calculée à l'aide de méthodes de Monte Carlo séquentielles. La façon d'effectuer ces calculs en ligne est détaillée dans l'Annexe 2.3.

2.2.2 Propriétés de convergence

Les propriétés de convergence des algorithmes BOEM se démontrent en plusieurs étapes que nous rappelons ici. Les énoncés des résultats et les preuves correspondantes sont détaillés au Chapitre 6. Ces résultats reposent sur des hypothèses de différents types précisées au Chapitre 6. Les principales sont rappelées ici :

i) **Paramètres** : l'espace des paramètres Θ est supposé compact. Cette hypothèse assure la stabilité de l'algorithme Monte Carlo BOEM. Elle peut être relâchée en combinant l'algorithme avec une procédure de type reprojection des estimations sur une suite croissante d'espaces compacts telle que celle proposée par [Chen *et al.*, 1988]. Cette procédure n'est pas détaillée au Chapitre 6 par souci de simplicité, mais les résultats que nous établissons peuvent être étendus aux algorithmes BOEM stabilisés par reprojection (voir [Le Corff et Fort, 2011b, Arxiv, version 2] et [Fort et Moulines, 2003] pour une application au cas d'un algorithme Monte Carlo EM),

ii) **Modèle de Markov caché** : nous supposons la condition de mélange fort,

$$\exists(\sigma_-,\sigma_+)\in\mathbb{R}_+^2\ ,\ \forall\theta\in\Theta\ ,\ \forall(x,x')\in\mathbb{X}^2\ ,\ \sigma_-\leq m_\theta(x,x')\leq\sigma_+\ .$$

Cette hypothèse permet d'établir l'oubli géométrique des filtres à la fois dans le sens *forward* et dans le sens *backward*, voir Proposition 6.5. Cet oubli de la chaîne cachée est un élément capital pour prouver la convergence de la quantité intermédiaire (2.3) lorsque $\tau\to+\infty$, voir Théorème 6.1-ii). Grâce à cette convergence, nous pouvons définir un algorithme limite dont les propriétés de convergence permettent l'analyse de l'algorithme Monte Carlo BOEM,

iii) **Observations** : les observations sont supposées stationnaires et β-mélangeantes, voir le Chapitre 6 pour une définition détaillée de la notion de β-mélange. La stationnarité permet d'appliquer un théorème ergodique, en particulier dans la preuve de la convergence de la quantité intermédiaire (2.3) lorsque $\tau\to+\infty$. L'hypothèse de β-mélange est quant à elle fondamentale pour contrôler la norme L_p de l'erreur d'approximation dans l'algorithme Monte Carlo BOEM,

iv) **Taille des blocs** : nous supposons que les tailles des blocs d'observations croissent de façon polynomiale : $\tau_n=\lfloor cn^a\rfloor$, avec $a>1$. Le fait que $\tau_n\to+\infty$ est naturel puisque cela permet de comparer l'algorithme Monte Carlo BOEM à l'algorithme limite mentionné plus haut. Une croissance logarithmique de la taille des blocs n'est pas suffisante pour assurer la convergence de l'algorithme Monte Carlo BOEM. Nous pourrions choisir une croissance géométrique en conservant les propriétés de convergence. Ce n'est cependant pas recommandé d'un point de vue

pratique (un nombre important d'observations est nécessaire à la production d'une nouvelle estimation),

v) **Approximation Monte Carlo** : il existe $p > 2$ tel que la norme L_p de l'erreur d'approximation Monte Carlo au sein de chaque bloc vérifie, pour tout $n \geq 0$,

$$\left\| \bar{S}^{\chi_n, T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \widetilde{S}_n \right\|_p = O(\tau_{n+1}^{-b}) \ ,$$

pour b suffisamment grand.

Algorithme limite

L'analyse de la convergence de nos algorithmes repose sur la comparaison de ces algorithmes à une procédure itérative déterministe que nous appelons *limiting EM*. Nous prouvons que la quantité intermédiaire $\bar{S}^{\chi,T}_{\tau}(\theta, \mathbf{Y})$ définie par (2.3), converge presque sûrement, lorsque $\tau \to +\infty$, vers une quantité limite notée $\bar{S}(\theta)$, la limite étant indépendante de T et de χ (voir Théorème 6.1). Ce résultat repose sur une application du théorème ergodique pour le processus stationnaire $\mathbf{Y}$, cette quantité limite ne dépend donc des observations qu'à travers leur loi.

Ce résultat permet de définir un algorithme limite, appelé *limiting EM*, donné par

$$\check{\theta}_n \overset{\text{def}}{=} \mathrm{R}(\check{\theta}_{n-1}) \overset{\text{def}}{=} \bar{\theta}\left(\bar{S}(\check{\theta}_{n-1})\right) \ . \tag{2.5}$$

Le premier résultat intermédiaire (donné par la Théorème 6.2) prouve que la séquence générée par le *limiting EM* converge vers l'ensemble des points stationnaires $\mathcal{L}$ de la mise à jour :

$$\mathcal{L} \overset{\text{def}}{=} \{\theta \in \Theta \ ; \ \mathrm{R}(\theta) = \theta\} \ .$$

La convergence repose sur l'existence d'une fonction de Lyapunov pour le *limiting EM* (voir Proposition 6.1). Pour tout $\theta \in \Theta$ et toute distribution χ sur $(\mathbb{X}, \mathcal{X})$, la log-vraisemblance normalisée $T^{-1}\ell^{\chi}_{\theta,T}(Y_{0:T})$, donnée par (1.4), converge $\mathbb{P}-$p.s. vers une quantité notée $\ell(\theta)$ lorsque $T \to +\infty$, cette limite étant indépendante de χ. La fonction de Lyapunov considérée est simplement définie par $\mathrm{W} : \theta \mapsto \exp(\ell(\theta))$ (l'utilisation de la fonction exponentielle assure la positivité nécessaire pour définir une fonction de Lyapunov).

Ce résultat nous permet aussi d'identifier les points stationnaires de l'ensemble $\mathcal{L}$ aux points stationnaires de cette log-vraisemblance limite ℓ. Cette remarque est fondamentale puisqu'elle permet, dans le cas où les observations utilisées correspondent aux observations d'un HMM stationnaire paramétré par $\theta_\star \in \Theta$, de prouver que $\theta_\star$ est un élément de cet ensemble de points limites. Elle nous permettra également de faire une analogie avec les algorithmes proposés dans le cas indépendant, voir Section 2.2.3.

Consistance des algorithmes Monte Carlo BOEM

La convergence du limiting EM étant établie, le Théorème 6.2 donne également la convergence de l'algorithme Monte Carlo BOEM. Il s'agit de comparer la mise à jour du paramère θ_n via une itération du *limiting EM*, donnée par $\bar{\theta}\left(\bar{S}(\theta_n)\right)$, à celle donnée par une itération de l'algorithme Monte Carlo BOEM, $\bar{\theta}\left(\widetilde{S}_n\right)$, où $\widetilde{S}_n$ est l'approximation Monte Carlo de $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y})$. En effet, la convergence de l'algorithme Monte Carlo BOEM vers l'ensemble $\mathcal{L}$ est assurée dès lors que nous prouvons que

$$\left|\mathrm{W}\left(\bar{\theta}\left(\widetilde{S}_n\right)\right)-\mathrm{W}\left(\bar{\theta}\left(\bar{S}(\theta_n)\right)\right)\right| \underset{n\to+\infty}{\longrightarrow} 0\ ,\mathbb{P}-\text{p.s.}$$

Sous nos hypothèses, $\bar{\theta}$ et W sont régulières de sorte qu'il suffit d'avoir un contrôle de la norme L_p de l'erreur $\widetilde{S}_n-\bar{S}(\theta_n)$ et d'appliquer le lemme de Borel-Cantelli. Ce contrôle est fourni par le Théorème 6.1 et provient de deux bornes intermédiaires :

i) du contrôle de la norme L_p de l'erreur entre $\bar{S}(\theta_n)$ et $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y})$ fourni par les lemmes 6.4, 6.5 et 6.6,

ii) de l'hypothèse de contrôle de la norme L_p de l'erreur entre $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y})$ et $\widetilde{S}_n$.

Nous verrons au Chapitre 3 comment l'hypothèse ii) de contrôle de l'erreur Monte Carlo effectuée sur chaque bloc peut être vérifiée dans le cadre de méthodes de Monte Carlo séquentielles.

Notons que la convergence de l'algorithme BOEM découle alors de celle de l'algorithme Monte Carlo BOEM, puisque dans ce cas seul le contrôle i) est nécessaire.

Le long de toute séquence $\{\theta_n\}_{n\geq 0}$ convergente, nous montrons aisément la convergence de la suite moyennisée $\{\widetilde{\theta}_n\}_{n\geq 0}$ en utilisant le théorème de Césaro.

Vitesses de convergence des algorithmes Monte Carlo BOEM

Nous étudions les vitesses de convergence de nos algorithmes le long d'une suite $\{\theta_n\}_{n\geq 0}$ convergeant vers un élément $\theta_\star\in\mathcal{L}$. Dans la suite nous écrivons $Z_n=O_{\mathrm{L}_p}(1)$ si $\limsup_n \mathbb{E}\left[|Z_n|^p\right]<\infty$ et $Z_n=O_{\mathrm{a.s}}(1)$ si $\sup_n|Z_n|<+\infty$, $\mathbb{P}-$p.s.

Pour la suite non moyennisée, le Théorème 6.3 donne

$$\sqrt{\tau_n}\ \left[\theta_n-\theta_\star\right]\mathbf{1}_{\lim_n\theta_n=\theta_\star}=O_{\mathrm{L}_{p/2}}(1)O_{\mathrm{a.s}}\left(1\right)\ ,$$

où $p>2$ provient du contrôle de l'erreur de l'approximation Monte Carlo évoqué en hypothèse. Ceci permet en particulier d'écrire

$$\lim_{M\to+\infty}\limsup_{n\to+\infty}\mathbb{P}\left\{\sqrt{\tau_n}\ \|\theta_n-\theta_\star\|\mathbf{1}_{\lim_n\theta_n=\theta_\star}\geq M\right\}=0\ .$$

Ce résultat prouve que la vitesse de convergence dépend alors fortement du choix fait pour la suite $\{\tau_n\}_{n\geq 0}$. De plus, il s'agit de la vitesse après n itérations de l'algorithme. Nous pouvons également nous intéresser à la vitesse de convergence après le traitement de n observations. Nous utilisons alors une autre suite d'estimateurs indexée par le nombre d'observations $\{\theta_k^i\}_{k\geq 0}$. Pour tout $k \geq 0$, θ_k^i est définie comme étant la valeur θ_n, où n est l'unique entier tel que $k \in [T_n + 1, T_{n+1}]$. La suite ainsi définie est constante par morceaux et évolue aux instants $\{T_k\}_{k\geq 0}$. Le résultat qui précède permet d'obtenir une vitesse de convergence égale à $k^{-a/(2(a+1))}$ (à une constante multiplicative près) pour la suite $\{\theta_k^i\}_{k\geq 0}$ dans le cas où $\tau_n = \lfloor cn^a \rfloor$. La vitesse de convergence pour $\{\theta_k^i\}_{k\geq 0}$ dépend donc de la vitesse de croissance des tailles des blocs (à travers le paramètre a) et est plus lente que $k^{-1/2}$.

Il est alors intéressant d'étudier la suite moyennisée $\{\widetilde{\theta}_n\}_{n\geq 0}$. Pour cette dernière, le Théorème 6.4 établit que

$$\sqrt{T_n}\left[\widetilde{\theta}_n - \theta_\star\right] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) \ ,$$

et par suite

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{\sqrt{T_n}\ \|\widetilde{\theta}_n - \theta_\star\| \mathbf{1}_{\lim_n \theta_n = \theta_\star} \geq M\right\} = 0 \ .$$

La vitesse de convergence est alors donnée par $T_n^{-1/2}$, autrement dit par l'inverse de la racine carrée du nombre total d'observations utilisées depuis le départ (et pas simplement depuis le début du bloc courant). Dans le cas de l'algorithme moyennisé, la vitesse de convergence de la suite $\{\widetilde{\theta}_k^i\}_{k\geq 0}$ associée est alors $k^{-1/2}$, quelle que soit la valeur de $a > 1$ lorsque $\tau_n = \lfloor cn^a \rfloor$. La vitesse est donc meilleure que pour la suite $\{\theta_k^i\}_{k\geq 0}$ et ne dépend plus de la vitesse de croissance de la taille des blocs.

2.2.3 Comparaison à la littérature

Il existe dans la littérature des résultats pour l'estimation en ligne au sens du maximum de vraisemblance dans le cas de données manquantes indépendantes (voir [Titterington, 1984] et [Cappé et Moulines, 2009]), dans le cas de modèles de Markov vachés à espace d'état fini (voir par exemple [Le Gland et Mevel, 1997], [Mongillo et Denève, 2008], [Tadić, 2010] ou encore [Cappé, 2011a]) et dans le cas de modèles de Markov cachés à espace d'état plus général (voir par exemple [Del Moral *et al.*, 2010b]).

Données manquantes indépendantes et identiquement distribuées

Nous nous plaçons tout d'abord dans le cas où les données manquantes $\{X_k\}_{k\geq 0}$ sont indépendantes et identiquement distribuées. Ainsi, pour tout

$\theta \in \Theta$ et tout $(x, x') \in \mathbb{X}^2$, $m_\theta(x, x') = m_\theta(x')$. Le modèle graphique dans une telle situation est donné par la Figure 2.1.

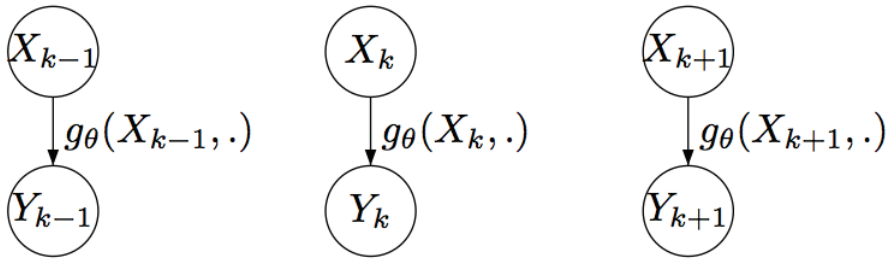


FIGURE 2.1 – Modèle graphique lorsque les données manquantes sont indépendantes.

Nous omettons la notation de la distribution initiale dans les espérances conditionnelles de cette section puisque les données manquantes sont i.i.d. Dans le cas d'un modèle exponentiel courbe, l'équation (2.1) s'écrit alors

$$m_\theta(x) g_\theta(x, y) = \exp\left\{\phi(\theta) + \langle S(x, y), \psi(\theta)\rangle\right\} \ .$$

L'équation (2.2) permet de comprendre le comportement limite de l'EM lorsque le nombre d'observations utilisées tend vers $+\infty$. La quantité intermédiaire de l'EM donnée par (2.2) devient, dans le cas indépendant,

$$Q(\theta, \theta_n) = \phi(\theta) + \left\langle \frac{1}{T} \sum_{k=0}^{T} \mathbb{E}_{\theta_n}\left[S(X_k, Y_k)|Y_k\right], \psi(\theta) \right\rangle \ , \tag{2.6}$$

où

$$\mathbb{E}_{\theta_n}\left[S(X_k, Y_k)|Y_k\right] \stackrel{\text{def}}{=} \frac{\int m_{\theta_n}(x_k) g_{\theta_n}(x_k, Y_k) S(x_k) \lambda(\mathrm{d}x_k)}{\int m_{\theta_n}(x_k) g_{\theta_n}(x_k, Y_k) \lambda(\mathrm{d}x_k)} \ .$$

De façon similaire à la démarche que nous adoptons pour les algorithmes BOEM, l'algorithme EM en ligne proposé par [Cappé et Moulines, 2009] est une perturbation d'un algorithme limite. Cet algorithme limite est obtenu en considérant le comportement limite de (2.6) lorsque le nombre d'observations T tend vers $+\infty$. Si nous notons π la densité commune des observations par rapport à la mesure μ, [Cappé et Moulines, 2009] prouve que la quantité intermédiaire de l'étape E de l'EM converge π-p.s., quand le nombre d'observations croît vers $+\infty$:

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}_{\theta_n}\left[S(X_k, Y_k)|Y_k\right] \underset{T \to +\infty}{\longrightarrow} \mathbb{E}_\pi\left[\mathbb{E}_{\theta_n}\left[S(X_0, Y_0)|Y_0\right]\right] \ , \quad \pi - \text{p.s.} \ ,$$

où pour toute fonction mesurable f, $\mathbb{E}_\pi[f] \stackrel{\text{def}}{=} \int f(y)\pi(\mathrm{d}y)$. Ce résultat permet de considérer un nouvel algorithme qui effectue la mise à jour suivante

$$\check{S}_n \stackrel{\text{def}}{=} G(\check{S}_{n-1}) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\mathbb{E}_{\bar{\theta}\left(\check{S}_{n-1}\right)} \left[S(X_0, Y_0) | Y_0 \right] \right] , \tag{2.7}$$

$$\check{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}(\check{S}_n) \stackrel{\text{def}}{=} \bar{\theta} \left(\mathbb{E}_\pi \left[\mathbb{E}_{\theta_{n-1}} \left[S(X_0, Y_0) | Y_0 \right] \right] \right) . \tag{2.8}$$

L'algorithme limite donné par l'équation (2.8) est l'équivalent dans le cas indépendant de l'algorithme limiting EM que nous proposons pour les HMM généraux défini par (2.5). L'équation (2.7) donne la mise à jour équivalente pour les statistiques qui est utile pour justifier la procédure d'approximation stochastique définie plus bas. [Cappé, 2011b] montre que les points fixes du limiting EM sont les points stationnaires de la divergence de *Kullback-Leibler* entre la vraie loi des observations et la loi paramétrée par θ :

$$\mathrm{KL}(\pi, g_\theta) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\log \pi(Y_0) \right] - \mathbb{E}_\pi \left[\ell_\theta(Y_0) \right] , \tag{2.9}$$

où $\ell_\theta(Y_0)$ est la log-vraisemblance de l'observation Y_0 définie par

$$\ell_\theta(Y_0) \stackrel{\text{def}}{=} \log \int m_\theta(x) g_\theta(x, Y_0) \lambda(\mathrm{d}x) . \tag{2.10}$$

D'autre part, les données $\{Y_k\}_{k\geq 0}$ étant indépendantes, la log-vraisemblance normalisée des observations vérifie

$$\frac{1}{n} \sum_{k=1}^{n} \ell_\theta(Y_k) \underset{n\to+\infty}{\longrightarrow} \mathbb{E}_\pi \left[\ell_\theta(Y_0) \right] , \pi - \text{p.s.}$$

Ainsi, minimiser la divergence de Kullback-Leibler définie par (2.9) équivaut à maximiser la log-vraisemblance limite $\mathbb{E}_\pi \left[\ell_\theta(Y_0) \right]$ et les points fixes de la mise à jour (2.8) sont aussi les points stationnaires de cette vraisemblance limite. Ce résultat est à mettre en lien avec l'ensemble des points limites de l'algorithme limiting EM que nous proposons dans le cas des HMM : ce sont les points stationnaires de la log-vraisemblance limite dans le cas HMM.

[Cappé et Moulines, 2009] propose donc un algorithme EM en ligne dans le cas i.i.d. dont l'objectif est de trouver les racines de l'équation

$$\mathbb{E}_\pi \left[\mathbb{E}_{\bar{\theta}(s)} \left[S(X_0, Y_0) | Y_0 \right] \right] - s = 0 , \tag{2.11}$$

qui correspondent aux points fixes du limiting EM défini par (2.7). Par application de la fonction $\bar{\theta}$ nous retrouvons alors aisément les points fixes de l'algorithme (2.7) défini sur l'espace des paramètres. Les variables $\{Y_k\}_{k\geq 0}$ étant indépendantes, nous nous trouvons dans une situation classique où nous pouvons appliquer un algorithme d'approximation stochastique (voir

par exemple [Kushner et Yin, 2003]). Ainsi, S_0 étant choisi arbitrairement, les solutions de (2.11) sont obtenues en calculant successivement

$$\begin{aligned} S_n &= (1-\gamma_n)S_{n-1} + \gamma_n \mathbb{E}_{\bar{\theta}(S_{n-1})}\left[S(X_n, Y_n)|Y_n\right] \ , \\ \theta_n &= \bar{\theta}(S_n) \ . \end{aligned}$$

Sous certaines hypothèses de régularité et avec des hypothèses classiques en approximation stochastique sur la suite $\{\gamma_n\}_{n\geq 1}$, [Cappé et Moulines, 2009, Théorème 5] prouve que la suite $\{\theta_n\}_{n\geq 0}$ converge vers l'ensemble des points stationnaires de la divergence de Kullback-Leibler définie par (2.9).

L'algorithme ainsi proposé produit donc une nouvelle estimation chaque fois qu'une nouvelle observation est disponible, ce qui constitue la différence majeure entre cet algorithme et nos algorithmes BOEM. Ceci provient du fait que du l'algorithme limite donné par (2.8) ne fait intervenir que la loi d'une observation. Dans le cas des HMM, l'algorithme limiting EM que nous introduisons dépend de tout le processus $\mathbf{Y}$ du fait de la dépendance entre les observations, ce qui nous conduit à utiliser des blocs d'observations.

[Cappé et Moulines, 2009, Théorème 7] prouve que l'estimateur proposé est asymptotiquement normal et que la convergence a lieu à la vitesse $\gamma_n^{-1/2}$ lorsque $\gamma_n = \gamma_0 n^{-\alpha}$ pour $\alpha \in (1/2, 1)$. Pour obtenir la même vitesse pour la suite $\{\theta_n^i\}_{n\geq 0}$ associée à l'algorithme BOEM, il nous faut donc choisir $\tau_n = \lfloor cn^a \rfloor$, avec $a = \alpha/(1-\alpha)$.

L'estimateur moyennisé proposé par [Cappé et Moulines, 2009] et défini, pour $n \geq n_0$ par (voir [Polyak, 1990, Polyak et Juditsky, 1992])

$$\widetilde{\theta}_n \stackrel{\text{def}}{=} \frac{1}{n-n_0+1} \sum_{k=n_0}^{n} \theta_n \ ,$$

converge à la vitesse $n^{-1/2}$ pour toute valeur de $\alpha \in (0, 1)$. La vitesse de convergence est donc indépendante du choix de la suite des pas $\{\gamma_n\}_{n\geq 0}$. Nous retrouvons donc les mêmes propriétés que pour l'estimateur moyennisé $\{\widetilde{\theta}_n^i\}_{n\geq 0}$ de l'algorithme BOEM. A noter toutefois que les vitesses données dans [Cappé et Moulines, 2009] sont valables pour des théorèmes de type Théorème Limite Central, de tels résultats peuvent constituer une extension future de nos travaux.

Les résultats de [Cappé et Moulines, 2009] sont similaires à ceux obtenus pour la procédure de type gradient de [Titterington, 1984] pour laquelle la mise à jour du paramètre prend la forme

$$\theta_{n+1} = \theta_n + \gamma_{n+1} I_c^{-1}(\theta_n) \nabla_{\theta=\theta_n} \ell_\theta(Y_{n+1}) \ , \tag{2.12}$$

où $\{\gamma_n\}_{n\geq 1}$ est une suite de réels strictement positifs, $\ell_\theta(Y_{n+1})$ est la log-vraisemblance de l'observation Y_{n+1} définie par (2.10) et où I_c est la matrice

d'information de Fisher associée aux données complètes, définie par

$$I_c^{-1}(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_\theta \left[\nabla_\theta^2 \log \left\{m_\theta(X_0) g_\theta(X_0, Y_0)\right\}\right] \\ = -\int m_\theta(x) g_\theta(x,y) \nabla_\theta^2 \log \left\{m_\theta(x) g_\theta(x,y)\right\} \lambda(\mathrm{d}x)\nu(\mathrm{d}y) \ .$$

Asymptotiquement et dans le cas de modèles bien spécifiés, la mise à jour proposée par [Cappé et Moulines, 2009] est équivalente à une mise à jour de type gradient similaire à (2.12), ce qui justifie les mêmes propriétés de convergence.

Modèles de Markov cachés généraux

L'hypothèse d'indépendance de la section précédente est assez restrictive et ne permet pas d'obtenir des algorithmes d'estimation en ligne pour les modèles de Markov cachés généraux. La dépendance entre les observations rend la tâche d'estimation bien plus délicate que dans le cas indépendant. Comme pour le cas des données manquantes indépendantes, il existe des algorithmes d'estimation en ligne basés sur des méthodes de gradient (voir par exemple [Le Gland et Mevel, 1997] et [Tadić, 2010]). En ce qui concerne les algorithmes de type EM, [Mongillo et Denève, 2008] propose un algorithme dans le cas où l'espace d'état et l'espace des observations sont finis. Cet algorithme a été généralisé par [Cappé, 2011a] au cas où les observations sont à valeurs dans un espace général. Nous discutons donc uniquement la procédure proposée par [Cappé, 2011a].

Cet algorithme s'appuie sur le travail de [Cappé et Moulines, 2009] pour les données manquantes i.i.d. La mise à jour proposée par [Cappé, 2011a] s'inspire de l'algorithme EM en ligne de [Cappé et Moulines, 2009] et le généralise au cas où les données ne sont pas indépendantes. Cependant, cette même dépendance ne permet pas l'analyse de la convergence de la procédure en utilisant les arguments classiques d'approximation stochastique. Seul [Cappé, 2011a, Théorème 1], donne quelques éléments pour comprendre le comportement asymptotique de l'algorithme. En se plaçant dans le cas où les observations $\mathbf{Y}$ sont stationnaires, ce théorème prouve, dans le cas d'espaces d'état finis, qu'il existe une fonction $\bar{S}$ telle que :

$$\bar{S}_T^{\chi,0}(\theta, \mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \mathbb{E}\left[\bar{S}(\theta)\right] \ , \quad \mathbb{P}-\text{p.s} \ .$$

Cela permet alors de définir un algorithme limiting EM défini par $\check{\theta}_{n+1} = \bar{\theta}(\bar{S}(\check{\theta}_n))$. L'algorithme limiting EM que nous proposons pour étudier la convergence de l'algorithme BOEM est une généralisation de ce résultat à des espaces d'état non nécessairement finis.

L'objectif est alors d'étudier la convergence de ce limiting EM et de comprendre l'algorithme EM en ligne comme une perturbation de ce dernier.

Il n'existe cependant pas de résultat de convergence pour l'algorithme donné par [Cappé, 2011a]. Si l'espace d'état n'est pas fini, [Del Moral *et al.*, 2010b] étend cet algorithme en ligne en utilisant des des Méthodes de Monte Carlo. Dans ce cas, l'algorithme proposé remplace, dans les mises à jour (2.13) et (2.14) données plus bas, les lois de filtrage par des approximations obtenues à l'aide de méthodes de Monte Carlo séquentielles. Ces calculs sont détaillés dans la section suivante.

L'algorithme d'estimation en ligne de [Cappé, 2011a] met en oeuvre les équations (2.16) et (2.17), données en Annexe 2.3, mais en changeant à chaque itération la valeur du paramètre. Si, à l'instant T, l'estimation du paramètre est θ_T, la loi de filtrage $\phi^{\chi}_{\theta_T,T}$ (donnée par (1.1)) est approchée par $\widehat{\phi}_T$ et $\widehat{\rho}^{\chi}_{\theta_T,T}$ est approché par $\widehat{\rho}_T$, lorsque l'observation Y_{T+1} est reçue l'algorithme procède de la façon suivante :

1. **Etape E :** Calculer pour tout $x \in \mathbb{X}$,

$$\widehat{\phi}_{T+1}(x) = \frac{\sum_{x' \in \mathbb{X}} \widehat{\phi}_T(x') m_{\theta_T}(x', x) g_{\theta_T}(x, Y_{T+1})}{\sum_{x', x'' \in \mathbb{X}} \widehat{\phi}_T(x') m_{\theta_T}(x', x'') g_{\theta_T}(x'', Y_{T+1})} , \tag{2.13}$$

et

$$\widehat{\rho}_{T+1}(x) = \sum_{x' \in \mathbb{X}} \left[\gamma_{T+1} S(x', x, Y_{T+1}) + (1 - \gamma_{T+1}) \widehat{\rho}_T(x')\right] \times \frac{\widehat{\phi}_T(x') m_{\theta_T}(x', x)}{\sum_{x'' \in \mathbb{X}} \widehat{\phi}_T(x'') m_{\theta_T}(x'', x)} . \tag{2.14}$$

2. **Etape M :** Définir

$$\theta_{T+1} = \bar{\theta} \left(\sum_{x \in \mathbb{X}} \widehat{\rho}_{T+1}(x) \widehat{\phi}_{T+1}(x) \right) .$$

2.3 Annexe : calculs récursifs pour les fonctionnelles additives dans les HMM

Nous justifions dans cette section le caractère en ligne de l'algorithme BOEM que nous proposons. Nous illustrons donc la façon dont la quantité intermédiaire de l'algorithme se calcule au fur et à mesure que les observations sont reçues dans le cas d'un espace d'état fini puis dans un cas plus général en utilisant des méthodes de Monte Carlo séquentielles.

2.3.1 Espace d'état fini

La mise en place de l'algorithme BOEM nécessite de pouvoir calculer, pour $\theta \in \Theta$, des quantités de la forme

$$\bar{S}_T^{\chi,0}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=1}^{n} \mathbb{E}_\theta^\chi \left[S(X_{k-1}, X_k, Y_k) | Y_{0:T} \right] ,$$

où $\mathbb{E}_\theta^\chi \left[\cdot | Y_{0:T} \right]$ est défini par (1.1). Le calcul récursif d'espérances conditionnelles de fonctionnelles additives dans les HMM à espace d'état fini a été étudié par [Zeitouni et Dembo, 1988], [Elliott et Krishnamurthy, 1999] ou encore [Cappé, 2001]. L'algorithme qui suit a été proposé par [Cappé, 2011a]. Soit

$$\rho_{\theta,T}^\chi(X_T) \stackrel{\text{def}}{=} \frac{1}{T} \sum_{k=1}^{T} \mathbb{E}_\theta^\chi \left[S(X_{k-1}, X_k, Y_k) | Y_{0:T}, X_T \right] ,$$

nous avons, de façon immédiate,

$$\bar{S}_T^{\chi,0}(\theta, \mathbf{Y}) = \mathbb{E}_\theta^\chi \left[\rho_{\theta,T}^\chi(X_T) \middle| Y_{0:T} \right] = \phi_{\theta,T}^\chi[\rho_{\theta,T}^\chi] , \tag{2.15}$$

où la loi de filtrage $\phi_{\theta,T}^\chi$ est définie par (1.1). Un calcul récursif de $\bar{S}_T^{\chi,0}(\theta, \mathbf{Y})$ nécessite donc deux étapes intermédiaires :

i) Un calcul récursif de $\{\rho_{\theta,t}^\chi(x)\}_{t=0}^T$ pour tout $x \in \mathbb{X}$.

ii) Un calcul récursif de $\{\phi_{\theta,t}^\chi(x)\}_{t=0}^T$ pour tout $x \in \mathbb{X}$.

Par les formules de Bayes, nous avons immédiatement

$$\phi_{\theta,t}^\chi(x) = \frac{\sum_{x' \in \mathbb{X}} \phi_{\theta,t-1}^\chi(x') m_\theta(x', x) g_\theta(x, Y_t)}{\sum_{x', x'' \in \mathbb{X}} \phi_{\theta,t-1}^\chi(x') m_\theta(x', x'') g_\theta(x'', Y_t)} . \tag{2.16}$$

D'autre part, par les propriétés des HMM (voir [Cappé *et al.*, 2005, Section 4]),

$$\rho_{\theta,t}^\chi(x) = \sum_{x' \in \mathbb{X}} \left[\frac{1}{t} S(x', x, Y_t) + \left(1 - \frac{1}{t} \right) \rho_{\theta,t-1}^\chi(x') \right] \times \frac{\phi_{\theta,t-1}^\chi(x') m_\theta(x', x)}{\sum_{x'' \in \mathbb{X}} \phi_{\theta,t-1}^\chi(x'') m_\theta(x'', x)} . \tag{2.17}$$

Nous concluons par (2.15) que

$$\bar{S}_T^{\chi,0}(\theta, \mathbf{Y}) = \sum_{x \in \mathbb{X}} \phi_{\theta,T}^\chi(x) \rho_{\theta,T}^\chi(x) .$$

Il est important de noter que l'algorithme ainsi décrit n'est valable que pour des modèles de Markov cachés à espaces d'état finis puisque les mises à jour nécessitent de sommer sur toutes les valeurs possibles des états.

2.3.2 Cas général

Lorsque l'espace d'état n'est pas fini ou lorsqu'il contient un grand nombre d'éléments, il est possible de tirer profit des équations (2.16) et (2.17) pour définir des approximations à l'aide de méthodes de Monte Carlo séquentielles. La méthode que nous présentons ici est donnée dans [Del Moral *et al.*, 2010b] et [Cappé, 2011a]. Dans ce cas, chaque loi de filtrage $\phi^{\chi}_{\theta,t}$, pour $t \in \{0, \cdots, T\}$, est approchée par un nuage de particules pondérées : $\left\{\left(\xi_t^{N,\ell}, \omega_t^{N,\ell}\right)\right\}_{\ell=1}^{N}$, voir Section 3.1 pour les détails sur la façon de produire ces particules pondérées. Nous présentons également en Section 5.3 un algorithme détaillé combinant le calcul récursif donné ci-dessous et la méthode particulaire de la Section 3.1.

Pour toute fonction mesurable h sur $\mathbb{X}$, la quantité $\phi^{\chi}_{t,\theta}[h]$ est approchée par $\phi^{N,ps}_{t,\theta}[h]$, où

$$\phi^{N,ps}_{t,\theta}[h] \stackrel{\text{def}}{=} \frac{1}{\sum_{\ell=1}^{N} \omega_t^{N,\ell}} \sum_{\ell=1}^{N} \omega_t^{N,\ell} h(\xi_t^{N,\ell}) \; .$$

Le calcul récursif souhaité est alors obtenu en remplaçant dans (2.17), la loi $\phi^{\chi}_{T-1,\theta}$ par son approximation particulaire $\phi^{N,ps}_{T-1,\theta}$. Nous définissons ainsi, à chaque instant $t \in \{0, \cdots, T\}$, une approximation $\widehat{\rho}_{t,\theta}(\xi_t^{N,\ell})$ de $\rho^{\chi}_{t,\theta}(\xi_t^{N,\ell})$. Nous obtenons, pour tout $\ell \in \{1, \cdots, N\}$,

$$\widehat{\rho}_{\theta,T}(\xi_T^{N,\ell}) = \sum_{i=1}^{N} \left[\frac{1}{T} S(\xi_{T-1}^{N,i}, \xi_T^{N,\ell}, Y_T) + \left(1 - \frac{1}{T}\right) \widehat{\rho}_{\theta,T-1}(\xi_{T-1}^{N,i})\right] \times \frac{\omega_{T-1}^{N,i} m_\theta(\xi_{T-1}^{N,i}, \xi_T^{N,\ell})}{\sum_{j=1}^{N} \omega_{T-1}^{N,j} m_\theta(\xi_{T-1}^{N,j}, \xi_T^{N,\ell})} \; .$$

Et nous concluons en remplaçant dans (2.15), $\phi^{\chi}_{T,\theta}$ par son approximation particulaire $\phi^{N,ps}_{T,\theta}$ et $\rho^{\chi}_{T,\theta}(\xi_T^{N,\ell})$ par $\widehat{\rho}_{T,\theta}(\xi_T^{N,\ell})$, pour $\ell \in \{1, \cdots, N\}$.

Chapitre 3

Contrôle de l'approximation particulaire pour le lissage de fonctionnelles additives (préambule)

Dans ce chapitre, nous nous intéressons au contrôle de l'erreur effectuée lorsque, dans un HMM, l'espérance conditionnelle d'une fonctionnelle additive dépendant des états cachés $X_{0:T}$, sachant les observations $Y_{0:T}$, est remplacée par une approximation particulaire. Nous proposons un contrôle de la norme L_p de l'erreur ainsi que des inégalités de déviation exponentielles. Les bornes données sont valables lorsque l'approximation particulaire est calculée à l'aide de différents algorithmes proposés dans la littérature. Les contrôles mettent en avant la dépendance explicite des bornes en fonction du nombre d'observations et du nombre de particules utilisées.

Dans ce chapitre, nous proposons différents types de contrôles pour l'erreur commise lorsque l'espérance conditionnelle d'une fonctionnelle additive est approchée à l'aide de méthodes de Monte Carlo séquentielles. L'objectif est d'obtenir une dépendance explicite en fonction du nombre de particules et du nombre d'observations utilisées.

De tels contrôles ont tout d'abord un intérêt pratique : ils permettent d'obtenir un critère pour choisir le nombre de particules à utiliser en fonction du nombre d'observations disponibles pour obtenir une précision souhaitée. De plus, bien que ne fournissant que des bornes supérieures, ils constituent

un outil de comparaison entre les différents algorithmes qui peuvent être utilisés pour réaliser les approximations recherchées. D'autre part, ces inégalités peuvent servir à démontrer des théorèmes limites lorsque le nombre de particules croît vers $+\infty$ (en choisissant par exemple un nombre de particules fonction du nombre d'observations). Enfin, ce type de majorations est indispensable pour démontrer la consistance de certains algorithmes d'estimation dans les HMM. Le cadre des fonctionnelles additives est particulièrement adapté à des algorithmes de type EM (voir les chapitres 2 et 6) ou de type gradient (utilisant une approximation du score).

Nous présentons en Section 3.1 le cadre général des méthodes de Monte Carlo séquentielles ainsi que les algorithmes pour lesquels nos bornes sont établies. La Section 3.2 contient notre contribution au contrôle de l'erreur d'approximation liée à ces algorithmes : nous proposons de nouvelles bornes de la norme L_p de l'erreur lorsque la quantité $\phi^{\chi}_{0:T|T}[h]$, définie par (1.1), est approchée par son approximation particulaire lorsque h est une fonctionnelle additive. Nous exposons également des inégalités de déviation exponentielles. Les résultats de la Section 3.2 ont fait l'objet de l'article de revue [Dubarry et Le Corff, 2011], sujet du Chapitre 7.

3.1 Algorithmes considérés

Nous rappelons tout d'abord le contexte dans lequel les algorithmes utilisés sont mis en oeuvre. Nous considérons une chaîne de Markov cachée $\{(X_k, Y_k)\}_{k\geq 0}$ de loi initiale χ sur $(\mathbb{X}, \mathcal{X})$ et dont les noyaux ont pour densités m et g, voir la Section 1.2. Dans la suite de ce chapitre, nous notons, pour tout $x_{0:T} \in \mathbb{X}^{T+1}$,

$$S_T(x_{0:T}) \stackrel{\text{def}}{=} \sum_{t=1}^{T} h_t(x_{t-1}, x_t) \;, \tag{3.1}$$

où les fonctions $\{h_t\}_{t=1}^T$ sont bornées sur $\mathbb{X}^2$ et à valeurs dans $\mathbb{R}^d$. Par la définition de la loi conditionnelle jointe donnée par (1.1), nous avons alors

$$\phi^{\chi}_{0:T|T}[S_T] = \mathbb{E}^{\chi}\left[S_T(X_{0:T})|Y_{0:T}\right] \;.$$

Dans la suite de ce chapitre, nous omettons la dépendance en χ, de manière à utiliser des notations similaires à celles du Chapitre 7. Nous travaillons également avec un jeu de données $Y_{0:T}$ fixe.

Cette section présente différents algorithmes pour approcher $\phi_{0:T|T}[S_T]$ à l'aide de nuages de particules associées à des poids. On présente en Section 3.1.1 un algorithme séquentiel permettant de produire des approximations des lois de filtrage $\{\phi_t\}_{t=0}^T$ et des lois de lissage $\{\phi_{u:v|s:t}\}$ pour

$0 \leq s \leq u \leq v \leq t \leq T$ dont les définitions sont données par (1.1). Cet algorithme est utilisé comme étape intermédiaire en Section 3.1.2 pour présenter les algorithmes pour lesquels les bornes données sont valables. La mise en place de ces algorithmes nécessite l'introduction de quelques quantités :

i) Une mesure ρ sur $(\mathbb{X}, \mathcal{X})$, telle que χ soit absolument continue par rapport à ρ. ρ permet de simuler les premières particules à l'instant 0.

ii) Une famille de fonctions $\{\vartheta_t\}_{t=1}^T$ définies sur $\mathbb{X}$ et à valeurs dans $\mathbb{R}_+^\star$. Les fonctions $\{\vartheta_t\}_{t=1}^T$ permettent d'ajuster les poids avec lesquels une particule de l'instant $t-1$ est choisie pour simuler une nouvelle particule à l'instant t.

iii) Une famille de noyaux de Markov $\{P_t\}_{t=1}^T$ sur $\mathbb{X} \times \mathcal{X}$ tels que, pour tout $x \in \mathbb{X}$, $P_t(x, \cdot)$ admette une densité $p_t(x, \cdot)$ par rapport à la mesure de référence λ, voir la Section 1.2. Le noyau P_t permet de simuler les particules de l'instant t.

3.1.1 Le path-space smoother

L'algorithme présenté dans cette section a été conçu, à l'origine, pour approcher les lois de filtrage $\{\phi_t\}_{t=0}^T$ puis a été modifié pour approcher les lois de lissage. Bien que nous nous intéressions au problème du lissage, les algorithmes de la Section 3.1.2 utilisent la version initiale de l'algorithme visant à approcher les lois de filtrage. Nous commençons donc par présenter le fonctionnement des méthodes de Monte Carlo séquentielles les plus répandues pour l'approximation des lois de filtrage $\{\phi_t\}_{t=0}^T$. Ces méthodes, introduites par [Gordon *et al.*, 1993] et [Kitagawa, 1996], permettent d'approcher chaque loi ϕ_t, pour $t \in \{0, \cdots, T\}$, par un ensemble de points $\{\xi_t^{N,\ell}\}_{\ell=1}^N$, appelés particules, associés à des poids $\{\omega_t^{N,\ell}\}_{\ell=1}^N$. A l'instant 0, on simule N particules indépendantes $\{\xi_0^{N,\ell}\}_{\ell=1}^N$ de même loi ρ et on leur associe les poids non normalisés $\{\omega_0^{N,\ell}\}_{\ell=1}^N$:

$$\omega_0^{N,\ell} \stackrel{\text{def}}{=} \frac{\mathrm{d}\chi}{\mathrm{d}\rho}(\xi_0^{N,\ell}) g(\xi_0^{N,\ell}, Y_0) \, .$$

On pose également

$$\Omega_0^N \stackrel{\text{def}}{=} \sum_{\ell=1}^N \omega_0^{N,\ell} \, .$$

Pour toute fonction mesurable h définie sur $\mathbb{X}$, l'espérance $\phi_0[h]$ est alors approchée par

$$\phi_0^{N,ps}[h] \stackrel{\text{def}}{=} \frac{1}{\Omega_0^N} \sum_{\ell=1}^N \omega_0^{N,\ell} h(\xi_0^{N,\ell}) \, .$$

Les particules simulées sont alors sélectionnées et propagées de façon à obtenir une approximation des lois de filtrage $\{\phi_t\}_{t=1}^T$. Nous présentons ici une

version du filtre auxiliaire, algorithme proposé par [Pitt et Shephard, 1999] et qui généralise les méthodes de [Gordon *et al.*, 1993] et de [Kitagawa, 1996]. Supposons alors que nous disposions de $\{\xi_t^{N,\ell}\}_{\ell=1}^N$ et $\{\omega_t^{N,\ell}\}_{\ell=1}^N$ permettant d'approcher la loi ϕ_t. Les nouvelles particules pondérées sont obtenues en simulant tout d'abord des paires $\{I_{t+1}^{N,\ell}, \xi_{t+1}^{N,\ell}\}_{\ell=1}^N$: pour tout $\ell \in \{1, \cdots, N\}$, $I_{t+1}^{N,\ell}$ représente l'indice de la particule à l'instant t permettant de simuler la particule $\xi_{t+1}^{N,\ell}$. Le couple $\{I_{t+1}^{N,\ell}, \xi_{t+1}^{N,\ell}\}_{\ell=1}^N$ est simulé suivant la loi instrumentale

$$\pi_{t+1|t+1}(i,h) \propto \omega_t^{N,i} \vartheta_{t+1}(\xi_t^{N,i}) P_{t+1}(\xi_t^{N,i}, h) \ ,$$

où les $\{\vartheta(\xi_{t+1}^{N,i})\}_{i=1}^N$ sont appelés poids d'ajustement. Le choix $\vartheta_{t+1}(\cdot) = 1$ et $P_{t+1}(x,\cdot) = M(x,\cdot)$ conduit au filtre proposé par [Gordon *et al.*, 1993] et [Kitagawa, 1996]. On définit alors le poids non normalisé associé à cette particule par

$$\omega_{t+1}^{N,\ell} \stackrel{\text{def}}{=} \frac{m(\xi_t^{N,I_{t+1}^{N,\ell}}, \xi_{t+1}^{N,\ell}) g(\xi_{t+1}^{N,\ell}, Y_{t+1})}{\vartheta_{t+1}(\xi_t^{N,\ell}) p_{t+1}(\xi_t^{N,I_{t+1}^{N,\ell}}, \xi_{t+1}^{N,\ell})} \ .$$

On introduit également

$$\Omega_{t+1}^N \stackrel{\text{def}}{=} \sum_{\ell=1}^N \omega_t^{N,\ell} \ .$$

Pour toute fonction mesurable h définie sur $\mathbb{X}$, l'espérance $\phi_{t+1}[h]$ est alors approchée par

$$\phi_{t+1}^{N,ps}[h] \stackrel{\text{def}}{=} \frac{1}{\Omega_{t+1}^N} \sum_{\ell=1}^N \omega_{t+1}^{N,\ell} h(\xi_{t+1}^{N,\ell}) \ .$$

L'algorithme présenté propose donc une méthode séquentielle pour approcher les lois de filtrage mais peut aussi être utilisé pour approcher les lois de lissage, voir [Kitagawa, 1996] ou [Del Moral, 2004, Section 3.4]. Si à chaque pas de temps nous écrivons

$$\xi_{0:t+1}^{N,\ell} \stackrel{\text{def}}{=} \left(\xi_{0:t}^{N,I_{t+1}^{N,\ell}}, \xi_{t+1}^{N,\ell} \right) \ ,$$

alors on peut produire une famille de N trajectoires pondérées qui permettent d'approcher l'espérance $\phi_{0:T|T}[h]$ par

$$\phi_{0:T|T}^{N,ps}[h] \stackrel{\text{def}}{=} \frac{1}{\Omega_T^N} \sum_{\ell=1}^N \omega_T^{N,\ell} h(\xi_{0:T}^{N,\ell}) \ ,$$

pour toute fonction mesurable h définie sur $\mathbb{X}^{T+1}$.

L'approximation particulaire ainsi obtenue est très peu efficace pour approcher les lois de lissage jointes. En effet, le *path-space smoother* souffre de la dégénérescence des trajectoires, voir par exemple [Kitagawa, 1996],

[Kitagawa et Sato, 2001] et [Fearnhead *et al.*, 2010]. A chaque instant, une nouvelle trajectoire est définie en adjoignant à la nouvelle particule simulée $\xi_{t+1}^{N,\ell}$ une trajectoire ancestrale $\xi_{0:t}^{N,I_{t+1}^{N,\ell}}$ choisie aléatoirement parmi les N trajectoires déjà simulées. Si une trajectoire ancestrale n'est pas sélectionnée, elle disparaît et n'est pas utilisée pour l'approximation particulaire. Ce processus étant répété pour chaque nouvelle observation, pour de grandes valeurs de T, les trajectoires simulées possèdent les mêmes ancêtres. Ce phénomène est à l'origine de la forte variance de l'estimation $\phi_{0:T|T}^{N,ps}[h]$ fournie par le *path-space smoother* (voir par exemple [Doucet *et al.*, 2011]). Une illustration de la dégénérescence est donnée par la Figure 3.1.

3.1.2 Les algorithmes FFBS et FFBSi

Les remarques de la section précédente nous orientent donc vers le choix d'autres algorithmes permettant un meilleur contrôle de l'erreur liée à l'approximation Monte Carlo. L'article présenté au Chapitre 7 se concentre sur deux algorithmes de lissage particulaire, le Forward Filtering Backward Smoothing algorithm (FFBS) et le Forward Filtering Backward Simulation algorithm (FFBSi). Pour ces deux algorithmes, on suppose dans un premier temps que l'on a réalisé un *path-space smoother* permettant d'obtenir une famille de particules pondérées $\{(\xi_t^{N,\ell}, \omega_t^{N,\ell}), 1 \leq \ell \leq N\}_{t=0}^{T}$ utilisées pour approcher les lois $\{\phi_t\}_{t=0}^{T}$.

L'algorithme FFBS

L'algorithme FFBS, introduit par [Doucet *et al.*, 2000] (voir également [Briers *et al.*, 2010]), associe de nouveaux poids aux particules déjà simulées. Toutes les particules simulées par le *path-space smoother* sont conservées, l'algorithme leur associe de nouveaux poids pour approcher les lois de lissage. Le fait de ne pas supprimer de nombreuses particules permet d'éviter le problème de dégénérescence évoqué plus haut. Comme on le verra au Chapitre 7, l'algorithme FFBS peut s'appliquer à des fonctionnelles plus complexes que celle définie par (3.1), par exemple des fonctionnelles de la forme $\sum_{t=r}^{T} h_t(X_{t-r:t})$. Cependant, la complexité de l'algorithme croît en $O(N^{r+1})$ et ne permet donc pas d'utiliser le FFBS de façon raisonnable pour de grandes valeurs de r. On présente donc dans cette section l'algorithme FFBS appliqué au calcul de l'approximation de $\phi_{0:T|T}[S_T]$, où S_T est une fonctionnelle additive de la forme (3.1) (c'est-à-dire dans le cas $r = 1$).

L'algorithme FFBS repose sur le calcul suivant :

$$\begin{aligned}\mathbb{E}\left[h_t(X_{t-1},X_t)|Y_{0:T}\right] &= \mathbb{E}\left[\mathbb{E}\left[h_t(X_{t-1},X_t)|X_t,Y_{0:T}\right]|Y_{0:T}\right] \ , \\ &= \mathbb{E}\left[\frac{\int \phi_{t-1}(\mathrm{d}x_{t-1})m(x_{t-1},X_t)h_t(x_{t-1},X_t)}{\int \phi_{t-1}(\mathrm{d}x_{t-1})m(x_{t-1},X_t)}\middle| Y_{0:T}\right] \ ,\end{aligned} \tag{3.2}$$

où l'on a utilisé le fait que, dans un HMM, la loi de X_{t-1} conditionnellement à X_t et $Y_{0:T}$ est donnée par

$$B_{\phi_{t-1}}(X_t,\mathrm{d}x_{t-1}) \stackrel{\text{def}}{=} \frac{m(x_{t-1},X_t)\phi_{t-1}(\mathrm{d}x_{t-1})}{\int \phi_{t-1}(\mathrm{d}x_{t-1})m(x_{t-1},X_t)} \ .$$

L'approximation souhaitée est alors obtenue de la façon suivante :

i) On remplace dans (3.2) ϕ_{t-1} par son approximation $\phi_{t-1}^{N,ps}$ fournie par le *path-space smoother*.

ii) On construit de façon récursive, en commençant en $t=T$, une approximation de la loi de lissage marginale $\phi_{t|T}$, voir la Section 1.2, utilisant les particules $\{\xi_t^{N,\ell}\}_{\ell=1}^N$ associées à de nouveaux poids $\{\omega_{t|T}^{N,\ell}\}_{\ell=1}^N$.

L'approximation donnée par l'algorithme FFBS s'écrit alors :

$$\phi_{0:T|T}^N[S_T] \stackrel{\text{def}}{=} \sum_{t=0}^{T}\sum_{i_{t-1},i_t=0}^{N} \omega_{t-1:t|T}^{N,i_{t-1:t}}\, h_t(\xi_{t-1}^{N,i_{t-1}},\xi_t^{N,i_t}) \ , \tag{3.3}$$

où les poids $\omega_{t-1:t}^{N,i_{t-1:t}}$ sont calculés de la façon suivante

i) A l'instant T,

$$\omega_{T-1:T|T}^{N,i_{T-1:T}} \stackrel{\text{def}}{=} \frac{\omega_{T-1}^{N,i_{T-1}} m(\xi_{T-1}^{N,i_{T-1}},\xi_T^{N,i_T})}{\sum_{\ell=1}^N \omega_{T-1}^{N,\ell} m(\xi_{T-1}^{N,\ell},\xi_T^{N,i_T})}\frac{\omega_T^{N,i_T}}{\Omega_T} \ .$$

ii) Pour tout $t\in\{T,\cdots,2\}$,

$$\begin{aligned}\omega_{t-1|T}^{N,i_{t-1}} &\stackrel{\text{def}}{=} \sum_{i_t=1}^{N}\omega_{t-1:t|T}^{N,i_{t-1:t}} \ , \\ \omega_{t-2:t-1|T}^{N,i_{t-2:t-1}} &\stackrel{\text{def}}{=} \frac{\omega_{t-2}^{N,i_{t-2}} m(\xi_{t-2}^{N,i_{t-2}},\xi_{t-1}^{N,i_{t-1}})}{\sum_{\ell=1}^N \omega_{t-2}^{N,\ell} m(\xi_{t-2}^{N,\ell},\xi_{t-1}^{N,i_{t-1}})}\omega_{t-1|T}^{N,i_{t-1}} \ .\end{aligned} \tag{3.4}$$

L'algorithme FFBSi

L'algorithme FFBSi, introduit par [Godsill *et al.*, 2004], simule de façon indépendante des trajectoires approximativement distribuées sous la loi de lissage $\phi_{0:T|T}$. Les trajectoires simulées sont construites en utilisant les particules produites par le *path-space smoother*. Contrairement à l'algorithme

FFBS, sa complexité ne dépend pas des fonctions h que nous voulons intégrer sous $\phi_{0:T|T}$. Pour chaque trajectoire, on commence par simuler un indice J_T^ℓ sur $\{1, \cdots, N\}$ avec des probabilités proportionnelles à $\{\omega_T^{N,\ell}\}_{\ell=1}^N$. Ensuite, conditionnellement aux indices $J_{t:T}^\ell$, pour $t \in \{1, \cdots, T\}$, on simule alors un indice J_{t-1}^ℓ sur $\{1, \cdots, N\}$ avec probabilités $\{\Lambda_{t-1}^N(J_t^\ell, k)\}_{k=1}^N$, où

$$\Lambda_{t-1}^N(i_t, i_{t-1}) \stackrel{\text{def}}{=} \frac{\omega_{t-1}^{N,i_{t-1}} m(\xi_{t-1}^{N,i_{t-1}}, \xi_t^{N,i_t})}{\sum_{\ell=1}^N \omega_{t-1}^{N,\ell} m(\xi_{t-1}^{N,\ell}, \xi_t^{N,i_t})} \, .$$

La loi de proposition ainsi définie provient de l'expression des poids de lissage (3.4). Lorsque la particule ξ_t^{N,J_t^ℓ} est choisie, nous associons les poids $\Lambda_{t-1}^N(J_t^\ell, i_{t-1})$ à tous les couples que l'on peut former avec les particules de l'instant précédent avant d'en choisir une aléatoirement suivant ces poids. Le processus de simulation de trajectoire est alors répété pour obtenir N (par commodité) trajectoires approximativement simulées sous la loi cible $\phi_{0:T|T}$ et le FFBSi fournit l'approximation

$$\widetilde{\phi}_{0:T|T}^N[h] \stackrel{\text{def}}{=} \frac{1}{N} \sum_{\ell=1}^N h(\xi_0^{N,J_0^\ell}, \cdots, \xi_T^{N,J_T^\ell}) \, , \tag{3.5}$$

pour h une fonction mesurable sur $\mathbb{X}^{T+1}$.

3.2 Contribution

Nous présentons dans cette section les résultats obtenus et faisant l'objet de l'article de revue inséré au Chapitre 7. Notre contribution prend deux formes différentes : en Section 3.2.1, nous nous intéressons au contrôle L_p de l'erreur obtenue en remplaçant $\phi_{0:T|T}$ par son approximation particulaire fournie par les algorithmes FFBS et FFBSi. En Section 3.2.2 nous donnons des inégalités de déviation exponentielles pour ces mêmes erreurs.

3.2.1 Contrôle de l'erreur L_p

Enoncé du résultat

Le premier résultat, donné par le Théorème 7.1 au Chapitre 7, fournit un contrôle de la norme L_p de l'erreur de l'approximation de l'espérance conditionnelle d'une fonctionnelle additive dans les HMM. Le résultat en question s'applique à des fonctions de la forme $\sum_{t=r}^T h_t(X_{t-r:t})$. Ceci nous permet d'obtenir des bornes pour des fonctionnelles additives de la forme (3.1) (en choisissant $r = 1$) aussi bien que pour des fonctions plus générales

dépendant de toute la trajectoire (en choisissant $r = T$). Nous ne présentons ici que le cas $r = 1$ par souci de clarté - voir les chapitres 5 et 6 pour des applications à des algorithmes de type EM. Dans le cas où S_T est de la forme (3.1), ce contrôle s'écrit, pour $p \geq 2$,

$$\left\| \phi^N_{0:T|T}\left[S_T\right] - \phi_{0:T|T}\left[S_T\right] \right\|_p \leq \frac{C}{\sqrt{N}} \left(1 + \sqrt{\frac{T}{N}}\right) \left(\sum_{t=1}^{T} \mathrm{osc}(h_t)^2 \right)^{1/2} , \quad (3.6)$$

où l'approximation $\phi^N_{0:T|T}\left[S_T\right]$ est donnée par (3.3) et où C ne dépend ni de N ni de T. On rappelle ici que nous supposons dans ce chapitre que les observations sont fixées. Dans ce cas, pour toute variable aléatoire Z, $\|Z\|_p \stackrel{\text{def}}{=} \mathbb{E}\left[|Z|^p\right]^{1/p}$ où l'espérance ne porte que sur les variables aléatoires simulées lors de l'étape d'approximation particulaire. Ce résultat est valable pour des fonctions $\{h_t\}_{t=1}^T$ bornées et sous certaines hypothèses données au Chapitre 7. Nous supposons en particulier la condition de mélange fort :

$$\exists(\sigma_-, \sigma_+) \in \mathbb{R}_+^2 \ , \ \forall (x, x') \in \mathbb{X}^2 \ , \ 0 < \sigma_- \leq m(x, x') \leq \sigma_+ < +\infty \ . \quad (3.7)$$

L'intérêt de cette inégalité provient de la dépendance en T/N (où T est le nombre d'observations et N le nombre de particules) de la borne supérieure. Contrairement aux résultats connus que nous présentons plus bas, la borne dépend de T et de N de façon homogène : l'erreur L_p peut être contrôlée en choisissant un nombre de particules de l'ordre de T.

La démarche adoptée pour obtenir ce résultat est donnée en Section 7.3. Elle repose sur l'oubli géométrique *forward* et *backward* de la chaîne de Markov, conséquence de l'hypothèse (3.7). La borne L_p découle ensuite d'une nouvelle décomposition de l'erreur : cette décomposition fait apparaître un premier terme martingale dont la norme L_p est contrôlée par des outils classiques (voir l'inégalité de Burkholder, [Hall et Heyde, 1980, Théorème 2.10, page 23]). L'autre terme, plus complexe, est traité en utilisant une décomposition plus fine (voir la Proposition 7.2).

Les résultats donnés au Chapitre 7 contiennent également une inégalité du même type pour l'approximation donnée par l'algorithme FFBSi définie par (3.5) (la dépendance en T et en N est la même que pour le FFBS, seules les constantes changent). Ce résultat provient du contrôle de l'erreur entre les algorithmes FFBS et FFBSi qui repose à nouveau sur les propriétés d'oubli géométrique *backward* de la chaîne de Markov.

Comparaison avec les résultats connus

Le résultat donné par le Théorème 7.1 peut être comparé avec certains contrôles proposés dans la littérature. Bien qu'il ne soit qu'une borne supérieure, il donne un critère de choix parmi plusieurs algorithmes d'approximation particulaire.

Par exemple, lorsqu'il est appliqué à des fonctions de la formes (3.1) et sous des hypothèses similaires aux notres (en particulier sous l'hypothèse (3.7)), [Del Moral et Doucet, 2003, Théorème 4] prouve que l'erreur L_p du *path-space smoother* vérifie

$$\left\| \phi^{N,ps}_{0:T|0:T}[S_T] - \phi_{0:T|0:T}[S_T] \right\|_p = O\left(\frac{T^2}{\sqrt{N}} \right) . \tag{3.8}$$

Dans ce cadre, la dépendance de la borne est explicite en T et N mais, pour un contrôle similaire à celui donné par (3.6) utilisé avec un nombre de particules de l'ordre de T, N doit être choisi dans (3.8) de l'ordre de T^4.

[Olsson *et al.*, 2008] propose une alternative dédiée aux approximations des quantités $\phi_{0:T|T}[S_T]$ lorsque S_T est une fonctionnelle additive de la forme (3.1). Puisque nous avons

$$\phi_{0:T|T}[S_T] = \sum_{t=1}^{T} \phi_{0:T|T}[h_t] ,$$

il s'agit donc d'approcher T espérances conditionnelles se rapportant aux fonctions h_t, chacune ne dépendant que des deux états X_{t-1} et X_t. L'idée est alors de

i) remplacer chaque espérance $\phi_{0:T|T}[h_t]$ par $\phi_{0:\kappa(t)|\kappa(t)}[h_t]$, où $\kappa(t) \stackrel{\text{def}}{=} T \wedge (t + \Delta_T)$ et où $\Delta_T > 0$ (voir aussi [Kitagawa et Sato, 2001]),

ii) choisir une approximation des espérances $\phi_{0:\kappa(t)|\kappa(t)}[h_t]$.

[Olsson *et al.*, 2008] applique simplement le *path-space smoother* pour calculer une valeur approchée des espérances $\phi_{0:\kappa(t)|\kappa(t)}[h_t]$. L'estimateur proposé est alors donné par

$$\phi^{N,lag}_{0:T|T}[S_T] \stackrel{\text{def}}{=} \sum_{t=1}^{T} \phi^{N,ps}_{0:\kappa(t)|\kappa(t)}[h_t] .$$

De cette façon, même pour les grandes valeurs de T, le calcul des espérances $\phi^{N,ps}_{0:\kappa(t)|\kappa(t)}[h_t]$, pour $t \ll T$, ne souffre plus du problème de dégénérescence, puisque le calcul se fait à l'aide des trajectoires particulaires pour lesquelles le rééchantillonnage s'arrête après l'instant $\kappa(t)$. Dans le cas où $\Delta_T = O(\log T)$ et avec l'hypothèse (3.7), [Olsson *et al.*, 2008, Théorème 3.3] prouve alors que l'erreur L_p vérifie

$$\left\| \phi^{N,lag}_{0:T|T}[S_T] - \phi_{0:T|T}[S_T] \right\|_p = O\left(\frac{T \log T}{\sqrt{N}} \right) . \tag{3.9}$$

Cette borne est meilleure que celle du *path-space smoother* de l'équation (3.8) mais fournit toujours un contrôle pour lequel l'exposant de T est plus élevé que celui de N.

La borne (3.6) nous permet également d'améliorer des inégalités L_p déjà existantes pour les algorithmes FFBS et FFBSi. En effet, [Douc *et al.*, 2011a, Théorème 5 et Corollaire 6] établissent une majoration de l'erreur L_p de la forme

$$\left\| \phi^N_{0:T|T}[h] - \phi_{0:T|T}[h] \right\|_p \leq C(p,T) \frac{\mathrm{osc}(h)}{\sqrt{N}} \ ,$$

pour n'importe quelle fonction bornée h sur $\mathbb{X}^{T+1}$, où $\phi^N_{0:T|T}[h]$ est l'approximation fournie par l'algorithme FFBS. [Douc *et al.*, 2011a] donne la même inégalité pour l'algorithme FFBSi. Ces résultats sont obtenus sans supposer la condition de mélange fort (3.7) mais, la valeur de C étant dépendante de T, la borne n'est pas directement exploitable. Si (3.7) est ajoutée, [Douc *et al.*, 2011a, Théorème 11] donne des bornes uniformes en T uniquement dans le cas des lois marginales (i.e. lorsque la fonction h ne dépend que d'un état, voir la Section 1.2), mais pas de nouveaux résultats pour des fonctions h dépendant de toute la trajectoire.

D'autre part, [Del Moral *et al.*, 2010a] fournit, sous (3.7) un contrôle de la forme

$$\left\| \phi^N_{0:T|T}[h] - \phi_{0:T|T}[h] \right\|_p \leq \frac{C(p,T)}{\sqrt{N}} \ , \tag{3.10}$$

où $\phi^N_{0:T|T}[h]$ est l'approximation fournie par l'algorithme FFBS. $C(p,T)$ est un $O(T)$ dans certains cas :

i) lorsque la fonction h est une fonctionnelle additive de la forme (3.1),

ii) lorsque la fonction h est une fonction bornée sur $\mathbb{X}^{T+1}$.

(3.10) fournit donc toujours un contrôle plus faible en T que la borne (3.6).

3.2.2 Inégalités de déviation exponentielles

Enoncé du résultat

La seconde contribution du Chapitre 7, donnée par le Théorème 7.2, concerne des inégalités de déviation exponentielles. Ces inégalités établissent que la probabilité d'obtenir une erreur supérieure à un $\epsilon > 0$ fixé, lorsque $\phi_{0:T|T}[S_T]$ est remplacé par son approximation Monte Carlo, décroît exponentiellement vite avec ϵ. Ces résultats sont toujours établis sous la condition de mélange fort (3.7) et sont énoncés de façon générale pour des fonctions de la forme $\sum_{t=r}^{T} h_t(X_{t-r:t})$, où les fonctions $\{h_t\}_{t=r}^{T}$ sont bornées sur $\mathbb{X}^{r+1}$. Dans le cas d'une fonctionnelle additive S_T de la forme (3.1) (r=1), nous avons, pour l'algorithme FFBS,

$$\mathbb{P}\left\{ \left| \phi_{0:T|T}[S_T] - \phi^N_{0:T|T}[S_T] \right| > \varepsilon \right\} \leq 2\exp\left(-\frac{CN\varepsilon^2}{\sum_{s=1}^{T} \mathrm{osc}(h_s)^2} \right) + 8\exp\left(-\frac{CN\varepsilon}{\sum_{s=1}^{T} \mathrm{osc}(h_s)} \right) \ .$$

Notre résultat permet également d'avoir une inégalité exponentielle pour une fonction h bornée sur $\mathbb{X}^{T+1}$:

$$\mathbb{P}\left\{\left|\phi_{0:T|T}\left[h\right]-\phi_{0:T|T}^{N}\left[h\right]\right|>\varepsilon\right\} \leq 2\exp\left(-\frac{CN\varepsilon^{2}}{(T+1)\mathrm{osc}(h)^{2}}\right)+8\exp\left(-\frac{CN\varepsilon}{(T+1)\mathrm{osc}(h)}\right) . \quad (3.11)$$

Nous obtenons les mêmes bornes pour l'approximation donnée par l'algorithme FFBSi (seules les constantes changent, la dépendance en T et en N est la même).

Comparaison avec les résultats connus

Notre borne précise les résultats connus dans la littérature à propos des inégalités de déviation exponentielles pour les algorithmes FFBS et FFBSi. En effet, [Douc *et al.*, 2011a, Théorème 5 et Corollaire 6] fournissent des inégalités de déviation exponentielles pour l'erreur commise lorsque la loi $\phi_{0:T|T}$ est remplacée par les approximations obtenues à l'aide du FFBS et du FFBSi (et d'une version modifiée du FFBSi). Pour tout $\epsilon > 0$, ils obtiennent, pour h bornée sur $\mathbb{X}^{T+1}$,

$$\mathbb{P}\left\{\left|\phi_{0:T|T}^{N}[h]-\phi_{0:T|T}^{N}[h]\right|>\epsilon\right\} \leq B(T)\exp\left(-\frac{C(T)N\epsilon^{2}}{\mathrm{osc}(h)^{2}}\right) .$$

Cependant, dans cette inégalité, les quantités B et C dépendent de T et nous n'avons donc pas de dépendance explicite en fonction du nombre d'observations, comme c'est le cas pour la borne (3.11).

3.2.3 Complément sur le calcul du biais

Tel qu'il est présenté, le Théorème 7.1 permet d'avoir un contrôle de la norme L_p de l'erreur. Bien que cela ne soit pas précisé au Chapitre 7, il est également possible d'obtenir de façon plus directe un contrôle du biais défini, pour l'algorithme FFBS, par

$$\mathbb{E}\left[\phi_{0:T|T}^{N}\left[S_{T}\right]\right]-\phi_{0:T|T}\left[S_{T}\right] .$$

On peut aisément obtenir un contrôle de ce biais en utilisant l'inégalité L_p donnée par (3.6). On a alors

$$\left|\mathbb{E}\left[\phi_{0:T|T}^{N}\left[S_{T}\right]\right]-\phi_{0:T|T}\left[S_{T}\right]\right| \leq \frac{C}{\sqrt{N}}\left(1+\sqrt{\frac{T}{N}}\right)\left(\sum_{t=1}^{T}\mathrm{osc}(h_{t})^{2}\right)^{1/2} .$$

On peut cependant obtenir un contrôle ne faisant intervenir qu'un terme impliquant T et N en utilisant notre décomposition de l'erreur. En effet, si l'on reprend la décomposition donnée par (7.16), on a

$$\phi_{0:T|T}^N [S_T] - \phi_{0:T|T} [S_T] = \sum_{t=0}^{T} D_{t,T}^N (S_T) + \sum_{t=0}^{T} C_{t,T}^N (S_T) \ ,$$

où $D_{t,T}^N$ est défini par (7.14) et $C_{t,T}^N$ par (7.15). La suite $\{D_{t,T}^N (S_T)\}_{t=0}^T$ est un incrément de martingale, on a donc

$$\mathbb{E} \left[\sum_{t=0}^{T} D_{t,T}^N (S_T) \right] = \mathbb{E} \left[D_{0,T}^N (S_T) \right] = 0 \ ,$$

où la seconde égalité provient du mécanisme de production des particules et des poids à l'instant 0. Par la Proposition 7.2 et l'inégalité de Cauchy-Schwarz, nous avons alors

$$\left| \mathbb{E} \left[\phi_{0:T|T}^N [S_T] \right] - \phi_{0:T|T} [S_T] \right| \leq \frac{C}{N} \sum_{t=1}^{T} \mathrm{osc}(h_t) \ .$$

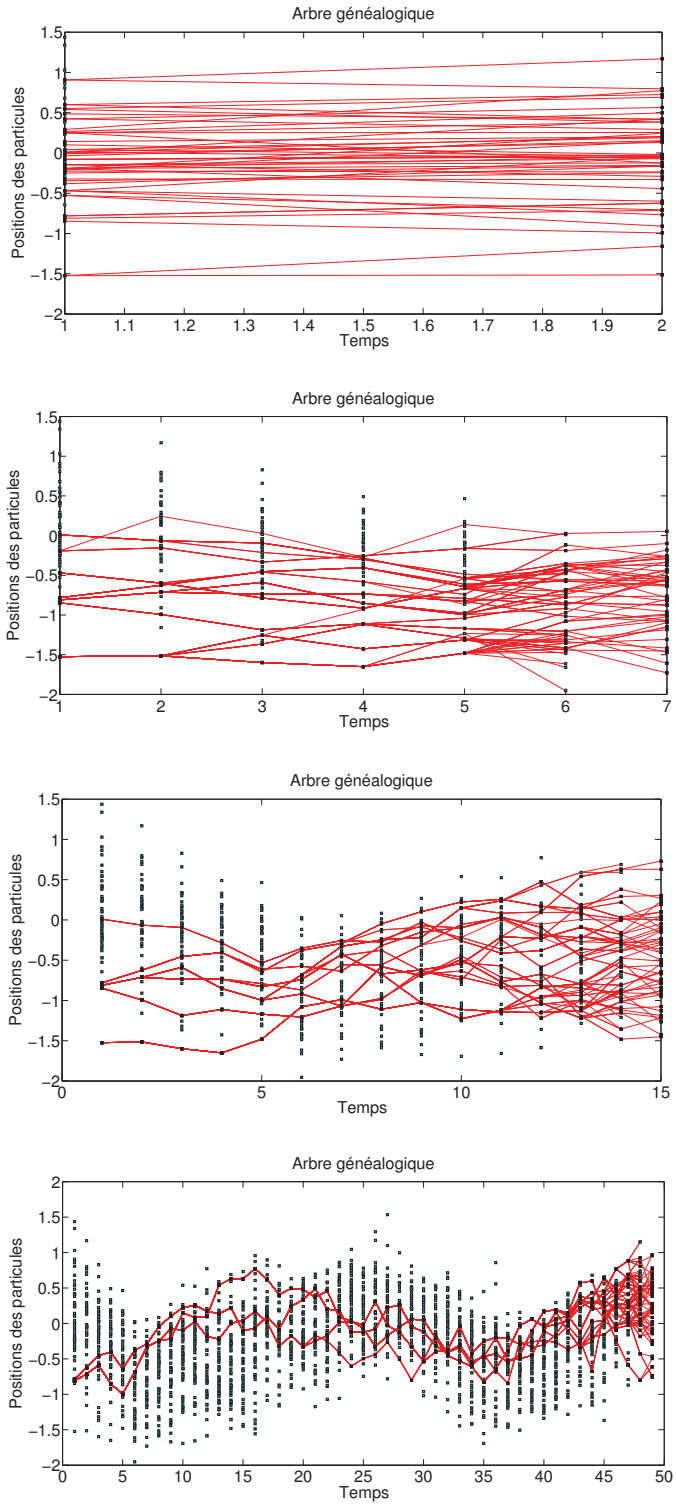


FIGURE 3.1 – Trajectoires particulaires obtenues avec l'algorithme *path-space smoother* dans le cas d'un modèle linéaire gaussien avec $N = 50$. Les trajectoires $\{\xi_{0:t}^{N,\ell}\}_{\ell=1}^{N}$ sont représentées (lignes rouges) ainsi que toutes les particules simulées depuis le départ (points). Les trajectoires ancestrales non sélectionnées sont supprimées au fur et à mesure.

Chapitre 4

Estimation non paramétrique dans les modèles de Markov cachés (préambule)

Dans ce chapitre, nous proposons une méthode d'estimation non paramétrique dans les modèles de Markov cachés. On considère une chaîne de Markov $\{X_k\}_{k\geq 0}$ observée au travers du processus $\{Y_k\}_{k\geq 0}$, où Y_k est une observation bruitée de $f_\star(X_k)$. La chaîne de Markov $\{X_k\}_{k\geq 0}$ est une marche aléatoire restreinte à un sous-espace compact de $\mathbb{R}^m$ dont la loi des incréments est connue à un facteur d'échelle $a_\star$ près. Nous souhaitons estimer $f_\star$ et $a_\star$ à l'aide d'un bloc d'observations. Nous discutons en premier lieu l'identifiabilité de ce modèle sous certaines hypothèses sur la chaîne $\{X_k\}_{k\geq 0}$ et sur la fonction $f_\star$. Nous proposons ensuite une estimation de type maximum de vraisemblance par paires. Nous montrons que la distance de Hellinger entre notre estimation de la loi des paires d'observations et la vraie loi tend vers 0 en probabilité. Ceci nous permet d'obtenir la consistance des estimateurs de $a_\star$ et de $f_\star$.

4.1 Introduction

Dans ce chapitre, nous nous intéressons à un nouveau problème d'estimation dans le cadre des chaînes de Markov cachées. La chaîne de Markov

$\{X_k\}_{k\geq 0}$ est une marche aléatoire stationnaire restreinte à un sous-espace compact de $\mathbb{R}^m$ dont la loi des incréments est connue à un facteur d'échelle $a_\star$ près. Cette chaîne de Markov n'est observée qu'à travers la suite d'observations $\{Y_k\}_{k\geq 0}$. Nous supposons que, pour tout $k \geq 0$, l'observation Y_k vérifie

$$Y_k \stackrel{\text{def}}{=} f_\star(X_k) + \epsilon_k \ , \tag{4.1}$$

où $f_\star$ est une fonction à valeurs dans $\mathbb{R}^\ell$ et où les variables aléatoires $\{\epsilon_k\}_{k\geq 0}$ sont i.i.d et de loi gaussienne connue. L'objectif de ce chapitre est l'estimation de la fonction $f_\star$ et du coefficient $a_\star$ à l'aide d'un bloc d'observations.

Nous insistons ici sur le fait que nous ne sommes plus dans le cadre de l'estimation en ligne du Chapitre 2 : l'estimation de $a_\star$ et $f_\star$ est réalisée à partir de toutes les observations disponibles sans imposer de contraintes sur la façon dont elles sont utilisées.

La difficulté de ce problème réside dans le fait que les points en lesquels la fonction $f_\star$ est évaluée ne sont pas observés et que la loi des $\{X_k\}_{k\geq 0}$ n'est pas connue. Les travaux existants dans la littérature apportent des solutions à ce type de problèmes lorsque les variables $\{X_k\}_{k\geq 0}$ sont observées.

On trouve tout d'abord des réponses dans les modèles à erreurs sur les variables, dans lesquels les régresseurs sont observés en présence de bruit : on dispose d'observations $\{(Y_k, Z_k)\}_{k\geq 0}$ telles que Y_k suit le modèle (4.1) avec un bruit $\{\epsilon_k\}_{k\geq 0}$ non nécessairement gaussien ; et Z_k suit le modèle

$$Z_k \stackrel{\text{def}}{=} X_k + \eta_k \ ,$$

où les erreurs $\{\eta_k\}_{k\geq 0}$ sont i.i.d et de loi connue. Il existe de nombreuses méthodes pour estimer la fonction $f_\star$ dans ce contexte. Un estimateur à noyau a été proposé et étudié par [Fan et Truong, 1993] dans le cas où les variables $\{X_k\}_{k\geq 0}$ sont i.i.d. Ce travail repose sur une estimation de la loi conditionnelle de Y_0 sachant X_0 en résolvant tout d'abord le problème de *déconvolution*, c'est-à-dire l'estimation à noyau de la densité de la loi de X_0 ; puis en proposant une estimation à noyau de la densité de la loi de (X_0, Y_0). La consistance de cet estimateur et les vitesses de convergence ont été étudiées par [Carroll et Hall, 1988], [Carroll et Stefanski, 1990], [Fan, 1991b] et [Fan, 1991a] sous des hypothèses de régularité sur la loi des variables aléatoires $\{X_k\}_{k\geq 0}$ et du bruit $\{\eta_k\}_{k\geq 0}$. Ces travaux ont été complétés sous des hypothèses plus générales par des méthodes de minimisation d'un contraste pénalisé dans [Comte *et al.*, 2006] et dans [Comte et Taupin, 2007]. Tous ces travaux ne peuvent cependant pas être utilisés dans notre situation puisque nous ne disposons pas d'observations de l'état $\{X_k\}_{k\geq 0}$ permettant de résoudre le problème de déconvolution.

Des réponses ont aussi été proposées dans le cas où les variables $\{X_k\}_{k\geq 0}$ forment une chaîne de Markov stationnaire et où le modèle d'observation (4.1) est simplifié en

$$Y_k = X_k + \epsilon_k \ ,$$

où les variables $\{\epsilon_k\}_{k\geq 0}$ sont i.i.d. et de loi connue non nécessairement gaussienne. Contrairement à notre modèle, ici c'est la chaîne $\{X_k\}_{k\geq 0}$ qui est observée et pas une transformation $\{f_\star(X_k)\}_{k\geq 0}$; l'inférence statistique ne porte donc que sur l'estimation de la loi invariante et de la densité du noyau de $\{X_k\}_{k\geq 0}$. [Lacour, 2008b] (resp. [Lacour, 2008a]) propose un estimateur lorsque la chaîne est observée sans erreurs (resp. avec erreurs) : l'estimation repose sur la minimisation d'un contraste pénalisé permettant une minimisation de la norme L_2 de l'erreur entre l'estimateur et les vraies densités. Nous proposons une alternative utilisant une estimation de la loi des paires d'observations.

Nous présentons en Section 4.2 notre contribution à ce problème d'estimation non paramétrique dans les HMM. Nous donnons les estimateurs de $f_\star$ et de $a_\star$ proposés et les théorèmes de convergence que nous avons établis. Les résultats présentés ont fait l'objet de l'article [Dumont et Le Corff, 2012a], soumis à une revue internationale et inséré au Chapitre 8. Nous renvoyons au Chapitre 8 pour les énoncés précis de ces résultats et les démonstrations qui leur sont associées.

4.2 Contribution

Nous introduisons tout d'abord quelques notations pour la définition de l'espace de fonctions auquel appartient notre estimateur de $f_\star$.

Soit K un sous-espace de $\mathbb{R}^m$. Pour toute fonction $f : K \longrightarrow \mathbb{R}^\ell$ et tout $j \in \{1, \cdots, \ell\}$, la j-ème composante de f est notée f_j. Pour $p > 0$, on définit l'espace L^p par

$$\mathrm{L}^p \stackrel{\text{def}}{=} \left\{ f : K \to \mathbb{R}^\ell \; ; \; \|f\|_{\mathrm{L}^p}^p \stackrel{\text{def}}{=} \int_K \|f(x)\|^p \lambda(\mathrm{d}x) < \infty \right\} ,$$

où λ est la mesure de Lebesgue sur $\mathbb{R}^m$ et $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^\ell$. Pour tout m-uplet $\alpha \stackrel{\text{def}}{=} \{\alpha_i\}_{i=1}^m$ d'entiers positifs, nous écrivons $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i$. Soit alors $W^{s,p}$ l'espace de Sobolev sur K de paramètres $s \in \mathbb{N}$ et $p \geq 1$:

$$W^{s,p} \stackrel{\text{def}}{=} \{ f \in \mathrm{L}^p ; \; D^\alpha f \in \mathrm{L}^p, \; \forall\, \alpha \in \mathbb{N}^m, \; |\alpha| \leq s \} \; , \tag{4.2}$$

où $D^\alpha f : K \to \mathbb{R}^\ell$ est le vecteur des dérivées partielles d'ordre α des composantes f_j, pour $j \in \{1, \cdots, \ell\}$. $W^{s,p}$ est muni de la norme $\|\cdot\|_{W^{s,p}}$ définie, pour tout $f \in W^{s,p}$, par

$$\|f\|_{W^{s,p}} \stackrel{\text{def}}{=} \left(\sum_{0 \leq |\alpha| \leq s} \|D^\alpha f\|_{\mathrm{L}^p}^p \right)^{1/p} . \tag{4.3}$$

Enfin, pour établir l'identifiabilité de notre modèle, nous devons choisir une distance sur l'ensemble des densités par rapport à la mesure de Lebesgue μ sur $\mathbb{R}^{2\ell}$. Nous utilisons la distance de Hellinger donnée, pour toutes densités p_1 et p_2 par rapport à μ, par

$$h(p_1,p_2) \stackrel{\text{def}}{=} \left[\frac{1}{2}\int \left(p_1^{1/2}(y)-p_2^{1/2}(y)\right)^2 \mu(\mathrm{d}y)\right]^{1/2} . \tag{4.4}$$

4.2.1 Modèle

Nous considérons une marche aléatoire $\{X_k\}_{k\geq 0}$ à valeurs dans un sous-espace compact K de $\mathbb{R}^m$. Nous supposons que le noyau de transition de cette marche aléatoire a une densité donnée, pour tout $x, x' \in K$, par

$$q_{a_\star}(x,x') \stackrel{\text{def}}{=} C_{a_\star}(x)\, q\left(\frac{\|x'-x\|}{a_\star}\right) , \tag{4.5}$$

par rapport à la mesure de Lebesgue λ sur K. $a_\star$ est un réel strictement positif inconnu et $q : \mathbb{R}_+ \to \mathbb{R}_+$ est une fonction continue et strictement monotone connue. Par suite,

$$C_{a_\star}(x) \stackrel{\text{def}}{=} \left(\int_K q\left(\frac{\|x'-x\|}{a_\star}\right)\lambda(\mathrm{d}x')\right)^{-1} .$$

Nous disposons d'observations $\{Y_k\}_{k\geq 0}$, modélisées par :

$$Y_k \stackrel{\text{def}}{=} f_\star(X_k) + \epsilon_k .$$

Nous nous plaçons dans le cas où $f_\star : K \to \mathbb{R}^\ell$ est inconnue et où les variables $\{\epsilon_k\}_{k\geq 0}$ sont des vecteurs gaussiens indépendants, indépendants des $\{X_k\}_{k\geq 0}$, de moyenne nulle et de matrice de variance $\sigma^2 I_\ell$ connue (I_ℓ est la matrice unité de taille $\ell \times \ell$, $\sigma^2 > 0$).

L'objectif est d'estimer la fonction $f_\star$ et le paramètre d'échelle $a_\star$ à partir d'un bloc d'observations.

L'identifiabilité du modèle proposé ainsi que les propriétés asymptotiques des estimateurs considérés reposent sur certaines hypothèses, dont nous rappelons les principales ici.

i) Nous supposons que $f_\star \in W^{s,p}$, où $W^{s,p}$ est l'espace de Sobolev défini par (4.2). Ceci nous permet de rechercher notre estimateur dans une classe de fonctions contenant $f_\star$. D'autre part, nous imposons une hypothèse reliant l'ordre de l'espace de Sobolev à la dimension de l'espace d'état m : $s > m/p + 1$. Cette hypothèse nous permet d'utiliser un plongement de $W^{s,p}$ dans $\mathcal{C}^1(K,\mathbb{R}^\ell)$, espace des fonctions continûment différentiables définies sur K et à valeurs dans $\mathbb{R}^\ell$. Dans les applications numériques, nous utilisons $p = 2$.

Nous supposons d'autre part que la fonction $f_\star$ est un difféomorphisme : en effet, la preuve de l'identifiabilité du modèle nous amène naturellement à considérer la fonction $\phi \stackrel{\text{def}}{=} (f_\star)^{-1} \circ f$ pour une fonction $f \in \mathcal{C}^1(K, \mathbb{R}^\ell)$. L'hypothèse mentionnée sur $f_\star$ nous assure que ϕ existe et est un élément de $\mathcal{C}^1(K, \mathbb{R}^\ell)$.

ii) L'espace d'état K est supposé compact et convexe. Ceci nous permet, en particulier, d'appliquer un théorème du point fixe à la fonction ϕ.

iii) Nous supposons enfin que $a_\star \in [a_-, +\infty[$ où la borne $a_- > 0$ est connue. Cette hypothèse nous permet d'assurer certaines propriétés de continuité et d'obtenir une inégalité de déviation uniforme pour le processus empirique.

iv) La chaîne $\{X_k\}_{k\geq 0}$ est supposée stationnaire. Pour tout $a \geq a_-$, nous noterons ν_a la densité de la probabilité invariante.

4.2.2 Estimateurs

Introduction

Nous présentons dans cette section les estimateurs de $f_\star$ et de $a_\star$ proposés. Le modèle considéré en Section 4.2.1 implique une structure complexe de la vraisemblance des observations : nous proposons donc une maximisation de la vraisemblance par paires des observations.

A partir de nos estimations de $f_\star$ et de $a_\star$ nous pouvons définir une estimation de la loi des paires d'observations. L'approche de vraisemblance par paires nous permet d'utiliser des inégalités reliant le processus empirique des observations à la distance de Hellinger entre la vraie loi des paires d'observations et notre estimation et d'établir la consistance des estimateurs proposés.

Vraisemblance par paires

Sous nos hypothèses, les variables aléatoires $\{X_k\}_{k\geq 0}$ et $\{Y_k\}_{k\geq 0}$ sont stationnaires. Pour considérer la vraisemblance par paires, il suffit alors de définir la densité de la loi des observations (Y_0, Y_1). Lorsque le modèle présenté en Section 4.2.1 est paramétré par une fonction mesurable $f : K \to \mathbb{R}^\ell$ et par $a \geq a_-$, cette densité est donnée, pour tout y_0, y_1 dans $\mathbb{R}^\ell$, par

$$p_{f,a}(y_0, y_1) \stackrel{\text{def}}{=} \int \varphi(y_0 - f(x_0))\varphi(y_1 - f(x_1))\nu_a(x_0)q_a(x_0, x_1)\lambda(\mathrm{d}x_0)\lambda(\mathrm{d}x_1) \,, \tag{4.6}$$

où φ est la densité de la loi gaussienne centrée et de variance $\sigma^2 I_\ell$. Nous considérons alors la log-vraisemblance basée sur les paires d'observations :

$$(f,a) \mapsto \frac{1}{n}\sum_{k=0}^{n-1} \log p_{f,a}(Y_{2k}, Y_{2k+1}) \; .$$

Cette expression correspond à la vraisemblance des observations $Y_{0:2n-1}$ lorsque les paires d'observations $\{(Y_{2k}, Y_{2k+1})\}_{k=0}^{n-1}$ sont indépendantes et lorsque la chaîne $\{X_k\}_{k\geq 0}$ est stationnaire. Elle permet d'utiliser les inégalités classiques pour les vraisemblances d'observations indépendantes.

Définition des estimateurs

La difficulté est de trouver un espace de fonctions pour la définition du modèle qui permette d'assurer l'identifiabilité tout en étant utilisable de façon numérique. Nous avons choisi d'imposer des propriétés de régularité en considérant l'espace des fonctions de Sobolev $W^{s,p}$ que l'on munit de la norme $\|\cdot\|_{W^{s,p}}$, voir (4.2) et (4.3), et en ajoutant un terme de pénalité à la log-vraisemblance par paires des observations. Ceci nous permet de donner la définition des estimateurs $(\widehat{f}_n, \widehat{a}_n)$ de $(f_\star, a_\star)$ calculés à l'aide de $2n$ observations $\{Y_k\}_{k=0}^{2n-1}$ par :

$$\left(\widehat{f}_n, \widehat{a}_n\right) \stackrel{\text{def}}{=} \operatorname*{argmax}_{f\in W^{s,p}, a\geq a^-} \left\{ \frac{1}{n}\sum_{k=0}^{n-1} \log p_{f,a}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 \|f\|_{W^{s,p}}^{\upsilon+1} \right\} \; ,$$

où υ est une constante strictement positive. Les quantités λ_n^2 et υ doivent également satisfaire quelques contraintes pour établir la consistance de nos estimateurs :

i) la puissance υ de la pénalité vérifie $\upsilon > 2\ell$. Cette hypothèse provient de l'inégalité maximale que nous avons obtenue pour le supremum du processus empirique des observations (voir Proposition 8.3) et permet d'obtenir une inégalité de déviation pour ce même supremum.

ii) λ_n^2 doit vérifier

$$\lambda_n \underset{n\to+\infty}{\longrightarrow} 0 \text{ et } \lambda_n^2 n^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty \; .$$

Ceci nous permet d'assurer la convergence en probabilité, au sens de la distance de Hellinger (voir (4.4)), de $p_{\widehat{f}_n, \widehat{a}_n}$ vers $p_{f_\star, a_\star}$.

4.2.3 Résultats

Identifiabilité

Le premier résultat que nous proposons, qui fait l'objet du Théorème 8.1, concerne l'identifiabilité de notre modèle. Pour $b > a_-$ et f une fonction de

$\mathcal{C}^1(K, \mathbb{R}^\ell)$, il s'agit d'étudier les conséquences de l'égalité des fonctions $p_{f,b}$ et $p_{f_\star,a_\star}$, définies par (4.6). Nous établissons que si $p_{f,b} = p_{f_\star,a_\star}$ alors $b = a_\star$, mais nous n'avons pas nécessairement égalité entre f et $f_\star$: on ne peut retrouver la fonction $f_\star$ qu'à une isométrie de K près.

Pour prouver ce résultat, nous considérons $\{(X'_k, Y'_k)\}_{k\geq 0}$, une chaîne de Markov cachée sur $K \times \mathbb{R}^\ell$ dont le noyau de transition de la chaîne cachée a pour densité q_b et pour loi initiale ν_b ; et une suite $\{\epsilon'_k\}_{k\geq 0}$ de variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2 I_\ell)$, indépendantes de $\{X'_k\}_{k\geq 0}$ telle que pour tout $k \geq 0$,

$$Y'_k \stackrel{\text{def}}{=} f(X'_k) + \epsilon'_k \ .$$

Supposer $p_{f,b} = p_{f_\star,a_\star}$ est donc équivalent à supposer que, pour tout $k \geq 0$, (Y_{2k}, Y_{2k+1}) et (Y'_{2k}, Y'_{2k+1}) ont la même loi. Cette propriété, combinée à la gaussianité de la loi des bruits $\{\epsilon_k\}_{k\geq 0}$ et $\{\epsilon'_k\}_{k\geq 0}$ implique que pour tout $k \geq 0$, (X_{2k}, X_{2k+1}) et $(\phi(X'_{2k}), \phi(X'_{2k+1}))$ ont la même loi, où $\phi \stackrel{\text{def}}{=} (f_\star)^{-1} \circ f$. On prouve alors que la fonction ϕ est une isométrie de K de la façon suivante :

1) on prouve en premier lieu que la fonction ϕ est un difféomorphisme. Cette fonction est $\mathcal{C}^1(K, \mathbb{R}^\ell)$ par les hypothèses sur $f_\star$ et f. Nous montrons donc qu'elle est injective et que son Jacobien est strictement positif en tout point.

   Nous prouvons l'injectivité en montrant, d'une part, que ϕ est un recouvrement de K et, d'autre part, en établissant que tout recouvrement de K à valeurs dans K est injectif. La stricte positivité du Jacobien provient de l'égalité en loi de (X_0, X_1) et de $(\phi(X'_0), \phi(X'_1))$,

2) ceci nous permet ensuite de combiner l'égalité des lois définies par $p_{f_\star,a_\star}$ et $p_{f,b}$ à un changement de variable utilisant la fonction ϕ. La définition des densités $q_{a_\star}$ et q_b et l'application d'un théorème du point fixe permettent de prouver que f est égale à $f_\star$ à isométrie près et que $b = a_\star$.

Consistance

Pour tout entier $n \geq 1$, nous utilisons les estimateurs $\widehat{f}_n$ et $\widehat{a}_n$ pour produire l'estimation $p_{\widehat{f}_n,\widehat{a}_n}$ de la densité $p_{f_\star,a_\star}$. La consistance des estimateurs $\{\widehat{f}_n\}_{n\geq 1}$ et $\{\widehat{a}_n\}_{n\geq 1}$ provient de la consistance au sens de Hellinger de la suite $\{p_{\widehat{f}_n,\widehat{a}_n}\}_{n\geq 1}$.

Le Théorème 8.2 prouve que la suite $\{\lambda_n^{-2} h(p_{\widehat{f}_n,\widehat{a}_n}, p_{f_\star,a_\star})\}_{n\geq 1}$ est bornée en probabilité, où h est la distance de Hellinger définie par (4.4). Ce théorème se démontre en utilisant la même démarche que dans le cas i.i.d. Dans le cas où les observations sont indépendantes, la convergence de l'estimateur du maximum de vraisemblance pénalisé repose sur un contrôle d'un processus empirique associé aux observations (voir [Van De Geer, 2000, Chapitre 10]).

Nous prouvons donc un contrôle du processus empirique donné, pour toute fonction $f \in W^{s,p}$ et tout $a \geq a_-$ par

$$\sqrt{n}\left\{\int \frac{1}{2}\ln\frac{p_{f,a}+p_{f_\star,a_\star}}{2p_{f_\star,a_\star}}\mathrm{d}\mathbb{P}_n \right. \\ \left. - \int \frac{1}{2}\ln\frac{p_{f,a}(y,y')+p_{f_\star,a_\star}(y,y')}{2p_{f_\star,a_\star}(y,y')}\, p_{f_\star,a_\star}(y,y')\mu(\mathrm{d}y)\mu(\mathrm{d}y')\right\} ,$$

où pour tout ensemble mesurable A de $\mathbb{R}^{2\ell}$,

$$\mathbb{P}_n(A) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A(Y_{2k}, Y_{2k+1}) .$$

La preuve du Théorème 8.2 requiert deux contrôles intermédiaires :

i) une inégalité de déviation pour le supremum du processus empirique défini ci-dessus. Cette inégalité est obtenue par une application des résultats de [Adamczak et Bednorz, 2012],

ii) une inégalité maximale pour le supremum du processus empirique défini ci-dessus, obtenue par une application de [Doukhan *et al.*, 1995].

Enfin, la consistance au sens de Hellinger pour la suite $\{p_{\widehat{f}_n,\widehat{a}_n}\}_{n\geq 0}$ et le résultat d'identifiabilité impliquent que la suite $\{\widehat{f}_n\}_{n\geq 1}$ converge en probabilité et dans $\mathcal{C}^1(K,\mathbb{R}^\ell)$ vers l'ensemble des fonctions égales à $f_\star$ à isométrie de K près. Ce résultat, ainsi que la convergence en probabilité de la suite $\{a_n\}_{n\geq 1}$ vers $a_\star$, est l'objet du Théorème 8.3.

Chapitre 5

Applications de l'algorithme BOEM au problème de cartographie et de localisation simultanées

Le problème de localisation et de cartographie simultanées (SLAM, pour Simultaneous Localization And Mapping), recouvre l'ensemble des situations dans lesquelles un mobile (un robot, un humain muni d'un appareil de détection) cherche à déterminer sa position et à construire une carte de son environnement. De nombreuses solutions ont été apportées au problème du SLAM suivant les hypothèses faites sur le modèle de transition du mobile ou sur son modèle d'observation (voir par exemple [Montemerlo *et al.*, 2003], [Burgard *et al.*, 2005] et [Martinez-Cantin, 2008]).

Nous nous intéressons au SLAM avec carte fixe et nous choisissons d'adopter la même formulation que celle proposée par [Martinez-Cantin, 2008]. Il s'agit de considérer le problème du SLAM comme un problème d'estimation dans les modèles de Markov cachés. L'état caché est constitué de l'état du mobile (par exemple sa position dans l'espace et son orientation) et les observations regroupent les informations reçues par le mobile au cours de son déplacement (par exemple distance et position angulaire d'obstacles). La carte est considérée comme un paramètre fixe et l'objectif est alors de fournir simultanément une estimation de l'état du robot et une estimation de la carte en utilisant les observations reçues par le mobile.

Considérer le problème du SLAM comme un problème d'estimation dans les modèles de Markov cachés nous permet d'envisager l'utilisation de l'algorithme Monte Carlo BOEM présenté au Chapitre 2. Cette procédure produit une estimation en ligne de la carte de l'environnement et fournit simul-

tanément une population de particules permettant une estimation de l'état du mobile.

Ce chapitre est organisé de la façon suivante. Nous commençons par étudier deux exemples d'applications de l'algorithme Monte Carlo BOEM à des problèmes de SLAM en Section 5.1 et en Section 5.2. Ces deux exemples sont les premiers cas d'application de l'algorithme Monte Carlo BOEM. La mise en oeuvre de l'algorithme Monte Carlo BOEM présenté au Chapitre 2 requiert l'utilisation d'une approximation Monte Carlo de la quantité intermédiaire de l'algorithme BOEM. Nous proposons ici d'utiliser une approximation particulaire détaillée en Annexe 5.3. Nous présentons en Section 5.1 un problème de SLAM basé sur des *landmarks*. Nous donnons différentes solutions apportées à ce problème du SLAM et nous expliquons en quoi elles diffèrent de celle que nous proposons. La Section 5.2 propose une application à un problème de SLAM indoor basé sur la propagation de signaux WiFi. L'Annexe 5.3 fournit également le résultat permettant de contrôler l'approximation Monte Carlo et de vérifier la convergence de l'algorithme Monte Carlo BOEM (voir Chapitre 2).

Les résultats présentés dans la Section 5.1 et l'Annexe 5.3 sont issus de l'article [Le Corff et Fort, 2011a], accepté pour publication dans une revue internationale. La Section 5.2 présente quant à elle des résultats issus de l'article [Dumont et Le Corff, 2012b], soumis dans une revue internationale.

5.1 SLAM basé sur des landmarks

L'état du robot est représenté par le vecteur $\{X_{t,i}\}_{i=1}^{3}$ où $X_{t,1}$ et $X_{t,2}$ sont les coordonnées cartésiennes du robot à l'instant t et où $X_{t,3}$ est son orientation à l'instant t. A chaque instant le mobile reçoit des commandes $u_t \stackrel{\text{def}}{=} (v_t, \psi_t)$ utilisées pour faire évoluer son état entre les instants t et $t + \mathrm{d}t$. v_t correspond à la commande de vitesse alors que ψ_t correspond à la commande d'orientation. Ces commandes sont déterministes et on supposera dans la suite qu'elles ne dépendent ni de l'état courant du robot ni de la carte. L'état du robot à l'instant t, étant donnés son état à l'instant précédent et les contrôles $(\widehat{V}_t, \widehat{\Psi}_t)$, s'écrit

$$x_t = f(x_{t-1}, \widehat{V}_t, \widehat{\Psi}_t) \,, \tag{5.1}$$

où $(\widehat{V}_t, \widehat{\Psi}_t)$ est un vecteur gaussien de moyenne (v_t, ψ_t) et de matrice de variance Q.

Nous supposons que la carte est représentée par un ensemble de points caractéristiques de l'environnement appelés *landmarks*. Le problème de cartographie consiste à estimer le vecteur $\theta \stackrel{\text{def}}{=} \{\theta_j\}_{j=1}^{q}$, où θ_j représente les coordonnées dans le plan du *landmark* numéro j. A l'instant t, le mobile

observe les *landmarks* situés dans son voisinage. Notons $c_t \subset \{1, \cdots, q\}$ l'ensemble des *landmarks* observés à l'instant t. Conditionnellement à c_t, les observations $\{Y_{t,i}\}_{i \in c_t}$ sont supposées indépendantes et vérifient

$$Y_{t,i} = h(X_t, \theta_i) + \delta_{t,i} \ , \tag{5.2}$$

où les $\{\delta_{t,i}\}_{i \in c_t}$ sont des vecteurs gaussiens indépendants centrés de même matrice de variance R.

Dans les premières approches proposées, le problème du SLAM n'est pas perçu comme un problème d'estimation ponctuel dans les HMM mais il s'agit d'estimer la distribution jointe de la carte et de l'état courant (ou de toute la trajectoire) du mobile conditionnellement à toutes les observations reçues. Nous présentons ici les solutions les plus populaires utilisées pour résoudre le problème du SLAM.

EKF-SLAM

Dans le cas de l'EKF-SLAM (pour *Extended Kalman Filter*-SLAM), l'objectif est d'approcher la distribution jointe de l'état du mobile à l'instant t et de la carte θ conditionnellement à toutes les observations reçues depuis le départ ; θ est donc intégré à l'état caché. L'EKF-SLAM suppose que cette distribution est gaussienne de moyenne μ_t et de matrice de variance Σ_t. L'algorithme repose sur une linéarisation des équations (5.1) et (5.2). Cette linéarisation permet d'obtenir un modèle linéaire gaussien. Ainsi, lorsque la commande u_t et les observations $\{Y_{t,i}\}_{i \in c_t}$ sont reçues on peut effectuer le calcul de μ_t et Σ_t en fonction de μ_{t-1} et Σ_{t-1} à l'aide du filtre de Kalman, voir [Kalman et Bucy, 1961].

La linéarisation des équations (5.1) et (5.2) et l'hypothèse de gaussianité des lois conditionnelles ont été analysées dans [Julier et Uhlmann, 2001] et [Bailey *et al.*, 2006a]. Ces travaux donnent des résultats expérimentaux justifiant la non consistance de l'EKF-SLAM lorsque $t \to +\infty$.

Fast-SLAM

L'algorithme Fast-SLAM, proposé par [Montemerlo *et al.*, 2003] propose d'approcher la distribution jointe de toute la trajectoire du mobile jusqu'à l'instant t et de la carte θ conditionnellement à toutes les observations reçues depuis le départ. Il propose d'approcher la trajectoire du mobile en utilisant une famille de particules pondérées, à l'aide de l'algorithme *path-space smoother* présenté au Chapitre 3. Puis, à chaque trajectoire particulaire est associée une estimation de la carte. Conditionnellement à la trajectoire du robot, les *landmarks* sont supposés indépendants et de loi gaussienne. La moyenne et la variance des gaussiennes associées à chaque *landmark* sont mises à jour en linéarisant le modèle d'observations (5.2).

Contrairement à l'EKF-SLAM, l'utilisation de filtres particulaires ne nécessite pas de linéarisation de l'équation (5.1). L'intérêt majeur de l'algorithme Fast-SLAM réside dans le fait qu'à chaque trajectoire particulaire est associée une estimation de la carte. Cependant, l'approximation particulaire utilisée dans [Montemerlo *et al.*, 2003] est basée sur le *path-space smoother* : la dégénérescence des trajectoires implique que toutes les trajectoires partagent les mêmes estimations pour tous les *landmarks* qui ne sont pas observés durant une longue période. Différentes méthodes pour améliorer la façon de propager les particules ont été proposées (voir par exemple [Burgard *et al.*, 2005]) mais ne permettent pas pour de longues trajectoires de résoudre le problème de dégénérescence. La divergence des estimations données par l'algorithme Fast-SLAM a été mise en évidence de façon numérique dans [Bailey *et al.*, 2006b].

Marginal-SLAM

L'algorithme Marginal-SLAM de [Martinez-Cantin, 2008] propose de produire une estimation ponctuelle de la carte et d'estimer l'état du mobile à l'instant t à l'aide de méthodes de Monte Carlo séquentielles. La carte est perçue comme un paramètre d'un modèle de Markov caché qui est estimé en ligne à l'aide d'une procédure de type maximum de vraisemblance récursif. La position du mobile est approchée à l'aide d'une famille de particules pondérées : l'avantage de cette approche provient de la stabilité en temps long des approximations particulaires des lois de filtrage puisque la méthode d'approximation stochastique utilisée ne nécessite plus de conserver toute la trajectoire particulaire comme dans le cas de l'algorithme Fast-SLAM.

Le calcul du maximum de vraisemblance récursif repose sur un calcul récursif de la log-vraisemblance et sur des calculs de gradient qui sont sensibles à l'implémentation (par exemple, sensibles à la taille des pas de mise à jour dans la direction de descente). Les méthodes basées sur l'algorithme EM ne sont pas sensibles à cette paramétrisation et sont donc plus robustes. Nous proposons donc en Section 5.1 une méthode basée sur l'algorithme BOEM. Cet algorithme partage les mêmes objectifs que l'algorithme Marginal-SLAM : une estimation ponctuelle de la carte est associée à une famille de particules pondérées approchant l'état du mobile à chaque instant. Nous comparons les performances de ces deux algorithmes.

5.1.1 Application de l'algorithme Monte Carlo BOEM

Nous utilisons ici le modèle cinématique du tricycle (voir par exemple [Bailey *et al.*, 2006a]) où la fonction f de l'équation (5.1) est donnée par

$$f(x_{t-1}, v_t, \psi_t) = x_{t-1} + \begin{pmatrix} \hat{v}_t dt \cos(x_{t-1,3} + \psi_t) \\ v_t dt \sin(x_{t-1,3} + \psi_t) \\ v_t dt B^{-1} \sin(\psi_t) \end{pmatrix} ,$$

où dt est l'intervalle de temps entre deux états du mobile et où B est son empattement.

A l'instant t, le mobile observe la distance et la position angulaire de tous les *landmarks* dans son voisinage ; la fonction h de l'équation (5.2) est donc donnée par

$$h(x, \tau) \stackrel{\text{def}}{=} \begin{pmatrix} \sqrt{(\tau_1 - x_1)^2 + (\tau_2 - x_2)^2} \\ \arctan \frac{\tau_2 - x_2}{\tau_1 - x_1} - x_3 \end{pmatrix} .$$

Les matrices de variance R et Q sont supposées connues.

Nous supposons en Section 5.1.2 qu'il n'y a pas d'erreurs d'association entre les observations et les *landmarks*, c'est-à-dire qu'à chaque instant t l'ensemble c_t est connu. Cette hypothèse est relâchée en Section 5.1.3.

Dans ce modèle, le noyau de Markov de la chaîne cachée ne dépend pas de la carte (voir (5.1)) et seule la fonction g_θ intervient dans le calcul de la vraisemblance complète. Or, la vraisemblance g_θ est telle que le modèle n'appartient pas à la famille exponentielle courbe :

$$\sum_{i \in c_t} \ln g_\theta(x_t, y_{t,i}) \propto \sum_{i \in c_t} \left[y_{t,i} - h(x_t, \theta_i) \right]^\star R^{-1} \left[y_{t,i} - h(x_t, \theta_i) \right] , \qquad (5.3)$$

où pour toute matrice A, $A^\star$ est la transposée de A. Par suite, pour appliquer l'algorithme BOEM, au début de chaque bloc d'observations, g_θ est approchée par une fonction dépendant des estimations courantes des *landmarks* : $\{\widehat{\theta}_i\}_{i=1}^q$. Cette approximation est choisie de manière à obtenir une vraisemblance complète appartenant à la famille exponentielle courbe. Par (5.3), approcher la fonction $\tau \mapsto h(x, \tau)$ par son développement de Taylor à l'ordre 1 au point $\tau = \widehat{\theta}_i$ donne une approximation de $\theta \mapsto \sum_{i \in c_t} \ln g_\theta(x_t, y_{t,i})$ quadratique en θ. Cette approche est fréquemment utilisée dans le cadre du SLAM de manière à profiter des outils concernant les modèles linéaires gaussiens, voir par exemple [Burgard *et al.*, 2005]. Nous obtenons alors une approximation de la vraisemblance complète donnée par

$$\theta \mapsto \sum_{i=1}^{q} \left(\theta_i - \widehat{\theta}_i \right)^\star A_i \left(\theta_i - \widehat{\theta}_i \right) - 2 B_i \left(\theta_i - \widehat{\theta}_i \right) + C_i , \qquad (5.4)$$

où q est le nombre total de *landmarks* et, pour tout $i \in \{1, \cdots, q\}$,

$$A_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbf{1}_{i \in c_t} \nabla_\tau h(x_t, \widehat{\theta}_i)^\star R^{-1} \nabla_\tau h(x_t, \widehat{\theta}_i) \ ,$$

$$B_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbf{1}_{i \in c_t} \left[y_{t,i} - h(x_t, \widehat{\theta}_i) \right]^\star R^{-1} \nabla_\tau h(x_t, \widehat{\theta}_i) \ ,$$

$$C_i \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbf{1}_{i \in c_t} \left[y_{t,i} - h(x_t, \widehat{\theta}_i) \right]^\star R^{-1} \left[y_{t,i} - h(x_t, \widehat{\theta}_i) \right] \ .$$

Cette forme exponentielle de la vraisemblance complète permet d'appliquer l'algorithme BOEM pour estimer la carte. Etant donnée la forme de (5.4), la mise à jour de l'estimation de la carte (lorsque toutes les observations du bloc ont été utilisées) se fait *landmark* par *landmark*.

Comme tous les *landmarks* ne sont pas observés à chaque instant, nous choisissons une suite $\{\tau_n \propto n^{1.1}\}_{n \geq 1}$ à croissance lente pour que le nombre total de mises à jour ne soit pas trop faible (dans l'expérience de la Section 5.1.2, nous avons 60 mises à jour pour un nombre total d'observations de 2000). Puisque le nombre d'observations n'est pas trop grand (le plus grand bloc contient 60 observations), le nombre de particules par bloc est constant : pour tout $n \geq 1$, $N_n = 50$.

5.1.2 Données simulées

Obtention des données

Dans l'expérience qui suit, $q = 15$ *landmarks* sont placés dans un carré de taille 45x45. La trajectoire du mobile est simulée en utilisant un ensemble connu de commandes de vitesse et d'orientation. En utilisant les *vraies* positions de tous les *landmarks* et la *vraie* trajectoire du mobile (points et ligne continue dans la Figure 5.1), des observations sont simulées en utilisant $R = \begin{pmatrix} \sigma_r^2 & \rho \\ \rho & \sigma_b^2 \end{pmatrix}$, où $\sigma_r = 0,5\,\text{m}$, $\sigma_b = \frac{\pi}{60}\text{rad}$ et $\rho = 0,01$. Nous choisissons $Q = \text{diag}(\sigma_v^2, \sigma_\phi^2)$ où $\sigma_v = 0,5\,\text{m.s}^{-1}$, $\sigma_\psi = \frac{\pi}{60}\,\text{rad}$ et $B = 1.5\,\text{m}$. La mission dure le temps d'acquérir $T = 2000$ observations. A chaque instant le mobile reçoit des observations de tous les *landmarks* présents dans un rayon de 10m.

Résultats

La position estimée à chaque instant est égale à la moyenne pondérée des positions données par les particules. Pour chaque simulation, la trajectoire estimée et l'estimation de la carte à la fin de la mission sont conservées. La Figure 5.1 représente les moyennes des trajectoires estimées et des cartes

estimées sur 50 simulations Monte Carlo indépendantes. Cette figure illustre le bon comportement de l'algorithme Monte Carlo BOEM appliqué à ce problème de SLAM.

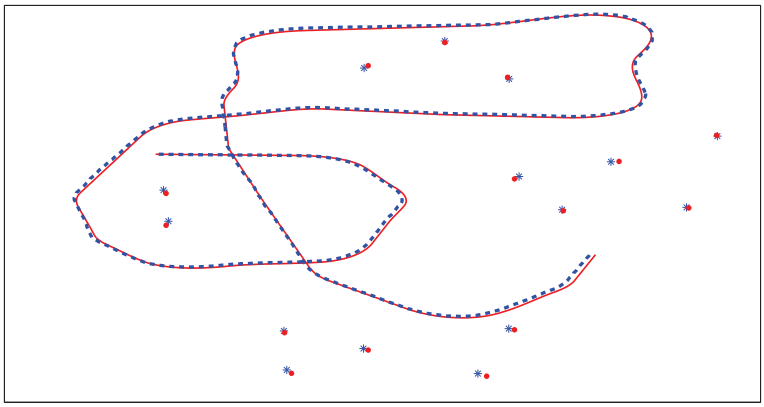


FIGURE 5.1 – Vraie trajectoire (ligne continue) et vraies positions des *landmarks* (boules) avec la trajectoire estimée (ligne pointillée) et les positions estimées des *landmarks* (étoiles) à la fin de la mission ($T = 2000$).

Nous comparons aussi notre algorithme à l'algorithme Marginal-SLAM proposé par [Martinez-Cantin, 2008] et présenté plus haut. La Figure 5.2 illustre l'estimation de la position de chaque *landmark*. L'algorithme Monte Carlo BOEM est appliqué dans le même cadre que précédemment et les paramètres d'implémentation de l'algorithme Marginal-SLAM sont choisis de telle sorte que les vitesses de convergence théoriques de l'algorithme BOEM et de la procédure de maximum de vraisemblance récursif du Marginal-SLAM soient les mêmes. Nous utilisons la version moyennisée de l'algorithme Monte Carlo BOEM et une procédure de moyennisation de type Polyak-Ruppert pour le Marginal-SLAM (voir [Polyak, 1990]). Pour chaque *landmark*, la dernière estimation de la position (à la fin de la mission) est conservée pour chacune des 50 simulations Monte Carlo. La Figure 5.2 présente la distance entre la position estimée et la vraie position pour chaque *landmark*. On constate dans cette expérience les meilleures performances de l'algorithme Monte Carlo BOEM tant en terme de biais qu'en terme de variance de l'estimation. La variabilité de l'estimation fournie par l'algorithme Marginal-SLAM provient du fait que la mise à jour du paramètre a lieu à chaque acquisition d'une nouvelle observation. On peut améliorer ses performances en ne mettant à jour l'estimation de la carte qu'après le traitement de blocs d'observations pour obtenir des estimations similaires à celles de l'algorithme Monte Carlo BOEM.

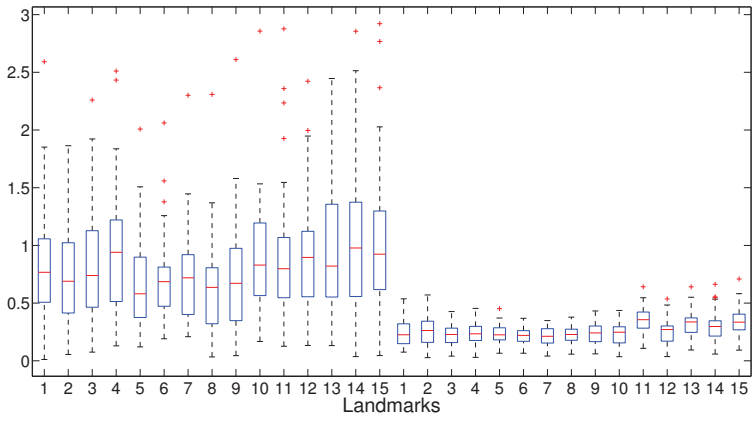


FIGURE 5.2 – Distance entre la position estimée à la fin de la mission et la vraie position pour chacun des 15 *landmarks* avec le Marginal-SLAM moyennisé (à gauche) et l'algorithme BOEM moyennisé (à droite).

5.1.3 Données réelles

La Figure 5.3 illustre les performances de l'algorithme Monte Carlo BOEM sur un jeu de données réelles. Nous utilisons le *car park data set*[1] qui correspond à une situation de SLAM basée sur des *landmarks* artificiels placés dans un parking. Dans ce cas, le processus d'association entre les observations et les *landmarks* n'est pas supposé connu. Ainsi, nous proposons le mécanisme d'association suivant (à la suite de l'algorithme fourni avec la base de données) : à chaque fois qu'une nouvelle observation $y_{t,i}$ est disponible, sa log-vraisemblance $\left[Y_{t,i} - h(\widehat{x}_t, \widehat{\theta}_i)\right]^\star R^{-1} \left[Y_{t,i} - h(\widehat{x}_t, \widehat{\theta}_i)\right]$ est calculée pour chaque *landmark* $\widehat{\theta}_i$ dans l'estimation courante de la carte et lorsque $\widehat{x}_t$ est l'estimation courante de l'état du mobile (moyenne pondérée des états donnés par les particules). Si toutes les vraisemblances sont sous un certain seuil, un nouveau *landmark* est créé. Sinon, l'observation est associée au *landmark* correspondant à la plus forte valeur de vraisemblance. L'algorithme est commencé avec une carte vide et des *landmarks* sont créés avec les premières observations. Nous associons à chaque *landmark* un compteur donnant le nombre de fois où on lui associe une observation. Si un *landmark* dans le voisinage du mobile n'est associé à aucune observation on fait décroître son compteur. Si le compteur d'un *landmark* est négatif, on l'ôte de la carte estimée. Dans la Figure 5.3, la trajectoire estimée et la carte estimée à la fin d'une mission ($T = 5565$) sont représentées. On utilise l'algorithme Monte Carlo BOEM avec les mêmes tailles de blocs que celles précisées en Section 5.1.1 mais un nombre de particules donné par $N_n = 100$.

1. Disponible à l'adresse `http://www.cas.kth.se/SLAM`

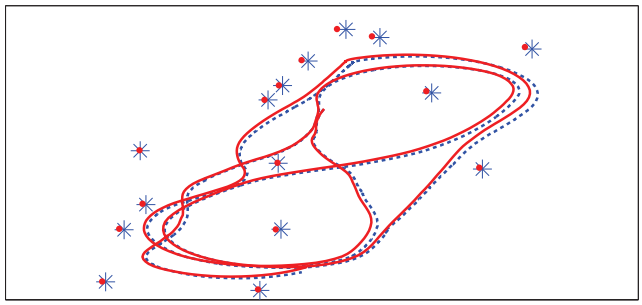


FIGURE 5.3 – Carte et trajectoire estimées sur le *car park dataset*. La trajectoire estimée (pointillés) et les positions estimées des *landmarks* (étoiles) données par l'algorithme BOEM sont comparées aux vraies valeurs (trait continu et points).

5.2 Application au SLAM indoor

Dans cette section, nous nous intéressons à la localisation d'un mobile dans un réseau de points d'accès WiFi. Le travail présenté dans cette section est l'objet de l'article [Dumont et Le Corff, 2012b], soumis pour publication dans une revue internationale.

Nous voulons estimer la position du mobile en utilisant la puissance des signaux WiFi reçus des points d'accès à proximité du mobile. L'estimation de la position du mobile repose sur la connaissance des cartes de propagation des signaux de chaque borne WiFi. A la différence de ce qui est proposé dans la littérature où l'on traite successivement la cartographie et la localisation, nous estimons simultanément la carte et la position du mobile. Nous trouvons différentes solutions dans la littérature :

- soit une phase de calibration préalable est effectuée en différents points de l'environnement permettant d'obtenir un ensemble de mesures de positions associées à des puissances reçues. Ensuite le problème de localisation est résolu en sélectionnant, parmi les données de calibration, les observations les plus proches de celle reçue par le mobile. La position estimée du mobile est alors définie comme un barycentre des positions associées aux mesures de calibration sélectionnées (voir par exemple [Evennou et Marx, 2006]).
- soit la cartographie est résolue de façon préalable en prenant en compte les obstacles (par exemple les murs) présents dans l'environnement (voir [Gorce *et al.*, 2007]). Néanmoins, les modèles statistiques envisagés se révèlent insuffisants pour prendre en compte des modifications dans la façon dont les signaux WiFi se propagent. Nous reviendrons plus en détail sur ces limitations en section 5.2.1.

Dans cette section, nous proposons une solution basée sur le SLAM com-

biné à un nouveau modèle pour représenter les cartes de propagation. Les données collectées par le mobile permettent d'estimer la carte puis de définir un estimateur de la position du mobile à l'aide d'une variante de l'algorithme Monte Carlo BOEM. L'avantage d'utiliser une telle procédure en ligne permet de prendre en compte toute modification dans la propagation des signaux WiFi. L'estimation des cartes de propagation est affectée par les mesures reçues par le mobile et l'estimation de la position n'est pas détériorée. Dans le cas d'une estimation fixe de la carte effectuée avant la procédure de localisation, ces changements ne sont pas pris en compte et la localisation se dégrade au fil du temps.

5.2.1 Modèle considéré

Nous notons $\{X_k\}_{k\geq 0}$ les coordonnées cartésiennes dans le plan du mobile. Pour des raisons pratiques, l'environnement continu est représenté par une grille $\mathcal{C}$. On suppose que $\{X_k\}_{k\geq 0}$ est une chaîne de Markov à valeurs dans l'ensemble fini $\mathcal{C}$, de distribution initiale χ et de matrice de transition donnée, pour tout $(x, x') \in \mathcal{C}^2$, par

$$m_{x,x'} \propto \mathrm{e}^{-\|x-x'\|^2/a} ,$$

où $a \in \mathbb{R}_+^\star$ (supposé connu) dépend de la vitesse moyenne du mobile et où $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^2$. Notons $|\mathcal{C}|$ le nombre d'éléments de $\mathcal{C}$.

Soit $G \stackrel{\text{def}}{=} \{G_{j,x}, 1 \leq j \leq B,\ x \in \mathcal{C}\}$ la matrice de taille $B \times |\mathcal{C}|$ telle que l'élément $G_{j,x}$ représente la puissance moyenne du signal reçu de la borne j quand le mobile est en position x. A chaque instant, le mobile reçoit l'observation Y_k à valeurs dans $\mathbb{R}^B$ modélisée par

$$Y_k \stackrel{\text{def}}{=} G_{.,X_k} + \varepsilon_k, \tag{5.5}$$

où $G_{.,x}$ est le vecteur $\{G_{j,x}\}_{j=1}^B$ et $\{\varepsilon_k\}_{k\geq 0}$ est une suite de vecteurs gaussiens indépendants, indépendants des $\{X_k\}_{k\geq 0}$, de moyenne 0 et de matrice de variance $\sigma^2 I_B$ (I_B est la matrice unité de taille $B \times B$).

Nous considérons un modèle dans lequel pour tout $j \in \{1, \cdots, B\}$ et tout $x \in \mathcal{C}$, $G_{j,x}$ est décomposé en deux termes :

$$G_{j,x} \stackrel{\text{def}}{=} \mu_{j,x} + \delta_{j,x} , \tag{5.6}$$

où

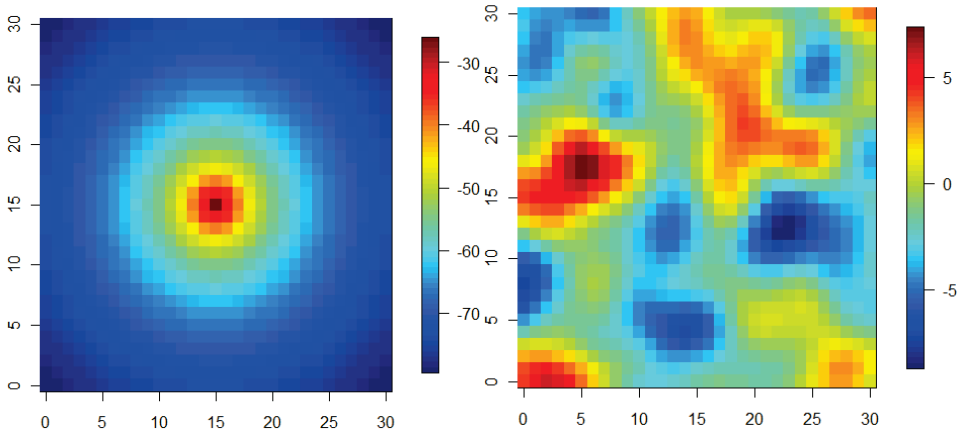
– le terme $\mu_{j,x}$ est issue de l'équation de transmission de Friis, voir [Friis, 1946] :

$$\mu_{j,x} \stackrel{\text{def}}{=} c_j + d_j \log \|x - O_j\| , \tag{5.7}$$

c_j, d_j étant des réels et, $\{O_j\}_{j=1}^B$ étant la position des B bornes WiFi. $\mu_{j,x}$ représente la propagation moyenne dépendant de la distance entre la position x du mobile et la borne j.

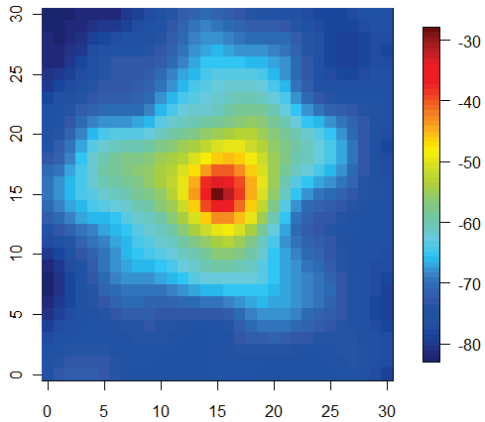
– le terme additif $\delta_{j,x}$ représente les perturbations dues à l'environnement, telles que la présence d'obstacles. Nous ne faisons aucune hypothèse a priori sur la relation entre j et x. L'originalité de notre modèle réside dans l'introduction de ce terme de perturbation $\delta_{j,x}$ dans les équations de Friis : les travaux antérieurs portant sur la cartographie omettent ce terme offrant donc moins de flexibilité quant à la modélisation de la carte.

La Figure 5.4 représente les fonctions $\mu_{j,.}$, $\delta_{j,.}$ et $G_{j,.}$ définies sur la grille $\mathcal{C} = \{0, \ldots, 30\} \times \{0, \ldots, 30\}$. Les paramètres utilisés pour obtenir ces représentations sont donnés en Section 5.2.3.



(a) $\mu_{j,.}$.

(b) $\delta_{j,.}$.



(c) $G_{j,.} = \mu_{j,.} + \delta_{j,.}$.

FIGURE 5.4 – Exemple de représentation spatiale de $\mu_{j,.}$, $\delta_{j,.}$ et $G_{j,.} = \mu_{j,.} + \delta_{j,.}$. Ces représentations correspondent aux signaux émis par la borne placée en position $(15, 15)$.

Dans la suite, l'objectif est d'estimer

$$c \stackrel{\text{def}}{=} \{c_j\}_{j=1}^B \,, \qquad d \stackrel{\text{def}}{=} \{d_j\}_{j=1}^B \,, \qquad \delta \stackrel{\text{def}}{=} \{\delta_j\}_{j=1}^B \,,$$

et σ^2. Nous notons $\theta \stackrel{\text{def}}{=} (c,\ d,\ \delta,\ \sigma^2)$.

5.2.2 Application de l'algorithme Monte Carlo BOEM

D'après (5.5), la vraisemblance de l'observation $y = (y_1, \cdots, y_B)$ sachant $x \in \mathcal{C}$ est donnée par

$$g_\theta(x, y) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-B/2} \prod_{j=1}^{B} \exp\left\{-\frac{1}{2\sigma^2}\left|y_j - G_{j,x}\right|^2\right\} .$$

La dynamique de l'état caché ne dépend pas du paramètre θ ainsi, la log-vraisemblance complète normalisée $T^{-1}\log p_\theta(X_{0:T}, Y_{0:T})$ (voir (1.3)) s'écrit, à un terme additif près ne dépendant pas de θ,

$$-\frac{B}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{B}\left\{\left\langle \frac{1}{T}\sum_{k=1}^{T} s_1(X_k), G_j^2\right\rangle + \frac{1}{T}\sum_{k=1}^{T} s_{3,j}(Y_k) - 2\left\langle \frac{1}{T}\sum_{k=1}^{T} s_{2,j}(X_k, Y_k), G_j\right\rangle\right\} , \quad (5.8)$$

où G_j et G_j^2 sont les vecteurs de $\mathbb{R}^{|\mathcal{C}|}$ donnés par

$$G_j \stackrel{\text{def}}{=} \{G_{j,x}\}_{x\in\mathcal{C}} \,, \qquad G_j^2 \stackrel{\text{def}}{=} \{G_{j,x}^2\}_{x\in\mathcal{C}} \,,$$

et les statistiques sont données par

$$\begin{aligned} s_1(X_k) &\stackrel{\text{def}}{=} \{1_x(X_k)\}_{x\in\mathcal{C}}, \\ s_{2,j}(X_k, Y_k) &\stackrel{\text{def}}{=} \{1_x(X_k)Y_{k,j}\}_{x\in\mathcal{C}}, \\ s_{3,j}(Y_k) &\stackrel{\text{def}}{=} Y_{k,j}^2 \,. \end{aligned}$$

La statistique $s_1(x)$ est le vecteur de la base canonique de $\mathbb{R}^{|\mathcal{C}|}$ qui vaut 1 en position x. Par suite, $\frac{1}{T}\sum_{k=1}^{T} s_1(X_k)$ représente le nombre moyen de visites à chaque point de la grille. Pour chaque j, la quantité $\frac{1}{T}\sum_{k=1}^{T} s_{2,j}(X_k, Y_k)$ représente la moyenne des puissances reçues de la borne j en chaque point de la grille.

Notons que G_j dépend de θ (voir Eqs (5.6 et (5.7)) ; pour simplifier la notation, nous omettons la dépendance.

L'algorithme Monte Carlo BOEM nécessite le calcul, pour tout $j \in \{1, \ldots, B\}$, des quantités suivantes :

$$\mathsf{S}_1^n \overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \mathbb{E}_{\theta_n}^{\chi, T_n} \left[\sum_{k=1}^{\tau_{n+1}} s_1(X_{k+T_n}) \middle| Y_{T_n+1:T_{n+1}} \right] ,$$

$$\mathsf{S}_{2,j}^n \overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \mathbb{E}_{\theta_n}^{\chi, T_n} \left[\sum_{k=1}^{\tau_{n+1}} s_{2,j}(X_{k+T_n}, Y_{k+T_n}) \middle| Y_{T_n+1:T_{n+1}} \right] ,$$

$$\mathsf{S}_{3,j}^n \overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{k=1}^{\tau_{n+1}} s_{3,j}(Y_{k+T_n}) .$$

Les espérances S_1^n, $\mathsf{S}_{2,j}^n$, $\mathsf{S}_{3,j}^n$ peuvent être calculées explicitement puisque l'espace d'état $\mathcal{C}$ est fini. Cependant, dans notre cas, $|\mathcal{C}|$ est élevé et les calculs en ligne sont trop coûteux pour être effectués de façon exacte. Nous choisissons donc d'appliquer l'algorithme Monte Carlo BOEM à l'aide de l'approximation particulaire présentée en Annexe 5.3 pour approcher ces espérances.

Néanmoins, pour la vraisemblance complète décrit par (5.8), l'étape de maximisation de l'algorithme Monte Carlo BOEM n'a pas (toujours) de solution du fait de la non-identifiabilité de notre modèle. En effet, le vecteur S_1^n est "creux" au sens où de nombreuses composantes sont nulles en pratique. Ce phénomène est d'autant plus fréquent que la taille des blocs τ_n est petite.

Par suite, nous considérons une modification de l'algorithme Monte Carlo BOEM obtenue en substituant la maximisation de la log-vraisemblance complète moyenne par une log-vraisemblance (complète) pénalisée moyenne : θ_{n+1} est alors définie comme la valeur θ maximisant la fonction

$$\theta \mapsto -\frac{B}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{B} \left\{ \left\langle \mathsf{S}_1^n, G_j^2 \right\rangle + \mathsf{S}_{3,j}^n - 2 \left\langle \mathsf{S}_{2,j}^n, G_j \right\rangle \right\} - \frac{1}{2\tau_{n+1}} \sum_{j=1}^{B} \delta_j^\star \Sigma_j^{-1} \delta_j ,$$

où le terme $\delta_j^\star \Sigma_j^{-1} \delta_j$ est une pénalité quadratique sur la perturbation δ_j de la propagation des signaux WiFi. Le choix de la pénalité quadratique permet d'obtenir un critère qui peut être maximisé de façon simple et explicite. D'autre part, nous affectons cette pénalité du poids τ_{n+1}^{-1} qui permet de l'atténuer pour les blocs contenant un grand nombre d'observations.

Il est aisé de voir que l'étape de maximisation de notre algorithme consiste à mettre à jour chaque triplet (c_j, d_j, δ_j) - indépendamment les uns des autres - puis à mettre à jour σ^2. Nous détaillons maintenant ces équations de mise à jour.

Notons L la matrice de taille $B \times |\mathcal{C}|$ donnée par $\mathrm{L}_{j,x} \stackrel{\text{def}}{=} \log|x - O_j|$; $\mathbf{1}$ le vecteur colonne de taille $|\mathcal{C}|$ dont toutes les composantes sont égales à 1 ; et $\mathrm{diag}(\mathsf{S}_1^n)$ la matrice diagonale dont les éléments diagonaux sont donnés par le vecteur S_1^n. Pour tout $j \in \{1, \ldots, B\}$, nous définissons

$$\begin{aligned}
M_{0,j} &\stackrel{\text{def}}{=} \left[\mathrm{diag}(\mathsf{S}_1^n) + \frac{\sigma^2}{\tau_{n+1}}\Sigma_j^{-1}\right] , \\
M_{1,j} &\stackrel{\text{def}}{=} \mathrm{diag}(\mathsf{S}_1^n)\left[I - M_{0,j}^{-1}\mathrm{diag}(\mathsf{S}_1^n)\right] , \\
M_{2,j} &\stackrel{\text{def}}{=} I - \mathrm{diag}(\mathsf{S}_1^n)M_{0,j}^{-1} , \\
W_{1,j} &\stackrel{\text{def}}{=} \mathbf{1}^\star M_{1,j}\mathbf{1} , \\
W_{2,j} &\stackrel{\text{def}}{=} \mathbf{1}^\star M_{1,j}\mathrm{L}_j , \\
W_{3,j} &\stackrel{\text{def}}{=} \mathrm{L}_j^\star M_{1,j}\mathrm{L}_j , \\
r_j &\stackrel{\text{def}}{=} W_{1,j}W_{3,j} - W_{2,j}^2 .
\end{aligned}$$

Les nouvelles estimations sont alors données, pour tout $j \in \{1, \ldots, B\}$, par

$$\begin{aligned}
\delta_j &= M_{0,j}^{-1}\left[\mathsf{S}_{2,j}^n - \mathrm{diag}(\mathsf{S}_1^n)(c_j\mathbf{1} + d_j\mathrm{L}_j)\right] , \\
c_j &= r_j^{-1}\left[W_{3,j}\mathbf{1}^\star - W_{2,j}\mathrm{L}_j^\star\right] M_{2,j}\mathsf{S}_{2,j}^n , \\
d_j &= r_j^{-1}\left[-W_{2,j}\mathbf{1}^\star + W_{1,j}\mathrm{L}_j^\star\right] M_{2,j}\mathsf{S}_{2,j}^n , \\
G_j &= c_j\mathbf{1} + d_j\mathrm{L}_j + \delta_j
\end{aligned}$$

et

$$\sigma^2 = \frac{1}{B}\sum_{j=1}^{B}\left\{(G_j^2)^\star\,\mathrm{diag}(\mathsf{S}_1^n)G_j^2 - 2(\mathsf{S}_{2,j}^n)^\star G_j + \mathsf{S}_{3,j}^n\right\} .$$

5.2.3 Expérience avec des données simulées

Obtention des données

Dans cette section, les performances de l'algorithme Monte Carlo BOEM sont illustrées avec des données simulées. Les expériences sont réalisées sur la grille $\mathcal{C} = \{0, \ldots, 30\} \times \{0, \ldots, 30\}$. On utilise $B = 17$ bornes WiFi, toutes les bornes étant modélisées avec les mêmes coefficients c et d, voir (5.7), pour tout $j \in \{1, \ldots, B\}$,

$$c_j = -26 \quad \text{et} \quad d_j = -17.5 .$$

Pour tout $j \in \{1, \ldots, B\}$, Σ_j est la matrice de variance-covariance donnée par $\Sigma_j(x, x') \stackrel{\text{def}}{=} v_1 * \exp(-\|x - x'\|^2/v_2)$ avec $v_1 = 10$ et $v_2 = 18$. La variance du bruit d'observation est fixée à $\sigma^2 = 25$. Le paramètre de la matrice de transition de la chaîne cachée est fixé à $a = 6$.

Mise en oeuvre de l'algorithme

Toutes les simulations sont initialisées en prenant :

$$\forall j \in \{1, \ldots, B\},\ c_j^0 = -10\ ,\ d_j^0 = -30\ ,\ \delta_j = 0 \quad \text{et} \quad \sigma_0^2 = 30\ .$$

Les tailles des blocs d'observations sont données, pour tout $n \geq 1$, par $\tau_n = 10n + 500$. Cette croissance lente pour la suite $\{\tau_n\}_{n\geq 0}$ a été fixée de façon à obtenir un nombre important de mises à jour de l'estimation de la carte qui ne nécessite pas un grand nombre d'observations. Le nombre de particules est fixé à $N_n = 25$ pour chaque bloc et la position initiale des particules est tirée aléatoirement sur $\mathcal{C}$. Nous donnons en Annexe 5.3 la méthode de Monte Carlo séquentielle utilisée pour approcher S_1^n, $\mathsf{S}_{2,j}^n$ et $\mathsf{S}_{3,j}^n$ sur chaque bloc.

i) *Estimation de la carte* : pour chaque carte G_j, nous calculons la norme L_1 de l'erreur :

$$\epsilon_j \stackrel{\text{def}}{=} \frac{1}{|\mathcal{C}|} \sum_{x\in\mathcal{C}} \left| G_{j,x} - \widehat{G}_{j,x} \right| \ ,$$

où $\widehat{G}_j$ est l'estimation courante de la j-ème carte de propagation. Nous représentons la moyenne sur toutes les cartes :

$$\bar{\epsilon} \stackrel{\text{def}}{=} \frac{1}{B} \sum_{j=1}^{B} \epsilon_j \ ,$$

dans le cas où $\widehat{G}_{j,x}$ est calculé par l'algorithme Monte Carlo BOEM et dans le cas où il est calculé par l'algorithme Monte Carlo BOEM moyennisé. Nous parlerons respectivement de l'erreur L_1 pour l'estimation de la carte et pour son estimation moyennisée.

ii) *Localisation du mobile* : la localisation est effectuée selon trois procédures utilisant trois systèmes de particules différents. Pour chaque système de particules, l'estimation de la position à un instant donné est définie comme étant la moyenne pondérée des particules.
   - *Localisation BOEM* : l'estimation de la position est donnée par le système de particules utilisé dans l'algorithme Monte Carlo BOEM.
   - *Localisation BOEM moyennisé* : l'estimation de la position est donnée par un second système de particules qui n'est utilisé que pour localiser le mobile. Les particules de ce système sont propagées avec $\vartheta_t^n = 1$ et $p_t^n = m$ et les poids sont calculés en fixant la carte à sa valeur estimée par l'algorithme Monte Carlo BOEM *moyennisé* (voir l'Annexe 5.3 pour les détails sur la propagation des systèmes de particules). Ce choix provient du fait que l'estimation moyennisée des cartes est meilleure que l'estimation non moyennisée (voir Figure 5.5).

– *Localisation optimale* : nous présentons également la meilleure performance possible à l'aide d'un troisième système de particules. Les particules de ce système sont propagées avec $\vartheta_t^n = 1$ et $p_t^n = m$ et les poids sont calculés en fixant la carte à sa vraie valeur (voir l'Annexe 5.3 pour les détails sur la propagation des systèmes de particules). Ainsi, dans l'incertitude sur l'estimation de la position du mobile, on élimine les fluctuations dues à l'estimation de la carte.

Sur chaque bloc, et pour chaque instant du bloc, nous calculons la distance entre la vraie position et la position estimée du mobile ; nous reportons le 0.8-quantile empirique de l'erreur de localisation calculé à partir de cet échantillon.

Résultats

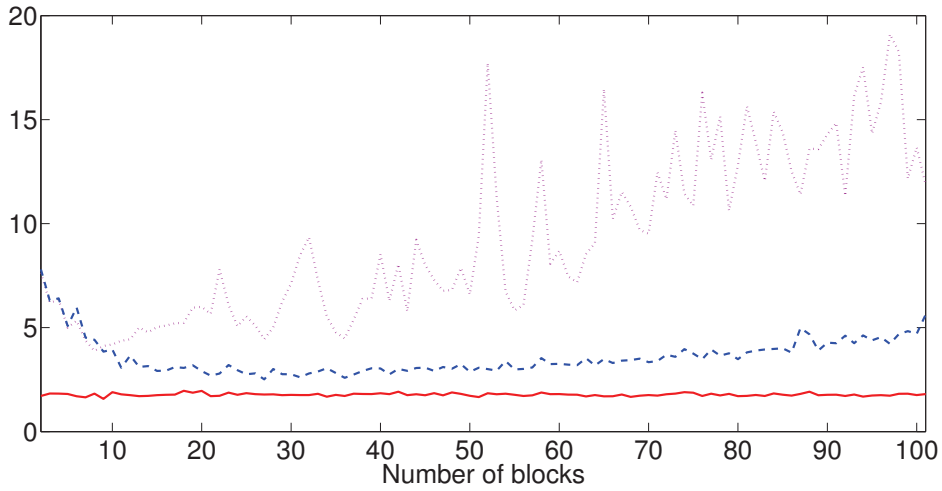
La Figure 5.5 représente l'erreur de localisation du mobile et l'erreur d'estimation de la carte. La Figure 5.5a montre que l'estimation de la position ne converge pas lorsque le nombre de blocs (soit le nombre d'observations) tend vers $+\infty$. Après 50 blocs (environ 40000 observations) la position est mal estimée et l'erreur de localisation accroît l'erreur d'estimation de la carte (donnée par la Figure 5.5b).

Pour surmonter cette difficulté, nous proposons d'utiliser l'estimation moyennisée de la carte : on remplace périodiquement l'estimation de la carte par son estimation moyennisée. Nous améliorons ainsi l'estimation moyennisée de la carte. L'estimation non moyennisée ne diverge plus mais augmente sur un cycle (entre deux substitutions de la carte moyennisée à la carte non moyennisée) dont la longueur est fixée par l'utilisateur. Cette procédure est illustrée en Figure 5.6, où l'étape de stabilisation est réalisée tous les 5 blocs.

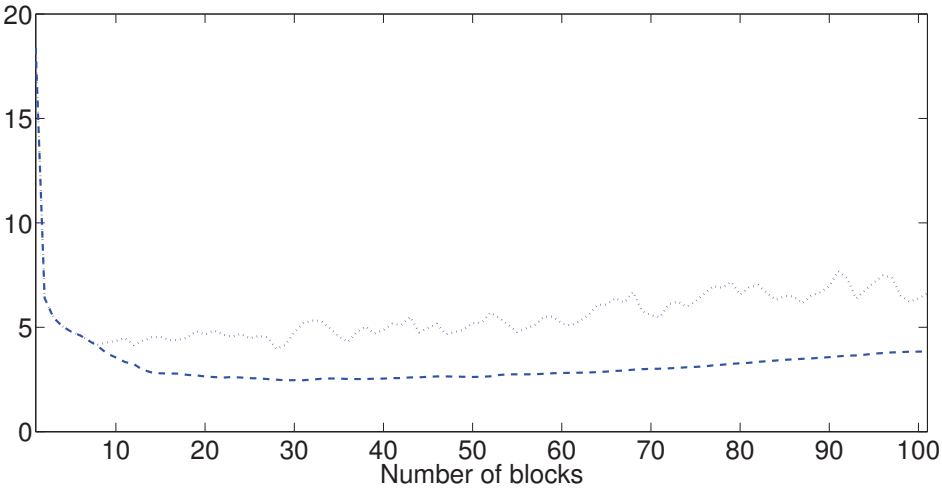
Les figures 5.7 et 5.8 représentent les erreurs de localisation du mobile sur 50 simulations Monte Carlo indépendantes. Dans ces simulations, l'estimation de la carte est remplacée tous les 5 blocs par son estimation moyennisée et nous utilisons la localisation BOEM moyennisée. Dans la Figure 5.7, l'erreur de localisation optimale (c'est-à-dire lorsque l'on suppose les cartes connues et que l'on se contente de localiser le mobile) est aussi donnée. La convergence de l'erreur de localisation vers la valeur optimale est presque atteinte après 100 blocs (environ 100000 observations).

5.2.4 Expérience avec des données réelles

Dans cette section, les performances de notre algorithme sont testées dans une situation réelle. 10 bornes WiFi ont été installées dans un entrepôt (la Figure 5.9 représente la carte de l'environnement avec la position des bornes WiFi).

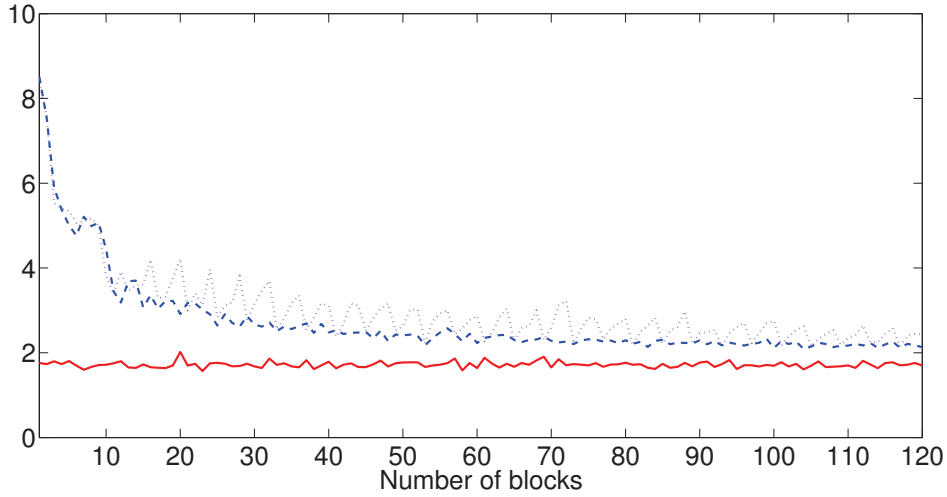


(a) 0.8-quantile de la distance entre la vraie position et la position estimée. L'erreur de localisation est donnée pour la localisation BOEM (pointillés), pour la localisation BOEM moyennisée (tirets) et pour la localisation optimale (ligne continue).

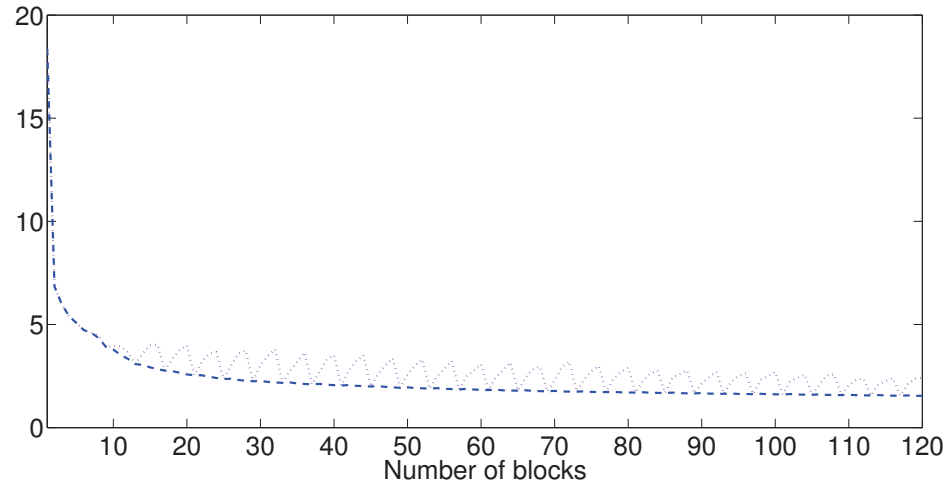


(b) Erreur L_1 sur l'estimation de la carte (pointillés) et l'estimation moyennisée de la carte (tirets).

FIGURE 5.5 – Estimation par l'algorithme Monte Carlo BOEM : erreurs sur la localisation du mobile et sur l'estimation de la carte.



(a) 0.8-quantile de la distance entre la vraie position et la position estimée. L'erreur de localisation est donnée pour la localisation BOEM (pointillés), pour la localisation BOEM moyennisée (tirets) et pour la localisation optimale (ligne continue).



(b) Erreur L_1 sur l'estimation de la carte (pointillés) et l'estimation moyennisée de la carte (tirets).

FIGURE 5.6 – Estimation par l'algorithme Monte Carlo BOEM avec étape de stabilisation : erreurs sur la localisation du mobile et sur l'estimation de la carte.

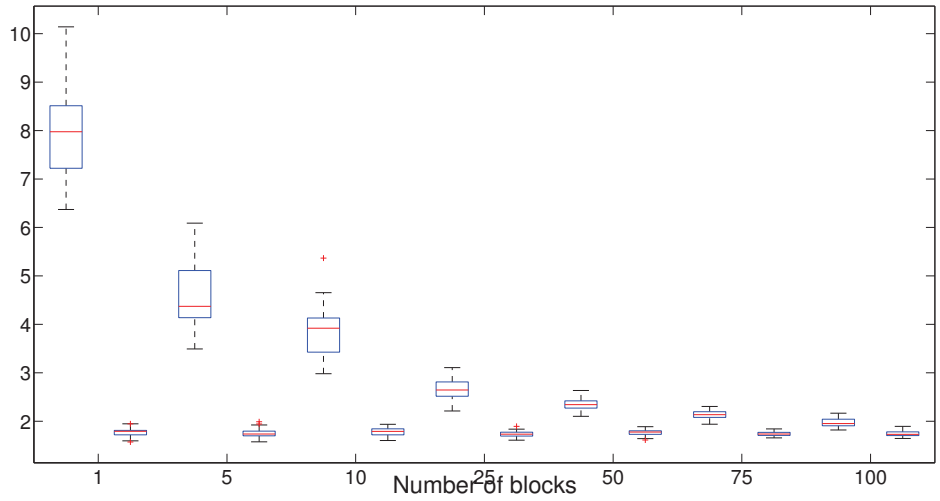


FIGURE 5.7 – Erreur de localisation donnée par l'algorithme stabilisé et erreur optimale après le traitement de 1, 5, 10, 25, 50, 75 et 100 blocs d'observations.

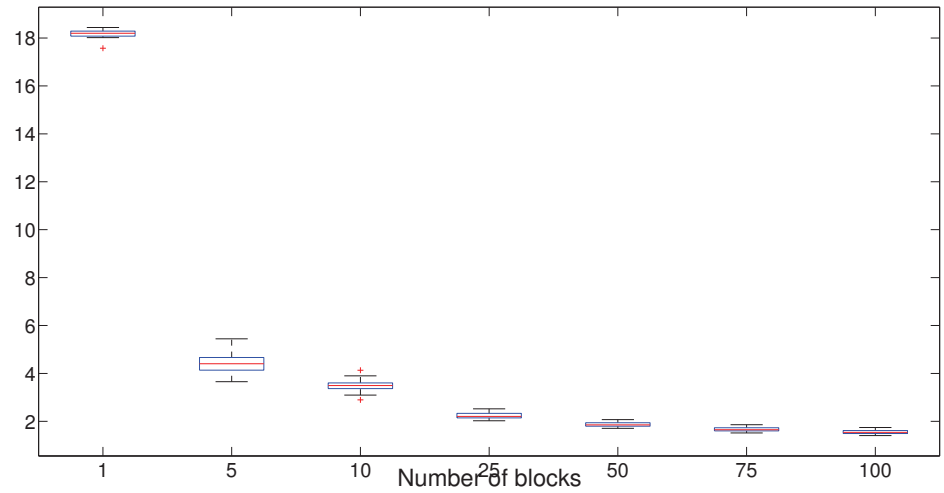


FIGURE 5.8 – Erreur L_1 sur l'estimation de la carte donnée par l'algorithme stabilisé et avec l'estimateur moyennisé après le traitement de 1, 5, 10, 25, 50, 75 et 100 blocs d'observations.

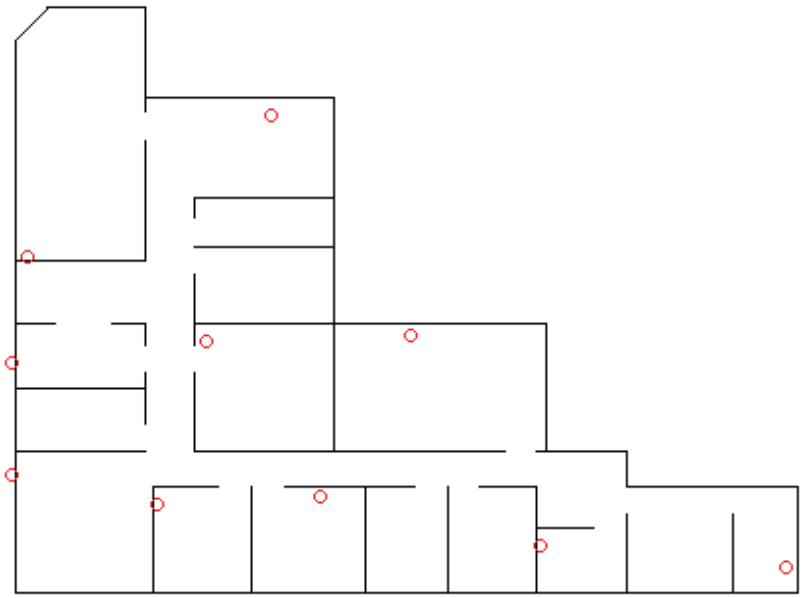


FIGURE 5.9 – Carte de l'environnement intérieur utilisé avec la position de toutes les bornes (cercles rouges).

Cette carte est discrétisée en utilisant une grille $\mathcal{C} = [0, 30] \times [0, 30]$. Dans ce cas la variance σ^2 est supposée connue et sa valeur ($\sigma^2 = 25$) a été calibrée par une série de mesures. Environ $T = 20000$ mesures ont été effectuées. Nous n'avons pas connaissance de la vraie position du mobile lors de l'acquisition de ces observations. On construit donc un échantillon de test en effectuant d'autres mesures tout en notant la vraie position du mobile lors de leur acquisition. Cet échantillon contient $T_{\text{test}} = 1100$ mesures : $\{X_t^{\text{test}}, Y_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$. Il sert à comparer la précision de la localisation effectuée à partir de différentes estimations de la carte obtenues à l'aide des 20000 premières mesures.

Il est important de noter une différence majeure avec le modèle présenté en Section 5.2.1 : pour chaque mesure, seules certaines bornes sont présentes dans Y_t, on ne dispose pas d'observations pour toutes les bornes à chaque instant. Nous supposons que nous savons identifier les bornes ayant émis les signaux reçus à chaque instant. Ainsi, on n'estime pas toutes les cartes simultanément. Nous introduisons alors des tailles de blocs et des compteurs de mesures spécifiques pour chaque borne.

Nous appliquons l'algorithme Monte Carlo BOEM avec procédure de stabilisation sur les 20000 mesures. A chaque fois qu'une estimation $\widehat{G}_j$ pour un $j \in \{1, \ldots, B\}$ est mise à jour (c'est-à-dire à chaque fois que l'on a parcouru toutes les observations d'un bloc associé à une borne), nous soumettons le

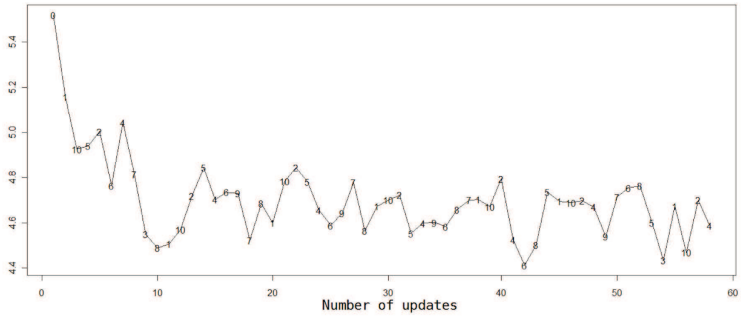


FIGURE 5.10 – Evolution du 0.8-quantile de la distance entre la vraie position et la position estimée. L'erreur de localisation est calculée sur l'échantillon de test chaque fois que l'une des cartes est estimée.

nouvel estimateur $\widehat{G}$ à un test avec les données $\{Y_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$. On utilise un système de particules propagé avec ces observations $\{Y_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$, $\vartheta_t = 1$, $p_t = m$ et où les poids sont calculés en fixant la carte à la valeur $\widehat{G}$ (voir l'Annexe 5.3 pour les détails sur la propagation des systèmes de particules). L'erreur de localisation est alors donnée par le 0.8-quantile empirique de la distance entre l'estimation fournie par ces particules et $\{X_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$. La Figure 5.10 représente cette erreur de localisation en fonction du nombre de mises à jour. Les numéros indiqués sur la Figure 5.10 indiquent quelles bornes ont servi à la mise à jour de l'estimation de la carte.

Malgré la petite taille de l'échantillon de test, la Figure 5.10 montre le bon comportement de l'erreur de localisation. Les cartes G_j ont été mises à jour entre 2 fois (la borne $j = 3$ n'est observée que 300 fois dans l'échantillon de test) et 7 fois ($j = 10$).

La Figure 5.11 représente l'estimation finale des cartes de propagation G_j, $j \in \{1, \cdots, 10\}$. Il est intéressant de remarquer que certaines de ces cartes (par exemple pour les bornes 1, 4 et 7) font apparaître la position des obstacles (les murs) responsables des perturbations.

5.3 Annexe : Contrôle de l'approximation particulaire

Nous commençons par rappeler l'expression de la quantité intermédiaire que nous cherchons à approcher. L'algorithme BOEM produit une suite d'estimations $\{\theta_n\}_{n\geq 0}$ en utilisant, pour tout $n \geq 0$, l'espérance $\bar{S}_{\tau_{n+1}}^{\chi_n, T_n}(\theta_n, \mathbf{Y})$

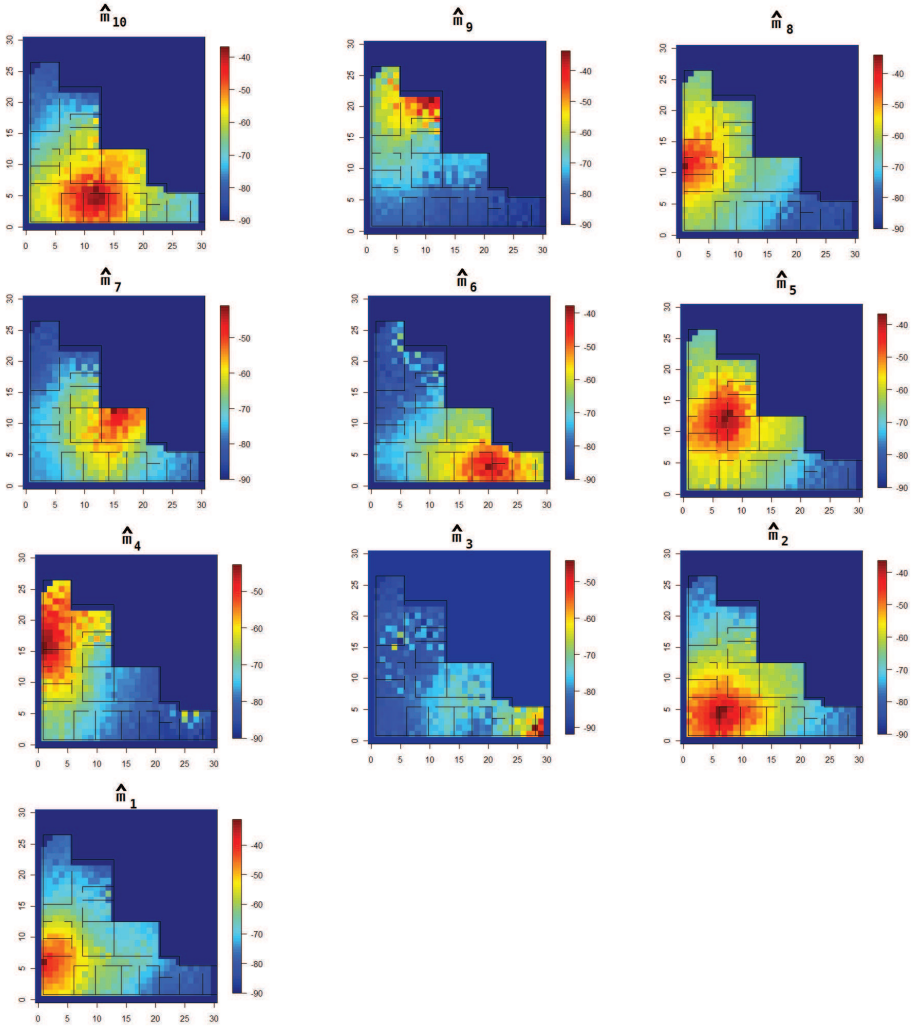


FIGURE 5.11 – Représentation graphique des cartes de propagation estimées par l'algorithme Monte Carlo BOEM moyennisé : la carte estimée associée à la borne j est donnée sur le graphe $\widehat{m}_j$.

définie par

$$\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y}) = \frac{1}{\tau_{n+1}}\sum_{k=1}^{\tau_{n+1}}\mathbb{E}_{\theta_n}^{\chi_n,T_n}\left[S(X_{T_n+k-1},X_{T_n+k},Y_{T_n+k})\middle|Y_{T_n+1:T_n+\tau_{n+1}}\right] ,$$

où χ_n est une loi sur $(\mathbb{X},\mathcal{X})$ et où $\mathbb{E}_{\theta_n}^{\chi_n,T_n}\left[\cdot\middle|Y_{T_n+1:T_n+\tau_{n+1}}\right]$ est définie par (1.2). $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y})$ est la quantité intermédiaire de l'algorithme BOEM et correspond à la quantité intermédiaire de l'algorithme EM, calculée avec les observations $Y_{T_n+1:T_n+\tau_{n+1}}$ du bloc n, lorsque les lois régissant le HMM sont paramétrées par θ_n et lorsque l'état initial est distribué sous la loi χ_n.

Cette quantité n'est calculable de façon explicite que dans certaines situations spécifiques : si l'espace d'état est fini ou lorsque l'on considère des modèles linéaires et gaussiens. Dans des cas plus généraux nous proposons au Chapitre 6 de remplacer cette quantité par une approximation Monte Carlo. Nous considérons ici la situation où $\bar{S}_{\tau_{n+1}}^{\chi_n,T_n}(\theta_n,\mathbf{Y})$ est approchée par l'approximation particulaire de l'algorithme FFBS calculée avec N_{n+1} particules, notée $\widetilde{S}_n$ (voir la Section 3.1.2). Nous effectuons ce choix pour deux raisons :

i) Le calcul de $\widetilde{S}_n$ avec l'algorithme FFBS est réalisable en ligne (sans stocker d'observations), comme cela est précisé en Section 2.3.

ii) On peut grâce au Chapitre 7 contrôler l'erreur L_p effectuée sur chaque bloc. Cela nous permet de choisir le nombre de particules par bloc pour vérifier les conditions suffisantes de convergence de l'algorithme Monte Carlo BOEM (voir l'hypothèse A6 du Chapitre 6).

Calcul en ligne

Pour plus de clarté, nous redonnons ici le mécanisme permettant d'obtenir l'approximation FFBS en ligne au sein du bloc n. On note η_n la loi de proposition des particules à l'instant T_n et $\{\vartheta_t^n\}_{t\leq\tau_{n+1}}$ et $\{p_t^n\}_{t\leq\tau_{n+1}}$ les poids d'ajustement et les noyaux de proposition utilisés dans notre algorithme FFBS (voir la Section 3.1.2 pour plus de précisions).

1) **Initialisation.**
Pour tout $\ell\in\{1,\cdots,N_{n+1}\}$, simuler de façon indépendante $\xi_0^{N,\ell}\sim\eta_n$ et définir $\omega_0^{N,\ell}\overset{\text{def}}{=}\frac{\mathrm{d}\chi_n}{\mathrm{d}\eta_n}(\xi_0^{N,\ell})g_{\theta_n}(\xi_0^{N,\ell},Y_{T_n})$.
Poser, pour tout $\ell\in\{1,\cdots,N_{n+1}\}$, $\rho_0(\xi_0^{N,\ell})\overset{\text{def}}{=}0$.

2) **Propagation : pour** $t\in\{0,\cdots,\tau_{n+1}-1\}$
Pour tout $\ell\in\{1,\cdots,N_{n+1}\}$, simuler $(I_{t+1}^{N_{n+1},\ell},\xi_{t+1}^{N_{n+1},\ell})$ suivant la loi instrumentale

$$\pi_{t+1|t+1}(i,\mathrm{d}x)\propto\omega_t^{N_{n+1},i}\vartheta_{t+1}^n(\xi_t^{N_{n+1},i})p_{t+1}^n(\xi_t^{N_{n+1},i},\mathrm{d}x) ,$$

Poser

$$\omega_{t+1}^{N_{n+1},\ell} \stackrel{\text{def}}{=} \frac{m_{\theta_n}(\xi_t^{N_{n+1},I_{t+1}^{N_{n+1},\ell}}, \xi_{t+1}^{N_{n+1},\ell}) g_{\theta_n}(\xi_{t+1}^{N_{n+1},\ell}, Y_{T_n+t+1})}{\vartheta_{t+1}^n(\xi_t^{N_{n+1},\ell}) p_{t+1}^n(\xi_t^{N_{n+1},I_{t+1}^{N_{n+1},\ell}}, \xi_{t+1}^{N_{n+1},\ell})} .$$

Poser, pour tout $\ell \in \{1, \cdots, N_{n+1}\}$,

$$\rho_{t+1}(\xi_t^{N_{n+1},\ell}) \stackrel{\text{def}}{=} \sum_{i=1}^{N_{n+1}} \left[\frac{1}{t+1} S(\xi_t^{N_{n+1},i}, \xi_{t+1}^{N_{n+1},\ell}, Y_{T_n+t+1}) + \left(1 - \frac{1}{t+1}\right) \rho_t(\xi_t^{N_{n+1},i}) \right] \times \frac{\omega_t^{N_{n+1},i} m_{\theta_n}(\xi_t^{N_{n+1},i}, \xi_{t+1}^{N_{n+1},\ell})}{\sum_{j=1}^{N_{n+1}} \omega_t^{N_{n+1},j} m_{\theta_n}(\xi_t^{N_{n+1},j}, \xi_{t+1}^{N_{n+1},\ell})} .$$

3) **Calcul de l'approximation.**
Définir

$$\widetilde{S}_n \stackrel{\text{def}}{=} \sum_{\ell=1}^{N_{n+1}} \omega_{\tau_{n+1}}^{N_{n+1},\ell} \rho_{\tau_{n+1}}(\xi_{\tau_{n+1}}^{N_{n+1},\ell}) .$$

L'approximation donnée par $\widetilde{S}_n$ est exactement celle fournie par l'algorithme FFBS introduite au Chapitre 3. Ceci nous assure que, dans le cas qui nous préoccupe, l'approximation particulaire donnée par l'algorithme FFBS peut se calculer en ligne. Elle ne nécessite donc pas d'effectuer un *path-space smoother* suivi d'un parcours des données à l'envers (de la dernière observation du bloc jusqu'à la première). D'autre part, les contrôles donnés pour l'algorithme FFBS sont applicables à notre approximation $\widetilde{S}_n$.

Contrôle de l'erreur d'approximation

Il nous reste maintenant à utiliser les résultats du Chapitre 7 pour obtenir un contrôle de l'erreur plus précis, faisant intervenir les observations dans les bornes (dans le Chapitre 7 nous travaillons conditionnellement à un jeu d'observations fixé).

Ces contrôles s'obtiennent en suivant les mêmes étapes que pour les preuves des propositions 7.1 et 7.2 du Chapitre 7. Nous ne donnons ici que le résultat, toutes les preuves étant détaillées par ailleurs dans l'article [Le Corff et Fort, 2011a]. Les preuves en question nécessitent l'introduction de nouvelles quantités liées aux observations ainsi que des hypothèses qui leur sont rattachées. Pour tout $y \in \mathbb{Y}$, on définit

$$\omega_+(y) = \sup_{\theta\in\Theta} \sup_{\substack{(x,x')\in\mathbb{X}\times\mathbb{X} \\ t\geq 0, n\geq 0}} \frac{m_\theta(x,x') g_\theta(x',y)}{\vartheta_t^n(x) p_t^n(x,x')}$$

et

$$b_-(y) \stackrel{\text{def}}{=} \inf_{\theta\in\Theta} \int g_\theta(x,y) \lambda(\mathrm{d}x) .$$

Pour effectuer ces contrôles, nous avons besoin

i) d'hypothèses sur le modèle HMM et sur le mécanisme de production des particules (similaires à celles données au Chapitre 7),

ii) d'hypothèses sur les observations (hypothèses de stationnarité et de contrôle de moments faisant intervenir les fonctions b_- et ω_+).

Sous ces hypothèses, on peut alors montrer qu'il existe $p > 2$ et une constante $C > 0$ tels que pour tout $n \geq 0$,

$$\left\| \widetilde{S}_n - \bar{S}^{\chi_n, T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) \right\|_p \leq C \left(\frac{1}{\tau_{n+1}^{1/2} N_{n+1}^{1/2}} + \frac{1}{N_{n+1}} \right) .$$

Ceci nous permet donc d'avoir un contrôle explicite de l'erreur L_p effectuée sur chaque bloc en fonction de la taille du bloc et du nombre de particules utilisées pour effectuer l'approximation FFBS. Si, comme au Chapitre 6 nous choisissons un nombre d'observations par bloc de la forme $\tau_n = \lfloor cn^a \rfloor$ avec $c > 0$ et $a > 1$, alors il est suffisant de choisir N_n de la forme $N_n = b\tau_n^d$ avec $d \geq (a+1)/2a$ pour obtenir l'hypothèse A6 du Chapitre 6 et avoir la convergence de l'algorithme Monte Carlo BOEM.

Chapitre 6

Algorithmes de type Expectation-Maximization en ligne pour l'estimation dans les modèles de Markov cachés (article)

The Expectation Maximization (EM) algorithm is a versatile tool for model parameter estimation in latent data models. When processing large data sets or data stream however, EM becomes intractable since it requires the whole data set to be available at each iteration of the algorithm. In this contribution, a new generic online EM algorithm for model parameter inference in general Hidden Markov Model is proposed. This new algorithm updates the parameter estimate after a block of observations is processed (online). The convergence of this new algorithm is established, and the rate of convergence is studied showing the impact of the block-size sequence. An averaging procedure is also proposed to improve the rate of convergence. Finally, practical illustrations are presented to highlight the performance of these algorithms in comparison to other online maximum likelihood procedures.

6.1 Introduction

A hidden Markov model (HMM) is a stochastic process $\{X_k, Y_k\}_{k\geq 0}$ in $\mathbb{X} \times \mathbb{Y}$, where the state sequence $\{X_k\}_{k\geq 0}$ is a Markov chain and where the observations $\{Y_k\}_{k\geq 0}$ are independent conditionally on $\{X_k\}_{k\geq 0}$. Moreover, the conditional distribution of Y_k given the state sequence depends only

on X_k. The sequence $\{X_k\}_{k\geq 0}$ being unobservable, any statistical inference task is carried out using the observations $\{Y_k\}_{k\geq 0}$. These HMM can be applied in a large variety of disciplines such as financial econometrics ([Mamon et Elliott, 2007]), biology ([Churchill, 1992]) or speech recognition ([Juang et Rabiner, 1991]).

The Expectation Maximization (EM) algorithm is an iterative algorithm used to solve maximum likelihood estimation in HMM. The EM algorithm is generally simple to implement since it relies on complete data computations. Each iteration is decomposed into two steps: the E-step computes the conditional expectation of the complete data log-likelihood given the observations and the M-step updates the parameter estimate based on this conditional expectation. In many situations of interest, the complete data likelihood belongs to the curved exponential family. In this case, the E-step boils down to the computation of the conditional expectation of the complete data sufficient statistic. Even in this case, except for simple models such as linear Gaussian models or HMM with finite state-spaces, the E-step is intractable and has to be approximated e.g. by Monte Carlo methods such as Markov Chain Monte Carlo methods or Sequential Monte Carlo methods (see [Carlin *et al.*, 1992] or [Cappé *et al.*, 2005, Doucet *et al.*, 2001] and the references therein).

However, when processing large data sets or data streams, the EM algorithm might become impractical. *Online* variants of the EM algorithm have been first proposed for independent and identically distributed (i.i.d.) observations, see [Cappé et Moulines, 2009]. When the complete data likelihood belongs to the cruved exponential family, the E-step is replaced by a stochastic approximation step while the M-step remains unchanged. The convergence of this online variant of the EM algorithm for i.i.d. observations is addressed by [Cappé et Moulines, 2009]: the limit points are the stationary points of the Kullback-Leibler divergence between the marginal distribution of the observation and the model distribution.

An online version of the EM algorithm for HMM when both the observations and the states take a finite number of values (resp. when the states take a finite number of values) was recently proposed by [Mongillo et Denève, 2008] (resp. by [Cappé, 2011a]). This algorithm has been extended to the case of general state-space models by substituting deterministic approximation of the smoothing probabilities for Sequential Monte Carlo algorithms (see for example [Cappé, 2009, Del Moral *et al.*, 2010a, Le Corff *et al.*, 2011b]). There do not exist convergence results for these online EM algorithms for general state-space models (some insights on the asymptotic behavior are nevertheless given in [Cappé, 2011a]): the introduction of many approximations at different steps of the algorithms makes the analysis quite challenging.

In this contribution, a new online EM algorithm is proposed for HMM with complete data likelihood belonging to the curved exponential family.

This algorithm sticks closely to the principles of the original batch-mode EM algorithm. The M-step (and thus, the update of the parameter) occurs at some deterministic times $\{T_k\}_{k\geq 1}$ i.e. we propose to keep a fixed parameter estimate for blocks of observations of increasing size. More precisely, let $\{T_k\}_{k\geq 0}$ be an increasing sequence of integers ($T_0 = 0$). For each $k \geq 0$, the parameter's value is kept fixed while accumulating the information brought by the observations $\{Y_{T_k+1}, \cdots, Y_{T_{k+1}}\}$. Then, the parameter is updated at the end of the block. This algorithm is an online algorithm since the sufficient statistics of the k-th block can be computed on the fly by updating an intermediate quantity when a new observation Y_t, $t \in \{T_k + 1, \ldots, T_{k+1}\}$ becomes available. Such recursions are provided in recent works on online estimation in HMM, see [Cappé, 2009, Cappé, 2011a, Del Moral *et al.*, 2010a].

This new algorithm, called *Block Online EM* (BOEM) is derived in Section 6.2 together with an *averaged* version. Section 6.3 is devoted to practical applications: the BOEM algorithm is used to perform parameter inference in HMM where the forward recursions mentioned above are available explicitly. In the case of finite state-space HMM, the BOEM algorithm is compared to a gradient-type recursive maximum likelihood procedure and to the online EM algorithm of [Cappé, 2011a]. The convergence of the BOEM algorithm is addressed in Section 6.4. The BOEM algorithm is seen as a perturbation of a deterministic *limiting EM* algorithm which is shown to converge to the stationary points of the limiting relative entropy (to which the true parameter belongs if the model is well specified). The perturbation is shown to vanish (in some sense) as the number of observations increases thus implying that the BOEM algorithms inherits the asymptotic behavior of the *limiting EM* algorithm. Finally, in Section 6.5, we study the rate of convergence of the BOEM algorithm as a function of the block-size sequence. We prove that the averaged BOEM algorithm is rate-optimal when the block-size sequence grows polynomially. All the proofs are postponed to Section 6.6; supplementary proofs and comments are provided in Appendix A.

6.2 The Block Online EM algorithms

6.2.1 Notations and Model assumptions

Our model is defined as follows. Let Θ be a compact subset of $\mathbb{R}^{d_\theta}$. We are given a family of transition kernels $\{M_\theta\}_{\theta\in\Theta}$, $M_\theta : \mathbb{X} \times \mathcal{X} \to [0, 1]$, a positive σ-finite measure μ on $(\mathbb{Y}, \mathcal{Y})$, and a family of transition densities with respect to μ, $\{g_\theta\}_{\theta\in\Theta}$, $g_\theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}_+$. For each $\theta \in \Theta$, define the

transition kernel K_θ on $\mathbb{X} \times \mathbb{Y}$ by

$$K_\theta\left[(x,y),C\right] \stackrel{\text{def}}{=} \int \mathbf{1}_C(x',y')\, g_\theta(x',y')\, \mu(\mathrm{d}y')\, M_\theta(x,\mathrm{d}x') \ .$$

Denote by $\{X_k, Y_k\}_{k\geq 0}$ the canonical coordinate process on the measurable space $\left((\mathbb{X}\times\mathbb{Y})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes\mathbb{N}}\right)$. For any $\theta \in \Theta$ and any probability distribution χ on $(\mathbb{X},\mathcal{X})$, let $\mathbb{P}_\theta^\chi$ be the probability distribution on $((\mathbb{X}\times\mathbb{Y})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes\mathbb{N}})$ such that $\{X_k,Y_k\}_{k\geq 0}$ is Markov chain with initial distribution $\mathbb{P}_\theta^\chi((X_0,Y_0)\in C) = \int \mathbf{1}_C(x,y)\, g_\theta(x,y)\, \mu(\mathrm{d}y)\, \chi(\mathrm{d}x)$ and transition kernel K_θ. The expectation with respect to $\mathbb{P}_\theta^\chi$ is denoted by $\mathbb{E}_\theta^\chi$. Throughout this paper, it is assumed that the Markov transition kernel K_θ has a unique invariant distribution π_θ (see below for further comments). For the stationary Markov chain with initial distribution π_θ, we write $\mathbb{P}_\theta$ and $\mathbb{E}_\theta$ instead of $\mathbb{P}_\theta^{\pi_\theta}$ and $\mathbb{E}_\theta^{\pi_\theta}$. Note also that the stationary Markov chain $\{X_k,Y_k\}_{k\geq 0}$ can be extended to a two-sided Markov chain $\{X_k,Y_k\}_{k\in\mathbb{Z}}$.

It is assumed that, for any $\theta\in\Theta$ and any $x\in\mathbb{X}$, $M_\theta(x,\cdot)$ has a density $m_\theta(x,\cdot)$ with respect to a finite measure λ on $(\mathbb{X},\mathcal{X})$. Define the complete data likelihood by

$$p_\theta(x_{0:T},y_{0:T}) \stackrel{\text{def}}{=} g_\theta(x_0,y_0) \prod_{i=0}^{T-1} m_\theta(x_i,x_{i+1}) g_\theta(x_{i+1},y_{i+1}) \ , \tag{6.1}$$

where, for any $u\leq s$, we will use the shorthand notation $x_{u:s}$ for the sequence $(x_u,\cdots,x_s)$. For any probability distribution χ on $(\mathbb{X},\mathcal{X})$, any $\theta\in\Theta$ and any $s\leq u\leq v\leq t$, we have

$$\mathbb{E}_\theta^\chi\left[f(X_{u:v})|Y_{s:t}\right] = \int f(x_{u:v})\phi^\chi_{\theta,u:v|s:t}(\mathrm{d}x_{u:v}) \ ,$$

where $\phi^\chi_{\theta,u:v|s:t}$ is the so-called fixed-interval smoothing distribution. We also define the fixed-interval smoothing distribution when $X_s\sim\chi$:

$$\mathbb{E}_\theta^{\chi,s}\left[f(X_{u:v})|Y_{s+1:t}\right] = \frac{\int \prod_{i=s+1}^t \{m_\theta(x_{i-1},x_i) g_\theta(x_i,Y_i)\} f(x_{u:v}) \chi(\mathrm{d}x_s)\lambda(\mathrm{d}x_{s+1:t})}{\int \prod_{i=s+1}^t \{m_\theta(x_{i-1},x_i) g_\theta(x_i,Y_i)\} \chi(\mathrm{d}x_s)\lambda(\mathrm{d}x_{s+1:t})} \ .$$

Given an initial distribution χ on $(\mathbb{X},\mathcal{X})$ and $T+1$ observations $Y_{0:T}$, the EM algorithm maximizes the so-called incomplete data log-likelihood $\theta\mapsto \ell^\chi_{\theta,T}$ defined by

$$\ell^\chi_{\theta,T}(\mathbf{Y}) \stackrel{\text{def}}{=} \log \int p_\theta(x_{0:T},Y_{1:T})\chi(\mathrm{d}x_0)\lambda(\mathrm{d}x_{1:T}) \ . \tag{6.2}$$

The central concept of the EM algorithm is that the intermediate quantity defined by

$$\theta\mapsto Q(\theta,\theta') \stackrel{\text{def}}{=} \mathbb{E}^\chi_{\theta'}\left[\log p_\theta(X_{0:T},Y_{0:T})|Y_{0:T}\right]$$

may be used as a surrogate for $\ell^{\chi}_{\theta,T}(Y_{0:T})$ in the maximization procedure. Therefore, the EM algorithm iteratively builds a sequence $\{\theta_n\}_{n\geq 0}$ of parameter estimates following the two steps:

i) Compute $\theta \mapsto Q(\theta, \theta_n)$.

ii) Choose θ_{n+1} as a maximizer of $\theta \mapsto Q(\theta, \theta_n)$.

In the sequel, it is assumed that there exist functions S, ϕ and ψ such that (see A1 for a more precise definition), for any $(x, x') \in \mathbb{X}^2$ and any $y \in \mathbb{Y}$,

$$m_\theta(x, x')g_\theta(x', y) = \exp\left\{\phi(\theta) + \left\langle S(x, x,', y), \psi(\theta)\right\rangle\right\} \ .$$

Therefore, the complete data likelihood belongs to the curved exponential family and the step i) of the EM algorithm amounts to computing

$$\theta \mapsto Q(\theta, \theta_n) = \phi(\theta) + \left\langle \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}^{\chi}_{\theta_n}\left[S(X_{t-1}, X_t, Y_t)|Y_{0:T}\right], \psi(\theta)\right\rangle \ ,$$

where $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbb{R}^d$. It is also assumed that for any $s \in \mathcal{S}$, where $\mathcal{S}$ is an appropriately defined set, the function $\theta \mapsto \phi(\theta) + \langle s, \psi(\theta)\rangle$ has a unique maximum denoted by $\bar{\theta}(s)$. Hence, a step of the EM algorithm writes

$$\theta_n = \bar{\theta}\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}^{\chi}_{\theta_{n-1}}\left[S(X_{t-1}, X_t, Y_t)|Y_{0:T}\right]\right) \ .$$

6.2.2 The Block Online EM (BOEM) algorithms

We now derive an online version of the EM algorithm. Define $\bar{S}^{\chi,T}_{\tau}(\theta, \mathbf{Y})$ as the intermediate quantity of the EM algorithm computed with the observations $Y_{T+1:T+\tau}$:

$$\bar{S}^{\chi,T}_{\tau}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau}\sum_{t=T+1}^{T+\tau} \mathbb{E}^{\chi,T}_{\theta}\left[S(X_{t-1}, X_t, Y_t)|Y_{T+1:T+\tau}\right] \ . \tag{6.3}$$

Let $\{\tau_n\}_{n\geq 1}$ be a sequence of positive integers such that $\lim_{n\to\infty} \tau_n = +\infty$ and set

$$T_n \stackrel{\text{def}}{=} \sum_{k=1}^{n} \tau_k \quad \text{and} \quad T_0 \stackrel{\text{def}}{=} 0 \ ; \tag{6.4}$$

τ_n denotes the length of the n-th block. Given an initial value $\theta_0 \in \Theta$, the BOEM algorithm defines a sequence $\{\theta_n\}_{n\geq 1}$ by

$$\theta_n \stackrel{\text{def}}{=} \bar{\theta}\left[S_{n-1}\right] \ , \text{ and } S_{n-1} \stackrel{\text{def}}{=} \bar{S}^{\chi_{n-1},T_{n-1}}_{\tau_n}(\theta_{n-1}, \mathbf{Y}) \ , \tag{6.5}$$

where $\{\chi_n\}_{n\geq 0}$ is a family of probability distributions on $(\mathbb{X}, \mathcal{X})$. By analogy to the regression problem, an estimator with reduced variance can be

obtained by averaging and weighting the successive estimates (see for example [Kushner et Yin, 1997, Polyak et Juditsky, 1992] for a discussion on the averaging procedures). Define $\Sigma_0 \stackrel{\text{def}}{=} 0$ and for $n \geq 1$,

$$\Sigma_n \stackrel{\text{def}}{=} \frac{1}{T_n} \sum_{j=1}^{n} \tau_j \, S_{j-1} \, . \tag{6.6}$$

Note that this quantity can be computed iteratively and does not require to store the past statistics $\{S_j\}_{j=0}^{n-1}$. Given an initial value $\widetilde{\theta}_0$, the averaged BOEM algorithm defines a sequence $\{\widetilde{\theta}_n\}_{n\geq 1}$ by

$$\widetilde{\theta}_n \stackrel{\text{def}}{=} \bar{\theta}\left(\Sigma_n\right) \, . \tag{6.7}$$

The algorithm above relies on the assumption that S_n can be computed in closed form. In the HMM case, this property is satisfied only for linear Gaussian models or when the state-space is finite. In all other cases, S_n cannot be computed explicitly and will be replaced by a Monte Carlo approximation $\widetilde{S}_n$. Several Monte Carlo approximations can be used to compute $\widetilde{S}_n$. The convergence properties of the Monte Carlo BOEM algorithms rely on the assumption that the Monte Carlo error can be controlled on each block. [Le Corff et Fort, 2011a] provides examples of applications when Sequential Monte Carlo algorithms are used. Hereafter, we use the same notation $\{\theta_n\}_{n\geq 0}$ and $\{\widetilde{\theta}_n\}_{n\geq 0}$ for the original BOEM algorithm or its Monte Carlo approximation.

Our algorithms update the parameter after processing a block of observations. Nevertheless, the intermediate quantity S_n can be either exactly computed or approximated in such a way that the observations are processed online. In this case, the intermediate quantity S_n or $\widetilde{S}_n$ is updated online for each observation. Such an algorithm is described in [Cappé, 2011a, Section 2.2] and [Del Moral *et al.*, 2010b, Proposition 2.1] and can be applied either to finite state-space HMM or to linear Gaussian models. A Sequential Monte Carlo approximation to compute $\widetilde{S}_n$ online for more complex models is proposed in [Del Moral *et al.*, 2010a] (see also [Le Corff et Fort, 2011a]).

The classical theory of maximum likelihood estimation often relies on the assumption that the "true" distribution of the observations belongs to the specified parametric family of distributions. In many cases, it is doubtful that this assumption is satisfied. It is therefore natural to investigate the convergence of the BOEM algorithms and to identify the possible limit for misspecified models i.e. when the observations $\mathbf{Y}$ are from an ergodic process which is not necessarily an HMM.

6.3 Application to inverse problems in Hidden Markov Models

In Section 6.3.1, the performance of the BOEM algorithm and its averaged version are illustrated in a truncated linear Gaussian model. In Section 6.3.2, the BOEM algorithm is compared to online maximum likelihood procedures in the case of finite state-space HMM.

Other applications of the Monte Carlo BOEM algorithm to more complex models with online Sequential Monte Carlo methods can be found in [Le Corff et Fort, 2011a].

6.3.1 Linear Gaussian Model

Consider the linear Gaussian model:

$$X_{t+1} = \phi X_t + \sigma_u U_t \, , \qquad Y_t = X_t + \sigma_v V_t \, ,$$

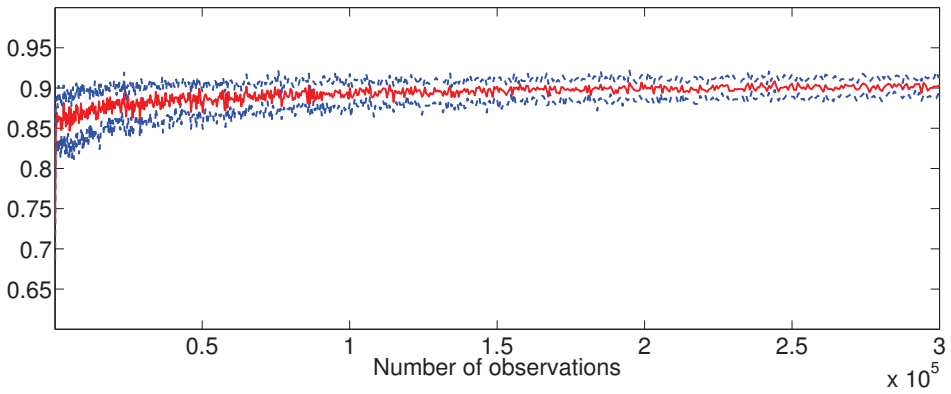
where $X_0 \sim \mathcal{N}\left(0, \sigma_u^2(1-\phi^2)^{-1}\right)$, $\{U_t\}_{t\geq 0}, \{V_t\}_{t\geq 0}$ are independent i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.9$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$.

We illustrate the convergence of the BOEM algorithms. We choose $\tau_n = n^{1.1}$. We display in Figure 6.1 the median and lower and upper quartiles for the estimation of ϕ obtained with 100 independent Monte Carlo experiments. Both the BOEM algorithm and its averaged version converge to the true value $\phi = 0.9$; the averaging procedure clearly improves the variance of the estimation.

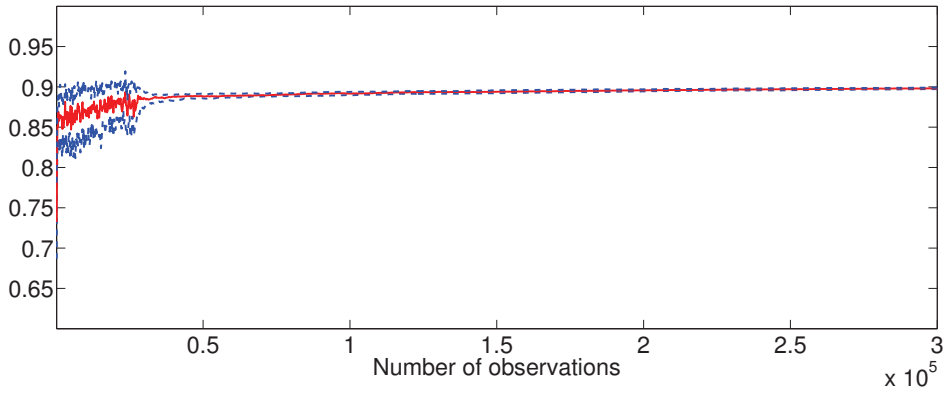
We now discuss the role of $\{\tau_n\}_{n\geq 0}$. Figure 6.2 displays the empirical variance, when estimating ϕ, computed with 100 independent Monte Carlo runs, for different numbers of observations and, for both the BOEM algorithm and its averaged version. We consider four polynomial rates $\tau_n \sim n^b$, $b \in \{1.2, 1.8, 2, 2.5\}$. Figure 6.2a shows that the choice of $\{\tau_n\}_{n\geq 0}$ has a great impact on the empirical variance of the (non averaged) BOEM path $\{\theta_n\}_{n\geq 0}$. To reduce this variability, a solution could consist in increasing the block sizes τ_n at a larger. The influence of the block size sequence τ_n is greatly reduced with the averaging procedure as shown in Figure 6.2b. We will show in Section 6.5 that averaging really improves the rate of convergence of the BOEM algorithm.

6.3.2 Finite state-space HMM

We consider a Gaussian mixture process with Markov dependence of the form: $Y_t = X_t + V_t$ where $\{X_t\}_{t\geq 0}$ is a Markov chain taking values in $\mathbb{X} \stackrel{\text{def}}{=} \{x_1, \ldots, x_d\}$, with initial distribution χ and a $d \times d$ transition matrix m. $\{V_t\}_{t\geq 0}$ are i.i.d. $\mathcal{N}(0, v)$ r.v., independent from $\{X_t\}_{t\geq 0}$, i.e., for all



(a) The BOEM algorithm without averaging.



(b) The BOEM algorithm with averaging.

Figure 6.1: Estimation of ϕ.

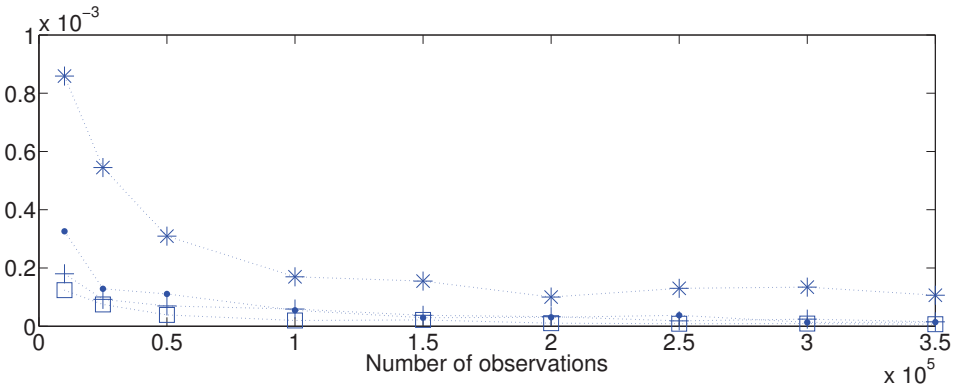
$(x, y) \in \mathbb{X} \times \mathbb{Y}$,

$$g_\theta(x, y) \stackrel{\text{def}}{=} (2\pi v)^{-1/2} \exp\left\{-\frac{(y-x)^2}{2v}\right\} ,$$

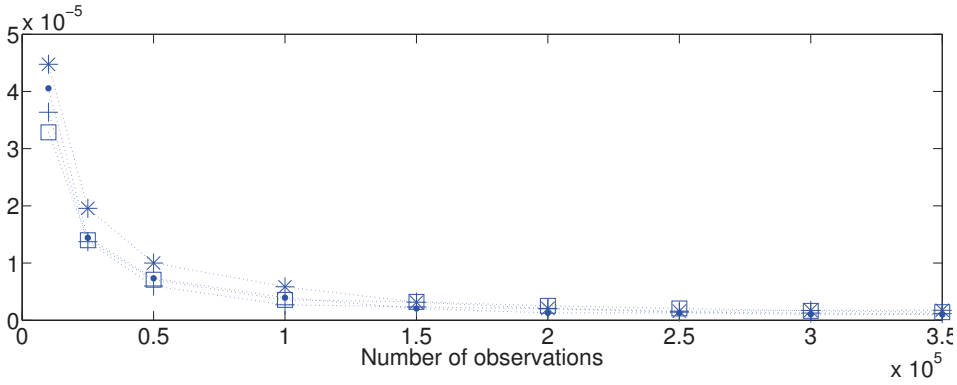
where $\theta \stackrel{\text{def}}{=} \left(v, x_{1:d}, (m_{i,j})_{i,j=1}^d\right)$. In the experiments below, the initial distribution below is chosen as the uniform distribution on $\mathbb{X}$. The statistics used to estimate θ are, for all $(i, j) \in \{1, \cdots, d\}$ and all $(x, x') \in \mathbb{X}^2$,

$$\begin{aligned} S^{i,0}(x, x', y) &= \mathbf{1}_{x_i}(x') , & S^{i,1}(x, x', y) &= y\mathbf{1}_{x_i}(x') , \\ S^{i,2}(x, x', y) &= y^2\mathbf{1}_{x_i}(x') , & S_{i,j}(x, x', y) &= \mathbf{1}_{x_i}(x)\mathbf{1}_{x_j}(x') . \end{aligned} \tag{6.8}$$

The online computation of these intermediate quantities can be found in [Cappé, 2011a, Section 2.2]. The computations below are performed for each statistic in (6.8). Define, for all $x \in \mathbb{X}$, $\phi_0(x) = \chi(x)$ and $\rho_0(x) = 0$.



(a) The BOEM algorithm, without averaging



(b) The BOEM algorithm, with averaging

Figure 6.2: The BOEM algorithm: empirical variance of the estimation of ϕ after $n = 0.5\ell \cdot 10^5$ observations ($\ell \in \{1, \cdots, 7\}$) for different block size schemes $\tau_n \sim n^{1.2}$ (stars), $\tau_n \sim n^{1.8}$ (dots), $\tau_n \sim n^2$ (crosses) and $\tau_n \sim n^{2.5}$ (squares).

i) For $t \in \{1, \cdots, \tau\}$, compute, for any $x \in \mathbb{X}$,

$$\phi_t(x) = \frac{\sum_{x' \in \mathbb{X}} \phi_{t-1}(x') m_{x',x} g_\theta(x, Y_{t+T})}{\sum_{x',x'' \in \mathbb{X}} \phi_{t-1}(x') m_{x',x''} g_\theta(x'', Y_{t+T})} ,$$

and

$$r_t(x, x') = \frac{\phi_{t-1}(x') m_{x',x}}{\sum_{x'' \in \mathbb{X}} \phi_{t-1}(x'') m_{x'',x}} .$$

$$\rho_t(x) = \sum_{x' \in \mathbb{X}} \left[\frac{1}{t} S(x, x', Y_{t+T}) + \left(1 - \frac{1}{t}\right) \rho_{t-1}(x') \right] r_t(x, x') .$$

ii) Set

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) = \sum_{x \in \mathbb{X}} \rho_\tau(x) \phi_\tau(x) .$$

At the end of the block, the new estimate is given, for all $(i,j) \in \{1,\cdots,d\}^2$ by (the dependence on $\mathbf{Y}$, θ, χ, T and τ is dropped from the notation)

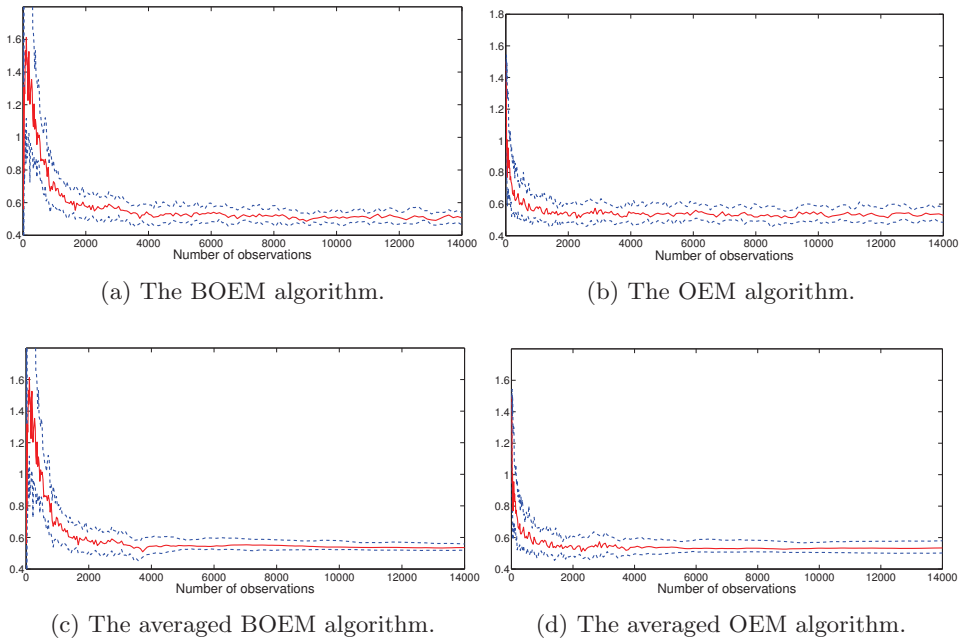
$$m_{i,j} = \frac{\bar{S}_{i,j}}{\sum_{j=1}^{d} \bar{S}_{i,j}} \ , \ x_i = \frac{\bar{S}^{i,1}}{\bar{S}^{i,0}}, \ v = \sum_{i=1}^{d} \bar{S}^{i,2} + \sum_{i=1}^{d} x_i^2 \bar{S}^{i,0} - 2\sum_{i=1}^{d} x_i \bar{S}^{i,1} \ .$$

Observations are sampled using $d = 6$, $v = 0.5$, $x_i = i \ , \forall i \in \{1,\ldots,d\}$ and the true transition matrix is given by

$$m = \begin{pmatrix} 0.5 & 0.05 & 0.1 & 0.15 & 0.15 & 0.05 \\ 0.2 & 0.35 & 0.1 & 0.15 & 0.05 & 0.15 \\ 0.1 & 0.1 & 0.6 & 0.05 & 0.05 & 0.1 \\ 0.02 & 0.03 & 0.1 & 0.7 & 0.1 & 0.05 \\ 0.1 & 0.05 & 0.13 & 0.02 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.13 & 0.12 & 0.1 & 0.45 \end{pmatrix} .$$

We first compare the averaged BOEM algorithm to the online EM (OEM) procedure of [Cappé, 2011a] combined with a Polyak-Ruppert averaging (see [Polyak et Juditsky, 1992]). Note that the convergence of the OEM algorithm is still an open problem. In this case, we want to estimate the variance v and the states $\{x_1,\ldots,x_d\}$. All the runs are started from $v = 2$ and from the initial states $\{-1;0;.5;2;3;4\}$. The algorithm in [Cappé, 2011a] follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$. It is expected that the rate of convergence in L_2 after n observations is $\gamma_n^{1/2}$ (and $n^{-1/2}$ for its averaged version) - this assertion relies on classical results for stochastic approximation. We prove in Section 6.5 that the rate of convergence of the BOEM algorithm is $n^{-b/(2(b+1))}$ (and $n^{-1/2}$ for its averaged version) when $\tau_n \propto n^b$. Therefore, we set $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. Figure 6.3 displays the empirical median and first and last quartiles for the estimation of v with both algorithms and their averaged versions as a function of the number of observations. These estimates are obtained over 100 independent Monte Carlo runs. Both the BOEM and the OEM algorithms converge to the true value of v and the averaged versions reduce the variability of the estimation. Figure 6.4 shows the similar behavior of both averaged algorithms for the estimation of x_1 in the same experiment. Some supplementary graphs on the estimation of the states can be found in Appendix A.3.

We now compare the averaged BOEM algorithm to a recursive maximum likelihood (RML) procedure (see [Le Gland et Mevel, 1997, Tadić, 2010]) combined with Polyak-Ruppert averaging (see [Polyak et Juditsky, 1992]). We want to estimate the variance v and the transition matrix m. All the runs are started from $v = 2$ and from a matrix m with each entry equal to $1/d$. The RML algorithm follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$ which is chosen in the same way

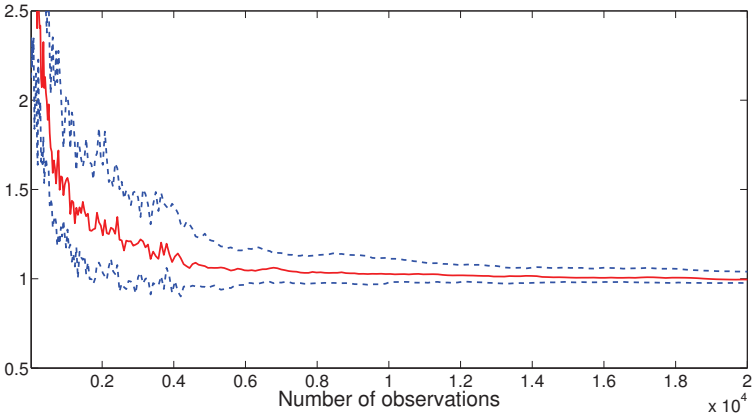


(a) The BOEM algorithm. (b) The OEM algorithm.

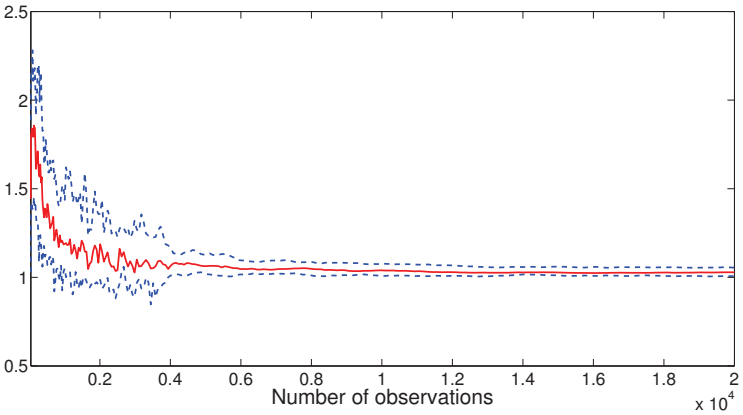
(c) The averaged BOEM algorithm. (d) The averaged OEM algorithm.

Figure 6.3: Estimation of v using the online EM and the BOEM algorithms (top) and their averaged versions (bottom). Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$.

as above. Therefore, for a fair comparison, the RML algorithm (resp. the BOEM algorithm) is run with $\gamma_n = n^{-0.53}$ (resp. $\tau_n = n^{1.1}$). Figure 6.5 displays the empirical median and empirical first and last quartiles of the estimation of $m(1,1)$ as a function of the number of observations over 100 independent Monte Carlo runs. For both algorithms, the bias and the variance of the estimation decrease as n increases. Nevertheless, the bias and/or the variance of the averaged BOEM algorithm decrease faster than those of the averaged RML algorithm (similar graphs have been obtained for the estimation of the other entries of the matrix m and for the estimation of v; see Appendix A.3). As a conclusion, it is advocated to use the averaged BOEM algorithm instead of the averaged RML algorithm.

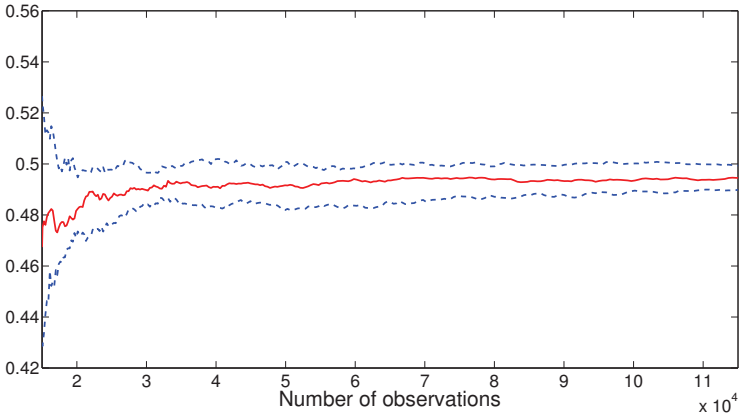


(a) The averaged BOEM algorithm.

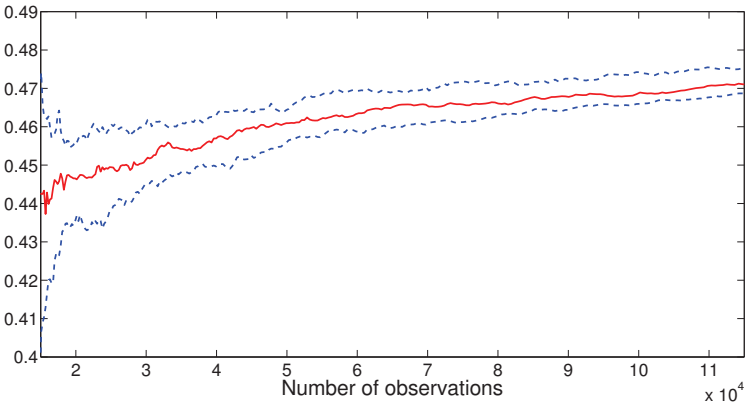


(b) The averaged OEM algorithm.

Figure 6.4: Estimation of x_1 using the averaged OEM and the averaged BOEM algorithms. Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. The first ten observations are omitted for a better visibility.



(a) The averaged BOEM algorithm.



(b) The averaged RML algorithm.

Figure 6.5: Empirical median (bold line) and first and last quartiles for the estimation of $m(1,1)$ using the averaged RML algorithm and the averaged BOEM algorithm (left). The true value is $m(1,1) = 0.5$ and the averaging procedure is starter after 10000 observations. The first 10000 observations are not displayed for a better clarity.

6.4 Convergence of the Block Online EM algorithms

Consider the following assumptions.

A1 (a) There exist continuous functions $\phi : \Theta \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}^d$ and $S : \mathbb{X} \times \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^d$ s.t.

$$\log m_\theta(x, x') + \log g_\theta(x', y) = \phi(\theta) + \left\langle S(x, x,', y), \psi(\theta) \right\rangle \ ,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product on $\mathbb{R}^d$.

(b) There exists an open subset $\mathcal{S}$ of $\mathbb{R}^d$ that contains the convex hull of $S(\mathbb{X} \times \mathbb{X} \times \mathbb{Y})$.

(c) There exists a continuous function $\bar{\theta} : \mathcal{S} \to \Theta$ s.t. for any $s \in \mathcal{S}$,

$$\bar{\theta}(s) = \mathrm{argmax}_{\theta \in \Theta} \ \{\phi(\theta) + \langle s, \psi(\theta) \rangle\} \ .$$

A2 There exist σ_- and σ_+ s.t. for any $(x, x') \in \mathbb{X}^2$ and any $\theta \in \Theta$, $0 < \sigma_- \leq m_\theta(x, x') \leq \sigma_+$. Set $\rho \stackrel{\text{def}}{=} 1 - (\sigma_-/\sigma_+)$.

A2, often referred to as the strong mixing condition, is commonly used to prove the forgetting property of the initial condition of the filter, see e.g. [Del Moral et Guionnet, 1998, Del Moral *et al.*, 2003]. This assumption holds for example if $\mathbb{X}$ is finite or for linear state-spaces with truncated gaussian state and measurement noises. More generally, this condition holds when $\mathbb{X}$ is compact.

We now introduce assumptions on the observation process $\mathbf{Y}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$\mathcal{F}_k^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\{Y_u\}_{u \leq k}\right) \quad \text{and} \quad \mathcal{G}_k^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\{Y_u\}_{u \geq k}\right) \tag{6.9}$$

be σ-fields associated to $\mathbf{Y}$. We also define the β-mixing coefficients by, see [Davidson, 1994],

$$\beta^{\mathbf{Y}}(n) = \sup_{u \in \mathbb{Z}} \sup_{B \in \mathcal{G}_{u+n}^{\mathbf{Y}}} \mathbb{E}\left[\left|\mathbb{P}(B|\mathcal{F}_u^{\mathbf{Y}}) - \mathbb{P}(B)\right|\right] \ , \forall \ n \geq 0 \ . \tag{6.10}$$

A3-(p) $\mathbb{E}\left[\sup_{x,x' \in \mathbb{X}^2} |S(x, x', Y_0)|^p\right] < +\infty$.

A4 (a) $\mathbf{Y}$ is a β-mixing stationary sequence such that there exist $C \in [0, 1)$ and $\beta \in (0, 1)$ satisfying, for any $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq C\beta^n$, where $\beta^{\mathbf{Y}}$ is defined in (6.10).

(b) $\mathbb{E}\left[|\log b_-(Y_0)| + |\log b_+(Y_0)|\right] < +\infty$ where

$$\begin{aligned} b_-(y) &\stackrel{\text{def}}{=} \inf_{\theta \in \Theta} \textstyle\int g_\theta(x, y) \lambda(\mathrm{d}x) \ , \\ b_+(y) &\stackrel{\text{def}}{=} \sup_{\theta \in \Theta} \textstyle\int g_\theta(x, y) \lambda(\mathrm{d}x) \ . \end{aligned}$$

Upon noting that, for all $n \geq 0$, $\beta^{\mathbf{Y}}(n) \leq \beta^{(\mathbf{X},\mathbf{Y})}(n)$, we can prove that A4(a) holds when $\mathbf{Y}$ is the observation process of a an HMM under classical geometric ergodicity conditions [Meyn et Tweedie, 1993, Chapter 15] and [Cappé *et al.*, 2005, Chapter 14].

A5 There exists $c > 0$ and $a > 1$ such that for all $n \geq 1$, $\tau_n = \lfloor cn^a \rfloor$.

For $p > 0$ and Z a random variable measurable with respect to the σ-algebra $\sigma(Y_n, n \in \mathbb{Z})$, set $\|Z\|_p \stackrel{\text{def}}{=} (\mathbb{E}[|Z|^p])^{1/p}$.

A6 -(p) There exists $b \geq (a+1)/2a$ (where a is defined in A5) such that, for any $n \geq 0$,

$$\left\| S_n - \widetilde{S}_n \right\|_p = O(\tau_{n+1}^{-b}) ,$$

where $\widetilde{S}_n$ is the Monte Carlo approximation of S_n defined by (6.5).

A6 gives a L_p control of the Monte Carlo error on each block. Such bounds are given for Sequential Monte Carlo algorithms in [Dubarry et Le Corff, 2011, Theorem 1]. [Le Corff et Fort, 2011a] provides practical conditions to ensure A6 in the case of Sequential Monte Carlo methods. In the sequel, $\mathcal{M}(\mathbb{X})$ denotes the set of all probability distributions on $(\mathbb{X}, \mathcal{X})$.

Theorem 6.1. *Let $\bar{p} > 2$. Assume that A1-2, A3-($\bar{p}$) and A4 hold.*

i) For any $\theta \in \Theta$, there exists a r.v. $\mathsf{S}(\theta, \mathbf{Y})$ s.t.

$$\sup_{\theta \in \Theta,\, \chi \in \mathcal{M}(\mathbb{X})} \left| \mathbb{E}_\theta^\chi \left[S(X_{-1}, X_0, Y_0) | Y_{-\tau:\tau} \right] - \mathsf{S}(\theta, \mathbf{Y}) \right| \leq C\rho^\tau \sup_{(x,x') \in \mathbb{X}^2} \left| S(x, x', Y_0) \right| , \quad \mathbb{P} - \text{a.s.} , \tag{6.11}$$

where C is a finite constant. Define for all $\theta \in \Theta$,

$$\bar{\mathrm{S}}(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathsf{S}(\theta, \mathbf{Y}) \right] . \tag{6.12}$$

ii) $\theta \mapsto \bar{\mathrm{S}}(\theta)$ is continuous on Θ and for any $T > 0$,

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \underset{\tau \to +\infty}{\longrightarrow} \bar{\mathrm{S}}(\theta) , \quad \mathbb{P} - \text{a.s.} , \tag{6.13}$$

where $\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y})$ is defined by (6.3).

iii) Assume in addition that A6-($\bar{p}$) holds. For any $p \in (2, \bar{p})$, there exists a constant C s.t. for any $n \geq 1$,

$$\left\| \widetilde{S}_n - \bar{\mathrm{S}}(\theta_n) \right\|_p \leq \frac{C}{\sqrt{\tau_{n+1}}} ,$$

where $\widetilde{S}_n$ is the Monte Carlo approximation of S_n defined by (6.5).

Theorem 6.1 allows to introduce the limiting EM algorithm, defined as the deterministic iterative algorithm $\check{\theta}_n = \mathrm{R}(\check{\theta}_{n-1})$ where

$$\mathrm{R}(\theta) \stackrel{\text{def}}{=} \bar{\theta}\left(\bar{\mathrm{S}}(\theta)\right) \ . \tag{6.14}$$

The limiting EM can be seen as an EM algorithm applied as if the whole trajectory $\mathbf{Y}$ was observed instead of $Y_{0:T}$. For this limiting EM, the so-called sufficient statistics depend on the observations only through the mean $\mathbb{E}\left[\mathsf{S}(\theta, \mathbf{Y})\right]$. The stationary points of the limiting EM are defined as

$$\mathcal{L} \stackrel{\text{def}}{=} \{\theta \in \Theta;\ \mathrm{R}(\theta) = \theta\} \ . \tag{6.15}$$

We show that there exists a Lyapunov function W w.r.t. to the map R and the set $\mathcal{L}$ *i.e.*, a continuous function W satisfying the two conditions:

(i) for all $\theta \in \Theta$,

$$\mathrm{W} \circ \mathrm{R}(\theta) - \mathrm{W}(\theta) \geq 0 \ , \tag{6.16}$$

(ii) for all compact set $\mathcal{K} \subset \Theta \setminus \mathcal{L}$,

$$\inf_{\theta \in \mathcal{K}} \{\mathrm{W} \circ \mathrm{R}(\theta) - \mathrm{W}(\theta)\} > 0 \ . \tag{6.17}$$

Recall that, for such a function, the sequence $\{\mathrm{W}(\check{\theta}_k)\}_{k\geq 0}$ is nondecreasing and $\{\check{\theta}_k\}_{k\geq 0}$ converges to $\mathcal{L}$.

Define, for any $m \geq 0$, $\theta \in \Theta$ and probability distribution χ on $(\mathbb{X}, \mathcal{X})$,

$$\begin{aligned} &p_\theta^\chi\left(Y_1 | Y_{-m:0}\right) \\ &\stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_{-m}) g_\theta(x_{-m}, Y_m) \prod_{i=-m+1}^{1} \{m_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i)\} \, \lambda(\mathrm{d}x_{-m+1:1})}{\int \chi(\mathrm{d}x_{-m}) g_\theta(x_{-m}, Y_m) \prod_{i=-m+1}^{0} \{m_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i)\} \, \lambda(\mathrm{d}x_{-m+1:0})} \ . \end{aligned}$$

By [Douc *et al.*, 2004b, Lemma 2 and Proposition 1], under A1-4, for any $\theta \in \Theta$, there exists a random variable $\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)$, such that for any probability distribution χ on $(\mathbb{X}, \mathcal{X})$, $\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)$ is the a.s. limit of $\log p_\theta^\chi\left(Y_1 | Y_{-m:0}\right)$ as $m \to +\infty$ and

$$T^{-1} \ell_{\theta,T}^\chi(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \ell(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\log p_\theta\left(Y_1 | Y_{-\infty:0}\right)\right] \ , \ \mathbb{P}-\text{a.s.} \ , \tag{6.18}$$

where $\ell_{\theta,T}^\chi(\mathbf{Y})$ is the log-likelihood defined by (6.2). The function $\theta \mapsto \ell(\theta)$ may be interpreted as the limiting log-likelihood. We consider the function W, given, for all $\theta \in \Theta$, by

$$\mathrm{W}(\theta) \stackrel{\text{def}}{=} \exp\{\ell(\theta)\} \ . \tag{6.19}$$

To identify the stationary points of the limiting EM algorithm as the stationary points of ℓ, we introduce an additional assumption.

A7 (a) For any $y \in \mathbb{Y}$ and for all $(x, x') \in \mathbb{X}^2$, $\theta \mapsto g_\theta(x, y)$ and $\theta \mapsto m_\theta(x, x')$ are continuously differentiable on Θ.

(b) $\mathbb{E}\left[\phi(\mathbf{Y}_0)\right] < +\infty$ where

$$\phi(y) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} \sup_{(x,x')\in\mathbb{X}^2} \left|\nabla_\theta \log m_\theta(x, x') + \nabla_\theta \log g_\theta(x', y)\right| \ .$$

Proposition 6.1. *Assume that A1-2, A3-*(1) *and A4 hold. Then, the function* W *given by* (6.19) *is a Lyapunov function for* $(\mathrm{R}, \mathcal{L})$*. Assume in addition that A7 holds. Then,* $\theta \mapsto \ell(\theta)$ *is continuously differentiable and*

$$\mathcal{L} = \{\theta \in \Theta;\ \mathrm{R}(\theta) = \theta\} = \{\theta \in \Theta;\ \nabla\ell(\theta) = 0\} \ .$$

Proposition 6.1 is proved in Section 6.6.2.

Remark. In the case where $\{Y_k\}_{k\geq 0}$ is the observation process of the stationary HMM $\{(X_k, Y_k)\}_{k\geq 0}$ parameterized by $\theta_\star \in \Theta$, we can build a two-sided stationary extension of this process to obtain a sequence of observations $\{Y_k\}_{k\in\mathbb{Z}}$. Following [Douc *et al.*, 2004b, Proposition 3], the quantity $\ell(\theta)$ can be written as

$$\begin{aligned}
\ell(\theta) &= \mathbb{E}_{\theta_\star}\left[\lim_{m\to+\infty} \log p_\theta(Y_1|Y_{-m:0})\right] \\
&= \lim_{m\to+\infty} \mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1|Y_{-m:0})\right] \\
&= \lim_{m\to+\infty} \mathbb{E}_{\theta_\star}\left[\mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1|Y_{-m:0})|Y_{-m:0}\right]\right] \ ,
\end{aligned}$$

where $p_\theta(Y_1|Y_{-m:0})$ is the conditional distribution under the stationary distribution. Since

$$\mathbb{E}_{\theta_\star}\left[\log p_{\theta_\star}(Y_1|Y_{-m:0})|Y_{-m:0}\right] - \mathbb{E}_{\theta_\star}\left[\log p_\theta(Y_1|Y_{-m:0})|Y_{-m:0}\right]$$

is the Kullback-Leibler divergence between $p_{\theta_\star}(Y_1|Y_{-m:0})$ and $p_\theta(Y_1|Y_{-m:0})$, for any $\theta \in \Theta$, $\ell(\theta_\star) - \ell(\theta) \geq 0$ and $\theta_\star$ is a maximizer of $\theta \mapsto \ell(\theta)$. If in addition $\theta_\star$ lies in the interior of Θ, then $\theta_\star \in \mathcal{L}$.

The following proposition gives sufficient conditions for the convergence of the limiting EM algorithm and the Monte Carlo BOEM algorithm to the set $\mathcal{L}$.

Theorem 6.2. *Let* $\bar{p} > 2$*. Assume that A1-2, A3-*$(\bar{p})$ *and A4 hold. Assume that* $\mathrm{W}(\mathcal{L})$ *has an empty interior. For any initial value* $\check{\theta}_0 \in \Theta$ *, there exists* $w_\star$ *s.t.* $\{\check{\theta}_k\}_{k\geq 0}$ *converges to* $\{\theta \in \mathcal{L};\ \mathrm{W}(\theta) = w_\star\}$*. If in addition A5 and A6-*$(\bar{p})$ *hold, then the sequence* $\{\theta_n\}_{n\geq 0}$ *converges* $\mathbb{P}-$a.s. *to the same stationary points.*

Theorem 6.2 is a direct application of Proposition 6.4 for the limiting EM algorithm. The proof for the Monte Carlo BOEM algorithm is detailed in Section 6.6.3.

By Sard's theorem if W is at least d_θ (where $\Theta \subset \mathbb{R}^{d_\theta}$) continuously differentiable, then $\mathrm{W}(\mathcal{L})$ has Lebesgue measure 0 and hence has an empty interior.

6.5 Rate of convergence of the Block Online EM algorithms

We address the rate of convergence of the Monte Carlo BOEM algorithms to a point $\theta_\star \in \mathcal{L}$. It is assumed that

A8 (a) $\bar{\mathrm{S}}$ and $\bar{\theta}$ are twice continuously differentiable on Θ and $\mathcal{S}$.

(b) There exists $0 < \gamma < 1$ s.t. the spectral radius of $\nabla_s(\bar{\mathrm{S}} \circ \bar{\theta})_{s=\bar{\mathrm{S}}(\theta_\star)}$ is lower than γ.

Hereafter, for any sequence of random variables $\{Z_n\}_{n\geq 0}$, write $Z_n = O_{\mathrm{L}_p}(1)$ if $\sup_n \mathbb{E}\left[|Z_n|^p\right] < \infty$ and $Z_n = O_{\mathrm{a.s}}(1)$ if $\sup_n |Z_n| < +\infty$ $\mathbb{P}-$a.s.

Theorem 6.3. *Let $\bar{p} > 2$. Assume that A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 hold. Then, for any $p \in (2, \bar{p})$,*

$$\sqrt{\tau_n}\ \left[\theta_n - \theta_\star\right] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{1}{\sqrt{\tau_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1)\ . \tag{6.20}$$

By a direct application of (6.20),

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{\sqrt{\tau_n}\ \|\theta_n - \theta_\star\| \mathbf{1}_{\lim_n \theta_n = \theta_\star} \geq M\right\} = 0\ .$$

The rate of convergence of the Monte Carlo BOEM algorithm is closely related to the choice of the number of observations per block. In (6.20), the rate is a function of the number of updates (i.e. the number of iteration of the algorithm). Theorem 6.4 shows that the averaging procedure reduces the influence of the block-size schedule: the rate of convergence is proportional to $T_n^{1/2}$ i.e. to the inverse of the square root of the total number of observations up to iteration n.

Theorem 6.4. *Let $\bar{p} > 2$. Assume that A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 hold. Then, for any $p \in (2, \bar{p})$,*

$$\sqrt{T_n}\left[\widetilde{\theta}_n - \theta_\star\right] \mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{n}{\sqrt{T_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1)\ . \tag{6.21}$$

In this case, (6.21) yields

$$\lim_{M\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{\sqrt{T_n}\ \|\widetilde{\theta}_n - \theta_\star\| \mathbf{1}_{\lim_n \theta_n = \theta_\star} \geq M\right\} = 0\ .$$

Theorems 6.3 and 6.4 give the rates of convergence as a function of the number of updates but they can also be studied as a function of the number of observations. Let $\{\theta_k^{\text{int}}\}_{k\geq 0}$ (resp. $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$) be such that, for any $k \geq 0$, θ_k^{int} (resp. $\widetilde{\theta}_k^{\text{int}}$) is the value θ_n (resp. $\widetilde{\theta}_n$), where n is the only integer such that $k \in [T_n + 1, T_{n+1}]$. The sequences $\{\theta_k^{\text{int}}\}_{k\geq 0}$ and $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$ are piecewise constant and their values are updated at times $\{T_n\}_{n\geq 1}$.

By Theorem 6.3, the rate of convergence of $\{\theta_k^{\text{int}}\}_{k\geq 0}$ is given (up to a multiplicative constant) by $k^{-a/(2(a+1))}$, where a is given by A5. This rates is slower than $k^{-1/2}$ and depends on the block-size sequence (through a). On the contrary, by Theorem 6.4, the rate of convergence of $\{\widetilde{\theta}_k^{\text{int}}\}_{k\geq 0}$ is given (up to a multiplicative constant) by $k^{-1/2}$, for any value of a. Therefore, this rate of convergence does not depend on the block-size sequence.

6.6 Proofs

Define, for any initial density χ on $(\mathbb{X}, \mathcal{X})$, any $\theta \in \Theta$, any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any $r < u \leq s \leq t$,

$$\Phi_{\theta,s,t}^{\chi,r}(h, \mathbf{y}) \stackrel{\text{def}}{=} \frac{\int \chi(x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1})\}\, h(x_{s-1}, x_s, y_s)\, \lambda(\mathrm{d}x_{r:t})}{\int \chi(x_r)\{\prod_{i=r}^{t-1} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, y_{i+1})\}\, \lambda(\mathrm{d}x_{r:t})}\,, \tag{6.22}$$

for any bounded function h on $\mathbb{X}^2 \times \mathbb{Y}$. Then, the intermediate quantity of the Block online EM algorithm is (see (6.3)),

$$\bar{S}_\tau^{\chi,T}(\theta, \mathbf{Y}) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=T+1}^{T+\tau} \Phi_{\theta,t,T+\tau}^{\chi,T}(S, \mathbf{Y})\,. \tag{6.23}$$

Lemma 6.1. *Assume A1-2. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ s.t. $\sup_{x,x'} |S(x, x', y_i)| < +\infty$ for any $i \in \mathbb{Z}$. Then for any $r > 0$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$, $\theta \mapsto \Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ is continuous on Θ.*

Proof. Set $K_\theta(x, x', y) \stackrel{\text{def}}{=} m_\theta(x, x') g_\theta(x', y)$. Let $r > 0$ and χ be a distribution on $(\mathbb{X}, \mathcal{X})$. By definition of $\Phi_{\theta,0,r}^{\chi,-r}(S, \mathbf{y})$ (see (6.22)) we have to prove that

$$\theta \mapsto \int \chi(\mathrm{d}x_{-r}) \left(\prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1}) \right) h(x_{-1}, x_0, y_0)\, \mathrm{d}\lambda(x_{-r+1:r})$$

is continuous for $h(x, x', y) = 1$ and $h(x, x', y) = S(x, x', y)$. By A1(a), the function $\theta \mapsto \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1})\, h(x_{-1}, x_0, y_0)$ is continuous. In

addition, under A1, for any $\theta \in \Theta$,

$$\left| \prod_{i=-r}^{r-1} K_\theta(x_i, x_{i+1}, y_{i+1})\, h(x_{-1}, x_0, y_0) \right| = |h(x_{-1}, x_0, y_0)| \exp\left(2r\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r}^{r-1} S(x_i, x_{i+1}, y_{i+1}) \right\rangle \right) .$$

Since Θ is compact, by A1, there exist constants C_1 and C_2 s.t. the supremum in $\theta \in \Theta$ of this expression is bounded above by

$$C_1 \sup_{x,x'} |h(x, x', y_0)| \exp\left(C_2 \sum_{i=-r}^{r-1} \sup_{x,x'} |S(x, x', y_{i+1})| \right) .$$

Since χ is a distribution and λ is a finite measure, the continuity follows from the dominated convergence theorem. □

Let us introduce the following shorthand $S_s(x, x') \stackrel{\text{def}}{=} S(x, x', Y_s)$. Define the shift operator ϑ onto $\mathbb{Y}^{\mathbb{Z}}$ by $(\vartheta \mathbf{y})_k = \mathbf{y}_{k+1}$ for any $k \in \mathbb{Z}$; and by induction, define the s-iterated shift operator $\vartheta^{s+1}\mathbf{y} = \vartheta(\vartheta^s \mathbf{y})$, with the convention that ϑ^0 is the identity operator. For a function h, define $\mathrm{osc}(h) \stackrel{\text{def}}{=} \sup_{z,z'} |h(z) - h(z')|$.

6.6.1 Proof of Theorem 6.1

The proof of Theorem 6.1 relies on auxiliary results about the forgetting properties of HMM. Most of them are close to published results and their proof is provided in the Appendix A. The main novelty is the forgetting property of the bivariate smoothing distribution.

Proof of i) Note that under A3-(1), $\mathbb{E}\left[\mathrm{osc}(S_0)\right] < +\infty$. Under A2, Proposition 6.5(ii) implies that for any $\theta \in \Theta$, there exists a r.v. $\Phi_\theta(S, \mathbf{Y})$ s.t. for any $r < s \leq T$,

$$\sup_{\theta \in \Theta} \left| \Phi^{\chi,r}_{\theta,s,T}(S, \mathbf{Y}) - \Phi_\theta(S, \vartheta^s \mathbf{Y}) \right| \leq \left(\rho^{T-s} + \rho^{s-r-1} \right) \mathrm{osc}(S_s) . \tag{6.24}$$

This concludes the proof of (6.11).

Proof of ii) We introduce the following decomposition: for all $T > 0$,

$$\bar{S}^{\chi,T}_\tau(\theta, \mathbf{Y}) = \frac{1}{\tau} \sum_{t=1}^{\tau} \left[\Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right) + \left\{ \Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^T \mathbf{Y}\right) - \Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right) \right\} \right] ,$$

upon noting that by (6.23), $\bar{S}^{\chi,T}_\tau(\theta, \mathbf{Y}) = \tau^{-1} \sum_{t=1}^{\tau} \Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^T \mathbf{Y}\right)$. By (6.22), (6.24) and A3-(1) $\mathbb{E}\left[|\Phi_\theta(S, \mathbf{Y})|\right] < +\infty$. Under A4, the ergodic

theorem (see e.g. [Billingsley, 1995, Theorem 24.1, p.314]) states that, for any fixed T,

$$\lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=1}^{\tau} \Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right) = \mathbb{E}\left[\Phi_\theta(S,\mathbf{Y})\right] \ , \quad \mathbb{P}-\text{a.s.}$$

By (6.24),

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \left|\Phi^{\chi,0}_{\theta,t,\tau}\left(S, \vartheta^{T}\mathbf{Y}\right) - \Phi_\theta\left(S, \vartheta^{t+T}\mathbf{Y}\right)\right| \leq \frac{1}{\tau}\sum_{t=1}^{\tau} \left(\rho^{\tau-t}+\rho^{t-1}\right)\text{osc}(S_{t+T}) \ . \tag{6.25}$$

Set $Z_t \stackrel{\text{def}}{=} \frac{1}{t}\sum_{s=1}^{t}\text{osc}(S_{s+T})$ and $Z_0 \stackrel{\text{def}}{=} 0$. Then, by an Abel transform,

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\rho^{t-1}\text{osc}(S_{t+T}) = \rho^{\tau-1}Z_\tau + \frac{1-\rho}{\tau}\sum_{t=1}^{\tau-1} t\rho^{t-1}Z_t \ . \tag{6.26}$$

By A3-(1) and A4, the ergodic theorem implies that $\lim_{\tau\to\infty} Z_\tau = \mathbb{E}\left[\text{osc}(S_0)\right]$, $\mathbb{P}-$a.s. Therefore, $\limsup_\tau Z_\tau < \infty$, $\mathbb{P}-$a.s. Since $\sum_{t\geq 1} t\rho^{t-1} < \infty$, this implies that $\tau^{-1}\sum_{t=1}^{\tau}\rho^{t-1}\text{osc}(S_{t+T}) \underset{\tau\to+\infty}{\longrightarrow} 0$, $\mathbb{P}-$a.s. Similarly,

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\rho^{\tau-t}\text{osc}(S_{t+T}) = Z_\tau - (1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}Z_t + \frac{1-\rho}{\tau}\sum_{t=1}^{\tau-1} t\rho^{t-1}Z_{\tau-t} \ .$$

Using the same arguments as for the second term in (6.26), we can state that $\lim_{\tau\to\infty}\tau^{-1}\sum_{t=1}^{\tau-1} t\rho^{t-1}Z_{\tau-t} = 0$, $\mathbb{P}-$a.s. Furthermore,

$$\left|\sum_{t=1}^{\tau-1}\frac{\rho^{\tau-t-1}}{1-\rho}Z_t - \mathbb{E}\left[\text{osc}(S_0)\right]\right| \leq \sum_{t=1}^{\tau-1}\frac{\rho^{\tau-t-1}}{1-\rho}\left|Z_t - \mathbb{E}\left[\text{osc}(S_0)\right]\right| + \mathbb{E}\left[\text{osc}(S_0)\right]\rho^{\tau-1} \ .$$

Since, $\mathbb{P}-$a.s., $Z_\tau \underset{\tau\to+\infty}{\longrightarrow} \mathbb{E}\left[\text{osc}(S_0)\right]$, the RHS converges $\mathbb{P}-$a.s. to 0 and

$$\lim_{\tau\to+\infty}\left|Z_\tau - (1-\rho)\sum_{t=1}^{\tau-1}\rho^{\tau-t-1}Z_t\right| = 0 \ , \quad \mathbb{P}-\text{a.s.}$$

Hence, the RHS in (6.25) converges $\mathbb{P}-$a.s. to 0 and this concludes the proof of (6.13). We now prove that the function $\theta \mapsto \mathbb{E}\left[\Phi_\theta\left(S,\mathbf{Y}\right)\right]$ is continuous by application of the dominated convergence theorem. By Proposition 6.5(ii), for any $\mathbf{y}$ s.t. $\text{osc}(S_0) < \infty$,

$$\lim_{r\to+\infty}\sup_{\theta\in\Theta}\left|\Phi^{\chi,-r}_{\theta,0,r}(S,\mathbf{y}) - \Phi_\theta\left(S,\mathbf{y}\right)\right| = 0 \ .$$

Then, by Lemma 6.1, $\theta \mapsto \Phi_\theta(S, \mathbf{y})$ is continuous for any $\mathbf{y}$ such that $\mathrm{osc}(S_0) < +\infty$. In addition, $\sup_{\theta\in\Theta} |\Phi_\theta(S, \mathbf{Y})| \leq \sup_{x,x'} |S(x, x', Y_0)|$. We then conclude by A3-(1).

Proof of iii) Let m_n, v_n be positive integers s.t. $1 \leq m_n \leq \tau_{n+1}$ and $\tau_{n+1} = 2v_n m_n + r_n$, where $0 \leq r_n < 2m_n$. Set $\Delta p \stackrel{\text{def}}{=} p^{-1} - \bar{p}^{-1}$. By the Minkowski inequality combined with Lemmas 6.5, 6.6 applied with $q_n \stackrel{\text{def}}{=} 2v_n m_n$, there exists a constant C s.t.

$$\left\|S_n - \bar{\mathrm{S}}(\theta_n)\right\|_p \leq C\left[\rho^{m_n} + \frac{m_n}{\tau_{n+1}} + \beta^{m_n \Delta p} + \frac{1}{\sqrt{\tau_{n+1}}}\right] .$$

The proof is concluded by choosing $m_n = \lfloor -\log \tau_{n+1} / (\log\rho \vee \Delta p \log\beta)\rfloor$ and by A6-($\bar{p}$) (since b in A6-($\bar{p}$) is such that $b \geq 1/2$).

6.6.2 Proof of Proposition 6.1

Continuity of R *and* W

By A1(c) and Theorem 6.1, the function R is continuous. Under A1-2 and A4, there exists a continuous function ℓ on Θ s.t. $\lim_T T^{-1}\ell_\chi\theta, T(\mathbf{Y}) = \ell(\theta)$ $\mathbb{P}-$a.s. for any distribution χ on $(\mathbb{X}, \mathcal{X})$ and any $\theta \in \Theta$, (see [Douc *et al.*, 2004b, Lemma 2 and Propositions 1 and 2], see also Theorem A.1). Therefore, W is continuous.

Proof of the Lyapunov property (6.16)

Under Assumption A1(a)

$$\frac{1}{T}\log p_\theta(x_{0:T}, Y_{1:T}) = \phi(\theta) + \left\langle \left\{\frac{1}{T}\sum_{t=1}^{T} S(x_{t-1}, x_t, Y_t)\right\}, \psi(\theta)\right\rangle ,$$

where $p_\theta(x_{0:T}, Y_{1:T})$ is defined by (6.1). Upon noting that

$$\int S(x_{t-1}, x_t, Y_t)\frac{p_\theta(x_{0:T}, Y_{1:T})}{\int p_\theta(z_{0:T}, Y_{1:T})\lambda(\mathrm{d}z_{1:T})\chi(\mathrm{d}z_0)}\lambda(\mathrm{d}x_{1:T})\chi(\mathrm{d}x_0) = \Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}) ,$$

the Jensen inequality gives, $\mathbb{P}-$a.s.,

$$\frac{1}{T}\ell^\chi_{\mathrm{R}(\theta),T}(\mathbf{Y}) - \frac{1}{T}\ell^\chi_{\theta,T}(\mathbf{Y}) \geq \phi(\mathrm{R}(\theta)) + \left\langle \frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}), \psi(\mathrm{R}(\theta))\right\rangle - \phi(\theta) - \left\langle \frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S, \mathbf{Y}), \psi(\theta)\right\rangle . \quad (6.27)$$

Under A1-4, it holds by Theorem 6.1 and [Douc *et al.*, 2004b, Lemma 2 and Proposition 1] (see also Theorem A.1) that for all $\theta \in \Theta$, $\mathbb{P}-$a.s.,

$$\frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \bar{\mathrm{S}}(\theta)\ , \qquad \frac{1}{T}\ell^{\chi}_{\theta,T}(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \ln \mathrm{W}(\theta)\ .$$

Therefore, when $T\to+\infty$, (6.27) implies

$$\ln\left(\mathrm{W}(\mathrm{R}(\theta))/\mathrm{W}(\theta)\right) \geq \phi(\mathrm{R}(\theta)) + \left\langle \bar{\mathrm{S}}(\theta), \psi(\mathrm{R}(\theta))\right\rangle - \phi(\theta) - \left\langle \bar{\mathrm{S}}(\theta), \psi(\theta)\right\rangle\ . \tag{6.28}$$

By definition of $\bar{\theta}$ and R (see A1(c) and (6.14)), the RHS is non negative. This concludes the proof of Proposition 6.1(6.16).

Proof of the Lyapunov property (6.17)

We prove that $\mathrm{W}\circ\mathrm{R}(\theta)-\mathrm{W}(\theta)=0$ if and only if $\theta\in\mathcal{L}$. Since $\mathrm{W}\circ\mathrm{R}-\mathrm{W}$ is continuous, this implies that $\inf_{\theta\in\mathcal{K}} \mathrm{W}\circ\mathrm{R}(\theta)-\mathrm{W}(\theta)>0$ for all compact set $\mathcal{K}\subset\Theta\setminus\mathcal{L}$. Let $\theta\in\Theta$ be s.t. $\mathrm{W}\circ\mathrm{R}(\theta)-\mathrm{W}(\theta)=0$. Then, the RHS in (6.28) is equal to zero. By definition of $\bar{\theta}$, $\mathrm{R}(\theta)=\theta$ and thus $\theta\in\mathcal{L}$. The converse implication is immediate from the definition of $\mathcal{L}$.

Stationary points If in addition A7 holds, Theorem A.2 proves that, for any initial distribution χ on $(\mathbb{X},\mathcal{X})$,

$$\frac{1}{T}\nabla_\theta \ell^{\chi}_{\theta,T}(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \nabla_\theta \ell(\theta) \quad \mathbb{P}-\text{a.s.}$$

Therefore,

$$\frac{1}{T}\nabla_\theta \ell^{\chi}_{\theta,T}(\mathbf{Y}) = \nabla_\theta\phi(\theta) + \nabla_\theta\psi'(\theta)\left\{\frac{1}{T}\sum_{t=1}^{T}\Phi^{\chi,0}_{\theta,t,T}(S,\mathbf{Y})\right\}\ ,$$

where A' is the transpose matrix of A. Theorem 6.1 yield,

$$\nabla_\theta \ell(\theta) = \nabla_\theta\phi(\theta) + \nabla_\theta\psi'(\theta)\bar{S}(\theta)\ .$$

The proof follows upon noting that by definition of $\bar{\theta}$, the unique solution to the equation $\nabla_\theta\phi(\tau)+\nabla_\theta\psi'(\tau)\bar{S}(\theta)=0$ is $\tau=\mathrm{R}(\theta)$.

6.6.3 Proof of Theorem 6.2

The proof of Theorem 6.2 relies on Proposition 6.4 applied with $T(\theta)\overset{\text{def}}{=}\mathrm{R}(\theta)$ and with $\theta_{n+1}=\bar{\theta}(\widetilde{S}_n)$. The key ingredient for this proof is the control of the L_p-mean error between the Monte Carlo Block Online EM algorithm and the *limiting EM*. The proof of this bound is derived in Theorem 6.1 and relies on preliminary lemmas given in Appendix 6.7. The proof of (6.38) is now close to the proof of [Fort et Moulines, 2003, Proposition 11] and is postponed to the Appendix A.

6.6.4 Proof of Theorem 6.3

Define $s_\star \stackrel{\text{def}}{=} \bar{\mathrm{S}}(\theta_\star)$ and write

$$\bar\theta(\widetilde S_n) - \bar\theta(s_\star) = \Upsilon(\widetilde S_n - s_\star) + \bar\theta(\widetilde S_n) - \bar\theta(s_\star) - \Upsilon(\widetilde S_n - s_\star) \;, \tag{6.29}$$

where $\Upsilon \stackrel{\text{def}}{=} \nabla\bar\theta(s_\star)$. We now derive the rate of convergence of the quantity $\widetilde S_n - s_\star$. Set $\mathrm{G}(s) \stackrel{\text{def}}{=} \bar{\mathrm{S}} \circ \bar\theta(s)$. Note that under A8(b), the spectral radius of Γ is lower than γ, where $\Gamma \stackrel{\text{def}}{=} \nabla\mathrm{G}(s_\star)$. Since $\mathrm{G}(s_\star) = s_\star$, we write

$$\widetilde S_n - s_\star = \Gamma\left(\widetilde S_{n-1} - s_\star\right) + \widetilde S_n - \mathrm{G}(\widetilde S_{n-1}) + \mathrm{G}(\widetilde S_{n-1}) - \mathrm{G}(s_\star) - \Gamma\left(\widetilde S_{n-1} - s_\star\right) \;.$$

Define $\{\mu_n\}_{n\ge0}$ and $\{\rho_n\}_{n\ge0}$ s.t. $\mu_0 = 0$, $\rho_0 = \widetilde S_0 - s_\star$ and

$$\mu_n \stackrel{\text{def}}{=} \Gamma\mu_{n-1} + e_n \;, \qquad \rho_n \stackrel{\text{def}}{=} \widetilde S_n - s_\star - \mu_n \;, \qquad n \ge 1 \;, \tag{6.30}$$

where,

$$e_n \stackrel{\text{def}}{=} \widetilde S_n - \bar{\mathrm{S}}(\theta_n) \;, \qquad n \ge 1 \;. \tag{6.31}$$

Proposition 6.2. *Assume A2, A3-($\bar p$), A4-5, A6-($\bar p$) and A8 for some $\bar p > 2$. Then for any $p \in (2, \bar p)$,*

$$\sqrt{\tau_n}\mu_n = O_{\mathrm{L}_p}(1) \quad \textit{and} \quad \tau_n\rho_n\mathbf{1}_{\lim_n S_n = s_\star} = O_{\mathrm{L}_{p/2}}(1)O_{\mathrm{a.s}}(1) \;.$$

The proof of Proposition 6.2 follows the same lines as the proof of [Fort et Moulines, 2003, Theorem 6]. The main ingredient is the control of $\|\mu_n\|_p$ which is a consequence of [Pólya et Szegő, 1976, Result 178, p. 39] and Theorem 6.1. The detailed proof is thus omitted and postponed to the Appendix A.

By Proposition 6.2, the first term in (6.29) gives

$$\sqrt{\tau_n}\Upsilon(S_n - s_\star)\mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{1}{\sqrt{\tau_n}} O_{\mathrm{L}_{p/2}}(1)O_{\mathrm{a.s}}(1) \;.$$

A Taylor expansion with integral remainder term gives the rate of convergence of the second term. This concludes the proof of Theorem 6.3, Eq. (6.20).

6.6.5 Proof of Theorem 6.4

In the sequel, for all function Ξ on $\Theta \times \mathbb{Y}^{\mathbb{Z}}$ and all $\upsilon \in \Theta$, we denote by $\mathbb{E}\left[\Xi(\theta, \mathbf{Y})\right]_{\theta=\upsilon}$ the function $\theta \mapsto \mathbb{E}\left[\Xi(\theta, \mathbf{Y})\right]$ evaluated at $\theta = \upsilon$. We preface the proof by the following lemma.

Lemma 6.2. *Assume A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 for some $\bar{p} > 2$. For any $p \in (2, \bar{p})$,*

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \left\| \sum_{k=1}^{n} \tau_{k+1} e_k \right\|_p < \infty \ ,$$

where e_n is given by (6.31).

Proof. By A5 and A6-($\bar{p}$), we have

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \sum_{k=1}^{n} \tau_{k+1} \left\| \widetilde{S}_k - S_k \right\|_p < \infty \ .$$

Then, it is sufficient to prove that

$$\limsup_{n\to+\infty} \frac{1}{\sqrt{T_{n+1}}} \left\| \sum_{k=1}^{n} \tau_{k+1} \left(\bar{S}(\theta_k) - S_k \right) \right\|_p < \infty \ .$$

Let $p \in (2, \bar{p})$. In the sequel, C is a constant independent on n and whose value may change upon each appearance. Let $1 \le m_n \le \tau_{n+1}$ and set $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$. By Lemma 6.6 applied with $q_k \stackrel{\text{def}}{=} 2v_k m_k$, we have,

$$\left\| \sum_{k=1}^{n} \tau_{k+1} \left(\bar{S}(\theta_k) - S_k \right) \right\|_p \le C \left(\sum_{k=1}^{n} \{ \tau_{k+1} \rho^{m_k} + m_k \} + \left\| \sum_{k=1}^{n} \{ \delta_k + \zeta_k \} \right\|_p \right) \ ,$$

where δ_k and ζ_k are defined by

$$\delta_k \stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right\} \ ,$$

$$\zeta_k \stackrel{\text{def}}{=} \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] - \mathbb{E}\left[\Phi^{\chi, -m_k}_{\theta, 0, m_k}(S, \mathbf{Y}) \right]_{\theta=\theta_k} \right\}$$

and where $F_{t,k}(\theta_k, \mathbf{Y}) \stackrel{\text{def}}{=} \Phi^{\chi, t-m_k}_{\theta_k, t, t+m_k}(S, \vartheta^{T_k}\mathbf{Y})$ and $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k}$ is given by (6.42). We will prove below that there exists C s.t.

$$\|\zeta_k\|_p \le C\, \beta^{m_k/pb} \tau_{k+1} \ , \qquad \forall k \ge 1 \tag{6.32}$$

$$\left\| \sum_{k=1}^{n} \delta_k \right\|_p \le C\sqrt{T_{n+1}} + C \sum_{k=1}^{n} \tau_{k+1} \beta^{m_k/pb} \ , \qquad \forall n \ge 1 \tag{6.33}$$

so that the proof is concluded by choosing $m_k = \lfloor \eta \log \tau_{k+1} \rfloor$, $\eta \stackrel{\text{def}}{=} (-1/\log \rho) \vee (-pb/\log \beta)$ and by using A5.

We turn to the proof of (6.32). By the Berbee Lemma (see [Rio, 1990, Chapter 5]) and A4, there exist $C \in [0,1)$ and $\beta \in (0,1)$ s.t. for all $k \geq 1$, there exists a random variable $Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ independent from $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k}$ with the same distribution as $Y_{T_k+m_k:T_{k+1}+m_k}$ and

$$\mathbb{P}\left\{ Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k} \neq Y_{T_k+m_k:T_{k+1}+m_k} \right\} \leq C\beta^{m_k} \ . \tag{6.34}$$

Upon noting that $\mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] = \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k}$, we have

$$\zeta_k = \sum_{t=2m_k}^{2v_k m_k} \left\{ \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] - \mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right\} \ . \tag{6.35}$$

Therefore, by setting $\mathcal{A}_k \stackrel{\text{def}}{=} \{ Y^{\star,(k)}_{T_k+m_k:T_{k+1}+m_k} \neq Y_{T_k+m_k:T_{k+1}+m_k} \}$,

$$|\zeta_k| \leq \sum_{t=2m_k}^{2v_k m_k} \mathbb{E}\left[\sup_{\theta\in\Theta} \left| F_{t,k}(\theta, \mathbf{Y}) - F_{t,k}(\theta, \mathbf{Y}^{\star,(k)}) \right| \mathbf{1}_{\mathcal{A}_k} \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \ .$$

Minkowski and Holder (with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$) inequalities, combined with (6.34), A4, Lemma 6.3 and A3-($\bar{p}$) yield (6.32).

We now prove (6.33). Upon noting that δ_k is $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_{k+1}}$-measurable and δ_k is a martingale increment, the Rosenthal inequality (see [Hall et Heyde, 1980, Theorem 2.12, p.23]) states that $\|\sum_{k=1}^n \delta_k\|_p \leq C \left(\sum_{k=1}^n I^{(1)}_k \right)^{1/p} + C I^{(2)}_n$ where

$$I^{(1)}_k \stackrel{\text{def}}{=} \mathbb{E}\left[|\delta_k|^p \right] \quad \text{and} \quad I^{(2)}_n \stackrel{\text{def}}{=} \left\| \left(\sum_{k=1}^n \mathbb{E}\left[|\delta_k|^2 \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] \right)^{1/2} \right\|_p \ .$$

Using again $\mathbb{E}\left[F_{t,k}(\theta_k, \mathbf{Y}^{\star,(k)}) \middle| \widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_k} \right] = \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k}$ and (6.35)

$$I^{(1)}_k \leq C \left\| \sum_{t=2m_k}^{2v_k m_k} \left\{ F_{t,k}(\theta_k, \mathbf{Y}) - \mathbb{E}\left[F_{t,k}(\theta, \mathbf{Y}) \right]_{\theta=\theta_k} \right\} \right\|^p_p + C \ \|\zeta_k\|^p_p \ .$$

By Lemma 6.5 and (6.32), there exists C s.t. for any $k \geq 1$

$$I^{(1)}_k \leq C \left(\tau^{p/2}_{k+1} + \tau^p_{k+1} \beta^{m_k/b} \right) \ , \tag{6.36}$$

and since $2/p < 1$, convex inequalities yield $\left(\sum_{k=1}^n I_k^{(1)}\right)^{1/p} \leq C\sqrt{T_{n+1}} + C\sum_{k=1}^n \tau_{k+1}\beta^{m_k/pb}$. By the Minkowski and Jensen inequalities, it holds $I_n^{(2)} \leq \left(\sum_{k=1}^n \{I_k^{(1)}\}^{2/p}\right)^{1/2}$. Hence, by (6.36),

$$I_n^{(2)} \leq C\sqrt{T_{n+1}} + C\sum_{k=1}^n \tau_{k+1}\beta^{m_k/pb} .$$

This concludes the proof of (6.33). □

We write $\Sigma_n - s_\star = \bar{\mu}_n + \bar{\rho}_n$ with

$$\bar{\mu}_n \stackrel{\text{def}}{=} \frac{1}{T_n}\sum_{k=1}^n \tau_k \mu_{k-1} \quad \text{and} \quad \bar{\rho}_n \stackrel{\text{def}}{=} \frac{1}{T_n}\sum_{k=1}^n \tau_k \rho_{k-1} . \tag{6.37}$$

Proposition 6.3. *Assume A2, A3-($\bar{p}$), A4-5, A6-($\bar{p}$) and A8 for some $\bar{p} > 2$. For any $p \in (2, \bar{p})$,*

$$\sqrt{T_n}\bar{\mu}_n = O_{\mathrm{L}_p}(1) , \qquad \frac{T_n}{n}\bar{\rho}_n \mathbf{1}_{\lim_n S_n = s_\star} = O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) .$$

Proof. Set $A \stackrel{\text{def}}{=} (I - \Gamma)$. Under A8, A^{-1} exists. By (6.30) and (6.37),

$$A\sqrt{T_n}\bar{\mu}_n = -\frac{\tau_{n+1}\mu_n}{\sqrt{T_n}} + \frac{1}{\sqrt{T_n}}\sum_{k=1}^n \tau_{k+1} e_k + \frac{1}{\sqrt{T_n}}\sum_{k=1}^n \tau_k \left(\frac{\tau_{k+1}}{\tau_k} - 1\right)\Gamma\mu_{k-1} .$$

The result now follows from Proposition 6.2, Lemma 6.2 and A5. The proof of the second assertion follows from (6.37) and Proposition 6.2. □

Upon noting that $\theta_\star = \bar{\theta}(s_\star)$, we may write, for the averaged sequence,

$$\widetilde{\theta}_n - \theta_\star = \Upsilon(\Sigma_n - s_\star) + \bar{\theta}(\Sigma_n) - \bar{\theta}(s_\star) - \Upsilon(\Sigma_n - s_\star) .$$

The first term in this decomposition gives

$$\sqrt{T_n}\Upsilon(\Sigma_n - s_\star)\mathbf{1}_{\lim_n \theta_n = \theta_\star} = O_{\mathrm{L}_p}(1) + \frac{n}{\sqrt{T_n}} O_{\mathrm{L}_{p/2}}(1) O_{\mathrm{a.s}}(1) .$$

By A8(b), as for the non averaged sequence, a Taylor expansion with integral remainder term gives the result for the second term. This concludes the proof of Theorem 6.4, Eq.(6.21).

6.7 Technical results

Proposition 6.4 is exactly [Fort et Moulines, 2003, Proposition 9] applied with a compact set Θ.

Proposition 6.4. *Let $T : \Theta \to \Theta$ and W be a continuous Lyapunov function relatively to T and to $\mathcal{L} \subset \Theta$. Assume $\mathrm{W}(\mathcal{L})$ has an empty interior and that $\{\theta_n\}_{n\geq 0}$ is a sequence lying in Θ such that*

$$\lim_{n\to+\infty} |\mathrm{W}(\theta_{n+1}) - \mathrm{W}\circ T(\theta_n)| = 0 \; . \tag{6.38}$$

Then, there exists $w_\star$ such that $\{\theta_n\}_{n\geq 0}$ converges to $\{\theta \in \mathcal{L};\ \mathrm{W}(\theta) = w_\star\}$.

The proof of Proposition 6.5 is given in Section A.2.

Proposition 6.5. *Assume A2. Let χ, $\widetilde{\chi}$ be two distributions on $(\mathbb{X}, \mathcal{X})$. For any measurable function $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\sup_{x,x'} |h(x, x', y_s)| < +\infty$ for any $s \in \mathbb{Z}$*

(i) For any $r < s \leq t$ and any $\ell_1, \ell_2 \geq 1$,

$$\sup_{\theta\in\Theta} \left|\Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t+\ell_2}(h,\mathbf{y})\right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \; . \tag{6.39}$$

(ii) For any $\theta \in \Theta$, there exists a function $\mathbf{y} \mapsto \Phi_\theta(h, \mathbf{y})$ s.t. for any distribution χ on $(\mathbb{X}, \mathcal{X})$ and any $r < s \leq t$

$$\sup_{\theta\in\Theta} \left|\Phi^{\chi,r}_{\theta,s,t}(h,\mathbf{y}) - \Phi_\theta\left(h, \vartheta^s\mathbf{y}\right)\right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \; . \tag{6.40}$$

Remark. (a) If $\chi = \widetilde{\chi}$, $\ell_1 = 0$ and $\ell_2 \geq 1$, (6.39) becomes

$$\sup_{\theta\in\Theta} \left|\Phi^{\chi,r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r}_{\theta,s,t+\ell_2}(h,\mathbf{y})\right| \leq \rho^{t-s}\mathrm{osc}(h_s) \; .$$

(b) if $\ell_2 = 0$ and $\ell_1 \geq 1$, (6.39) becomes

$$\sup_{\theta\in\Theta} \left|\Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t}(h,\mathbf{y})\right| \leq \rho^{s-1-r}\mathrm{osc}(h_s) \; .$$

Lemma 6.3 is a consequence of (6.22) and of Proposition 6.5(ii).

Lemma 6.3. *Assume A2. Let $r < s \leq t$ be integers, $\theta \in \Theta$ and $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, and $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ s.t. for any $s \in \mathbb{Z}$, $\sup_{x,x'} |h(x, x', y_s)| < \infty$. Then*

$$\left|\Phi^{\chi,r}_{\theta,s,t}(h,\mathbf{y})\right| \leq \sup_{(x,x')\in\mathbb{X}^2} \left|h(x,x',y_s)\right| \; , \; \left|\Phi_\theta\left(h, \vartheta^s\mathbf{y}\right)\right| \leq \sup_{(x,x')\in\mathbb{X}^2} \left|h(x,x',y_s)\right| \; .$$

For any $L \geq 1$, $m \geq 1$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$, define

$$\kappa^{\chi}_{L,m}(\boldsymbol{\theta}, \mathbf{Y}) \stackrel{\text{def}}{=} \Phi^{\chi, L-m}_{\boldsymbol{\theta}, L, L+m}(S, \mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon=\boldsymbol{\theta}} . \tag{6.41}$$

We introduce the σ-algebra $\widetilde{\mathcal{F}}_{T_n}$ defined by

$$\widetilde{\mathcal{F}}_{T_n} \stackrel{\text{def}}{=} \sigma\{\mathcal{F}^{\mathbf{Y}}_{T_n}, \mathcal{H}_{T_n}\} , \tag{6.42}$$

where $\mathcal{F}_{T_n}$ is given by (6.9) and where $\mathcal{H}_{T_n}$ is independent from $\mathbf{Y}$ (the σ-algebra $\mathcal{H}_{T_n}$ is generated by the random variables independent from the observations $\mathbf{Y}$ used to produce the Monte Carlo approximation of $\{S_{k-1}\}_{k=1}^{n}$). Hence, for any positive integer m and any $B \in \mathcal{G}^{\mathbf{Y}}_{T_n+m}$, since $\mathcal{H}_{T_n}$ is independent from B and from $\mathcal{F}^{\mathbf{Y}}_{T_n}$, $\mathbb{P}(B|\widetilde{\mathcal{F}}_{T_n}) = \mathbb{P}(B|\mathcal{F}^{\mathbf{Y}}_{T_n})$. Therefore, the mixing coefficients defined in (6.10) are such that

$$\beta(\mathcal{G}^{\mathbf{Y}}_{T_n+m}, \widetilde{\mathcal{F}}_{T_n}) = \beta(\mathcal{G}^{\mathbf{Y}}_{T_n+m}, \mathcal{F}_{T^{\mathbf{Y}}_n}) .$$

Note that θ_n is $\widetilde{\mathcal{F}}_{T_n}$- measurable and that $\widetilde{S}_n$ is $\widetilde{\mathcal{F}}_{T_{n+1}}$-measurable.

Lemma 6.4. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any distribution χ on $(\mathbb{X}, \mathcal{X})$, any $m \geq 1$, $k, \ell \geq 0$ and any Θ-valued $\widetilde{\mathcal{F}}^{\mathbf{Y}}_0$-measurable r.v. $\boldsymbol{\theta}$,*

$$\left\| \sum_{u=1}^{k} \kappa^{\chi}_{2um+\ell,m}(\boldsymbol{\theta}, \mathbf{Y}) \right\|_p \leq C \left[\sqrt{\frac{k}{m}} + k\beta^{m\,\Delta p} \right] ,$$

where $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$ and β is given by A4.

Proof. For ease of notation χ is dropped from the notation $\kappa^{\chi}_{2um,m}$. By the Berbee Lemma (see [Rio, 1990, Chapter 5]), for any $m \geq 1$, there exists a Θ-valued r.v. $\boldsymbol{\upsilon}^{\star}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ independent from $\mathcal{G}^{\mathbf{Y}}_m$ (see (6.9)) s.t.

$$\mathbb{P}\left\{\boldsymbol{\theta} \neq \boldsymbol{\upsilon}^{\star}\right\} = \sup_{B \in \mathcal{G}^{\mathbf{Y}}_m} |\mathbb{P}(B|\sigma(\boldsymbol{\theta})) - \mathbb{P}(B)| . \tag{6.43}$$

Set $L_u \stackrel{\text{def}}{=} 2um + \ell$. We write

$$\begin{aligned}
\sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\theta}, \mathbf{Y}) &= \sum_{u=1}^{k} \left\{ \Phi^{\chi, L_u-m}_{\boldsymbol{\theta}, L_u, L_u+m}(S, \mathbf{Y}) - \Phi^{\chi, L_u-m}_{\boldsymbol{\upsilon}^{\star}, L_u, L_u+m}(S, \mathbf{Y}) \right\} \\
&+ \sum_{u=1}^{k} \kappa_{L_u,m}(\boldsymbol{\upsilon}^{\star}, \mathbf{Y}) + k \left\{ \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon=\boldsymbol{\upsilon}^{\star}} - \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S, \mathbf{Y})\right]_{\upsilon=\boldsymbol{\theta}} \right\} .
\end{aligned} \tag{6.44}$$

By the Holder's inequality with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\left\|\Phi^{\chi,L-m}_{\boldsymbol{\theta},L,L+m}(S,\mathbf{Y}) - \Phi^{\chi,L-m}_{\boldsymbol{\upsilon}^\star,L,L+m}(S,\mathbf{Y})\right\|_p \leq \left\|\Phi^{\chi,L-m}_{\boldsymbol{\theta},L,L+m}(S,\vartheta^T\mathbf{Y}) - \Phi^{\chi,L-m}_{\boldsymbol{\upsilon}^\star,L,L+m}(S,\mathbf{Y})\right\|_{\bar{p}} \mathbb{P}\left\{\boldsymbol{\theta} \neq \boldsymbol{\upsilon}^\star\right\}^{\Delta p} .$$

By A3-($\bar{p}$), A4, (6.10) and (6.43), there exists a constant C_1 s.t. for any $m, L \geq 1$, any distribution χ and any Θ-valued $\widetilde{\mathcal{F}}_0^{\mathbf{Y}}$-measurable r.v. $\boldsymbol{\theta}$,

$$\left\|\Phi^{\chi,L-m}_{\boldsymbol{\theta},L,L+m}(S,\mathbf{Y}) - \Phi^{\chi,L-m}_{\boldsymbol{\upsilon}^\star,L,L+m}(S,\mathbf{Y})\right\|_{\bar{p}} \leq C_1\beta^{m\Delta p} .$$

Similarly, there exists a constant C_2 s.t. for any $m \geq 1$, any distribution χ and any Θ-valued $\widetilde{\mathcal{F}}_0^{\mathbf{Y}}$-measurable r.v. $\boldsymbol{\theta}$,

$$\left\|\mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S,\mathbf{Y})\right]_{\upsilon=\boldsymbol{\upsilon}^\star} - \mathbb{E}\left[\Phi^{\chi,-m}_{\upsilon,0,m}(S,\mathbf{Y})\right]_{\upsilon=\boldsymbol{\theta}}\right\|_p \leq C_2\beta^{m\Delta p} .$$

Let us consider the second term in (6.44). For any $u \geq 1$ and any $\upsilon \in \Theta$, the r.v. $\kappa_{L_u,m}(\upsilon,\mathbf{Y})$ is a measurable function of $\mathbf{Y}_i$ for all $L_u - m + 1 \leq i \leq L_u + m$. Since $L_u \geq 2um$, for any $\upsilon \in \Theta$, $\sum_{u=1}^k \kappa_{L_u,m}(\upsilon,\mathbf{Y})$ is $\mathcal{G}_m^{\mathbf{Y}}$-measurable. $\boldsymbol{\upsilon}^\star$ is independent from $\mathcal{G}_m^{\mathbf{Y}}$ so that:

$$\left\|\sum_{u=1}^k \kappa_{L_u,m}(\boldsymbol{\upsilon}^\star,\mathbf{Y})\right\|_p = \mathbb{E}\left[\mathbb{E}\left[\left|\sum_{u=1}^k \kappa_{L_u,m}(\upsilon,\mathbf{Y})\right|^p\right]_{\upsilon=\boldsymbol{\upsilon}^\star}\right]^{1/p} .$$

Define the strong mixing coefficient (see [Davidson, 1994])

$$\alpha^{\mathbf{Y}}(r) \stackrel{\text{def}}{=} \sup_{u\in\mathbb{Z}} \sup_{(A,B)\in\mathcal{F}_u^{\mathbf{Y}}\times\mathcal{G}_{u+r}^{\mathbf{Y}}} \left|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)\right| , r \geq 0 .$$

Then, [Davidson, 1994, Theorem 14.1, p.210] implies that for any $m \geq 1$, the strong mixing coefficients of the sequence $\kappa_{(\mathbf{m})} \stackrel{\text{def}}{=} \{\kappa_{L_u,m}(\upsilon,\mathbf{Y})\}_{u\geq 1}$ satisfies $\alpha^{\kappa_{(\mathbf{m})}}(i) \leq \alpha^{\mathbf{Y}}(2(i-1)m+1)$. Furthermore, by [Rio, 1990, Theorem 2.5],

$$\left\|\sum_{u=1}^k \kappa_{L_u,m}(\upsilon,\mathbf{Y})\right\|_p \leq (2kp)^{1/2}\left(\int_0^1 \left[N_{(m)}(t)\wedge k\right]^{p/2} \mathcal{Q}^p_{\upsilon,m}(t)\mathrm{d}t\right)^{1/p} ,$$

where $N_{(m)}(t) \stackrel{\text{def}}{=} \sum_{i\geq 1} \mathbf{1}_{\alpha^{\kappa_{(\mathbf{m})}}(i)>t}$ and $\mathcal{Q}_{\upsilon,m}$ denotes the inverse of the tail function $t \mapsto \mathbb{P}(|\kappa_{L_u,m}(\upsilon,\mathbf{Y})| \geq t)$. The sequence $\mathbf{Y}$ being stationary, this inverse function does not depend on u. By A4 and the inequality $\alpha^{\mathbf{Y}}(r) \leq \beta^{\mathbf{Y}}(r)$ (see e.g. [Davidson, 1994, Chapter 13]), there exist $\beta \in [0,1)$ and $C \in (0,1)$ s.t. for any $u, m \geq 1$,

$$N_{(m)}(u) \leq \sum_{i\geq 1} \mathbf{1}_{\alpha^{\mathbf{Y}}(2(i-1)m+1)>u} \leq \sum_{i\geq 1} \mathbf{1}_{C\beta^{2(i-1)m}>u} \leq \left(\frac{\log u - \log C}{2m\log\beta}\right) \vee 0 .$$

Let U be a uniform r.v. on $[0,1]$. Observe that $C\beta^{2mk} < 1$. Then, by the Holder inequality applied with $a \stackrel{\text{def}}{=} \bar{p}/p$ and $b^{-1} \stackrel{\text{def}}{=} 1 - a^{-1}$,

$$\begin{aligned}
\left\| \left[N_{(m)}(U) \wedge k\right]^{1/2} \mathcal{Q}_{\upsilon,m}(U) \right\|_p &\stackrel{\text{def}}{=} \left(\int_0^1 \left[N_{(m)}(u) \wedge k\right]^{p/2} \mathcal{Q}^p_{\upsilon,m}(u) \mathrm{d}u \right)^{1/p} \\
&\leq \left[\frac{-1}{2m \log \beta} \right]^{1/2} \left\| \mathcal{Q}_{\upsilon,m}(U) \left(-\log \frac{U}{C} \right)^{1/2} \mathbf{1}_{(C\beta^{Cmk},C)}(U) \right\|_p \\
&\qquad + k^{1/2} \left\| \mathcal{Q}_{\upsilon,m}(U) \mathbf{1}_{U \leq C\beta^{2mk}} \right\|_p \ , \\
&\leq \left\{ (C\beta^{2mk})^{\Delta p} k^{1/2} + \left[\frac{-1}{2m \log \beta} \right]^{1/2} \left\| \left(-\log \frac{U}{C} \right)^{1/2} \mathbf{1}_{(C\beta^{Cmk},C)}(U) \right\|_{pb} \right\} \\
&\qquad \times \left\| \mathcal{Q}_{\upsilon,m}(U) \right\|_{\bar{p}} \ .
\end{aligned}$$

Since U is uniform on $[0,1]$, $\mathcal{Q}_{\upsilon,m}(U)$ and $|\kappa_{L_u,m}(\upsilon, \mathbf{Y})|$ have the same distribution, see [Rio, 1990]. Then, by Lemma 6.3 and A3-($\bar{p}$), there exists a constant C s.t. for any $\upsilon \in \Theta$, any $m \geq 1$,

$$\sup_{\upsilon \in \Theta} \left\| \mathcal{Q}_{\upsilon,m}(U) \right\|_{\bar{p}} \leq C \left\| \sup_{x,x' \in \mathbb{X}^2} |S(x,x',\mathbf{Y}_0)| \right\|_{\bar{p}} \ ,$$

which concludes the proof. □

Lemma 6.5. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. Let $p \in (2, \bar{p})$. There exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$,*

$$\left\| \frac{1}{\tau_{n+1}} \sum_{t=2m_n}^{2v_n m_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_p \leq C \left[\frac{1}{\sqrt{\tau_{n+1}}} + \beta^{m_n \Delta p} \right] \ ,$$

where $\kappa^{\chi}_{L,m}$ and β are defined by (6.41) *and A4, $v_n \stackrel{\text{def}}{=} \left\lfloor \frac{\tau_{n+1}}{2m_n} \right\rfloor$ and $\Delta p \stackrel{\text{def}}{=} \frac{\bar{p}-p}{p\bar{p}}$.*

Proof. We write,

$$\left\| \sum_{t=2m_n}^{2v_n m_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_p \leq \sum_{\ell=0}^{2m_n-1} \left\| \sum_{u=1}^{v_n-1} \kappa^{\chi}_{2um_n+\ell,m_n}(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_p \ .$$

Observe that by definition θ_n is $\widetilde{\mathcal{F}}^{\mathbf{Y}}_{T_n}$-measurable. Then, by Lemma 6.4, there exists a constant C s.t. for any $m_n \geq 1$ and any $\ell \geq 0$,

$$\left\| \sum_{u=1}^{v_n-1} \kappa^{\chi}_{2um_n+\ell,m_n}(\theta_n, \vartheta^{T_n} \mathbf{Y}) \right\|_p \leq C \left[\sqrt{\frac{v_n}{m_n}} + v_n \beta^{m_n \Delta p} \right] \ .$$

The proof is concluded upon noting that $\tau_{n+1} \geq 2m_n v_n$. □

Lemma 6.6. *Assume A2, A3-($\bar{p}$) and A4 for some $\bar{p} > 2$. For any $p \in (2, \bar{p}]$, there exists a constant C s.t. for any $n \geq 1$, any $1 \leq m_n \leq q_n \leq \tau_{n+1}$ and any distribution χ on $(\mathbb{X}, \mathcal{X})$,*

$$\left\| \bar{S}^{\chi, T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \bar{\mathrm{S}}(\theta_n) - \widetilde{\rho}_n \right\|_p \leq C \left[\rho^{m_n} + \frac{m_n}{\tau_{n+1}} + \frac{\tau_{n+1} - q_n}{\tau_{n+1}} \right] ,$$

where $\widetilde{\rho}_n \overset{\text{def}}{=} \tau_{n+1}^{-1} \sum_{t=2m_n}^{q_n} \kappa^{\chi}_{t,m_n}(\theta_n, \vartheta^{T_n}\mathbf{Y})$ and $\kappa^{\chi}_{L,m}$ is defined by (6.41).

Proof. By (6.3) and (6.22), $\bar{S}^{\chi,T_n}_{\tau_{n+1}}(\theta_n, \mathbf{Y}) - \bar{\mathrm{S}}(\theta_n) - \widetilde{\rho}_n = \sum_{i=1}^{4} g_{i,n}$ where

$$\begin{aligned}
g_{1,n} &\overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{\tau_{n+1}} \left(\Phi^{\chi,0}_{\theta_n,t,\tau_{n+1}}(S, \vartheta^{T_n}\mathbf{Y}) - \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) \right) , \\
g_{2,n} &\overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=1}^{2m_n-1} \left(\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right) , \\
g_{3,n} &\overset{\text{def}}{=} \frac{1}{\tau_{n+1}} \sum_{t=q_n+1}^{\tau_{n+1}} \left(\Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right) , \\
g_{4,n} &\overset{\text{def}}{=} \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} - \bar{\mathrm{S}}(\theta_n) .
\end{aligned}$$

In the case $\tau_{n+1} > 2m_n$, it holds

$$\begin{aligned}
\tau_{n+1} \left| g_{1,n} \right| \leq & \sum_{t=\tau_{n+1}-m_n+1}^{\tau_{n+1}} \left(\rho^{m_n-1} + \rho^{\tau_{n+1}-t} \right) \operatorname{osc}(S_{t+T_n}) \\
& + \sum_{t=1}^{m_n} \left(\rho^{m_n} + \rho^{t-1} \right) \operatorname{osc}(S_{t+T_n}) + 2\rho^{m_n-1} \sum_{t=m_n+1}^{\tau_{n+1}-m_n} \operatorname{osc}(S_{t+T_n}) ,
\end{aligned}$$

where we used Proposition 6.5(i) and Remark 6.7 in the last inequality. By A3-($\bar{p}$) and A4, there exists C s.t. $\|g_{1,n}\|_p \leq C\left(\rho^{m_n} + \tau_{n+1}^{-1}\right)$. The same bound hold in the case $\tau_{n+1} \leq 2m_n$. For $g_{2,n}$ and $g_{3,n}$, we use the bounds

$$\begin{aligned}
& \left| \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right| \\
& \qquad \leq \sup_{(x,x') \in \mathbb{X}^2} \left| S(x, x', Y_{T_n+t}) \right| + \mathbb{E}\left[\sup_{(x,x') \in \mathbb{X}^2} \left| S(x, x', Y_0) \right| \right] .
\end{aligned}$$

Then, by A4,

$$\begin{aligned}
& \left\| \Phi^{\chi,t-m_n}_{\theta_n,t,t+m_n}(S, \vartheta^{T_n}\mathbf{Y}) - \mathbb{E}\left[\Phi^{\chi,-m_n}_{\theta,0,m_n}(S, \mathbf{Y}) \right]_{\theta=\theta_n} \right\|_p \\
& \qquad \leq 2 \left\| \sup_{(x,x') \in \mathbb{X}^2} \left| S(x, x', Y_0) \right| \right\|_p ,
\end{aligned}$$

and the RHS is finite under A3-($\bar{p}$). Finally,

$$|g_{4,n}| \leq 2\rho^{m_n-1}\mathbb{E}\left[\mathrm{osc}(S_0)\right] \ ,$$

where we used Theorem 6.1. This concludes the proof. □

Chapitre 7

Inégalités de déviation non asymptotiques pour l'estimation de fonctionnelles additives lissées dans les modèles de Markov cachés (article)

The approximation of fixed-interval smoothing distributions is a key issue in inference for general state-space hidden Markov models (HMM). This contribution establishes non-asymptotic bounds for the Forward Filtering Backward Smoothing (FFBS) and the Forward Filtering Backward Simulation (FFBSi) estimators of fixed-interval smoothing functionals. We show that the rate of convergence of the L_q-mean errors of both methods depends on the number of observations T and the number of particles N only through the ratio T/N for additive functionals. In the case of the FFBS, this improves recent results providing bounds depending on $T/\sqrt{N}$.

7.1 Introduction

State-space models play a key role in statistics, engineering and econometrics; see for example [Cappé *et al.*, 2005], [Durbin et Koopman, 2000], and [West et Harrison, 1989]. Consider a process $\{X_t\}_{t\geq 0}$ taking values in a general state-space $\mathbb{X}$. This hidden process can be observed only through the observation process $\{Y_t\}_{t\geq 0}$ taking values in $\mathbb{Y}$. Statistical inference in general state-space models involves the computation of expectations of

additive functionals of the form

$$S_T = \sum_{t=1}^{T} h_t(X_{t-1}, X_t) \ ,$$

conditionally to $\{Y_t\}_{t=0}^T$, where T is a positive integer and $\{h_t\}_{t=1}^T$ are functions defined on $\mathbb{X}^2$. These smoothed additive functionals appear naturally for maximum likelihood parameter inference in hidden Markov models. The computation of the gradient of the log-likelihood function (Fisher score) or of the intermediate quantity of the Expectation Maximization algorithm involves the estimation of such smoothed functionals, see [Cappé *et al.*, 2005, Chapter 10 and 11] and [Doucet *et al.*, 2011].

Except for linear Gaussian state-spaces or for finite state-spaces, these smoothed additive functionals cannot be computed explicitly. In this paper, we consider Sequential Monte Carlo algorithms, henceforth referred to as particle methods, to approximate these quantities. These methods combine sequential importance sampling and sampling importance resampling steps to produce a set of random particles with associated importance weights to approximate the fixed-interval smoothing distributions.

The most straightforward implementation is based on the so-called path-space method. The complexity of this algorithm per time-step grows only linearly with the number N of particles, see [Del Moral, 2004]. However, a well-known shortcoming of this algorithm is known in the literature as the path degeneracy; see [Doucet *et al.*, 2011] for a discussion.

Several solutions have been proposed to solve this degeneracy problem. In this paper, we consider the Forward Filtering Backward Smoothing algorithm (FFBS) and the Forward Filtering Backward Simulation algorithm (FFBSi) introduced in [Doucet *et al.*, 2000] and further developed in [Godsill *et al.*, 2004]. Both algorithms proceed in two passes. In the forward pass, a set of particles and weights is stored. In the Backward pass of the FFBS the weights are modified but the particles are kept fixed. The FFBSi draws independently different particle trajectories among all possible paths. Since they use a backward step, these algorithms are mainly adapted for batch estimation problems. However, as shown in [Del Moral *et al.*, 2010a], when applied to additive functionals, the FFBS algorithm can be implemented forward in time, but its complexity grows quadratically with the number of particles. As shown in [Douc *et al.*, 2011a], it is possible to implement the FFBSi with a complexity growing only linearly with the number of particles.

The control of the L_q-norm of the deviation between the smoothed additive functional and its particle approximation has been studied recently in [Del Moral *et al.*, 2010a, Del Moral *et al.*, 2010b]. In the unpublished paper by [Del Moral *et al.*, 2010b], it is shown that the FFBS estimator variance of any smoothed additive functional is upper bounded by terms

depending on T and N only through the ratio T/N. Furthermore, in [Del Moral *et al.*, 2010a], for any $q > 2$, a L_q-mean error bound for smoothed functionals computed with the FFBS is established. When applied to strongly mixing kernels, this bound amounts to be of order $T/\sqrt{N}$ either for

(i) uniformly bounded in time general path-dependent functionals,

(ii) unnormalized additive functionals (see [Del Moral *et al.*, 2010a, Equation (3.8), pp. 957]).

In this paper, we establish L_q-mean error and exponential deviation inequalities of both the FFBS and FFBSi smoothed functionals estimators. We show that, for any $q \geq 2$, the L_q-mean error for both algorithms is upper bounded by terms depending on T and N only through the ratio T/N under the strong mixing conditions for (i) and (ii). We also establish an exponential deviation inequality with the same functional dependence in T and N.

This paper is organized as follows. Section 7.2 introduces further definitions and notations and the FFBS and FFBSi algorithms. In Section 7.3, upper bounds for the L_q-mean error and exponential deviation inequalities of these two algorithms are presented. In Section 7.4, some Monte Carlo experiments are presented to support our theoretical claims. The proofs are presented in Sections 7.5 and 7.6.

7.2 Framework

Let $\mathbb{X}$ and $\mathbb{Y}$ be two general state-spaces endowed with countably generated σ-fields $\mathcal{X}$ and $\mathcal{Y}$. Let M be a Markov transition kernel defined on $\mathbb{X} \times \mathcal{X}$ and $\{g_t\}_{t \geq 0}$ a family of functions defined on $\mathbb{X}$. It is assumed that, for any $x \in \mathbb{X}$, $M(x, \cdot)$ has a density $m(x, \cdot)$ with respect to a reference measure λ on $(\mathbb{X}, \mathcal{X})$. For any integers $T \geq 0$ and $0 \leq s \leq t \leq T$, any measurable function h on $\mathbb{X}^{t-s+1}$, and any probability distribution χ on $(\mathbb{X}, \mathcal{X})$, define

$$\phi_{s:t|T}[h] \stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_0) g_0(x_0) \prod_{u=1}^{T} M(x_{u-1}, \mathrm{d}x_u) g_u(x_u) h(x_{s:t})}{\int \chi(\mathrm{d}x_0) g_0(x_0) \prod_{u=1}^{T} M(x_{u-1}, \mathrm{d}x_u) g_u(x_u)} , \qquad (7.1)$$

where $a_{u:v}$ is a short-hand notation for $\{a_s\}_{s=u}^{v}$. The dependence on $g_{0:T}$ is implicit and is dropped from the notations.

Remark. Note that this equation has a simple interpretation in the particular case of hidden Markov models. Indeed, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{X_t\}_{t \geq 0}$ a Markov chain on $(\Omega, \mathcal{F}, \mathbb{P})$ with transition kernel M and initial distribution χ (which we denote $X_0 \sim \chi$). Let $\{Y_t\}_{t \geq 0}$ be a sequence of observations on $(\Omega, \mathcal{F}, \mathbb{P})$ conditionally independent given $\sigma(X_t, t \geq 0)$ and such that the conditional distribution of Y_u given $\sigma(X_t, t \geq 0)$ has a

density given by $g(X_u, \cdot)$ with respect to a reference measure on $\mathcal{Y}$ and set $g_u(x) = g(x, Y_u)$. Then, the quantity $\phi_{s:t|T}[h]$ defined by (7.1) is the conditional expectation of $h(X_{s:t})$ given $Y_{0:T}$:

$$\phi_{s:t|T}[h] = \mathbb{E}\left[h(X_{s:t})|Y_{0:T}\right] \ , \quad X_0 \sim \chi \ .$$

In its original version, the FFBS algorithm proceeds in two passes. In the forward pass, each filtering distribution $\phi_t \stackrel{\text{def}}{=} \phi_{t:t}$, for any $t \in \{0, \ldots, T\}$, is approximated using weighted samples $\left\{(\omega_t^{N,\ell}, \xi_t^{N,\ell})\right\}_{\ell=1}^N$, where T is the number of observations and N the number of particles: all sampled particles and weights are stored. In the backward pass of the FFBS, these importance weights are then modified (see [Doucet *et al.*, 2000, Hürzeler et Künsch, 1998, Kitagawa, 1996]) while the particle positions are kept fixed. The importance weights are updated recursively backward in time to obtain an approximation of the fixed-interval smoothing distributions $\left\{\phi_{s:T|T}\right\}_{s=0}^T$. The particle approximation is constructed as follows.

Forward pass Let $\{\xi_0^{N,\ell}\}_{\ell=1}^N$ be i.i.d. random variables distributed according to the instrumental density ρ_0 and set the importance weights $\omega_0^{N,\ell} \stackrel{\text{def}}{=} \mathrm{d}\chi/\mathrm{d}\rho_0(\xi_0^{N,\ell})\, g_0(\xi_0^{N,\ell})$. The weighted sample $\{(\xi_0^{N,\ell}, \omega_0^{N,\ell})\}_{\ell=1}^N$ then targets the initial filter ϕ_0 in the sense that

$$\phi_0^N[h] \stackrel{\text{def}}{=} \frac{1}{\sum_{\ell=1}^N \omega_0^{N,\ell}} \sum_{\ell=1}^N \omega_0^{N,\ell} h(\xi_0^{N,\ell})$$

is a consistent estimator of $\phi_0[h]$ for any bounded and measurable function h on $\mathbb{X}$.
Let now $\{(\xi_{s-1}^{N,\ell}, \omega_{s-1}^{N,\ell})\}_{\ell=1}^N$ be a weighted sample targeting ϕ_{s-1}. We aim at computing new particles and importance weights targeting the probability distribution ϕ_s. Following [Pitt et Shephard, 1999], this may be done by simulating pairs $\{(I_s^{N,\ell}, \xi_s^{N,\ell})\}_{\ell=1}^N$ of indices and particles from the instrumental distribution:

$$\pi_{s|s}(\ell, h) \propto \omega_{s-1}^{N,\ell} \vartheta_s(\xi_{s-1}^{N,\ell}) P_s(\xi_{s-1}^{N,\ell}, h) \ ,$$

on the product space $\{1, \ldots, N\} \times \mathbb{X}$, where $\{\vartheta_s(\xi_{s-1}^{N,\ell})\}_{\ell=1}^N$ are the adjustment multiplier weights and P_s is a Markovian proposal transition kernel. In the sequel, we assume that $P_s(x, \cdot)$ has, for any $x \in \mathbb{X}$, a density $p_s(x, \cdot)$ with respect to the reference measure λ. For any $\ell \in \{1, \ldots, N\}$ we associate to the particle $\xi_s^{N,\ell}$ its importance weight defined by:

$$\omega_s^{N,\ell} \stackrel{\text{def}}{=} \frac{m(\xi_{s-1}^{N,I_s^{N,\ell}}, \xi_s^{N,\ell}) g_s(\xi_s^{N,\ell})}{\vartheta_s(\xi_{s-1}^{N,I_s^{N,\ell}}) p_s(\xi_{s-1}^{N,I_s^{N,\ell}}, \xi_s^{N,\ell})} \ .$$

Backward smoothing For any probability measure η on $(\mathbb{X}, \mathcal{X})$, denote by B_η the backward smoothing kernel given, for all bounded measurable function h on $\mathbb{X}$ and for all $x \in \mathbb{X}$, by:

$$\mathrm{B}_\eta(x, h) \stackrel{\text{def}}{=} \frac{\int \eta(\mathrm{d}x')\, m(x', x) h(x')}{\int \eta(\mathrm{d}x')\, m(x', x)} ,$$

For all $s \in \{0, \ldots, T-1\}$ and for all bounded measurable function h on $\mathbb{X}^{T-s+1}$, $\phi_{s:T|T}[h]$ may be computed recursively, backward in time, according to

$$\phi_{s:T|T}[h] = \int \mathrm{B}_{\phi_s}(x_{s+1}, \mathrm{d}x_s)\, \phi_{s+1:T|T}(\mathrm{d}x_{s+1:T})\, h(x_{s:T}) .$$

7.2.1 The forward filtering backward smoothing algorithm

Consider the weighted samples $\left\{(\xi_t^{N,\ell}, \omega_t^{N,\ell})\right\}_{\ell=1}^N$, drawn for any $t \in \{0, \ldots, T\}$ in the forward pass. An approximation of the fixed-interval smoothing distribution can be obtained using

$$\phi_{s:T|T}^N[h] = \int \mathrm{B}_{\phi_s^N}(x_{s+1}, \mathrm{d}x_s)\, \phi_{s+1:T|T}^N(\mathrm{d}x_{s+1:T})\, h(x_{s:T}) , \tag{7.2}$$

and starting with $\phi_{T:T|T}^N[h] = \phi_T^N[h]$. Now, by definition, for all $x \in \mathbb{X}$ and for all bounded measurable function h on $\mathbb{X}$,

$$\mathrm{B}_{\phi_s^N}(x, h) = \sum_{i=1}^N \frac{\omega_s^{N,i} m(\xi_s^{N,i}, x)}{\sum_{\ell=1}^N \omega_s^{N,\ell} m(\xi_s^{N,\ell}, x)} h\left(\xi_s^{N,i}\right) ,$$

and inserting this expression into (7.2) gives the following particle approximation of the fixed-interval smoothing distribution $\phi_{0:T|T}[h]$

$$\phi_{0:T|T}^N[h] = \sum_{i_0=1}^N \cdots \sum_{i_T=1}^N \left(\prod_{u=1}^T \Lambda_u^N(i_u, i_{u-1}) \right) \times \frac{\omega_T^{N,i_T}}{\Omega_T^N} h\left(\xi_0^{N,i_0}, \ldots, \xi_T^{N,i_T}\right) , \tag{7.3}$$

where h is a bounded measurable function on $\mathbb{X}^{T+1}$,

$$\Lambda_t^N(i, j) \stackrel{\text{def}}{=} \frac{\omega_t^{N,j} m(\xi_t^{N,j}, \xi_{t+1}^{N,i})}{\sum_{\ell=1}^N \omega_t^{N,\ell} m(\xi_t^{N,\ell}, \xi_{t+1}^{N,i})} , \quad (i, j) \in \{1, \ldots, N\}^2 , \tag{7.4}$$

and

$$\Omega_t^N \stackrel{\text{def}}{=} \sum_{\ell=1}^N \omega_t^{N,\ell} . \tag{7.5}$$

The estimator of the fixed-interval smoothing distribution $\phi^N_{0:T|T}$ might seem impractical since the cardinality of its support is N^{T+1}. Nevertheless, for additive functionals of the form

$$S_{T,r}(x_{0:T}) = \sum_{t=r}^{T} h_t(x_{t-r:t}) \ , \tag{7.6}$$

where r is a non negative integer and $\{h_t\}_{t=r}^T$ is a family of bounded measurable functions on $\mathbb{X}^{r+1}$, the complexity of the FFBS algorithm is reduced to $O(N^{r+2})$. Furthermore, the smoothing of such functions can be computed forward in time as shown in [Del Moral *et al.*, 2010a]. This forward algorithm is exactly the one presented in [Doucet *et al.*, 2011] as an alternative to the use of the path-space method. Therefore, the results outlined in Section 7.3 hold for this method and confirm the conjecture mentioned in [Doucet *et al.*, 2011].

7.2.2 The forward filtering backward simulation algorithm

We now consider an algorithm whose complexity grows only linearly with the number of particles for any functional on $\mathbb{X}^{T+1}$. For any $t \in \{1, \dots, T\}$, we define

$$\mathcal{F}^N_t \stackrel{\text{def}}{=} \sigma\left\{(\xi^{N,i}_s, \omega^{N,i}_s); 0 \le s \le t, 1 \le i \le N\right\} \ .$$

The transition probabilities $\{\Lambda^N_t\}_{t=0}^{T-1}$ defined in (7.4) induce an inhomogeneous Markov chain $\{J_u\}_{u=0}^T$ evolving backward in time as follows. At time T, the random index J_T is drawn from the set $\{1, \dots, N\}$ with probability proportional to $(\omega^{N,1}_T, \dots, \omega^{N,N}_T)$. For any $t \in \{0, \dots, T-1\}$, the index J_t is sampled in the set $\{1, \dots, N\}$ according to $\Lambda^N_t(J_{t+1}, \cdot)$. The joint distribution of $J_{0:T}$ is therefore given, for $j_{0:T} \in \{1, \dots, N\}^{T+1}$, by

$$\mathbb{P}\left[J_{0:T} = j_{0:T} \,\middle|\, \mathcal{F}^N_T\right] = \frac{\omega^{N,j_T}_T}{\Omega^N_T} \Lambda^N_{T-1}(j_T, j_{T-1}) \dots \Lambda^N_0(j_1, j_0) \ . \tag{7.7}$$

Thus, the FFBS estimator (7.3) of the fixed-interval smoothing distribution may be written as the conditional expectation

$$\phi^N_{0:T|T}[h] = \mathbb{E}\left[h\left(\xi^{N,J_0}_0, \dots, \xi^{N,J_T}_T\right) \,\middle|\, \mathcal{F}^N_T \right] \ ,$$

where h is a bounded measurable function on $\mathbb{X}^{T+1}$. We may therefore construct an unbiased estimator of the FFBS estimator given by

$$\widetilde{\phi}^N_{0:T|T}[h] = N^{-1} \sum_{\ell=1}^{N} h\left(\xi^{N,J^\ell_0}_0, \dots, \xi^{N,J^\ell_T}_T\right) \ , \tag{7.8}$$

where $\{J_{0:T}^{\ell}\}_{\ell=1}^{N}$ are N paths drawn independently given $\mathcal{F}_T^N$ according to (7.7) and where h is a bounded measurable function on $\mathbb{X}^{T+1}$. This practical estimator was introduced in [Godsill *et al.*, 2004] (Algorithm 1, p. 158). An implementation of this estimator whose complexity grows linearly in N is introduced in [Douc *et al.*, 2011a].

7.3 Non-asymptotic deviation inequalities

In this Section, the L_q-mean error bounds and exponential deviation inequalities of the FFBS and FFBSi algorithms are established for additive functionals of the form (7.6). Our results are established under the following assumptions.

A9 (i) There exists $(\sigma_-, \sigma_+) \in (0,\infty)^2$ such that $\sigma_- < \sigma_+$ and for any $(x, x') \in \mathbb{X}^2$, $\sigma_- \leq m(x, x') \leq \sigma_+$ and we set $\rho \stackrel{\text{def}}{=} 1 - \sigma_-/\sigma_+$.

(ii) There exists $c_- \in \mathbb{R}_+^*$ such that $\int \chi(\mathrm{d}x) g_0(x) \geq c_-$ and for any $t \in \mathbb{N}^*$, $\inf_{x\in\mathbb{X}} \int M(x, \mathrm{d}x') g_t(x') \geq c_-$.

A10 (i) For all $t \geq 0$ and all $x \in \mathbb{X}$, $g_t(x) > 0$.

(ii) $\sup_{t\geq 0} |g_t|_\infty < \infty$.

A11 $\sup_{t\geq 1} |\vartheta_t|_\infty < \infty$, $\sup_{t\geq 0} |p_t|_\infty < \infty$ and $\sup_{t\geq 0} |\omega_t|_\infty < \infty$ where

$$\omega_0(x) \stackrel{\text{def}}{=} \frac{d\chi}{d\rho_0}(x) g_0(x), \quad \omega_t(x, x') \stackrel{\text{def}}{=} \frac{m(x, x') g_t(x')}{\vartheta_t(x) p_t(x, x')}, \forall t \geq 1 \, .$$

Assumptions A9 and A10 give bounds for the model and assumption A11 for quantities related to the algorithm. A9(i), referred to as the strong mixing condition, is crucial to derive time-uniform exponential deviation inequalities and a time-uniform bound of the variance of the marginal smoothing distribution (see [Del Moral et Guionnet, 2001] and [Douc *et al.*, 2011a]). For all function h from a space E to $\mathbb{R}$, $\mathrm{osc}(h)$ is defined by:

$$\mathrm{osc}(h) \stackrel{\text{def}}{=} \sup_{(z,z')\in \mathrm{E}^2} |h(z) - h(z')| \, .$$

Theorem 7.1. *Assume A9–11. For all $q \geq 2$, there exists a constant C (depending only on q, σ_-, σ_+, c_-, $\sup_{t\geq 1} |\vartheta_t|_\infty$ and $\sup_{t\geq 0} |\omega_t|_\infty$) such that for any $T < \infty$, any integer r and any bounded and measurable functions $\{h_s\}_{s=r}^{T}$,*

$$\left\| \phi_{0:T|T}^N [S_{T,r}] - \phi_{0:T|T} [S_{T,r}] \right\|_q \leq \frac{C}{\sqrt{N}} \Upsilon_{r,T}^N \left(\sum_{s=r}^{T} \mathrm{osc}(h_s)^2 \right)^{1/2} ,$$

where $S_{T,r}$ *is defined by* (7.6), $\phi^N_{0:T|T}$ *is defined by* (7.3) *and where*

$$\Upsilon^N_{r,T} \stackrel{\text{def}}{=} \sqrt{r+1}\left(\sqrt{1+r} \wedge \sqrt{T-r+1} + \frac{\sqrt{1+r}\sqrt{T-r+1}}{\sqrt{N}}\right) .$$

Similarly,

$$\left\|\widetilde{\phi}^N_{0:T|T}\left[S_{T,r}\right] - \phi_{0:T|T}\left[S_{T,r}\right]\right\|_q \leq \frac{C}{\sqrt{N}}\Upsilon^N_{r,T}\left(\sum_{s=r}^{T}\mathrm{osc}(h_s)^2\right)^{1/2} ,$$

where $\widetilde{\phi}^N_{0:T|T}$ *is defined by* (7.8).

Remark. In the particular cases where $r = 0$ and $r = T$, $\Upsilon^N_{0,T} = 1 + \sqrt{T + 1/N}$ and $\Upsilon^N_{T,T} = \sqrt{T+1}(1 + \sqrt{T + 1/N})$. Then, Theorem 7.1 gives

$$\left\|\phi^N_{0:T|T}\left[S_{T,0}\right] - \phi_{0:T|T}\left[S_{T,0}\right]\right\|_q \leq C\frac{\left(\sum_{s=0}^{T}\mathrm{osc}(h_s)^2\right)^{1/2}}{\sqrt{N}}\left(1 + \sqrt{\frac{T+1}{N}}\right) ,$$

and

$$\left\|\phi^N_{0:T|T}\left[S_{T,T}\right] - \phi_{0:T|T}\left[S_{T,T}\right]\right\|_q \leq C\sqrt{\frac{T+1}{N}}\left(1 + \sqrt{\frac{T+1}{N}}\right)\mathrm{osc}(h_T)^2 .$$

As stated in Section 7.1, theses bounds improve the results given for the FFBS estimator in [Del Moral *et al.*, 2010a].

Remark. The dependence on $1/\sqrt{N}$ is hardly surprising. Under the stated strong mixing condition, it is known that the L_q-norm of the marginal smoothing estimator $\phi^N_{t-r:t|T}[h]$, $t \in \{r, \ldots, T\}$ is uniformly bounded in time by $\left\|\phi^N_{t-r:t|T}[h]\right\|_q \leq C\mathrm{osc}(h)N^{-1/2}$ (where C depends only on q, σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$). The dependence in $\sqrt{T}$ instead of T reflects the forgetting property of the filter and the backward smoother. As for $r \leq s < t \leq T$, the estimators $\phi^N_{s-r:s|T}[h_s]$ and $\phi^N_{t-r:t|T}[h_t]$ become asymptotically independent as $(t-s)$ gets large, the L_q-norm of the sum $\sum_{t=r}^{T}\phi^N_{t-r:t|T}[h_t]$ scales as the sum of a mixing sequence (see [Davidson, 1994]).

Remark. It is easy to see that the scaling in $\sqrt{T/N}$ cannot in general be improved. Assume that the kernel m satisfies $m(x, x') = m(x')$ for all $(x, x') \in \mathbb{X} \times \mathbb{X}$. In this case, for any $t \in \{0, \ldots, T\}$, the filtering distribution is

$$\phi_t[h_t] = \frac{\int m(x)g_t(x)h_t(x)\mathrm{d}x}{\int m(x)g_t(x)\mathrm{d}x} ,$$

and the backward kernel is the identity kernel. Hence, the fixed-interval smoothing distribution coincides with the filtering distribution. If we assume that we apply the bootstrap filter for which $p_s(x, x') = m(x')$ and $\vartheta_s(x) = 1$, the estimators $\{\phi^N_{t|T}[h_t]\}_{t\in\{0,\ldots,T\}}$ are independent random variables corresponding to importance sampling estimators. It is easily seen that

$$\left\| \sum_{t=0}^{T} \phi_t^N[h_t] - \phi_t[h_t] \right\|_q \leq C \max_{0\leq t\leq T} \{\mathrm{osc}(h_t)\} \sqrt{\frac{T}{N}} \, .$$

Remark. The independent case also clearly illustrates why the path-space methods are sub-optimal (see also [Briers *et al.*, 2010] for a discussion). When applied to the independent case (for all $(x, x') \in \mathbb{X} \times \mathbb{X}$, $m(x, x') = m(x')$ and $p_s(x, x') = m(x')$), the asymptotic variance of the path-space estimators is given in [Del Moral, 2004] by

$$\begin{aligned}
&\Gamma_{0:T|T}[S_{T,0}] \\
&\quad \stackrel{\mathrm{def}}{=} \sum_{t=0}^{T-1} \frac{m(g_T^2)}{m(g_T)^2} \frac{m(g_t[h_t - \phi_t(h_t)]^2)}{m(g_t)} + \frac{m(g_T^2[h_T - \phi_T(h_T)]^2)}{m(g_T)^2} \\
&\qquad + \sum_{t=1}^{T-1} \left\{ \sum_{s=0}^{t-1} \frac{m(g_t^2)}{m(g_t)^2} \frac{m(g_s[h_s - \phi_s(h_s)]^2)}{m(g_s)} + \frac{m(g_t^2[h_t - \phi_t(h_t)]^2)}{m(g_t)^2} \right\} \\
&\qquad + \frac{\chi(g_0^2[h_0 - \phi_0(h_0)]^2)}{\chi(g_0)^2} \, .
\end{aligned}$$

The asymptotic variance thus increases as T^2 and hence, under the stated assumptions, the variance of the path-space methods is of order T^2/N. It is believed (and proved in some specific scenarios) that the same scaling holds for path-space methods for non-degenerated Markov kernel (the result has been formally established for strongly mixing kernel under the assumption that σ_-/σ_+ is sufficiently close to 1).

We provide below a brief outline of the main steps of the proofs (a detailed proof is given in Section 7.5). Following [Douc *et al.*, 2011a], the proofs rely on a decomposition of the smoothing error. For all $0 \leq t \leq T$ and all bounded and measurable function h on $\mathbb{X}^{T+1}$ define the kernel $\mathrm{L}_{t,T} : \mathbb{X}^{t+1} \times \mathcal{X}^{\otimes T+1} \to [0, 1]$ by

$$\mathrm{L}_{t,T} h(x_{0:t}) \stackrel{\mathrm{def}}{=} \int \prod_{u=t+1}^{T} M(x_{u-1}, \mathrm{d}x_u) g_u(x_u) h(x_{0:T}) \, .$$

The fixed-interval smoothing distribution may then be expressed, for all bounded and measurable function h on $\mathbb{X}^{T+1}$, by

$$\phi_{0:T|T}[h] = \frac{\phi_{0:t|t}\left[\mathrm{L}_{t,T} h\right]}{\phi_{0:t|t}\left[\mathrm{L}_{t,T} \mathbf{1}\right]} \, ,$$

and this suggests to decompose the smoothing error as follows

$$\begin{aligned}\Delta_T^N[h] &\stackrel{\text{def}}{=} \phi_{0:T|T}^N[h] - \phi_{0:T|T}[h] \\ &= \sum_{t=0}^{T} \frac{\phi_{0:t|t}^N[\mathrm{L}_{t,T}h]}{\phi_{0:t|t}^N[\mathrm{L}_{t,T}\mathbf{1}]} - \frac{\phi_{0:t-1|t-1}^N[\mathrm{L}_{t-1,T}h]}{\phi_{0:t-1|t-1}^N[\mathrm{L}_{t-1,T}\mathbf{1}]} ,\end{aligned} \tag{7.9}$$

where we used the convention

$$\frac{\phi_{0:-1|-1}^N[\mathrm{L}_{-1,T}h]}{\phi_{0:-1|-1}^N[\mathrm{L}_{-1,T}\mathbf{1}]} = \frac{\phi_0[\mathrm{L}_{0,T}h]}{\phi_0[\mathrm{L}_{0,T}\mathbf{1}]} = \phi_{0:T|T}[h] .$$

Furthermore, for all $0 \le t \le T$,

$$\begin{aligned}\phi_{0:t|t}^N[\mathrm{L}_{t,T}h] &= \int \phi_{0:t|t}^N(\mathrm{d}x_{0:t})\mathrm{L}_{t,T}h(x_{0:t}) \\ &= \int \phi_t^N(\mathrm{d}x_t)\mathrm{B}_{\phi_{t-1}^N}(x_t,\mathrm{d}x_{t-1})\cdots\mathrm{B}_{\phi_0^N}(x_1,\mathrm{d}x_0)\mathrm{L}_{t,T}h(x_{0:t}) \\ &= \int \phi_t^N(\mathrm{d}x_t)\mathcal{L}_{t,T}^N h(x_t) ,\end{aligned}$$

where $\mathcal{L}_{t,T}^N$ and $\mathcal{L}_{t,T}$ are two kernels on $\mathbb{X}\times\mathcal{X}^{\otimes(T+1)}$ defined for all $x_t\in\mathbb{X}$ by

$$\mathcal{L}_{t,T}h(x_t) \stackrel{\text{def}}{=} \int \mathrm{B}_{\phi_{t-1}}(x_t,\mathrm{d}x_{t-1})\cdots\mathrm{B}_{\phi_0}(x_1,\mathrm{d}x_0)\mathrm{L}_{t,T}h(x_{0:t}) \tag{7.10}$$

$$\mathcal{L}_{t,T}^N h(x_t) \stackrel{\text{def}}{=} \int \mathrm{B}_{\phi_{t-1}^N}(x_t,\mathrm{d}x_{t-1})\cdots\mathrm{B}_{\phi_0^N}(x_1,\mathrm{d}x_0)\mathrm{L}_{t,T}h(x_{0:t}) . \tag{7.11}$$

For all $1\le t\le T$ we can write

$$\begin{aligned}\frac{\phi_{0:t|t}^N[\mathrm{L}_{t,T}h]}{\phi_{0:t|t}^N[\mathrm{L}_{t,T}\mathbf{1}]} - \frac{\phi_{0:t-1|t-1}^N[\mathrm{L}_{t-1,T}h]}{\phi_{0:t-1|t-1}^N[\mathrm{L}_{t-1,T}\mathbf{1}]} &= \frac{\phi_t^N[\mathcal{L}_{t,T}^N h]}{\phi_t^N[\mathcal{L}_{t,T}^N\mathbf{1}]} - \frac{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N h]}{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N\mathbf{1}]} \\ &= \frac{1}{\phi_t^N[\mathcal{L}_{t,T}^N\mathbf{1}]}\left(\phi_t^N[\mathcal{L}_{t,T}^N h] - \frac{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N h]}{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N\mathbf{1}]}\phi_t^N[\mathcal{L}_{t,T}^N\mathbf{1}]\right) ,\end{aligned}$$

and then,

$$\Delta_T^N[h] = \sum_{t=0}^{T} \frac{N^{-1}\sum_{\ell=1}^N \omega_t^{N,\ell} G_{t,T}^N h(\xi_t^{N,\ell})}{N^{-1}\sum_{\ell=1}^N \omega_t^{N,\ell}\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})} , \tag{7.12}$$

with $G_{t,T}^N$ is a kernel on $\mathbb{X}\times\mathcal{X}^{\otimes(T+1)}$ defined, for all $x_t\in\mathbb{X}$ and all bounded and measurable function h on $\mathbb{X}^{T+1}$, by

$$G_{t,T}^N h(x_t) \stackrel{\text{def}}{=} \mathcal{L}_{t,T}^N h(x_t) - \frac{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N h]}{\phi_{t-1}^N[\mathcal{L}_{t-1,T}^N\mathbf{1}]}\mathcal{L}_{t,T}^N\mathbf{1}(x_t) ,$$

where, by the same convention as above,

$$G_{0,T}^N h(x_0) \stackrel{\text{def}}{=} \mathrm{L}_{0,T} h(x_0) - \frac{\phi_0[\mathcal{L}_{0,T} h]}{\phi_0[\mathcal{L}_{0,T}\mathbf{1}]} \mathcal{L}_{0,T}\mathbf{1}(x_0) \; .$$

Two families of random variables $\left\{C_{t,T}^N(f)\right\}_{t=0}^T$ and $\left\{D_{t,T}^N(f)\right\}_{t=0}^T$ are now introduced to transform (7.12) into a suitable decomposition to compute an upper bound for the L_q-mean error. As shown in Lemma 7.1, the random variables $\{\omega_t^{N,\ell} G_{t,T}^N f(\xi_t^{N,\ell})\}_{\ell=1}^N$ are centered given $\mathcal{F}_{t-1}^N$. The idea is to replace $N^{-1}\sum_{\ell=1}^N \omega_t^{N,\ell} \mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})$ in (7.12) by its conditional expectation given $\mathcal{F}_{t-1}^N$ to get a martingale difference. This conditional expectation is computed using the following intermediate result. For any measurable function h on $\mathbb{X}$ and any $t \in \{0, \ldots, T\}$,

$$\mathbb{E}\left[\omega_t^{N,1} h(\xi_t^{N,1}) \middle| \mathcal{F}_{t-1}^N\right] = \frac{\phi_{t-1}^N\left[M g_t h\right]}{\phi_{t-1}^N[\vartheta_t]} \; . \tag{7.13}$$

Indeed,

$$\begin{aligned}
&\mathbb{E}\left[\omega_t^{N,1} h(\xi_t^{N,1}) \middle| \mathcal{F}_{t-1}^N\right] \\
&= \mathbb{E}\left[\frac{m(\xi_{t-1}^{N,I_t^{N,1}}, \xi_t^{N,1}) g_t(\xi_t^{N,1})}{\vartheta_t(\xi_{t-1}^{N,I_t^{N,1}}) p_t(\xi_{t-1}^{N,I_t^{N,1}}, \xi_t^{N,1})} h(\xi_t^{N,1}) \middle| \mathcal{F}_{t-1}^N\right] \\
&= \left(\sum_{i=1}^N \omega_{t-1}^{N,i} \vartheta_t(\xi_{t-1}^{N,i})\right)^{-1} \sum_{i=1}^N \int \omega_{t-1}^{N,i} \vartheta_t(\xi_{t-1}^{N,i}) p_t(\xi_{t-1}^{N,i}, x) \frac{M(\xi_{t-1}^{N,i}, \mathrm{d}x) g_t(x)}{\vartheta_t(\xi_{t-1}^{N,i}) p_t(\xi_{t-1}^{N,i}, x)} h(x) \\
&= \left(\sum_{i=1}^N \omega_{t-1}^{N,i} \vartheta_t(\xi_{t-1}^{N,i})\right)^{-1} \sum_{i=1}^N \int \omega_{t-1}^{N,i} M(\xi_{t-1}^{N,i}, \mathrm{d}x) g_t(x) h(x) \\
&= \frac{\phi_{t-1}^N\left[M g_t h\right]}{\phi_{t-1}^N[\vartheta_t]} \; .
\end{aligned}$$

This result, applied with the function $h = \mathcal{L}_{t,T}\mathbf{1}$, yields

$$\mathbb{E}\left[\omega_t^{N,1} \mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,1}) \middle| \mathcal{F}_{t-1}^N\right] = \frac{\phi_{t-1}^N\left[M g_t \mathcal{L}_{t,T}\mathbf{1}\right]}{\phi_{t-1}^N[\vartheta_t]} = \frac{\phi_{t-1}^N\left[\mathcal{L}_{t-1,T}\mathbf{1}\right]}{\phi_{t-1}^N[\vartheta_t]} \; .$$

For any $0 \le t \le T$, define for all bounded and measurable function h on $\mathbb{X}^{T+1}$,

$$\begin{aligned}
D_{t,T}^N(h) &\stackrel{\text{def}}{=} \mathbb{E}\left[\omega_t^{N,1} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,1})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \middle| \mathcal{F}_{t-1}^N\right]^{-1} N^{-1} \sum_{\ell=1}^N \omega_t^{N,\ell} \frac{G_{t,T}^N h(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \qquad (7.14) \\
&= \frac{\phi_{t-1}^N[\vartheta_t]}{\phi_{t-1}^N\left[\frac{\mathcal{L}_{t-1,T}\mathbf{1}}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}\right]} N^{-1} \sum_{\ell=1}^N \omega_t^{N,\ell} \frac{G_{t,T}^N h(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \; ,
\end{aligned}$$

$$C_{t,T}^N(h) \stackrel{\text{def}}{=} \left[\frac{1}{N^{-1}\sum_{i=1}^N \omega_t^{N,i} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,i})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}} - \frac{\phi_{t-1}^N[\vartheta_t]}{\phi_{t-1}^N\left[\frac{\mathcal{L}_{t-1,T}\mathbf{1}}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}\right]} \right] \times N^{-1} \sum_{\ell=1}^N \omega_t^{N,\ell} \frac{G_{t,T}^N h(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \,. \quad (7.15)$$

Using these notations, (7.12) can be rewritten as follows:

$$\Delta_T^N[h] = \sum_{t=0}^T D_{t,T}^N(h) + \sum_{t=0}^T C_{t,T}^N(h) \,. \quad (7.16)$$

For any $q \geq 2$, the derivation of the upper bound relies on the triangle inequality:

$$\left\|\Delta_T^N[S_{T,r}]\right\|_q \leq \left\| \sum_{t=0}^T D_{t,T}^N(S_{T,r}) \right\|_q + \sum_{t=0}^T \left\| C_{t,T}^N(S_{T,r}) \right\|_q \,,$$

where $S_{T,r}$ is defined in (7.6). The proof for the FFBS estimator $\phi_{0:T|T}^N$ is completed by using Proposition 7.1 and Proposition 7.2. According to (7.16), the smoothing error can be decomposed into a sum of two terms which are considered separately. The first one is a martingale whose L_q-mean error is upper-bounded by $\sqrt{(T+1)/N}$ as shown in Proposition 7.1. The second one is a sum of products, L_q-norm of which being bounded by $1/N$ in Proposition 7.2.

The end of this section is devoted to the exponential deviation inequality for the error $\Delta_T^N[S_{T,r}]$ defined by (7.9). We use the decomposition of $\Delta_T^N[S_{T,r}]$ obtained in (7.16) leading to a similar dependence on the ratio $(T+1)/N$. The martingale term $D_{t,T}^N(S_{T,r})$ is dealt with using the Azuma-Hoeffding inequality while the term $C_{t,T}^N(S_{T,r})$ needs a specific Hoeffding-type inequality for ratio of random variables.

Theorem 7.2. *Assume A9–11. There exists a constant C (depending only on σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$) such that for any $T < \infty$, any $N \geq 1$, any $\varepsilon > 0$, any integer r, and any bounded and measurable functions $\{h_s\}_{s=r}^T$,*

$$\mathbb{P}\left\{ \left| \phi_{0:T|T}\left[S_{T,r}\right] - \phi_{0:T|T}^N\left[S_{T,r}\right] \right| > \varepsilon \right\} \leq 2\exp\left(-\frac{CN\varepsilon^2}{\Theta_{r,T}\sum_{s=r}^T \mathrm{osc}(h_s)^2} \right) + 8\exp\left(-\frac{CN\varepsilon}{(1+r)\sum_{s=r}^T \mathrm{osc}(h_s)} \right) \,,$$

where $S_{T,r}$ *is defined by* (7.6), $\phi^N_{0:T|T}$ *is defined by* (7.3) *and where*

$$\Theta_{r,T} \stackrel{\text{def}}{=} (1+r)\left\{(1+r) \wedge (T-r+1)\right\} \ . \tag{7.17}$$

Similarly,

$$\mathbb{P}\left\{\left|\phi_{0:T|T}\left[S_{T,r}\right] - \widetilde{\phi}^N_{0:T|T}\left[S_{T,r}\right]\right| > \varepsilon\right\}$$
$$\leq 4\exp\left(-\frac{CN\varepsilon^2}{\Theta_{r,T}\sum_{s=r}^{T}\mathrm{osc}(h_s)^2}\right) + 8\exp\left(-\frac{CN\varepsilon}{(1+r)\sum_{s=r}^{T}\mathrm{osc}(h_s)}\right) \ ,$$

where $\widetilde{\phi}^N_{0:T|T}$ *is defined by* (7.8).

7.4 Monte-Carlo Experiments

In this section, the performance of the FFBSi algorithm is evaluated through simulations and compared to the path-space method.

7.4.1 Linear gaussian model

Let us consider the following model:

$$\begin{cases} X_{t+1} &= \phi X_t + \sigma_u U_t \ , \\ Y_t &= X_t + \sigma_v V_t \ , \end{cases}$$

where X_0 is a zero-mean random variable with variance $\frac{\sigma_u^2}{1-\phi^2}$, $\{U_t\}_{t\geq 0}$ and $\{V_t\}_{t\geq 0}$ are two sequences of independent and identically distributed standard gaussian random variables (independent from X_0). The parameters $(\phi, \sigma_u, \sigma_v)$ are assumed to be known. Observations were generated using $\phi = 0.9$, $\sigma_u = 0.6$ and $\sigma_v = 1$. Table 7.1 provides the empirical variance of the estimation of the unnormalized smoothed additive functional $\mathcal{I}_T \stackrel{\text{def}}{=} \sum_{t=0}^{T}\mathbb{E}\left[X_t|Y_{0:T}\right]$ given by the path-space and the FFBSi methods over 250 independent Monte Carlo experiments. We display in Figure 7.1 the empirical variance for different values of N as a function of T for both estimators. These estimates are represented by dots and a linear regression (resp. quadratic regression) is also provided for the FFBSi algorithm (resp. for the path-space method).

In Figure 7.2 the FFBSi algorithm is compared to the path-space method to compute the smoothed value of the empirical mean $(T+1)^{-1}\mathcal{I}_T$. For the purpose of comparison, this quantity is computed using the Kalman smoother. We display in Figure 7.2 the box and whisker plots of the estimations obtained with 100 independent Monte Carlo experiments. The FFBSi

Table 7.1: Empirical variance for different values of T and N.

Path-space

T \ N	300	500	750	1000	1500	5000	10000	15000	20000
300	137.8	119.4	63.7	46.1	36.2	12.8	7.1	3.8	3.0
500	290.0	215.3	192.5	161.9	80.3	30.1	14.9	11.3	7.4
750	474.9	394.5	332.9	250.5	206.8	71.0	35.6	24.4	21.7
1000	673.7	593.2	505.1	483.2	326.4	116.4	70.8	37.9	34.6
1500	1274.6	1279.7	916.7	804.7	655.1	233.9	163.1	89.7	80.0

FFBSi

T \ N	300	500	750	1000	1500
300	5.1	3.1	2.3	1.4	1.0
500	9.7	5.1	3.7	2.6	2.2
750	11.2	7.1	4.9	3.7	2.6
1000	16.5	10.5	6.7	5.1	3.4
1500	25.6	14.1	7.8	6.8	5.1

algorithm clearly outperforms the other method for comparable computational costs. In Table 7.2, the mean CPU times over the 100 runs of the two methods are given as a function of the number of particles (for $T = 500$ and $T = 1000$).

Table 7.2: Average CPU time to compute the smoothed value of the empirical mean in the LGM

$T = 500$	FFBSi	Path-space method		
N	500	500	5000	10000
CPU time (s)	4.87	0.24	2.47	4.65

$T = 1000$	FFBSi	Path-space method		
N	1000	1000	10000	20000
CPU time (s)	16.5	0.9	8.5	17.2

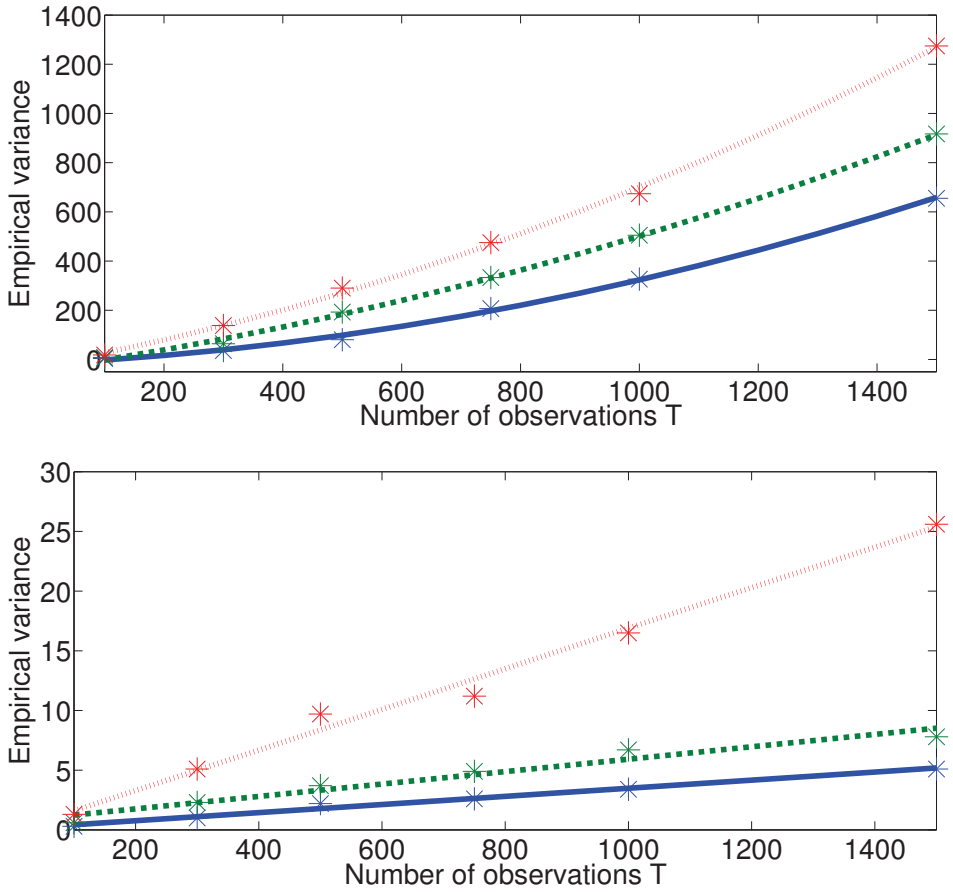


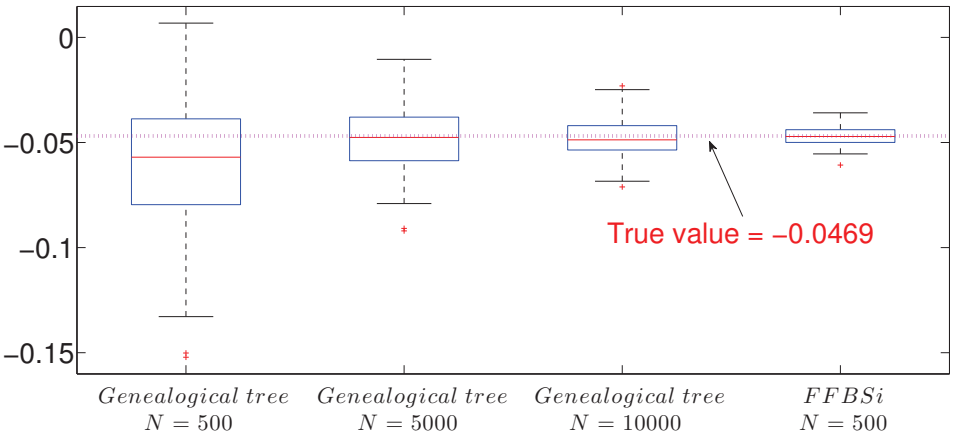
Figure 7.1: Empirical variance of the path-space (top) and FFBSi (bottom) for $N = 300$ (dotted line), $N = 750$ (dashed line) and $N = 1500$ (bold line).

7.4.2 Stochastic Volatility Model

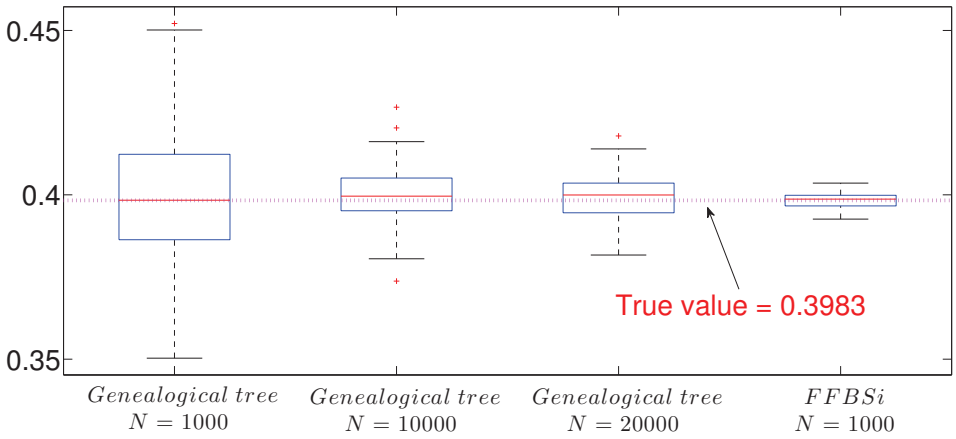
Stochastic volatility models (SVM) have been introduced to provide better ways of modeling financial time series data than ARCH/GARCH models ([Hull et White, 1987]). We consider the elementary SVM model introduced by [Hull et White, 1987]:

$$\begin{cases} X_{t+1} = \phi X_t + \sigma U_{t+1} \ , \\ Y_t = \beta \mathrm{e}^{\frac{X_t}{2}} V_t \ , \end{cases}$$

where X_0 is a zero-mean random variable with variance $\frac{\sigma_u^2}{1-\phi^2}$, $\{U_t\}_{t\geq 0}$ and $\{V_t\}_{t\geq 0}$ are two sequences of independent and identically distributed standard gaussian random variables (independent from X_0). This model was used to generate simulated data with parameters $(\phi = 0.3, \sigma = 0.5, \beta = 1)$ assumed to be known in the following experiments. The empirical variance



(a) Time $T = 500$



(b) Time $T = 1000$

Figure 7.2: Computation of smoothed additive functionals in a linear gaussian model. The variance of the estimation given by the FFBSi algorithm is the smallest one in both cases.

of the estimation of $\mathcal{I}_T$ given by the path-space and the FFBSi methods over 250 independent Monte Carlo experiments is displayed in Table 7.3. We display in Figure 7.3 the empirical variance for different values of N as a function of T for both estimators.

Table 7.3: Empirical variance for different values of T and N in the SVM.

Path-space method

T \ N	300	500	750	1000	1500	5000	10000	15000	20000
300	52.7	33.7	22.0	17.8	12.3	3.8	2.0	1.4	1.2
500	116.3	84.8	64.8	53.5	30.7	11.4	6.8	4.1	2.8
750	184.7	187.6	134.2	120.0	65.8	29.1	12.8	7.3	7.7
1000	307.7	240.4	244.7	182.8	133.2	43.6	24.5	15.6	11.6
1500	512.1	487.5	445.5	359.9	249.5	90.9	52.0	32.6	29.3

FFBSi

T \ N	300	500	750	1000	1500
300	1.2	0.6	0.5	0.4	0.2
500	2.1	1.2	0.8	0.6	0.4
750	3.7	1.8	1.4	0.9	0.6
1000	4.0	2.7	1.8	1.3	0.9
1500	7.3	3.8	3.1	1.6	1.4

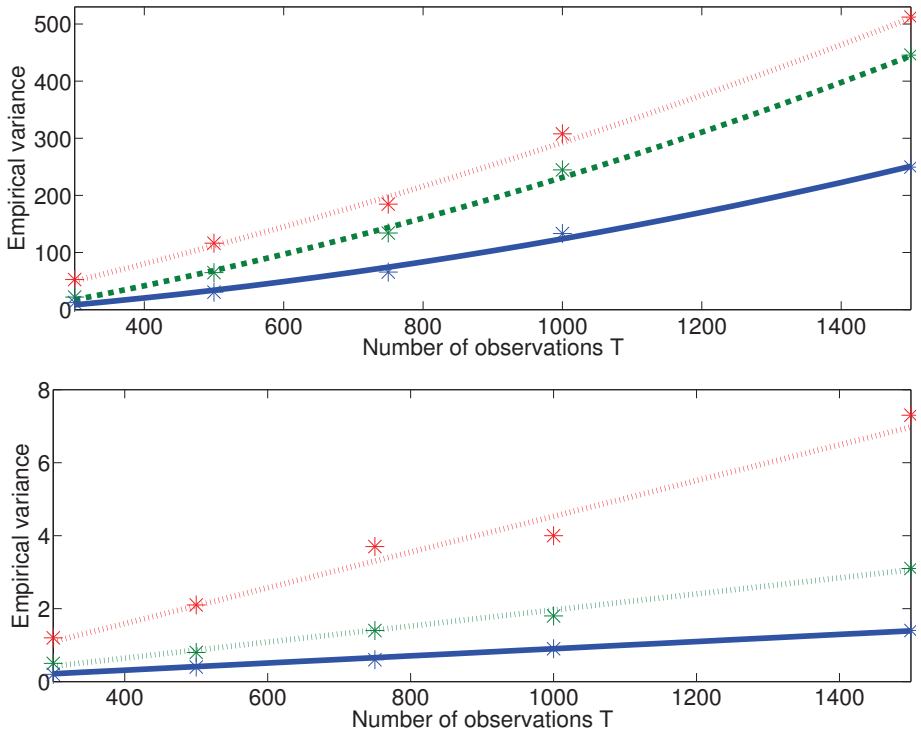


Figure 7.3: Empirical variance of the path-space (top) and FFBSi (bottom) for $N = 300$ (dotted line), $N = 750$ (dashed line) and $N = 1500$ (bold line) in the SVM.

7.5 Proof of Theorem 7.1

We preface the proof of Proposition 7.1 by the following Lemma:

Lemma 7.1. *Under assumptions A9–11, we have, for any $t \in \{0, \dots, T\}$ and any measurable function h on $\mathbb{X}^{T+1}$:*

(i) The random variables $\left\{ \omega_t^{N,\ell} \dfrac{G_{t,T}^N h(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \right\}_{\ell=1}^{N}$ *are, for all $N \in \mathbb{N}$:*

(a) conditionally independent and identically distributed given $\mathcal{F}_{t-1}^N$,

(b) centered conditionally to $\mathcal{F}_{t-1}^N$.

where $G_{t,T}^N h$ is defined in (7.3) *and $\mathcal{L}_{t,T}^N$ is defined in* (7.11).

(ii) For any integers r, t and N:

$$\left| \frac{G_{t,T}^N S_{T,r}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \right| \leq \sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s) \ , \tag{7.18}$$

where $S_{T,r}$ and ρ are respectively defined in (7.6) *and in A9*(i).

(iii) For all $x \in \mathbb{X}$, $\dfrac{\mathcal{L}_{t,T}\mathbf{1}(x)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \geq \dfrac{\sigma_-}{\sigma_+}$ *and* $\dfrac{\mathcal{L}_{t-1,T}\mathbf{1}(x)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \geq c_- \dfrac{\sigma_-}{\sigma_+}$.

Proof. The proof of (i) is given by [Douc *et al.*, 2011a, Lemma 3].

Proof of (ii). Let $\Pi_{s-r:s,T}$ be the operator which associates to any bounded and measurable function h on $\mathbb{X}^{r+1}$ the function $\Pi_{s-r:s,T} h$ given, for any $(x_0, \dots, x_T) \in \mathbb{X}^{T+1}$, by

$$\Pi_{s-r:s,T} h(x_{0:T}) \stackrel{\text{def}}{=} h(x_{s-r:s}) \ .$$

Then, we write $S_{T,r} = \sum_{s=r}^{T} \Pi_{s-r:s,T} h_s$ and $G_{t,T}^N S_{T,r} = \sum_{s=r}^{T} G_{t,T}^N \Pi_{s-r:s,T} h_s$. By (7.3), we have

$$\frac{G_{t,T}^N \Pi_{s-r:s,T} h_s(x_t)}{\mathcal{L}_{t,T}^N \mathbf{1}(x_t)} = \frac{\mathcal{L}_{t,T}^N \Pi_{s-r:s,T} h_s(x_t)}{\mathcal{L}_{t,T}^N \mathbf{1}(x_t)} - \frac{\phi_{t-1}^N [\mathcal{L}_{t-1,T}^N \Pi_{s-r:s,T} h_s]}{\phi_{t-1}^N [\mathcal{L}_{t-1,T}^N \mathbf{1}]} \ ,$$

and, following the same lines as in [Douc *et al.*, 2011a, Lemma 10],

$$|G_{t,T}^N \Pi_{s-r:s,T} h_s|_\infty \leq \rho^{s-r-t} \mathrm{osc}(h_s) |\mathcal{L}_{t,T}\mathbf{1}|_\infty \quad \text{if} \quad t \leq s - r \ ,$$

$$|G_{t,T}^N \Pi_{s-r:s,T} h_s|_\infty \leq \rho^{t-s} \mathrm{osc}(h_s) |\mathcal{L}_{t,T}\mathbf{1}|_\infty \quad \text{if} \quad t > s \ ,$$

where ρ is defined in A9(i). Furthermore, for any $s - r < t \leq s$,

$$|G_{t,T}^N \Pi_{s-r:s,T} h_s|_\infty \leq \mathrm{osc}(h_s) |\mathcal{L}_{t,T}\mathbf{1}|_\infty \ ,$$

which shows (ii).

Proof of (iii). From the definition (7.10), for all $x \in \mathbb{X}$ and all $t \in \{1, \ldots, T\}$,

$$\mathcal{L}_{t,T}\mathbf{1}(x) = \int m(x, x_{t+1}) g_{t+1}(x_{t+1}) \prod_{u=t+2}^{T} M(x_{u-1}, \mathrm{d}x_u) g_u(x_u) \lambda(\mathrm{d}x_{t+1}) \; ,$$

hence, by assumption A9,

$$|\mathcal{L}_{t,T}\mathbf{1}|_\infty \leq \sigma_+ \int g_{t+1}(x_{t+1}) \mathcal{L}_{t+1,T}\mathbf{1}(x_{t+1}) \lambda(\mathrm{d}x_{t+1})$$
$$\mathcal{L}_{t,T}\mathbf{1}(x) \geq \sigma_- \int g_{t+1}(x_{t+1}) \mathcal{L}_{t+1,T}\mathbf{1}(x_{t+1}) \lambda(\mathrm{d}x_{t+1}) \; ,$$

which concludes the proof of the first statement. By construction, for any $x \in \mathbb{X}$ and any $t \in \{1, \ldots, T\}$,

$$\mathcal{L}_{t-1,T}\mathbf{1}(x) = \int M(x, \mathrm{d}x') g_t(x') \mathcal{L}_{t,T}\mathbf{1}(x') \; ,$$

and then, by assumption A9,

$$\frac{\mathcal{L}_{t-1,T}\mathbf{1}(x)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} = \int M(x, \mathrm{d}x') g_t(x') \frac{\mathcal{L}_{t,T}\mathbf{1}(x')}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \geq c_- \frac{\sigma_-}{\sigma_+} \; .$$

□

Proposition 7.1. *Assume A9–11. For all $q \geq 2$, there exists a constant C (depending only on q, σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$) such that for any $T < \infty$, any integer r and any bounded and measurable functions $\{h_s\}_{s=r}^T$ on $\mathbb{X}^{r+1}$,*

$$\left\| \sum_{t=0}^{T} D_{t,T}^N(S_{T,r}) \right\|_q \leq \frac{C}{\sqrt{N}} \sqrt{1+r} \left(\sqrt{1+r} \wedge \sqrt{T-r+1} \right) \left(\sum_{s=r}^{T} \mathrm{osc}(h_s)^2 \right)^{1/2} , \tag{7.19}$$

where $D_{t,T}^N$ is defined in (7.14).

Proof. Since $\left\{ D_{t,T}^N(S_{T,r}) \right\}_{0\leq t\leq T}$ is a is a forward martingale difference and $q \geq 2$, Burkholder's inequality (see [Hall et Heyde, 1980, Theorem 2.10, page 23]) states the existence of a constant C depending only on q such that:

$$\mathbb{E}\left[\left| \sum_{t=0}^{T} D_{t,T}^N(S_{T,r}) \right|^q \right] \leq C\mathbb{E}\left[\left| \sum_{t=0}^{T} D_{t,T}^N(S_{T,r})^2 \right|^{\frac{q}{2}} \right] \; .$$

Moreover, by application of the last statement of Lemma 7.1(iii),

$$\frac{\phi_{t-1}^N[\vartheta_t]}{\phi_{t-1}^N\left[\frac{\mathcal{L}_{t-1,T}\mathbf{1}}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}\right]} \leq \frac{\sigma_+ \sup_{t\geq 0} |\vartheta_t|_\infty}{\sigma_- c_-} \; ,$$

and thus,

$$\mathbb{E}\left[\left|\sum_{t=0}^{T} D_{t,T}^N(S_{T,r})^2\right|^{\frac{q}{2}}\right] \leq \left(\frac{\sigma_+ \sup_{t\geq 0}|\vartheta_t|_\infty}{\sigma_- c_-}\right)^q \mathbb{E}\left[\left|\sum_{t=0}^{T}\left(N^{-1}\sum_{\ell=1}^{N} a_{t,T}^{N,\ell}\right)^2\right|^{\frac{q}{2}}\right],$$

where $a_{t,T}^{N,\ell} \stackrel{\text{def}}{=} \omega_t^{N,\ell} \frac{G_{t,T}^N S_{T,r}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}$. By the Minkowski inequality,

$$\left\|\sum_{t=0}^{T} D_{t,T}^N(S_{T,r})\right\|_q \leq C\left\{\sum_{t=0}^{T}\left(\mathbb{E}\left[\left|N^{-1}\sum_{\ell=1}^{N} a_{t,T}^{N,\ell}\right|^q\right]\right)^{2/q}\right\}^{1/2}. \tag{7.20}$$

Since for any $t \geq 0$ the random variables $\left\{a_{t,T}^{N,\ell}\right\}_{\ell=1}^{N}$ are conditionally independent and centered conditionally to $\mathcal{F}_{t-1}^N$, using again the Burkholder and the Jensen inequalities we obtain

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{\ell=1}^{N} a_{t,T}^{N,\ell}\right|^q \middle| \mathcal{F}_{t-1}^N\right] &\leq CN^{q/2-1}\sum_{\ell=1}^{N}\mathbb{E}\left[\left|a_{t,T}^{N,\ell}\right|^q \middle| \mathcal{F}_{t-1}^N\right] \\ &\leq C\left[\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)\right]^q N^{q/2}, \end{aligned} \tag{7.21}$$

where the last inequality comes from (7.18). Finally, by (7.20) and (7.21) we get

$$\left\|\sum_{t=0}^{T} D_{t,T}^N(S_{T,r})\right\|_q \leq CN^{-1/2}\left\{\sum_{t=0}^{T}\left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)\right)^2\right\}^{1/2}.$$

By the Holder inequality, we have

$$\begin{aligned}&\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s) \\ &\qquad\leq \left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\right)^{1/2} \times \left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)^2\right)^{1/2} \\ &\qquad\leq C\sqrt{1+r}\left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)^2\right)^{1/2},\end{aligned}$$

which yields

$$\left\|\sum_{t=0}^{T} D_{t,T}^N(S_{T,r})\right\|_q \leq CN^{-1/2}(1+r)\left(\sum_{s=r}^{T}\mathrm{osc}(h_s)^2\right)^{1/2}.$$

We obtain similarly

$$\left\| \sum_{t=0}^{T} D_{t,T}^N(S_{T,r}) \right\|_q \leq CN^{-1/2}(1+r)^{1/2} \sum_{s=r}^{T} \mathrm{osc}(h_s) \ ,$$

which concludes the proof. □

Proposition 7.2. *Assume A9–11. For all $q \geq 2$, there exists a constant C (depending only on q, σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$) such that for any $T < +\infty$, any $0 \leq t \leq T$, any integer r, and any bounded and measurable functions $\{h_s\}_{s=r}^T$ on $\mathbb{X}^{r+1}$,*

$$\left\| C_{t,T}^N(S_{T,r}) \right\|_q \leq \frac{C}{N} \sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s) \ , \tag{7.22}$$

where $C_{t,T}^N$ is defined in (7.15).

Proof. According to (7.15), $C_{t,T}^N(S_{T,r})$ can be written

$$C_{t,T}^N(S_{T,r}) = U_{t,T}^N V_{t,T}^N W_{t,T}^N \ , \tag{7.23}$$

where

$$\begin{aligned}
U_{t,T}^N &= \frac{N^{-1} \sum_{\ell=1}^N \omega_t^{N,\ell} \frac{G_{t,T}^N S_{T,r}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}}{N^{-1}\Omega_t^N} \ , \\
V_{t,T}^N &= N^{-1} \sum_{\ell=1}^N \left(\mathbb{E}\left[\omega_t^{N,1} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,1})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \middle| \mathcal{F}_{t-1} \right] - \omega_t^{N,\ell} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \right) \ , \\
W_{t,T}^N &= \frac{N^{-1}\Omega_t^N}{\mathbb{E}\left[\omega_t^{N,1} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,1})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \middle| \mathcal{F}_{t-1} \right] N^{-1} \sum_{\ell=1}^N \omega_t^{N,\ell} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}} \ ,
\end{aligned}$$

and where Ω_t^N is defined by (7.5). Using the last statement of Lemma 7.1, we get the following bound:

$$\begin{aligned}
\mathbb{E}\left[\omega_t^{N,1} \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,1})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \middle| \mathcal{F}_{t-1} \right] &= \frac{\phi_{t-1}^N \left[\mathcal{L}_{t-1,T}\mathbf{1} / |\mathcal{L}_{t,T}\mathbf{1}|_\infty \right]}{\phi_{t-1}^N[\vartheta_t]} \\
&\geq \frac{c_-\sigma_-}{|\vartheta_t|_\infty \, \sigma_+} \\
\frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} &\geq \frac{\sigma_-}{\sigma_+} \ ,
\end{aligned}$$

which implies

$$\left| W_{t,T}^N \right| \leq \left(\frac{\sigma_+}{\sigma_-} \right)^2 \frac{|\vartheta_t|_\infty}{c_-} \ . \tag{7.24}$$

Then, $\left|C^N_{t,T}(S_{T,r})\right| \leq C \left|U^N_{t,T}\right| \left|V^N_{t,T}\right|$ and we can use the decomposition

$$U^N_{t,T} V^N_{t,T} = V^N_{t,T} \left[\frac{N^{-1}\sum_{\ell=1}^N a^{N,\ell}_{t,T}}{\mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right]} + \frac{N^{-1}\sum_{\ell=1}^N a^{N,\ell}_{t,T}}{\widetilde{\Omega}^N_t \mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right]} \left(\mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right] - \widetilde{\Omega}^N_t \right) \right] ,$$

where $a^{N,\ell}_{t,T} \stackrel{\text{def}}{=} \omega^{N,\ell}_t \frac{G^N_{t,T} S_{T,r}(\xi^{N,\ell}_t)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}$ and $\widetilde{\Omega}^N_t \stackrel{\text{def}}{=} N^{-1}\Omega^N_t$. By (7.13), we have $\mathbb{E}\left[\omega^{N,1}_t \middle| \mathcal{F}^N_{t-1}\right] = \frac{\phi^N_{t-1}[Mg_t]}{\phi^N_{t-1}[\vartheta_t]}$ and then, by A9(ii), A11 and (7.18),

$$\frac{1}{\mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right]} \leq \frac{|\vartheta_t|_\infty}{c_-}$$

and

$$\frac{N^{-1}\sum_{\ell=1}^N a^{N,\ell}_{t,T}}{\widetilde{\Omega}^N_t \mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right]} \leq C \frac{|\vartheta_t|_\infty}{c_-} \sum_{s=r}^T \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s) .$$

Therefore,

$$\left|C^N_{t,T}(S_{T,r})\right| \leq C \left(C^{1,N}_{t,T} + \sum_{s=r}^T \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s) C^{2,N}_{t,T} \right) ,$$

where

$$C^{1,N}_{t,T} \stackrel{\text{def}}{=} V^N_{t,T} \cdot N^{-1} \sum_{\ell=1}^N a^{N,\ell}_{t,T} \quad \text{and} \quad C^{2,N}_{t,T} \stackrel{\text{def}}{=} V^N_{t,T} \left| \mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right] - \widetilde{\Omega}^N_t \right| .$$

The random variables $\left\{ \omega^{N,\ell}_t \frac{\mathcal{L}_{t,T}\mathbf{1}(\xi^{N,\ell}_t)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \right\}_{\ell=1}^N$ being bounded and conditionally independent given $\mathcal{F}^N_{t-1}$, following the same steps as in the proof of Proposition 7.1, there exists a constant C (depending only on q, σ_-, σ_+, c_- and $\sup_{t\geq 0} |\omega_t|_\infty$) such that $\left\| V^N_{t,T} \right\|_{2q} \leq C N^{-1/2}$. Similarly

$$\left\| N^{-1} \sum_{\ell=1}^N a^{N,\ell}_{t,T} \right\|_{2q} \leq C \frac{\sum_{s=r}^T \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s)}{N^{1/2}} ,$$

and

$$\left\| \mathbb{E}\left[\widetilde{\Omega}^N_t \middle| \mathcal{F}_{t-1}\right] - \widetilde{\Omega}^N_t \right\|_{2q} \leq \frac{C}{N^{1/2}} .$$

The Cauchy-Schwarz inequality concludes the proof of (7.22). □

The proof of Theorem 7.1 is now concluded for the FFBS estimator $\phi^N_{0:T|T}[S_{T,r}]$ and we can proceed to the proof for the FFBSi estimator. We preface the proof of Theorem 7.1 for the FFBSi estimator $\widetilde{\phi}^N_{0:T|T}$ by the following Lemma. We first define the backward filtration $\left\{\mathcal{G}^N_{t,T}\right\}_{t=0}^{T+1}$ by

$$\begin{cases} \mathcal{G}^N_{T+1,T} \stackrel{\text{def}}{=} \mathcal{F}^N_T \ , \\ \mathcal{G}^N_{t,T} \stackrel{\text{def}}{=} \mathcal{F}^N_T \vee \sigma\left\{J^\ell_u, 1 \le \ell \le N, t \le u \le T\right\}, \quad \forall\, t \in \{0,\ldots,T\} \ . \end{cases}$$

Lemma 7.2. *Assume A9–11. Let $\ell \in \{1,\ldots,N\}$ and $T < +\infty$. For any bounded measurable function h on $\mathbb{X}^{r+1}$ we have,*

(i) for all u,t such that $r \le t \le u \le T$,

$$\left|\mathbb{E}\left[h\left(\xi^{N,J^\ell_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u,T}\right] - \mathbb{E}\left[h\left(\xi^{N,J^\ell_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u+1,T}\right]\right| \le \rho^{u-t}\mathrm{osc}(h) \ ,$$

where ρ is defined in A9(i).

(ii) for all u,t such that $t-r \le u \le t-1 \le T$,

$$\left|\mathbb{E}\left[h\left(\xi^{N,J^\ell_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u,T}\right] - \mathbb{E}\left[h\left(\xi^{N,J^\ell_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u+1,T}\right]\right| \le \mathrm{osc}(h) \ .$$

Proof. According to Section 7.2.2, for all $\ell \in \{1,\ldots,N\}$, $\{J^{N,\ell}_u\}_{u=0}^T$ is an inhomogeneous Markov chain evolving backward in time with backward kernel $\{\Lambda^N_u\}_{u=0}^{T-1}$. For any $r \le t \le u \le T$, we have

$$\begin{aligned} &\mathbb{E}\left[h\left(\xi^{N,J^{N,\ell}_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u,T}\right] - \mathbb{E}\left[h\left(\xi^{N,J^{N,\ell}_{t-r:t}}_{t-r:t}\right)\middle|\mathcal{G}^N_{u+1,T}\right] \\ &\qquad = \sum_{j_{t:u}}\left[\delta_{J^{N,\ell}_u}(j_u) - \left(\Lambda_u(J^{N,\ell}_{u+1}, j_u)\mathbf{1}_{u<T} + \frac{\omega^{N,j_u}_T}{\Omega_u}\mathbf{1}_{u=T}\right)\right] \\ &\qquad\qquad \times \prod_{\ell=u}^{t+1}\Lambda^N_{\ell-1}(j_\ell, j_{\ell-1}) \sum_{j_{t-r:t-1}} \prod_{\ell=t}^{t-r+1}\Lambda^N_{\ell-1}(j_\ell, j_{\ell-1}) h\left(\xi^{N,j_{t-r:t}}_{t-r:t}\right) \ . \end{aligned}$$

The RHS of this equation is the difference between two expectations started with two different initial distributions. Under A9(i), the backward kernel satisfies the uniform Doeblin condition,

$$\forall (i,j) \in \{1,\ldots,N\}^2 \quad \Lambda^N_s(i,j) \ge \frac{\sigma_-}{\sigma_+}\frac{\omega^i_s}{\Omega^N_s} \ ,$$

and the proof is completed by the exponential forgetting of the backward kernel (see [Cappé *et al.*, 2005, Del Moral et Guionnet, 2001]). The proof of (ii) follows exactly the same lines. □

To compute an upper-bound for the L_q-mean error of the FFBSi algorithm, we may define the difference between the FFBS and the FFBSi estimators:

$$\delta_T^N\left[S_{T,r}\right] = \widetilde{\phi}_{0:T|T}^N\left[S_{T,r}\right] - \phi_{0:T|T}^N\left[S_{T,r}\right] \ . \tag{7.25}$$

Proof of Theorem 7.1 for the FFBSi estimator. The difference between the FFBS and the FFBSi estimators, δ_T^N, defined in (7.25), can be written

$$\begin{aligned}
\delta_T^N\left[S_{T,r}\right] &= \frac{1}{N}\sum_{\ell=1}^{N}\sum_{t=r}^{T} h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right) - \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,1}}\right)\middle|\mathcal{F}_T^N\right] \\
&= \frac{1}{N}\sum_{\ell=1}^{N}\sum_{t=r}^{T}\sum_{u=t-r}^{T} \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right)\middle|\mathcal{G}_{u,T}^N\right] - \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right)\middle|\mathcal{G}_{u+1,T}^N\right] \\
&= \frac{1}{N}\sum_{\ell=1}^{N}\sum_{u=0}^{T}\zeta_u^{N,\ell} \ ,
\end{aligned}$$

where

$$\zeta_u^{N,\ell} \stackrel{\text{def}}{=} \sum_{t=r}^{(u+r)\wedge T} \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right)\middle|\mathcal{G}_{u,T}^N\right] - \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right)\middle|\mathcal{G}_{u+1,T}^N\right] \ .$$

For all $\ell \in \{1,\ldots,N\}$ and all $u \in \{0,\ldots,T\}$, the random variable $\zeta_u^{N,\ell}$ is $\mathcal{G}_{u,T}^N$-measurable and $\mathbb{E}\left[\zeta_u^{N,\ell}\middle|\mathcal{G}_{u+1,T}^N\right] = 0$ so that $\zeta_u^{N,\ell}$ can be seen as the increment of a backward martingale. Hence, since $q \geq 2$, using the Burkholder inequality (see [Hall et Heyde, 1980, Theorem 2.10, page 23]), there exists a constant C (depending only on q, σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$) such that:

$$\left\|\delta_T^N\left[S_{T,r}\right]\right\|_q \leq C\left\{\sum_{u=0}^{T}\mathbb{E}\left[\left|N^{-1}\sum_{\ell=1}^{N}\zeta_u^{N,\ell}\right|^q\right]^{2/q}\right\}^{1/2} \ . \tag{7.26}$$

Then, since the random variables $\{\zeta_u^{N,\ell}\}_{\ell=1}^N$ are conditionally independent and centered conditionally to $\mathcal{G}_{u+1,T}^N$, using the Burkholder inequality once again implies:

$$\mathbb{E}\left[\left|\sum_{\ell=1}^{N}\zeta_u^{N,\ell}\right|^q\middle|\mathcal{G}_{u+1,T}^N\right] \leq CN^{q/2-1}\sum_{\ell=1}^{N}\mathbb{E}\left[\left|\zeta_u^{N,\ell}\right|^q\middle|\mathcal{G}_{u+1,T}^N\right] \ . \tag{7.27}$$

Furthermore, according to Lemma 7.2(i),

$$\left|\zeta_u^{N,\ell}\right| \leq \sum_{t=r}^{u} \left| \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right) \middle| \mathcal{G}_{u,T}^N \right] - \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right) \middle| \mathcal{G}_{u+1,T}^N \right] \right|$$
$$+ \sum_{t=u+1}^{(u+r)\wedge T} \left| \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right) \middle| \mathcal{G}_{u,T}^N \right] - \mathbb{E}\left[h_t\left(\xi_{t-r:t}^{N,J_{t-r:t}^{N,\ell}}\right) \middle| \mathcal{G}_{u+1,T}^N \right] \right|$$
$$\leq \sum_{t=r}^{u} \rho^{u-t}\mathrm{osc}(h_t) + \sum_{t=u+1}^{(u+r)\wedge T} \mathrm{osc}(h_t)\ . \quad (7.28)$$

Putting (7.26), (7.27) and (7.28) together leads to

$$\left\|\delta_T^N\left[S_{T,r}\right]\right\|_q \leq \frac{C}{\sqrt{N}} \left\{ \sum_{u=0}^{T} \left(\sum_{t=r}^{(u+r)\wedge T} \rho^{(u-t)\vee 0}\mathrm{osc}(h_t) \right)^2 \right\}^{1/2} .$$

Using the Holder inequality as in the proof of Proposition 7.1 yields

$$\left\|\delta_T^N\left[S_{T,r}\right]\right\|_q \leq \frac{C}{\sqrt{N}} \sqrt{1+r}\left(\sqrt{1+r} \wedge \sqrt{T-r+1}\right) \left(\sum_{s=r}^{T} \mathrm{osc}(h_s)^2 \right)^{1/2} ,$$

and the proof of Theorem 7.1 for the FFBSi estimator is derived from the triangle inequality:

$$\left\| \phi_{0:T|T}\left(S_{T,r}\right) - \widetilde{\phi}_{0:T|T}^N\left(S_{T,r}\right) \right\|_q \leq \left\|\Delta_T^N[S_{T,r}]\right\|_q + \left\|\delta_T^N\left[S_{T,r}\right]\right\|_q ,$$

where $\Delta_T^N[S_{T,r}]$ is defined by (7.9) and $\delta_T^N\left[S_{T,r}\right]$ is defined by (7.25). □

7.6 Proof of Theorem 7.2

We preface the proof of the Theorem by showing that the martingale term of the error $\Delta_T^N[S_{T,r}]$ (which is defined by (7.9)) satisfies an exponential deviation inequality in the following Proposition.

Proposition 7.3. *Assume A9–11. There exists a constant C (depending only on σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$) such that for any $T<\infty$, any $N\geq 1$, any $\varepsilon>0$, any integer r and any bounded and measurable functions $\{h_s\}_{s=r}^T$ on $\mathbb{X}^{r+1}$,*

$$\mathbb{P}\left\{ \left| \sum_{t=0}^{T} D_{t,T}^N(S_{T,r}) \right| > \varepsilon \right\} \leq 2\exp\left(-\frac{CN\varepsilon^2}{\Theta_{r,T}\sum_{s=r}^{T}\mathrm{osc}(h_s)^2} \right) , \quad (7.29)$$

where $D_{t,T}^N$ is defined in (7.14) *and $\Theta_{r,T}$ is defined by* (7.17).

Proof. According to the definition of $D^N_{t,T}(S_{T,r})$ given in (7.14), we can write

$$\sum_{t=0}^{T} D^N_{t,T}(S_{T,r}) = \sum_{k=1}^{N(T+1)} \upsilon^N_k \ ,$$

where for all $t \in \{0,\dots,T\}$ and $\ell \in \{1,\dots,N\}$, $\upsilon^N_{Nt+\ell}$ is defined by

$$\upsilon^N_{Nt+\ell} = \frac{\phi^N_{t-1}[\vartheta_t]}{\phi^N_{t-1}\left[\frac{\mathcal{L}_{t-1,T}\mathbf{1}}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}\right]} N^{-1}\omega^{N,\ell}_t \frac{G^N_{t,T} S_{T,r}(\xi^{N,\ell}_t)}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty} \ ,$$

and is bounded by (see (7.18))

$$\left|\upsilon^N_{Nt+\ell}\right| \le CN^{-1}\sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s) \ .$$

Furthermore, we define the filtration $\left\{\mathcal{H}^N_k\right\}_{k=1}^{N(T+1)}$, for all $t \in \{0,\dots,T\}$ and $\ell \in \{1,\dots,N\}$, by:

$$\mathcal{H}^N_{Nt+\ell} \stackrel{\text{def}}{=} \mathcal{F}^N_{t-1} \vee \sigma\left\{\left(\omega^{N,i}_t, \xi^{N,i}_t\right), 1 \le i \le \ell\right\} \ ,$$

with the convention $\mathcal{F}^N_{-1} = \sigma(Y_{0:T})$. Then, according to Lemma 7.1, $\{\upsilon_k\}_{k=1}^{N(T+1)}$ is a martingale increment for $\{\mathcal{H}^N_k\}_{k=1}^{N(T+1)}$ and the Azuma-Hoeffding inequality completes the proof. □

Proposition 7.4. *Assume A9–11. There exists a constant C (depending only on σ_-, σ_+, c_-, $\sup_{t\ge1}|\vartheta_t|_\infty$ and $\sup_{t\ge0}|\omega_t|_\infty$) such that for any $T < \infty$, any $N \ge 1$, any $\varepsilon > 0$, any integer r and any bounded and measurable functions $\{h_s\}_{s=r}^T$ on $\mathbb{X}^{r+1}$,*

$$\mathbb{P}\left\{\left|\sum_{t=0}^{T} C^N_{t,T}(S_{T,r})\right| > \varepsilon\right\} \le 8\exp\left(-\frac{CN\varepsilon}{(1+r)\sum_{s=r}^T \mathrm{osc}(h_s)}\right) \ . \tag{7.30}$$

where $C^N_{t,T}(F)$ is defined in (7.15).

Proof. In order to apply Lemma 7.4 in the appendix, we first need to find an exponential deviation inequality for $C^N_{t,T}(S_{T,r})$ which is done by using the decomposition $C^N_{t,T}(S_{T,r}) = U^N_{t,T}V^N_{t,T}W^N_{t,T}$ given in (7.23). First, the ratio $U^N_{t,T}$ is dealt with through Lemma 7.3 in the appendix by defining

$$\begin{cases} a_N \stackrel{\text{def}}{=} N^{-1}\sum_{\ell=1}^N \omega^{N,\ell}_t G^N_{t,T} S_{T,r}(\xi^{N,\ell}_t)/|\mathcal{L}_{t,T}\mathbf{1}|_\infty \ , \\ b_N \stackrel{\text{def}}{=} N^{-1}\sum_{\ell=1}^N \omega^{N,\ell}_t \ , \\ b \stackrel{\text{def}}{=} \mathbb{E}[\omega^1_t|\mathcal{F}^N_{t-1}] = \phi^N_{t-1}\left[Mg_t\right]/\phi^N_{t-1}[\vartheta_t] \ , \\ \beta \stackrel{\text{def}}{=} c_-/|\vartheta_t|_\infty \ . \end{cases}$$

Assumption A9(ii) and A11 shows that $b \geq \beta$ and (7.18) shows that $|a_N/b_N| \leq C(1+r)\max_{0\leq t\leq T}\{\mathrm{osc}(h_t)\}$. Therefore, Condition (I) of Lemma 7.3 is satisfied. The bounds $0 < \omega_t^l \leq |\omega_t|_\infty$ and the Hoeffding inequality lead to

$$\mathbb{P}[|b_N - b| \geq \varepsilon] = \mathbb{E}\left[\mathbb{P}\left[\left|N^{-1}\sum_{\ell=1}^{N}\left(\omega_t^{N,\ell} - \mathbb{E}[\omega_t^{N,1}|\mathcal{F}_{t-1}^N]\right)\right| \geq \varepsilon \middle| \mathcal{F}_{t-1}^N\right]\right] \leq 2\exp\left(-\frac{2N\varepsilon^2}{|\omega_t|_\infty^2}\right) ,$$

establishing Condition (ii) in Lemma 7.3. Finally, Lemma 7.1(i) and the Hoeffding inequality imply that

$$\mathbb{P}\left[|a_N| \geq \varepsilon\right] = \mathbb{E}\left[\mathbb{P}\left[\left|N^{-1}\sum_{\ell=1}^{N}\omega_t^{N,\ell}G_{t,T}^N S_{T,r}(\xi_t^{N,\ell})/|\mathcal{L}_{t,T}\mathbf{1}|_\infty\right| \geq \varepsilon \middle| \mathcal{F}_{t-1}^N\right]\right] \leq 2\exp\left(-\frac{N\varepsilon^2}{2|\omega_t|_\infty^2\left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)\right)^2}\right) .$$

Lemma 7.3 therefore yields

$$\mathbb{P}\left\{\left|U_{t,T}^N\right| \geq \varepsilon\right\} \leq 2\exp\left(-\frac{CN\varepsilon^2}{\left(\sum_{s=r}^{T}\rho^{\max(t-s,s-r-t,0)}\mathrm{osc}(h_s)\right)^2}\right) .$$

Then $V_{t,T}^N$ is dealt with by using again the Hoeffding inequality and the bounds $0 < b_{t,T}^{N,\ell} \leq |\omega_t|_\infty$, where $b_{t,T}^{N,\ell} \stackrel{\mathrm{def}}{=} \omega_t^{N,\ell}\frac{\mathcal{L}_{t,T}\mathbf{1}(\xi_t^{N,\ell})}{|\mathcal{L}_{t,T}\mathbf{1}|_\infty}$:

$$\mathbb{P}\left[\left|N^{-1}\sum_{\ell=1}^{N}b_{t,T}^{N,\ell} - \mathbb{E}\left[b_{t,T}^{N,1}\middle|\mathcal{F}_{t-1}\right]\right| \geq \varepsilon\right]$$
$$= \mathbb{E}\left[\mathbb{P}\left[\left|N^{-1}\sum_{\ell=1}^{N}\left(b_{t,T}^{N,\ell} - \mathbb{E}\left[b_{t,T}^{N,\ell}\middle|\mathcal{F}_{t-1}^N\right]\right)\right| \geq \varepsilon \middle| \mathcal{F}_{t-1}^N\right]\right] \leq 2\exp\left(-CN\varepsilon^2\right) .$$

Finally, $W_{t,T}^N$ has been shown in (7.24) to be bounded by a constant depending only on σ_-, σ_+, c_-, $\sup_{t\geq 1}|\vartheta_t|_\infty$ and $\sup_{t\geq 0}|\omega_t|_\infty$: $\left|W_{t,T}^N\right| \leq C$ so that

$$\mathbb{P}\left\{\left|C_{t,T}^N(S_{T,r})\right| > \varepsilon\right\} \leq \mathbb{P}\left\{\left|U_{t,T}^N V_{t,T}^N\right| > \varepsilon/C\right\} \leq \mathbb{P}\left\{\left|U_{t,T}^N\right| > \varepsilon_u\right\} + \mathbb{P}\left\{\left|V_{t,T}^N\right| > \varepsilon_v\right\} ,$$

where

$$\varepsilon_u \stackrel{\text{def}}{=} \sqrt{\varepsilon \sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s)/C} \ ,$$
$$\varepsilon_v \stackrel{\text{def}}{=} \sqrt{\frac{\varepsilon}{C \sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s)}} \ .$$

Therefore,

$$\mathbb{P}\left\{\left|C_{t,T}^N(S_{T,r})\right| > \varepsilon\right\} \leq 4 \exp\left(-\frac{CN\varepsilon}{\sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s)}\right) \ .$$

The proof of (7.30) is finally completed by applying Lemma 7.4 with

$$X_t = C_{t,T}^N(S_{T,r}) \ , \quad A = 4 \ , \quad B_t = \frac{CN}{\sum_{s=r}^{T} \rho^{\max(t-s,s-r-t,0)} \mathrm{osc}(h_s)} \ , \quad \gamma = 1/2 \ .$$

□

Proof of Theorem 7.2 for the FFBS estimator. The result is obtained by writing

$$\mathbb{P}\left\{\left|\Delta_T^N[S_{T,r}]\right| > \varepsilon\right\} \leq \mathbb{P}\left\{\left|\sum_{t=0}^{T} C_{t,T}^N(S_{T,r})\right| > \varepsilon/2\right\} + \mathbb{P}\left\{\left|\sum_{t=0}^{T} D_{t,T}^N(S_{T,r})\right| > \varepsilon/2\right\} \ ,$$

and using (7.29) and (7.30). □

Proof of Theorem 7.2 for the FFBSi estimator. We recall the decomposition used in the proof of Theorem 7.1 for the FFBSi estimator:

$$\delta_T^N\left[S_{T,r}\right] = \frac{1}{N} \sum_{\ell=1}^{N} \sum_{u=0}^{T} \zeta_u^{N,\ell} \ ,$$

where $\delta_T^N\left[S_{T,r}\right]$ is defined by (7.25). Since $\left\{\zeta_u^{N,\ell}\right\}_{\ell=1}^{N}$ are $\mathcal{G}_{u,T}^N$ measurable and centered conditionally to $\mathcal{G}_{u+1,T}^N$ using the same steps as in the proof of Proposition 7.3, we get

$$\mathbb{P}\left\{\left|\delta_T^N\left[S_{T,r}\right]\right| > \varepsilon\right\} \leq 2 \exp\left(-\frac{CN\varepsilon^2}{\Theta_{r,T} \sum_{s=r}^{T} \mathrm{osc}(h_s)^2}\right) \ ,$$

where $\Theta_{r,T}$ is defined by (7.17). The proof is finally completed by writing

$$\phi_{0:T|T}\left[S_{T,r}\right] - \widetilde{\phi}_{0:T|T}^N\left[S_{T,r}\right] = \Delta_T^N[S_{T,r}] + \delta_T^N\left[S_{T,r}\right] \ ,$$

and by using Theorem 7.2 for the FFBS estimator. □

7.7 Technical results

Lemma 7.3. *Assume that a_N, b_N, and b are random variables defined on the same probability space such that there exist positive constants β, B, C, and M satisfying*

(i) $|a_N/b_N| \leq M$, $\mathbb{P}$*-a.s. and* $b \geq \beta$, $\mathbb{P}$*-a.s.,*

(ii) For all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left[|b_N - b| > \epsilon\right] \leq B\mathrm{e}^{-CN\epsilon^2}$,

(iii) For all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left[|a_N| > \epsilon\right] \leq B\mathrm{e}^{-CN(\epsilon/M)^2}$.

Then,

$$\mathbb{P}\left\{\left|\frac{a_N}{b_N}\right| > \epsilon\right\} \leq B\exp\left(-CN\left(\frac{\epsilon\beta}{2M}\right)^2\right).$$

Proof. See [Douc *et al.*, 2011a, Lemma 4]. □

Lemma 7.4. *For $T \geq 0$, let $\{X_t\}_{t=0}^T$ be $(T+1)$ random variables. Assume that there exists a constants $A \geq 1$ and for all $0 \leq t \leq T$, there exists a constant $B_t > 0$ such that and all $\varepsilon > 0$*

$$\mathbb{P}\{|X_t| > \varepsilon\} \leq Ae^{-B_t\varepsilon}.$$

Then, for all $0 < \gamma < 1$ and all $\varepsilon > 0$, we have

$$\mathbb{P}\left\{\left|\sum_{t=0}^T X_t\right| > \varepsilon\right\} \leq \frac{A}{1-\gamma}e^{-\gamma B\varepsilon/(T+1)},$$

where

$$B \stackrel{\text{def}}{=} \left(\frac{1}{T+1}\sum_{t=0}^T B_t^{-1}\right)^{-1}.$$

Proof. By the Bienayme-Tchebychev inequality, we have

$$\begin{aligned}\mathbb{P}\left\{\left|\sum_{t=0}^T X_t\right| > \varepsilon\right\} &= \mathbb{P}\left\{\exp\left[\frac{\gamma B}{T+1}\left|\sum_{t=0}^T X_t\right|\right] > e^{\gamma B\varepsilon/(T+1)}\right\} \\ &\leq e^{-\gamma B\varepsilon/(T+1)}\mathbb{E}\left[\exp\left[\frac{\gamma B}{T+1}\left|\sum_{t=0}^T X_t\right|\right]\right]. \quad (7.31)\end{aligned}$$

It remains to bound the expectation in the RHS of (7.31) by $A(1-\gamma)^{-1}$. First, by the Minkowski inequality,

$$\begin{aligned}\mathbb{E}\left[\exp\left[\frac{\gamma B}{T+1}\left|\sum_{t=0}^T X_t\right|\right]\right] &= \sum_{q=0}^\infty \frac{\gamma^q B^q}{q!(T+1)^q}\mathbb{E}\left[\left|\sum_{t=0}^T X_t\right|^q\right] \\ &\leq 1 + \sum_{q=1}^\infty \frac{\gamma^q B^q}{q!(T+1)^q}\left(\sum_{t=0}^T \|X_t\|_q\right)^q.\end{aligned}$$

Moreover, for $q \geq 1$, $\mathbb{E}\left[|X_t|^q\right]$ can be bounded by

$$\mathbb{E}\left[|X_t|^q\right] = \int_0^\infty \mathbb{P}\{|X_t| > \varepsilon^{1/q}\}\mathrm{d}\varepsilon \leq A \int_0^\infty e^{-B_t\varepsilon^{1/q}}\mathrm{d}\varepsilon = \frac{Aq!}{B_t^q} \ ,$$

Finally,

$$\mathbb{E}\left[\exp\left[\frac{\gamma B}{T+1}\left|\sum_{t=0}^{T} X_t\right|\right]\right] \leq A\sum_{q=0}^{\infty}\gamma^q = \frac{A}{(1-\gamma)} \ .$$

□

Chapitre 8

Estimation non paramétrique dans les modèles de Markov cachés (article)

This paper outlines a new procedure to perform nonparametric estimation in hidden Markov models. It is assumed that a Markov chain $\{X_k\}_{k\geq 0}$ is observed only through a process $\{Y_k\}_{k\geq 0}$, where Y_k is a noisy observation of $f_\star(X_k)$, where $f_\star$ is an unknown function. We prove the identifiability of the model when the state-space of the Markov chain is compact and has a transition kernel which is known up to a scaling parameter, and under smoothness assumptions on the function $f_\star$. We propose a penalized maximum pairwise likelihood estimator of the function $f_\star$ which we prove to be consistent as the number of observations grows to infinity. Finally, we provide numerical experiments to support our claims.

8.1 Introduction

A bivariate stochastic process $\{(X_k, Y_k)\}_{k\geq 0}$ is said to be a hidden Markov model (HMM) if the state sequence $\{X_k\}_{k\geq 0}$ is a Markov chain, if the observations $\{Y_k\}_{k\geq 0}$ are independent conditionally on $\{X_k\}_{k\geq 0}$ and if the conditional distribution of Y_k given the state sequence depends only on X_k. These models can be applied in a large variety of disciplines such as financial econometrics ([Mamon et Elliott, 2007]), biology ([Churchill, 1992]) or speech recognition ([Juang et Rabiner, 1991]).

In this paper, the state-space of the Markov chain $\{X_k\}_{k\geq 0}$ is assumed to be a convex and compact subset of $\mathbb{R}^m$. This Markov chain is a random

walk with increment distribution known up to a scaling factor $a_\star$. The observations are given, for any $k \geq 0$, by $Y_k = f_\star(X_k) + \epsilon_k$, where $f_\star$ is a function on K taking values in $\mathbb{R}^\ell$ and the measurement noise $\{\epsilon_k\}_{k\geq 0}$ is an i.i.d. sequence of i.i.d. Gaussian on $\mathbb{R}^\ell$ with known covariance matrix. The aim of this paper is to estimate the function $f_\star$ using only the observations $\{Y_k\}_{k\geq 0}$.

In classical regression models such as *errors-in-variables* models, the variables $\{X_k\}_{k\geq 0}$ are observed through a sequence $\{Z_k\}_{k\geq 0}$ given by $Z_k \stackrel{\text{def}}{=} X_k + \eta_k$, where the random variables $\{\eta_k\}_{k\geq 0}$ are i.i.d. with known distribution. Many solutions have been proposed to solve this regression problem using an estimation of the probability density of X_0 (this is the *deconvolution* problem), see [Carroll et Hall, 1988], [Carroll et Stefanski, 1990] and [Fan et Truong, 1993] for an estimation based on kernel density estimators; see also [Comte et Taupin, 2007] for an estimation based on the minimization of a penalized contrast. Nevertheless, all these works rely on the assumption that the process $\{X_k\}_{k\geq 0}$ is directly observed, which is not the case in our model.

When $\{X_k\}_{k\geq 0}$ is a Markov chain, [Lacour, 2008b] proposed an estimation of the density of the invariant probability and of the Markov kernel of $\{X_k\}_{k\geq 0}$ when the chain is observed. The estimation procedure amounts to minimizing a penalized contrast in order to minimize the empirical L_2-norm of the error. [Lacour, 2008a] provided an extension of this work in the HMM framework when the observations are given by

$$Y_k = X_k + \epsilon_k \ ,$$

where the random variables $\{\epsilon_k\}_{k\geq 0}$ are i.i.d. with known distribution. These works provide estimation procedures on the Markov chain $\{X_k\}_{k\geq 0}$ but there does not exist any result on the nonparametric estimation problem studied in this paper.

This problem is motivated by an application to localization using radio measurements (see [Dumont et Le Corff, 2012b]). In this case, at each time step k, a mobile device observes the power of signals transmitted by ℓ antennas; this measurement is denoted by Y_k. The localization of the device is denoted by X_k and is assumed to be a Markov chain on a subset of $\mathbb{R}^2$. The problem consists in estimating the localizations $\{X_k\}_{k\geq 0}$ only observing the signal powers $\{Y_k\}_{k\geq 0}$. In this application, $f_\star$ represents the average propagation model, which means that the variable Y_k follows the normal distribution on $\mathbb{R}^\ell$, $\mathcal{N}(f_\star(X_k), \sigma^2 I_\ell)$. An accurate estimation of the positions $\{X_k\}_{k\geq 0}$, using particle filtering for instance, relies on a good estimation of $f_\star$.

The main result of this paper is the identifiability of the model. We assume that the Markov chain $\{X_k\}_{k\geq 0}$ is stationary with known (up to a scaling factor $a_\star$) transition kernel, and that $f_\star$ is a diffeomorphism on

its image (which necessarily implies that $m \leq \ell$). We assume in addition that $f_\star$ is smooth in the sense that it belongs to some Sobolev space $W^{s,p}$ (see (8.5)). Provided that f is continuously differentiable and is such that $(f(X_0), f(X_1))$ and $(f_\star(X_0), f_\star(X_1))$ have the same distribution we show that there exists an isometric transformation ϕ on the state-space K such that $f = f_\star \circ \phi$. A key step is to show that $(f_\star)^{-1} \circ f$ is necessarily a diffeomorphism on its image, which is done using algebraic topology and measure theoretic arguments.

Our estimator $\widehat{f}_n$ is defined as a maximizer of a penalized pairwise likelihood on the Sobolev space $W^{s,p}$. The parameters s and p of the Sobolev space are assumed to satisfy $s > m/p+1$ and K is assumed to be compact to allow the use of classical Sobolev embeddings into the space of continuously differentiable functions on K. This estimator of $f_\star$ is associated to an estimator $\widehat{p}_n$ of the marginal distribution of a pair of observations (see (8.11)). We prove that the Hellinger distance between $\widehat{p}_n$ and the true distribution of a pair of observations under $(f_\star, a_\star)$ vanishes as the number of observations grows to infinity. More precisely, we prove that the rate of convergence of $\widehat{p}_n$, in Hellinger distance, can be chosen as close as possible to $n^{-1/2}$. The consistency of $(\widehat{f}_n, \widehat{a}_n)$ follows as a consequence together with the identifiability result and continuity properties. To analyze the asymptotic properties of our estimators, we need, as it is now well understood, deviation inequalities for the empirical process of the observations. To that purpose, we use the concentration inequality for additive functionals of Markov chains proved in [Adamczak et Bednorz, 2012] and the maximal inequality for dependent processes of [Doukhan *et al.*, 1995] to have a control on the supremum of a function-indexed empirical process.

Our results are supported by numerical experiments: in the case where the scaling parameter $a_\star$ is known and $m = 1$, we provide an Expectation-Maximization based algorithm to compute $\widehat{f}_n$, see [Dempster *et al.*, 1977]. We show that the estimation procedure can be solved using a differential equation. We provide several simulations that show the efficiency of our method.

In Section 8.2 the model, the estimators and the assumptions are presented. The main results are displayed in Section 8.3: the identifiability of the model in Section 8.3.1 and the consistency of the estimator along with a rate of convergence in Section 8.3.2. The algorithm and numerical experiments are given in Section 8.4. Section 8.5 gathers important proofs on the identifiability and consistency needed to state the main results. Additional technical results are provided in Section 8.6.

8.2 Model and definitions

Let ℓ and m be positive integers and K be a subset of $\mathbb{R}^m$. The main statistical problem considered in this paper is the estimation of an unknown target function $f_\star : K \to \mathbb{R}^\ell$ when observing a process $\{Y_k\}_{k\in\mathbb{N}}$ such that for any $k \geq 0$, Y_k belongs to $\mathbb{R}^\ell$ and satisfies

$$Y_k \stackrel{\text{def}}{=} f_\star(X_k) + \epsilon_k \ .$$

$\{\epsilon_k\}_{k\in\mathbb{N}}$ is assumed to be an i.i.d Gaussian process with common distribution $\mathcal{N}(0, \sigma^2 I_\ell)$, I_ℓ being the identity matrix of size ℓ and σ^2 a fixed positive parameter. Denote by φ the probability distribution of ϵ_0, *i.e.*

$$\forall z \in \mathbb{R}^\ell,\ \varphi(z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-\ell/2} \exp\left\{-\frac{\|z\|^2}{2\sigma^2}\right\} \ ,$$

where $\|\cdot\|$ is the euclidean norm on $\mathbb{R}^m$ (we use the same notation for the euclidian norm on $\mathbb{R}^\ell$). $\{X_k\}_{k\in\mathbb{N}}$ is assumed to be a non observed Markov chain, taking its values in K and independent of $\{\epsilon_k\}_{k\in\mathbb{N}}$. In the sequel, all the density functions are with respect to the Lebesgue measure on K, denoted by μ. For any $a \in \mathbb{R}_+^\star$, denote by q_a the transition density on K defined, for all $x, x' \in K$, by

$$q_a(x, x') \stackrel{\text{def}}{=} C_a(x) q\left(\frac{\|x' - x\|}{a}\right) \ , \tag{8.1}$$

where q is a known, continuous and strictly monotone function on $\mathbb{R}_+$ and where

$$C_a(x) \stackrel{\text{def}}{=} \left(\int_K q\left(\frac{\|x' - x\|}{a}\right) \mathrm{d}x'\right)^{-1} \ , \tag{8.2}$$

where $\mathrm{d}x'$ is a shorthand notation for $\mu(\mathrm{d}x')$. In our numerical application in Section 8.4.2, the Gaussian kernel $q(x) = \exp(-x^2/2)$ is chosen. The Markov transition kernel associated with q_a is denoted by Q_a. Assume the existence of an unknown parameter $a_\star > 0$ such that

H1 $\{X_k\}_{k\in\mathbb{Z}}$ is a stationary Markov chain with transition kernel $Q_{a_\star}$.

It follows from H1 that $\{Y_k\}_{k\in\mathbb{N}}$ is stationary. Assume the following statement on the set K:

H2 (i) K is compact.

(ii) K is convex.

(iii) K satisfies the *cone condition*.

The cone condition states the existence of a finite cone Δ in $\mathbb{R}^m$ such that for any x in K there exists an isometric transformation Δ_x of Δ in $\mathbb{R}^m$, with vertex x and included in K (see [Adams et Fournier, 2003]). Any compact set with $\mathcal{C}^2$-regular boundary satisfies the cone condition. Note that this condition may also hold for compact sets with non-regular boundaries such as cubes in $\mathbb{R}^m$ for instance.

As an immediate consequence of the compactness of K and of the continuity of q, there exists $0 < \sigma_-(a) < \sigma_+(a) < +\infty$ such that, for all $x, x' \in K$,

$$\sigma_-(a) \leq q_a(x, x') \leq \sigma_+(a) \ . \tag{8.3}$$

For any $a > 0$, Q_a is a ψ-irreducible and recurrent Markov kernel and then, it has a unique invariant probability distribution (see [Meyn et Tweedie, 1993, Theorem 10.0.1]). By the symmetry of the kernel $(x, x') \to q\left(\frac{\|x-x'\|}{a}\right)$, the finite measure on K with density function $x \mapsto C_a^{-1}(x)$ is Q_a-invariant. Therefore, the unique invariant probability of Q_a has a density given by

$$\forall x \in K,\ \nu_a(x) \stackrel{\text{def}}{=} \frac{\int_K q\left(\frac{\|x'-x\|}{a}\right) \mathrm{d}x'}{\int_{K^2} q\left(\frac{\|x'-x''\|}{a}\right) \mathrm{d}x' \mathrm{d}x''} \ . \tag{8.4}$$

For any $f : K \longrightarrow \mathbb{R}^\ell$ and any $j \in \{1, \cdots, \ell\}$, the j^{th} component of f is denoted by f_j. Set

$$\mathrm{L}^p \stackrel{\text{def}}{=} \left\{ f : K \to \mathbb{R}^\ell \ ; \ \|f\|_{\mathrm{L}^p}^p = \int_K \|f(x)\|^p \mathrm{d}x < \infty \right\} \ .$$

For any m-tuple $\alpha \stackrel{\text{def}}{=} \{\alpha_i\}_{i=1}^m$ of non-negative integers, we write $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i$. Denote by $W^{s,p}$ the Sobolev space on K with parameter $s \in \mathbb{N}$ and $p \geq 1$, *i.e.*,

$$W^{s,p} \stackrel{\text{def}}{=} \{ f \in \mathrm{L}^p;\ D^\alpha f \in \mathrm{L}^p, \alpha \in \mathbb{N}^m \text{ and } |\alpha| \leq s \} \ , \tag{8.5}$$

where $D^\alpha f : K \to \mathbb{R}^\ell$ is the vector of partial derivatives of order α of the components f_j, for $j \in \{1, \cdots, \ell\}$. $W^{s,p}$ is equipped with the norm $\|\cdot\|_{W^{s,p}}$ defined, for any $f \in W^{s,p}$, by

$$\|f\|_{W^{s,p}} \stackrel{\text{def}}{=} \left(\sum_{0 \leq |\alpha| \leq s} \|D^\alpha f\|_{\mathrm{L}^p}^p \right)^{1/p} \ . \tag{8.6}$$

Let $s \in \mathbb{N}$ and $p \geq 1$. For any $f : K \mapsto \mathbb{R}^\ell$, denote by $\mathrm{Im}(f)$ the image in $\mathbb{R}^\ell$ of f, $\mathrm{Im}(f) \stackrel{\text{def}}{=} f(K)$.

H3 (i) $f_\star \in W^{s,p}$.

(ii) $f_\star : K \to \mathrm{Im}(f_\star)$ is a diffeomorphism.

For any $k \geq 0$, let $\mathcal{C}^k(K)$ be the vector space of all the functions $f : K \to \mathbb{R}$ which, together with all their partial derivatives $D^\alpha f$ of order $|\alpha| \leq k$, are continuous on K. Let $\|\cdot\|_{\mathcal{C}^k(K)}$ be the norm on $\mathcal{C}^k(K)$ defined by $\|f\|_{\mathcal{C}^k(K)} \stackrel{\text{def}}{=} \sup_{|\alpha| \leq k} \|D^\alpha f\|_\infty$.

Remark. Note that, for any $j \in \{1, \cdots, \ell\}$ and $f \in W^{s,p}$, f_j belongs to $W^{s,p}(K, \mathbb{R})$, the Sobolev space of real-valued functions with parameters s and p. Let $k \geq 0$, by [Adams et Fournier, 2003, Theorem 6.3], assuming that K satisfies H2(i) and H2(iii) and $s > m/p + k$, $W^{s,p}(K, \mathbb{R})$ is compactly embedded into $\left(\mathcal{C}^k(K), \|\cdot\|_{\mathcal{C}^k(K)}\right)$. Provided that $s > m/p + 1$, and arguing component by component, $W^{s,p}$ is compactly embedded into $\mathcal{C}^1(K, \mathbb{R}^\ell)$, shortly denoted by $\mathcal{C}^1$ and equipped with the norm $\|\cdot\|_{\mathcal{C}^1}$ given, for any $f \in \mathcal{C}^1$, by $\|f\|_{\mathcal{C}^1} = \sum_{j=1}^{\ell} \|f_j\|_{\mathcal{C}^1(K)}$. Moreover, the identity function $id : W^{s,p} \to \mathcal{C}^1$ being linear and continuous, there exists a positive coefficient denoted by κ such that, for any $f \in W^{s,p}$,

$$\|f\|_{\mathcal{C}^1} \leq \kappa \|f\|_{W^{s,p}} \ . \tag{8.7}$$

H4 $s > m/p + 1$.

For any $f : K \to \mathbb{R}^\ell$ and any $x \in \mathbb{R}^m$ where f is differentiable, the Jacobian of f at x, is defined by

$$J_f^2(x) \stackrel{\text{def}}{=} \text{Det}\left[D_f(x)^T D_f(x)\right] \ ,$$

where $D_f(x)$ is the $\ell \times m$ gradient matrix of f at x defined, for any $j \in \{1, \ldots, \ell\}$ and any $i \in \{1, \ldots, m\}$, by

$$D_f(x)_{j,i} \stackrel{\text{def}}{=} \frac{\partial f_j}{\partial x_i}(x) \ .$$

Remark. i) Assuming H3(i) and H4, $f_\star \in \mathcal{C}^1$ and thus $J_{f_\star}$ exists. Moreover, H3(ii) holds whenever $f_\star$ is a one to one function and $J_{f_\star}(x) > 0$ for all x in K.

ii) By H3(ii), for any x in K, the linear application $D_{f_\star}(x)$ is injective and thus, $m \leq \ell$.

We now give the definition of the estimators $\left(\widehat{f}_n, \widehat{a}_n\right)$ of $(f_\star, a_\star)$ given $2n$ observations $\{Y_k\}_{k=0}^{2n-1}$. For practical reasons, we assume that $a_\star \in [a_-, +\infty[$, for a known $a_- > 0$. For all integer $n \geq 1$, define $\left(\widehat{f}_n, \widehat{a}_n\right)$ by

$$\left(\widehat{f}_n, \widehat{a}_n\right) \stackrel{\text{def}}{=} \text{argmax}_{f \in W^{s,p}, a \geq a_-} \left\{ \frac{1}{n} \sum_{k=0}^{n-1} \ln p_{f,a}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 I^2(f) \right\} \ , \tag{8.8}$$

where, for all y_0, y_1 in $\mathbb{R}^\ell$,

$$p_{f,a}(y_0, y_1) \stackrel{\text{def}}{=} \int \varphi(y_0 - f(x_0)) \varphi(y_1 - f(x_1)) \nu_a(x_0) q_a(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 \tag{8.9}$$

and, for some positive υ,

$$I^2(f) \stackrel{\text{def}}{=} ||f||_{W^{s,p}}^{\upsilon+1} \ . \tag{8.10}$$

Remark. By the dominated convergence theorem, the function

$$(f,a) \mapsto \frac{1}{n}\sum_{k=0}^{n-1} \ln p_{f,a}(Y_{2k}, Y_{2k+1})$$

is continuous on $\mathcal{C}^1 \times [a_-, \infty[$, thus, by (8.8) and (8.10) $\widehat{f}_n$ exists and belongs to $\mathcal{C}^1$.

Consider the following assumption on υ.

H5 $\upsilon > 2\ell$.

Note that $(f,a) \to \frac{1}{n}\sum_{k=0}^{n-1} \ln p_{f,a}(Y_{2k}, Y_{2k+1})$ does not represent the likelihood of the observations $\{Y_k\}_{k=0}^{2n-1}$ but what we call the pairwise pseudo-likelihood of the observations.

By (8.8), $\widehat{a}_n$ could be equal to ∞ so that we shall extend our definitions to this case. By the dominated convergence theorem, for any $x_0, x_1 \in K$, any $y_0, y_1 \in \mathbb{R}^\ell$ and any measurable function f, $q_a(x_0,x_1)$, $\nu_a(x_0)$ and $p_{f,a}(y_0,y_1)$ converge as $a \to \infty$ to $q_\infty(x_0,x_1)$, $\nu_\infty(x_0)$ and $p_{f,\infty}(y_0,y_1)$, defined by:

$$\nu_\infty(x_0) \stackrel{\text{def}}{=} \mu(K)^{-1}\ ,\ q_\infty(x_0,x_1) \stackrel{\text{def}}{=} \mu(K)^{-1}\ ,$$
$$p_{f,\infty}(y_0,y_1) \stackrel{\text{def}}{=} \mu(K)^{-2} \int \varphi(y_0 - f(x_0))\mathrm{d}x_0 \int \varphi(y_1 - f(x_1))\mathrm{d}x_1\ .$$

Let $\widehat{p}_n$ denote the maximum penalized likelihood estimator (MLE) of the density on $\mathbb{R}^{2\ell}$ of (Y_0, Y_1), defined by

$$\widehat{p}_n \stackrel{\text{def}}{=} p_{\widehat{f}_n, \widehat{a}_n}\ . \tag{8.11}$$

The convergence properties of this estimator will be analyzed with the Hellinger metric, defined, for any probability densities p_1 and p_2 on $\mathbb{R}^{2\ell}$, by

$$h(p_1,p_2) \stackrel{\text{def}}{=} \left[\frac{1}{2}\int \left(p_1^{1/2}(y) - p_2^{1/2}(y)\right)^2 \mathrm{d}y\right]^{1/2}\ . \tag{8.12}$$

Remark. The reason we use the Sobolev framework instead of directly considering the space $\mathcal{C}^1$ is, first of all, computational. Indeed, as we will see in Section 8.4, the Sobolev norm chosen in penalty (8.10) can be easily manipulated compared with the $\mathcal{C}^1$ norm. Moreover, Theorem 8.2 ensures that $\|\widehat{f}_n\|_{W^{s,p}}$ stays bounded and thus, that $\{\widehat{f}_n\}_{n\geq 1}$ lies in a compact subset of $\mathcal{C}^1$. This plays a key role in the proof of Theorem 8.3.

Section 8.3 provides the main results of the paper. Theorem 8.1 establishes the identifiability of our model. Then, the Hellinger consistency of the MLE (8.11) is shown in Theorem 8.2. This result does not imply, *a priori*, the consistency of the estimators $\left(\widehat{f}_n, \widehat{a}_n\right)$ defined by (8.8). However, by Theorem 8.1, whenever the MLE is consistent, so is $\left(\widehat{f}_n, \widehat{a}_n\right)$ up to an

isometric transformation on the state space K. The consistency of $\left(\widehat{f}_n, \widehat{a}_n\right)$ is given by Theorem 8.3.

8.3 Main results

8.3.1 Identifiability

We denote by $\mathcal{I}$ the set of all the isometries of K. For any functions f and h defined on K we write $f \overset{\mathcal{I}}{\sim} h$ and say that f and h are in the same equivalence class modulo the isometric transformations of K, if and only if there exists an isometry $\tilde{\phi}$ on K such that $f = h \circ \tilde{\phi}$. In the sequel, for any random variables X and Y, we write $X \overset{\mathcal{D}}{=} Y$ if X and Y have the same distribution. The proof of Theorem 8.1 relies on intermediate lemmas. The proofs of these lemmas are postponed to Section 8.5.1.

Theorem 8.1. *Assume H1-3. Let $f : K \to \mathbb{R}^\ell$ be $\mathcal{C}^1$ and $0 < b \leq \infty$. Assume also that $h(p_{f,b}, p_{f_\star,a_\star}) = 0$ where $p_{f,b}$ and $p_{f_\star,a_\star}$ are defined by* (8.9)*. Then, $b = a_\star$ and $f \overset{\mathcal{I}}{\sim} f_\star$.*

Proof. The proof of the intermediate lemmas are postponed to Section 8.5.1. Let $0 < b \leq \infty$ and $f \in \mathcal{C}^1$ such that $h(p_{f,b}, p_{f_\star,a_\star}) = 0$. Let $\{X'_k\}_{k\geq 0}$ be a Markov chain with initial distribution ν_b and transition kernel Q_b. Consider also $\{\epsilon'_k\}_{k\geq 0}$ a sequence of independent $\mathcal{N}(0, \sigma^2 I_\ell)$ random variables, independent from $\{X'_k\}_{k\geq 0}$. It is also assumed that $\{X'_k, \epsilon'_k\}_{k\geq 0}$ are independent from $\{X_k, \epsilon_k\}_{k\geq 0}$. Define, for any $k \geq 0$, $Y'_k = f(X'_k) + \epsilon'_k$. If $h(p_{f,b}, p_{f_\star,a_\star}) = 0$, then, for any $k \geq 0$, $(Y_k, Y_{k+1}) \overset{\mathcal{D}}{=} (Y'_k, Y'_{k+1})$. The density φ being known, this yields

$$(f(X'_k), f(X'_{k+1})) \overset{\mathcal{D}}{=} (f_\star(X_k), f_\star(X_{k+1})) \,. \tag{8.13}$$

(8.13) and the irreducibility of the Markov chains $\{X_k\}_{k\geq 0}$ and $\{X'_k\}_{k\geq 0}$ imply that $\mathrm{Im}(f) = \mathrm{Im}(f_\star)$. By H3, $f_\star$ is a diffeomorphism. Let $(f_\star)^{-1}$ denotes its inverse function and define

$$\phi \overset{\text{def}}{=} (f_\star)^{-1} \circ f \,. \tag{8.14}$$

Since $f_\star$ is a diffeomorphism and f is $\mathcal{C}^1$, ϕ is $\mathcal{C}^1$. The first step consists in proving that (8.13) implies that ϕ is a diffeomorphism. To do so, we show in Lemma 8.1 that $J_\phi > 0$ and thus that ϕ is a local diffeomorphism.

Lemma 8.1. *Assume H2*(i) *and H3. For all $x \in K$, $J_\phi(x) > 0$, where ϕ is defined by* (8.14)*.*

Then, we show that ϕ is necessarily a covering map of K (see definition below) and that, under H2(ii), any covering map of K is a one to one function. These results are established in Lemma 8.2 and Lemma 8.3.

$\phi : K \to K$ is said to be a covering map if and only if (see [Lee, 2000, Chapter 11])

(i) ϕ is continuous.

(ii) ϕ is surjective.

(iii) For every $y \in K$, there exists an open neighbourhood V of y and a family $(O_i)_{i\in I}$ of disjoint open subsets of K such that $\phi^{-1}(V) = \bigcup_{i\in I} O_i$, with O_i mapped homeomorphically onto V by ϕ, for all $i \in I$.

Lemma 8.2. *Assume H2*(i) *and H3. Then, the function ϕ defined by* (8.14) *is a covering map.*

Lemma 8.3. *Assume H2*(ii). *Then, every covering map $\phi : K \to K$ is a one to one function.*

By Lemma 8.2 and Lemma 8.3, ϕ, defined by (8.14) is a one to one function, moreover Lemma 8.1 below states that $J_\phi > 0$ on K and thus achieves the proof that ϕ is a diffeomorphism.

Denote by ϕ^{-1} the inverse function of ϕ. By (8.13), for all $x \in K$ and all positive measurable function h on K,

$$\begin{aligned} Q_a(x,h) = \mathbb{E}\left[h(X_k)|X_{k-1}=x\right] &= \mathbb{E}\left[h\circ\phi(X'_k)|X'_{k-1}=\phi^{-1}(x)\right] \ , \\ &= Q_b(\phi^{-1}(x), h\circ\phi) \ . \end{aligned}$$

Moreover,

$$\begin{aligned} Q_b(\phi^{-1}(x), h\circ\phi) &= \int_K h\circ\phi(u) Q_b(\phi^{-1}(x),u)\mathrm{d}u \\ &= \int_K h(u) Q_b(\phi^{-1}(x),\phi^{-1}(u))|J_{\phi^{-1}}(u)|\mathrm{d}u \ . \end{aligned}$$

Then, for all $(x,x') \in K^2$,

$$Q_{a_\star}(x,x') = Q_b(\phi^{-1}(x),\phi^{-1}(x'))|J_{\phi^{-1}}(x')| \ .$$

This equation directly leads to $b < \infty$. Indeed, if $b = \infty$, the left side of the equation depends on x (since $a_\star < \infty$) whereas the right side does not. We can now suppose $0 < b < \infty$. By (8.1),

$$C_{a_\star}(x) q\left(\frac{\|x'-x\|}{a_\star}\right) = C_b(\phi^{-1}(x)) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x)\|}{b}\right)|J_{\phi^{-1}}(x')| \ . \tag{8.15}$$

Therefore, for all $x \in K$, applying (8.15) with $x' = x$ yields

$$|J_{\phi^{-1}}(x)| = \frac{C_{a_\star}(x)}{C_b(\phi^{-1}(x))} \ . \tag{8.16}$$

By H2(i-ii), Schauder's theorem (see [Smart, 1980]) states that there exists $x_0 \in K$ such that $\phi^{-1}(x_0) = x_0$. By (8.15), there exists a constant C such that, for all $x \in K$

$$|J_{\phi^{-1}}(x)| = C \frac{q\left(\frac{\|x-x_0\|}{a_\star}\right)}{q\left(\frac{\|\phi^{-1}(x)-\phi^{-1}(x_0)\|}{b}\right)} .$$

Plugging this expression and (8.16) in (8.15) yields

$$\begin{aligned} &q\left(\frac{\|x-x_0\|}{a_\star}\right) q\left(\frac{\|x'-x\|}{a_\star}\right) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x_0)\|}{b}\right) \\ &\quad = q\left(\frac{\|\phi^{-1}(x)-\phi^{-1}(x_0)\|}{b}\right) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x)\|}{b}\right) q\left(\frac{\|x'-x_0\|}{a_\star}\right) . \end{aligned} \tag{8.17}$$

Applied with $x' = x_0$, we have, for all $x \in K$,

$$q\left(\frac{\|x-x_0\|}{a_\star}\right) = q\left(\frac{\|\phi^{-1}(x)-\phi^{-1}(x_0)\|}{b}\right)$$

and then, since q is a one to one function by assumption,

$$\frac{\|x-x_0\|}{a_\star} = \frac{\|\phi^{-1}(x)-x_0\|}{b} .$$

Considering the supremum of the last inequality for $x \in K$ yields $b = a_\star$. Then, (8.17) gives, for all $x, x' \in K$

$$\|\phi^{-1}(x') - \phi^{-1}(x)\| = \|x' - x\| .$$

Therefore, ϕ is an isometry and $f = f_\star \circ \phi$ which concludes the proof. □

8.3.2 Convergence results

Theorem 8.2 states the Hellinger consistency of the MLE $\widehat{p}_n$ and ensures that the Sobolev norm of the estimator $\widehat{f}_n$ is bounded. Theorem 8.1 and Theorem 8.2 lead to the second main result, Theorem 8.3, which guarantees that $(\widehat{f}_n, \widehat{a}_n)$ is also consistent. The proof of Theorem 8.2 uses the same classical proof scheme as in the independent case, see [Van De Geer, 2000, Section 10.2] for an illustration of such a proof. This proof relies on the control of the empirical process, it requires both a result on the concentration of the empirical process and a maximal inequality. Unfortunately, the tools used in the independent case such as the Bernstein or the Hoeffding inequalities do not hold in our model and similar results in the dependent case have

to be used, see [Adamczak et Bednorz, 2012]. Denote by $\mathbb{P}$ the distribution of $\{Y_k\}_{k\geq 0}$ under the true parameters $(f_\star, a_\star)$. For any sequence of random variables $\{Z_n\}_{n\geq 0}$ and any sequence of positive numbers $\{\alpha_n\}_{n\geq 0}$, we write $Z_n = O_{\mathbb{P}}(\alpha_n)$ if

$$\lim_{T\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\left\{|Z_n| > T\alpha_n\right\} = 0 \; .$$

Theorem 8.2. *Assume H1-2, H3*(i) *and H4-5. Let* $\left(\widehat{f}_n, \widehat{a}_n\right)$ *be defined by* (8.8) *and* $I(f)$ *by* (8.10)*. Then, provided that*

$$\lambda_n \underset{n\to+\infty}{\longrightarrow} 0 \text{ and } \lambda_n^2 n^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty \; , \tag{8.18}$$

we have

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) = O_{\mathbb{P}}(\lambda_n^2) \quad \text{and} \quad I^2(\widehat{f}_n) = O_{\mathbb{P}}(1) \; . \tag{8.19}$$

Proof. The proof relies on a *basic inequality* which controls the Hellinger risk $h^2(\widehat{p}_n, p_{f_\star,a_\star})$ and the complexity of the estimator $I^2(\widehat{f}_n)$ by the empirical process, see [Van De Geer, 2000]. The control of this empirical process will be done in Proposition 8.1. We set, for any density function p on $\mathbb{R}^{2\ell}$,

$$g_p \stackrel{\text{def}}{=} \frac{1}{2} \ln \frac{p + p_{f_\star,a_\star}}{2p_{f_\star,a_\star}} \; . \tag{8.20}$$

Let $\mathbb{P}_n$ be the empirical distribution based on the observations $\{Y_{2k}, Y_{2k+1}\}_{k=0}^{n-1}$, *i.e.*, for any measurable set A of $\mathbb{R}^{2\ell}$,

$$\mathbb{P}_n(A) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_A(Y_{2k}, Y_{2k+1}) \; .$$

By (8.8) and (8.11), the basic inequality of [Van De Geer, 2000, Lemma 10.5], states that:

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) + 4\lambda_n^2 I^2(\widehat{f}_n) \leq 16 \int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}^\star) + 4\lambda_n^2 I^2(f_\star) \; . \tag{8.21}$$

Therefore, a control of the term $\int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P})$ in the right hand side of (8.21) will provide simultaneously a bound on the growth of $h^2(\widehat{p}_n, p_{f_\star,a_\star})$ and $I^2(\widehat{f}_n)$. The empirical process indexed by $W^{s,p}$ is defined, for any $f \in W^{s,p}$ and any $a \geq a_-$, by

$$\nu_n(g_{p_{f,a}}) \stackrel{\text{def}}{=} \sqrt{n} \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}) \; ,$$

where $g_{p_{f,a}}$ is defined by (8.20). Proposition 8.1 provides a deviation inequality for the supremum of the normalized empirical process.

Proposition 8.1. *Assume H1-2, H3*(i) *and H4-H5. There exist some positive constants K, Σ and T such that, for any $x > 0$,*

$$\mathbb{P}\left\{\sup_{f\in W^{s,p},\ a\geq a_-} \frac{\nu_n(g_{p_{f,a}})}{I^2(f)\vee 1} \geq T + x\right\} \leq Ke^{-\Sigma x} . \tag{8.22}$$

Proposition 8.1 is proved in Section 8.5.2 below. It ensures that

$$\sup_{f\in W^{s,p},\ a\geq a_-} \frac{\int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P})}{I^2(f)\vee 1} = O_{\mathbb{P}}(n^{-1/2}) .$$

Plugging this bound into (8.21) gives

$$(4 + O_{\mathbb{P}}(n^{-1/2}\lambda_n^{-2}))I^2(\widehat{f}_n) \leq 4I^2(f_\star) + O_{\mathbb{P}}(n^{-1/2}\lambda_n^{-2}) .$$

By (8.18), this establishes the second statement of (8.19). Combining this result with (8.21) gives:

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq O_{\mathbb{P}}(n^{-1/2}) + O_{\mathbb{P}}(\lambda_n^2)$$

which proves the first statement of (8.19) and concludes the proof of Theorem 8.2. □

Equations (8.18) and (8.19) give a rate of convergence of $h^2(\widehat{p}_n, p_{f_\star,a_\star})$. This rate of convergence is slower than $n^{-1/2}$ but can be chosen as close as wanted to $n^{-1/2}$, *e.g.* we can choose $\lambda_n^2 = n^{-1/2}\ln n$.

On the other hand, $I^2(\widehat{f}_n) = O_{\mathbb{P}}(1)$ and the Sobolev embedding ensures that $\widehat{f}_n$ belongs to some compact subset of $\mathcal{C}^1$ with probability converging to 1 as n tends to ∞. Let $d_{\mathcal{C}^1}$ denotes the distance function on $\mathcal{C}^1$ associated with the norm $\|\cdot\|_{\mathcal{C}^1}$. Let also $\mathcal{F}_\star$ be the set of all the functions f in the same equivalence class as $f^\star$ modulo the isometric transformations of K, *i.e.*

$$\mathcal{F}_\star \stackrel{\text{def}}{=} \{f;\ f \stackrel{\mathcal{I}}{\sim} f_\star\} . \tag{8.23}$$

Theorem 8.3. *Assume H1-5. Let $\left(\widehat{f}_n, \widehat{a}_n\right)$ be defined by* (8.8) *and $I(f)$ by* (8.10)*. Then, provided that*

$$\lambda_n \underset{n\to+\infty}{\longrightarrow} 0 \text{ and } \lambda_n^2 n^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty ,$$

we have,

$$d_{\mathcal{C}^1}(\hat{f}_n, \mathcal{F}_\star) \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{and} \quad \widehat{a}_n \underset{n\to+\infty}{\longrightarrow} a_\star \text{ in } \mathbb{P}-\text{probability} , \tag{8.24}$$

where $\mathcal{F}_\star$ is defined by (8.23)*.*

Proof. We prove (8.24) introducing the Alexandroff compactification $[a_-, \infty]$ of $[a_-, \infty[$ and a distance function on this set such that $[a_-, \infty]$ is compact and metric. Moreover, for any $I > 0$, the set $\mathcal{B}_{W^{s,p}}(0, I)$, defined as the closure in $\mathcal{C}_1$ of $\{f \in W^{s,p};\ I(f) \leq I\}$, is a compact subset of $\mathcal{C}^1$. Thus, $\mathcal{B}_{W^{s,p}}(0, I) \times [a_-, \infty]$ is a compact subset of $\mathcal{C}^1 \times [a_-, \infty]$ and (8.24) will result from Theorem 8.2 and continuity arguments on the function $(f, a) \mapsto h^2(p_{f,a}, p_{f_\star,a_\star})$.

By Theorem 8.2, for any $\gamma > 0$, there exist $\epsilon > 0$ and $I > 0$ such that:

$$\limsup_{n\to+\infty} \mathbb{P}\left\{h^2(\widehat{p}_n, p_{f_\star,a_\star}) > \epsilon\lambda_n^2\right\} \leq \frac{\gamma}{2}\,, \tag{8.25}$$

$$\limsup_{n\to+\infty} \mathbb{P}\left\{I(\widehat{f}_n) > I\right\} \leq \frac{\gamma}{2}\,. \tag{8.26}$$

Denote by d the distance on $\mathcal{C}^1 \times [a_-, \infty]$ defined, for all $((f, a), (f', a')) \in \left(\mathcal{C}^1 \times [a_-, \infty]\right)^2$ by

$$d\left((f, a), \left(f', a'\right)\right) = d_{\mathcal{C}_1}(f, f') + |\arctan(a) - \arctan(a')|\,,$$

with $\arctan(\infty) = \frac{\pi}{2}$. The distance on $[a_-, \infty]$ defined for any a and a' in $[a_-, \infty]$ by $|\arctan(a) - \arctan(a')|$ ensures its compactness. Therefore $E \stackrel{\text{def}}{=} \mathcal{B}_{W^{s,p}}(0, I) \times [a_-, \infty]$ is a compact subset of $\left(\mathcal{C}^1 \times [a_-, \infty], d\right)$. We also set

$$d\left((f, a), (\mathcal{F}_\star, a_\star)\right) = \inf_{f' \in \mathcal{F}_\star} d\left((f, a), \left(f', a_\star\right)\right)\,.$$

For any $\eta > 0$, denote by E_η the following set

$$E_\eta \stackrel{\text{def}}{=} E \setminus \bigcup_{f' \in \mathcal{F}_\star} \left\{(f, a) \in \mathcal{C}^1 \times [a_-, \infty];\ d\left((f, a), \left(f', a_\star\right)\right) < \eta\right\}\,,$$

E_η is a non-empty and closed subset of E which is compact in $\mathcal{C}_1 \times [a_-, \infty]$, thus E_η is also a compact subset of $\mathcal{C}_1 \times [a_-, \infty]$. By the dominated convergence theorem, the function defined on E, by

$$(f, a) \mapsto h^2(p_{f,a}, p_{f_\star,a_\star})$$

is continuous relatively to the topology defined by the distance d on $\mathcal{C}^1 \times [a_-, \infty]$. The compactness of E_η implies that $h^2(p_{f,a}, p_{f_\star,a_\star})$ reaches its minimum on E_η. Let ϵ_η be this minimum. By Theorem 8.1 and since, for any f in $\mathcal{B}_{W^{s,p}}(0, I)$, $h^2\left(p_{f,\infty}, p_{f_\star,a_\star}\right) > 0$, $\epsilon_\eta > 0$. Moreover,

$$\begin{aligned}\mathbb{P}\{d((\widehat{f}_n, \widehat{a}_n), (\mathcal{F}_\star, a_\star)) > \eta\} &\leq \mathbb{P}\{I(\widehat{f}_n) > I\} + \mathbb{P}\left\{h^2(\widehat{p}_n, p_{f_\star,a_\star}) > \epsilon\lambda_n^2\right\} \\ &\quad + \mathbb{P}\left\{I(\widehat{f}_n) \leq I,\, h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq \epsilon\lambda_n^2, d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\}\,.\end{aligned}$$

However, if $I(\widehat{f}_n) \leq I$ and $d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta$, then $\widehat{f}_n$ belongs to E_η and $h^2(\widehat{p}_n, p_{f_\star,a_\star}) \geq \epsilon_\eta$. Choosing n big enough such that $\epsilon\lambda_n^2 < \epsilon_\eta$,

$$\mathbb{P}\left\{I(\widehat{f}_n) \leq I,\, h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq \epsilon\lambda_n^2,\, d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\} = 0\ .$$

and, by (8.25) and (8.26),

$$\limsup_{n\to+\infty} \mathbb{P}\left\{d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\} \leq \gamma\ .$$

Since γ can be chosen arbitrarily small, for any $\eta > 0$,

$$\limsup_{n\to+\infty} \mathbb{P}\left\{d_{\mathcal{C}_1}(\widehat{f}_n, \mathcal{F}_\star) + |\arctan(\widehat{a}_n) - \arctan(a_\star)| > \eta\right\} = 0\ ,$$

and $\lim_{n\to\infty} d_{\mathcal{C}_1}(\widehat{f}_n, \mathcal{F}_\star) = 0$ in probability. Moreover, the function tan being continuous on $[0, \frac{\pi}{2}[$ and since $\arctan(a_\star) \neq \frac{\pi}{2}$, $\lim_{n\to\infty} |\widehat{a}_n - a_\star| = 0$ in probability.

□

8.4 Numerical experiments

In this section, we suppose the parameter $a_\star$ to be known and illustrate the performance of the estimator $\widehat{f}_n$ defined by (8.8). For practical considerations, we choose $\upsilon = 1$ in (8.10) and $p = 2$. The theoretical results provided in Section 8.3 rely on the assumption that $\upsilon > 2\ell$. However, choosing $\upsilon = 1$ allows to define an algorithm easy to implement with good convergence behavior. Using $\upsilon > 1$ would imply more involved numerical procedures to obtain parameter estimates. Let n be a positive integer, in this section, we denote by $\widehat{f}$ the estimator defined by (8.8) that maximizes the function T defined by

$$\begin{array}{rcccl} T & : & W^{s,2} & \to & \mathbb{R} \\ & & f & \mapsto & \frac{1}{n}\sum_{k=0}^{n-1} \ln p_{f,a_\star}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 ||f||^2_{W^{s,2}}\ . \end{array}$$

The HMM framework suggests to use an Expectation-Maximization (EM) type procedure, see [Dempster *et al.*, 1977]. This algorithm iteratively produces a sequence of estimates $\{\widehat{f}^p\}_{p\geq 0}$. Assume the current parameter estimate is given by $\widehat{f}^p$. The estimate $\widehat{\widehat{f}}^{p+1}$ is defined as one of the maximizer of the function Q defined by

$$f \mapsto Q(f, \widehat{f}^p) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p}\left[\ln p_{f,a_\star}\left(X_{2k}, Y_{2k}, X_{2k+1}, Y_{2k}\right) \middle| Y_{2k},\ Y_{2k+1}\right] - \lambda_n^2 ||f||^2_{W^{s,2}}\ ,$$

where $\mathbb{E}_{\widehat{f}^p}[\cdot]$ denotes the expectation under the law of the stationary HMM parameterized by $\widehat{f}^p$ and where

$$p_{f,a_\star}\left(x,y,x',y'\right) = \nu_{a_\star}(x) q_{a_\star}(x,x') \varphi(y-f(x)) \varphi(y-f(x')) \; .$$

The differential of $f \mapsto Q(f, \widehat{f}^p)$ is given, for any $f, h \in W^{s,2}$, by

$$\mathrm{d}_f Q(\cdot, \widehat{f}^p)(h) = S_{n,1}(\widehat{f}^p, f, h) + S_{n,2}(\widehat{f}^p, f, h) - 2\lambda_n^2 \sum_{0 \le |\alpha| \le s} \langle D^\alpha f, D^\alpha h \rangle_{\mathrm{L}_2} \; ,$$

where

$$S_{n,1}(\widehat{f}^p, f, h) \stackrel{\mathrm{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p}\left[\langle h(X_{2k}), f(X_{2k}) - Y_{2k} \rangle | Y_{2k:2k+1}\right] \; ,$$

$$S_{n,2}(\widehat{f}^p, f, h) \stackrel{\mathrm{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p}\left[\langle h(X_{2k+1}), f(X_{2k+1}) - Y_{2k+1} \rangle | Y_{2k:2k+1}\right] \; .$$

$\widehat{f}^{p+1}$ is then defined as the function $f \in W^{s,2}$ such that for any $h \in W^{s,2}$, $\mathrm{d}_f Q(\widehat{f}^p, \cdot)(h) = 0$. In the sequel, we choose $s = 2$ and $K = [0,1]$, therefore, this implies, for any $h \in W^{2,2}([0,1], \mathbb{R})$,

$$S_{n,1}(\widehat{f}^p, f, h) + S_{n,2}(\widehat{f}^p, f, h) - 2\lambda_n^2 \sum_{\alpha=0}^{2} \left\langle f^{(\alpha)}, h^{(\alpha)} \right\rangle_{\mathrm{L}_2} = 0 \; . \tag{8.27}$$

This equation can be applied to any function h in $W_0^{2,2} \stackrel{\mathrm{def}}{=} \{h \in W([0,1], \mathbb{R});\ h(0) = h(1) = 0\}$. Using integration by parts, this yields, for any component f_j and any $x \in [0,1]$,

$$\begin{aligned} &\left(1 + \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \right) f_j(x) \\ &- f_j^{(2)}(x) + f_j^{(4)}(x) = \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ Y_{2k} \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + Y_{2k+1} \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \; , \end{aligned} \tag{8.28}$$

where $\phi^{\widehat{f}^p,a_\star}_{2k|2k:2k+1}$ and $\phi^{\widehat{f}^p,a_\star}_{2k+1|2k:2k+1}$ are the filtering distributions defined by

$$\phi^{\widehat{f}^p,a_\star}_{2k|2k:2k+1}(x) \stackrel{\mathrm{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x,x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x'}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \; ,$$

$$\phi^{\widehat{f}^p,a_\star}_{2k+1|2k:2k+1}(x') \stackrel{\mathrm{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x,x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \; .$$

8.4.1 Numerical approximations

Let $N \geq 1$ be an integer. The differential system (8.28) is solved using a discretization of the state space $[0,1]$ by $\left\{\frac{i}{N}\right\}_{i=0}^{N}$. The filtering distributions $\phi_{2k|2k:2k+1}^{\widehat{f}^p,a_\star}$ and $\phi_{2k+1|2k:2k+1}^{\widehat{f}^p,a_\star}$ are approximated by piecewise constant functions $\underline{\phi}_k^{\widehat{f}^p,a_\star}$ and $\overline{\phi}_k^{\widehat{f}^p,a_\star}$, defined by

$$\underline{\phi}_k^{\widehat{f}^p,a_\star}(x) \stackrel{\text{def}}{=} \sum_{i=0}^{N-1} \mathbf{1}_{\left[\frac{i}{N},\frac{i+1}{N}\right[}(x)\, \underline{\varphi}_{i,k}^{\widehat{f}^p} \quad \text{and} \quad \overline{\phi}_k^{\widehat{f}^p,a_\star}(x) \stackrel{\text{def}}{=} \sum_{i=0}^{N-1} \mathbf{1}_{\left[\frac{i}{N},\frac{i+1}{N}\right[}(x)\, \overline{\varphi}_{i,k}^{\widehat{f}^t} \,,$$

where, for any $i \in \{0,\dots,N-1\}$, $\underline{\varphi}_{i,k}^{\widehat{f}^p}$ (resp. $\overline{\varphi}_{i,k}^{\widehat{f}^p}$) is the approximation of $\underline{\phi}_k^{\widehat{f}^p,a_\star}(\frac{i}{N})$ (resp. $\overline{\phi}_k^{\widehat{f}^p,a_\star}(\frac{i}{N})$) obtained with an Euler scheme. The equation (8.28) is solved on each interval $[\frac{i}{N},\frac{i+1}{N}[$, $i \in \{0,\cdots,N-1\}$, which is straightforward since the coefficients are constant and the equation is linear. For any $i \in \{0,\dots,N-1\}$ and any $j \in \{0,\dots,\ell\}$, the solution $f_{j,i}$ on the interval $[\frac{i}{N},\frac{i+1}{N}[$ belongs to some affine space of dimension 4. Thus, $4N$ parameters have to be chosen to uniquely determine the solution $\widehat{f}_j^{p+1} = \sum_{i=0}^{N-1} \mathbf{1}_{[\frac{i}{N},\frac{i+1}{N}[} f_{j,i}$. The $\mathcal{C}^3$-regularity conditions for each boundary provides $4(N-1)$ equations and solving (8.27) with $h(x)=1$, $h(x)=x$, $h(x)=x^2$ and $h(x)=x^3$ leads to four other linear equations which conclude the computation of $\widehat{f}_j^{p+1}$.

Algorithm 1 One iteration of the algorithm

Require: N, $\widehat{f}^p$, $a_\star$, $Y_{0:2n-1}$.
Ensure: $\widehat{f}^{p+1}$
  for $i \in \{0,\cdots,N\}$ **do**
    for $k \in \{0,\cdots,n-1\}$ **do**
      Compute $\underline{\varphi}_{i,k}^{\widehat{f}^p}$ and $\overline{\varphi}_{i,k}^{\widehat{f}^p}$.
    end for
  end for
  for $j \in \{1,\cdots,\ell\}$ **do**
    for $i \in \{0,\cdots,N-1\}$ **do**
      Compute $f_{j,i}$ by solving (8.28).
    end for
    Set $\widehat{f}_j^{p+1} = \sum_{i=0}^{N-1} \mathbf{1}_{[\frac{i}{N},\frac{i+1}{N}[} f_{j,i}$.
  end for

8.4.2 Experimental results

The Algorithm 1 is applied with the Gaussian kernel ($a_\star = 1$):

$$\forall x \in \mathbb{R}\ ,\ q(x) = \exp\left\{-\frac{1}{2}x^2\right\}\ .$$

The aim is first to estimate the function (in this case $\ell = 3$)

$$\begin{array}{rccc} f_\star & : & [0,1] & \to & \mathbb{R}^3 \\ & & x & \mapsto & (3x, 30(x-1/4)(x-1/2)(x-3/4), 2\cos(5x))\ , \end{array}$$

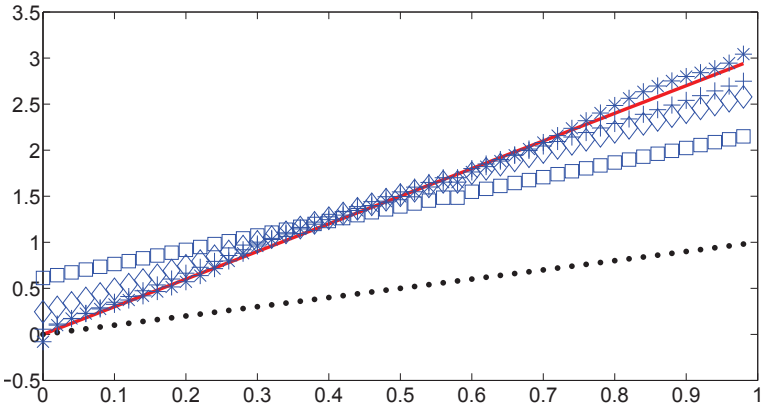
We use $\sigma^2 = 1$ and $N = 50$ to sample observations from the discretized model. The estimation is started with the estimate

$$\begin{array}{rccc} \widehat{f}^0 & : & [0,1] & \to & \mathbb{R}^3 \\ & & x & \mapsto & (x,0,0)\ . \end{array}$$

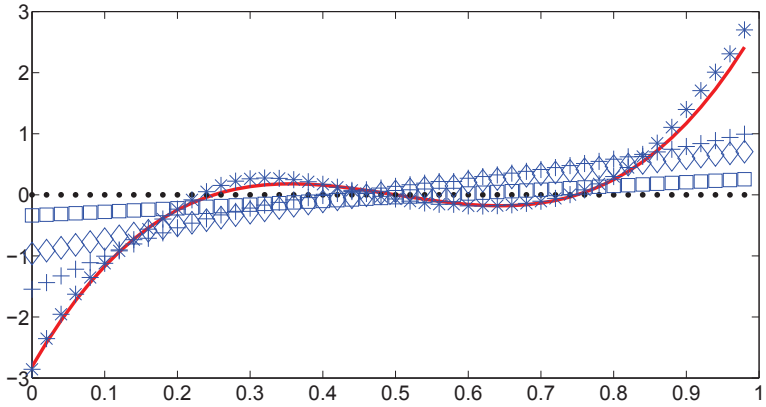
The Algorithm 1 is run with $\lambda_n^2 = \lfloor c\ln(n)/\sqrt{n}\rfloor$. Figure 8.1 displays the estimate after 1, 2, 3 and 25 iterations with $n = 50000$ observations along with the true functions for each coordinate. Figure 8.1 shows that after few iterations of the algorithm, the estimate can recover the curvature of the function $f_\star$, even with a flat initial estimate. Figure 8.2 gives the evolution of the error as a function of the number of observations. We consider the L_2-error defined, for $h_1, h_2 : [0,1] \to \mathbb{R}$, by

$$\|h_1 - h_2\|_2 \stackrel{\text{def}}{=} \left(\frac{1}{N}\sum_{i=1}^{N}\left|h_1\left(\frac{i}{N}\right) - h_2\left(\frac{i}{N}\right)\right|^2\right)^{1/2}\ .$$

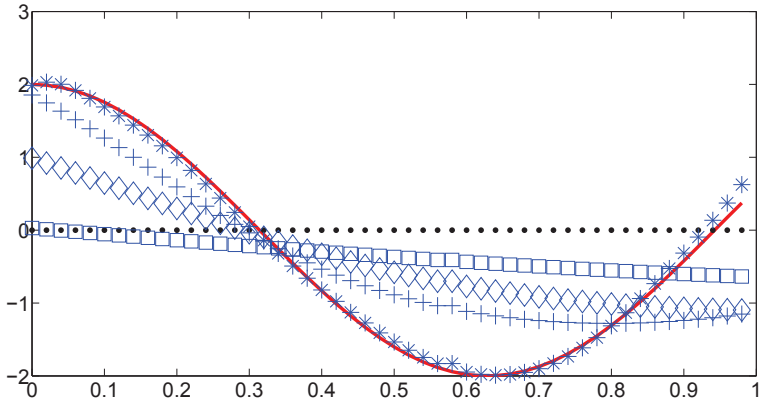
For each number of observations 50 independent Monte Carlo runs are used to compute the L_2-error after 25 iterations of the algorithm. Figure 8.2 shows the median and the lower and upper quartiles over the 50 independent Monte Carlo runs.



(a) f_1 and its estimates.

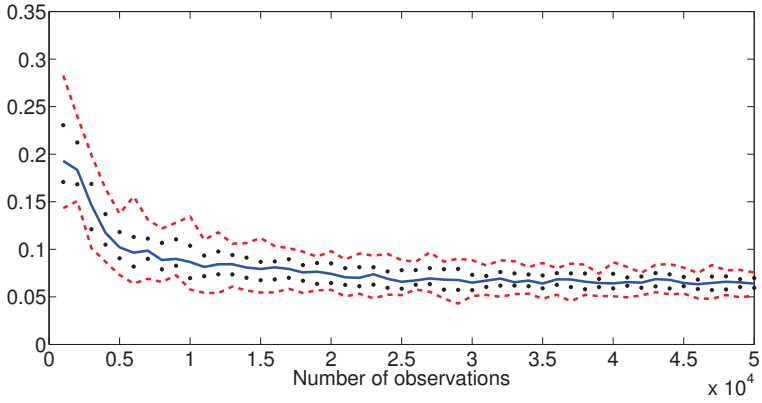


(b) f_2 and its estimates.

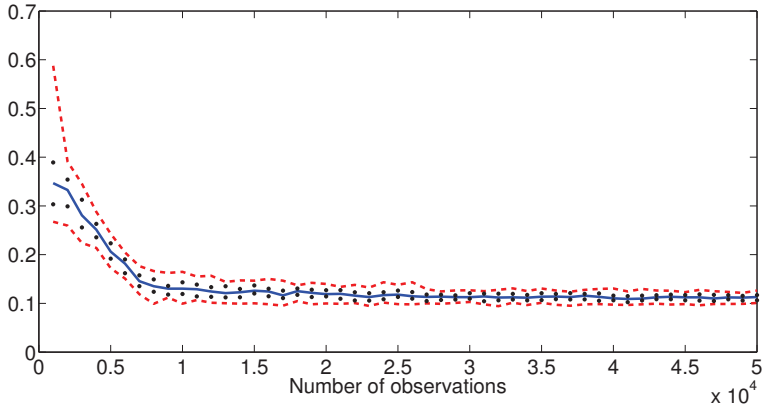


(c) f_3 and its estimates.

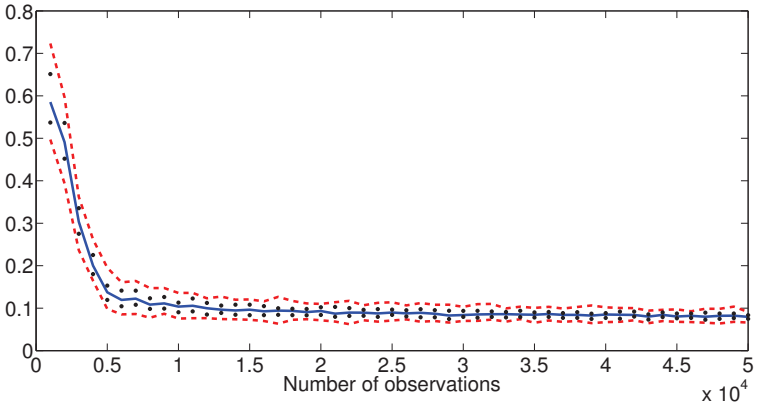
Figure 8.1: Estimation of f_1, f_2 and f_3 after 25 iterations of the algorithm. The true function (bold line) and the initial estimate (dots) are displayed along with the estimates after 1 (squares), 2 (diamonds), 3 (crosses) and 25 (stars) iterations.



(a) L_2-error for f_1



(b) L_2-error for f_2



(c) L_2-error for f_3

Figure 8.2: L_2 error for each coordinate. The median (bold line), .25 and .75 quantiles (dotted lines) and .05 and .95 quantiles (balls) over 50 independent Monte Carlo runs are represented.

8.5 Proofs

8.5.1 Identifiability

Proof of Lemma 8.1. By (8.14), (8.13) becomes

$$(\phi(X_0'), \phi(X_1')) \stackrel{\mathcal{D}}{=} (X_0, X_1) \ .$$

We now give an expression of the density of these two random vectors on $K \times K$. Let h be a bounded measurable function of $K \times K$. We have

$$\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] = \int h(\phi(x_0), \phi(x_1))\nu_b(x_0)q_b(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1 \ . \quad (8.29)$$

We introduce the set

$$A \stackrel{\text{def}}{=} \{z \in K;\ \forall x \in K \text{ s.t. } \phi(x) = z,\ J_\phi(x) > 0\} \ .$$

Let assume h is of the form

$$h(x_0, x_1) \stackrel{\text{def}}{=} h_2(x_0, x_1)\mathbf{1}_A(x_0)\mathbf{1}_A(x_1) \ , \quad (8.30)$$

where h_2 is any bounded measurable function. We have

$$\begin{aligned}\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] &= \int h_2(\phi(x_0), \phi(x_1))\nu_b(x_0)q_b(x_0, x_1) \\ &\qquad \times \mathbf{1}_A(\phi(x_0))\mathbf{1}_A(\phi(x_1))\mathrm{d}x_0\mathrm{d}x_1 \\ &= \int h_2(\phi(x_0), \phi(x_1))\frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)}\mathbf{1}_A(\phi(x_0))\mathbf{1}_A(\phi(x_1)) \\ &\qquad \times J_\phi(x_0)J_\phi(x_1)\mathrm{d}x_0\mathrm{d}x_1 \ .\end{aligned}$$

By [Evans et Gariepy, 1992, Theorem 2, p.99] and the area formula, for almost every $z \in K$, $\phi^{-1}(\{z\})$ is at most countable and we can apply the change of variable $z_0 = \phi(x_0)$, $z_1 = \phi(x_1)$.

$$\begin{aligned}\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] &= \int h_2(z_0, z_1)\mathbf{1}_A(z_0)\mathbf{1}_A(z_1) \\ &\qquad \times \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)}\mathrm{d}z_0\mathrm{d}z_1 \ .\end{aligned}$$

Moreover,

$$\mathbb{E}\left[h(X_0, X_1)\right] = \int h_2(z_0, z_1)\nu_{a_\star}(z_0)q_{a_\star}(z_0, z_1)\mathbf{1}_A(z_0)\mathbf{1}_A(z_1)\mathrm{d}z_0\mathrm{d}z_1 \ .$$

Therefore, for almost any $(z_0, z_1) \in K \times K$,

$$\nu_{a_\star}(z_0) q_{a_\star}(z_0, z_1) \mathbf{1}_A(z_0) \mathbf{1}_A(z_1) = \mathbf{1}_A(z_0) \mathbf{1}_A(z_1) \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0) q_b(x_0, x_1)}{J_\phi(x_0) J_\phi(x_1)} \, .$$

By Sard Theorem (see [Bröcker et Lander, 1975]), since ϕ is $\mathcal{C}^1$,

$$\mu\left(\{z \in K;\ \exists x \in K,\ \phi(x) = z \text{ and } J_\phi(x) = 0\}\right) = 0 \, ,$$

Therefore, the function $z \mapsto \mathbf{1}_A(z)$ equals 1 almost everywhere in K. Finally, for almost any $(z_0, z_1) \in K \times K$,

$$\nu_{a_\star}(z_0) q_{a_\star}(z_0, z_1) = \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0) q_b(x_0, x_1)}{J_\phi(x_0) J_\phi(x_1)} \, . \tag{8.31}$$

Let us assume that there exists $x_0 \in K$ such that $J_\phi(x_0) = 0$. There also exists $x \in K$ such that $J_\phi(x) > 0$ (otherwise $\mu(K) = \mu(\phi(J_\phi^{-1}(\{0\}))) = 0$ by Sard Theorem). By the mean value theorem, for all large enough $k \in \mathbb{N}^\star$, there exists $x_k \in K$ such that $J_\phi(x_k) = \frac{1}{k}$. By the inverse function theorem, there exists U_k neighbourhood of x_k in K such that $\phi|_{U_k}$ is a diffeomorphism. J_ϕ being continuous, there also exits a neighbourhood V_k such that, for all $x \in V_k$, $|J_\phi(x) - J_\phi(x_k)| \leq \frac{1}{2k}$. Therefore, for all x in V_k,

$$\frac{3}{2k} \geq J_\phi(x) \geq J_\phi(x_k) - \frac{1}{2k} = \frac{1}{2k} \, .$$

Let $W_k = U_k \cap V_k$, $\phi|_{W_k}$ is a diffeomorphism and $\mu(\phi(W_k)) > 0$. Therefore, there exists $(z_{k,0}, z_{k,1}) \in \phi(W_k) \times \phi(W_k)$ such that (8.31) is true. We denote by $x_{k,0}$ and $x_{k,1}$ the unique elements of W_k such that $z_{k,0} = \phi(x_{k,0})$ and $z_{k,1} = \phi(x_{k,1})$. Then,

$$\begin{aligned} \nu_{a_\star}(z_{k,0}) q_{a_\star}(z_{k,0}, z_{k,1}) &= \sum_{\substack{x_0 \in \phi^{-1}(\{z_{k,0}\}) \\ x_1 \in \phi^{-1}(\{z_{k,1}\})}} \frac{\nu_b(x_0) q_b(x_0, x_1)}{J_\phi(x_0) J_\phi(x_1)} \\ &\geq \frac{\nu_b(x_{k,0}) q_b(x_{k,0}, x_{k,1})}{J_\phi(x_{k,0}) J_\phi(x_{k,1})} \geq \frac{2k}{3} \nu_b(x_{k,0}) q_b(x_{k,0}, x_{k,1}) \, . \end{aligned}$$

By H2(i), $(x_0, x_1) \mapsto \nu_c(x_0) q_c(x_0, x_1)$ is bounded for any $0 < c \leq \infty$: there exists $0 < C_c^- < C_c^+$ such that, for any $(x_0, x_1) \in K^2$, $0 < C_c^- \leq \nu_c(x_0) q_c(x_0, x_1) \leq C_c^+$, we have, for any $k \geq 1$ large enough,

$$C_a^+ \geq \nu_{a_\star}(z_{k,0}) q_{a_\star}(z_{k,0}, z_{k,1}) \geq C_b^- \frac{2k}{3} \, ,$$

which is absurd and concludes the proof. □

Proof of Lemma 8.2. (i) comes from the continuity of $(f_\star)^{-1}$ and f and (ii) is true since $\mathrm{Im}(f_\star) = \mathrm{Im}(f)$. For (iii), let $z \in K$ and assume the set $\phi^{-1}(\{z\}) \stackrel{\text{def}}{=} \{x \in K; \phi(x) = z\}$ is infinite. By Lemma 8.1, ϕ is of full rank, then,by the inverse function theorem, for each $x \in \phi^{-1}(\{z\})$, there exists an open neighborhood V_x of x such that the function $\phi : V_x \to \phi(V_x)$ is a diffeomorphism and such that the $\{V_x\}_{x\in\phi^{-1}(\{z\})}$ are pairwise disjoint. By H2(i), there exists $n \in \mathbb{N}$ such that $\bigcup_{x\in\phi^{-1}(\{z\})} V_x$ can be covered with only n subsets of the form V_{x_i}, $x_i \in \phi^{-1}(\{z\})$, $i \in \{1, \ldots, n\}$. Therefore, for any $x \in \phi^{-1}(\{z\})$, there exists $i \in \{1, \ldots n\}$ such that $x \in V_{x_i}$. If $x \notin \{x_i\}_{i=1}^n$, then $x \in V_x \cap V_{x_i}$ which is absurd since V_x and V_{x_i} are disjoint. Therefore, $\phi^{-1}(\{z\}) \stackrel{\text{def}}{=} \{x \in K; \phi(x) = z\}$ is finite and denoted by $\{x_i\}_{i=1}^n$, $n \in \mathbb{N}^\star$. Let $\{V_i\}_{i=1}^n$ be disjoint open subsets of K such that $x_i \in V_i$ and define $V \stackrel{\text{def}}{=} \cap_{i=1}^n \phi(V_i)$, $O_i = \phi^{-1}_{|V_i}(V)$. Then, V is an open neighborhood of z which concludes the proof. □

Proof of Lemma 8.3. Assume there exist x_1 and x_2 in K such that $x_1 \neq x_2$ and $\phi(x_1) = \phi(x_2) = y$. By H2(ii), K is path-connected and there exists a continuous path $\gamma : [0, 1] \to K$ such that $\gamma(0) = x_1$ and $\gamma(1) = x_2$. Then $\phi\circ\gamma$ is a continuous path taking values in K such that $\phi\circ\gamma(0) = \phi\circ\gamma(1) = y$. If $\widetilde{\gamma}$ denotes the path defined by, for all $t \in [0, 1]$, $\widetilde{\gamma}(t) = y$, then $\phi \circ \gamma$ and $\widetilde{\gamma}$ are two paths in K with the same initial and terminal values. By H2(ii), K is simply connected and $\phi \circ \gamma$ and $\widetilde{\gamma}$ are path homotopic (see [Lee, 2000, p.151]). The function $u : [0, 1] \to K$ such that, for all $t \in [0, 1]$, $u(t) = x_1$ is a lift (see [Lee, 2000, p.237]) of $\widetilde{\gamma}$ for the covering map ϕ. Moreover, γ is a lift of $\phi \circ \gamma$ for the covering map ϕ. By the homotopy lifting property (see [Lee, 2000, Proposition 11.11, p.238]), since $u(0) = \gamma(0) = x_1$, then u and γ are path homotopic and have the same extremity: $x_1 = x_2$. This is absurd. □

8.5.2 Proof of Proposition 8.1

Proposition 8.1 provides a deviation inequality on the empirical process renormalized by $I^2(f)$. First of all, for any $M \geq 1$ the Sobolev ball of radius M centred in 0 is denoted by $W_M^{s,p}$. Define the following collections of functions on $\mathbb{R}^{2\ell}$:

$$\mathcal{G}_M \stackrel{\text{def}}{=} \left\{ g_{p_{f,a}};\ f \in W_M^{s,p},\ a \geq a_- \right\} \text{ and } \overline{\mathcal{G}}_M \stackrel{\text{def}}{=} \left\{ g - \mathbb{E}_\star\left[g(Y_0, Y_1)\right];\ g \in \mathcal{G}_M \right\},$$

where $\mathbb{E}_\star$ is the expectation under the distribution $\mathbb{P}$.

The first step of the proof establishes a deviation inequality on the empirical process restricted to the Sobolev balls $W_M^{s,p}$, $\sup_{g\in\mathcal{G}_M} |\nu_n(g)|$. The dependency in M of this inequality allows the determination of a lower bound on υ in the penalty (8.10) sufficient to establish Proposition 8.1. The

second step and conclusion of the proof consists in using the peeling device with the decomposition:

$$W^{s,p} = W_1^{s,p} \cup \bigcup_{k\geq 0} \left\{ W_{2^{k+1}}^{s,p} \setminus W_{2^k}^{s,p} \right\} ,$$

in order to apply the deviation inequalities on $\sup_{g\in\mathcal{G}_M} |\nu_n(g)|$, to each band $\left\{ W_{2^{k+1}}^{s,p} \setminus W_{2^k}^{s,p} \right\}$. Proposition 8.2 gives a concentration inequality for the empirical processes restricted to the Sobolev ball $W_M^{s,p}$.

Proposition 8.2. *Assume H1, H2*(i), *H3*(i) *and H2*(iii), *H4. There exist some positive constants K_1, K_2, C and c, depending on $f_\star$ and $a_\star$ such that, for any $M \geq 1$, any $n \geq 1$ and any $t \geq Cn^{-1/2}$,*

$$\mathbb{P}\left\{ \sup_{g\in\mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g\in\mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{8.32}$$

The proof of Proposition 8.2 relies on the concentration results for Markov chains of [Adamczak et Bednorz, 2012] and is given in the Section 8.6.1.

It remains to control $\mathbb{E}_\star \left[\sup_{g\in\mathcal{G}_M} |\nu_n(g)| \right]$ for any $M \geq 1$.

Proposition 8.3. *Assume H1, H2*(i), *H2*(iii), *H3*(i) *and H4-5. There exists a positive constant K depending on υ, such that, for any $M \geq 1$,*

$$\mathbb{E}_\star \left[\sup_{g\in\mathcal{G}_M} |\nu_n(g)| \right] \leq K M^{\upsilon+1} . \tag{8.33}$$

The proof of Proposition 8.3 is given in Section 8.6.2. We now combine Proposition 8.2 and Proposition 8.3 to obtain a deviation inequality on the empirical process restricted to the truncated collection of functions $\mathcal{G}_M$. Let $\eta > 0$. There exist K_1, K_2 and K_3 such that for any $M \geq 1$, any $n \geq 1$ and any $t \geq \frac{C}{\sqrt{n}}$,

$$\mathbb{P}\left\{ \sup_{g\in\mathcal{G}_M} |\nu_n(g)| \geq K_3 M^{\upsilon+1} + Mt \right\} \quad \leq \quad K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) \quad . \tag{8.34}$$

Proposition 8.1 is obtained applying the peeling device (see for example [Van De Geer, 2000, Lemma 5.14]). Let $\{x_k\}_{k\in\mathbb{N}^\star}$ be some chosen weights such that,

$$\sum_{k\geq 1} \mathrm{e}^{-x_k} < +\infty \quad \text{and, for any } k \geq 1, \quad C \vee 1 \leq x_k \leq 2^{k\upsilon} .$$

Let $k \geq 0$, for any positive x, if $t \stackrel{\text{def}}{=} x + x_k$, for any $n \geq 1$, we have $t \geq C \geq \frac{C}{\sqrt{n}}$. Since $t \geq x_k \geq 1$ and $x_k \leq 2^{k\upsilon}$, we have $e^{-K_2 t^2} \leq e^{-K_2 t}$ and

$t \leq 2^{k\upsilon}(x+1)$. Plugging these relations into (8.34) leads to

$$\mathbb{P}\left\{\sup_{f\in W_{2^k}^{s,p},\ a\geq a_-}\left|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})\right| \geq \frac{2^{k(\upsilon+1)}}{\sqrt{n}}(K_3'+x)\right\} \leq K_1' e^{-K_2(x+x_k)}\,, \tag{8.35}$$

where $K_1' \stackrel{\text{def}}{=} 2K_1$ and $K_3' \stackrel{\text{def}}{=} K_3+1$. If $T \stackrel{\text{def}}{=} 2^{\upsilon+1}K_3'$,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f\in W^{s,p},\ a\geq a_-}\frac{|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})|}{I^2(f)\vee 1} \geq \frac{T+x}{\sqrt{n}}\right\}\\
&\quad\leq \mathbb{P}\left\{\sup_{f\in W_1^{s,p},\ a\geq a_-}\left|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})\right| \geq \frac{T+x}{\sqrt{n}}\right\}\\
&\quad+\sum_{k=0}^{\infty}\mathbb{P}\left\{\sup_{f\in W_{2^{k+1}}^{s,p},\ a\geq a_-}\left|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})\right| \geq 2^{k(\upsilon+1)}\frac{T+x}{\sqrt{n}}\right\}\,.
\end{aligned}$$

However, since $T \geq K_3'$ and $x_1 \geq 0$, by (8.35) applied with $k=0$,

$$\mathbb{P}\left\{\sup_{f\in W_1^{s,p},\ a\geq a_-}\left|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})\right| \geq \frac{T+x}{\sqrt{n}}\right\} \leq K_1' e^{-K_2 x}\,.$$

Therefore, by the definition of T and by (8.35),

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{f\in W^{s,p},\ a\geq a_-}\frac{|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})|}{I^2(f)\vee 1} \geq \frac{T+x}{\sqrt{n}}\right\}\\
&\qquad\leq K_1' e^{-K_2 x}+\sum_{k=0}^{\infty}\mathbb{P}\left\{\sup_{f\in W_{2^{k+1}}^{s,p},\ a\geq a_-}\left|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})\right|\right.\\
&\qquad\qquad\left.\geq (2^{k+1})^{\upsilon+1}\frac{K_3'+x/2^{\upsilon+1}}{\sqrt{n}}\right\}\\
&\qquad\leq K_1' e^{-K_2 x}+\sum_{k=0}^{\infty}K_1' e^{-K_2\left(x/2^{\upsilon+1}+x_{k+1}\right)}\,.
\end{aligned}$$

This last equation ensures the existence of some positive constants K and Σ such that

$$\mathbb{P}\left\{\sup_{f\in W^{s,p},\ a\geq a_-}\frac{|\int g_{p_{f,a}}\mathrm{d}(\mathbb{P}_n-\mathbb{P})|}{I^2(f)\vee 1} \geq \frac{T+x}{\sqrt{n}}\right\} \leq K e^{-\Sigma x}\,.$$

8.6 Additional proofs

8.6.1 Proof of Proposition 8.2

For the sake of simplicity, for any $f \in W^{s,p}$ and $\mathbf{x} = (x_0, x_1) \in K \times K$, we set, for any $a > 0$,

$$\boldsymbol{f}(\mathbf{x}) \stackrel{\text{def}}{=} (f(x_0), f(x_1)) \in \mathbb{R}^{2\ell} \quad \text{and} \quad \boldsymbol{\nu}_a(\mathbf{x}) \stackrel{\text{def}}{=} \nu_a(x_0) q_a(x_0, x_1) \; . \tag{8.36}$$

This section is devoted to the proof of Proposition 8.2 and relies on an intermediate lemma on the envelope functions of the sets $\mathcal{G}_M$ and $\overline{\mathcal{G}}_M$ defined, for any $\mathbf{y} \in \mathbb{R}^{2\ell}$, by

$$\mathrm{G}_M(\mathbf{y}) \stackrel{\text{def}}{=} \sup_{g \in \mathcal{G}_M} g(\mathbf{y}) \quad \text{and} \quad \overline{\mathrm{G}}_M(\mathbf{y}) \stackrel{\text{def}}{=} \sup_{g \in \overline{\mathcal{G}}_M} g(\mathbf{y}) \; .$$

Lemma 8.4. *Assume H2*(i)*, H2*(iii)*, H3*(i) *and H4. There exists a constant* $C_{\mathrm{G}} > 0$ *such that, for any* $\mathbf{y} \in \mathbb{R}^{2\ell}$*,*

$$\mathrm{G}_M(\mathbf{y}) \leq C_G \left(1 + M \|\mathbf{y}\|\right) \; .$$

Proof. For any $\mathbf{y} \in \mathbb{R}^{2\ell}$, any $f \in W_M^{s,p}$ and any $a \geq a_-$,

$$\begin{aligned} g_{p_{f,a}}(\mathbf{y}) &= \frac{1}{2} \ln \frac{1}{2} + \frac{1}{2} \ln \left(1 + \frac{p_{f,a}(\mathbf{y})}{p_{f_\star,a_\star}(\mathbf{y})}\right) \\ &\leq \frac{1}{2} \ln \frac{1}{2} + \frac{1}{2} \ln \left(1 + \sup_{\mathbf{x} \in K^2} \frac{\boldsymbol{\nu}_a(\mathbf{x}) \exp\left(-\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)}{\boldsymbol{\nu}_{a_\star}(\mathbf{x}) \exp\left(-\|\boldsymbol{f}^\star(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)}\right) \; . \end{aligned}$$

By H2(i), (8.3), (8.4) and (8.36), there exists a constant $c_\nu > 1$ such that

$$\sup_{\mathbf{x} \in K^2} \frac{\boldsymbol{\nu}_a(\mathbf{x})}{\boldsymbol{\nu}_{a_\star}(\mathbf{x})} \leq c_\nu \; .$$

Therefore,

$$g_{p_{f,a}}(\mathbf{y}) = \frac{1}{2} \ln \frac{1}{2} + \frac{1}{2} \ln \left(1 + c_\nu \sup_{\mathbf{x} \in K^2} \frac{\exp\left(-\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)}{\exp\left(-\|\boldsymbol{f}^\star(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)}\right) \; .$$

By H4 and H3(i) $f_\star$ is bounded and there exists a constant c such that

$$\frac{\exp\left(-\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)}{\exp\left(-\|\boldsymbol{f}^\star(\mathbf{x}) - \mathbf{y}\|^2 / 2\sigma^2\right)} \leq \exp\left(c(1 + \|\boldsymbol{f}(\mathbf{x})\| \cdot \|\mathbf{y}\|)\right) \; .$$

Then, there exists a constant c such that

$$g_{p_{f,a}}(\mathbf{y}) \leq c(1 + \|\boldsymbol{f}(\mathbf{x})\| \cdot \|\mathbf{y}\|) \; ,$$

and the proof is concluded by (8.7). □

Lemma 8.4 implies that there exists a constant $C > 0$ such that, for any $\mathbf{y} \in \mathbb{R}^{2\ell}$,

$$\overline{\mathrm{G}}_M(\mathbf{y}) \leq C\,(1 + M\|\mathbf{y}\|)\ . \tag{8.37}$$

Let $\boldsymbol{Z} \stackrel{\text{def}}{=} \{\boldsymbol{Z}_k\}_{k\geq 0}$ be the Markov chain, defined, for any $k \geq 0$, by $\boldsymbol{Z}_k \stackrel{\text{def}}{=} (X_{2k}, Y_{2k}, X_{2k+1}, Y_{2k+1})$. For any $\mathbf{z} = (x_0, y_0, x_1, y_1)$ in $(K \times \mathbb{R}^\ell)^2$, we denote by $\mathbb{P}_{\mathbf{z}}$ the conditional version of $\mathbb{P}$ where the starting distribution of the Markov chain $\boldsymbol{Z}$ is the Dirac distribution in $\mathbf{z}$ and by $\mathbb{E}_{\mathbf{z}}$ the associated expectation. The proof of Proposition 8.2 is obtained by integration of the following result with respect to the invariant distribution of the Markov chain $\boldsymbol{Z}$.

Proposition 8.4. *Assume that H2*(i)*, H3*(i) *and H4 hold. There exist some positive constants K_1, K_2, C and c, depending on $f_\star$ and $a_\star$ such that, for any $M > 1$, any $\mathbf{z} \in (K \times \mathbb{R}^\ell)^2$, any $n \geq 1$ and any $t \geq Cn^{-1/2}$,*

$$\mathbb{P}_{\mathbf{z}}\left\{\sup_{g\in\mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star\left[\sup_{g\in\mathcal{G}_M} |\nu_n(g)|\right] + Mt\right\} \leq K_1\left(e^{-K_2 t^2} + e^{-K_2 t}\right)\ . \tag{8.38}$$

Proposition 8.4 is an application of [Adamczak et Bednorz, 2012, Theorem 7] to the class $\{\bar{g}/M;\ \bar{g} \in \overline{\mathcal{G}}_M\}$. Indeed, Lemma 8.4 gives an upper bound for $\overline{\mathrm{G}}_M/M$ which is independent from M. This allows us to apply [Adamczak et Bednorz, 2012, Theorem 7] where all the constants in the upper bound of $\mathbb{P}_{\mathbf{z}}\left\{\sup_{g\in\mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star\left[\sup_{g\in\mathcal{G}_M} |\nu_n(g)|\right] + Mt\right\}$ not depend on M.

By [Adamczak et Bednorz, 2012, Section 3.2], it is sufficient to prove that there exists a small set D, see [Meyn et Tweedie, 1993, Section 5.2], such that

i) there exists $\kappa > 1$ satisfying $\sup_{\mathbf{z}\in D} \mathbb{E}_{\mathbf{z}}\left[\kappa^{\tau_D}\right] < +\infty$, with $\tau_D \stackrel{\text{def}}{=} \min\{k \geq 1;\ \boldsymbol{Z}_k \in D\}$,

ii) the extended chain satisfies a drift condition: there exists a function $V : (K \times \mathbb{R}^\ell)^2 \to \mathbb{R}_+$ and $b > 0$ such that

$$\mathbf{Q}_{a_\star} V(\mathbf{z}) - V(\mathbf{z}) \leq -\exp\left(\overline{\mathrm{G}}_M(\mathbf{y})/M\right) + b\mathbf{1}_D(\mathbf{z})\ ,$$

where $\mathbf{Q}_{a_\star}$ is the Markov transition kernel of the extended chain $\boldsymbol{Z}$. By [Meyn et Tweedie, 1993, Theorem 14.2.3 and Theorem 14.2.4], ii) is satisfied if

$$\sup_{\mathbf{z}\in D} \mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D - 1} \exp\left\{\frac{\overline{\mathrm{G}}_M}{M}(\mathbf{Y}_k)\right\}\right] < +\infty\ , \tag{8.39}$$

where $\mathbf{Y}_k \stackrel{\text{def}}{=} (Y_{2k}, Y_{2k+1})$. In this case, we can choose

$$V(\mathbf{z}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\sigma_D} \exp\left\{\frac{\overline{\mathrm{G}}_M}{M}(\mathbf{Y}_k)\right\}\right]\ ,$$

where $\sigma_D \stackrel{\text{def}}{=} \min\{k \geq 0;\ \boldsymbol{Z}_k \in D\}$. By Lemma 8.4 there exists $K > 0$ (independent from M) such that the function V is upper bounded by K on D. Therefore, [Adamczak et Bednorz, 2012, Theorem 7] states the existence of constants K_1, K_2, c and C such that, for any $t \geq Cn^{-1/2}$, any $\mathbf{z} \in (K \times \mathbb{R}^\ell)^2$ and any $n > 1$,

$$\mathbb{P}_{\mathbf{z}} \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t\sqrt{n}/\log(n)} + e^{-K_2 t\sqrt{n}} + e^{-K_2 t} \right) ,$$

which concludes the proof of Proposition 8.4.

We now turn to the proof of i) and (8.39). By (8.3), it can be proved that the transition kernel $\mathbf{Q}_{a_\star}$ of the extended chain $\boldsymbol{Z}$ also satisfies a strong mixing condition. Therefore any subset of $(K \times \mathbb{R}^\ell)^2$ is a small set for this extended chain. i) and (8.39) can be established by a proper choice of D. By H3(i), there exists $M_\star < +\infty$ such that $\|f_\star\|_{W^{s,p}} = M_\star$. Furthermore, we have, for all $x \in K$, $\|f_\star(x)\| \leq \sqrt{\ell}\kappa M_\star$. Consider the set

$$D \stackrel{\text{def}}{=} K \times K \times \mathrm{B}(0, \sqrt{\ell}\kappa M_\star + \rho) \times \mathrm{B}(0, \sqrt{\ell}\kappa M_\star + \rho) ,$$

where $\mathrm{B}(0, \sqrt{\ell}\kappa M_\star + \rho) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^\ell;\ \|y\| \leq \sqrt{\ell}\kappa M_\star + \rho\}$ and where $\rho > 0$ is a constant to be chosen later. Lemma 8.5 holds easily since, for all $k \geq 0$, ϵ_k is $\mathcal{N}_\ell(0, \sigma^2 I_\ell)$

Lemma 8.5. *For all $k \geq 0$ and all $\mathbf{z} \in D$,*

$$\mathbb{P}_{\mathbf{z}} \{\tau_C > k\} \leq \exp\{-\lambda(\rho)k\} ,$$

where

$$\lambda(\rho) \underset{\rho \to +\infty}{\longrightarrow} +\infty .$$

Then, for $\kappa > 1$ and $\mathbf{z} \in D$,

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} [\kappa^{\tau_D}] &= \sum_{k \geq 1} \mathbb{P}_{\mathbf{z}} \{\tau_D = k\} \kappa^k \\ &\leq \mathrm{e}^{\lambda(\rho)} \sum_{k \geq 1} \mathrm{e}^{-(\lambda(\rho) - \ln \kappa)k} . \end{aligned}$$

The right hand side of the last equation is finite if ρ is chosen sufficiently large. This concludes the proof of i).

Proof of (8.39). Let $\mathbf{z} \in D$. By Lemma 8.4, there exists a constant C such that

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D-1}\exp\left\{\frac{\overline{\mathrm{G}}_M}{M}(\mathbf{Y}_k)\right\}\right] &\leq \mathrm{e}^C\mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D-1}\exp\left\{C\|\mathbf{Y}_k\|\right\}\right]\\ &\leq \mathrm{e}^C\mathbb{E}_{\mathbf{z}}\left[\tau_D\exp\left\{C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right]\\ &\leq \mathrm{e}^C\mathbb{E}_{\mathbf{z}}\left[\tau_D^2\right]^{1/2}\mathbb{E}_{\mathbf{z}}\left[\exp\left\{2C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right]^{1/2}.\end{aligned}$$

By Lemma 8.5, $\sup_{\mathbf{z}\in D}\mathbb{E}_{\mathbf{z}}\left[\tau_D^2\right] < +\infty$. For the second term we write

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}}\left[\exp\left\{2C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right]\\ &\qquad=\sum_{p\geq 1}\mathbb{E}_{\mathbf{z}}\left[\mathbf{1}_{\tau_D=p}\exp\left\{2C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right]\\ &\qquad\leq\sum_{p\geq 1}\mathbb{P}_{\mathbf{z}}\left\{\tau_D=p\right\}^{1/2}\mathbb{E}_{\mathbf{z}}\left[\exp\left\{4C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right]^{1/2}, \qquad (8.40)\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_z\left[\exp\left\{4C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right] &\leq \exp\left\{4C\|\mathbf{y}\|\right\}\\ &\times\left(\left(2\pi\sigma^2\right)^{-\ell}\int\exp\left\{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}+\|\mathbf{y}\|\left(\frac{\sqrt{2\ell}\kappa M_\star}{\sigma^2}+4C\right)\right\}\mathrm{d}\mathbf{y}\right)^{p-1}.\end{aligned}$$

Let $\mathcal{H}$ be the Hausdorff measure on $\mathbb{R}^{2\ell}$ of order $2\ell-1$ restricted to $S^{2\ell-1}$, where $S^{2\ell-1} \stackrel{\text{def}}{=} \{x\in\mathbb{R}^{2\ell};\ \|x\|=1\}$. Then,

$$\begin{aligned}&\left(2\pi\sigma^2\right)^{-\ell}\int\exp\left\{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}+\|\mathbf{y}\|\left(\frac{\sqrt{2\ell}\kappa M_\star}{\sigma^2}+4C\right)\right\}\mathrm{d}\mathbf{y}\\ &\quad\leq\left(2\pi\sigma^2\right)^{-\ell}\int_{\mathbb{R}_+^\star\times S^{2\ell-1}}\exp\left\{-\frac{\|ru\|^2}{2\sigma^2}+\|ru\|\left(\frac{\sqrt{2\ell}\kappa M_\star}{\sigma^2}+4C\right)\right\}\mathcal{H}(\mathrm{d}u)r^{2\ell-1}\mathrm{d}r\\ &\quad=\mathcal{H}(S^{2\ell-1})\left(2\pi\sigma^2\right)^{-\ell}\int_{\mathbb{R}_+^\star}\exp\left\{-\frac{r^2}{2\sigma^2}+r\left(\frac{\sqrt{2\ell}\kappa M_\star}{\sigma^2}+4C\right)\right\}r^{2\ell-1}\mathrm{d}r\\ &\quad\leq\mathcal{H}(S^{2\ell-1})\exp\left\{\frac{\left(\sqrt{2\ell}\kappa M_\star+4C\sigma^2\right)^2}{2\sigma^2}\right\}I_{2\ell-1}\left(\sqrt{2\ell}\kappa M_\star+4C\sigma^2\right),\end{aligned}$$

where, for any $c \in \mathbb{R}$, the sequence $\{I_k(c)\}_{k=1}^{\infty}$ is given by :

$$I_k(c) = \left(2\pi\sigma^2\right)^{-\ell} \int_{\mathbb{R}_+^\star} \exp\left\{-\frac{1}{2\sigma^2}\left[r-c\right]^2\right\} r^k \mathrm{d}r \ .$$

If ξ denotes a Gaussian random variable with mean c and variance σ^2, we have

$$I_k(c) = (2\pi\sigma^2)^{-\ell+1/2}\mathbb{E}\left[\xi^k \mathbf{1}_{\xi>0}\right] \leq \mathbb{E}\left[|\xi|^k\right] \ .$$

Then,

$$I_k(c) \leq (2\pi\sigma^2)^{-\ell+1/2}\mathbb{E}\left[|\xi-c+c|^k\right] \leq (2\pi\sigma^2)^{-\ell+1/2}\sum_{i=0}^{k}\binom{k}{i}c^i\mathbb{E}\left[|\xi-c|^{k-i}\right] \ .$$

If $B(k) \stackrel{\text{def}}{=} (2\pi\sigma^2)^{-\ell+1/2}\max_{0\leq i\leq k}\mathbb{E}\left[|\xi-c|^{k-i}\right]$ (which is independent from c), then

$$I_k(c) \leq B(k)\sum_{i=0}^{k}\binom{k}{i}c^i \leq B(k)(1+c)^k \ .$$

This yields,

$$\mathbb{E}_z\left[\exp\left\{4C\sum_{i=0}^{p-1}\|\mathbf{Y}_i\|\right\}\right] \leq \exp\left\{4C\|\mathbf{y}\|\right\}\left[B(2\ell-1)\mathcal{H}(S^{2\ell-1})\right]^{p-1} \times \exp\left\{\frac{(\sqrt{2\ell}\kappa M_\star + 4C\sigma^2)^2}{2\sigma^2}p\right\}(1+\sqrt{2\ell}\kappa M_\star + 4C\sigma^2)^{(2\ell-1)p} \ .$$

Finally, for all $p \geq 1$ and all $z \stackrel{\text{def}}{=} (\mathbf{x},\mathbf{y}) \in C$,

$$\mathbb{E}_z\left[\exp\left\{4C\sum_{i=0}^{p-1}\|\mathbf{Y}_i\|\right\}\right] \leq \exp\left\{4C\|\mathbf{y}\|\right\}\exp\left\{\eta(4C)(p-1)\right\} \ ,$$

where

$$\eta(4C) \stackrel{\text{def}}{=} \ln(\kappa(\ell)) + \frac{\left(\sqrt{2\ell}\kappa M_\star + 4C\sigma^2\right)^2}{2\sigma^2} + (2\ell-1)\ln\left(1+\sqrt{2\ell}\kappa M_\star + 4C\sigma^2\right) \tag{8.41}$$

and where $\kappa(\ell)$ is a constant depending only on ℓ. Therefore, by (8.40) and Lemma 8.5, this concludes the proof for a sufficiently large ρ. □

8.6.2 Proof of Proposition 8.3

We prove Proposition 8.3 using entropy with bracketing arguments on the class of functions $\mathcal{G}_M$. Define the class of function

$$\mathcal{P}_M \stackrel{\text{def}}{=} \left\{ p_{f,a} : \ f \in W_M^{s,p}, \ a \geq a_- \right\} .$$

Let $\|\cdot\|$ be a norm on $\mathcal{G}$, the entropy with bracketing for the norm $\|\cdot\|$ is defined as follows:

Definition 8.1. *Let $\mathcal{G}$ be some class of functions. For any positive δ, let $N_{[]}(\delta, \mathcal{G}, \|\cdot\|)$ be the smallest N such that there exist a set of brackets $\left\{ \left[g_i^L, g_i^U\right] \right\}_{i=1}^N$ for which $\|g_i^U - g_i^L\| \leq \delta$ for all $i \in \{1, \cdots, N\}$, and for any g in $\mathcal{G}$, there exist $i \in \{1, \cdots, N\}$ such that*

$$g_i^L \leq g \leq g_i^U .$$

$N_{[]}(\delta, \mathcal{G}, \|\cdot\|)$ is called the δ-number with bracketing of $\mathcal{G}$, and $H_{[]}(\delta, \mathcal{G}, \|\cdot\|) = \ln N_{[]}(\delta, \mathcal{G}, \|\cdot\|)$ is the δ-entropy with bracketing of $\mathcal{G}$.

Let $\mathbf{Y} \stackrel{\text{def}}{=} \{\mathbf{Y}_k\}_{k\in\mathbb{Z}}$ be the observations process defined, for all $k \in \mathbb{Z}$, by $\mathbf{Y}_k \stackrel{\text{def}}{=} (Y_{2k}, Y_{2k+1})$. A way of measuring the dependency of the process $\mathbf{Y}$ is the determination of its β-mixing coefficients defined, for any $n \geq 1$,

$$\beta_n \stackrel{\text{def}}{=} \sup_{u>0} \sup_{A \in \mathcal{G}_{u+n}^{\mathbf{Y}}} \left| \mathbb{P}\left(A \middle| \mathcal{H}_u^{\mathbf{Y}}\right) - \mathbb{P}_\star(A) \right| , \tag{8.42}$$

where $\mathcal{H}_u^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\mathbf{Y}_k, k \leq u\right)$ and $\mathcal{G}_{u+n}^{\mathbf{Y}} \stackrel{\text{def}}{=} \sigma\left(\mathbf{Y}_k, k \geq u+n\right)$. Let $\{\beta_n\}_{n\geq 1}$ be defined by (8.42), then, by combining [Rio, 1990, Chapter 9] and the results on the control of the ergodicity of Markov chains by coupling techniques of [Douc *et al.*, 2004a], it can be proved that there exist β in $(0,1)$ and $C > 0$ such that, for any $n \geq 1$,

$$\beta_n \leq C\beta^n . \tag{8.43}$$

Define the mixing rate function $\beta(\cdot)$, by $\beta(t) \stackrel{\text{def}}{=} \beta_{\lfloor t \rfloor}$ if $t \geq 1$ and $\beta(t) = 1$ otherwise. For any numerical function g, we denote by $\mathcal{Q}_g$ the quantile function of $|g(\mathbf{Y}_0)|$ and define the norm $\|g\|_{2,\beta}$ as in [Doukhan *et al.*, 1995] by

$$\|g\|_{2,\beta} \stackrel{\text{def}}{=} \left[\int_0^1 \beta^{-1}(u) \left[\mathcal{Q}_g(u) du\right]^2 \right]^{1/2} ,$$

where β^{-1} denotes the càdlàg inverse of the function $\beta(\cdot)$. We also denote by $\mathcal{L}_{2,\beta}(\mathbb{P})$ the class of numerical functions g such that $\|g\|_{2,\beta} < \infty$.

Proposition 8.5. *Assume H2*(i), *H2*(iii) *and H4. For any* $p' > 1$, $s' > 2\ell/p'$, *any integer* $r > 1$ *and any even number* b *such that* $b > s' + 2\ell(1-1/p')$, *there exists a positive constant* C *such that:*

$$\forall \epsilon > 0,\ M \geq 1, \quad H_{[]}(\epsilon, \mathcal{G}_M, \|\cdot\|_{2,\beta}) \leq C \left(\frac{M^{s'+b+\frac{2}{p'}\ell}}{\epsilon^{2r}} \right)^{2\ell/s'} .$$

The proof of Proposition 8.5 is given in Section 8.6.3. Proposition 8.5 allows to apply [Doukhan *et al.*, 1995, Theorem 3] to the class of functions $\mathcal{G}_M$. Let B be the function defined on $\mathbb{R}_+$ by $B(x) \stackrel{\text{def}}{=} \int_0^x \beta^{-1}(t)\mathrm{d}t$ and, for any $\epsilon > 0$, $\delta_M(\epsilon) \stackrel{\text{def}}{=} \sup_{t \leq \epsilon} \mathcal{Q}_{G_M}(t)\sqrt{B(t)}$. The following lemma is an application of [Doukhan *et al.*, 1995, Lemma 2] it allows to bound the $\|\cdot\|_{2,\beta}$-norm by $\|\cdot\|_{\mathrm{L}_{2r}}$ for all $r > 1$. For any g in $\mathcal{L}_{2,\beta}$ and any $r > 1$,

$$\|g\|_{2,\beta} \leq \|g\|_{\mathrm{L}_{2r}(\mathbb{P})} \sqrt{\int_0^1 u^{-1/r}\beta^{-1}(u)\mathrm{d}u} . \tag{8.44}$$

Moreover, by [Massart et Picard, 2007, Lemma 7.26], for any natural number $r > 1$ there exist a positive constant C such that for any f in $W^{s,p}$ and $a > 0$,

$$\|g_{p_{f,a}}\|^{2r}_{\mathrm{L}_{2r}(\mathbb{P})} \leq C h(p_{f,a}, p_{f_\star,a_\star}) . \tag{8.45}$$

The Hellinger distance being bounded, (8.45) and (8.44) state the existence of a positive number d such that $\|g_{p_{f,a}}\|_{2,\beta} \leq d$ for all f in $W^{s,p}$ and $a \geq a_-$. Define for any $M \geq 1$, $\varphi_M \stackrel{\text{def}}{=} \int_0^d \sqrt{H_{[]}(u, \mathcal{G}_M, \|\cdot\|_{2,\beta})}\mathrm{d}u$. Thus, by [Doukhan *et al.*, 1995, Theorem 3], provided that $\delta_M(\epsilon) \underset{\epsilon\to 0}{\longrightarrow} 0$, there exists a constant C such that

$$\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] \leq C\varphi_M \left(1 + \frac{\delta_M(1 \wedge \epsilon_{n,M})}{d} \right) , \tag{8.46}$$

where $\epsilon_{n,M}$ is the unique solution on $\mathbb{R}_+$ of the equation:

$$\frac{x^2}{B(x)} = \frac{\varphi_M^2}{nd^2} .$$

In the sequel, we control the quantities appearing in (8.46). By Proposition 8.5 and the definition of φ_M, for any $p' > 1$, $s' > 2\ell/p'$, $r > 1$ and any even number b such that $b > s' + 2\ell(1-1/p')$, there exists a constant C depending on p', s', r and b, such that

$$\varphi_M \leq C \left(M^{s'+b+\frac{2}{p'}\ell} \right)^{\ell/s'} \int_0^d u^{-2r\ell/s'}\mathrm{d}u , \tag{8.47}$$

with $\int_0^d u^{-2r\ell/s'}\mathrm{d}u < \infty$ whenever $s' > 2r\ell$. If b is the unique even number such that $s' + 2\ell(1-1/p') < b \leq \lceil s' + 2\ell(1-1/p')\rceil + 1$ and if s' tends to infinity in (8.47), then $\left(M^{s'+b+\frac{2}{p'}\ell}\right)^{\ell/s'} \underset{s'\to\infty}{\longrightarrow} M^{2\ell}$ and it follows that,

$$\forall \eta > 0,\ \exists C > 0,\ \forall M \geq 1,\ \varphi_M \leq C M^{2\ell+\eta}\ .$$

By H5, there exists a constant C such that

$$\varphi_M \leq C M^{\upsilon}\ . \tag{8.48}$$

Lemma 8.6. *Assume H3*(i) *and H4. There exists $C > 0$, such that, for any $M \geq 1$ and any $t \in (0,1)$,*

$$\mathcal{Q}_{G_M}(t) \leq CM\left(1 + \ln^{1/2}\left(\frac{1}{t}\right)\right)\ .$$

Proof. Set $\boldsymbol{\epsilon}_0 = (\epsilon_0, \epsilon_1)$, set $u > C_G$, where C_G is defined in Lemma 8.4,

$$\begin{aligned}
\mathbb{P}\left\{G_M(\mathbf{Y}_0) \geq u\right\} &\leq \mathbb{P}\left\{C_G(1 + M\|\mathbf{Y}_0\|) \geq u\right\} \\
&\leq \mathbb{P}\left\{\|\mathbf{Y}_0\| \geq \frac{u/C_G - 1}{M}\right\} \\
&\leq \mathbb{P}\left\{\|\boldsymbol{f}_\star(\mathbf{X}_0)\| + \|\boldsymbol{\epsilon}_0\| \geq \frac{u/C_G - 1}{M}\right\} \\
&\leq \mathbb{P}\left\{\|\boldsymbol{\epsilon}_0\| \geq \frac{u/C_G - 1}{M} - c_\infty\right\}\ ,
\end{aligned}$$

where $\|\boldsymbol{f}_\star(\mathbf{x})\| \leq c_\infty$ for all $\mathbf{x}$ in K^2 ($f_\star$ is bounded by H4 and H3(i)). Using Cirelson-Ibragimov-Sudakov inequality, see [Massart et Picard, 2007, Section 1.2.1], for any $x > 0$

$$\mathbb{P}\left\{\frac{1}{\sigma}\left(\|\boldsymbol{\epsilon}_0\| - \mathbb{E}(\|\boldsymbol{\epsilon}_0\|)\right) \geq x\right\} \leq \mathrm{e}^{-\frac{x^2}{2}}\ .$$

Hence,

$$\begin{aligned}
\mathbb{P}\left\{G_M(\mathbf{Y}_0) \geq u\right\} &\leq \exp\left(-\frac{\left(\frac{u/C_G-1}{M} - c_\infty - \mathbb{E}(\|\boldsymbol{\epsilon}_0\|)\right)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\left(\frac{c_1 u-1}{M} - c_2\right)^2}{2\sigma^2}\right)\ ,
\end{aligned}$$

where $c_1 \overset{\text{def}}{=} \frac{1}{C_G}$ and $c_2 \overset{\text{def}}{=} c_\infty + \mathbb{E}_\star\left[\|\boldsymbol{\epsilon}_0\|\right]$. Setting $1 \geq t > 0$, let u be such that $t = \mathbb{P}\left\{G_M(\mathbf{Y}_0) \geq u\right\}$, then,

$$\exp\left(-\frac{\left(\frac{c_1 u-1}{M} - c_2\right)^2}{2\sigma^2}\right) \geq t$$

implies

$$u \leq \frac{1}{c_1}\left(Mc_2 + 1 + M\left(2\sigma^2 \ln\left(\frac{1}{t}\right)\right)^{1/2}\right) ,$$

which concludes the proof. □

By (8.43), there exists a constant $C > 0$ such that,

$$\forall x \in (0,1),\ B(x) \leq Cx\left(1 + \ln\left(\frac{1}{x}\right)\right) . \tag{8.49}$$

Lemma 8.7. *Assume H3*(i) *and H4. There exists $C > 0$ such that for any $0 < \epsilon \leq 1$,*

$$\delta_M(\epsilon) \leq CM\left(\epsilon^{1/2}\ln\left(\frac{1}{\epsilon}\right)\mathbf{1}_{\epsilon \leq \mathrm{e}^{-2}} + \mathbf{1}_{\epsilon > \mathrm{e}^{-2}}\right) \leq CM .$$

Proof. By Lemma 8.6,

$$\mathcal{Q}_{G_M}(t) \leq CM\left(1 + \ln^{1/2}\left(\frac{1}{t}\right)\right) .$$

Therefore, by (8.49)

$$\mathcal{Q}_{G_M}(t)\sqrt{B(t)} \leq CM\left(1 + \ln^{1/2}\left(\frac{1}{t}\right)\right)\sqrt{t\left\{1 + \ln\left(\frac{1}{t}\right)\right\}} .$$

For $t \geq e^{-1}$, $\ln(t^{-1}) \leq 1$ and $\mathcal{Q}_{G_M}(t)\sqrt{B(t)} \leq CM$.
For $t \leq e^{-1}$, $\ln(t^{-1}) \geq 1$, this yields, for $t \leq \mathrm{e}^{-1}$,

$$\mathcal{Q}_{G_M}(t)\sqrt{B(t)} \leq CM \ln\left(\frac{1}{t}\right)\sqrt{t} .$$

The proof is concluded upon noting that the function $t \mapsto t^{1/2}\ln\left(t^{-1}\right)$ reaches is maximum at e^{-2}. □

Finally, Lemma 8.7 ensures that $\delta_M(\epsilon) \underset{\epsilon \to 0}{\longrightarrow} 0$ for any $M \geq 1$, and Proposition 8.3 results from (8.46), Lemma 8.7, and (8.48).

8.6.3 Proof of Proposition 8.5

The aim of this section is to prove Proposition 8.5. The computation of $H_{[]}(\epsilon, \mathcal{G}_M, \|\cdot\|_{2,\beta})$ is not an easy task as the dependency of $\|g\|_{2,\beta}$ in g only appears trough the quantile function $\mathcal{Q}_g$. Moreover, the dependency in M of the entropy $H_{[]}(\epsilon, \mathcal{G}_M, \|\cdot\|_{2,\beta})$ is not straightforward. The next lemma allows to control the bracketing entropy of $\mathcal{G}_M$ relatively to the $\|\cdot\|_{2,\beta}$-norm by the entropy of $\mathcal{P}_M$ relatively to the $\|\cdot\|_{\mathrm{L}_1(\mathbb{R}^{2\ell})}$-norm .

Lemma 8.8. *For any integer $r > 1$, there exists a constant C such that:*

$$H_{[]}\left(\epsilon, \mathcal{G}_M, \|\cdot\|_{2,\beta}\right) \leq C H_{[]}\left(\epsilon^{2r}, \mathcal{P}_M, \|\cdot\|_{\mathrm{L}_1(\mathbb{R}^{2\ell})}\right) .$$

Proof. The function ln being increasing, if $[P_U, P_L]$ is a bracket for $\mathcal{P}_M$, then $[g_{P_U}, g_{P_L}]$ is a bracket for $\mathcal{G}_M$. Moreover, by [Massart et Picard, 2007, Lemma 7.26], there exists a positive constant C such that

$$\|g_{P_U} - g_{P_L}\|^{2r}_{\mathrm{L}_{2r}(\mathbb{P})} \leq C \|\sqrt{P_U} - \sqrt{P_L}\|^2_{\mathrm{L}_2(\mathbb{R}^{2\ell})} .$$

Moreover it is straightforward that $\|\sqrt{P_U} - \sqrt{P_L}\|^2_{\mathrm{L}_2(\mathbb{R}^{2\ell})} \leq \|P_U - P_L\|_{\mathrm{L}_1(\mathbb{R}^{2\ell})}$. The proof is concluded using (8.44). □

[Nickel et Potscher, 2001] provides results on the entropy rates for function classes of Besov or Sobolev-type. Therefore, to control the entropy rate of $\mathcal{P}_M$ we prove that it is included in some weighted Sobolev Space. Define the polynomial weighting function $\langle \mathbf{y} \rangle^b \stackrel{\text{def}}{=} \left(1 + \|\mathbf{y}\|^2\right)^{b/2}$ parametrized by $b \in \mathbb{R}$ where $\mathbf{y} \in \mathbb{R}^{2\ell}$. Furthermore, define for $p' \geq 1$, and $s' > 2\ell/p'$ the weighted Sobolev space

$$W^{s',p'}\left(\mathbb{R}^{2\ell}, \langle \mathbf{y} \rangle^b\right) \stackrel{\text{def}}{=} \left\{ f : \ f \cdot \langle \mathbf{y} \rangle^b \in W^{s',p'}\left(\mathbb{R}^{2\ell}, \mathbb{R}\right) \right\} .$$

Lemma 8.9. *Assume H2*(i), *H2*(iii) *and H4. For any $p' \geq 1$, $s' > 2\ell/p'$ and any even and positive number b, there exists a positive constant C such that*

$$\forall f \in W^{s,p}, \ \forall a \geq a_-, \ \|p_{f,a} \cdot \langle \mathbf{y} \rangle^b\|_{W^{s',p'}(\mathbb{R}^{2\ell},\mathbb{R})} \leq C \left(1 \vee \|f\|_{W^{s,p}}\right)^{s'+b+\frac{2}{p'}\ell} .$$

Proof. Let f be a function in $W^{s,p}$, for any $a \geq a_-$,

$$\|p_{f,a} \cdot \langle \mathbf{y} \rangle^b\|^{p'}_{W^{s',p'}(\mathbb{R}^{2\ell},\mathbb{R})} = \sum_{|\alpha| \leq s'} \|D^\alpha \left(p_{f,a} \cdot \langle \mathbf{y} \rangle^b\right)\|^{p'}_{\mathrm{L}_{p'}} .$$

Applying the general Leibniz rule component by component, for any $\alpha \in \mathbb{N}^{2\ell}$,

$$D^\alpha \left(p_{f,a} \cdot \langle \mathbf{y} \rangle^b\right) = \sum_{\alpha' \leq \alpha} \binom{\alpha}{\alpha'} D^{\alpha'}(\langle \mathbf{y} \rangle^b) D^{\alpha - \alpha'}(p_{f,a}) , \tag{8.50}$$

where $\binom{\alpha}{\alpha'} \stackrel{\text{def}}{=} \prod_{j=1}^{2\ell} \binom{\alpha_j}{\alpha'_j}$. Thus, Lemma 8.9 results from the control of $\|D^{\alpha^{(1)}}(\langle \mathbf{y} \rangle^b) D^{\alpha^{(2)}}(p_{f,a})\|_{\mathrm{L}_{p'}}$ for any given $\alpha^{(1)}$ and $\alpha^{(2)}$ in $\mathbb{N}^{2\ell}$. It is straightforward that, for any α in $\mathbb{N}^{2\ell}$, there exists a polynomial function P_α whose degree does not exceed $|\alpha|$ such that, for any $\mathbf{y} \in \mathbb{R}^{2\ell}$,

$$D^\alpha p_{f,a}(\mathbf{y}) = \int_{\mathbf{x} \in K^2} P_\alpha(\boldsymbol{f}(\mathbf{x}) - \mathbf{y}) \exp\left\{ -\frac{\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2}{2\sigma^2} \right\} \boldsymbol{\nu}_a(\mathbf{x}) \mathrm{d}\mathbf{x} . \tag{8.51}$$

Moreover, since b is an even number, that for any $\alpha \in \mathbb{R}^{2\ell}$ such that $|\alpha| \leq b$, $D^\alpha \langle \mathbf{y} \rangle^b$ is a polynomial function denoted by $P_{b,\alpha}$ whose degree does not exceed $b - |\alpha|$. In the case where $|\alpha| > b$, $D^\alpha \langle \mathbf{y} \rangle^b = 0$. Since $P_{\alpha^{(2)}}$ and $P_{b,\alpha^{(1)}}$ are both polynomial functions, and since (8.7) ensures that, for any $\mathbf{x}$ in K^2, $\|\boldsymbol{f}(\mathbf{x})\| \leq \sqrt{2}\kappa \|f\|_{W^{s,p}} \leq \sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}})$, there exist a constant C depending on $\alpha^{(1)}, \alpha^{(2)}$ and b such that, for any $\mathbf{y}$ in $\mathbb{R}^{2\ell}$ and any $\mathbf{x}$ in K^2,

$$\begin{aligned}
&\left| P_{b,\alpha^{(1)}}(\mathbf{y}) P_{\alpha^{(2)}}(\boldsymbol{f}(\mathbf{x}) - \mathbf{y}) \right| \\
&\qquad \leq C(1 + \|\mathbf{y}\|)^{b - |\alpha^{(1)}|} \mathbf{1}_{|\alpha^{(1)}| \leq b} \times \left(\sqrt{2}\kappa \left(1 \vee \|f\|_{W^{s,p}}\right) + \|\mathbf{y}\| \right)^{|\alpha^{(2)}|} .
\end{aligned}$$

Define the following subset of $\mathbb{R}^{2\ell}$

$$A_f \stackrel{\text{def}}{=} \left\{ \mathbf{y} \in \mathbb{R}^{2\ell};\ \|\mathbf{y}\| \leq \sqrt{2}\kappa \left(1 \vee \|f\|_{W^{s,p}}\right) \right\} .$$

$\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|$ can be lower bounded by 0 when $\mathbf{y}$ belongs to A_f and by $\left| \sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}}) - \|\mathbf{y}\| \right|$ when $\mathbf{y}$ belongs to A_f^c. Therefore, uniformly in $\mathbf{x} \in K^2$,

$$\exp\left\{ -\frac{\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2}{2\sigma^2} \right\} \leq \mathbf{1}_{A_f}(\mathbf{y}) + \mathbf{1}_{A_f^c}(\mathbf{y}) e^{-\frac{1}{2\sigma^2}\left(\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}}) - \|y\|\right)^2} .$$

Thus, there exists a constant $C > 0$, independent from a, such that, for any $\mathbf{y}$ in $\mathbb{R}^{2\ell}$,

$$\begin{aligned}
&\left| D^{\alpha^{(1)}}(\langle \mathbf{y} \rangle^b) D^{\alpha^{(2)}}(p_{f,a})(\mathbf{y}) \right| \\
&\leq C(1 \vee \|f\|_{W^{s,p}})^{\alpha^{(2)}} \cdot (1 + \|\mathbf{y}\|)^{b - |\alpha^{(1)}|} \left(1 + \frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}})} \right)^{|\alpha^{(2)}|} \\
&\qquad \times \left[\mathbf{1}_{A_f}(\mathbf{y}) + \mathbf{1}_{A_f^c}(\mathbf{y}) e^{-\frac{1}{2\sigma^2}\left(\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}}) - \|y\|\right)^2} \right] .
\end{aligned}$$

Therefore, for any $p' \geq 1$,

$$\|D^{\alpha^{(1)}}(\langle \mathbf{y} \rangle^b) D^{\alpha^{(2)}}(p_{f,a})\|_{\mathrm{L}_{p'}}^{p'} \leq C(1 \vee \|f\|_{W^{s,p}})^{p'\alpha^{(2)}} \left(I_1 + I_2\right) ,$$

where,

$$\begin{aligned}
I_1 &\stackrel{\text{def}}{=} \int_{A_f} (1 + \|\mathbf{y}\|)^{p'(b - |\alpha^{(1)}|)} \left(1 + \frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}})} \right)^{p'|\alpha^{(2)}|} \mathrm{d}\mathbf{y} , \\
I_2 &\stackrel{\text{def}}{=} \int_{A_f^c} (1 + \|\mathbf{y}\|)^{p'(b - |\alpha^{(1)}|)} \left(1 + \frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}})} \right)^{p'|\alpha^{(2)}|} \\
&\qquad \times \mathrm{e}^{-\frac{p'}{2\sigma^2}\left(\sqrt{2}\kappa(1 \vee \|f\|_{W^{s,p}}) - \|y\|\right)^2} \mathrm{d}\mathbf{y} .
\end{aligned}$$

By applying the change of variable $\mathbf{y}' = \frac{1}{\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})}\mathbf{y}$ in I_1 and I_2, and noting that $\mathrm{e}^{-\frac{p'\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})}{2\sigma^2}(1-\|y'\|)^2} \leq \mathrm{e}^{-\frac{\sqrt{2}\kappa p'}{2\sigma^2}(1-\|y'\|)^2}$, there exists a constant C such that

$$\|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^b)D^{\alpha^{(2)}}(p_{f,a})\|_{\mathrm{L}_{p'}}^{p'} \leq C(1\vee\|f\|_{W^{s,p}})^{p'(|\alpha^{(2)}|-|\alpha^{(1)}|+b)+2\ell} . \quad (8.52)$$

Using (8.52) in (8.50) with $\alpha^{(1)} = \alpha'$ and $\alpha^{(2)} = \alpha - \alpha'$ for any $|\alpha| \leq s'$ and $\alpha' \leq \alpha$ concludes the proof. □

Hence Lemma 8.9 ensures that, for any $p' \geq 1$, $s' > 2\ell/p'$ any even integer b, the renormalized classes of functions $\mathcal{P}_M/M^{s'+b+\frac{2}{p'}\ell}$, $M \geq 1$ belong to the same bounded subspace of $W^{s',p'}(\mathbb{R}^{2\ell}, \langle\mathbf{y}\rangle^b)$. By [Nickel et Potscher, 2001, Corollary 4], for any $p' \geq 1$, and any $s' > 2\ell/p'$, provided that $b > s' + 2\ell(1-\frac{1}{p'})$, there exists a constant C such that

$$\forall\epsilon > 0,\ H_{[]}\left(\epsilon, \mathcal{P}_M/M^{s'+b+\frac{2}{p'}\ell}, \|\cdot\|_{\mathrm{L}_1(\mathbb{R}^{2\ell})}\right) \leq C\epsilon^{-2\ell/s'} . \quad (8.53)$$

Lemma 8.8 and (8.53) conclude the proof of Proposition 8.5.

Acknowledgments

The authors are grateful to Elisabeth Gassiat and Eric Moulines for their fruitful remarks. They also wish to thank Aurélien Poiret and Frédéric Leroux for their advice.

Annexe A

Résultats supplémentaires pour le Chapitre 6

This is a supplementary material to the paper [Le Corff et Fort, 2011b].

It contains technical discussions and/or results adapted from published papers. In Sections A.1 and A.2, we provide results - useful for the proofs of some theorems in [Le Corff et Fort, 2011b] - which are close to existing results in the literature.

It also contains, in Section A.3, additional plots for the numerical analyses in [Le Corff et Fort, 2011b].

A.1 Detailed proofs of Chapter 6

A.1.1 Proof of Theorem 6.2

Proof. By Proposition 6.4, it is sufficient to prove that

$$\left|\mathrm{W}\circ\mathrm{R}(\theta_n)-\mathrm{W}\circ\bar{\theta}(\widetilde{S}_n)\right| \underset{n\to+\infty}{\longrightarrow} 0\ , \quad \mathbb{P}-\text{a.s.}\ . \tag{A.1}$$

By Theorem 6.1, the function $\bar{\mathrm{S}}$ given by (6.12) is continuous on Θ and then $\bar{\mathrm{S}}(\Theta) \overset{\text{def}}{=} \{s\in\mathcal{S};\exists\theta\in\Theta, s=\bar{\mathrm{S}}(\theta)\}$ is compact and, for any $\delta>0$, we can define the compact subset $\bar{\mathrm{S}}(\Theta,\delta) \overset{\text{def}}{=} \left\{s\in\mathbb{R}^d;\mathrm{d}(s,\bar{\mathrm{S}}(\Theta))\leq\delta\right\}$ of $\mathcal{S}$, where $\mathrm{d}(s,\bar{\mathrm{S}}(\Theta)) \overset{\text{def}}{=} \inf_{s'\in\bar{\mathrm{S}}(\Theta)}|s-s'|$. Let $\delta>0$ (small enough) and $\varepsilon>0$. Since $\mathrm{W}\circ\bar{\theta}$ is continuous (see A1(c) and Proposition 6.1) and $\bar{\mathrm{S}}(\Theta,\delta)$ is compact, $\mathrm{W}\circ\bar{\theta}$ is uniformly continuous on $\bar{\mathrm{S}}(\Theta,\delta)$ and there exists $\eta>0$ s.t.,

$$\forall x,y\in\bar{\mathrm{S}}(\Theta,\delta)\ , \quad |x-y|\leq\eta\Rightarrow|\mathrm{W}\circ\bar{\theta}(x)-\mathrm{W}\circ\bar{\theta}(y)|\leq\varepsilon\ . \tag{A.2}$$

Set $\alpha \stackrel{\text{def}}{=} \delta \wedge \eta$ and $\Delta S_n \stackrel{\text{def}}{=} |\bar{\mathrm{S}}(\theta_n) - \widetilde{S}_n|$. We write,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\mathrm{W} \circ \bar{\theta}(\bar{\mathrm{S}}(\theta_n)) - \mathrm{W} \circ \bar{\theta}(\widetilde{S}_n)\right| \geq \varepsilon\right\} \\
&\qquad = \mathbb{P}\left\{\left|\mathrm{W} \circ \bar{\theta}(\bar{\mathrm{S}}(\theta_n)) - \mathrm{W} \circ \bar{\theta}(\widetilde{S}_n)\right| \geq \varepsilon; \Delta S_n > \delta\right\} \\
&\qquad\qquad + \mathbb{P}\left\{\left|\mathrm{W} \circ \bar{\theta}(\bar{\mathrm{S}}(\theta_n)) - \mathrm{W} \circ \bar{\theta}(\widetilde{S}_n)\right| \geq \varepsilon; \Delta S_n \leq \delta\right\} \\
&\qquad \leq \mathbb{P}\left\{\Delta S_n > \delta\right\} + \mathbb{P}\left\{\Delta S_n > \eta\right\} \leq 2\mathbb{P}\left\{\Delta S_n > \alpha\right\} \ .
\end{aligned}$$

By the Markov inequality and Theorem 6.1, for all $p \in (2, \bar{p})$, there exists a constant C s.t.

$$\mathbb{P}\left\{\left|\mathrm{W} \circ \bar{\theta}(\bar{\mathrm{S}}(\theta_n)) - \mathrm{W} \circ \bar{\theta}(\widetilde{S}_n)\right| \geq \varepsilon\right\} \leq \frac{2}{\alpha^p}\mathbb{E}\left[|\bar{\mathrm{S}}(\theta_n) - \widetilde{S}_n|^p\right] \leq C\tau_{n+1}^{-p/2} \ .$$

(A.1) follows from A5 and the Borel-Cantelli lemma (since $p > 2$ and $a > 1$). □

Proposition A.1 shows that we can address equivalently the convergence of the statistics $\{\widetilde{S}_n\}_{n\geq 0}$ to some fixed point of G and the convergence of the sequence $\{\theta_n\}_{n\geq 0}$ to some fixed point of R.

Proposition A.1. *Assume A1-2, A3-($\bar{p}$), A4*(a)*, A5 and A6-($\bar{p}$) hold for some $\bar{p} > 2$.*

*(i) Let $\theta_\star \in \mathcal{L}$. Set $s_\star \stackrel{\text{def}}{=} \bar{\mathrm{S}}(\theta_\star) = \mathrm{G}(s_\star)$. Then, $\mathbb{P}-$*a.s.*,*

$$\lim_{n\to+\infty} \left|\widetilde{S}_n - s_\star\right| \mathbf{1}_{\lim_n \theta_n = \theta_\star} = 0 \ .$$

*(ii) Let $s_\star \in \mathcal{S}$ s.t. $\mathrm{G}(s_\star) = s_\star$. Set $\theta_\star \stackrel{\text{def}}{=} \bar{\theta}(s_\star) = \mathrm{R}(\theta_\star)$. Then $\mathbb{P}-$*a.s.*,*

$$\lim_{n\to+\infty} |\theta_n - \theta_\star| \, \mathbf{1}_{\lim_n \widetilde{S}_n = s_\star} = 0 \ .$$

Proof. Let $\bar{\mathrm{S}}$ be given by (6.12). By Theorem 6.1 and A5,

$$\lim_n \left(\widetilde{S}_n - \bar{\mathrm{S}}(\theta_n)\right) = 0 \quad \mathbb{P}-\text{a.s.}$$

By Theorem 6.1, $\bar{\mathrm{S}}$ is continuous. Hence,

$$\lim_n \left|\widetilde{S}_n - \bar{\mathrm{S}}(\theta_\star)\right| \mathbf{1}_{\lim_n \theta_n = \theta_\star} = 0 \quad \mathbb{P}-\text{a.s.}$$

and the proof of (i) follows. Since $\bar{\theta}$ is continuous, (ii) follows. □

A.1.2 Proof of Proposition 6.2

We start with rewriting some definitions and assumptions introduced in [Le Corff et Fort, 2011b]. Define the sequences μ_n and ρ_n, $n \geq 0$ by $\mu_0 = 0$, $\rho_0 = \widetilde{S}_0 - s_\star$ and

$$\mu_n \stackrel{\text{def}}{=} \Gamma\mu_{n-1} + e_n \;, \qquad \rho_n \stackrel{\text{def}}{=} \widetilde{S}_n - s_\star - \mu_n \;, \qquad n \geq 1 \;, \tag{A.3}$$

where,

$$e_n \stackrel{\text{def}}{=} \widetilde{S}_n - \bar{\mathrm{S}}(\theta_n) \;, \qquad n \geq 1 \;, \tag{A.4}$$

and $\bar{\mathrm{S}}$ is given by (6.12).

Proof. Let $p \in (2, \bar{p})$. By (A.3), for all $n \geq 1$, $\mu_n = \sum_{k=0}^{n-1} \Gamma^k e_{n-k}$. By A8 and the Minkowski inequality, for all $n \geq 1$, $\|\mu_n\|_p \leq \sum_{k=0}^{n-1} \gamma^k \|e_{n-k}\|_p$. By Theorem 6.1, there exists a constant C s.t. for any $n \geq 1$,

$$\|\mu_n\|_p \leq C \sum_{k=0}^{n-1} \gamma^k \sqrt{\frac{1}{\tau_{n+1-k}}} \;.$$

By [Pólya et Szegő, 1976, Result 178, p. 39] and A5, this yields $\sqrt{\tau_n}\mu_n = O_{\mathrm{L}_p}(1)$.

By A8, using a Taylor expansion with integral form of the remainder term,

$$\begin{aligned}
&\mathrm{G}(\widetilde{S}_{n-1}) - \mathrm{G}(s_\star) - \Gamma\left(\widetilde{S}_{n-1} - s_\star\right) \\
&\qquad = \sum_{i,j=1}^{d} \left(\widetilde{S}_{n-1,i} - s_{\star,i}\right)\left(\widetilde{S}_{n-1,j} - s_{\star,j}\right) R_{n-1}(i,j) \\
&\qquad = \sum_{i,j=1}^{d} (\mu_{n-1,i} + \rho_{n-1,i})(\mu_{n-1,j} + \rho_{n-1,j}) R_{n-1}(i,j) \;,
\end{aligned}$$

where $x_{n,i}$ denotes the i-th component of $x_n \in \mathbb{R}^d$ and

$$R_n(i,j) \stackrel{\text{def}}{=} \int_0^1 (1-t) \frac{\partial^2 \mathrm{G}}{\partial s_i \partial s_j}\left(s_\star + t(\widetilde{S}_n - s_\star)\right) \mathrm{d}t \;, \qquad n \in \mathbb{N}, 1 \leq i,j \leq d \;.$$

Observe that under A8, $\limsup_n |R_n| \mathbf{1}_{\lim_n \theta_n = \theta_\star} < \infty$ w.p.1. Define for $n \geq 1$ and $k \leq n$,

$$H_n \stackrel{\text{def}}{=} \sum_{i=1}^{d} (2\mu_{n,i} + \rho_{n,i}) R_n(i,\cdot) \;, \qquad r_n \stackrel{\text{def}}{=} \sum_{i,j=1}^{d} R_n(i,j)\mu_{n,i}\mu_{n,j} \;, \tag{A.5}$$

$$\psi(n,k) \stackrel{\text{def}}{=} (\Gamma + H_n) \cdots (\Gamma + H_k) \;, \tag{A.6}$$

with the convention $\psi(n,n+1) \stackrel{\text{def}}{=} \text{Id}$. By (A.3),

$$\rho_n = \psi(n-1,0)\rho_0 + \sum_{k=0}^{n-1} \psi(n-1,k+1) r_k \ . \tag{A.7}$$

Since $\sqrt{\tau_n}\mu_n = O_{\mathrm{L}_p}(1)$, A5 and $p>2$ imply that $\mu_n \underset{n\to+\infty}{\longrightarrow} 0$, $\mathbb{P}-$a.s. Then, by (A.3), $\rho_n \mathbf{1}_{\lim_n \theta_n=\theta_\star} \underset{n\to+\infty}{\longrightarrow} 0$, $\mathbb{P}-$a.s. and by (A.5) $\lim_{n\to+\infty} |H_n|\, \mathbf{1}_{\lim_n \theta_n=\theta_\star} = 0$, $\mathbb{P}-$a.s. Let $\widetilde{\gamma} \in (\gamma,1)$, where γ is given by A8. Since $\lim_{n\to+\infty} |H_n|\, \mathbf{1}_{\lim_n \theta_n=\theta_\star} = 0$, there exists a $\mathbb{P}-$a.s. finite random variable Z_1 s.t., for all $0 \le k \le n-1$,

$$|\psi(n-1,k)|\, \mathbf{1}_{\lim_n \theta_n=\theta_\star} \le \widetilde{\gamma}^{n-k} Z_1 \mathbf{1}_{\lim_n \theta_n=\theta_\star} \ . \tag{A.8}$$

Therefore, $|\psi(n-1,0)\rho_0|\, \mathbf{1}_{\lim_n \theta_n=\theta_\star} \le \widetilde{\gamma}^n Z_1 \ |\rho_0| \quad \mathbb{P}-$a.s., and, by A3-($\bar{p}$), (6.3), $\mathbb{E}\left[|\rho_0|^{\bar{p}}\right] < +\infty$ which implies that $\rho_0 < +\infty$ $\mathbb{P}-$a.s. Since $\widetilde{\gamma} < 1$, the first term in the RHS of (A.7) is $\tau_n^{-1} O_{\mathrm{L}_p}(1) O_{\mathrm{a.s}}(1)$.

We now consider the second term in the RHS of (A.7). From equation (A.8),

$$\left| \sum_{k=0}^{n-1} \psi(n-1,k+1) r_k \right| \mathbf{1}_{\lim_n \theta_n=\theta_\star} \le Z_1 \ \sum_{k=0}^{n-1} \widetilde{\gamma}^{n-k-1} \, |r_k|\, \mathbf{1}_{\lim_n S_n=s_\star} \ , \quad \mathbb{P}-\text{a.s.}$$

By (A.5) and A8, there exists a $\mathbb{P}-$a.s. finite random variable Z_2 s.t.

$$|r_k|\, \mathbf{1}_{\lim_n \theta_n=\theta_\star} \le Z_2 \sum_{i,j=1}^{d} \mu_{k,i}\mu_{k,j} \ , \quad \mathbb{P}-\text{a.s.}$$

In addition, since $\sqrt{\tau_n}\mu_n = O_{\mathrm{L}_p}(1)$, there exists a constant C s.t.

$$\left\| \sum_{k=0}^{n-1} \widetilde{\gamma}^{n-k-1} \sum_{i,j=1}^{d} \mu_{k,i}\mu_{k,j} \right\|_{p/2} \le C \ \sum_{k=0}^{n-1} \frac{\widetilde{\gamma}^{n-k-1}}{\tau_k} \ .$$

Applying again [Pólya et Szegő, 1976, Result 178, p. 39] yields that the second term in the RHS of (A.7) is $\tau_n^{-1} O_{\mathrm{a.s}}(1) O_{\mathrm{L}_{p/2}}(1)$. □

A.2 General results on HMM

In this section, we derive results on the forgetting properties of HMM (Section A.2.1), on their applications to bivariate smoothing distributions (Section A.2.2), on the asymptotic behavior of the normalized log-likelihood (Section A.2.3) and on the normalized score (Section A.2.4).

For any sequence $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any function $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}$, denote by h_s the function on $\mathbb{X}^2 \to \mathbb{R}$ given by

$$h_s(x,x') \stackrel{\text{def}}{=} h(x,x',y_s) \ . \tag{A.9}$$

A.2.1 Forward and Backward forgetting

In this section, the dependence on θ is dropped from the notation for better clarity. For any $s \in \mathbb{Z}$ and any $A \in \mathcal{X}$, define

$$L_s(x, A) \stackrel{\text{def}}{=} \int m(x, x') g(x', y_{s+1}) \mathbb{1}_A(x') \lambda(\mathrm{d}x') \;, \tag{A.10}$$

and, for any $s \leq t$ denote by $L_{s:t}$ the composition of the kernels defined by

$$L_{s:s} \stackrel{\text{def}}{=} L_s \;, \qquad L_{s:u+1}(x, A) \stackrel{\text{def}}{=} \int L_{s:u}(x, \mathrm{d}x') L_{u+1}(x', A) \;.$$

By convention, $L_{s:s-1}$ is the identity kernel: $L_{s:s-1}(x, A) = \delta_x(A)$. For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, any probability distribution χ on $(\mathbb{X}, \mathcal{X})$ and for any integers such that $r \leq s < t$, let us define two Markov kernels on $(\mathbb{X}, \mathcal{X})$ by

$$\mathrm{F}_{s,t}(x, A) \stackrel{\text{def}}{=} \frac{\int L_s(x, \mathrm{d}x_{s+1}) 1_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X})}{L_{s:t-1}(x, \mathbb{X})} \;, \tag{A.11}$$

$$\mathrm{B}_s^{\chi,r}(x, A) \stackrel{\text{def}}{=} \frac{\int \phi_{s|r:s}^{\chi,r}(\mathrm{d}x_s) 1_A(x_s) m(x_s, x)}{\int \phi_{s|r:s}^{\chi,r}(\mathrm{d}x_s) m(x_s, x)} \;, \tag{A.12}$$

where

$$\phi_{s|r:s}^{\chi,r}(A) \stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_r) L_{r:s-1}(x_r, \mathrm{d}x_s) 1_A(x_s)}{\int \chi(\mathrm{d}x_r) L_{r:s-1}(x_r, \mathbb{X})} \;.$$

Finally, the Dobrushin coefficient of a Markov kernel $F : (\mathbb{X}, \mathcal{X}) \longrightarrow [0, 1]$ is defined by:

$$\delta(F) \stackrel{\text{def}}{=} \frac{1}{2} \sup_{(x,x') \in \mathbb{X}^2} \left|\left| F(x, \cdot) - F(x', \cdot) \right|\right|_{\mathrm{TV}} \;.$$

Lemma A.1. *Assume that there exist positive numbers σ_-, σ_+ such that $\sigma_- \leq m(x, x') \leq \sigma_+$ for any $x, x' \in \mathbb{X}$. Then for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $\delta(\mathrm{F}_{s,t}) \leq \rho$ and $\delta(\mathrm{B}_s^{\chi,r}) \leq \rho$ where $\rho \stackrel{\text{def}}{=} \sigma_-/\sigma_+$.*

Proof. Let r, s, t be such that $r \leq s < t$. Under the stated assumptions,

$$\int L_s(x_s, \mathrm{d}x_{s+1}) 1_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \geq \sigma_- \int g(x_{s+1}, y_{s+1}) 1_A(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(\mathrm{d}x_{s+1})$$

and

$$L_{s:t-1}(x_s, \mathbb{X}) \leq \sigma_+ \int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(\mathrm{d}x_{s+1}) \;.$$

This yields to

$$\mathrm{F}_{s,t}(x_s, A) \geq \frac{\sigma_-}{\sigma_+} \frac{\int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \mathbb{1}_A(x_{s+1}) \lambda(\mathrm{d}x_{s+1})}{\int g(x_{s+1}, y_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathbb{X}) \lambda(\mathrm{d}x_{s+1})} \;.$$

Similarly, the assumption implies

$$\mathrm{B}_s^{\chi,r}(x_{s+1}, A) \geq \frac{\sigma_-}{\sigma_+} \phi_{s|r:s}^{\chi,r}(A) \ ,$$

which gives the upper bound for the Dobrushin coefficients, see for example [Cappé *et al.*, 2005, Lemma 4.3.13]. □

Lemma A.2. *Assume that there exist positive numbers σ_-, σ_+ such that $\sigma_- \leq m(x, x') \leq \sigma_+$ for any $x, x' \in \mathbb{X}$. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$.*

(i) for any bounded function h, any probability distributions χ and $\widetilde{\chi}$ and any integers $r \leq s \leq t$

$$\left| \frac{\int \chi(\mathrm{d}x_r) L_{r:s-1}(x_r, \mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathbb{X})}{\int \chi(\mathrm{d}x_r) L_{r:t-1}(x_r, \mathbb{X})} - \frac{\int \widetilde{\chi}(\mathrm{d}x_r) L_{r:s-1}(x_r, \mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathbb{X})}{\int \widetilde{\chi}(\mathrm{d}x_r) L_{r:t-1}(x_r, \mathbb{X})} \right| \leq \rho^{s-r} \mathrm{osc}(h) \ , \tag{A.13}$$

(ii) for any bounded function h, for any non-negative functions f and $\widetilde{f}$ and any integers $r \leq s \leq t$

$$\left| \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)} - \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \widetilde{f}(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \widetilde{f}(x_t)} \right| \leq \rho^{t-s} \mathrm{osc}(h) \ . \tag{A.14}$$

Proof of (i). See [Cappé *et al.*, 2005, Proposition 4.3.23].

Proof of (ii) When $s = t$, then (ii) is equal to

$$\left| \frac{\int \chi(\mathrm{d}x_t) h(x_t) f(x_t)}{\int \chi(\mathrm{d}x_t) f(x_t)} - \frac{\int \chi(\mathrm{d}x_t) h(x_t) \widetilde{f}(x_t)}{\int \chi(\mathrm{d}x_t) \widetilde{f}(x_t)} \right| \ .$$

This is of the form $(\eta - \widetilde{\eta})\, h$ where η and $\widetilde{\eta}$ are probability distributions on $(\mathbb{X}, \mathcal{X})$. Then,

$$|(\eta - \widetilde{\eta})\, h| \leq \frac{1}{2} \, ||\eta - \widetilde{\eta}||_{\mathrm{TV}} \, \mathrm{osc}\,(h) \leq \mathrm{osc}(h) \ .$$

Let $s < t$. By definition of the backward smoothing kernel, see (A.12),

$$\mathrm{B}_s^{\chi,s}(x_{s+1}, A) = \frac{\int \chi(\mathrm{d}x_s) 1_A(x_s) m(x_s, x_{s+1})}{\int \chi(\mathrm{d}x_s) m(x_s, x_{s+1})} \ .$$

Therefore,

$$\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)$$
$$= \int \chi(\mathrm{d}x_s) L_s(x_s, \mathrm{d}x_{s+1}) \mathrm{B}_s^{\chi,s} h(x_{s+1}) L_{s+1:t-1}(x_{s+1}, \mathrm{d}x_t) f(x_t) \ .$$

By repeated application of the backward smoothing kernel we have

$$\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)$$
$$= \int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \mathrm{B}_{t-1:s}^{\chi,s} h(x_t) f(x_t) \ ,$$

where we denote by $\mathrm{B}_{t-1:s}^{\chi,s}$ the composition of the kernels defined by induction for $s \leq u$

$$\mathrm{B}_{s:s}^{\chi,s} \stackrel{\text{def}}{=} \mathrm{B}_s^{\chi,s} \ , \qquad \mathrm{B}_{u:s}^{\chi,s}(x, A) \stackrel{\text{def}}{=} \int \mathrm{B}_u^{\chi,s}(x, \mathrm{d}x') \mathrm{B}_{u-1:s}^{\chi,s}(x', A) \ .$$

Finally, by definition of $\phi_{t|s:t}^{\chi,s}$

$$\left| \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) f(x_t)} - \frac{\int \chi(\mathrm{d}x_s) h(x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \widetilde{f}(x_t)}{\int \chi(\mathrm{d}x_s) L_{s:t-1}(x_s, \mathrm{d}x_t) \widetilde{f}(x_t)} \right|$$
$$= \left| \frac{\phi_{t|s:t}^{\chi,s} \left[\left(\mathrm{B}_{t-1:s}^{\chi,s} h \right) f \right]}{\phi_{t|s:t}^{\chi,s} \left[f \right]} - \frac{\phi_{t|s:t}^{\chi,s} \left[\left(\mathrm{B}_{t-1:s}^{\chi,s} h \right) \widetilde{f} \right]}{\phi_{t|s:t}^{\chi,s} \left[\widetilde{f} \right]} \right| \ .$$

This is of the form $(\eta - \widetilde{\eta}) \, \mathrm{B}_{t-1:s}^{\chi,s} h$ where η and $\widetilde{\eta}$ are probability distributions on $(\mathbb{X}, \mathcal{X})$. The proof of the second statement is completed upon noting that

$$\left| \eta \mathrm{B}_{t-1:s}^{\chi,s} h - \widetilde{\eta} \mathrm{B}_{t-1:s}^{\chi,s} h \right| \leq \frac{1}{2} \left| \left| \eta - \widetilde{\eta} \right| \right|_{\mathrm{TV}} \mathrm{osc} \left(\mathrm{B}_{t-1:s}^{\chi,s} h \right)$$
$$\leq \frac{1}{2} \left| \left| \mu - \widetilde{\mu} \right| \right|_{\mathrm{TV}} \delta \left(\mathrm{B}_{t-1:s}^{\chi,s} \right) \mathrm{osc}(h) \leq \rho^{t-s} \mathrm{osc}(h) \ ,$$

where we used Lemma A.1 in the last inequality. □

A.2.2 Bivariate smoothing distribution

Proposition A.2. *Assume A2. Let χ, $\widetilde{\chi}$ be two distributions on $(\mathbb{X}, \mathcal{X})$. For any measurable function $h : \mathbb{X}^2 \times \mathbb{Y} \to \mathbb{R}^d$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\sup_{x,x'} |h(x, x', y_s)| < +\infty$ for any $s \in \mathbb{Z}$*

(i) For any $r < s \leq t$ and any $\ell_1, \ell_2 \geq 1$,

$$\sup_{\theta\in\Theta} \left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t+\ell_2}(h,\mathbf{y}) \right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \ . \quad \text{(A.15)}$$

(ii) For any $\theta \in \Theta$, there exists a function $\mathbf{y} \mapsto \Phi_\theta(h,\mathbf{y})$ s.t. for any distribution χ on $(\mathbb{X},\mathcal{X})$ and any $r < s \leq t$

$$\sup_{\theta\in\Theta} \left| \Phi^{\chi,r}_{\theta,s,t}(h,\mathbf{y}) - \Phi_\theta(h,\vartheta^s\mathbf{y}) \right| \leq \left(\rho^{s-1-r} + \rho^{t-s}\right) \mathrm{osc}(h_s) \ . \quad \text{(A.16)}$$

Remark. (a) If $\chi = \widetilde{\chi}$, $\ell_1 = 0$ and $\ell_2 \geq 1$, (A.15) becomes

$$\sup_{\theta\in\Theta} \left| \Phi^{\chi,r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r}_{\theta,s,t+\ell_2}(h,\mathbf{y}) \right| \leq \rho^{t-s}\mathrm{osc}(h_s) \ .$$

(b) if $\ell_2 = 0$ and $\ell_1 \geq 1$, (A.15) becomes

$$\sup_{\theta\in\Theta} \left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t}(h,\mathbf{y}) \right| \leq \rho^{s-1-r}\mathrm{osc}(h_s) \ .$$

Proof. (i) Let r, s, t such that $r < s \leq t$, $\ell_1, \ell_2 \geq 1$, and $\theta \in \Theta$. Define the distribution $\chi_{\theta,r-\ell_1:r}$ on $(\mathbb{X},\mathcal{X})$ by

$$\chi_{\theta,r-\ell_1:r}(A) \stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_{r-\ell_1}) L_{\theta,r-\ell_1:r-1}(x_{r-\ell_1},\mathrm{d}x_r) 1_A(x_r)}{\int \chi(\mathrm{d}x_{r-\ell_1}) L_{\theta,r-\ell_1:r-1}(x_{r-\ell_1},\mathbb{X})} \ , \quad \forall A \in \mathcal{X} \ .$$

We write $\left| \Phi^{\widetilde{\chi},r}_{\theta,s,t}(h,\mathbf{y}) - \Phi^{\chi,r-\ell_1}_{\theta,s,t+\ell_2}(h,\mathbf{y}) \right| \leq \widetilde{T}_1 + \widetilde{T}_2$ where,

$$\widetilde{T}_1 \stackrel{\text{def}}{=} \left| \frac{\int \widetilde{\chi}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) h_s(x_{s-1},x_s) L_{\theta,s-1}(x_{s-1},\mathrm{d}x_s) L_{\theta,s:t-1}(x_s,\mathbb{X})}{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r,\mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) h_s(x_{s-1},x_s) L_{\theta,s-1}(x_{s-1},\mathrm{d}x_s) L_{\theta,s:t-1}(x_s,\mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r,\mathbb{X})} \right| \ ,$$

and

$$\widetilde{T}_2 \stackrel{\text{def}}{=} \left| \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) h_s(x_{s-1},x_s) L_{\theta,s-1}(x_{s-1},\mathrm{d}x_s) L_{\theta,s:t-1}(x_s,\mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r,\mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) h_s(x_{s-1},x_s) L_{\theta,s-1}(x_{s-1},\mathrm{d}x_s) L_{\theta,s:t+\ell_2-1}(x_s,\mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t+\ell_2-1}(x_r,\mathbb{X})} \right| \ .$$

Set $\bar{h}_{s,t} : x \mapsto \int F_{\theta,s-1,t}(x,\mathrm{d}x_s) h_s(x,x_s)$ where $F_{\theta,s-1,t}$ is the forward smoothing kernel (see (A.11)). Then,

$$\widetilde{T}_1 = \left| \frac{\int \widetilde{\chi}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) \bar{h}_{s,t}(x_{s-1}) L_{\theta,s-1:t-1}(x_{s-1},\mathbb{X})}{\int \widetilde{\chi}_r(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r,\mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:s-2}(x_r,\mathrm{d}x_{s-1}) \bar{h}_{s,t}(x_{s-1}) L_{\theta,s-1:t-1}(x_{s-1},\mathbb{X})}{\int \chi_{\theta,r-\ell_1:r}(\mathrm{d}x_r) L_{\theta,r:t-1}(x_r,\mathbb{X})} \right| \ .$$

By Lemma A.2(i),

$$\widetilde{T}_1 \leq \rho^{s-1-r}\mathrm{osc}(\bar{h}_{s,t}) \leq 2\rho^{s-1-r}\sup_{x\in\mathbb{X}}|\bar{h}_{s,t}(x)| \leq 2\rho^{s-1-r}\sup_{(x,x')\in\mathbb{X}^2}|h_s(x,x')| \,.$$

Set $\widetilde{h}_s : x \mapsto \int \mathrm{B}_{\theta,s-1}^{\chi_\theta, r-\ell_{1:s-1}, s-1}(x, \mathrm{d}x_{s-1}) h_s(x_{s-1}, x)$, where $\mathrm{B}_{\theta,s-1}^{\chi_\theta, r-\ell_{1:s-1}, s-1}$ is the backward smoothing kernel (see (A.12)). Then,

$$\widetilde{T}_2 = \left| \frac{\int \chi_{\theta,r-\ell_{1:s}}(\mathrm{d}x_s)\widetilde{h}_s(x_s)L_{\theta,s:t-1}(x_s,\mathrm{d}x_t)L_{\theta,t:t+\ell_2-1}(x_t,\mathbb{X})}{\int \chi_{\theta,r-\ell_{1:s}}(\mathrm{d}x_s)L_{\theta,s:t-1}(x_s,\mathrm{d}x_t)L_{\theta,t:t+\ell_2-1}(x_t,\mathbb{X})} \right.$$
$$\left. - \frac{\int \chi_{\theta,r-\ell_{1:s}}(\mathrm{d}x_s)\widetilde{h}_s(x_s)L_{\theta,s:t-1}(x_s,\mathbb{X})}{\int \chi_{\theta,r-\ell_{1:s}}(\mathrm{d}x_s)L_{\theta,s:t-1}(x_s,\mathbb{X})} \right| \,.$$

Then, by Lemma A.2(ii),

$$\widetilde{T}_2 \leq \rho^{t-s}\mathrm{osc}(\widetilde{h}_s) \leq 2\rho^{t-s}\sup_{x\in\mathbb{X}}|\widetilde{h}_s(x)| \leq 2\rho^{t-s}\sup_{(x,x')\in\mathbb{X}^2}|h_s(x,x')| \,.$$

The proof is concluded upon noting that, for any constant c, $\mathrm{osc}(h) = 2\inf_{c\in\mathbb{R}}\left\{\sup_{(x,x')\in\mathbb{X}^2}|h_s(x,x') - c|\right\}$.

(ii) By (A.15), for any increasing sequence of non negative integers $(r_\ell)_{\ell\geq 0}$, $(t_\ell)_{\ell\geq 0}$ s.t. $\lim r_\ell = \lim t_\ell = +\infty$, the sequence $\{\Phi_{\theta,0,t_\ell}^{\chi,-r_\ell}(h,\mathbf{y})\}_{\ell\geq 0}$ is a Cauchy sequence uniformly in θ and χ. Then, there exists a limit $\Phi_\theta(h,\mathbf{y})$ s.t.

$$\lim_{\ell\to+\infty}\sup_{\chi}\sup_{\theta\in\Theta}\left|\Phi_{\theta,0,t_\ell}^{\chi,-r_\ell}(h,\mathbf{y}) - \Phi_\theta(h,\mathbf{y})\right| = 0 \,. \tag{A.17}$$

We write, for any $r < s \leq t$ and any $\ell \geq 1$

$$\left|\Phi_{\theta,s,t}^{\chi,r}(h,\mathbf{y}) - \Phi_\theta(h,\vartheta^s\mathbf{y})\right|$$
$$\leq \left|\Phi_{\theta,s,t}^{\chi,r}(h,\mathbf{y}) - \Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h,\mathbf{y})\right| + \left|\Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h,\mathbf{y}) - \Phi_\theta(h,\vartheta^s\mathbf{y})\right| \,.$$

Since $\Phi_{\theta,s,t+\ell}^{\chi,r-\ell}(h,\mathbf{y}) = \Phi_{\theta,0,t+\ell-s}^{\chi,r-\ell-s}(h,\vartheta^s\mathbf{y})$, Proposition A.2(i) yields

$$\left|\Phi_{\theta,s,t}^{\chi,r}(h,\mathbf{y}) - \Phi_\theta(h,\vartheta^s\mathbf{y})\right| \leq \left(\rho^{s-r-1} + \rho^{t-s}\right)\mathrm{osc}(h_s)$$
$$+ \left|\Phi_{\theta,0,t+\ell-s}^{\chi,r-\ell-s}(h,\vartheta^s\mathbf{y}) - \Phi_\theta(h,\vartheta^s\mathbf{y})\right| \,.$$

The proof is concluded by (A.17). □

A.2.3 Limiting normalized log-likelihood

Define for any $r \leq s$,

$$\delta_{\theta,s}^{\chi,r}(\mathbf{y}) \stackrel{\mathrm{def}}{=} \ell_{\theta,s+1}^{\chi,r}(\mathbf{y}) - \ell_{\theta,s}^{\chi,r}(\mathbf{y}) \,, \tag{A.18}$$

where $\ell^{\chi,r}_{\theta,s+1}(\mathbf{y})$ is defined by

$$\ell_{\chi,r}\theta, s+1(\mathbf{Y}) \stackrel{\text{def}}{=} \log \int \chi(\mathrm{d}x_r) \prod_{u=r+1}^{s+1} m_\theta(x_{u-1}, x_u) g_\theta(x_u, \mathbf{Y}_u)\ \lambda(\mathrm{d}x_{r+1:s+1}) \ . \tag{A.19}$$

For any $T > 0$ and any probability distribution χ on $(\mathbb{X}, \mathcal{X})$, we thus have

$$\ell^{\chi,0}_{\theta,T}(\mathbf{y}) = \sum_{s=0}^{T-1} \left(\ell^{\chi,0}_{\theta,s+1}(\mathbf{y}) - \ell^{\chi,0}_{\theta,s}(\mathbf{y}) \right) = \sum_{s=0}^{T-1} \delta^{\chi,0}_{\theta,s}(\mathbf{y}) \ . \tag{A.20}$$

It is established in Lemma A.3 that for any $\theta \in \Theta$, $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $s \geq 0$ and any initial distribution χ, the sequence $\{\delta^{\chi,s-r}_{\theta,s}(\mathbf{y})\}_{r\geq 0}$ is a Cauchy sequence and its limit does not depend upon χ. Regularity conditions on this limit are given in Lemmas A.4 and A.5. Finally, Theorem A.1 shows that for any θ, $\lim_T T^{-1}\ell^{\chi,0}_{\theta,T}(\mathbf{Y})$ exists w.p.1. and this limit is a (deterministic) continuous function in θ.

Lemma A.3. *Assume A2.*

(i) For any $\ell, r, s \geq 0$, any initial distributions χ, χ' on $\mathbb{X}$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$

$$\sup_{\theta\in\Theta} \left| \delta^{\chi,s-r}_{\theta,s}(\mathbf{y}) - \delta^{\chi',s-r-\ell}_{\theta,s}(\mathbf{y}) \right| \leq \frac{2}{1-\rho}\rho^r \ .$$

(ii) For any $\theta \in \Theta$, there exists a function $\mathbf{y} \mapsto \delta_\theta(\mathbf{y})$ such that for any initial distribution χ, any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ and any $r, s \geq 0$,

$$\sup_{\theta\in\Theta} \left| \delta^{\chi,s-r}_{\theta,s}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \leq \frac{2}{1-\rho}\rho^r \ .$$

Proof. Proof of (i). Let $s \geq 0$ and r and r' be such that $r' > r$. By (A.18) and (A.19), we have $|\delta^{\chi,s-r}_{\theta,s}(\mathbf{y}) - \delta^{\chi',s-r'}_{\theta,s}(\mathbf{y})| = |\log\alpha - \log\beta|$ where

$$\begin{aligned} \alpha &\stackrel{\text{def}}{=} \frac{\int \chi(\mathrm{d}x_{s-r}) \prod_{i=s-r+1}^{s+1} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(\mathrm{d}x_i)}{\int \chi(\mathrm{d}x_{s-r}) \prod_{i=s-r+1}^{s} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(\mathrm{d}x_i)} \ , \\ \beta &\stackrel{\text{def}}{=} \frac{\int \chi'(\mathrm{d}x_{s-r'}) \prod_{i=s-r'+1}^{s+1} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(\mathrm{d}x_i)}{\int \chi'(\mathrm{d}x_{s-r'}) \prod_{i=s-r'+1}^{s} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(\mathrm{d}x_i)} \ . \end{aligned} \tag{A.21}$$

We prove that

$$\alpha \wedge \beta \geq \sigma_- \int g_\theta(x_{s+1}, y_{s+1}) \lambda(\mathrm{d}x_{s+1}) \ , \tag{A.22}$$

$$|\alpha - \beta| \leq 2\rho^r \sigma_+ \int g_\theta(x_{s+1}, y_{s+1}) \lambda(\mathrm{d}x_{s+1}) \ , \tag{A.23}$$

and the proof is concluded since $|\log\alpha - \log\beta| \leq |\alpha - \beta|/(\alpha \wedge \beta)$.

The minorization on α and β is a consequence of A2 upon noting that α and β are of the form $\int \mu(\mathrm{d}x_s) m_\theta(x_s, x_{s+1}) g_\theta(x_{s+1}, y_{s+1}) \lambda(\mathrm{d}x_{s+1})$ for some probability measure μ. The upper bound on $|\alpha - \beta|$ is a consequence of Lemma A.2(i) applied with

$$\tilde{\chi}(\mathrm{d}x_{s-r}) \leftarrow \int_{\mathbb{X}^{r'-r}} \chi'(\mathrm{d}x_{s-r'}) \left\{ \prod_{i=s-r'}^{s-r-1} g_\theta(x_i, y_i) m_\theta(x_i, x_{i+1}) \right\} \lambda(\mathrm{d}x_{s-r'+1:s-r})$$

and $h(u) \leftarrow \int g_\theta(x_{s+1}, y_{s+1}) m_\theta(u, x_{s+1}) \lambda(\mathrm{d}x_{s+1})$.

Proof of (ii). By (i), for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, the sequence $\{\delta_{\theta,0}^{\chi,-r}(\mathbf{y})\}_{r \geq 0}$ is a Cauchy sequence uniformly in θ: there exists a limit denoted by $\delta_\theta(\mathbf{y})$ - which does not depend upon χ - such that

$$\lim_{r \to +\infty} \sup_{\theta \in \Theta} \left| \delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \delta_\theta(\mathbf{y}) \right| = 0 \, . \tag{A.24}$$

We write for $r \leq r'$

$$\left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \leq \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_{\theta,s}^{\chi,s-r'}(\mathbf{y}) \right| + \left| \delta_{\theta,s}^{\chi,s-r'}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \, .$$

Observe that by definition, $\delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) = \delta_{\theta,0}^{\chi,-r}(\vartheta^s \circ \mathbf{y})$. This property, combined with Lemma A.3(i), yield

$$\sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \leq \frac{2}{1-\rho} \rho^r + \sup_{\theta \in \Theta} \left| \delta_{\theta,0}^{\chi,-r'}(\vartheta^s \circ \mathbf{y}) - \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \, .$$

When $r' \to +\infty$, the second term in the rhs tends to zero by (A.24) - for fixed $\mathbf{y}, s$ and χ -. This concludes the proof. □

Lemma A.4. *Assume A2. For any* $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ *and* $s \geq 0$,

$$\sup_{r \geq 0} \sup_{\theta \in \Theta} \left| \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) \right| \leq |\log \sigma_+ b_+(\mathbf{y}_{s+1})| + |\log \sigma_- b_-(\mathbf{y}_{s+1})| \, ,$$

and, for any $r \geq 0$,

$$\sup_{\theta \in \Theta} |\delta_\theta(\mathbf{y})| \leq \frac{2}{(1-\rho)} \rho^r + |\log \sigma_+ b_+(\mathbf{y}_1)| + |\log \sigma_- b_-(\mathbf{y}_1)| \, .$$

Proof. For any $0 < m \leq A/B \leq M$, $|\log(A/B)| \leq |\log M| + |\log m|$. Note that by definition, $\delta_{\theta,s}^{\chi,0}(\mathbf{y})$ is of the form $\log(A/B)$ and under A4(b), $\sigma_- b_-(y_{s+1}) \leq A/B \leq \sigma_+ b_+(y_{s+1})$. The second upper bound is a consequence of Lemma A.3(ii). □

Lemma A.5. *Assume A1-2 and A4. Then,* $\theta \mapsto \mathbb{E}\left[\delta_\theta(\mathbf{Y})\right]$ *is continuous on* Θ *and*

$$\lim_{\eta \to 0} \mathbb{E}\left[\sup_{\{\theta, \theta' \in \Theta; |\theta - \theta'| < \eta\}} |\delta_\theta(\mathbf{Y}) - \delta_{\theta'}(\mathbf{Y})| \right] = 0 \, , \quad \mathbb{P}-\text{a.s.} \tag{A.25}$$

Proof. By the dominated convergence theorem, Lemma A.4 and A4(b), $\theta \mapsto \mathbb{E}_\star[\delta_\theta(\mathbf{Y})]$ is continuous if $\theta \mapsto \delta_\theta(\mathbf{y})$ is continuous for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$. Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$. By Lemma A.3(ii), $\lim_{r\to+\infty} \sup_{\theta\in\Theta} |\delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \delta_\theta(\mathbf{y})| = 0$. Therefore, $\theta \mapsto \delta_\theta(\mathbf{y})$ is continuous provided for any $r \geq 0$, $\theta \mapsto \delta_{\theta,0}^{\chi,-r}(\mathbf{y})$ is continuous (for fixed $\mathbf{y}$ and χ). By definition of $\delta_{\theta,0}^{\chi,-r}(\mathbf{y})$, see (A.18), it is sufficient to prove that $\theta \mapsto \ell_{\theta,s}^{\chi,-r}(\mathbf{y})$ is continuous for $s \in \{0,1\}$. By definition of $\ell_{\theta,s}^{\chi,-r}(\mathbf{y})$, see (A.19),

$$\ell_{\theta,s}^{\chi,-r}(\mathbf{y}) = \log \int \chi(\mathrm{d}x_{-r}) \prod_{i=-r+1}^{s} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i) \lambda(\mathrm{d}x_i) \; .$$

Under A1(a), $\theta \mapsto \prod_{i=-r+1}^{s} m_\theta(x_{i-1}, x_i) g_\theta(x_i, y_i)$ is continuous on Θ, for any $x_{-r:s}$ and $\mathbf{y}$. In addition, under A1, for any $\theta \in \Theta$,

$$\left| \prod_{i=-r+1}^{s} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1}) \right| = \exp\left((s+r)\phi(\theta) + \left\langle \psi(\theta), \sum_{i=-r+1}^{s} S(x_i, x_{i+1}, \mathbf{y}_{i+1}) \right\rangle \right) \; .$$

Since, by A1, ϕ and ψ are continuous, and since Θ is compact, there exist constants C_1 and C_2 such that,

$$\sup_{\theta\in\mathcal{K}} \left| \prod_{i=-r+1}^{s} m_\theta(x_i, x_{i+1}) g_\theta(x_{i+1}, \mathbf{y}_{i+1}) \right| \leq C_1 \exp\left(C_2 \sum_{i=-r+1}^{s} \sup_{x,x'} |S(x, x,', \mathbf{y}_{i+1})| \right) \; .$$

Since the measure $\chi(\mathrm{d}x_{-r}) \prod_{i=-r+1}^{s} \lambda(\mathrm{d}x_i)$ is finite, the dominated convergence theorem now implies that $\ell_{\theta,s}^{\chi,-r}(\mathbf{y})$ is continuous on Θ.

For the proof of (A.25), let us apply the dominated convergence theorem again. Since Θ is compact, for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$, $\theta \mapsto \delta_\theta(\mathbf{y})$ is uniformly continuous and $\lim_{\eta\to 0} \sup_{|\theta-\theta'|<\eta} |\delta_\theta(\mathbf{y}) - \delta_{\theta'}(\mathbf{y})| = 0$. In addition, we have by Lemma A.4

$$\sup_{\{\theta,\theta'\in\Theta; |\theta-\theta'|<\eta\}} |\delta_\theta(\mathbf{y}) - \delta_{\theta'}(\mathbf{y})| \leq 2 \sup_{\theta\in\Theta} |\delta_\theta(\mathbf{y})| \leq \frac{4}{(1-\rho)} + 2\left\{ |\log \sigma_+ \, b_+(y_1)| + |\log \sigma_- \, b_-(y_1)| \right\} \; .$$

Under A4, this upper bound is $\mathbb{P}$-integrable. This concludes the proof. □

Theorem A.1. *Assume A1-2 and A4. Define the function* $\ell : \Theta \to \mathbb{R}$ *by* $\ell(\theta) \stackrel{\text{def}}{=} \mathbb{E}\left[\delta_\theta(\mathbf{Y})\right]$, *where* $\delta_\theta(\mathbf{y})$ *is defined in Lemma A.4.*

(i) The function $\theta \mapsto \ell(\theta)$ *is continuous on* Θ.

(ii) For any initial distribution χ *on* $(\mathbb{X}, \mathcal{X})$

$$\left|\frac{1}{T}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) - \ell(\theta)\right| \underset{T\to+\infty}{\longrightarrow} 0\ , \quad \mathbb{P}-\text{a.s.} \tag{A.26}$$

where $\ell^{\chi,0}_{\theta,T}(\mathbf{Y})$ *is defined in* (A.19).

(iii) For any initial distribution χ *on* $(\mathbb{X}, \mathcal{X})$

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) - \ell(\theta)\right| \underset{T\to+\infty}{\longrightarrow} 0\ , \quad \mathbb{P}-\text{a.s.} \tag{A.27}$$

Proof. (i) is proved in Lemma A.5.

(ii) By (A.20), for any $T > 0$, we have, for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$:

$$\begin{aligned}\frac{1}{T}\ell^{\chi,0}_{\theta,T}(\mathbf{y}) &= \frac{1}{T}\sum_{s=0}^{T-1}\delta^{\chi,0}_{\theta,s}(\mathbf{y}) \\ &= \frac{1}{T}\sum_{s=0}^{T-1}\left(\delta^{\chi,0}_{\theta,s}(\mathbf{y}) - \delta_\theta(\vartheta^s\circ\mathbf{y})\right) + \frac{1}{T}\sum_{s=0}^{T-1}\delta_\theta(\vartheta^s\circ\mathbf{y})\ .\end{aligned}$$

By Lemma A.3(ii), for any $0 \le s \le T-1$, $\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta_\theta(\vartheta^s\circ\mathbf{Y})\right| \le 2\frac{\rho^s}{1-\rho}$. Since $\rho \in (0,1)$,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{s=0}^{T-1}\left(\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta_\theta(\vartheta^s\circ\mathbf{Y})\right) = 0 \quad \mathbb{P}-\text{a.s.}\ .$$

By Lemma A.4

$$\mathbb{E}\left[\delta_\theta(\mathbf{Y})\right] \le \frac{2}{(1-\rho)} + \mathbb{E}\left[\left|\log\sigma_+ b_+(\mathbf{Y}_1)\right| + \left|\log\sigma_- b_-(\mathbf{Y}_1)\right|\right]\ ,$$

and the rhs is finite under assumption A4(b). By A4(a), the ergodic theorem, see [Billingsley, 1995, Theorem 24.1, p.314], concludes the proof.

(iii) Since Θ is compact, (A.27) holds if for any $\varepsilon > 0$, any $\theta' \in \Theta$, there exists $\eta > 0$ such that

$$\lim_{T\to+\infty}\sup_{\{\theta;|\theta-\theta'|<\eta\}\cap\Theta}\left|T^{-1}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) - T^{-1}\ell^{\chi,0}_{\theta',T}(\mathbf{Y})\right| \le \varepsilon\ , \quad \mathbb{P}-\text{a.s.} \tag{A.28}$$

Let $\varepsilon > 0$ and $\theta' \in \Theta$. Choose $\eta > 0$ such that

$$\mathbb{E}\left[\sup_{\{\theta\in\Theta;|\theta-\theta'|<\eta\}}\left|\delta_\theta(\mathbf{Y}) - \delta_{\theta'}(\mathbf{Y})\right|\right] \le \varepsilon\ ; \tag{A.29}$$

such an η exists by Lemma A.5. By (A.20), we have, for any $\theta \in \Theta$ such that $|\theta - \theta'| < \eta$

$$\left|\frac{1}{T}\ell^{\chi,0}_{\theta,T}(\mathbf{Y}) - \frac{1}{T}\ell^{\chi,0}_{\theta',T}(\mathbf{Y})\right| \leq \frac{1}{T}\sum_{s=0}^{T-1}\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta^{\chi,0}_{\theta',s}(\mathbf{Y})\right| . \tag{A.30}$$

In addition, by Lemma A.3(ii)

$$\begin{aligned}
&\sum_{s=0}^{T-1}\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta^{\chi,0}_{\theta',s}(\mathbf{Y})\right| \\
&\quad\leq 2\sum_{s=0}^{T-1}\sup_{\theta\in\Theta}\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta_\theta(\vartheta^s\circ\mathbf{Y})\right| + \sum_{s=0}^{T-1}|\delta_\theta(\vartheta^s\circ\mathbf{Y}) - \delta_{\theta'}(\vartheta^s\circ\mathbf{Y})| \\
&\qquad\qquad\leq \frac{4}{(1-\rho)^2} + \sum_{s=0}^{T-1}\Xi(\vartheta^s\circ\mathbf{Y})
\end{aligned}$$

where $\Xi(\mathbf{y}) \stackrel{\text{def}}{=} \sup_{\{\theta\in\Theta;|\theta-\theta'|<\eta\}}|\delta_\theta(\mathbf{y} - \delta_{\theta'}(\mathbf{y})|$. This implies that

$$\lim_{T\to+\infty}\sup_{\{\theta\in\Theta;|\theta-\theta'|<\eta\}}\frac{1}{T}\sum_{s=0}^{T-1}\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta^{\chi,0}_{\theta',s}(\mathbf{Y})\right| \leq \lim_{T\to+\infty}\frac{1}{T}\sum_{s=0}^{T-1}\Xi(\vartheta^s\circ\mathbf{Y}) .$$

Under A4, the ergodic theorem implies that the rhs converges $\mathbb{P}-$a.s. to $\mathbb{E}[\Xi(\mathbf{Y})]$, see [Billingsley, 1995, p.314]. Then, using again (A.29),

$$\lim_{T\to+\infty}\sup_{\{\theta\in\Theta;|\theta-\theta'|<\eta\}}\frac{1}{T}\sum_{s=0}^{T-1}\left|\delta^{\chi,0}_{\theta,s}(\mathbf{Y}) - \delta^{\chi,0}_{\theta',s}(\mathbf{Y})\right| \leq \varepsilon , \quad \mathbb{P}-\text{a.s.}$$

Then, (A.28) holds and this concludes the proof. □

A.2.4 Limit of the normalized score

This section is devoted to the proof of the $\mathbb{P}-$a.s. convergence of the normalized score $T^{-1}\nabla_\theta\ell^{\chi,0}_{\theta,T}(\mathbf{Y})$ to $\nabla_\theta\ell(\theta)$. This result is established under additional assumptions on the model.

S1 (a) For any $y \in \mathbb{Y}$ and for all $(x, x') \in \mathbb{X}^2$, $\theta \mapsto g_\theta(x, y)$ and $\theta \mapsto m_\theta(x, x')$ are continuously differentiable on Θ.

(b) We assume that $\mathbb{E}[\phi(\mathbf{Y}_0)] < +\infty$ where

$$\phi(y) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta}\sup_{(x,x')\in\mathbb{X}^2}\left|\nabla_\theta\log m_\theta(x, x') + \nabla_\theta\log g_\theta(x', y)\right| . \tag{A.31}$$

Lemma A.6. *Assume S1. For any initial distribution χ, any integers $s, r \geq 0$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, the function $\theta \mapsto \ell_{\theta,s}^{\chi,s-r}(\mathbf{y})$ is continuously differentiable on Θ and*

$$\nabla_\theta \ell_{\theta,s}^{\chi,s-r}(\mathbf{y}) = \sum_{u=s-r}^{s} \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_\theta, \mathbf{y}) \; ,$$

where Υ_θ is the function defined on $\mathbb{X}^2 \times \mathbb{Y}$ by

$$\Upsilon_\theta : (x, x', y) \mapsto \nabla_\theta \log \left\{ m_\theta(x, x') g_\theta(x', y) \right\} \; .$$

Proof. Under S1, the dominated convergence theorem implies that the function $\theta \mapsto \ell_{\theta,s}^{\chi,s-r}(\mathbf{y})$ is continuously differentiable and its derivative is obtained by permutation of the gradient and integral operators. □

Lemma A.7. *Assume A2 and S1.*

(i) There exists a function $\xi : \mathbb{Y}^{\mathbb{Z}} \to \mathbb{R}_+$ such that for any $s \geq 0$ and any $r, r' \geq s$, any initial distribution χ, χ' on $\mathbb{X}$ and any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$,

$$\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \right| \leq \frac{16 \rho^{-1/4}}{1-\rho} \rho^{(r' \wedge r)/4} \; \xi(\mathbf{y}) \; ,$$

where

$$\xi(\mathbf{y}) \stackrel{\text{def}}{=} \sum_{u \in \mathbb{Z}} \phi(\mathbf{y}_u) \rho^{|u|/4} \; . \tag{A.32}$$

(ii) For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ satisfying $\xi(\mathbf{y}) < +\infty$, the function $\theta \mapsto \delta_\theta(\mathbf{y})$ given by Lemma A.3(ii) *is continuously differentiable on Θ; and, for any $\theta \in \Theta$, any initial distribution χ and any integers $r \geq s \geq 0$,*

$$\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_\theta(\mathbf{y} \circ \vartheta^s) \right| \leq \frac{16 \rho^{-1/4}}{1-\rho} \rho^{r/4} \; \xi(\mathbf{y}) \; .$$

Proof. (i) By definition of $\delta_{\theta,s}^{\chi,s-r}(\mathbf{y})$, see (A.18) and Lemma A.6,

$$\begin{aligned}
&\nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \\
&\quad = \nabla_\theta \ell_{\theta,s+1}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \ell_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \ell_{\theta,s+1}^{\chi',s-r'}(\mathbf{y}) + \nabla_\theta \ell_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \\
&\quad = \sum_{u=s-r}^{s} \left(\Phi_{\theta,u,s+1}^{\chi,s-r-1}(\Upsilon_\theta, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi,s-r-1}(\Upsilon_\theta, \mathbf{y}) \right) \\
&\quad - \sum_{u=s-r'}^{s} \left(\Phi_{\theta,u,s+1}^{\chi',s-r'-1}(\Upsilon_\theta, \mathbf{y}) - \Phi_{\theta,u,s}^{\chi',s-r'-1}(\Upsilon_\theta, \mathbf{y}) \right) \\
&\quad + \Phi_{\theta,s+1,s+1}^{\chi,s-r-1}(\Upsilon_\theta, \mathbf{y}) - \Phi_{\theta,s+1,s+1}^{\chi',s-r'-1}(\Upsilon_\theta, \mathbf{y}) \; .
\end{aligned}$$

We can assume without loss of generality that $r' \leq r$ so that

$$\begin{aligned}
&\nabla_\theta \delta^{\chi,s-r}_{\theta,s}(\mathbf{y}) - \nabla_\theta \delta^{\chi',s-r'}_{\theta,s}(\mathbf{y}) \\
&= \sum_{u=s-r}^{s-r'-1} \left\{ \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right\} + \Phi^{\chi,s-r-1}_{\theta,s+1,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,s+1,s+1}(\Upsilon_\theta, \mathbf{y}) \\
&+ \sum_{u=s-r'}^{s} \left\{ \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) + \Phi^{\chi',s-r'-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right\} .
\end{aligned}$$

Under A2 and S1, Remark A.2.2 can be applied and for any $s - r \leq u \leq s - r' - 1$,

$$\left| \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| \leq 2\rho^{s-u} \phi(\mathbf{y}_u),$$

where $\phi_u(\mathbf{y})$ is defined in (A.31). Similarly, by Remark A.2.2

$$\left| \Phi^{\chi,s-r-1}_{\theta,s+1,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,s+1,s+1}(\Upsilon_\theta, \mathbf{y}) \right| \leq 2\rho^{r'+1} \phi(\mathbf{y}_{s+1}) .$$

For any $s - r' \leq u \leq s$, by Remark A.2.2,

$$\begin{aligned}
&\left| \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) + \Phi^{\chi',s-r'-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| \\
&\leq \left| \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) \right| + \left| \Phi^{\chi',s-r'-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| \\
&\leq 4\rho^{u+r'-s} \phi(\mathbf{y}_u)
\end{aligned}$$

and by Remark A.2.2,

$$\begin{aligned}
&\left| \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) + \Phi^{\chi',s-r'-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| \\
&\leq \left| \Phi^{\chi,s-r-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi,s-r-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| + \left| \Phi^{\chi',s-r'-1}_{\theta,u,s+1}(\Upsilon_\theta, \mathbf{y}) - \Phi^{\chi',s-r'-1}_{\theta,u,s}(\Upsilon_\theta, \mathbf{y}) \right| \\
&\leq 4\rho^{s-u} \phi(\mathbf{y}_u) .
\end{aligned}$$

Hence,

$$\left| \nabla_\theta \delta^{\chi,s-r}_{\theta,s}(\mathbf{y}) - \nabla_\theta \delta^{\chi',s-r'}_{\theta,s}(\mathbf{y}) \right| \leq 2 \sum_{u=s-r}^{s-r'-1} \rho^{s-u} \phi(\mathbf{y}_u) + 4 \sum_{u=s-r'}^{s+1} \left(\rho^{u+r'-s} \wedge \rho^{s-u} \right) \phi(\mathbf{y}_u) .$$

Furthermore,

$$\begin{aligned}
&\sum_{u=s-r'}^{s+1} \phi(\mathbf{y}_u) \left(\rho^{u+r'-s} \wedge \rho^{s-u}\right) \\
&\qquad \le \sum_{s-r'\le u\le \lfloor s-r'/2\rfloor} \rho^{s-u}\phi(\mathbf{y}_u) + \sum_{u\ge \lfloor s-r'/2\rfloor} \rho^{u+r'-s}\phi(\mathbf{y}_u) \\
&\qquad \le \rho^{r'/2} \sum_{u\in\mathbb{Z}} \phi(\mathbf{y}_u)\rho^{|u|/4} \ \ldots \\
&\qquad \times \left(\sum_{u\le \lfloor s-r'/2\rfloor} \rho^{s-u-r'/2-|u|/4} + \sum_{\lfloor s-r'/2\rfloor+1\le u\le s+1} \rho^{u+r'/2-s-|u|/4} \right) \\
&\qquad \le 2\frac{\rho^{(r'-1)/4}}{1-\rho} \sum_{u\in\mathbb{Z}} \phi(\mathbf{y}_u)\rho^{|u|/4} \,,
\end{aligned}$$

where we used that $\sup_{s-r'\le u\le \lfloor s-r'/2\rfloor} |u| \le r'$ and $\sup_{\lfloor s-r'/2\rfloor+1\le u\le s+1} |u| \le r'+1$. Moreover, upon noting that $-u/2+(s+1)/2 \le s-u-r'/2$ when $u \le s-r'-1$,

$$\begin{aligned}
\sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u)\rho^{s-u} &\le \rho^{r'/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u)\rho^{s-u-r'/2} \\
&\le \rho^{r'/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u)\rho^{-u/2+(s+1)/2} \\
&\le \rho^{r'/2}\rho^{(s+1)/2} \sum_{u=s-r}^{s-r'-1} \phi(\mathbf{y}_u)\rho^{|u|/2} \,,
\end{aligned}$$

where we used that $s-r'-1 \le 0$ in the last inequality.

Hence,

$$\sup_{\theta\in\Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_{\theta,s}^{\chi',s-r'}(\mathbf{y}) \right| \le \frac{16}{1-\rho}\rho^{(r'-1)/4} \sum_{u\in\mathbb{Z}} \phi(\mathbf{y}_u)\rho^{|u|/4} \,. \tag{A.33}$$

(ii) Let $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$. Then for any $u \in \mathbb{Z}$, $\phi(\mathbf{y}_u) < +\infty$. By Lemma A.6 and Eq. (A.18), the functions $\{\theta \mapsto \delta_{\theta,0}^{\chi,-r}(\mathbf{y})\}_{r\ge 0}$ are C^1 functions on Θ. By (i), there exists a function $\theta \mapsto \tilde{\delta}_\theta(\mathbf{y})$ such that

$$\lim_{r\to+\infty} \sup_{\theta\in\Theta} \left| \nabla_\theta \delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \tilde{\delta}_\theta(\mathbf{y}) \right| = 0 \,.$$

Furthermore, by Lemma A.3,

$$\lim_{r\to+\infty} \sup_{\theta\in\Theta} \left| \delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \delta_\theta(\mathbf{y}) \right| = 0 \,.$$

Then, $\theta \mapsto \delta_\theta(\mathbf{y})$ is C^1 on Θ and for any $\theta \in \Theta$, $\tilde{\delta}_\theta(\mathbf{y}) = \nabla_\theta \delta_\theta(\mathbf{y})$.

We thus proved that for any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$ and for any initial distribution χ,

$$\lim_{r \to +\infty} \sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,0}^{\chi,-r}(\mathbf{y}) - \nabla_\theta \delta_\theta(\mathbf{y}) \right| = 0 \; . \tag{A.34}$$

Observe that by definition, $\nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) = \nabla_\theta \delta_{\theta,0}^{\chi,-r}(\vartheta^s \circ \mathbf{y})$. This property, combined with Lemma A.7(i), yields

$$\begin{aligned} &\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \\ &\qquad \leq \frac{16\rho^{-1/4}}{1-\rho} \rho^{r/4} \, \xi(\mathbf{y}) + \sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,0}^{\chi,-r'}(\vartheta^s \circ \mathbf{y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{y}) \right| \; . \end{aligned}$$

Since $\xi(\vartheta^s \circ \mathbf{y}) < +\infty$, when $r' \to +\infty$, the second term tends to zero by (A.34) - for fixed $\mathbf{y}, s$ and χ -. This concludes the proof. □

Lemma A.8. *(i) Assume S1. For any $\mathbf{y} \in \mathbb{Y}^{\mathbb{Z}}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, for any integers $r, s \geq 0$,*

$$\sup_{\theta \in \Theta} \left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) \right| \leq 2 \sum_{u=s-r}^{s+1} \phi(\mathbf{y}_u) \; .$$

(ii) Assume A2 and S1. Then, for any $y \in \mathbb{Y}^{\mathbb{Z}}$ such that $\xi(\mathbf{y}) < +\infty$ and for any $r \geq 0$,

$$\sup_{\theta \in \Theta} |\nabla_\theta \delta_\theta(\mathbf{y})| \leq 2 \sum_{u=-r}^{1} \phi(\mathbf{y}_u) + \frac{16\rho^{-1/4}}{1-\rho} \xi(\mathbf{y}) \rho^{r/4} \; ,$$

where $\xi(\mathbf{y})$ is defined in Lemma A.7.

Proof. (i) By (A.18) and Lemma A.6,

$$\begin{aligned} &\left| \nabla_\theta \delta_{\theta,s}^{\chi,s-r}(\mathbf{y}) \right| = \left| \nabla_\theta \ell_{\theta,s+1}^{\chi,s-r}(\mathbf{y}) - \nabla_\theta \ell_{\theta,s}^{\chi,s-r}(\mathbf{y}) \right| \\ &\leq 2 \sum_{u=s-r}^{s+1} \left| \frac{\int \chi(\mathrm{d}x_{s-r}) L_{\theta,s-r:u-1}(x_{s-r}, \mathrm{d}x_u) \nabla_\theta \log \left[m_\theta(x_{u-1}, x_u) g_\theta(x_u, \mathbf{y}_u) \right] L_{\theta,u:s-1}(x_u, \mathbb{X})}{\int \chi(\mathrm{d}x_{s-r}) L_{\theta,s-r:s-1}(x_{s-r}, \mathbb{X})} \right| \; . \end{aligned}$$

The proof is concluded upon noting that for any $s - r \leq u \leq s + 1$,

$$\left| \frac{\int \chi(\mathrm{d}x_{s-r}) g_\theta(x_{s-r}, y_{s-r}) L_{\theta,s-r:u-1}(x_{s-r}, \mathrm{d}x_u) \nabla_\theta \log g_\theta(x_u, \mathbf{y}_u) L_{\theta,u:s-1}(x_u, \mathbb{X})}{\int \chi(\mathrm{d}x_{s-r}) g_\theta(x_{s-r}, \mathbf{y}_{s-r}) L_{\theta,s-r:s-1}(x_{s-r}, \mathbb{X})} \right|$$

is upper bounded by $\phi(\mathbf{y}_u)$.

(ii) is a consequence of Lemma A.7(ii) and Lemma A.8(i). □

Theorem A.2. *Assume A2, A4(a) and S1.*

(i) For any $T \geq 0$ and any distribution χ on $\mathbb{X}$, the functions $\theta \mapsto \ell_{\theta,T}^{\chi,0}(\mathbf{Y})$ and $\theta \mapsto \ell(\theta)$ are continuously differentiable $\mathbb{P}-$a.s.

(ii) For any initial distribution χ on $(\mathbb{X}, \mathcal{X})$,

$$\frac{1}{T}\nabla_\theta \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) \underset{T\to+\infty}{\longrightarrow} \nabla_\theta \ell(\theta) \quad \mathbb{P}-\text{a.s.} \tag{A.35}$$

Proof. By (A.20) and Lemma A.6, for any $\mathbf{y}$ such that $\phi(\mathbf{y}_u) < +\infty$ for any $u \in \mathbb{Z}$, $\ell_{\theta,T}^{\chi,0}(\mathbf{y})$ and $\delta_{\theta,s}^{\chi,0}(\mathbf{y})$ are continuously differentiable and (A.20) implies

$$\nabla_\theta \ell_{\theta,T}^{\chi,0}(\mathbf{y}) = \sum_{s=0}^{T-1} \nabla_\theta \delta_{\theta,s}^{\chi,0}(\mathbf{y}) \; .$$

This decomposition leads to

$$\frac{1}{T}\nabla_\theta \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) = \frac{1}{T}\sum_{s=0}^{T-1}\Big(\nabla_\theta \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y})\Big) + \frac{1}{T}\sum_{s=0}^{T-1} \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y}) \; . \tag{A.36}$$

Consider the first term of the rhs of (A.36). Since $\mathbf{Y}$ is a stationary process, assumption S1(b) implies that $\mathbb{E}\left[\xi(\mathbf{Y})\right] < +\infty$, where ξ is defined by (A.32). Then, $\xi(\mathbf{Y}) < +\infty \quad \mathbb{P}-$a.s. and by Lemma A.7(ii), for any $0 \leq s \leq T-1$,

$$\left|\nabla_\theta \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y})\right| \leq \xi(\mathbf{Y}) \frac{16\rho^{-1/4}}{1-\rho}\rho^{s/4} \; .$$

Therefore

$$\frac{1}{T}\sum_{s=0}^{T-1}\left|\nabla_\theta \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y})\right| \leq \frac{1}{T}\xi(\mathbf{Y}) \frac{16\rho^{-1/4}}{1-\rho}\frac{1}{1-\rho^{1/4}} \; ,$$

and

$$\lim_{T\to\infty}\frac{1}{T}\sum_{s=0}^{T-1}\Big(\nabla_\theta \delta_{\theta,s}^{\chi,0}(\mathbf{Y}) - \nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y})\Big) = 0 \; , \quad \mathbb{P}-\text{a.s.}$$

Finally, consider the second term of the rhs of (A.36). By Lemma A.8 (applied with $r = 1$), $\mathbb{E}\left[|\nabla_\theta \delta_\theta(\mathbf{Y})|\right] < +\infty$. Under H**??**, the ergodic theorem (see [Billingsley, 1995, Theorem 24.1, p.314]) states that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{s=0}^{T-1}\nabla_\theta \delta_\theta(\vartheta^s \circ \mathbf{Y}) = \mathbb{E}\left[\nabla_\theta \delta_\theta(\mathbf{Y})\right] \; , \quad \mathbb{P}-\text{a.s.}$$

Then, by (A.36) and the above discussion,

$$\lim_{T\to\infty}\frac{1}{T}\nabla_\theta \ell_{\theta,T}^{\chi,0}(\mathbf{Y}) = \mathbb{E}\left[\nabla_\theta \delta_\theta(\mathbf{Y})\right] \; , \quad \mathbb{P}-\text{a.s.}$$

By Lemma A.8, applied with $r = 0$,

$$\sup_{\theta\in\Theta} |\nabla_\theta \delta_\theta(\mathbf{Y})| \leq 2\left[\phi(Y_0) + \phi(Y_1)\right] + \xi(\mathbf{Y})\rho^{1/2} ,$$

and the rhs is integrable under the stated assumptions. Therefore, by the dominated convergence theorem, $\mathbb{E}\left[\nabla_\theta \delta_\theta(\mathbf{Y})\right] = \nabla_\theta \mathbb{E}\left[\delta_\theta(\mathbf{Y})\right] = \nabla_\theta \ell(\theta)$. This concludes the proof. □

A.3 Additional experiments

In this section, we provide additional plots for the applications studied in Section 6.3.

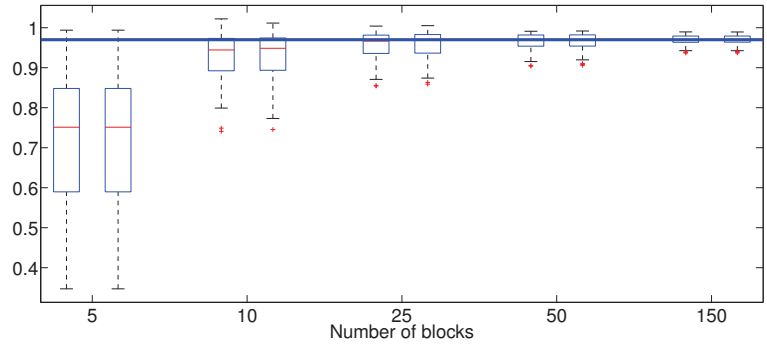
A.3.1 Linear Gaussian model

Figure A.1 illustrates the fact that the convergence properties of the BOEM do not depend on the initial distribution χ used in each block. Data are sampled using $\phi = 0.97$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$. Figure A.1 displays the estimation of ϕ by the averaged BOEM algorithm with $\tau_n \sim n$ and $\tau_n \sim n^{1.5}$, over 100 independent Monte Carlo runs as a function of the number of blocks. We consider first the case when χ is the stationary distribution of the hidden process i.e. $\chi \equiv \mathcal{N}(0, (1-\phi^2)^{-1}\sigma_u^2)$, and the case when χ is the filtering distribution obtained at the end of the previous block, computed with the Kalman filter. The estimation error is similar for both initialization schemes, even when ϕ is close to 1 and for any choice of $\{\tau_n\}_{n\geq 1}$.

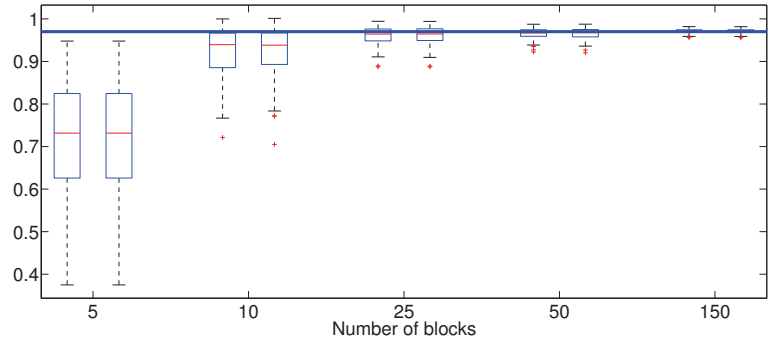
The theoretical analysis of BOEM says that a sufficient condition for convergence is the increasing size of the blocks. On Figure A.2, we compare different strategies for the definition of $\tau_n \stackrel{\text{def}}{=} T_n - T_{n-1}$. A slowly increasing sequence $\{\tau_n\}_{n\geq 0}$ is compared to different strategies using the same number of observations within each block. We consider the Linear Gaussian model:

$$X_{t+1} = \phi X_t + \sigma_u U_t , \qquad Y_t = X_t + \sigma_v V_t ,$$

where $X_0 \sim \mathcal{N}\left(0, \sigma_u^2(1-\phi^2)^{-1}\right)$, $\{U_t\}_{t\geq 0}, \{V_t\}_{t\geq 0}$ are i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.9$, $\sigma_u^2 = 0.6$ and $\sigma_v^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma_u^2 = 1$ and $\sigma_v^2 = 2$. Figure A.2 shows the estimation of ϕ over 100 independent Monte Carlo runs (same conclusions could be drawn for σ_u^2 and σ_v^2). For each choice of $\{\tau_n\}_{n\geq 0}$, the median and first and last quartiles of the estimation are represented as a function of the number of observations.



(a) $\tau_n \sim n$



(b) $\tau_n \sim n^{1.5}$

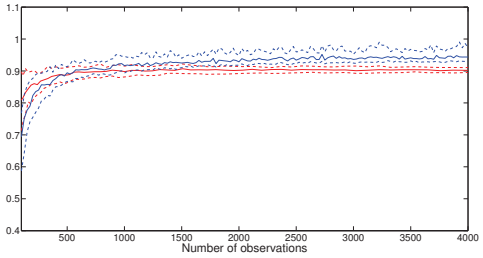
Figure A.1: Estimation of ϕ after $5, 10, 25, 50$ and 150 blocks, with two different initialization schemes: the stationary distribution (left) and the filtering distribution at the end of the previous block (right). The boxplots are computed with 100 Monte Carlo runs.

We observe that BOEM does not converge when the block size sequence is constant and small: as shown in Figure A.2, if the number of observations is too small ($\tau_n = 25$), the algorithm is a poor approximation of the *limiting EM* recursion and does not converge. With greater block sizes ($\tau_n = 100$ or $\tau_n = 350$), the algorithm converges but the convergence is slower because it is initialized far from the true value and many observations are needed to get several estimations. BOEM with slowly increasing block sizes has a better behavior since many estimations are produced at the beginning and, once the estimates are closer to the true value, the bigger block sizes reduce the variance of the estimation.

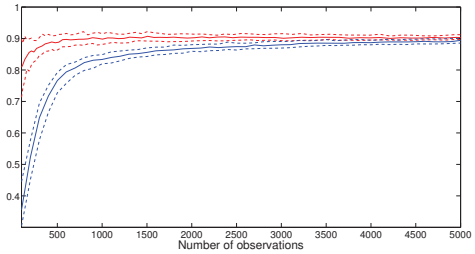
Moreover, our convergence rates are given up to a multiplicative constant : the theory says that $\sum_n \tau_n^{-\gamma/2} < \infty$ where γ is related to the ergodic behavior of the HMM (see assumption A5).

Even if the sequence is chosen to increase at a polynomial rate, we can have $\tau_n \sim c\ n^\alpha$ ($\alpha > 1$) with a constant c such that the first blocks are

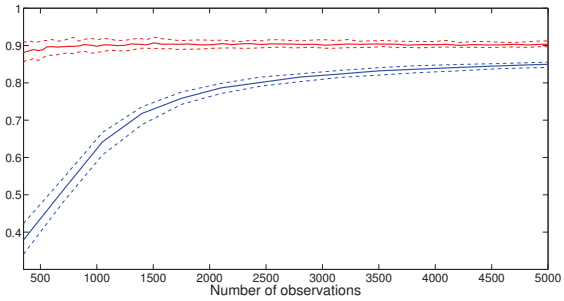
quite small to allow a sufficiently large number of updates of the parameters $\{\theta_n, n \geq 1\}$. During a (deterministic) "burn-in" period, the first blocks can even be of a fixed length before beginning the "increasing" procedure.



(a) $\tau_n = n^{1.1}$ (red) and $\tau_n = 25$ (blue).



(b) $\tau_n = n^{1.1}$ (red) and $\tau_n = 100$ (blue).



(c) $\tau_n = n^{1.1}$ (red) and $\tau_n = 350$ (blue).

Figure A.2: Estimation of ϕ with different block size schemes: the median (bold line) and the first and last quartiles (dotted line) are shown for $\tau_n = n^{1.1}$ (red), $\tau_n = 100$ (black) and $\tau_n = 350$ (purple). The quantities are computed with 100 Monte Carlo runs.

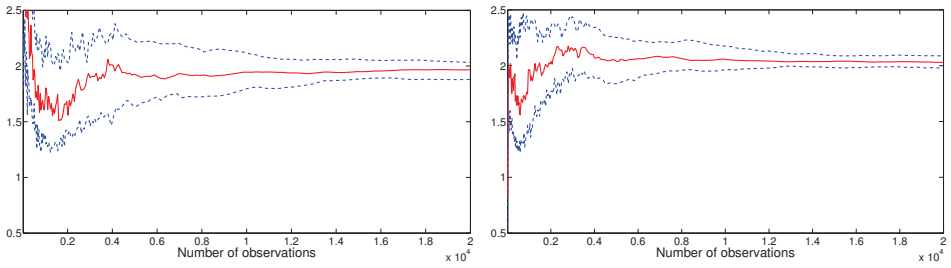
A.3.2 Finite state-space HMM

Observations are sampled using $d = 6$, $v = 0.5$, $x_i = i$, $\forall i \in \{1, \ldots, d\}$ and the true transition matrix is given by

$$m = \begin{pmatrix} 0.5 & 0.05 & 0.1 & 0.15 & 0.15 & 0.05 \\ 0.2 & 0.35 & 0.1 & 0.15 & 0.05 & 0.15 \\ 0.1 & 0.1 & 0.6 & 0.05 & 0.05 & 0.1 \\ 0.02 & 0.03 & 0.1 & 0.7 & 0.1 & 0.05 \\ 0.1 & 0.05 & 0.13 & 0.02 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.13 & 0.12 & 0.1 & 0.45 \end{pmatrix}.$$

Comparison to an online EM based procedure

In this case, we want to estimate the states $\{x_1, \ldots, x_d\}$. All the runs are started from $v = 2$ and from the initial states $\{-1; 0; .5; 2; 3; 4\}$. The experiment is the same as the one in Section 6.3.2. The averaged BOEM is compared to an online EM procedure (see [Cappé, 2011a]) combined with Polyak-Ruppert averaging (see [Polyak, 1990]). This online EM based algorithm follows a stochastic approximation update and depends on a step-size sequence $\{\gamma_n\}_{n\geq 0}$ which is chosen in the same way as in Section 6.3.2. Figure A.3 displays the empirical median and first and last quartiles for the estimation of x_2 with both averaged algorithms as a function of the number of observations. These estimates are obtained over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. Both algorithms converge to the true value $x_2 = 2$ and these plots confirm the similar behavior of BOEM and the online EM of [Cappé, 2011a].

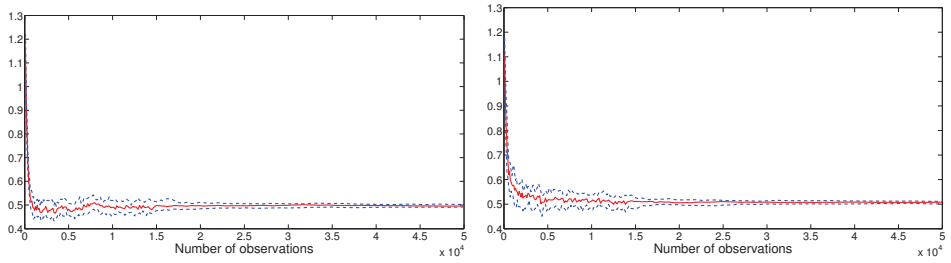


(a) Estimation of x_2 with averaged BOEM. (b) Estimation of x_2 with averaged OEM.

Figure A.3: Estimation of x_2 using the averaged online EM and averaged BOEM. Each plot displays the empirical median (bold line) and the first and last quartiles (dotted lines) over 100 independent Monte Carlo runs with $\tau_n = n^{1.1}$ and $\gamma_n = n^{-0.53}$. The first ten observations are omitted for a better visibility.

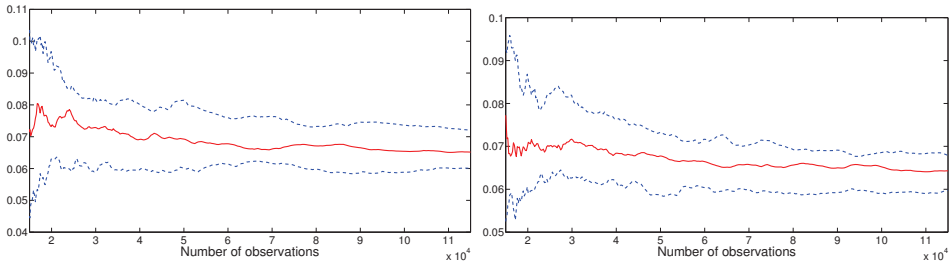
Comparison to a recursive maximum likelihood procedure

In the numerical applications below, we give supplementary graphs to compare the convergence of the averaged BOEM with the convergence of the Polyak-Ruppert averaged RML procedure. The experiment is the same as the one in Section 6.3.2. Figure A.4 and A.5 displays the empirical median and first and last quartiles of the estimation of v and $m(1, 2)$ over 100 independent Monte Carlo runs. Both algorithms have a similar behavior for the estimation of these parameters.



(a) Averaged BOEM. (b) Averaged RML.

Figure A.4: Empirical median (bold line) and first and last quartiles (dotted line) for the estimation of v using the averaged RML algorithm (right) and the averaged BOEM algorithm (left). The true values is $v = 0.5$ and the averaging procedure is starter after 10000 observations. The first 10000 observations are not displayed for a better clarity.



(a) Averaged BOEM. (b) Averaged RML.

Figure A.5: Empirical median (bold line) and first and last quartiles (dotted line) for the estimation of $m(1,2)$ using the averaged RML algorithm (right) and the averaged BOEM algorithm (left). The true values is $m(1,2) = 0.05$ and the averaging procedure is starter after 10000 observations. The first 10000 observations are not displayed for a better clarity.

A.3.3 Stochastic volatility model

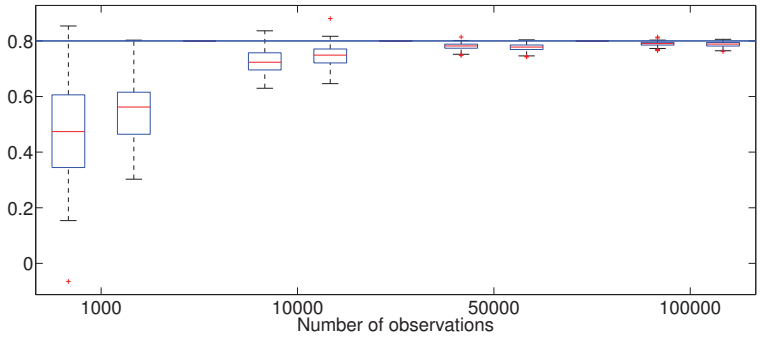
Consider the following stochastic volatility model:

$$X_{t+1} = \phi X_t + \sigma U_t \,, \qquad Y_t = \beta \mathrm{e}^{\frac{X_t}{2}} V_t \,,$$

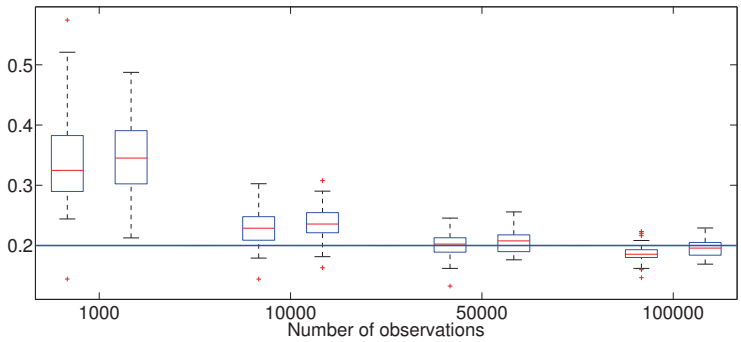
where $X_0 \sim \mathcal{N}\left(0, (1-\phi^2)^{-1}\sigma^2\right)$ and $(U_t)_{t\geq 0}$ and $(V_t)_{t\geq 0}$ are two sequences of i.i.d. standard Gaussian r.v., independent from X_0. Data are sampled using $\phi = 0.8$, $\sigma^2 = 0.2$ and $\beta^2 = 1$. All runs are started with $\phi = 0.1$, $\sigma^2 = 0.6$ and $\beta^2 = 2$.

In this model, the smoothed sufficient statistics $\{\bar{S}_{\tau_n}^{\chi,T_{n-1}}(\theta_{n-1}, \mathbf{Y})\}_{n\geq 1}$ can not be computed explicitly. We thus propose to replace the exact computation by a Monte Carlo approximation based on particle filtering. The

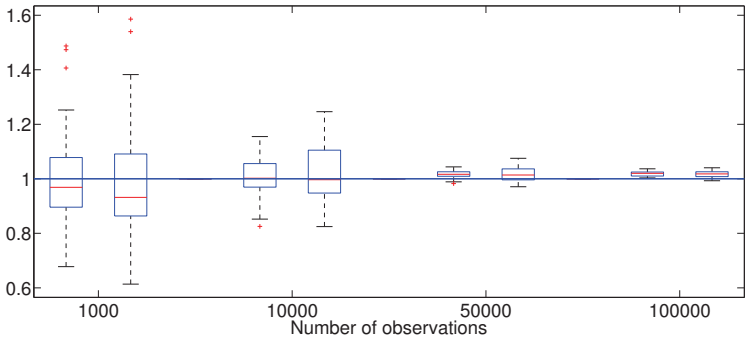
performance of the Stochastic BOEM is compared to the online EM algorithm given in [Cappé, 2011a] (see also [Del Moral *et al.*, 2010b]). To our best knowledge, there do not exist results on the asymptotic behavior of the algorithms by [Cappé, 2011a, Del Moral *et al.*, 2010b]; these algorithms rely on many approximations that make the proof quite difficult (some insights on the asymptotic behavior are given in [Cappé, 2011a]). Despite there are no results in the literature on the rate of convergence of the Online EM algorithm by [Cappé, 2011a] we choose the step size γ_n in [Cappé, 2011a] and the block size τ_n s.t. $\gamma_n = n^{-0.6}$ and $\tau_n \propto n^{3/2}$ (see Section 6.3.2 for a discussion on this choice). 50 particles are used for the approximation of the filtering distribution by Particle filtering. We report in Figure A.6, the boxplots for the estimation of the three parameters (β, ϕ, σ^2) for the Polyak-Ruppert [Polyak, 1990] averaged Online EM and the averaged BOEM. Both averaged versions are started after 20000 observations. Figure A.6 displays the estimation of ϕ, σ^2 and β^2. This figure shows that both algorithms have the same behavior. Similar conclusions are obtained by considering other true values for ϕ (such as $\phi = 0.95$). Therefore, the intuition is that online EM and Stochastic BOEM have the same asymptotic behavior. The main advantage of the second approach is that it relies on approximations which can be controlled in such a way that we are able to show that the limiting points of the particle version of the Stochastic BOEM algorithms are the stationary points of the limiting normalized log-likelihood of the observations.



(a) Estimation of ϕ.



(b) Estimation of σ^2.



(c) Estimation of β^2.

Figure A.6: Estimation of ϕ, σ^2 and β^2 using the averaged online EM algorithm (left) and the averaged BOEM (right), after $n = \{1000, 10k, 50k, 100k\}$ observations. The true value of ϕ is 0.8.

Bibliographie

[Adamczak et Bednorz, 2012] Adamczak, R. et Bednorz, W. (2012). Exponential concentration inequalities for additive functionals of markov chains. arXiv :1201.3569v1.

[Adams et Fournier, 2003] Adams, R. et Fournier, J. (2003). *Sobolev Spaces*. Numéro vol. 140 de Pure and Applied Mathematics. Academic Press.

[Bailey *et al.*, 2006a] Bailey, T., Nieto, J., Guivant, J., Stevens, M. et Nebot, E. (2006a). Consistency of the EKF-SLAM algorithm. *In IEEE International Conference on Intelligent Robots and Systems*, pages 3562–3568.

[Bailey *et al.*, 2006b] Bailey, T., Nieto, J. et Nebot, E. (2006b). Consistency of the FastSLAM algorithm. *IEEE Intl. Conf. on Robotics and Automation*.

[Baum et Petrie, 1966] Baum, L. E. et Petrie, T. P. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37:1554–1563.

[Benaim, 1999] Benaim, M. (1999). Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33:1–68.

[Bickel *et al.*, 1998] Bickel, P. J., Ritov, Y. et Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26:1614–1635.

[Billingsley, 1995] Billingsley, P. (1995). *Probability and Measure*. Wiley, New York, 3rd édition.

[Briers *et al.*, 2010] Briers, M., Doucet, A. et Maskell, S. (2010). Smoothing algorithms for state-space models. *Annals Institute Statistical Mathematics*, 62(1):61–89.

[Bröcker et Lander, 1975] Bröcker, T. et Lander, L. (1975). *Differentiable Germs and Catastrophes*. London Mathematical Society Lecture Note Series. Cambridge University Press.

[Burgard *et al.*, 2005] Burgard, W., Fox, D. et Thrun, S. (2005). *Probabilistic robotics*. Cambridge, MA :MIT Press.

[Cappé, 2001] Cappé, O. (2001). Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods Appl.*, 7(1–2):81–92.

[Cappé, 2009] Cappé, O. (2009). Online sequential Monte Carlo EM algorithm. *In IEEE Workshop on Statistical Signal Processing (SSP)*.

[Cappé, 2011a] Cappé, O. (2011a). Online EM algorithm for hidden Markov models. *J. Comput. Graph. Statist.*, 46(2):728–749.

[Cappé, 2011b] Cappé, O. (2011b). Online Expectation Maximization. *In* Mengersen, K., Titterington, M. et Robert, C., éditeurs : *Mixtures : Estimation and Applications*. Wiley.

[Cappé et Moulines, 2009] Cappé, O. et Moulines, E. (2009). Online Expectation-Maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613.

[Cappé *et al.*, 2005] Cappé, O., Moulines, E. et Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

[Carlin *et al.*, 1992] Carlin, B., N.G., P. et D.S., S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association*, 87:493–500.

[Carroll et Hall, 1988] Carroll, R. et Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, pages 1184–1186.

[Carroll et Stefanski, 1990] Carroll, R. et Stefanski, L. (1990). Deconvolving kernel density estimators. *Statistics*, 21:169–184.

[Chen *et al.*, 1988] Chen, H., Gao, A. et Guo, L. (1988). Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stoch. Proc. Appl.*, 27:217–231.

[Churchill, 1992] Churchill, G. (1992). Hidden Markov chains and the analysis of genome structure. *Computers & Chemistry*, 16(2):107–115.

[Comte *et al.*, 2006] Comte, F., Rozenholc, Y. et Taupin, M.-L. (2006). Penalized contrast estimator for density deconvolution. *The Canadian Journal of Statistics*, 34:431–452.

[Comte et Taupin, 2007] Comte, F. et Taupin, M.-L. (2007). Nonparametric estimation of the regression function in an errors-in-variables model. *Statistica sinica*, 17(3):1065–1090.

[Davidson, 1994] Davidson, J. (1994). *Stochastic Limit Theory : An Introduction for Econometricians*. Oxford University Press.

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

[Del Moral et Doucet, 2003] Del Moral, P. et Doucet, A. (2003). On a class of genealogical and interacting Metropolis models. *Seminaire de Probabilites XXXVII, Lecture notes in Math.*, pages 415–446.

[Del Moral *et al.*, 2010a] Del Moral, P., Doucet, A. et Singh, S. (2010a). A Backward Particle Interpretation of Feynman-Kac Formulae. *ESAIM M2AN*, 44(5):947–975.

[Del Moral *et al.*, 2010b] Del Moral, P., Doucet, A. et Singh, S. (2010b). Forward smoothing using sequential Monte Carlo. Rapport technique, arXiv :1012.5390v1.

[Del Moral et Guionnet, 1998] Del Moral, P. et Guionnet, A. (1998). Large deviations for interacting particle systems : applications to non-linear filtering. *Stoch. Proc. App.*, 78:69–95.

[Del Moral et Guionnet, 2001] Del Moral, P. et Guionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. *Annales de l'Institut Henri Poincaré*, 37:155–194.

[Del Moral *et al.*, 2003] Del Moral, P., Ledoux, M. et Miclo, L. (2003). On contraction properties of Markov kernels. *Probab. Theory Related Fields*, 126(3):395–420.

[Delyon *et al.*, 1999] Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1).

[Dempster *et al.*, 1977] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion).

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

[Douc et Moulines, 2012] Douc, R. et Moulines, E. (2012). Asymptotic properties of the maximum likelihood estimation in misspecified hidden markov models. Rapport technique, arXiv :1110.0356.

[Douc *et al.*, 2011b] Douc, R., Moulines, E., Olsson, J. et Van Handel, R. (2011b). Consistency of the maximum likelihood estimator for general hidden Markov models. *ann. Statist.*, 39:474–513.

[Douc *et al.*, 2004a] Douc, R., Moulines, E. et Rosenthal, J. (2004a). Quantitative bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 14(4):1643–1665.

[Douc *et al.*, 2004b] Douc, R., Moulines, E. et Rydén, T. (2004b). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5):2254–2304.

[Doucet *et al.*, 2001] Doucet, A., De Freitas, N. et Gordon, N., éditeurs (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.

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

[Doucet *et al.*, 2011] Doucet, A., Poyiadjis, G. et Singh, S. (2011). Particle approximations of the score and observed information matrix in state-space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

[Doukhan *et al.*, 1995] Doukhan, P., Massart, P. et Rio, E. (1995). Invariance principle for absolutely regular processes. *Annales de l'Institut Henri Poincaré*, 31:393–427.

[Dubarry et Le Corff, 2011] Dubarry, C. et Le Corff, S. (2011). Fast computation of smoothed additive functionals in general state-space models. *In IEEE Workshop on Statistical Signal Processing (SSP)*.

[Dubarry et Le Corff, 2011] Dubarry, C. et Le Corff, S. (2011). Non-asymptotic deviation inequalities for smoothed additive functionals in non-linear state-space models. *Accepté pour publication dans Bernoulli, arXiv :1012.4183v1*.

[Dumont et Le Corff, 2012a] Dumont, T. et Le Corff, S. (2012a). Non-parametric estimation in hidden Markov models. Rapport technique.

[Dumont et Le Corff, 2012b] Dumont, T. et Le Corff, S. (2012b). Simultaneous localization and mapping problem in wireless sensor networks. Rapport technique.

[Durbin et Koopman, 2000] Durbin, J. et Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *J. Roy. Statist. Soc. B*, 62:3–29.

[Elliott et Krishnamurthy, 1999] Elliott, R. J. et Krishnamurthy, V. (1999). New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models. *IEEE Trans. Automat. Control*, 44(5).

[Evans et Gariepy, 1992] Evans, L. et Gariepy, R. (1992). *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press.

[Evennou et Marx, 2006] Evennou, F. et Marx, F. (2006). Advanced integration of WiFi and inertial navigation systems for indoor mobile positioning. *EURASIP J. Appl. Signal Process.*, 2006:164–164.

[Fan, 1991a] Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *The indian journal of statistics*, 53:97–110.

[Fan, 1991b] Fan, J. (1991b). On the optimal rates of convergence for non-parametric deconvolution problems. *Ann. Statist.*, 19:1257–1272.

[Fan et Truong, 1993] Fan, J. et Truong, Y. (1993). Nonparametric regression with errors in variables. *Ann. Statist.*, 21:1900–1925.

[Fearnhead *et al.*, 2010] Fearnhead, P., Wyncoll, D. et Tawn, J. (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika*, 97(2):447–464.

[Fort et Moulines, 2003] Fort, G. et Moulines, E. (2003). Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259.

[Friis, 1946] Friis, H. T. (1946). Noise figures of radio receivers. *In Proc. IRE*, volume 34, pages 419–423.

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

[Gorce *et al.*, 2007] Gorce, J. M., Jaffres-Runser, K. et Roche, G. D. L. (2007). Deterministic approach for fast simulations of indoor radio wave propagation.

[Gordon *et al.*, 1993] Gordon, N., Salmond, D. et Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, 140:107–113.

[Hall et Heyde, 1980] Hall, P. et Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York, London.

[Hull et White, 1987] Hull, J. et White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance*, 42:281–300.

[Hürzeler et Künsch, 1998] Hürzeler, M. et Künsch, H. R. (1998). Monte Carlo approximations for general state-space models. *J. Comput. Graph. Statist.*, 7:175–193.

[Jensen et Petersen, 1999] Jensen, J. L. et Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.*, 27:514–535.

[Juang et Rabiner, 1991] Juang, B. et Rabiner, L. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33:251–272.

[Julier et Uhlmann, 2001] Julier, S. et Uhlmann, J. (2001). A counter example to the theory of simultaneous localization and map building. *In IEEE international conference on robotics and automation*, volume 4, pages 4238–4243. Citeseer.

[Kalman et Bucy, 1961] Kalman, R. E. et Bucy, R. (1961). New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME, Series D*, 83(3):95–108.

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

[Kitagawa et Sato, 2001] Kitagawa, G. et Sato, S. (2001). Monte carlo smoothing and self-organizing state-space model. *In* Doucet, A., De Freitas, N. et Gordon, N., éditeurs : *Sequential Monte Carlo methods in Practice*. Springer.

[Kushner et Yin, 1997] Kushner, H. J. et Yin, G. G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer.

[Kushner et Yin, 2003] Kushner, H. J. et Yin, G. G. (2003). *Stochastic Approximation and recursive Algorithms and Applications*, volume 35. Springer.

[Lacour, 2008a] Lacour, C. (2008a). Adaptive estimation of the transition density of a particular hidden Markov chain. *Journal of Multivariate Analysis*, 99(5):787–814.

[Lacour, 2008b] Lacour, C. (2008b). Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications*, 118(2):232–260.

[Laruelle, 2011] Laruelle, S. (2011). *Analyse d'Algorithmes Stochastiques Appliqués à la Finance*. Thèse de doctorat, Université Pierre et Marie Curie.

[Laruelle et Pagès, 2012] Laruelle, S. et Pagès, G. (2012). Stochastic approximation with averaging innovation applied to finance. *Monte Carlo Methods and Applications*, 18:1–52.

[Le Corff et Fort, 2011a] Le Corff, S. et Fort, G. (2011a). Convergence of a particle-based approximation of the block online Expectation Maximization algorithm. *Accepté pour publication dans Transactions on Modeling and Computer Simulation, arXiv :1111.1307*.

[Le Corff et Fort, 2011b] Le Corff, S. et Fort, G. (2011b). Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Rapport technique, arXiv :1108.3968v1.

[Le Corff *et al.*, 2011a] Le Corff, S., Fort, G. et Moulines, E. (2011a). Block online EM for hidden Markov models with general state-space. *In 14th Applied Stochastic Models and Data Analysis International Conference*.

[Le Corff *et al.*, 2011b] Le Corff, S., Fort, G. et Moulines, E. (2011b). Online EM algorithm to solve the SLAM problem. *In IEEE Workshop on Statistical Signal Processing (SSP)*.

[Le Corff *et al.*, 2012] Le Corff, S., Fort, G. et Moulines, E. (2012). New online EM algorithms for general hidden Markov models. application to the slam problem. *In 10th international conference, LVA/ICA*.

[Le Gland et Mevel, 1997] Le Gland, F. et Mevel, L. (1997). Recursive estimation in HMMs. *In Proc. IEEE Conf. Decis. Control*, pages 3468–3473.

[Lee, 2000] Lee, J. (2000). *Introduction to Topological Manifolds*. Graduate Texts in Mathematics. Springer.

[Leroux, 1992] Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, 40:127–143.

[Mac Donald et Zucchini, 2009] Mac Donald, I. et Zucchini, W. (2009). *Hidden Markov Models for time series : an introduction using R.* CRC press.

[Mamon et Elliott, 2007] Mamon, R. et Elliott, R. (2007). *Hidden Markov Models in Finance*, volume 104 de *International Series in Operations Research & Management Science*. Springer, Berlin.

[Martinez-Cantin, 2008] Martinez-Cantin, R. (2008). *Active Map Learning for Robots : Insights into Statistical Consistency.* Thèse de doctorat, University of Zaragoza.

[Massart et Picard, 2007] Massart, P. et Picard, J. (2007). *Concentration inequalities and model selection : Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003.* Numéro vol. 1896 de Ecole d'Eté de Probabilités de Saint-Flour. Springer-Verlag.

[Meyn et Tweedie, 1993] Meyn, S. P. et Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, London.

[Mongillo et Denève, 2008] Mongillo, G. et Denève, S. (2008). Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716.

[Montemerlo *et al.*, 2003] Montemerlo, M., Thrun, S., Koller, D. et Wegbreit, B. (2003). FastSLAM 2.0 : An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. *In Proceedings of the Sixteenth IJCAI*, Mexico.

[Nickel et Potscher, 2001] Nickel, R. et Potscher, B. (2001). Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov and Sobolev type. *J. Theor. Probab.*, 20:177–199.

[Olsson *et al.*, 2008] Olsson, J., Cappe, O., Douc, R. et Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179.

[Pitt et Shephard, 1999] Pitt, M. K. et Shephard, N. (1999). Filtering via simulation : Auxiliary particle filters. *J. Am. Statist. Assoc.*, 94(446):590–599.

[Pólya et Szegő, 1976] Pólya, G. et Szegő, G. (1976). *Problems and Theorems in Analysis. Vol. II.* Springer.

[Polyak, 1990] Polyak, B. T. (1990). A new method of stochastic approximation type. *Autom. Remote Control*, 51:98–107.

[Polyak et Juditsky, 1992] Polyak, B. T. et Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855.

[Poyiadjis *et al.*, 2005] Poyiadjis, G., Doucet, A. et Singh, S. S. (2005). Particle methods for optimal filter derivative : application to parameter

estimation. *In Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages v/925–v/928.

[Rio, 1990] Rio, E. (1990). *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer.

[Robbins et Monro, 1951] Robbins, H. et Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407.

[Smart, 1980] Smart, D. (1980). *Fixed point theorems*. Cambridge University Press.

[Tadić, 2010] Tadić, V. B. (2010). Analyticity, convergence, and convergence rate of recursive maximum-likelihood estimation in hidden Markov models. *IEEE Trans. Inf. Theor.*, 56:6406–6432.

[Titterington, 1984] Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. *J. Roy. Statist. Soc. B*, 20(3):257–267.

[Van De Geer, 2000] Van De Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[West et Harrison, 1989] West, M. et Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer.

[Wu, 1983] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103.

[Zeitouni et Dembo, 1988] Zeitouni, O. et Dembo, A. (1988). Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE Trans. Inform. Theory*, 34(4):890–893.