UNIVERSITÉ PARIS-SUD XI

Préparée à l'École Doctorale de Mathématiques de la région Paris-Sud

Laboratoire de Mathématiques de la Faculté des Sciences d'Orsay

THÈSE DE DOCTORAT SUR TRAVAUX

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD XI

Spécialité : Mathématiques

par **Thierry Dumont**

Contributions à la localisation intra-muros. De la modélisation à la calibration théorique et pratique d'estimateurs.

Soutenue le 13 Décembre 2012 devant la commission d'examen :

M.	Ramon	**Van Handel**	Princeton University	Rapporteur
M.	François	**Le Gland**	INRIA Rennes	Rapporteur
Mme.	Élisabeth	**Gassiat**	Université Paris XI	Directrice de thèse
M.	Loïc	**Guillemard**	Id Services	Examinateur
M.	Pascal	**Massart**	Université Paris XI	Examinateur
M.	Éric	**Moulines**	Télécom ParisTech	Examinateur

Thierry Dumont
thierry.dumont@math.u-psud.fr
`www.math.u-psud.fr/~dumont`

Thèse préparée sous la direction d'Élisabeth Gassiat
au Département de Mathématiques d'Orsay
Laboratoire de mathématiques (UMR 8628), Bât 425
Université Paris-Sud 11
91405 Orsay Cedex

Thèse financée par l'entreprise ID Services
22/24 rue Jean Rostand
Parc club de l'université
91400 Orsay

Abstract

Foreshadowing the next big step in the field of navigation, indoor geolocation has been a very active field of research in the last few years. While geolocation entered the life of many individuals and professionals, particularly through assisted navigation systems on roads, needs to extend the applications inside the buildings are more and more present. However, existing systems face many more technical constraints than those encountered outside, including the chaotic propagation of electromagnetic waves in confined and inhomogeneous environments. In this manuscript, we propose a statistical approach to the problem of geolocation of a mobile device inside a building, using the WiFi surrounding waves. This manuscript focuses on two central issues: the determination of WiFi wave propagation maps inside a building and the construction of estimators of the mobile's positions using these propagation maps.

The statistical framework used in this thesis to answer these questions is that of hidden Markov models. We propose, in a parametric framework, an inference method for the online estimation of the propagation maps, on the basis of the informations reported by the mobile. In a nonparametric framework, we investigated the possibility of estimating the propagation maps considered as a single regular function on the environment that we wish to geolocate. Our results on the nonparametric estimation in hidden Markov models make it possible to produce estimators of the propagation functions whose consistency is established in a general framework. The last part of the manuscript deals with the estimation of the context tree in variable length hidden Markov models.

Key Words : Indoor localization, WiFi, hidden Markov models, statistical inference, nonparametric estimation, online estimation.

Résumé

Préfigurant la prochaine grande étape dans le domaine de la navigation, la géolocalisation *intra-muros* est un domaine de recherche très actif depuis quelques années. Alors que la géolocalisation est entrée dans le quotidien de nombreux professionnels et particuliers avec, notamment, le guidage routier assisté, les besoins d'étendre les applications à l'intérieur se font de plus en plus pressants. Cependant, les systèmes existants se heurtent à des contraintes techniques bien supérieures à celles rencontrées à l'extérieur, la faute, notamment, à la propagation chaotique des ondes électromagnétiques dans les environnements confinés et inhomogènes. Nous proposons dans ce manuscrit une approche statistique du problème de géolocalisation d'un mobile à l'intérieur d'un bâtiment utilisant les ondes WiFi environnantes. Ce manuscrit s'articule autour de deux questions centrales : celle de la détermination des cartes de propagation des ondes WiFi dans un bâtiment donné et celle de la construction d'estimateurs des positions du mobile à l'aide de ces cartes de propagation.

Le cadre statistique utilisé dans cette thèse afin de répondre à ces questions est celui des modèles de Markov cachés. Nous proposons notamment, dans un cadre paramétrique, une méthode d'inférence permettant l'estimation en ligne des cartes de propagation, sur la base des informations relevées par le mobile. Dans un cadre non-paramétrique, nous avons étudié la possibilité d'estimer les cartes de propagation considérées comme simple fonction régulière sur l'environnement à géolocaliser. Nos résultats sur l'estimation non paramétrique dans les modèles de Markov cachés permettent d'exhiber un estimateur des fonctions de propagation dont la consistance est établie dans un cadre général. La dernière partie du manuscrit porte sur l'estimation de l'arbre de contextes dans les modèles de Markov cachés à longueur variable.

Mots-clés : Localisation intra-muros, WiFi, modèles de Markov cachés, inférence statistique, estimation non-paramétrique, estimation en ligne.

Remerciements

Ce manuscrit conclut trois années de travail, je tiens ici à exprimer ma reconnaissance envers tous ceux qui de près ou de loin y ont contribué.

Je tiens tout d'abord à remercier chaleureusement Élisabeth Gassiat, ma directrice de thèse. Élisabeth, merci d'avoir encadré mon travail durant ces trois années de thèse. Merci d'avoir mis à ma disposition ton savoir, ta vivacité d'esprit, ton audace et tes relations. J'ai ainsi pu aborder les problématiques mathématiques explorées durant cette thèse sous les meilleurs angles, essentiellement grâce à tes conseils avisés, mais aussi avec l'aide des spécialistes que tu m'as présentés. Merci de m'avoir épaulé dans les moments de doutes et d'avoir dépensé tant d'énergie lors de nos rendez-vous hebdomadaires à m'enseigner le métier de chercheur. J'ai été honoré d'être sous ta direction pendant ces trois années, cette expérience restera j'en suis sûr parmi les plus enrichissantes que je connaitrai et ceci en grande partie grâce à toi.

Un grand merci aussi à Loïc Guillemard et, à travers lui, à l'entreprise Id Services qui a supporté ce travail. Merci de m'avoir proposé ce projet passionnant et de m'avoir accordé ta confiance pour le réaliser. Ces trois années m'ont permis de découvrir une multitude de domaines diverses comme la programmation ou l'élaboration et la gestion de projet en équipe. J'ai pu participer durant ma thèse à plusieurs phases de la réalisation d'un produit industriel, de la recherche au développement en passant par la protection de nos innovations et par la mise en place de bancs de tests sur site pilote. Merci aussi pour tes encouragements, ton support et tes conseils, essentiellement sur le plan humain.

Je tiens à exprimer ma gratitude à Ramon Van Handel et à François Le Gland pour l'intérêt qu'ils ont porté à mes travaux en acceptant à la fois de rapporter cette thèse et de faire partie de mon jury. Merci aussi à Eric Moulines et Pascal Massart qui m'ont fait l'honneur de participer à mon jury. Je voudrais par ailleurs aussi remercier Pascal Massart d'une part de m'avoir incité à effectuer ma thèse sous la direction d'Élisabeth qui, selon ses dires, est une "excellente directrice de thèse" et d'autre part d'avoir su, avec Marie Sauvé que je remercie aussi, soulever mon intérêt pour les statistiques.

Un grand merci à Liliane Bel et à Michel Prenat pour l'attention permanente qu'ils ont portée à mes travaux.

Merci à Sylvain avec qui j'ai collaboré pendant ma thèse. J'ai apprécié nos nombreux échanges mathématiques qui se sont toujours déroulés dans la bonne humeur.

Je profite de l'opportunité que m'offre cette page de remerciements pour témoigner ma reconnaissance aux enseignants de l'école Jean Jaures, du collège Molière, du lycée Saint-Rémi, du lycée Faidherbe, de l'ENS Cachan et de la faculté des Sciences d'Orsay qui ont réussi, tout au long de ma scolarité, à me transmettre leurs passions, notamment en sciences.

Je n'aurais pas su arriver au terme de cette thèse sans le support moral de mon entourage.

Julie, thank you for your love, the strongest support I could receive to keep going through these three years. Thanks for your courage and patience during these years that have been especially difficult for you. Thank you so much for encouraging me during the stressful moments, and for your numerous and precious advice that helped me overcome many situations. Finally, I would like to thank you for the kind attention you give me when I regularly expose my mathematical theories to you.

Merci à mon frère Julien, que je prends régulièrement en modèle. Le simple fait d'énumérer l'ensemble de ses qualités humaines et intellectuelles prendrait plus de pages que n'en contient ce

manuscrit. Merci pour ta gentillesse, d'être toujours là pour moi et de m'encourager dans tout ce que j'entreprends.

Un immense merci à mes parents, Agnès et Jean-Michel, pour tout ce qu'ils font pour moi, à ma grand-mère Denise pour son courage exemplaire, ainsi qu'aux autres membres de ma famille, oncles, tantes, cousines, d'être si gentils et attentionnés au quotidien, permettant ainsi à chacun de s'épanouir.

Merci au support de mes amis, en particulier de Aurélien avec qui j'ai eu le bonheur d'effectuer ces neuf années d'études, du charlot Maxence à qui je dois de nombreuses épopées hilarantes, toutes plus marquantes les unes que les autres. Merci à François, ami d'enfance et second petit frère ainsi qu'à Monique et Fernand ses parents. Takk fyrir Cyrille, sans ton expérience et tes conseils je n'aurais peut être pas envisagé de faire une thèse, merci aussi pour nos aventures sportives sur les chemins de France et de Navarre.

Et finalement un grand merci à mes collègues et amis, Laure, Caroline, Lucie, Lionel, Louis, Olivier, Chou-fleur, Christophe, Anne Claire, et à Poug le Roubaisien, pour tous les bons moments partagés ensemble.

Table des matières

Introduction

Depuis les années 90, les études scientifiques portant sur la géolocalisation à l'intérieur des bâtiments (ou *intra muros*) se sont multipliées. Les systèmes satellitaires de géolocalisation permettent actuellement une localisation précise à l'extérieur des bâtiments, permettant ainsi son utilisation dans un grand nombre d'applications. Leur utilisation à l'intérieur des bâtiments s'avère, quant à elle, beaucoup plus complexe. La première raison à cela réside dans la précision offerte par les systèmes satellitaires. Actuellement d'environ 10 mètres pour le système GPS public, cette précision empêche la réalisation de nombreuses applications à l'intérieur des bâtiments pour lesquelles une précision plus fine est nécessaire. La deuxième raison, et la plus importante, est que les signaux GPS parviennent difficilement à traverser les murs et les toits de nos bâtiments, rendant ainsi impossible son utilisation à l'intérieur. Beaucoup de technologies peuvent alors être mises à contribution pour supplanter le GPS pour la localisation *intra muros*, nous nous intéressons dans cette thèse à la localisation par signaux WiFi. Les ondes WiFi sont des ondes radios possédant des propriétés intéressantes. Outre le fait que de nombreux bâtiments sont déjà équipés en WiFi, facilitant ainsi la mise en place d'un système de localisation basé sur cette technologie, les ondes WiFi ont une portée de plusieurs dizaines de mètres et sont capables de traverser la plupart des matériaux que l'on peut rencontrer dans les bâtiments. Cependant, ces ondes sont soumises aux mêmes perturbations que toute autre onde radio lorsqu'elles rencontrent un obstacle, comme les effets d'absorption ou de réflexion. Alors que les techniques de localisation satellitaire exploitent essentiellement le fait que les signaux se propagent en ligne droite entre le satellite et le récepteur à localiser, à l'intérieur des bâtiments, les signaux sont constamment déviés par des obstacles et les techniques classiques de triangulation ou de trilatération peuvent difficilement être mises en oeuvre. L'information sur les signaux WiFi que nous avons choisi d'exploiter à des fins de localisation est la puissance des ondes à leur réception, cette information est fournie par la plupart des appareils communiquant en WiFi (appelés par la suite terminaux mobiles). Lorsqu'un signal radio est émis en champ libre, la puissance (en dBm) du signal décroît de manière logarithmique avec la distance entre le récepteur et l'émetteur de l'onde, dans ce cas, la localisation peut se faire en déterminant, grâce aux puissances reçues par le terminal, sa distance à chacun des points d'accès environnants. À l'intérieur des bâtiments la puissance évolue de manière plus complexe, formant ainsi des *cartes de propagation* épousant, d'une certaine manière, l'architecture du bâtiment. Tout l'enjeu se situe dans la détermination de ces cartes de propagation.

Nous aborderons le sujet d'un point du vue statistique, plusieurs aspects du problème seront discutés. Le premier porte sur la modélisation des phénomènes physiques intervenant dans la propagation des ondes à l'intérieur des bâtiments, ainsi que la modélisation des déplacements humains. Le deuxième porte sur l'estimation des quantités intervenant dans nos modèles (soit les cartes de propagation des signaux et, le cas échéant, les paramètres de déplacement du mobile) que nous trai-

terons de manière théorique et appliquée. Le troisième objectif est la localisation en elle même, soit la détermination des positions à l'aide des informations de puissance reçues. Nous nous plaçons dans le cas où un nombre ℓ de points d'accès (émetteurs WiFi) sont installés dans l'environnement K à géolocaliser. Nous supposons qu'à intervalles de temps réguliers le terminal mobile, que l'on souhaite géolocaliser, mesure la puissance des ondes émises par les ℓ points d'accès environnants, nous notons alors $\{Y_t\}_{t\in\mathbb{N}}$ la suite de ces mesures de puissances (en dBm), avec $Y_t \in \mathbb{R}^\ell$ et $\{X_t\}_{t\in\mathbb{N}}$ la suite des positions du terminal mobile correspondantes. Dans les résultats d'application présentés dans cette thèse, nous supposerons que la suite des positions $\{X_t\}_{t\in\mathbb{N}}$ appartient à un compact de $\mathbb{R}^2$.

Ce mémoire est composé de quatre parties, chaque partie traitant d'un des aspects de la problématique.

La première partie (partie I) de cette thèse est dédiée en particulier à la modélisation générale du problème. Nous introduisons notamment dans la section 2.3 ce que nous appellerons le *modèle spatio-temporel à états latents*, qui sera étudié dans le reste de la thèse et qui est décrit par l'équation suivante,

$$\forall t \in \mathbb{N},\ Y_t = f_\star(X_t) + \epsilon_t\ , \tag{1}$$

où ϵ_t est un bruit de mesure. La fonction $f_\star$ est alors une fonction représentant la propagation "moyenne" des ondes dans le bâtiment à géolocaliser, pour toute position x donnée, $f_\star(x)$ représente la puissance moyenne reçue par le terminal lorsqu'il se trouve en x. L'un des objectifs principaux de notre étude est donc l'estimation cette fonction $f_\star$. En effet, une telle estimation nous permettra de construire des estimateurs des positions X_t ayant observé les vecteurs de puissances reçues Y_t, ceci en tentant d'approcher la position x telle que $f_\star(x)$ soit "la plus proche" de la puissance observée Y_t. La seconde hypothèse importante émise dans ce mémoire est que la suite de positions prises par le terminal à localiser $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov. L'intuition derrière cette hypothèse est que la suite de positions est corrélée : si le pas de temps entre deux mesures faites par le terminal mobile Y_t et Y_{t+1} est faible, les positions associées X_t et X_{t+1} seront alors proches géographiquement. Le processus aléatoire $\{X_t, Y_t\}_{t\in\mathbb{N}}$, où seules les variables $Y_t, t \in \mathbb{N}$ sont observées, est alors un *modèle de Markov caché* (ou HMM pour *Hidden Markov Model*). Les HMM sont des outils puissants d'analyse, couramment utilisés dans des domaines divers comme en finance économétrique (Mamon and Elliott [2007]), en biologie (Churchill [1992]) ou reconnaissance vocale (Juang and Rabiner [1991]). De nombreux résultats ont été démontrés sur ces modèles, nous pouvons citer en particulier Cappé et al. [2005] qui en regroupe les principaux. Nous nous intéresserons dans la suite à des problématiques d'inférence dans les HMM que nous traiterons au sens du maximum de vraisemblance, ainsi qu'à des techniques de filtrage, notamment aux techniques de filtrage particulaire.

Dans la deuxième partie de cette thèse nous appliquerons des techniques classiques de filtrage pour les HMM à notre problème. En particulier, deux méthodes de localisation seront présentées dans cette partie. Ces deux méthodes sont basées sur une discrétisation de l'environnement, la fonction $f_\star$ est alors estimée grâce à une campagne de mesures préalable des puissances en tout ou partie des points issus de la discrétisation. La première de ces méthodes (présentée dans le chapitre 4) utilise une discrétisation grossière de l'environnement, l'espace d'états K de la chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$ (*i.e.* l'ensemble des valeurs pouvant être prises par le processus $\{X_t\}_{t\in\mathbb{N}}$) est alors fini. La fonction $f_\star$ est estimée grâce à une campagne de mesures préalable en tous les points de K. À tout instant t donné, un estimateur $\widehat{X}_t$ de la position X_t, que nous appellerons prédicteur de maximum *a posteriori*, est construit sur la base des mesures passées $Y_0, \ldots, Y_t$. Pour cela, nous utiliserons des

techniques classiques de filtrage pour les modèles de Markov cachés à espace d'états finis. Cependant, les résultats obtenus en terme de précision de la géolocalisation se sont révélés imprécis. Dans le but de les améliorer, la deuxième méthode (présentée dans le chapitre 5) utilise une discrétisation fine de l'environnement. Bien que l'espace d'états K soit toujours fini, son cardinal est très élevé, une phase de mesure préalable en chacune des positions de K est donc inconcevable. La phase de mesures est donc effectuée en certaines positions de K, la première étape de cette deuxième méthode consiste donc à construire un prédicteur de la fonction $f_\star$ sur K grâce à ces mesures. Pour ce faire, il est indispensable de donner une structure de dépendance spatiale à la fonction $f_\star$, nous émettons alors l'hypothèse que $\{f_\star(x)\}_{x\in K}$ est la réalisation d'un processus Gaussien sur K dont l'espérance est une fonction de x représentant la propagation théorique d'une onde radio radio en champ libre. Cette modélisation de $f_\star$ ainsi que notre volonté de l'estimer grâce à une campagne de mesures préalable en certains points de K nous conduit tout naturellement à considérer les techniques de statistiques spatiales (ou *krigeage*) (Cressie [1993]). Nous présenterons alors une méthode d'estimation de $\{f_\star(x)\}_{x\in K}$, inspirée de ces techniques. Une fois la fonction $f_\star$ estimée, la localisation se fait alors grâce à des techniques de *filtrage particulaire*, permettant le calcul approché des lois de filtrage *a posteriori* (*i.e.* les probabilités de présences de X_t en x conditionnellement à $Y_0, \ldots, Y_t$) :

$$\mathbb{P}\left(X_t = x \mid Y_0, \ldots, Y_t\right),\ x \in K\ .$$

Ces techniques de filtrage particulaire permettent alors le calcul du prédicteur de maximum *a posteriori* $\widehat{X}_t$, défini comme le x de K réalisant le maximum des lois de filtrage *a posteriori* et choisi comme estimateur de la position X_t. Devant la grande variabilité de cet estimateur, nous proposons une technique de lissage des positions, prenant en compte la régularité des déplacements humains, qui nous permettra de réduire l'erreur de positionnement. Nous avons soumis dans la section 5.3.2 notre système de géolocalisation aux données réelles dans un environnement de type entrepôt, la précision obtenue à l'issue de ces tests est de l'ordre de 5 à 7 mètres dans 80% des cas, selon le nombre et l'emplacement des points d'accès présents dans l'environnement. Bien que la précision de notre système soit acceptable, la phase de calibration préalable des cartes de propagation est contraignante en pratique puisqu'elle nécessite un effort humain. De plus, le modèle spatio-temporel à états latents décrit par l'équation (1) suppose que la fonction $f_\star$ ne dépend pas du temps, cependant, la manière dont les ondes WiFi se propagent dans les bâtiments (et donc la fonction $f_\star$) peut être modifiée lorsque des obstacles sont ajoutés ou déplacés. Les méthodes utilisant une phase de calibration souffrent alors d'une perte de précision dans le temps due aux modifications successives de la fonction $f_\star$.

Dans la troisième partie (partie III) nous étudierons la possibilité de se passer de cette phase de calibration en estimant la fonction $f_\star$ grâce uniquement aux mesures $\{Y_t\}_{t\in\mathbb{N}}$. Deux approches de ce problème seront étudiées. La première, présentée dans le chapitre 7 et basée sur une discrétisation fine de l'environnement, utilise un algorithme de localisation et de cartographie simultanée (ou *SLAM* pour *Simultaneous Localisation And Mapping*) permettant le calcul récursif des prédicteurs de maximum *a posteriori* $\widehat{X}_n$, $n \in \mathbb{N}$, simultanément aux mises à jour, elles aussi récursives, des cartes de propagations $f_\star$. Cette méthode fait l'objet d'un dépôt de brevet (Dumont and Gassiat [2012]) ainsi que d'un article (Dumont and Le Corff [2012a]), soumis pour publication dans une revue internationale et présenté dans le chapitre 7. Nous supposerons toujours que $f_\star$ est un processus gaussien permettant ainsi de lui donner une structure de dépendance spatiale. Notre méthode est basée sur un algorithme dérivé de l'algorithme EM (pour *Expectation-Maximisation*) en ligne (Cappé

[2011]), elle permet de mettre à jours régulièrement notre estimation de $f_\star$ à mesure que les observations $\{Y_t\}_{t\in\mathbb{N}}$ sont effectuées par le mobile. Lorsque l'évolution dans le temps de la propagation des ondes est lente (lorsque la convergence de notre algorithme est atteinte avant que la fonction $f_\star$ soit modifiée à nouveau), une ré-initialisation régulière de notre algorithme permet alors de prendre en compte cette évolution.

La deuxième approche, présentée dans le chapitre 8, est de ne pas discrétiser l'environnement et de supprimer l'hypothèse de gaussianité sur $f_\star$. L'espace d'états K est alors considéré comme un compact d'intérieur non vide de $\mathbb{R}^2$. Cette approche, portant sur la construction théorique et la démonstration de convergence d'un estimateur non-paramétrique de $f_\star$, a conduit à la rédaction de l'article *Nonparametric estimation in hidden Markov models* (Dumont and Le Corff [2012b]) présenté dans le chapitre 8 et soumis pour publication dans une revue internationale. Cet article est consacré à une nouvelle méthode d'estimation non paramétrique dans les modèles de Markov cachés. L'originalité de cette étude porte sur le fait que l'on arrive à approcher la fonction $f_\star$ alors que les valeurs prises par le processus $\{X_t\}_{t\in\mathbb{N}}$ ne sont pas observées. Nous supposons que la chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$ est isotrope de noyau connu à un facteur d'échelle près, noté $a_\star$. Grâce à cette condition sur la structure de dépendance de $\{X_t\}_{t\in\mathbb{N}}$, et sous certaines conditions sur K et $f_\star$, le modèle décrit par l'équation (1) est rendu identifiable. Ce résultat d'identifiabilité utilise des outils de géométrie différentielle et de topologie pour prouver que s'il existe un couple (f, b) de paramètres tel que le processus $\{Y'_t\}_{t\in\mathbb{N}}$, décrit par l'équation $Y'_t = f(X'_t) + \epsilon'_t$, a même distribution que $\{Y_t\}_{t\in\mathbb{N}}$, avec $\{X'_t\}_{t\in\mathbb{N}}$ chaîne de Markov isotrope de même noyau que $\{X_t\}_{t\in\mathbb{N}}$ et de facteur d'échelle b, et $\{\epsilon'_t\}_{t\in\mathbb{N}}$ a même distribution que $\{\epsilon_t\}_{t\in\mathbb{N}}$, alors $b = a_\star$ et il existe une isométrie ϕ de K telle que $f = f_\star \circ \phi$. L'estimation de $f_\star$ et du paramètre $a_\star$ se fait en considérant $(\widehat{f}_n, \widehat{a}_n)$ maximisant un critère de vraisemblance pénalisée sur un espace de Sobolev. La consistance de cet estimateur est vérifiée grâce à un contrôle du processus empirique pour les processus faiblement dépendants.

La dernière partie (partie IV) de cette thèse est consacrée à une classe de modèles peu étudiée jusqu'à présent, les modèles de Markov cachés à longueur variable. Nous supposons alors que l'espace d'états est fini. Lorsque le processus $\{X_t\}_{t\in\mathbb{N}}$ est modélisé par une chaîne de Markov, la distribution de X_t conditionnellement à tout le passé $X_{t-1}, X_{t-2}, \ldots, X_0$ dépend uniquement de X_{t-1}. Dans cette partie nous supposons que la suite $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov à longueur variable (ou VLMC pour *Variable Length Markov Chains*). Un arbre de contextes $\tau_\star$ est associé à la VLMC $\{X_t\}_{t\in\mathbb{N}}$. $\tau_\star$ est un ensemble de suites (que l'on considère finies dans notre étude) de K, tel que pour tout élément (contexte) $s = (s_1, \ldots, s_l)$ de $\tau_\star$, et pour tous éléments $x_{t-1}, \ldots, x_0$ de K vérifiant $x_{t-1} = s_1, \ldots, x_{t-l} = s_l$, la distribution de X_t, conditionnellement à $X_{t-1} = x_{t-1}, \ldots, X_0 = x_0$, est égale à la distribution de X_t conditionnellement à $X_{t-1} = x_{t-1}, \ldots, X_{t-l} = x_{t-l}$. La distribution de la VLMC est alors déterminée par son arbre de contextes $\tau_\star$ et par les probabilité de passages $P_{s,x}$, $s = (s_1, \ldots, s_l) \in \tau_\star,\ x \in K$ où $P_{s,x} = \mathbb{P}\left(X_t = x | x_{t-1} = s_1, \ldots, x_{t-l} = s_l\right)$. Ces modèles, introduits par Rissanen dans les années 80 (Rissanen [1986]) sont largement utilisés, notamment en théorie du codage (Willems et al. [1995]). L'objectif de cette partie est d'affiner notre compréhension de la dynamique de $\{X_t\}_{t\in\mathbb{N}}$ grâce à la détermination de son arbre de contextes $\tau_\star$ ainsi que des probabilités de passages $\{P_{s,x}\}_{s\in\tau_\star,\ x\in K}$. Ce problème d'estimation de l'arbre de contexte dans les VLMC a déjà été étudié (voir par exemple Csiszar and Shields [2000], Garivier [2006]). Cependant, dans notre cas, la VLMC $\{X_t\}_{t\in\mathbb{N}}$ n'est pas observée directement mais à travers le processus d'observations $\{Y_t\}_{t\in\mathbb{N}}$. Nous dirons alors que $\{X_t, Y_t\}_{t\in\mathbb{N}}$ est un modèle de Markov caché à longueur variable (ou VLHMM pour *Variable Length Hidden Markov Models*). L'enjeu de cette quatrième partie est l'estimation de

l'arbre de contexte dans les VLHMM. Notre étude à donné lieu à la rédaction d'un article (*Context tree estimation in variable length hidden Markov models*, Dumont [2011]), à ce jour en révision pour le journal *IEEE Transactions on Information Theory*, qui sera présenté dans le chapitre 11. Nous y construirons un estimateur $\widehat{\tau}_n$ de $\tau_\star$ maximisant un critère de vraisemblance pénalisée, dont nous prouverons la consistance presque-sûre grâce à un contrôle des fluctuations de la vraisemblance dans l'esprit des travaux de Chambaz et al. [2009] sur l'estimation d'ordre (nombre d'états) dans les HMM.

Bibliographie

O. Cappé. Online EM algorithm for Hidden Markov Models. *To appear in J. Comput. Graph. Statist.*, 2011.

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2 ; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

A. Chambaz, A. Garivier, and E. Gassiat. A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification. *Journal of Stat. Planning and Inf.*, 139 :962–977, 2009.

G. Churchill. Hidden Markov Chains and the Analysis of Genome Structure. *Computers & Chemistry*, 16(2) :107–115, 1992.

N. A. C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience, revised Edition edition, January 1993. ISBN 0471002550.

I. Csiszar and P. C. Shields. The consistency of BIC Markov order estimator. *Annals of Stat.*, 28 : 1601–1619, 2000.

T. Dumont. Context tree estimation in variable length hidden Markov models. Technical report, 2011.

T. Dumont and E. Gassiat. Procédé de mise à jour continu d'un paramètre représentatif d'une grandeur physique dépendant de sa localisation, et dispositif associé. Patent Provis. app. num. : 1000146721, 2012.

T. Dumont and S. Le Corff. Simultaneous localization and mapping problem in wireless sensor networks. Technical report, 2012a.

T. Dumont and S. Le Corff. Nonparametric estimation in hidden Markov models. Technical report, 2012b.

A. Garivier. Consistency of the unlimited BIC Context Tree estimator. *IEEE Trans. Inform. Theory*, 52 :4630–4635, 2006.

B. Juang and L. Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33 : 251–272, 1991.

R.S. Mamon and R.J. Elliott. *Hidden Markov Models in Finance*, volume 104 of *International Series in Operations Research & Management Science*. Springer, Berlin, 2007.

J. Rissanen. Stochastic complexity and modeling. *Ann. Statist.*, 14 :1080–1100, 1986.

F. M. J. Willems, Y.i M. Shtarkov, and T. J. Tjalkens. The Context Tree Weighting Method : Basic Properties. *IEEE Transactions on Information Theory*, 41 :653–664, 1995.

Première partie

État de l'art et modélisation des phénomènes physiques intervenant dans la propagation des ondes

Chapitre 1

État de l'art de la géolocalisation

Sommaire

En quelques années, les services de géolocalisation se sont développés de manière fulgurante, que ce soit via l'explosion des smartphones pour le grand public, équipés de GPS et de systèmes inertiels, ou par l'utilisation de plus en plus fréquente de balises satellites sur les flottes de véhicules dans les entreprises. L'utilisation de cartes papier perd notamment de plus en plus de terrain face à l'arrivée des nouvelles technologies en matière de localisation. Dans cette partie seront mis en avant les avantages et inconvénients d'un monde de plus en plus "géolocalisé". Nous aborderons notamment les problèmes portant sur la vie privée et les droit des personnes géolocalisées, que ce soit au travail ou à l'échelle de la personne notamment sur les réseaux sociaux. Nous parlerons aussi des avancées dans la vie courante que peuvent apporter les services de localisation comme leur utilisation pour la sécurité et le bien-être des personnes : balises Argos pour localiser une victime prise dans une avalanches, aides aux personnes à mobilité réduite, optimisation du temps d'intervention et de l'organisation dans les hôpitaux, coordination des services d'intervention (pompiers, policiers,...) ou tout simplement l'aide au guidage proposé par les smartphones qui évitent de se déplacer avec les cartes des endroits à visiter. Nous aborderons aussi l'apport de la géolocalisation dans l'industrie notamment pour la logistique avec l'optimisation des parcours sur la chaîne de traitement d'un produit, ou pour la sécurité.

1.1 Définition de la géolocalisation et cadre légal de son utilisation

Définition Wikipedia :

La géolocalisation ou géoréférencement est un procédé permettant de positionner un objet (une personne, une information,) sur un plan ou une carte à l'aide de ses coordonnées géographiques.

1.1.1 L'utilisation de la géolocalisation

Les premières techniques de localisation apparaissent très tôt durant l'antiquité avec les premières cartographies. S'ensuivent les inventions comme la boussole (par la chine au IIe Siècle av. J.C.) ou, beaucoup plus tard (au XVIIIe Siècle), le sextant permettant de se repérer grâce aux astres. Depuis très tôt dans l'histoire, la nécessité de se repérer dans l'espace se fait ressentir. Le *Global Positioning System* (GPS) inventé en 1978 par le département de la défense américaine marque un tournant dans le positionnement sur le globe. Depuis, beaucoup d'autres technologies sont utilisées à des fins de localisation. Nous nous intéresserons à certaines d'entre elles dans la section 1.2 et la section 1.3. Les applications de la géolocalisation sont aussi bien professionnelles que privées.

Dans la sphère professionnelle tout d'abord, la localisation en temps réel ou différé des personnes ou objets a depuis toujours représenté un gain de productivité, de temps, et de carburant. Une des inventions les plus anciennes et des plus utilisée à ces fins aujourd'hui est l'invention du code à barres en 1952. Ce procédé permet de traquer les produits et/ou le matériel sur toute la chaîne de production permettant ainsi son optimisation. Cependant, l'arrivée du GPS pour le grand public et les entreprises a, elle aussi, transformé la manière dont certaines entreprises gèrent leurs marchandises. La gestion des flottes de véhicules assistée par GPS, notamment, a considérablement été modifiée dans les entreprises de transport (transports en commun, transport de marchandises,...) et a ainsi permis de nombreuses économies pour ces entreprises. La technologie GSM (*Global System for Mobile Communications*) est, elle aussi, largement utilisée pour géolocaliser. La position d'un téléphone portable peut, par trilatération, être déterminée, ce qui permet notamment d'assister les services d'urgence à la personne (pompiers, samu, police ou gendarmerie) en leur permettant d'identifier la position d'une urgence dans un laps de temps restreint. La société Ekahau a déployé dans certains hôpitaux une solution de localisation afin d'optimiser les déplacements du personnel, d'identifier rapidement l'emplacement des équipements hospitaliers dans un souci de gain de temps durant les interventions et de suivre les déplacements de patients présentant un danger pour eux même ou pour les autres.

Dans la sphère privée, et avec l'avènement des nouvelles technologies, les moyens de localisation ont explosé. Il est dès lors possible d'avoir accès au positionnement GPS depuis sa voiture, permettant une navigation bien plus aisée qu'avec l'utilisation de cartes routières, mais aussi depuis son *smartphone*, facilitant ainsi les déplacements, en particulier dans une ville inconnue. Certaines applications permettent même de localiser un proche portant son téléphone uniquement via un accès à Internet. Récemment, le système BlueEyes (REF), basé sur la technologie RFID, est expérimenté dans certaines stations du métro parisien. Ce système propose un système de guidage pour Personnes à Mobilité Réduite (PMR). Il permet ainsi aux personnes de se déplacer de manière autonome dans les couloirs des stations de métro leur étant peu accessibles jusqu'à présent. Depuis peu, les compagnies françaises Insiteo et Polestar proposent des solutions de localisation, utilisant les signaux envoyés par une infrastructure WiFi, pour faciliter le guidage des personnes dans certains centres commerciaux et salons.

1.1.2 Les risques inhérents à la géolocalisation

Malgré les avantages apportés par les systèmes de localisation, le fait de localiser des personnes à leur insu ou non pose un problème évident de liberté individuelle. En France la commission nationale de l'informatique et de libertés (CNIL) est chargée de veiller à ce que les principes relatifs à la protection de données à caractère personnel soient bien respectés. Cet organisme a mis en place un guide de la géolocalisation des salariés (Guide), imposant certaines règles aux employeurs désireux de suivre les déplacements de leurs salariés ou du véhicule qu'ils occupent.

Depuis l'arrivée des *smartphones*, équipés de moyens de localisation, sur le marché, des questions inhérentes à la protection des données à caractère personnel se posent. La mise en mémoire du parcours effectué par le mobile, ou la publication régulière de sa position sur les réseaux sociaux, par exemple, peuvent entrer en conflit avec les règles imposées par la CNIL dès lors que l'existence même de ces données peut ne pas être connue de l'utilisateur.

1.2 La géolocalisation en extérieur

1.2.1 Les systèmes satellitaires, principe et acteurs

Les systèmes satellitaires reposent tous sur le même principe : la trilatération. La trilatération est l'estimation de positions utilisant les mesures de distances à des points référents dont la position est supposée connue. Le principe consiste à mesurer le temps (*Time Of Arrival* - TOA) que met une onde radio pour se propager entre un émetteur d'onde et le récepteur dont on veut connaitre la position. La vitesse de propagation de l'onde étant connue (celle de la lumière), cela permet de construire un estimateur de la distance entre l'émetteur et le récepteur. La figure 1.1 illustre le principe de la trilatération. Les systèmes satellitaires se basent sur un déploiement de satellites en orbite autour de la terre. Chacun de ces satellites ayant une trajectoire bien déterminée (sur son orbite), la position d'un satellite sur son orbite est aisément déterminée. Cette position consiste en un jeu de paramètres appelé éphéméride. Les éphémérides de tous les satellites sont donc calculés par une station de contrôle, qui envoie régulièrement leur éphéméride, et donc leur position, à chacun des satellites en orbite. Chaque satellite transmet alors en continu un signal, codant notamment l'éphéméride et le temps de départ du signal, à destination des récepteurs terrestres désireux de se localiser. A la réception, le récepteur décode le signal, déduit la position du satellite, et grâce à son horloge interne, détermine le temps de parcours de l'onde. En principe seuls 3 signaux provenant de 3 satellites différents sont nécessaires pour effectuer le calcul de position. En effet, supposons que l'ensemble $\{1,2,3\}$ représente les 3 satellites, si pour $i \in \{1,2,3\}$, t_i représente le temps de parcours de l'onde et (x_i, y_i, z_i) la position relative au satellite i dans un référentiel donné. En notant c la vitesse de la lumière, alors $r_i = ct_i$ estime la distance entre le satellite i et le récepteur. De plus, en notant (x, y, z) la position inconnue du récepteur dans le même référentiel, alors (x, y, z) vérifie le système à 3 équations

$$\forall i \in \{1,2,3\},\ r_i = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2}\ ,$$

qui peut être résolu utilisant des méthodes de linéarisation (Evennou [2007]). Malheureusement, contrairement aux satellites dont les horloges (atomiques) sont synchronisées, l'horloge du récepteur

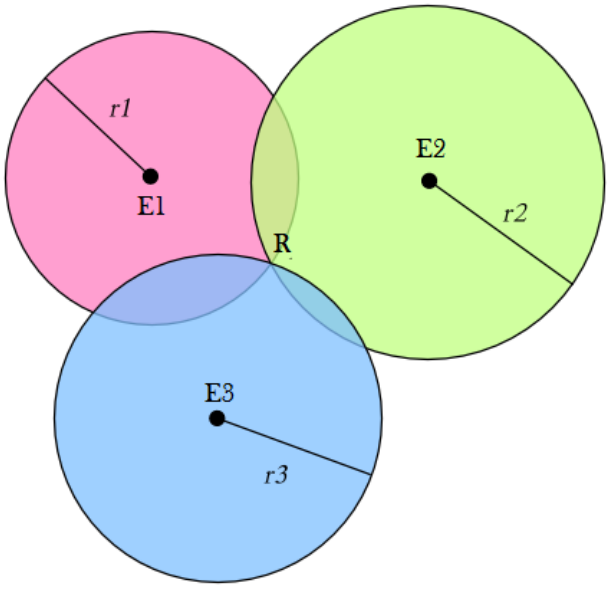


FIGURE 1.1 – Trilatération sur un plan : La détermination de la distance d'un récepteur (R) à 3 émetteurs (E1,E2 et E3) dont les positions sont connues permet de calculer la position de ce récepteur

ne l'est pas. Il existe donc un biais t_b dans la mesure du temps de parcours t_i de sorte que $c(t_i - t_b)$ correspond à la "vraie" distance entre le satellite i et le récepteur. Si $b = ct_b$, alors

$$\forall i \in \{1,2,3\},\ r_i = \sqrt{(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2} + b\ ,$$

de sorte que 4 paramètres au lieu de 3 doivent être estimés, d'où la nécessité pour le récepteur de recevoir le signal d'un quatrième satellite afin de rendre le système résoluble. Les trois principaux acteurs de la géolocalisation par satellite sont les suivants :

Le système américain GPS est constitué de 24 satellites au minimum, il est disponible pour le grand public depuis 1995 mais les restrictions imposées à des fins militaires sur le signal à usage civil ne permet une localisation précise qu'à une centaine de mètres près. Ce n'est qu'en 2000 que les États-Unis lèvent cette restriction permettant alors une localisation entre 10 et 20 mètres de précision.

Glonass est un système de positionnement par satellites d'origine soviétique. Il est opérationnel en 1995 avec 24 satellites en orbite offrant une précision comparable à celle du GPS. Cependant, par manque de budget alloué au projet, et face à la faible durée de vie des satellites soviétiques (environ 3-4 ans), le nombre de satellites diminue considérablement jusqu'en 2000 où seuls 6 satellites sont restés en orbite. Depuis, la Russie s'est lancée dans plusieurs programmes de renouvellement de ses satellites avec pour objectifs l'augmentation de leur durée de vie et l'amélioration de la précision.

Galileo est l'équivalent européen des systèmes Américains et Russes, il devrait être mis en service en 2020 et permettrait d'obtenir une précision de l'ordre de 4 à 5 mètres pour le service gratuit

et de l'ordre du mètre pour le service commercial.

1.2.2 Les systèmes de localisation terrestres

Basés sur le même principe de trilatération que les systèmes satellitaires, les solutions suivantes permettent une géolocalisation utilisant des émetteurs terrestres :

Le DGPS (système GPS différentiel) utilise des stations de réception des signaux GPS terrestres. Situés à des emplacements connus, ils sont chargés d'analyser les signaux GPS provenant des satellites et d'estimer les erreurs entachant l'information apportée par ces signaux. Quatre sources de fluctuation des signaux ont été répertoriées : erreurs dues au décalage d'horloge, les erreurs dans l'estimation de l'éphéméride des satellites par la station de base, les effets relativistes (dues à la vitesse excessive des satellites par rapport au récepteur provoquant des décalages temporelles) et les erreurs dues à l'entrée dans l'atmosphère des signaux dont le comportement (vitesse, direction,...) s'en trouve modifié. Les corrections ainsi apportées sont alors transmises aux récepteurs compatibles DGPS à proximité qui peuvent alors corriger eux-même les signaux qu'ils reçoivent des mêmes satellites et ainsi se positionner de manière plus précise. Ce système permettrait de passer des 20 mètres de précision offerts par le système GPS classique à environ 4 mètres de précision.

Le système Egnos est destiné à fonctionner de paire avec le système satellitaire Galileo, il fonctionne sur le même principe que le DGPS.

Le système GSM : depuis l'apparition des téléphones portables, il est plus difficile (pour les services d'urgence par exemple) de déterminer d'où ont été émis les appels. Le système de localisation GSM fonctionne sur le principe de trilatération illustré sur la figure 1.1. Il a pour objectif de géolocaliser les téléphones portables (GSM) en utilisant les signaux qu'ils échangent avec les antennes relais servant à la communication. Contrairement aux systèmes satellitaires, l'information utilisée sur un signal reçu est sa puissance et non plus son temps de parcours qui nécessiterait d'équiper les antennes (et les téléphones portables) d'horloges très précises qu'il faudrait de plus synchroniser. Cependant le principe reste le même. Comme nous le verrons dans la section 2.2, il existe un lien entre la puissance d'un signal reçu et la distance entre l'antenne émettrice du signal et l'endroit où il a été reçu. Pour localiser un appareil GSM, les antennes avoisinantes relèvent la puissance des signaux envoyés par cet appareil. Un calculateur traite alors ces données afin de déterminer les distances séparant l'appareil GSM aux antennes recevant ses signaux. Ces distances permettant alors d'estimer par trilatération la position du GSM. La précision obtenue par cette technique est de l'ordre de 10 à 50 mètres.

1.3 La géolocalisation intra-muros

Depuis les années 90, les recherches se sont intensifiées dans le domaine de la géolocalisation intra-muros. Préfigurant la prochaine grande étape dans le domaine de la navigation, ce domaine vise à lever l'impossibilité pour les techniques satellitaires à fournir leurs services à l'intérieur des bâtiments. Les domaines d'applications pour la géolocalisation intra-muros sont très vaste. Que se

soit dans la santé, la sécurité, le commerce, ou la logistique, les attentes pour une solution de géolocalisation intra-muros se font de plus en plus grandes. Il faut savoir que parmi les solutions existantes, toute précision aussi fine que désirée peut à ce jour être atteinte, cela grâce à un choix adéquat de la technologie utilisée. Certaines de ces technologies seront discutées dans cette section. Cependant, celles permettant d'obtenir une localisation précision très fine demandent une infrastructure lourde et souvent coûteuse. Ce qui rend la géolocalisation intra-muros si différente de la localisation en extérieur est la densité importante d'obstacles à l'intérieur des bâtiments. Contrairement aux systèmes satellitaires en extérieur qui, mis à part dans certains milieux très urbanisés, arrivent à envoyer des signaux en ligne directe entre un émetteur et un récepteur, à l'intérieur des bâtiments, les obstacles (murs, mobiliers, objets, personnes,...) perturbent la propagation des signaux. Se baser sur le temps de parcours du signal, en plus des coûts imposés par la mise en place de la solution (nécessité d'une horloge très précise et de systèmes de synchronisations tout aussi précis), est rendu très difficile du fait que le parcours que le signal a emprunté n'est pas forcément la ligne droite et est inconnu. Dans ce qui suit, différentes technologies permettant la géolocalisation intra-muros sont mises en avant.

- La technologie infra-rouge (IR). Cette technologie est surtout utilisée en robotique. En particulier, Abrate et al. [2007] propose un algorithme de localisation et de cartographie simultanée (abrévier SLAM pour Simultenous Localization And Tracking, nous étudierons une technique similaire dans le cadre de la géolocalisation WiFi dans la section 7) afin de localiser un robot grâce à un système IR permettant de détecter les obstacles à proximité. Grâce à ses capteurs, le robot construit une cartographie de l'environnement et de ses obstacles et se localise sur cette même carte. Cette technique suppose donc a priori que les obstacles et leur emplacement ne varient pas dans le temps. D'autres systèmes utilisent un principe d'émetteurs-récepteurs où l'objectif est de localiser un émetteur IR grâce à des capteurs disséminés dans l'environnement. Les systèmes IR ont cependant leur limite étant donné que les rayons IR sont stoppés par le moindre obstacle, le champ doit être dégagé pour que ces rayons puissent être utilisés.
- La localisation par données inertielles. Une centrale inertielle est capable de délivrer des informations relatives à son accélération (linéaire et angulaire). La localisation se fait alors par intégration de ces données de manière à déterminer sa vitesse et sa position. Cette détermination n'est d'ailleurs possible que lorsque l'on connait la vitesse et la position de départ de la centrale. Cependant ces systèmes sont soumis à dérives, car les données de la centrale seront nécessairement bruitées. Une calibration régulière de la position, ou l'hybridation de ce système de localisation avec une autre technologie, est nécessaire pour minimiser les effets de ces dérives.
- localisation par ondes radio (WiFi, bluetooth, Ultra large bande (ULB),....). Les ondes radio ont la propriété de pouvoir traverser les obstacles les plus communs dans les bâtiments, de sorte qu'une localisation peut être envisagée sans avoir accès, en toute position de l'environnement, à une vision directe avec les émetteurs.

L'ULB : La caractéristique de l'ULB (pour Ultra Large bande) réside dans sa bande de fréquence qui, comme son nom l'indique, est relativement large, cette largeur devant être au minimum 20 pour cent de la fréquence centrale (voir Gezici et al. [2005]). Cette largeur de bande offre de grands avantages que ce soit pour la communication ou pour la localisation. En effet, le signal est constitué d'une grande gamme de fréquences ce qui augmente

la probabilité que l'une d'entre elles traverse ou contourne les obstacles présents dans l'environnement. De plus, les systèmes ultra large bande offrent une précision temporelle élevée, les techniques utilisant le TOA (trilatération) peuvent alors être mis en oeuvre. En plaçant des émetteurs ULB au plafond, par exemple, la précision obtenue par trilatération est de quelques centimètres seulement. Cependant, avec une portée de quelque dizaines de mètres en champ libre (20 mètre environ), l'infrastructure à déployer pour couvrir un bâtiment est lourde et coûteuse. De plus, les équipements capables de communiquer en ULB ne sont pas très nombreux à ce jour.

Le RFID : la première utilisation du RFID (de l'anglais Radio Frequency IDentification) est, à l'instar des codes à barres, destinée à l'identification de produits. Le principe du RFID est basé sur la lecture du champ magnétique d'une étiquette RFID passive par un lecteur RFID. Son utilisation dans la localisation consiste à placer dans l'environnement à localiser des lecteurs RFID (sur les pas de portes par exemple pour détecter les entrées et sorties). Au passage d'une étiquette RFID, le lecteur identifiera le champ magnétique de l'étiquette et pourra alors communiquer à un serveur l'information. Comme pour le ULB, le problème de la portée se pose aussi puisqu'elle est de l'ordre du mètre. Cependant, comme c'est le cas dans la solution de localisation de l'entreprise Ekahau, cette technologie peut facilement être combinée à d'autres systèmes de localisation pour affiner la localisation à certains endroits.

Le GPS Les difficultés du GPS à passer à travers les matériaux ainsi que sa faible précision rend impossible son utilisation à l'intérieur de locaux. Cependant, des solutions existent, l'idée, développée en particulier par le département électronique et physique de l'école d'ingénieurs Télécom Sup Paris, est de placer une antenne GPS à l'extérieur du bâtiment à géolocaliser afin d'acheminer les signaux provenant des satellites à l'intérieur des bâtiments via plusieurs antennes relais. La difficulté ici réside dans le fait que les signaux GPS ainsi retransmis sont, comme toute onde radio, perturbés par les obstacles de l'environnement. Le signal émis "ricoche" alors sur ces obstacles et arrive en plusieurs répliques au récepteur. L'objectif est alors d'identifier la première de ces répliques correspondant en théorie au chemin direct entre l'émetteur et le récepteur.

Le WiFi Les ondes WiFi peuvent, elles aussi, être utilisées à des fins de localisation. Plusieurs arguments sont en faveur de l'utilisation des ondes WiFi pour la localisation. Le premier avantage concerne l'installation d'un système de localisation utilisant le WiFi, puisqu'elle est avant tout logicielle. En effet, de nombreux bâtiments et de lieux publics sont déjà équipés en points d'accès WiFi (Access Point ou AP) que l'on peut utiliser à des fins de géolocalisation et, de la même manière, la plupart des terminaux portatifs sont déjà capables de communiquer en WiFi. Outre les faibles besoins matériels nécessaires à son implémentation, ces ondes, d'une portée de plusieurs dizaines de mètres (environ 50m en intra-muros) sont capables de traverser la plupart des obstacles. Ainsi, il n'est pas nécessaire d'installer un point d'accès dans chaque pièce ou derrière chaque obstacle de l'environnement. Finalement le coût d'installation de points d'accès supplémentaires (au cas où la couverture n'est pas assez dense) est relativement faible.

Bibliographie

F. Abrate, B. Bona, and M. Indri. Experimental EKF-based SLAM for Mini-rovers with IR Sensors Only. In *EMCR*, 2007.

F. Evennou. *Techniques et technologies de localisation avancées pour terminaux mobiles dans les environnements indoor*. 2007. Ph.D. Thesis, Univ. Joseph Fourier, Grenoble, France.

S. Gezici, Z. Tian, G.B. Giannakis, H. Kobayashi, A.F. Molisch, H.V. Poor, and Z. Sahinoglu. Localization via ultra-wideband radios : a look at positioning aspects for future sensor networks. *Signal Processing Magazine, IEEE*, 22(4) :70–84, 2005.

CNIL Guide. Guide pour les employeurs et les salariés. http ://cdekeyser.com/data/files/cdekeyser.com-cnil-guide-geolocalisation.pdf.

Chapitre 2

Etude de la localisation par mesure du RSSI

Sommaire

2.1 Principes de la localisation par ondes WiFi

Dans cette section nous aborderons l'architecture (basique) d'un réseau WiFi, puis nous étudierons les propriétés du RSSI, information sur la puissance des signaux WiFi reçus par le terminal, que l'on utilisera pour la géolocalisation. Nous terminerons ce chapitre par la description du modèle statistique étudié dans le reste de ce mémoire.

Commençons tout d'abord par présenter l'infrastructure WiFi illustrée sur la figure 2.1. Le système est composé de terminaux mobiles (ordinateurs portables, appareils mobiles miniaturisés, *smartphone*, ...) communiquant, grâce au réseau, avec un serveur d'applications. Ce serveur sera alors chargé de géolocaliser ces terminaux. La communication entre les terminaux mobiles et le réseau se fait en WiFi par l'intermédiaire de points d'accès (*access points* ou APs). Certains acteurs de la géolocalisation par signaux WiFi intègrent le module de calcul de la position dans les terminaux. Cependant la capacité de calcul de certains terminaux étant limitée, les méthodes de localisation nécessitant des calculs lourds comme celles présentées dans cette thèse nécessiteront l'utilisation d'un serveur plus puissant pour les effectuer.

Nous supposons qu'à intervalle de temps régulier, chaque terminal mobile mesure la puissance des signaux (*Received Signal Strength Indicator* ou RSSI) provenant des APs environnants (voir la figure 2.2). Il transmet alors, sous la forme d'un tableau AP/RSSI, ces informations au serveur, via le réseau. À la réception, le serveur traite ces informations afin de localiser chaque terminal mobile. Chaque position est alors enregistrée dans une base de données, envoyée au terminal, ou traitée par une application tiers. L'idée sous-jacente à la localisation par RSSI est que plus le RSSI d'un AP est fort, plus le terminal est proche de ce point d'accès. La mesure faite par le terminal des

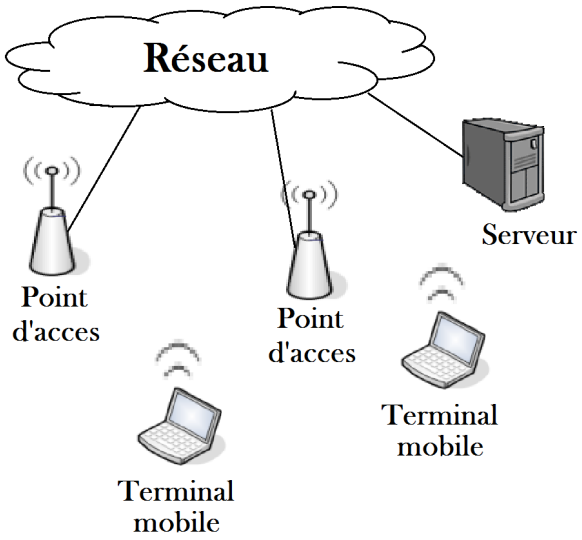


FIGURE 2.1 – Représentation basique d'une infrastructure WiFi

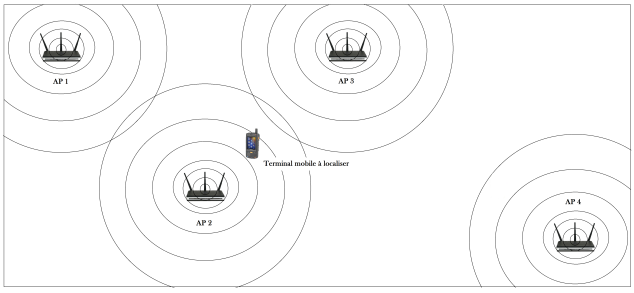


FIGURE 2.2 – Principe de la localisation par ondes WiFi

RSSI peut varier d'un appareil à un autre étant donné qu'il s'agit seulement d'un indicateur de la puissance du signal. Pour de nombreux terminaux WiFi, le RSSI est mesuré en dBm qui représente le rapport entre la puissance mesurée (en Watt) et un milliwatt : $P(dBm) = 10\log_{10}(1000P(W))$. Cette unité de mesure absolue n'est donc plus un simple indicateur et sera alors désignée par RSS (*Received Signal Strength*). Par ailleurs, en comparant les mesures du RSSI produites par différents constructeurs de terminaux, nous pensons pouvoir établir une règle de conversion linéaire entre le RSSI mesuré par un terminal WiFi quelconque et son équivalent en dBm. Nous parlerons donc dans la suite de géolocalisation par mesure du RSS.

La géolocalisation par mesures du RSS est difficile pour plusieurs raisons. Déterminer la position d'un terminal grâce au RSS suppose de connaitre le RSS "attendu" correspondant à n'importe quelle position, ce que l'on appellera : cartes de propagation des signaux. La section 2 décrira la nature des problèmes rencontrés lorsque l'on entreprend d'établir de telles cartes de propagation dans un milieu confiné. Une autre raison à cette difficulté est la grande incertitude lors de la mesure du RSS. La figure 2.3 est extraite de Pan et al. [2008], elle représente la variation du RSS pour une position

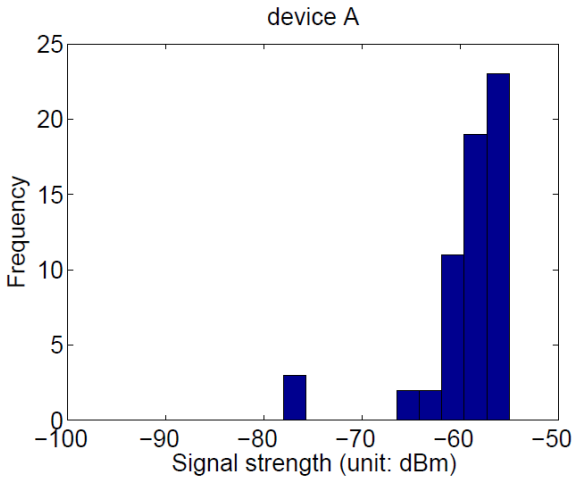


FIGURE 2.3 – Variation du RSS mesuré par un terminal pour une position fixe.

fixe. Ainsi, pour une position donnée et entre deux mesures consécutives d'un terminal, le RSS peut varier de plus de 10dBm ce qui représente une incertitude importante.

2.2 Un peu de physique

Les ondes WiFi sont des ondes radio dont la fréquence se situe autour de 2.4GHz. Comme toute onde radio, les ondes WiFi se propagent en ligne droite dans un milieu homogène et isotrope, mais, à la rencontre d'un obstacle, l'onde est soumise aux mêmes phénomènes de perturbation que la lumière : la réflexion, la réfraction, la diffraction, la diffusion et les interférences. Lorsque l'onde ne rencontre aucun obstacle, sa puissance évolue en fonction de la distance qu'elle a parcourue selon l'équation des télécommunications (ou équation de Friis (Friis [1946])) donnée par :

$$Y_R(d) = Y_E + 10\log_{10}(G_E G_R) + 20\log_{10}(\frac{\lambda}{4\pi d}) \ , \tag{2.1}$$

où $Y_R(d)$ est la puissance reçue (en dBm) à une distance d de l'émetteur de l'onde, Y_E est la puissance d'émission de l'onde (en dBm), G_E (resp. G_R) est le gain de l'antenne d'émission (resp. de réception) et λ est la longueur d'onde. Cependant cette modélisation est trop "simpliste" pour prendre en compte les multiples perturbations subies par les ondes WiFi en milieux confinés. La figure 2.4 représente la carte de propagation d'une onde WiFi partant d'un point d'accès dans un environnement de bureaux. Cette carte a été construite en mesurant le RSS du point d'accès concerné en chacun des points d'une grille recouvrant l'environnement. Environ 10 mesures du RSS par position ont été effectuées, la figure 2.4 représente alors la moyenne de ces mesures pour chaque position. On remarque que le RSS en un point donné ne dépend pas uniquement de la distance qui le sépare du point d'accès, mais aussi des obstacles environnants perturbant ainsi le trajet direct de l'onde. Plusieurs modèles de propagation *déterministes* ont pour objectif de considérer les obstacles dans la

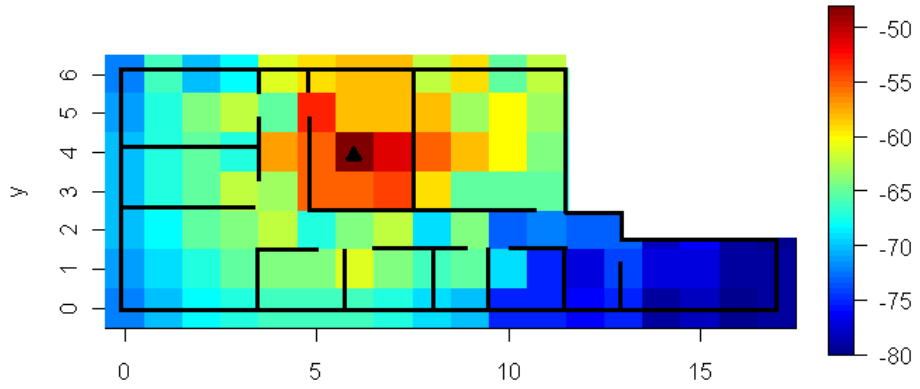


FIGURE 2.4 – Mesures du RSS (en dBm) dans un environnement de bureau. Le triangle noir symbolise la position du point d'accès, et les segments noirs représentent les murs.

construction des cartes de propagation. Deux catégories de méthodes déterministes sont présentées ci-dessous :

Les modèles géométriques ou multi-trajets (voir Roche [2007]). Ces modèles utilisent les lois de l'optique géométrique de Descartes (ou "lancer de rayons") en considérant un faisceau de rayons partant de l'émetteur (AP) aux récepteurs. La rencontre d'un rayon avec un obstacle donne naissance à un rayon transmis et un rayon réfléchi dont la puissance dépend du type de matériau rencontré (les coefficients d'absorption et de réflexion du matériau influent sur ces puissances). Certains modèles prennent en compte les phénomènes de diffraction. Pour cela, la puissance reçue en chaque coin de mur est calculée. Puis chaque coin de mur est considéré comme source annexe d'émission desquelles sont lancés d'autres faisceaux de rayons dont les puissances de départ ont été calculées. Un coefficient de diffraction est ensuite affecté à chacun de ces rayons en fonction de l'angle qu'il forme avec le rayon incident. Cependant, ces méthodes nécessitent beaucoup de calculs, et demande de répertorier tous les obstacles susceptibles de modifier la propagation des ondes.

Les modèles numériques. Basées sur les équations de Maxwell, ces méthodes comme les méthodes aux différences finies (voir Roche [2007]) ou le *parflow* (Roche [2007], Gorce and Ubeda [2001]), permettent elles aussi la simulation de la propagation des ondes en milieux confinés, elles sont cependant très complexes à implémenter. Roche [2007] décrit en détails la méthode *parflow* ainsi que son implémentation. Cependant, ici aussi, l'ensemble des obstacles ne peut être considéré dans le calcul, et l'influence du mobilier, par exemple, n'est pas prise en compte (alors qu'un meuble en métal, par exemple, à proximité de l'AP peut perturber l'onde WiFi de manière significative).

Ces méthodes déterministes peuvent s'avérer très utiles lors de l'agencement de l'infrastructure WiFi. Elles permettent notamment en pratique d'identifier, pour une configuration donnée, les zones de faible couverture WiFi qui pourraient poser un problème de connectivité au réseau. De telles méthodes permettent alors, pour un environnement donné, de connaitre le nombre et les emplacements optimaux d'APs nécessaire pour couvrir l'environnement. D'autres méthodes déterministes existent,

cependant, leur mise en oeuvre est très complexe et ne peut prendre en compte de manière réaliste tous les obstacles, ni les êtres humains (étant composé principalement d'eau, le corps humain a pour effet d'atténuer fortement une onde WiFi la traversant). Nous proposons dans la section 2.3, un modèle statistique qui nous permettra par la suite d'estimer, de manière empirique, les cartes de propagation des ondes.

2.3 Modélisation statistique du problème

Notons tout d'abord K l'environnement à géolocaliser appelé dans la suite espace d'états, dans ce travail, nous nous restreignons au cas où K est un compact de $\mathbb{R}^2$. Notons ℓ le nombre d'APs présents dans l'environnement. L'ensemble des APs est alors identifié à l'ensemble $\{1,\ldots,\ell\}$. Pour tout x de K, notons $Z(x) = (Z_1(x),\ldots,Z_\ell(x))$ le RSS mesuré par le terminal mobile au point x. Pour tout j de $\{1,\ldots,\ell\}$, $Z_j(x)$ représente alors le RSS relatif à l'AP j. Nous supposons alors, pour tout j de $\{1,\ldots,j\}$,

$$Z_j(x) \stackrel{\text{def}}{=} f_{\star,j}(x) + \epsilon_j \ , \tag{2.2}$$

où $f_{\star,j}$ est une fonction de K dans $\mathbb{R}$ appelée "carte de propagation moyenne pour l'AP j". Notons $f_\star = (f_{\star,1},\ldots,f_{\star,\ell})$. Nous supposons par ailleurs, que $\{\epsilon_j\}_{j=1}^{\ell}$ est un bruit de mesure indépendant, identiquement distribué (i.i.d.) de moyenne nulle et indépendant de la position x.

Nous supposons que le terminal mobile relève le RSS, à intervalle de temps régulier, et le communique au serveur. Notons alors, pour tout pas de temps $t \in \mathbb{N}$, $Y_t = (Y_{t,1},\ldots,Y_{t,\ell})$ les puissances relevées par le mobile à l'instant t (le vecteur Y_t est alors appelé "observation à l'instant t"). Pour tout $t \in \mathbb{N}$, nous désignons par $X_t \in K$ la position inconnue du terminal à l'instant t. $\{Y_t\}_{t\in\mathbb{N}}$ vérifie alors :

$$Y_t \stackrel{\text{def}}{=} f_\star(X_t) + \epsilon_t \ , \tag{2.3}$$

où le processus aléatoire $\{\epsilon_t\}_{t\in\mathbb{N}}$ est i.i.d. sur $\mathbb{R}^\ell$, indépendant de $\{X_t\}_{t\in\mathbb{N}}$, de moyenne nulle et tel que pour tout $t \in \mathbb{N}$, $\epsilon_t = \{\epsilon_{t,1},\ldots,\epsilon_{t,\ell}\}$ soit, lui aussi, i.i.d. . Dans ce modèle, les observations sont à valeurs dans $\mathbb{R}^\ell$, $\mathbb{R}^\ell$ est alors appelé espace d'observations du modèle. Nous appellerons *modèle spatio-temporel à états latents* le modèle décrit par les équations (2.3) et (2.2) lorsque les valeurs du processus $\{X_t\}_{t\in\mathbb{N}}$ sont non observées. Ce modèle est courant dans la littérature sur le sujet, Bahl and Padmanabhan [2000] discrétise l'environnement en un ensemble de cellules de sorte que K est approximé par un ensemble fini de positions $\{x_1,\ldots,x_{N_c}\}$, avec N_c le nombre total de cellules considérées. La fonction $\{f_\star(x_i)\}_{i=1}^{N_c}$ est alors estimée en effectuant une ou plusieurs mesures en chacune des cellules x_i. L'une des méthodes étudiée par Bahl and Padmanabhan [2000] consiste à construire un estimateur de la position à l'instant t, X_t, sur la base de la mesure du RSS Y_t, comme étant la cellule x_i telle que $f_\star(x_i)$ soit la plus proche, en distance euclidienne, de Y_t. Cette modélisation équivaut alors dans notre cas à supposer $K = \{x_1,\ldots,x_{N_c}\}$, que les positions X_t sont aléatoires, indépendantes de loi uniforme sur K, et que, pour tout AP $j \in \{1,\ldots,\ell\}$, $\epsilon_{t,j}$ suit une loi normale $\mathcal{N}(0,\sigma_\star^2)$. L'estimateur de la position proposée par Bahl and Padmanabhan [2000] est alors le prédicteur de maximum *a posteriori* dans ce modèle communément appelé dans ce cas le plus proche voisin. Ce modèle sera présenté dans le chapitre 4. Evennou and Marx [2006] propose une modélisation markovienne du processus $\{X_t\}_{t\in\mathbb{N}}$, comme Bahl and Padmanabhan [2000], $f_\star$ est estimé en un nombre fini de positions $\{x_1,\ldots,x_{N_c}\}$ sans pour autant discrétiser l'espace d'état K.

La modélisation markovienne des positions ajoutant de la dépendance dans la suite de positions $\{X_t\}_{t\in\mathbb{N}}$, un estimateur de X_t, construit sur la base des mesures passées $\{Y_1, \ldots, Y_t\}$ (noté dans la suite $Y_{1:t}$), est calculé par techniques de filtrage particulaire (voir le chapitre 5). Le modèle utilisé par Evennou and Marx [2006] régissant les mesures du terminal mobile est lui aussi donné par (2.2). Cependant, la valeur de $f_\star(x)$ étant estimée uniquement pour les positions x de l'ensemble $\{x_1, \ldots, x_{N_c}\}$, les techniques de filtrage particulaire classiques sont alors modifiées pour pallier à ce manque d'informations. Des modélisations similaires seront étudiées dans la partie III et le chapitre 5.

Bibliographie

P. Bahl and V.N. Padmanabhan. RADAR : An In-Building RF-Based User Location and Tracking System. In *INFOCOM*, pages 775–784, 2000.

F. Evennou and F. Marx. Advanced integration of WIFI and inertial navigation systems for indoor mobile positioning. *EURASIP J. Appl. Signal Process.*, 2006 :164–164, January 2006. ISSN 1110-8657. doi : 10.1155/ASP/2006/86706.

H. T. Friis. A Note on a Simple Transmission Formula. *Proceedings of the IRE*, 34(5) :254–256, September 1946.

J.-M. Gorce and S. Ubeda. Propagation simulation with the ParFlow method : fast computation using a multi-resolution scheme. In *Vehicular Technology Conference, 2001. VTC 2001 Fall. IEEE VTS 54th*, volume 3, pages 1603 –1607 vol.3, 2001. doi : 10.1109/VTC.2001.956469.

S.J. Pan, V.W. Zheng, Q. Yang, and D. H. Hu. Transfer Learning for Wifi-based Indoor Localization. In *AAAI-08 Workshop on Transfer Learning for Complex Task of the 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence*, pages –, 2008.

G. De La Roche. *Simulation de la propagation des ondes radio en environnement multi-trajets pour l'étude des réseaux sans fil.* PhD thesis, INSA de Lyon - INRIA Rhône-Alpes, 2007.

Chapitre 3

Éléments de théorie sur les modèles de Markov cachés

Sommaire

Dans cette section, nous achevons la description générale du modèle entamée dans la section 2.3 et fixons le cadre mathématique formel utilisé dans la suite. Dans les modèles étudiés dans cette thèse, nous supposerons que le processus $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov homogène sur l'espace d'états K. L'équation (2.3) fixe les distributions conditionnelles $Y_t|X_t = x$ pour tout x de K, de plus, les conditions imposées sur le processus $\{\epsilon_t\}_{t\in\mathbb{N}}$ impliquent que, pour tout entier naturel n et pour tous $x_0, \ldots, x_n$ éléments de K, les variables $Y_1, \ldots, Y_n$ sont indépendantes conditionnellement à l'événement $\{X_0 = x_0, \ldots, X_n = x_n\}$. Le processus bivarié $\{X_t, Y_t\}_{t\in\mathbb{N}}$ est alors un modèle de Markov caché (*hidden Markov model* ou HMM) dont la définition et les résultats principaux sont donnés dans cette section. Ces résultats sont extraits de Cappé et al. [2005].

3.1 Définitions

Définition 1 (Noyaux de transition). *Soit $(\mathbb{X}, \mathcal{X})$ et $(\mathbb{Y}, \mathcal{Y})$ deux espaces mesurables. Un noyau de transition de $(\mathbb{X}, \mathcal{X})$ à $(\mathbb{Y}, \mathcal{Y})$ est une fonction Q de $(\mathbb{X} \times \mathcal{Y}) \to [0, \infty]$, telle que*

- *pour tout $x \in \mathbb{X}$, $Q(x, \cdot)$ est une mesure positive sur $(\mathbb{Y}, \mathcal{Y})$,*
- *pour tout $A \in \mathcal{Y}$, la fonction $x \mapsto Q(x, A)$ est mesurable,*
- *pour tout $x \in \mathbb{X}$, $Q(x, \mathbb{Y}) = 1$.*

Dans le cas où $(\mathbb{Y}, \mathcal{Y}) = (\mathbb{X}, \mathcal{X})$, Q est alors un noyau de transition markovien sur $(\mathbb{X}, \mathcal{X})$. On dit que Q admet une densité par rapport à une mesure μ sur $(\mathbb{Y}, \mathcal{Y})$ s'il existe une fonction $(\mathcal{X}, \mathcal{Y})$ -

mesurable positive $q : (\mathbb{X}, \mathbb{Y}) \to [0, \infty]$ telle que :

$$Q(x, A) = \int_A q(x, y)\mu(\mathrm{d}y) \ .$$

q est alors appelé densité de transition. Soit μ une mesure positive sur l'espace $(\mathbb{X}, \mathcal{X})$, nous désignons par μQ la mesure positive sur $(\mathbb{Y}, \mathcal{Y})$ définie, pour tout $A \in \mathcal{Y}$, par

$$\mu Q(A) = \int \mu(dx) Q(x, A)$$

Si Q est un noyau de transition markovien sur $(\mathbb{X}, \mathcal{X})$, on dit qu'une mesure positive μ sur $(\mathbb{X}, \mathcal{X})$ est invariante par Q si $\mu Q = \mu$. Si de plus μ est une probabilité, alors μ est une probabilité invariante pour le noyau Q.

Définition 2 (Chaîne de Markov). *Soit $(\Omega, \mathcal{F}, F, \mathbb{P})$ un espace de probabilité filtré (de filtration F) et Q un noyau de transition Markovien sur un espace mesurable $(\mathbb{X}, \mathcal{X})$. Un processus stochastique $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov sous $\mathbb{P}$, à valeurs dans $\mathbb{X}$ et relativement à la filtration F, si $\{X_t\}_{t\in\mathbb{N}}$ est F-adaptée et, pour tout $t \in \mathbb{N}$ et $A \in \mathcal{X}$,*

$$\mathbb{P}\left(X_{t+1} \in A \mid F_t\right) = Q(X_t, A) \ .$$

Remarque 1. *Si $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov relativement à une filtration F, alors $\{X_t\}_{t\in\mathbb{N}}$ est nécessairement F-adaptée et est aussi une chaîne de Markov relativement à sa filtration naturelle. Par la suite, une chaîne de Markov par rapport à sa filtration naturelle s'appellera simplement chaîne de Markov.*

La distribution d'une chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$ est alors déterminée par la connaissance de sa distribution initiale ν (loi de X_0) et de son noyau de transition Q. Dans la suite, l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ est fixé et nous noterons $\mathbb{E}$ l'espérance sous $\mathbb{P}$.

Proposition 3.1.1. *Soit $\{X_t\}_{t\in\mathbb{N}}$ une Chaîne de Markov à valeurs dans $(\mathbb{X}, \mathcal{X})$, de distribution initiale ν et de noyau de transition Q. Pour tout entier naturel n et pour toute fonction f $\mathcal{X}^{n+1}$ mesurable de $\mathbb{X}^{n+1}$,*

$$\mathbb{E}\left(f(X_0, \ldots, X_n)\right) = \int f(x_0, \ldots, x_n)\nu(\mathrm{d}x_0) \prod_{i=1}^{n} Q(x_{i-1}, \mathrm{d}x_i) \ .$$

Définition 3 (Modèles de Markov cachés). *Soit $(\mathbb{X}, \mathcal{X})$ et $(\mathbb{Y}, \mathcal{Y})$ deux espaces mesurables. Soit Q (resp. G) un noyau de transition markovien sur $(\mathbb{X}, \mathcal{X})$ (resp. de $(\mathbb{X}, \mathcal{X})$ sur $(\mathbb{Y}, \mathcal{Y})$). Considérons le noyau de transition markovien défini sur l'espace mesurable produit $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \otimes \mathcal{Y})$ par*

$$T\left((x, y), C\right) = \int \int_C Q(x, \mathrm{d}x')G(x', \mathrm{d}y'), \quad (x, y) \in \mathbb{X} \times \mathbb{Y}, \ C \in \mathcal{X} \otimes \mathcal{Y} \ . \tag{3.1}$$

La chaîne de Markov $\{X_t, Y_t\}_{t\in\mathbb{N}}$ de noyau de transition T et de distribution initiale $\nu \otimes G$, où ν est une mesure de probabilité sur $(\mathbb{X}, \mathcal{X})$ est appelée, dans le cas où le processus $\{X_t\}_{t\in\mathbb{N}}$ n'est pas observé, chaîne de Markov cachée.

Dans la suite, nous utiliserons l'abréviation HMM (*Hidden Markov Model*) pour chaîne de Markov cachée. Nous supposerons de même que le noyau de transition G admet une densité de transition g par rapport à une mesure λ sur $\mathbb{Y}$, on dit alors que la HMM $\{X_t, Y_t\}_{t\in\mathbb{N}}$ est partiellement dominée.

3.2 Vraisemblance, lois de filtrage, lois de lissage

Les résultats de cette section ont été introduits par Baum et al. [1970] dans le cas où $\mathbb{X}$ est fini. Ils traitent en particulier des résultats de filtrage qui consistent à déterminer la loi conditionnelle de X_n ayant observé $Y_0, \ldots, Y_n$ et dont on se servira pour construire un estimateur de la position X_n. D'après (3.1), pour tout entier naturel n, toute fonction mesurable f de Y^{n+1}, et toute mesure initiale ν sur $\mathbb{X}$,

$$\mathbb{E}_\nu\left(f(Y_0,\ldots,Y_n)\right) = \int_{y_{0:n}\in\mathbb{Y}^{n+1}} f(y_0,\ldots,y_n) \int_{x_{0:n}\in\mathbb{X}^{n+1}} \nu(x_0)g(x_0,y_0) \cdot \prod_{t=1}^{n} Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)\ \lambda^{\otimes(n+1)}(\mathrm{d}y_0,\ldots,\mathrm{d}y_n)\ .$$

Nous définissons alors ce que l'on appellera la vraisemblance des observations comme la densité de probabilité de $Y_{0:n}$ relativement à la mesure $\lambda^{\otimes(n+1)}$ par,

$$L_{\nu,n}(y_{1:n}) \stackrel{\text{def}}{=} \int_{x_{0:n}\in\mathbb{X}^{n+1}} \nu(x_0)g(x_0,y_0)\cdot\prod_{t=1}^{n} Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t), \quad \forall y_{1:n}\in\mathbb{Y}^{n+1}\ . \tag{3.2}$$

La fonction $\ell_{\nu,n}$, définie par

$$\ell_{\nu,n} = \log L_{\nu,n} \tag{3.3}$$

est la fonction de log-vraisemblance des observations.

Définition 4. *Pour tous entier naturels k,l et n tels que $k \le l$, $\phi_{\nu,k:l|n}$ désigne la distribution conditionnelle de $X_{k:l}$ conditionnellement à $Y_{0:n}$. $\phi_{\nu,k:l|n}$ vérifie alors les points suivants :*

i) $\phi_{\nu,k:l|n}$ est un noyau de transition de $\mathbb{Y}^{(n+1)}$ à $\mathbb{X}^{l-k+1}$ tel que
- *$\forall A\in\mathcal{X}^{\otimes(n+1)}$, la fonction $y_{0:n}\mapsto\phi_{\nu,k:l|n}(y_{0:n},A)$ est $\mathcal{Y}^{\otimes(n+1)}$ - mesurable.*
- *$\forall y_{1:n}\in\mathbb{Y}^{n+1}$, $A\mapsto\phi_{\nu,k:l|n}(y_{0:n},A)$ est une distribution de probabilité sur $\left(\mathbb{X}^{l-k+1},\mathcal{X}^{\otimes(n+1)}\right)$.*

ii) $\phi_{\nu,k:l|n}$ vérifie que, $\mathbb{P}_\nu$ presque sûrement, pour toute fonction f mesurable bornée de $\mathbb{X}^{l-k+1}$,

$$\mathbb{E}_\nu\left(f(X_{k:l})\mid Y_{0:n}\right) = \int_{x_{k:l}\in\mathbb{X}^{l-k+1}} \phi_{\nu,k:l|n}(Y_{0:n},\mathrm{d}x_{k:l})\ .$$

Pour tout entier n positif, nous définissons en particulier
- $\phi_{\nu,0:n|n}$, la distribution de lissage jointe,
- pour tout entier k tel que $0\le k<n$, $\phi_{\nu,k|n}$, la distribution de lissage marginale,
- $\phi_{\nu,n|n}$, la distribution de filtrage.
- pour tout entier $p>0$, $\phi_{\nu,n+p|n}$, la distribution de prédiction à p pas.

Proposition 3.2.1 (Proposition 3.1.4 de Cappé et al. [2005]). *Soit n une entier naturel, et $y_{0:n}\in\mathbb{Y}^{n+1}$ tel que $L_{\nu,n}(y_{0:n})>0$. La distribution de lissage jointe satisfait, pour toute fonction $\mathcal{X}^{\otimes(n+1)}$-mesurable bornée f :*

$$\phi_{\nu,0:n|n}(y_{0:n},f) = L_{\nu,n}(y_{0:n})^{-1}\int_{x_{0:n}\in\mathbb{X}^{n+1}} f(x_{0:n})\nu(\mathrm{d}x_0)g(x_0,y_0)\prod_{t=1}^{n}Q(x_{t-1},\mathrm{d}x_t)g(x_t,y_t)\ .$$

De la même manière, pour tout indice $p \geq 0$, et toute fonction f $\mathcal{X}^{\otimes(n+p+1)}$-mesurable bornée,

$$\phi_{\nu,0:n+p|n}(y_{0:n}, f) = \int_{x_{0:n+p} \in \mathbb{X}^{n+p+1}} f(x_{0:n+p}) \phi_{\nu,0:n|n}(y_{0:n}, \mathrm{d}x_{0:n}) \prod_{t=n+1}^{n+p} Q(x_{t-1}, \mathrm{d}x_t) \ .$$

La Proposition 3.2.1 permet notamment d'exprimer, par marginalisation de l'équation (3.4), les distributions de lissage marginales $\phi_{\nu,k|n}$ pour tout indice $0 \leq k \leq n$. Pour toute fonction f, $\mathcal{X}$-mesurable bornée et pour tout $y_{0:n} \in \mathbb{Y}^{n+1}$ tel que $L_{\nu,n}(y_{0:n}) > 0$,

$$\phi_{\nu,k|n}\left(y_{0:n}, f\right) = \int_{x_{0:n} \in \mathbb{X}^{n+1}} f(x_k) \phi_{\nu,0:n|n}(y_{0:n}, \mathrm{d}x_{0:n}) \tag{3.4}$$

Remarque 2. *Dans la suite, nous effectuons les simplifications de notation suivantes.*

- *La dépendance en les observations $y_{1:n}$ des quantités $\phi_{\nu,k:l|n}$, $0 \leq k \leq l$ sera désormais implicite de sorte que pour toute fonction f de $\mathbb{X}^{l-k+1}$, mesurable bornée, $\phi_{\nu,k:l|n}(f)$ représente $\phi_{\nu,k:l|n}(y_{0:n}, f)$. De la même manière, l'écriture de la vraisemblance $L_{\nu,n}(y_{0:n})$ sera désormais $L_{\nu,n}$.*
- *Pour tout indice n et toute mesure initiale ν, nous noterons $\phi_{\nu,n} = \phi_{\nu,n|n}$.*

3.3 Décomposition *Forward-Backward*

L'objectif de cette section est de fournir une technique permettant de calculer efficacement les lois de lissage et de filtrage $\phi_{\nu,k|n}$ pour tout indice $0 \leq k \leq n$. Cette technique, intitulée "décomposition *Forward-backward*" permet d'exprimer les distributions de lissage $\{\phi_{\nu,k|n}\}_{0 \leq k \leq n}$, en fonction de deux quantités calculables récursivement dont voici une définition.

Définition 5 (Variables *Forward* et *Backward*)**.** *Soit un indice $k \leq n$ et ν une mesure de probabilité sur $(\mathbb{X}, \mathcal{X})$, alors la distribution de filtrage $\phi_{\nu,k}$ est aussi appelée noyau de transition forward. La fonction positive $\beta_{k|n}$ définie, pour tout $x \in \mathbb{X}$, par :*

$$\beta_{k|n}(y_{k+1:n}, x) \stackrel{\text{def}}{=} \frac{L_{\nu,k}}{L_{\nu,n}} \int_{x_{k+1:n} \in \mathbb{X}^{n-k}} Q(x, \mathrm{d}x_{k+1}) g(x_{k+1}, y_{k+1}) \prod_{t=k+2}^{n} Q(x_{t-1}, \mathrm{d}x_t) g(x_t, y_t) \ , \tag{3.5}$$

est appelée fonction backward.

Par convention, le produit $\prod_{t=k+2}^{n} Q(x_{t-1}, \mathrm{d}x_t) g(x_t, y_t)$ est égale à 1 pour $k = n-1$. $\beta_{n|n}$ est défini comme la fonction constante égale à $\frac{L_{\nu,k}}{L_{\nu,n}}$ sur $\mathbb{Y}^{n-k} \times \mathbb{X}$.

Remarque 3. *Comme lors de la simplification de notations opérée Remarque 2, la dépendance en $y_{1:n}$ de $\beta_{k|n}(y_{k+1:n}, x)$ sera désormais implicite.*

En développant l'expression de $\phi_{\nu,n|n}$ dans l'équation (3.4), la distribution de lissage $\phi_{\nu,k|n}$ peut s'exprimer en fonction des quantités $\phi_{\nu,k}$ et $\beta_{k|n}$ de la manière suivante :

$$\phi_{\nu,k|n}(f) = \int_{x \in \mathbb{X}} f(x) \phi_{\nu,k}(\mathrm{d}x) \beta_{k|n}(x) \tag{3.6}$$

Remarque 4. *L'introduction des quantités Forward et Backward pour le calcul des lois de lissage, ainsi que la méthode de calcul récursive de ces quantités (algorithme forward-backward) ont été introduits en 1970 par Baum et al. [1970] dans le cas où $\mathbb{X}$ est fini. Notre définition des quantité forward et backward $\phi_{\nu,k}$ et $\beta_{k|n}$ correspondent à des versions normalisées de ces quantités.*

La proposition 3.3.1 reprend la proposition 3.2.5 de Cappé et al. [2005], elle permet de calculer de manière récursive les quantités $\phi_{\nu,k}$ et $\beta_{k|n}$.

Proposition 3.3.1 (Récursion *Forward- Backward*)**.** *Les mesures de filtrage peuvent être obtenues, pour toute fonction mesurable bornée f de $(\mathbb{X}, \mathcal{X})$, de la manière suivante : récursivement, pour $k = 1, \ldots, n$,*

$$\begin{aligned} c_{\nu,k} &= \int_{(x,x')\in\mathbb{X}^2} \phi_{\nu,k-1}(\mathrm{d}x) Q(x, \mathrm{d}x') g(x', y_k) \ , \\ \phi_{\nu,k}(f) &= c_{\nu,k}^{-1} \int_{(x,x')\in\mathbb{X}^2} f(x') \phi_{\nu,k-1}(\mathrm{d}x) Q(x, \mathrm{d}x') g(x', y_k) \ , \end{aligned} \tag{3.7}$$

avec, comme condition initiale,

$$\begin{aligned} c_{\nu,0} &= \int_{(x)\in\mathbb{X}} g(x, y_0) \nu(\mathrm{d}x), \\ \phi_{\nu,0}(f) &= c_{\nu,0}^{-1} \int_{x\in\mathbb{X}} f(x) g(x, y_0) \nu(\mathrm{d}x) \ . \end{aligned}$$

Les fonctions Backward sont obtenues par la formule récursive descendante suivante : pour tout $x \in \mathbb{X}$ et tout $k \in \{0, \ldots, n-1\}$,

$$\beta_{k|n}(x) = c_{\nu,k+1}^{-1} \int_{x'\in\mathbb{X}} Q(x, \mathrm{d}x') g(x', y_{k+1}) \beta_{k+1|n}(x'), \tag{3.8}$$

avec comme condition initiale, $\beta_{n|n}(x) = 1$.

3.4 Algorithme EM pour l'inférence dans les modèles de Markov cachés complètement dominés.

Nous avions supposé dans la section 3.1 que le noyau de transition G possédait une densité de transition notée g relativement à une mesure de probabilité λ sur $(\mathbb{Y}, \mathcal{Y})$. Dans cette section, nous supposons également que le noyau markovien Q possède une densité de transition q par rapport à une mesure de probabilité μ sur $(\mathbb{X}, \mathcal{X})$, on dit alors que la chaîne de Markov cachée $(X_t, Y_t)_{t\in\mathbb{N}}$ est complètement dominée. On suppose de plus que notre modèle est paramétrique : soit $n \in \mathbb{N}$, la densité jointe de $(X_t, Y_t)_{0\leq t\leq n}$ par rapport à la mesure $\mu^{\otimes(n+1)} \otimes \lambda^{\otimes(n+1)}$ appartient à une famille paramétrique $\{f_n(\cdot;\theta)\}_{\theta\in\Theta}$ où la notation $(\cdot;\theta)$ indique la dépendance en θ et, pour tout $x_{0:n} \in \mathbb{X}^{n+1}$, $y_{0:n} \in \mathbb{Y}^{n+1}$ et $\theta \in \Theta$, $f_n(\cdot\ ;\theta)$ est donné par :

$$f_n(x_{0:n}, y_{0:n}; \theta) = \nu(x_0; \theta) g(x_0, y_0; \theta) \prod_{t=1}^{n} q(x_{t-1}, x_t; \theta) g(x_t, y_t; \theta) \ . \tag{3.9}$$

Θ est alors appelé espace de paramètres du modèle. On supposera que Θ est un ouvert de $\mathbb{R}^{d_\theta}$.

Notons $L_n(\theta)$ la vraisemblance des observations $y_{1:n}$ sous le paramètre θ, et $\ell_n(\theta)$ la log-vraisemblance. Sous l'hypothèse que $L_n(\theta) > 0$ pour tout $\theta \in \Theta$, définissons pour tout couple de paramètres $(\theta, \theta') \in \Theta^2$, $\mathcal{Q}(\theta; \theta')$, appelée quantité intermédiaire, par :

$$\mathcal{Q}(\theta; \theta') \stackrel{\text{def}}{=} \frac{1}{L_n(\theta')} \int \log\left(f_n(x_{0:n}, y_{0:n}; \theta)\right) f_n(x_{0:n}, y_{0:n}; \theta') \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n}) \,. \tag{3.10}$$

où l'on utilise la convention $0 \log 0 = 0$.

Proposition 3.4.1. *Si ∇_θ désigne la fonction gradient par rapport à la variable θ, supposons que, pour tout $(\theta, \theta') \in \Theta^2$,*

$$\int \left| \nabla_\theta \log \frac{f_n(x_{0:n}, y_{0:n}; \theta)}{L_n(\theta)} \right| f_n(x_{0:n}, y_{0:n}; \theta') \mu^{\otimes(n+1)}(\mathrm{d}x_{0:n})$$

est fini pour tout $(\theta, \theta') \in \Theta^2$, alors,

$$\ell_n(\theta) - \ell_n(\theta') \geq \mathcal{Q}(\theta; \theta') - \mathcal{Q}(\theta'; \theta') \,,$$

où l'inégalité est stricte, sauf dans le cas où $f_n(\cdot; \theta)/L_n(\theta)$ et $f_n(\cdot; \theta')/L_n(\theta')$ sont égaux presque sûrement. Si de plus les fonctions $\theta \mapsto \mathcal{Q}(\theta; \theta')$ et L_n sont différentiables, alors :

$$\nabla_\theta \ell_n(\theta') = \nabla_\theta \mathcal{Q}(\theta; \theta')|_{\theta=\theta'}$$

La proposition 3.4.1 assure que, pour un paramètre θ' donné, tout paramètre θ augmentant la quantité intermédiaire par rapport à la valeur $\mathcal{Q}(\theta'; \theta')$ augmente aussi la valeur de la log-vraisemblance. Ainsi, l'algorithme *Expectation-Maximisation* (EM) proposé par Dempster et al. [1977], construit séquentiellement une suite d'estimateurs $\{\theta^i\}_{i \geq 1}$ sur la base d'une valeur initiale θ_0 permettant d'augmenter, à chaque pas la valeur de la log-vraisemblance :

Étape E Détermine la fonction $\mathcal{Q}(\theta; \theta^i)$,

Étape M Choisit θ^{i+1} un des maximiseur de $\theta \mapsto \mathcal{Q}(\theta; \theta^i)$.

La proposition 3.4.1 garantit alors la croissance de la suite $\{L_n(\theta_i)\}_{i \geq 0}$ et, dans le cas où l'algorithme "stagne" sur un paramètre $\theta_\star$, la fonction $\theta \mapsto \mathcal{Q}(\theta; \theta_\star)$ est maximale en $\theta_\star$. Nécessairement, $\nabla_\theta \ell_n(\theta_\star) = 0$ et $\theta_\star$ est alors un point stationnaire de la vraisemblance. L'algorithme EM peut être utilisé dans de nombreux modèles dès lors que l'on cherche à maximiser une vraisemblance. L'expression de la quantité intermédiaire (3.10) est propre aux chaînes de Markov cachées, une description plus générale de l'algorithme EM peut être trouvée dans Cappé et al. [2005].

D'après l'équation (3.9), si $Y_{0:n}$ représente le vecteur aléatoire des observations jusqu'au temps n, l'expression de $\mathcal{Q}$ peut être réécrite de la manière additive suivante :

$$\begin{aligned} \mathcal{Q}(\theta, \theta') = \mathbb{E}_{\theta'}\left[\log \nu(X_0; \theta) \mid Y_{0:n}\right] + \sum_{t=1}^{n} \mathbb{E}_{\theta'}\left[\log q(X_{t-1}, X_t; \theta) \mid Y_{0:n}\right] \\ + \sum_{t=0}^{n} \mathbb{E}_{\theta'}\left[\log g(X_t, Y_t; \theta) \mid Y_{0:n}\right] \,. \end{aligned}$$

L'étape E de l'algorithme EM requière donc les distributions de lissages $\{\phi^{\theta'}_{\nu,t|n}\}_{0\leq t\leq n}$, sous le paramètre θ' ainsi que les distributions de lissages bivariées $\{\phi^{\theta'}_{\nu,t:t+1|n}\}_{0\leq t\leq n-1}$ pour la partie transition de la quantité intermédiaire. Le modèle étant complètement dominé, ces distributions possèdent une densité de probabilité toujours notée $\phi^{\theta'}_{\nu,t|n}$ (resp. $\phi^{\theta'}_{\nu,t:t+1|n}$) par rapport à la mesure μ sur $\mathbb{X}$ (resp. $\mu\otimes\mu$ sur $\mathbb{X}^2$). Les densités $\phi^{\theta'}_{\nu,t|n}$ peuvent alors être calculées à l'aide de l'algorithme *Forward - Backward* décrit dans la section 3.3, tandis que les densités de lissages bivariées $\phi^{\theta'}_{\nu,t:t+1|n}$ peuvent être déterminées grâce aux lois de lissages $\phi^{\theta'}_{\nu,t|n}$ suivant l'équation suivante : pour tous x et x' de $\mathbb{X}$,

$$\phi^{\theta'}_{\nu,t:t+1|n}(x,x') = \frac{\phi^{\theta'}_{\nu,t|n}(x)q(x,x';\theta)g(x',Y_{t+1};\theta')}{\int \phi^{\theta'}_{\nu,t|n}(x'')q(x'',x''';\theta')g(x''',Y_{t+1};\theta')\mu(\mathrm{d}x'')\mu(\mathrm{d}x''')} \ .$$

3.5 Modélisation générale du processus $\{X_t, Y_t\}_{t\in\mathbb{N}}$

Comme mentionné au début de la section 3, nous supposons désormais que la suite de positions $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov sur $(\mathbb{X},\mathcal{X}) = (K,\mathcal{K})$, où $K\subset\mathbb{R}^2$ et $\mathcal{K}$ est une σ-algèbre sur K, de noyau de transition noté Q. On supposera par ailleurs que le noyau de transition Q admet une densité de transition $q : K\times K\to\mathbb{R}_+$ par rapport à une mesure μ représentant, soit la mesure de Lebesgue lorsque K est supposé d'intérieur non vide dans $\mathbb{R}^2$, soit la mesure de comptage lorsque K est supposé fini. Nous supposons que le bruit de mesure $\{\epsilon_t\}_{t\geq 0}$ de l'équation (2.3) est i.i.d. de loi commune $\mathcal{N}\left(0,\sigma_\star^2 I_\ell\right)$, où $\sigma_\star^2>0$ et où I_ℓ représente la matrice identité de taille ℓ. D'après l'équation (2.3), si G est le noyau de transition de $(K,\mathcal{K})$ à $(\mathbb{R}^\ell,\mathcal{B}(\mathbb{R}^\ell))$, où $\mathcal{B}(\mathbb{R}^\ell)$ désigne les boréliens de $\mathbb{R}^\ell$, déterminant la distribution de $\{Y_t\}_{t\geq 0}$ conditionnellement à $\{X_t\}_{t\geq 0}$, alors G possède une densité de transition notée g, par rapport à la mesure de Lebesgue λ sur $\mathbb{Y}=\mathbb{R}^\ell$, définie, pour tout $x\in K$, pour tout $y\in\mathbb{R}^\ell$ par

$$g(x,y) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma_\star^2}^\ell}\prod_{j=1}^{\ell}\exp\left(-\frac{1}{2\sigma_\star^2}\|y-f_\star(x)\|^2_{\mathbb{R}^\ell}\right) \ ,$$

où $\|\cdot\|_{\mathbb{R}^\ell}$ désigne la norme euclidienne sur $\mathbb{R}^\ell$. $\{X_t,Y_t\}_{t\in\mathbb{N}}$ est donc une HMM de noyau de transition T, a densité de transition $t : (K\times\mathbb{R}^\ell)^2\to\mathbb{R}_+$ par rapport à $\mu\otimes\lambda$ définie, pour tous (x,y) et (x',y') de $K\times\mathbb{R}^\ell$, par

$$t\left((x,y),(x',y')\right) \stackrel{\text{def}}{=} q(x,x')g(x',y') \ . \tag{3.11}$$

Bibliographie

L.E. Baum, T.Petrie, G.Soules, and N.Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41 (1) :164–171, 1970. ISSN 00034851. doi : 10.2307/2239727.

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2 ; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series B*, 39(1) :1–38, 1977.

Deuxième partie

Localisation dans des modèles spatio-temporels à états latents avec calibration préalable des cartes de propagation.

Chapitre 4

Discrétisation grossière de l'environnement.

Sommaire

Le premier modèle étudié utilise une discrétisation grossière de l'espace d'états comme illustré sur la figure 4.1. Nous supposons donc l'existence d'un entier naturel N_c, et d'éléments $x_1, \dots, x_{N_c}$ de $\mathbb{R}^B$ (centres des cellules de la grille) tels que $K = \{x_1, \dots, x_{N_c}\}$. Afin d'alléger les notations futures, la notation $K = 1 : N_c$ sera utilisée.

4.1 Généralités

On considère dans cette section un espace d'états $\mathbb{X}$ fini muni de la σ-algèbre $\mathcal{X} = \mathcal{P}(\mathbb{X})$ où $\mathcal{P}(\mathbb{X})$ représente l'ensemble de parties de $\mathbb{X}$. Par la suite nous adopterons la notation $\mathbb{X} = \{1, \dots, N_c\} = 1 : N_c$. Le noyau de transition Q de la chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$ est alors déterminé par sa matrice dite de passage, notée aussi Q, de taille $N_c \times N_c$ et définie par :

$$\forall (i,j) \in (1 : N_c)^2,\ Q_{i,j} = \mathbb{P}(X_1 = j | X_0 = i)\ .$$

La densité de transition q est alors relative à la mesure de comptage μ sur $(\mathbb{X}, \mathcal{X})$ et est définie, pour tout $(i,j) \in (1 : N_c)^2$ par $q(i,j) = Q_{i,j}$. La fonction $f_\star$ est déterminée par la valeur qu'elle prend en chacun des éléments $x_1, \dots, x_{N_c}$ de K, que l'on notera par la suite $m_{\star,1}, \dots, m_{\star,N_c}$.

Dans le cas où les valeurs des paramètres Q, $m_{\star,1}, \dots, m_{\star,N_c}$ et $\sigma_\star^2$ sont connues, l'algorithme *Forward-Backward* permet la détermination des densités de lissage $\{\phi_{\nu,k|n}(x)\}_{x\in 1:N_c}$. Les densités de filtrage $\{\phi_{\nu,n}(x)\}_{x\in 1:N_c}$ peuvent alors facilement être calculées grâce à la partie *Forward* de l'algorithme. Le prédicteur de maximum a posteriori $\widehat{X}_n$ de X_n construit sur la base des observations $Y_{0:n}$ jusqu'au temps n est donc défini comme le x de $1 : N_c$ maximisant $\{\phi_{\nu,n}(x)\}_{x\in 1:N_c}$:

$$\widehat{X}_n \overset{\text{def}}{=} \underset{x\in 1:N_c}{\operatorname{argmax}}\, \phi_{\nu,n}(x)\ . \tag{4.1}$$

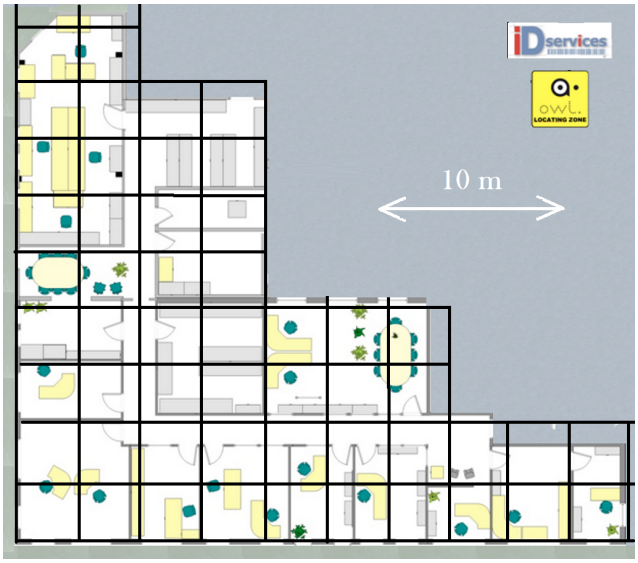


FIGURE 4.1 – Discrétisation grossière d'un environnement de bureau.

L'algorithme de Viterbi (voir Rabiner [1990]), construit sur la base de l'algorithme *Forward-Backward* permet, quant à lui, de calculer le chemin de $\widehat{X}_{0:n}$ maximisant la loi de lissage jointe $\phi_{\nu,0:n|n}$:

$$\widehat{X}_{0:n} \overset{\text{def}}{=} \underset{x_{0:n}\in(1:N_c)^{n+1}}{\operatorname{argmax}} \phi_{\nu,0:n|n}(x_{0:n}) \, .$$

Dans le cas où les valeurs des paramètres Q, $m_{\star,1},\dots,m_{\star,N_c}$ et $\sigma_\star^2$ sont inconnus, l'algorithme EM de la section 3.4 permet la construction d'un estimateur de maximum de vraisemblance pour ces paramètres. Cependant, dans le cas où $\mathbb{X}$ est fini ces estimateurs sont soumis au problème classique dans les HMM de *label switching*. Le phénomène de *label switching* est un problème de non identifiabilité des états cachés. Si l'on pose $m_\star = (m_{\star,1},\dots,m_{\star,N_c})$ et $\theta_\star = (Q, m_\star, \sigma_\star^2)$, alors, pour tout entier naturel n et pour toute permutation p de $1 : N_c$, si $\theta_p = (Q\circ p, m_\star \circ p, \sigma_\star^2)$ où $(Q\circ p)_{i,j} = Q_{p(i),p(j)}$ et $(m_\star \circ p)_i = m_{\star,p(i)}$ alors il y a égalité entre la vraisemblance sous le paramètre $\theta_\star$ et la vraisemblance sous le paramètre θ_p :

$$L_n(\theta_p) = L_n(\theta_\star)$$

L'algorithme EM a donc pour objectif d'approcher $\theta_\star$ à permutation près. Dans certaines applications le *label switching* n'est pas un inconvénient, cependant, dans le cadre de la géolocalisation, chaque élément i est associé à une position physique x_i et le *label switching* a alors pour effet de mélanger de manière complètement inconnue les positions x_i. L'utilisation de l'algorithme EM dans ce cadre n'est donc pas étudiée dans cette section, nous verrons dans le chapitre 5 et dans la section 7 d'autres modèles prenant en compte l'aspect spatial de $f_\star$ de manière à s'affranchir du problème de *label switching* en rendant le modèle identifiable.

4.2 Application

Afin de mettre en oeuvre l'algorithme *Forward Backward* et de construire les estimateurs $\widehat{X}_n$ ou $\widehat{X}_{0:n}$, une campagne préliminaire de mesures est effectuée avec l'objectif de déterminer un estimateur de $m_\star$ et de $\sigma^2_\star$. Ainsi, pour chaque position x_i, $i = 1, \ldots, N_c$, de K, une série de N mesures $Z_{i,1}, \ldots, Z_{i,N}$ est effectuée. À i fixé dans $1 : N_c$, la valeur de $m_{\star,i}$ est alors approchée par la moyenne des $Z_{i,1}, \ldots, Z_{i,N}$ en définissant :

$$\widehat{m}_i \stackrel{\text{def}}{=} \frac{1}{N} \sum_{k=1}^{N} Z_{i,k} \ .$$

De manière similaire, $\sigma^2_\star$ est approché par $\widehat{\sigma}^2$ défini par :

$$\widehat{\sigma}^2 \stackrel{\text{def}}{=} \frac{1}{N_c(N-1)} \sum_{i=1}^{N_c} \sum_{k=1}^{N} (Z_{i,k} - \widehat{m}_i)^2 \ . \tag{4.2}$$

La distribution du processus $\{X_t, Y_t\}_{t \in \mathbb{N}}$ est alors achevée en fixant les valeurs des transitions $Q_{i,j}$. Pour cela, nous déterminons la distance maximale d_{max} que le terminal mobile peut parcourir entre deux mesures t et $t+1$. Cette distance dépend à la fois de la vitesse v du terminal mobile (environ $3m/s$ si le terminal est porté par un humain) et de l'intervalle de temps Δ_t entre deux mesures. Ainsi, $d_{max} = v\Delta_t$. De plus, en fonction de l'architecture du bâtiment à géolocaliser, le passage d'une position i à une position j peut être empêché par un obstacle (un mur par exemple). Pour tout $i \in K$, nous définissons n_i comme le nombre d'éléments j de K tels que i et j ne sont pas séparés par un obstacle et tels que $d(x_i, x_j) \leq d_{max}$, ou d désigne ici la distance euclidienne dans $\mathbb{R}^2$.

Nous définissons alors $\widehat{Q}$ un estimateur de "bon-sens" de la matrice de transition $Q_\star$ par : pour tout i et j dans K,

$$\widehat{Q}_{i,j} = \begin{cases} 0 & \text{si } i \text{ et } j \text{ sont séparés par un obstacle,} \\ 0 & \text{si } d(x_i, x_j) > d_{max}, \\ 1/n_i & \text{sinon.} \end{cases}$$

La précision de l'estimateur $\widehat{X}_n$ définit par Équation (4.1), obtenue après un découpage de l'environnement en cellules de taille 2,50 mètres (voir la figure 4.1) est de l'ordre de 3 à 4 mètres en moyenne (le nombre de point d'accès utilisés étant $\ell = 9$). La précision obtenue est comparable à celle de l'algorithme Radar proposé dans Bahl and Padmanabhan [2000] qui utilise la méthode du plus proche voisin qui équivaut à choisir $\widehat{Q}_{i,j} = 1/N_c$ pour tout (i, j) de $(1 : N_c)^2$.

Bibliographie

P. Bahl and V.N. Padmanabhan. RADAR : An In-Building RF-Based User Location and Tracking System. In *INFOCOM*, pages 775–784, 2000.

R.L. Rabiner. Readings in speech recognition. chapter A tutorial on hidden Markov models and selected applications in speech recognition, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4.

Chapitre 5

Discrétisation fine de l'environnement et prise en compte de l'aspect spatial de la propagation des signaux

Sommaire

Dans le chapitre précédent une discrétisation grossière de l'environnement est effectuée de manière à ce que la fonction $f_\star$ puisse être estimée grâce à une campagne de mesures préalable en chacune des cellules issues de la discrétisation. Cependant, cette approche discrète d'un problème, à la base continu, constitue une approximation que les algorithmes d'inférence ou de localisation les plus développés ne pourront rattraper. En effet, la modélisation du déplacement du terminal (le modèle de transition) est très approximative puisqu'elle suppose que le terminal "saute" de cellule en cellule. L'objectif de ce chapitre est de s'approcher du modèle de base continu en diminuant considérablement le pas de discrétisation (en cellules de moins de 25cm de côté), comme illustré sur la figure 5.1. L'espace d'états K est alors fini, et la fonction $f_\star$ doit alors, comme dans la section précédente, être estimée en chacune des positions x de l'environnement K. Cependant, une campagne de mesures préalable en chacune des cellules de la grille K obtenue après une discrétisation aussi fine est inconcevable, nous proposerons alors une modélisation de l'évolution spatiale de $f_\star$ permettant de s'affranchir

FIGURE 5.1 – Discrétisation fine de l'environnement

de ce problème. Nous présenterons une technique inspirée des techniques de krigeage (voir Cressie [1993], Guyon [2007]) extrapolant les mesures effectuées en certaines positions de l'environnement à l'ensemble des cellules de la grille. De même, les techniques d'estimation propres aux HMM à espace d'états fini ne pourront être mises en oeuvre dans ce chapitre, leur complexité dépendant linéairement du nombre d'états, le temps de calcul serait alors dépendant de la taille de la grille. Nous proposerons alors une adaptation des filtres particulaire au cas où l'espace d'états est discret. Nous présenterons alors les résultats en terme de précision de la géolocalisation et d'estimation de la fonction $f_\star$ obtenus grâce à cette méthode.

Ici, un estimateur de $f_\star$ est construit sur la base d'une campagne de mesures préalable en un nombre restreint de points de la grille que l'on notera S. Ainsi, en chaque point s de S est effectuée une série de N mesures de la puissance des signaux, que l'on notera $Z(s,1),\ldots,Z(s,N)$. Ces mesures seront désignées par mesures d'entrainement.

Les travaux décrits dans cette partie ont permis la mise en place d'un démonstrateur, sous la forme d'une application logicielle de type client-serveur. Cette application permet de localiser, en temps réel, un terminal mobile envoyant régulièrement, au serveur de localisation, les informations de puissances des signaux WiFi qu'il reçoit. Les performances de ce démonstrateur ont été testées dans un environnement de type entrepôt, les résultats de ces tests sont donnés dans la section 5.3.2.

5.1 Détermination du modèle d'émission

Les techniques de krigeage sont des méthodes d'estimation géostatistiques introduites par l'ingénieur Daniel Gerhardus Krige dans les années 50 pour des applications minières (voir Krige [1951]). Le krigeage permet l'étude des phénomènes spatiaux et possède des applications dans des domaines tels que la météorologie ou les sciences environnementales. Notre objectif ici consiste à s'inspirer des techniques de krigeage en émettant une hypothèse de gaussianité sur la fonction $f_\star$, vue comme processus spatial.

Nous cherchons à construire un interpolateur linéaire de $f_\star$ sur K, construit sur la base d'observations $\{Z(s,i)\}_{i\in 1:N,\ s\in S}$ tel que, pour tout s de S et pour tout i, $Z(s,i) = f_\star(s) + \epsilon_{s,i}$ où

$\{\epsilon_{s,i}\}_{i\in 1:N,\ s\in S}$ est i.i.d. de loi commune $\mathcal{N}(0,\sigma^2)$. Pour cela, on suppose que $f_\star$ est un processus Gaussien sur $\mathbb{R}^2$:

Définition 6 (Processus gaussien). *Un processus aléatoire Z indexé par un ensemble* $\mathbb{X}$ *est un processus gaussien si, pour toute partie finie* $S \subset \mathbb{X}$*, et pour toute suite réelle* $\{a_s\}_{s\in S}$*,* $\sum_{s\in S} a_s Z_s$ *est une variable gaussienne.*

Si Z est à valeurs réelles et si la matrice de covariance Σ_S de $\mathrm{Z}_S = \{\mathrm{Z}_s\}_{s\in S}$, définie, pour tous s et s' de S, par $\Sigma_S(s,s') = \mathrm{cov}(\mathrm{Z}_s, \mathrm{Z}_{s'})$ est inversible, Z_S possède une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^{|S|}$ définie, pour toute suite réelle $z_S = \{z_s\}_{z\in S}$, par :

$$p_S(z_S) = \sqrt{2\pi}^{-|S|} \det(\Sigma_S)^{-1/2} \exp\left(-\frac{1}{2}(z_S-\mu_S)^T \Sigma_S^{-1}(z_S-\mu_S)\right) ,$$

où $\mu_S = \mathbb{E}(\mathrm{Z}_S)$, et z^T désigne la transposée du vecteur z.

Nous supposons que le processus gaussien $f_\star$, à valeurs dans $\mathbb{R}^\ell$, possède une espérance $\mu_\star$ paramétrée de telle sorte que $\mu_\star$ représente la propagation moyenne des ondes WiFi dans le bâtiment, dictée par la formule de Friis (2.1). Posons tout d'abord, pour tout $j \in \{1\ldots,\ell\}$, $O_j \in \mathbb{R}^2$ la position supposée connue de l'AP j. Nous supposons alors l'existence de deux paramètres $c^\star_{1,j}$ et $c^\star_{2,j}$ tels que la j-ème composante de $\mu_\star$, notée $\mu_{\star,j}$, soit définie de la manière suivante : pour tout x dans $\mathbb{R}^2$,

$$\mu_{\star,j}(x) \stackrel{\text{def}}{=} c^\star_{1,j} + c^\star_{2,j} \log(\|O_j - x\|_{\mathbb{R}^2}) ,$$

où log désigne le logarithme népérien. La formule de Friis fournit une valeur déterministe pour les paramètres $c^\star_{1,j}$ et $c^\star_{2,j}$, mais ces paramètres correspondent à la propagation des ondes en champ libre, nous supposerons donc ces paramètres inconnus par la suite. Supposons la décomposition suivante sur $f_\star$: pour tout x de $\mathbb{R}^2$,

$$f_\star(x) = \mu_\star(x) + \delta_\star(x) , \tag{5.1}$$

où $\delta_\star = \{\delta_{\star,j}\}_{j=1}^\ell$ est une collection de processus gaussiens indépendants de moyennes nulles. Cette modélisation de la propagation des ondes à l'intérieur des bâtiments peut alors être considérée comme semi-déterministe, $\mu_\star$ représentant la propagation moyenne et déterministe de l'onde, basée sur un modèle physique de propagation et $\delta_\star$ regroupant les perturbations de la propagation dues aux obstacles, comme la réfraction ou la diffraction, présentés dans la section 2.2.

Nous supposons par ailleurs que le processus $\delta_\star$ est stationnaire au second ordre et isotrope : pour tout $j = 1,\ldots,\ell$, il existe une fonction réelle C_j sur $\mathbb{R}_+$, appelée fonction de covariance stationnaire, isotrope, telle que, pour tous x et x' de $\mathbb{R}^2$,

$$\mathrm{cov}(\delta_{\star,j}(x), \delta_{\star,j}(x')) = C_j(\|x-x'\|_{\mathbb{R}^2}) .$$

En 2006, Ferris et al. [2006] utilise une modélisation similaire de la propagation des ondes WiFi par des processus gaussiens, en supposant le processus $f_\star$ stationnaire d'ordre 2 et isotrope de moyenne nulle.

La fonction d'autovovariance de $\delta_{\star,j}$ est donc invariante par isométrie de $\mathbb{R}^2$ et, pour tout $h > 0$, le processus des accroissements $I^h \stackrel{\text{def}}{=} \left\{\delta_{\star,j}(x) - \delta_{\star,j}(x+hu);\ x\in\mathbb{R}^2,\ u\in\mathbb{R}^2,\ \|u\|_{\mathbb{R}^2}=1\right\}$ est

stationnaire d'ordre 2. La fonction de covariance C_j peut alors être définie, pour tout $h > 0$, par

$$\begin{aligned} C_j(h) &= \mathrm{cov}(\delta_{\star,j}(x), \delta_{\star,j}(x+hu)) \\ &= \mathbb{E}\left(\delta_{\star,j}(x)\delta_{\star,j}(x+hu)\right) \ , \end{aligned}$$

avec $x \in \mathbb{R}^2$ et $u \in \mathbb{R}^2$ tel que $\|u\|_{\mathbb{R}^2} = 1$ fixés, quelconques.

Finalement, nous introduisons une hypothèse supplémentaire sur la fonction d'autocovariance : il existe deux constantes strictement positives v_1 et v_2 telles que, pour tout $h > 0$,

$$C_j(h) \stackrel{\text{def}}{=} v_1 \exp\left(-\frac{h^2}{2v_2}\right) \ . \tag{5.2}$$

On dit alors que le modèle de covariance du processus δ_j est Gaussien, les constantes v_1 et v_2 seront appelées dans la suite *hyperparamètres de covariance*

5.1.1 Prédiction de $f_\star$ à hyperparamètres de covariance et variance σ^2 connus

Le but de cette section est la construction d'un estimateur $\widehat{f}(x)$ de $f_\star(x)$ pour tout x de la grille K construit sur la base des mesures d'entrainement $\{Z(s,i)\}_{s\in S,\ i=1:N}$. Nous supposons ici que $f_\star$ est à valeurs réelles et oublions volontairement les indices relatifs aux points d'accès, l'extension au cas où $f_\star$ est à valeurs dans $\mathbb{R}^\ell$ se fait en raisonnant composante par composante. Nous noterons alors Σ la matrice d'autocovariance du processus $\delta_\star$ que l'on suppose inversible (condition vérifiée dès lors que v_1 et v_2 sont strictement positifs). Introduisons tout d'abord les notations suivantes :

$$\begin{aligned} LD &\stackrel{\text{def}}{=} \{\log(\|O-x\|_{\mathbb{R}^2})\}_{x\in K} \ , \\ LD_S &\stackrel{\text{def}}{=} \{\mathbf{1}_{x\in S}\log(\|O-x\|_{\mathbb{R}^2})\}_{x\in K} \ , \\ \mathbb{1}_S &\stackrel{\text{def}}{=} \{\mathbf{1}_{x\in S}\}_{x\in K} \ , \\ \mathrm{diag}_S &\stackrel{\text{def}}{=} \mathrm{diag}\left(\mathbb{1}_S\right) \ , \\ Z_{\cdot,i} &\stackrel{\text{def}}{=} \{\mathbf{1}_{x\in S} Z_{x,i}\}_{x\in K},\ \forall i \in 1:N \ , \\ \overline{Z} &\stackrel{\text{def}}{=} \frac{1}{N|S|}\sum_{s\in S}\sum_{i=1}^{N} Z_{s,i} \ , \\ \overline{LD} &\stackrel{\text{def}}{=} \frac{1}{|S|}\sum_{s\in S} LD_s \ , \\ \overline{LD^2} &\stackrel{\text{def}}{=} \frac{1}{|S|}\sum_{s\in S} LD_s^2 \ , \\ \overline{LDZ} &\stackrel{\text{def}}{=} \frac{1}{N|S|}\sum_{s\in S} LD_s \sum_{i=1}^{N} Z_{s,i} \ . \end{aligned}$$

L'approche choisie pour estimer la fonction $f_\star$ consiste, tout d'abord, à construire un estimateur $\widehat{\theta} = (\widehat{c}_1, \widehat{c}_2, \widehat{\delta})$, du vecteur de paramètres $\theta_\star = (c_1^\star, c_2^\star, \delta_\star)$. Nous introduisons alors ce que l'on

désignera par log-vraisemblance pénalisée des mesures d'entrainement $\{Z_{\cdot,i}\}_{i=1}^N$, définie, pour tout $\theta = (c_1, c_2, \delta)$, par

$$\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N) \stackrel{\text{def}}{=} \frac{1}{N}\log L_\theta(\{Z_{\cdot,i}\}_{i=1}^N) + \text{pen}(N,\delta)\ , \tag{5.3}$$

avec

$$L_\theta(\{Z_{\cdot,i}\}_{i=1}^N) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}^N}\prod_{i=1}^N \exp\left[-\frac{\|Z_{\cdot,i} - (c_1\mathbb{1}_S + c_2 LD_S + \text{diag}_S\delta)\|^2}{2\sigma^2}\right]\ ,$$

représentant la vraisemblance des observations $\{Z_{\cdot,i}\}_{i=1}^N$ ($\|\cdot\|$ désigne ici la norme euclidienne sur $\mathbb{R}^{|K|}$) et

$$\text{pen}(N,\delta) \stackrel{\text{def}}{=} -\frac{1}{2N}\delta^T\Sigma^{-1}\delta\ ,$$

représentant une fonction de pénalité induite par l'*a priori* gaussien sur le processus $\delta_\star$ décrit dans la section précédente (équation (5.1)). Nous définissons alors $\widehat{\theta} = (\widehat{c}_1, \widehat{c}_2, \widehat{\delta})$, estimateur du maximum de vraisemblance pénalisée du paramètre $\theta_\star$, comme l'un des θ maximisant $\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N)$. L'expression de $\widehat{\theta}$ se fait alors par annulation du gradient de $\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N)$ par rapport à θ. Ré-écrivons tout d'abord l'équation (5.3) à l'aide des notations introduites au début de cette section.

$$\begin{aligned}\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N) = -\frac{1}{2N\sigma^2}\Bigg[&\sum_{i=1}^N \|Z_{\cdot,i}\|^2 - 2c_1N|S|\overline{Z} - 2c_2N|S|\overline{LDZ} - 2\delta^T\left(\sum_{i=1}^N Z_{\cdot,i}\right)\\ &+ N|S|c_1^2 + Nc_2^2\overline{LD^2}|S| + N\delta^T\text{diag}_S\delta\\ &+ 2N|S|c_1c_2\overline{LD} + 2Nc_1\delta^T\mathbb{1}_S + 2Nc_2\delta^T LD_S\Bigg]\\ &- \frac{1}{2N}\delta^T\Sigma^{-1}\delta\ .\end{aligned} \tag{5.4}$$

où $\langle\cdot,\cdot\rangle$ désigne le produit scalaire usuel de $\mathbb{R}^{|K|}$. Étant donné que $\widehat{c}_1$, $\widehat{c}_2$ et $\widehat{\delta}$ réalisent le maximum de $\frac{1}{N}\ell_\theta$, ils annulent alors, nécessairement, les dérivées partielle de (5.4) par rapport à c_1, c_2 et δ :

$$\frac{\partial}{\partial c_1}\left(\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N)\right)\Bigg|_\theta = -\frac{1}{2N\sigma^2}\left[-2N|S|\overline{Z} + 2N|S|c_1 + 2N|S|\overline{LD}c_2 + 2N\delta^T\mathbb{1}_S\right]\ , \tag{5.5}$$

$$\frac{\partial}{\partial c_2}\left(\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N)\right)\Bigg|_\theta = -\frac{1}{2N\sigma^2}\left[-2N|S|\overline{LDZ} + 2N|S|\overline{LD^2}c_2 + 2N|S|\overline{LD}c_1 + 2N\delta^T LD_S\right]\ , \tag{5.6}$$

$$\begin{aligned}\frac{\partial}{\partial\delta}\left(\frac{1}{N}\ell_\theta(\{Z_{\cdot,i}\}_{i=1}^N)\right)\Bigg|_\theta = -\frac{1}{2N\sigma^2}\Bigg[&-2\sum_{i=1}^N Z_{\cdot,i} + 2N\text{diag}_S\delta\\ &+ 2Nc_1\mathbb{1}_S + 2Nc_2LD_S\Bigg] - \frac{1}{N}\Sigma^{-1}\delta\ .\end{aligned} \tag{5.7}$$

Si l'on défini

$$M \stackrel{\text{def}}{=} \left[\text{diag}_S + \frac{\sigma^2}{N}\Sigma^{-1}\right]\ ,$$

alors, M est inversible, M^{-1} est symétrique et, si δ annule l'expression (5.7), alors

$$\delta = M^{-1}\left[\frac{\sum_{i=1}^{N} Z_{\cdot,i}}{N} - c_1 \mathbb{1}_S - c_2 LD_S\right] . \tag{5.8}$$

Nous pouvons alors remplacer δ par son expression en fonction de c_1 et c_2 dans les équations (5.5) et (5.6). L'annulation des expressions ainsi obtenues équivaut au système d'équations sur c_1 et c_2 suivant :

$$\begin{cases} \left(1 - \frac{1}{|S|}\mathbb{1}_S^T M^{-1}\mathbb{1}_S\right) c_1 & + \left(\overline{LD} - \frac{1}{|S|} LD_S^T M^{-1}\mathbb{1}_S\right) c_2 & = \overline{Z} - \frac{1}{|S|N}\sum_{i=1}^{N} Z_{\cdot,i}^T M^{-1}\mathbb{1}_S \\ \left(\overline{LD} - \frac{1}{|S|}\mathbb{1}_S^T M^{-1} LD_S\right) c_1 & + \left(\overline{LD^2} - \frac{1}{|S|} LD_S^T M^{-1} LD_S\right) c_2 & = \overline{LDZ} - \frac{1}{|S|N}\sum_{i=1}^{N} Z_{\cdot,i}^T M^{-1} LD_S . \end{cases} \tag{5.9}$$

Dès que $|S| \geq 2$ et qu'il existe s_1 et s_2 dans S tels que $LD_{s_1} \neq LD_{s_2}$, le système (5.9) est inversible et $(\widehat{c_1}, \widehat{c_2})$ est alors l'unique couple de solutions de ce système. $\widehat{\delta}$ est alors défini par (5.8) en substituant $\widehat{c_1}$ à c_1 et $\widehat{c_2}$ à c_2. Finalement, nous définissons $\widehat{f}$, estimateur de $f_\star$ sur la grille K par :

$$\forall x \in K,\ \widehat{f}(x) \overset{\text{def}}{=} \widehat{c_1} + \widehat{c_2}\log(\|O - x\|_{\mathbb{R}^2}) + \widehat{\delta}_x . \tag{5.10}$$

5.1.2 Estimation des hyperparamètres de covariance

Comme pour la section précédente, nous supposons ici que $\ell = 1$ et omettons volontairement les indices relatifs aux points d'accès. Dans la section précédente, nous avons construit un estimateur $\{\widehat{f}(x)\}_{x\in K}$ de la fonction $f_\star$ sur K construit, sur la base des mesures d'entraînement $\{Z_{s,i}\}_{s\in S,\ i\in 1:N}$ et lorsque la matrice de covariance Σ est connue. Ici, nous nous intéressons à l'estimation des hyperparamètres v_1 et v_2 intervenant dans l'expression (5.2). Nous rappelons la définition de la fonction d'autocovariance isotrope : pour tout u, vecteur unitaire quelconque, la fonction d'autovovariance C est donnée, pour tout h positif, par

$$C(h) = \mathbb{E}\left(\delta_\star(x)\delta_\star(x + hu)\right) .$$

Malheureusement, la partie déterministe $\mu_\star$ étant inconnue, nous n'avons pas accès aux réalisations du processus $\delta_\star$. Cette problématique, soulevée par Matheron dans Matheron [1970], est classique dans la littérature sur le krigeage, la fonction d'autocovariance doit alors être calculée sur la base des résidus $e_{s,i} \overset{\text{def}}{=} Z_{s,i} - \widehat{\mu}(s)$ où $\widehat{\mu}(s)$ est un estimateur de la fonction $\mu_\star$. Cependant, comme nous l'avons vu dans l'équation (5.9), une bonne estimation de $\mu_\star$ requiert la connaissance de la structure de dépendance spatiale de $\delta_\star$. Cette structure de dépendance ne pouvant être estimée qu'avec la connaissance de $\mu_\star$... Nous procédons alors à une première estimation de $\mu_\star$ par moindres carrés ordinaires :

$$(\widehat{c_1}, \widehat{c_2}) = \underset{c_1,c_2}{\operatorname{argmin}} \sum_{s\in S}\sum_{i=1}^{N} \left[Z_{s,i} - c_1 - c_2 LD_s\right]^2 .$$

La construction d'un estimateur de v_1 et v_2 peut alors se faire en introduisant ce que l'on appellera le semi-variogramme γ défini, quelque soit u vecteur unitaire, et pour tout $h > 0$ par :

$$\gamma(h) \overset{\text{def}}{=} \frac{1}{2}\operatorname{var}(\delta_\star(x) - \delta_\star(x + hu)) .$$

L'équation suivante relie la fonction d'autocovariance et le semi-variogramme.

$$\gamma(h) = C(0) - C(h) \ .$$

L'estimation du semi-variogramme est plus aisée que celle de la fonction d'autocovariance puisqu'elle ne requiert pas d'estimation de l'espérance de $\delta_\star(x)$, de plus, contrairement à la fonction d'autocovariance, son existence ne suppose pas a priori l'existence de la variance de $\delta_\star(x)$. D'après (5.2), le semi-variogramme est donné par

$$\gamma(h) = v_1(1 - \exp(-h^2/2v_2)) \ ,$$

et des estimateurs de v_1 et de v_2 pourront alors être construits suite à une première estimation du semi-variogramme γ. Nous posons $\widehat{\gamma}$, estimateur de γ défini, pour tout h positif, par :

$$\widehat{\gamma}(h) \stackrel{\text{def}}{=} \frac{1}{2N(h)N^2} \sum_{(s,s')\in V(h)} \sum_{i=1}^{N} \sum_{i'=1}^{N} (e_{s,i} - e_{s',i'})^2 \ ,$$

où $V(h) \stackrel{\text{def}}{=} \{(s,s') \in S^2 \mid \|s - s'\|_{\mathbb{R}^2} = h\}$ et $N(h) \stackrel{\text{def}}{=} |V(h)|$. Si les points de S sont répartis sur une grille, γ ne peut être estimé que pour un nombre restreint de distances (les distances h telles que $N(h) \neq 0$), nous introduisons alors une tolérance sur les distances en modifiant l'ensemble $V(h)$ de sorte que

$$V(h) = \{(s,s') \in S^2 \mid h - \Delta/2 \leq \|s - s'\|_{\mathbb{R}^2} < h + \Delta/2\} \ ,$$

où Δ est un pas de tolérance fixé. Nous construisons $\{\widehat{\gamma}(h_k)\}_{k=1}^{k_{max}}$, où $h_k \stackrel{\text{def}}{=} k\Delta$ et k_{max} est défini par $k_{max} \stackrel{\text{def}}{=} \max\{k \in \mathbb{N} \mid \ N(h_k) \neq 0\}$. Nous posons alors, comme estimateurs de v_1 et v_2, l'estimateur des moindres carrés pondérés suivant :

$$(\widehat{v}_1, \widehat{v}_2) \stackrel{\text{def}}{=} \underset{v_1>0, v_2>0}{\operatorname{argmin}} \sum_{k=0}^{k_{max}} w_k \left(v_1(1 - e^{-\frac{h^2}{2v_2}}) - \widehat{\gamma}(h_k) \right)^2 \ , \tag{5.11}$$

où les w_k sont des poids fixés. En particulier, Arnaud and Emery [2000] affirme que le semi-variogramme expérimental n'est pas fiable pour de grandes distances, on pourra alors annuler les poids w_k correspondant aux distances h_k supérieurs à un seuil, si l'on souhaite accorder plus d'importances aux distances h_k telles que $N(h_k)$ soit grand, on peut alors choisir $w_k = N(h_k)$ dans (5.11). Le package *geoR* du logiciel de statistiques R met à disposition de l'utilisateur toutes les fonctionnalités permettant le calcul des hyperparamètres, notamment la méthode d'ajustement désirée (choix des poids w_k, seuillage de la distance,...), ainsi que des méthodes de calcul des autres quantités intervenant en statistiques spatiales.

Les estimateurs $\widehat{v}_1$ et $\widehat{v}_2$ de v_1 et de v_2 permettent alors la construction de $\widehat{\Sigma}$, estimateur de la matrice de covariance Σ. Comme mentionné au début de cette section, l'estimation du semi-variogramme, basé sur les résidus $e_{s,i}$, est biaisée puisqu'elle est basée sur une estimation $\widehat{\mu}$ de $\mu_\star$ ne prenant pas en compte la structure de dépendance du processus $\delta_\star$, ce qui introduit de la dépendance dans les erreurs. Une méthode de réduction de ce biais, introduit par Cressie [1993], consiste à construire un premier estimateur de Σ, comme décrit dans cette section, sur la base des résidus $e_{s,i}$

calculés à l'aide d'une première estimation des paramètres c_1 et c_2 par moindres carrés ordinaires, puis d'estimer à nouveau les paramètres c_1 et c_2 grâce à l'équation (5.9) en substituant son estimateur à Σ et en construisant un premier estimateur de σ^2 suivant l'équation (4.2). La matrice de covariance Σ est ensuite ré-estimée sur la base de ce nouvel estimateur de $\mu_\star$, ce processus pouvant être itéré plusieurs fois jusqu'à convergence des estimateurs, qui, en pratique, arrive dès la première itération.

5.2 Méthodes de Monte-Carlo Sequentielles (SMC) sur la grille de discrétisation

Un estimateur de $f_\star$ pouvant être construit grâce aux techniques de statistiques spatiales décrites dans les sections précédentes nous supposons, dans cette section, les paramètres d'émission et de transition connus. Nous noterons alors g la densité du noyau d'émission, Q la matrice de transition et ν la mesure initiale de la chaîne de Markov $\{X_t\}_{t\geq 0}$. Nous nous intéressons ici à la construction d'estimateurs $\widehat{X}_n$ de X_n, $n \geq 0$, sur la base des observations $Y_{0:n}$. Comme dans le cas où l'environnement est grossièrement discrétisé, $|K|$ est fini, et pour toute mesure de probabilité initiale ν sur K les lois de filtrages $\phi_{\nu,n}$ possèdent une densité par rapport à la mesure de comptage sur K, toujours notées $\phi_{\nu,n}$. Ces densités vérifient alors la formule de récursion *forward*, dérivant de l'équation (3.7), suivante :

$$\begin{aligned} c_{\nu,k} &= \sum_{(x,x')\in K^2} \phi_{\nu,k-1}(x')Q(x',x)g(x,y_k), \\ \phi_{\nu,k}(x) &= c_{\nu,k}^{-1} \sum_{x'\in K} \phi_{\nu,k-1}(x')Q(x',x)g(x,y_k) \ . \end{aligned}$$

La construction des prédicteurs de maximum *a posteriori*, $\widehat{X}_n = \mathrm{argmax}_{x\in K}\, \phi_{\nu,n}(x),\ n \geq 0$, nécessite donc le calcul de $\phi_{\nu,n}(x)$ pour tout x de K, le temps de calcul de la densité $\phi_{\nu,n}$ dépend alors linéairement du nombre d'éléments d'états dans K, ainsi, ce temps de calcul augmentera dès lors que la taille de l'environnement augmente ou que le pas de discrétisation diminue. Afin de réduire et d'homogénéiser ce temps de calcul, nous introduisons ici des méthodes d'approximation de Monte Carlo séquentielles communément appelées *techniques de filtrage particulaire*. Ces techniques ont originellement pour objectif d'approximer les lois de lissages ou de filtrages lorsqu'elles ne peuvent être calculées explicitement (lorsque l'espace d'états est continu par exemple), dans notre cas, elles permettront à la fois de réduire le temps de calcul des positions mais aussi de le rendre indépendant du nombre d'états considérés. L'idée principale développée dans cette section est d'approcher les distributions de lissages $\{\phi_{\nu,n}\}_{n\geq 0}$ par des quantités aléatoires $\{\widehat{\phi}_{\nu,n}\}_{n\geq 0}$.

5.2.1 Échantillonage et ré-échantillonage d'importance pour l'estimation d'intégrales.

Dans un premier temps, nous nous intéresserons à une méthode d'approximation d'intégrales de la forme,

$$\phi(f) = \int_{\mathbb{X}} f(x)\phi(\mathrm{d}x) \ ,$$

lorsque f est une fonction réelle $\mathcal{X}$-mesurable sur l'espace $(\mathbb{X}, \mathcal{X})$ et ϕ une mesure de probabilité d'intérêt. Une méthode d'approximation de $\phi(f)$ consiste à simuler un N-échantillon $\{\xi^i\}_{i=1}^N$ de la mesure de probabilité ϕ puis d'approximer $\phi(f)$ par la moyenne $\frac{1}{N}\sum_{i=1}^N f(\xi^i)$. L'objectif des techniques d'échantillonage d'importance (ou IS pour *important sampling*) et de ré-échantillonage d'importance (ou SIR pour *sampling importance resampling*) est d'approximer $\phi(f)$, lorsque l'échantillonage sous ϕ n'est pas possible, ou difficile. Supposons que la distribution ϕ est absolument continue par rapport à une mesure ψ sur $\mathbb{X}$, noté $\phi << \psi$, où ψ est choisie telle que l'échantillonage sous ψ est possible, ψ est alors appelée *mesure de probabilité d'importance*. Nous notons alors $d\phi/d\psi$ la dérivée de Radon-Nikodym (densité de probabilité par rapport à la mesure ψ) de ϕ par rapport à ψ. $\phi(f)$ peut alors être réécrit de la manière suivante :

$$\phi(f) = \int_{\mathbb{X}} f(x)\frac{\mathrm{d}\phi}{\mathrm{d}\psi}(x)\psi(\mathrm{d}x) \ .$$

Ainsi, dans le cas où $d\phi/d\psi$ est connue à constante près, si $\{\xi^i\}_{i=1}^N$ est un N-échantillon de même loi ψ, $\phi(f)$ peut être approchée par :

$$\widehat{\phi}_N^{\mathrm{IS}}(f) = \frac{\sum_{i=1}^N f(\xi^i)\frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\xi^i)}{\sum_{i=1}^N \frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\xi^i)} \ .$$

Par la loi forte des grands nombres appliquée aux quantités $\frac{\sum_{i=1}^N f(\xi^i)\frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\xi^i)}{N}$ et $\frac{\sum_{i=1}^N \frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\xi^i)}{N}$, l'estimateur $\widehat{\phi}_N^{\mathrm{IS}}(f)$ de $\phi(f)$ est fortement consistent.

En plus d'estimer des quantités de la forme $\phi(f)$, l'échantillonage d'importance peut aussi être utilisé pour échantilloner, de manière approximative, selon la distribution ϕ. Ceci peut être fait grâce à la méthode de ré-échantillonage d'importance SIR qui consiste à simuler un premier M-échantillon $\{\widetilde{\xi}^i\}_{i=1}^M$ selon la distribution ψ, puis de calculer les quantités $\{\omega^i\}_{i=1}^M$, appelés poids d'importance et définis, pour tout $i = 1, \ldots, M$, par,

$$\omega_i = \frac{\frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\widetilde{\xi}^i)}{\sum_{j=1}^M \frac{\mathrm{d}\phi}{\mathrm{d}\psi}(\widetilde{\xi}^j)} \ .$$

Une fois les poids d'importance calculés, une étape, dite de ré-échantillonage, est effectuée : si N est un nombre positif, nous définissons $\{I_1, \ldots, I_N\}$, N variables aléatoires à valeurs dans $\{1, \ldots, M\}$, indépendantes conditionnellement à $\{\tilde{\xi}^i, \omega^i\}_{i=1}^M$ et telles que, pour tout $i = 1, \ldots, N$ et pour tout $j = 1, \ldots, M$,

$$\mathbb{P}(I^i = j \mid \{\tilde{\xi}^l, \omega^l\}_{l=1}^M) = \omega^j \ ,$$

l'étape de ré-échantillonage consiste alors à poser, pour tout $i = 1, \ldots, N$,

$$\xi^i = \widetilde{\xi}^{I^i} \ .$$

Bien que les variables ξ^i et ξ^j, $(i, j) \in (1 : N)^2$ ne soient pas indépendantes ([Cappé et al., 2005, Section 9.2]), elles le sont asymptotiquement dans le sens où, pour toutes fonctions mesurables bornées f et g de $\mathbb{X}$, $\mathbb{E}\left(f(\xi^i)g(\xi^j)\right)$ tend vers $\phi(f)\phi(g)$ lorsque M, taille du premier échantillon $\{\widetilde{\xi}^i\}_{i=1}^M$,

tend vers l'infini, de sorte que $\xi^1, \ldots, \xi^N$ peut être vu par la suite comme un N-échantillon de la distribution ϕ. On définit alors $\widehat{\phi}^{SIR}_{M,N}(f)$, un autre estimateur de $\phi(f)$, par la moyenne empirique :

$$\widehat{\phi}^{SIR}_{M,N}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi^i) \; .$$

$\widehat{\phi}^{SIR}_{M,N}(f)$ est alors un estimateur sans biais de $\phi(f)$.

5.2.2 Échantillonage et ré-échantillonage d'importance pour l'estimation des lois de filtrage.

Lorsque le modèle est complètement dominé, les distributions de filtrage $\{\phi_{\nu,n}\}_{n\in\mathbb{N}}$ possèdent une densité par rapport à la mesure μ sur $\mathbb{X}$, nous noterons alors indifféremment $\{\phi_{\nu,n}\}_{n\in\mathbb{N}}$ pour désigner les lois de filtrage ou leur densité par rapport à μ. Dans cette section, nous nous intéressons au calcul effectif des lois de filtrages $\phi_{\nu,n}$. Par la suite, nous omettons les mentions à la distribution initiale ν dans l'expression des quantités de lissage et de filtrage. Les lois de filtrages peuvent alors être exprimées de manière récursive : pour toute fonction réelle mesurable bornée f,

$$\phi_0(f) = \frac{\int f(x) g(x, Y_0) \nu(\mathrm{d}x)}{\int g(x, Y_0) \nu(\mathrm{d}x)} \; ,$$

et pour tout $t \geq 0$,

$$\phi_{t+1}(f) = \int \int f(x') \phi_t(\mathrm{d}x) T_t(x, \mathrm{d}x') \; ,$$

où T_t est le noyau de transition sur $(\mathbb{X}, \mathcal{X})$ défini, pour tout $x \in \mathbb{X}$ et toute fonction réelle, mesurable et bornée f, par

$$T_t(x, f) = \frac{L_t}{L_{t+1}} \int f(x') Q(x, \mathrm{d}x') g(x', Y_{t+1}) \; .$$

Il est important de remarquer que T_t n'est pas un noyau de transition de probabilité et que l'échantillonage selon le noyau de transition T_t renormalisé peut être difficile sauf dans des cas simples. Nous présentons ici des méthodes d'échantillonage d'importance permettant l'estimation récursive des lois de filtrages $\{\phi_t\}_{t\in\mathbb{N}}$, utilisant les méthodes d'échantillonage d'importance étudiées précédemment et appliquées à l'estimation des noyaux de transition $\{T_t\}_{t\in\mathbb{N}}$.

Soir $\{R_t\}_{t\in\mathbb{N}}$ une famille de noyaux de transition markoviennes sur $(\mathbb{X}, \mathcal{X})$. Si l'on note $\nu_0 = \nu$ la distribution initiale de la chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$, notons $\{\nu_{0:t}\}_{t\in\mathbb{N}}$ la famille de mesures de probabilité associée à la chaîne de Markov inhomogène (*i.e.* dont le noyau de transition dépend du temps t) de distribution initiale ν_0 et de noyaux de transition $\{R_t\}_{t\in\mathbb{N}}$ définie, pour toute fonction $\mathcal{X}^{\otimes(t+1)}$-mesurable et bornée f, par :

$$\nu_{0:t}(f) \stackrel{\mathrm{def}}{=} \int_{x_{0:t}\in\mathbb{X}^{t+1}} f(x_{0:t}) \nu(\mathrm{d}x_0) \prod_{k=0}^{t-1} R_k(x_k, \mathrm{d}x_{k+1}) \; .$$

Les noyaux R_t sont alors appelés *noyaux de transition d'importance*. La distribution initiale $\nu_0 = \nu$ du processus $\{X_t\}_{t\in\mathbb{N}}$ étant supposée connue, la loi de filtrage ϕ_0 est absolument continue par rapport

à ν_0 (car le modèle d'émission est dominé). On notera alors $\frac{\mathrm{d}\phi_0}{\mathrm{d}\nu_0}$ la dérivée de Radon-Nikodym de ϕ_0 par rapport à ν_0. Nous supposons par ailleurs que pour tout $t \geq 0$, et pour tout $x \in \mathbb{X}$, la mesure $T_t(x,\cdot)$ est absolument continue par rapport à $R_t(x,\cdot)$, et nous noterons $\frac{\mathrm{d}T_t(x,\cdot)}{\mathrm{d}R_t(x,\cdot)}$ la dérivée de Radon-Nikodym de $T_t(x,\cdot)$ par rapport à $R_t(x,\cdot)$. Ainsi, pour tout $t \in \mathbb{N}$, les lois de filtrages peuvent être exprimées de la manière suivante :

$$\phi_t(f) = \int_{x_{0:t}\in\mathbb{X}^{t+1}} f(x_t)\frac{\mathrm{d}\phi_0}{\mathrm{d}\nu_0}(x_0)\left\{\prod_{k=0}^{t-1}\frac{\mathrm{d}T_k(x_k,\cdot)}{\mathrm{d}R_k(x_k,\cdot)}(x_{k+1})\right\}\nu_{0:t}(\mathrm{d}x_{0:t}) .$$

Ainsi, en considérant $\nu_{0:t}$ comme mesure de probabilité d'importance, nous pouvons simuler un N-échantillon $\{\xi^i_{0:t}\}_{i=1}^N$ de la distribution $\nu_{0:t}$. L'estimateur IS de ϕ_t est alors définit, pour toute fonction mesurable bornée f, par

$$\widehat{\phi}^{IS}_{t,N} = \frac{\sum_{i=1}^N \omega^i_t f(\xi^i_t)}{\sum_{i=1}^N \omega^i_t} , \tag{5.12}$$

où $\{\omega^i_t\}_{t\in\mathbb{N}}$ est défini, pour tout $i = 1,\ldots,N$, de manière récursive, par

$$\omega^i_0 = \frac{\mathrm{d}\phi_0}{\mathrm{d}\nu_0}(\xi^i_0) ,$$
$$\omega^i_{t+1} = \omega^i_t\frac{\mathrm{d}T_t(\xi^i_t,\cdot)}{\mathrm{d}R_t(\xi^i_t,\cdot)}(\xi^i_{t+1}) , \ \forall t \geq 0 .$$

En développant l'expression du noyau $T_t(\xi^i_t,\cdot)$, l'algorithme SIS (*Sequential Important Sampling*), décrit dans l'algorithme 1 permet la simulation des variables $\{\xi^i_{0:t}\}_{i=1}^N$ ainsi que le calcul des poids $\{\omega^i_t\}_{i=1}^N$ de manière récursive. Nous désignerons désormais, pour tout $t \geq 0$ la suite des couples

Algorithm 1 Sequential Important Sampling (SIS)

1: **Initialisation** : Simuler un N-échantillon $\{\xi^i_0\}_{i=1}^N$ de la distribution ν_0, et poser, pour tout $i = 1,\ldots,N$,
$$\omega^i_0 = g(\xi^i_0, Y_0) .$$
2: **Récursion** :
3: Pour $t \geq 0$,
4: Simuler $\{\xi^i_{t+1}\}_{i=1}^N$, indépendant conditionnellement à $\{\xi^i_{0:t}\}_{i=1}^N$, de sorte que, pour tout $i = 1,\ldots,N$, $\xi^i_{t+1} \sim R_t(\xi^i_t,\cdot)$ et définir $\xi^i_{0:t+1} = (\xi^i_{0:t}, \xi^i_{t+1})$.
5: Calculer, pour tout $i = 1,\ldots,N$, les poids d'importance :
$$\omega^i_{t+1} = \omega^i_t g(\xi^i_{t+1}, Y_{t+1})\frac{\mathrm{d}Q(\xi^i_t,\cdot)}{\mathrm{d}R_t(\xi^i_t,\cdot)}(\xi^i_{t+1}) .$$

$\{(\xi^i_t,\omega^i_t)\}_{i=1}^N$, par *particules au temps t*. La mise en application de l'algorithme SIS nécessite le choix d'une suite de noyaux de transition d'importance $\{R_t\}_{t\geq 0}$. Le premier choix naturel qui s'impose (et celui que l'on fera par la suite) est de poser, pour tout indice t positif, $R_t = Q$, dans ce cas, la récursion sur les poids d'importance s'écrit, pour tout $t \geq 0$:

$$\omega^i_{t+1} = \omega^i_t g(\xi^i_{t+1}, Y_{t+1}) .$$

Cependant, l'utilisation de l'algorithme SIS en pratique possède un inconvénient : la dégénérescence des poids d'importance ω_t^i. Lorsque t est grand, la majorité des poids ω_t^i devient négligeable par rapport à $\sum_{i=1}^N \omega_t^i$ de sorte que, même lorsque N est grand, seules quelques particules impactent réellement sur les distributions approchées $\widehat{\phi}_{t,N}^{IS}$. L'effort computationnel de l'algorithme SIS devient alors obsolète dès lors que l'on met à jour des particules qui n'interviennent pas ou très peu dans la construction de $\widehat{\phi}_{t,N}^{IS}$. De plus, si l'on pose, pour tout $t \geq 0$, $i_t^{\max} = \text{argmax}_{i=1,\dots,N} \omega_t^i$, et $\widehat{X}_t = \xi_t^{i_t^{\max}}$ comme estimateur de la position au temps t, il existera alors en pratique un indice $i_0 \in \{1,\dots,N\}$ tel que $i_t^{\max} = i_0$ pour tout t assez grand, et $\widehat{X}_t$ sera alors égale systématiquement à $\xi_t^{i_0}$. Nous définissons alors l'estimateur SISR de ϕ_t en procédant régulièrement à une étape de ré-échantillonage des particules sur le même principe que le ré-échantillonage intervenant dans la construction de l'estimateur $\widehat{\phi}_{M,N}^{SIR}$ dans la section précédente. Ce ré-échantillonage intervient alors dans le calcul récursif des particules suivant l'algorithme SISR décrit dans l'algorithme 2. Dans le cas

Algorithm 2 Sequential Important Sampling with Resampling (SISR)

1: **Initialisation** : Simuler un N-échantillon $\{\xi_0^i\}_{i=1}^N$ de la distribution ν_0, et poser, pour tout $i = 1,\dots,N$,
$$\omega_0^i = g(\xi_0^i, Y_0) \ .$$

2: **Récursion** : Pour $t \geq 0$, Simuler $\{\widetilde{\xi}_{t+1}^i\}_{i=1}^N$, indépendant conditionnellement à $\{\xi_{0:t}^i\}_{i=1}^N$, de sorte que, pour tout $i = 1,\dots,N$, $\widetilde{\xi}_{t+1}^i \sim R_t(\xi_t^i, \cdot)$.

3: Calculer, pour tout $i = 1,\dots,N$, les poids d'importance :
$$\widetilde{\omega}_{t+1}^i = \omega_t^i g(\widetilde{\xi}_{t+1}^i, Y_{t+1}) \frac{\mathrm{d}Q(\xi_t^i, \cdot)}{\mathrm{d}R_t(\xi_t^i, \cdot)}(\widetilde{\xi}_{t+1}^i) \ .$$

4: **Ré-échantillonage (optionnel)** : Simuler de manière indépendante, conditionnellement à $\{(\xi_{0:t}^j, \omega_{0:t}^j, \widetilde{\xi}_{t+1}^j, \widetilde{\omega}_{t+1}^j),\ j = 1,\dots,N\}$, $I_{t+1}^1,\dots,I_{t+1}^N$ de sorte que, pour tous $i,j = 1,\dots,N$,
$$\mathbb{P}(I_{t+1}^i = j \mid \{(\xi_{0:t}^j, \omega_{0:t}^j, \widetilde{\xi}_{t+1}^j, \widetilde{\omega}_{t+1}^j)\}_{l=1}^N) = \frac{\widetilde{\omega}_{t+1}^j}{\sum_{l=1}^N \widetilde{\omega}_{t+1}^l} \ .$$
Poser, pour tout $i = 1,\dots,N$, $\omega_{t+1}^i = 1$.

5: **Dans le cas où le ré-échantillonage n'est pas effectué** : poser, pour tout $i = 1,\dots,N$, $I_{t+1}^i = i$, $\omega_{t+1}^i = \widetilde{\omega}_{t+1}^i$ et $\xi_{0:t+1}^i = (\xi_{0:t}^i, \widetilde{\xi}_{t+1}^i)$.

où Q est choisi comme noyau de transition d'importance, et lorsque le ré-échantillonage est effectué systématiquement, l'algorithme SISR est communément appelé filtre *bootstrap* ou filtre particulaire.

5.3 Filtre *bootstrap* pour la géolocalisation WiFi

5.3.1 Application du filtre *bootstrap*

Dés lors que la phase préliminaire d'estimation de la fonction $f_\star$ est effectuée grâce aux résultats de la section 5.1, nous choisissons d'effectuer le calcul de la suite de prédicteurs de maximum *a*

posteriori $\{\widehat{X}_n\}_{n\geq 0}$ à l'aide du filtre *bootstrap*. Pour cela, nous supposons préalablement l'existence d'un paramètre a supposé connu tel que le noyau de transition Q est à densité $q_a(.,.)$ par rapport à la mesure de comptage μ sur K, où pour tout $x \in K$, il existe une constante $C_a(x)$ telle que pour tout x' de K,

$$q_a(x,x') \stackrel{\text{def}}{=} C_a(x) \exp(-\frac{\|x-x'\|^2_{\mathbb{R}^2}}{a}) \ .$$

La constante $C_a(x)$ est alors déterminée par le fait que $q_a(x,\cdot)$ doit être une densité de probabilité par rapport à la mesure de comptage μ, ainsi,

$$C_a(x)^{-1} = \sum_{x' \in K} \exp(-\frac{\|x-x'\|^2_{\mathbb{R}^2}}{a}) \ .$$

Le choix Gaussien pour le noyau de transition Q a été fait de manière à prendre en compte le fait que le terminal mobile, dont la vitesse est limitée (surtout à l'intérieur des bâtiments), ne peut parcourir une distance trop grande dans l'intervalle de temps séparant deux mesures du RSS consécutives. En particulier, la valeur du paramètre a n'a pas de réel sens physique. Cependant, en pratique, nous faisons dépendre sa valeur de la vitesse moyenne ou maximale du terminal mobile à géolocaliser, plus cette vitesse est faible, plus la valeur du paramètre a devra être petite de manière à ce que les transitions $q_a(x,x')$ soient relativement faibles pour des positions x et x' éloignées. *A contrario*, lorsque le terminal mobile peut en pratique passer d'une position x de K à n'importe quelle autre position x' de K dans le temps séparant deux envois de mesures Y_t et Y_{t+1}, la valeur choisie pour le paramètre a devra être élevée, ceci ayant pour effet de rapprocher la loi de transition $Q_a(x,\cdot)$ de la loi uniforme sur K.

Afin de prendre en considération l'emplacement des obstacles dans le modèle de déplacement, nous fixons $q_a(x,x') = 0$ lorsque x et x' sont séparés par un obstacle, empêchant alors le passage du terminal de la position x à la position x'. Cependant, ceci a pour effet de bloquer la progression des particules dans l'environnement : lorsque les particules, à l'instant t, $\{\xi_t^i\}_{i=1}^N$, se trouvent toutes du même côté d'un mur alors que le terminal se trouve en réalité de l'autre côté, les particules à l'instant $t+1$ ne pourront alors pas traverser le mur afin de rejoindre la bonne position. L'ajout d'une étape de ré-initialisation régulière des particules dans l'algorithme *bootstrap* permet de contourner ce problème.

Une fois le noyau de transition Q fixé, à chaque nouvelle mesure Y_t envoyée par le terminal au serveur, les particules sont mises à jour selon le filtre *bootstrap* (Algorithme 2 avec ré-échantillonage systématique). L'estimation de la position à l'instant t se fait alors à l'aide du prédicteur de maximum *a posteriori*,

$$\widehat{X}_t = \xi_t^{i_t^{\max}} \ , \tag{5.13}$$

où

$$i_t^{\max} = \underset{i=1,\ldots,N}{\operatorname{argmax}}\, \omega_t^i \ .$$

5.3.2 Résultats obtenus en pratique

Dans cette section, nous soumettons notre système de localisation à un test sur données réelles. L'objectif est de géolocaliser un terminal dans un environnement illustré par la figure 5.2. L'environnement est un entrepôt, les obstacles (fixes) présents dans l'environnement et illustrés sur la figure

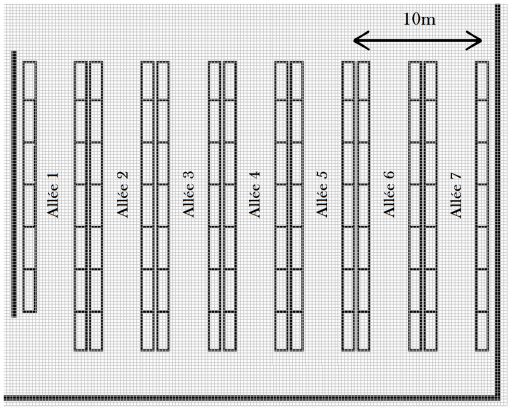


FIGURE 5.2 – Plan de l'environnement à géolocaliser finement discrétisé.

5.2 sont des supports en acier sur lesquels sont entreposées des palettes de marchandises. Nous procédons à une fine discrétisation de la carte (pas de discrétisation : 25cm). La première partie du test consiste à placer les points d'accès WiFi : au total, cinq configurations différentes de l'infrastructure WiFi seront étudiées, ces configurations sont représentées sur la figure 5.3.

Estimation des cartes de propagation Pour chaque configuration représentée sur la figure 5.3, nous commençons la phase de test par la construction des cartes de propagation WiFi (estimation de $f_\star$). Une campagne de mesures préalable est alors effectuée en plusieurs points de la carte. Nous désignons par S l'ensemble de ces points de mesures, au total 146 points de la carte ont été mesurés ($|S| = 146$), et, pour chaque point de mesure s de S, $M = 5$ mesures ont été faites. Une fois cette phase de mesures terminée, et dans l'objectif de tester la méthode de prédiction de la section 5.1, nous retirons de l'ensemble S un nombre N_{test} de points de mesures de manière aléatoire et désignons par S_{test} l'ensemble de ces points (techniques du *leave* N_{test}*-out*).

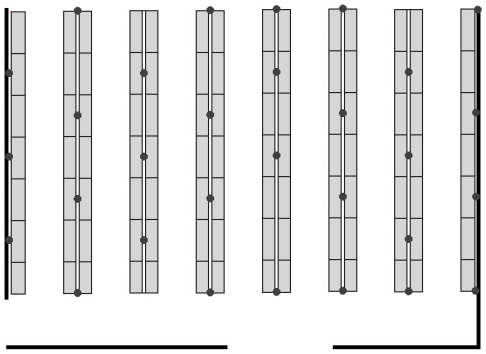
Nous procédons alors à l'estimation de la fonction $f_\star$ comme décrit dans la section 5.1. Notons, pour tout point d'accès $j = 1, \ldots, \ell$, $\widehat{f^j}$, prédicteur de $f_{\star_j}$ sur K, construit sur la base des mesures $\{Z^j_{s,i}\}_{s \in S \setminus S_{\text{test}}}$, et défini par l'équation (5.10). Posons aussi, pour tout $j = 1, \ldots, \ell$, $\text{err}^j(dBm) = \frac{1}{M|S_{\text{test}}|} \sum_{s \in S_{\text{test}}} \sum_{i=1}^{M} |Z^j_{s,i} - \widehat{f^j}(s)|$, estimateur de l'erreur L_1 entre $\widehat{f^j}$ et $f_{\star_j}$, et

$$\overline{\text{err}}(dBm) = \frac{1}{\ell} \sum_{j=1}^{\ell} \text{err}^j(dBm) \; .$$

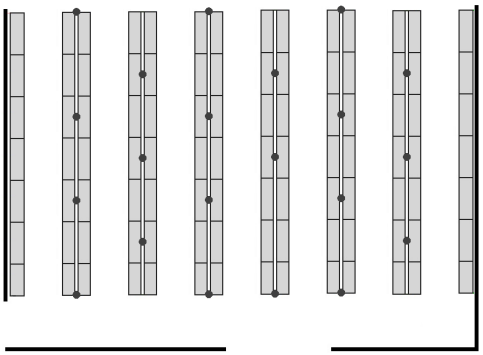
En particulier, pour la configuration 2 représentée sur la figure 5.3 (b), En prenant $|S_{\text{test}}| = 20$ la valeur observée de $\overline{\text{err}}(dBm)$ à l'issue de la construction de $\{\widehat{f^j}\}_{j=1}^{\ell}$ est de

$$\overline{\text{err}}(dBm) = 3.93 \; dBm \; .$$

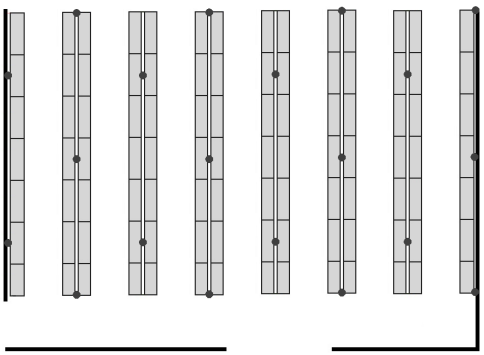
Il est important de souligner que cette valeur dépend du type d'environnement dans lequel on travaille, ici un entrepôt, mais aussi du nombre et de l'emplacement des points de mesures dans la phase de calibration préalable.



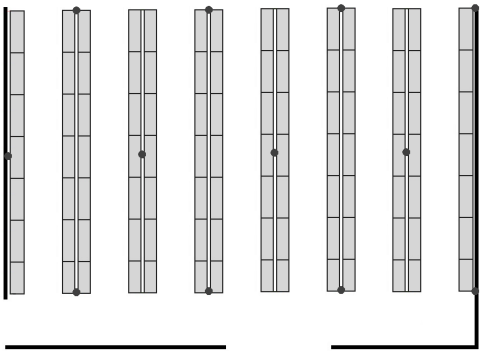
(a) Configuration 1, $\ell = 30$.



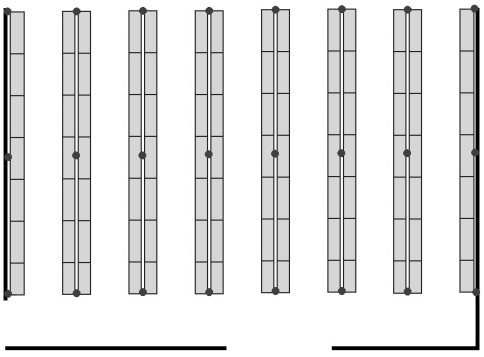
(b) Configuration 2 $\ell = 21$.



(c) Configuration 3 $\ell = 20$.



(d) Configuration 4 $\ell = 12$.



(e) Configuration 5 $\ell = 24$.

FIGURE 5.3 – Les cinq configurations testées, les points représentants les emplacements des points d'accès.

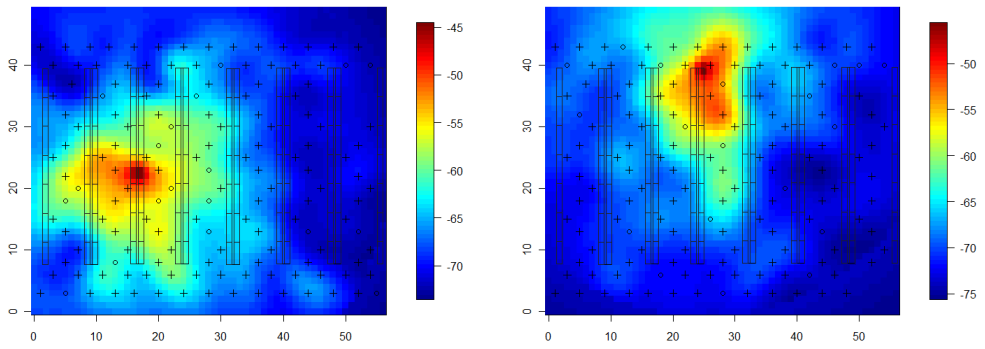


FIGURE 5.4 – Estimation des cartes de propagations moyennes de deux points d'accès, l'échelle, à droite de chaque graphique, est en dBm, les croix symbolisent les points de $S \setminus S_{\text{test}}$, les ronds symbolisent les points de S_{test}.

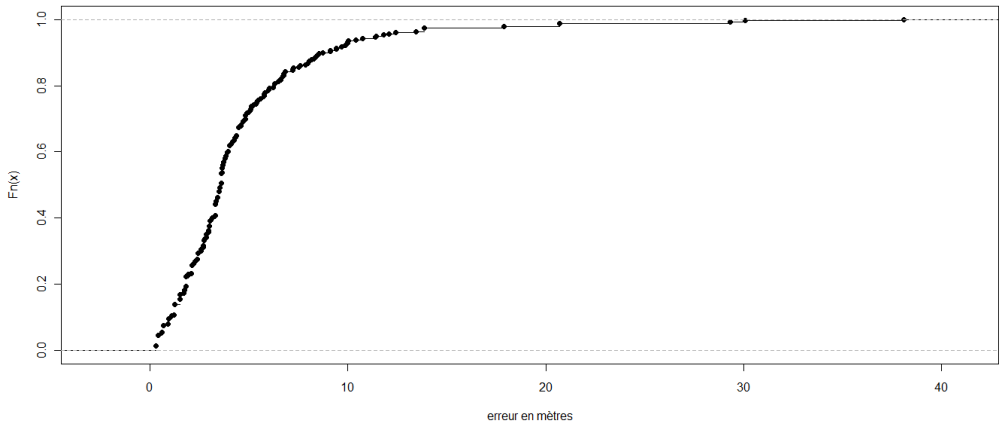


FIGURE 5.5 – Fonction de répartition des erreurs pour la configuration 1.

Mesure de la précision Une fois la construction des estimateurs $\{\widehat{f}^j\}_{j=1}^{\ell}$ terminée, nous testons la précision du filtre *bootstrap* à l'aide d'un système de codes à barres. Ceux-ci sont disséminés dans l'environnement, à des emplacements connus, et peuvent être lus par le terminal mobile. La précision du système est alors mesurée en comparant, à chaque lecture de code à barres par le terminal, la dernière position estimée $\widehat{X}_t$ avec la position, X_t, connue du code à barres lu par le terminal. Si l'on note T_{test} les indices t pour lesquelles une lecture de code à barre a été effectuée, les écarts $\{\|\widehat{X}_t - X_t\|_{\mathbb{R}^2}\}_{t \in T_{\text{test}}}$ sont sauvegardés dans une base de données sur le serveur de localisation. La figure 5.5 donne la fonction de répartition des écarts mesurés à l'issu du test de précision pour la configuration 1.

Le tableau 5.1 résume les résultats des tests de précision pour chaque configuration, il indique les quantiles à 80%, 70% et 60% des écarts $\{\|\widehat{X}_t - X_t\|_{\mathbb{R}^2}\}_{t \in T_{\text{test}}}$ obtenus pour chaque configuration. La précision à 80% (quantiles à 80%) varie de $6.3m$ à $10m$ selon la configuration. La précision semble dépendante de la configuration considérée et du nombre de points d'accès associés. En particulier,

la configuration 4 (la configuration avec le moins d'APs : $\ell = 12$) possède l'erreur en localisation la plus élevée par rapport aux autres configurations (dont le nombre d'APs varie de 20 à 30).

Afin d'améliorer les résultats présentés dans le tableau 5.1, nous proposons dans le paragraphe suivant une méthode de lissage des positions brutes $\{\widehat{X}_t\}_{t\in\mathbb{N}}$.

Quantiles / Configuration	1	2	3	4	5
80%	$6.3m$	$7.9m$	$7.6m$	$9.9m$	$6.3m$
70%	$4.8m$	$6.8m$	$6.1m$	$8.1m$	$4.9m$
60%	$4m$	$5.8m$	$5.1m$	$6.7m$	$4.2m$

TABLE 5.1 – tableau récapitulatif des précisions obtenues pour chaque configuration.

Lissage des positions par moyenne mobile Écrivons tout d'abord, pour tout $t \geq 0$, $\widehat{X}_t = X_t + W_t$ avec $W_t = \widehat{X}_t - X_t$. La technique de lissage des positions décrite dans cette section a pour objectif d'améliorer la précision en tenant compte du caractère régulier de la suite de positions $\{X_t\}_{t\in\mathbb{N}}$. Afin de comprendre l'idée sous-jacente à cette technique, plaçons nous dans l'hypothèse où $\{W_t\}_{t\in\mathbb{N}}$ est un bruit blanc (cette hypothèse n'est, bien entendu, pas réaliste car cela entrainerait en particulier que $\{\widehat{X}_t\}_{t\in\mathbb{N}}$ est non biaisé), tel que pour tout $t \geq 0$, $W_t = (W_{1,t}, W_{2,t})$ avec $W_{1,t}$ et $W_{2,t}$ bruits blancs indépendants de variance commune s^2 connue. Le modèle de transition utilisé dans la section 5.3.1 suppose que le mobile se déplace de manière aléatoire suivant un noyau de transition donné. Dans la pratique ceci est rarement vérifié : le terminal, si porté par un humain, va suivre une trajectoire, en général, régulière. Plaçons nous dans le cas où le terminal mobile possède, localement, une accélération constante (pouvant bien sûr être nulle en cas de vitesse constante ou de position fixe), cette supposition n'est pas absurde, les déplacements humains comportent rarement des accélérations brutales et s'effectuent même souvent à vitesse constante. Nous supposons alors que, pour tout $t_0 \geq 0$, il existe un voisinage V de t_0 ainsi que trois coefficients $a = (a_1, a_2)$, $b = (b_1, b_2)$ et $c = (c_1, c_2)$ tel que pour tout $t \in V$,

$$X_t = at^2 + bt + c\ .$$

Les estimateurs des positions $\{\widehat{X}_t\}_{t\in V}$ vérifient alors

$$\forall t \in V,\ \widehat{X}_t = at^2 + bt + c + W_t\ .$$

Nous optons pour un lissage de type moyenne mobile et définissons $\widetilde{X}_t$ estimateur lissé de la position X_t en posant :

$$\widetilde{X}_t \stackrel{\text{def}}{=} \sum_{k=0}^{p} \theta_k \widehat{X}_{t-k}\ , \tag{5.14}$$

où p est un indice tel que $t - p \in V$ et $\{\theta_k\}_{k=0}^p$ est une suite de coefficients de pondérations.

Nous cherchons alors la suite de coefficients de pondération $\{\theta_k\}_{k=0}^p$ satisfaisant que la transformation de $\widehat{X}_t$ décrite par l'équation (5.14) préserve la tendance polynomiale d'ordre 2 $at^2 + bt + c$ et réduise au maximum la variance d'estimation. Ainsi, la conservation de la tendance se traduit par la

condition suivante sur les coefficients de pondération :

$$\forall t \in V,\ at^2 + bt + c = \mathbb{E}(\widetilde{X}_t) = \sum_{k=0}^{p} \theta_k \left[a(t-k)^2 + b(t-k) + c \right] \ .$$

Ce qui équivaut à :

$$\sum_{k=0}^{p} \theta_k = 1 \ , \ \sum_{k=0}^{p} k\theta_k = 0 \ , \ \sum_{k=0}^{p} k^2\theta_k = 0$$

À ces trois conditions sur les coefficients $\{\theta_k\}_{k=0}^{p}$ s'ajoute la condition de minimisation de la variance,

$$\begin{aligned} \mathrm{var}(\widetilde{X}_t - X_t) &\stackrel{\mathrm{def}}{=} \mathbb{E}\left[(\widetilde{X}_{1,t} - \big(a_1t^2 + b_1t + c_1\big))^2 + (\widetilde{X}_{2,t} - \big(a_2t^2 + b_2t + c_2\big))^2 \right] \ , \\ &= 2s^2 \sum_{k=0}^{p} \theta_k^2 \end{aligned}$$

Nous résolvons alors ce problème de minimisation sous contrainte par l'introduction du lagrangien $\mathcal{L}$, que l'on cherche alors à minimiser :

$$\mathcal{L}(\{\theta_k\}_{k=0}^{p}, \lambda_1, \lambda_2, \lambda_3) = \mathrm{var}(\widetilde{X}_t - X_t) + \lambda_1 \left(\sum_{k=0}^{p} \theta_k - 1 \right) + \lambda_2 \left(\sum_{k=0}^{p} k\theta_k \right) + \lambda_3 \left(\sum_{k=0}^{p} k^2\theta_k \right) \ . \quad (5.15)$$

L'annulation des dérivées partielles $\frac{\partial}{\partial \theta_k}\mathcal{L}$ pour tous les k ainsi que des dérivées partielles $\frac{\partial}{\partial \lambda_i}\mathcal{L}$, $i = 1, 2, 3$ conduit alors au système inversible suivant :

$$\left\{ \begin{array}{ll} 4s^2\theta_k + \lambda_1 + k\lambda_2 + k^2\lambda_3 & = 0 \ , \ \forall k = 0, \ldots, p \\ \sum_{k=0}^{p} \theta_k & = 1 \\ \sum_{k=0}^{p} k\theta_k & = 0 \\ \sum_{k=0}^{p} k^2\theta_k & = 0 \end{array} \right. \ . \quad (5.16)$$

Posons alors

$$M = \begin{pmatrix} 4s^2 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 \\ 0 & 4s^2 & 0 & \ldots & 0 & 1 & 1 & 1 \\ 0 & 0 & 4s^2 & \ldots & 0 & 1 & 2 & 2^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 4s^2 & 1 & p & p^2 \\ 1 & 1 & 1 & \ldots & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & \ldots & p & 0 & 0 & 0 \\ 0 & 1 & 2^2 & \ldots & p^2 & 0 & 0 & 0 \end{pmatrix} \ ,$$

si l'on pose

$$v = (0, 0, 0, \ldots, 0, 1, 0, 0)^T \ ,$$

vecteur de taille $p + 4$, alors, l'unique solution $u = (\theta_0, \cdots, \theta_p, \lambda_1, \lambda_2, \lambda_3)^T$ du système (5.3.2) est donnée par $u = M^{-1}v$. Les $p + 1$ premières coordonnées de u représentent alors le vecteur de coefficients de pondération recherché.

En choisissant $p = 40$, et se fixant $4s^2 = 20$, le tableau 5.2 nous donne la précision obtenue à l'issue de ce lissage, soit les quantiles à 80%, 70% et 60% des erreurs $\left\{\|\widetilde{X}_t - X_t\|_{\mathbb{R}^2}\right\}_{t\in T_{\text{test}}}$ pour chaque configuration. Notre technique de lissage a ainsi permis de réduire l'erreur de positionnement d'un mètre environ pour chaque configuration.

Quantiles / Configuration	1	2	3	4	5
80%	$4.8m$	$6.8m$	$6.7m$	$8.4m$	$5.5m$
70%	$4.1m$	$5.9m$	$5.8m$	$7.0m$	$4.3m$
60%	$3.5m$	$5.2m$	$4.7m$	$5.8m$	$3.6m$

TABLE 5.2 – tableau récapitulatif des précisions obtenues après lissage pour chaque configuration.

Il est important de souligner que les écarts $\{\|\widehat{X}_t - X_t\|_{\mathbb{R}^2}\}_{t\in T_{\text{test}}}$ ou $\{\|\widetilde{X}_t - X_t\|_{\mathbb{R}^2}\}_{t\in T_{\text{test}}}$ ne sont pas homogènes sur l'environnement K. En effet, nous remarquons une disparité des résultats de précision, pour une configuration donnée, en fonction de l'endroit où se trouvent les positions tests X_t, $t \in T_{\text{test}}$. Les résultats des tableaux 5.1 et 5.2 sont donc relatifs à la suite de positions $\{X_t\}_{t\in T_{\text{test}}}$. Pour illustrer ces propos, le tableau 5.3 regroupe les résultats de précisions obtenus, après lissage des positions, pour chaque allée considérée (c.f. Figure 5.2 pour la numérotation des allées).

Configuration / Allée	1	2	3	4	5	6	7
1	$8.0m$	$3.1m$	$4.1m$	$4.6m$	$3.7m$	$3.9m$	$5.3m$
2	$8.3m$	$5.4m$	$5.4m$	$7.9m$	$7.0m$	$8.9m$	$6.5m$
3	$6.4m$	$4.9m$	$6.3m$	$6.8m$	$7.3m$	$6.3m$	$9.9m$
4	$9.0m$	$10m$	$5.3m$	$7.8m$	$9.0m$	$9.5m$	$9.2m$
5	$6.9m$	$4.9m$	$4.0m$	$3.6m$	$6.4m$	$5.1m$	$4.0m$

TABLE 5.3 – Précision (quantiles à 80%) par allée obtenue après lissage pour chaque configuration.

5.4 Conclusion

Au vu des résultats, la précision de notre système est dépendante du nombre et de l'emplacement des points d'accès utilisés, la précision optimale obtenue à l'issue des tests de précision ($4.8m$ dans 80% des cas) correspondant à la configuration ayant un nombre de points d'accès maximal. Nous avons aussi soulevé des problèmes de disparité de la précision en fonction de l'endroit de la carte considéré. Une augmentation du nombre de points d'accès dans les zones où la précision est la moins bonne pourrait permettre d'améliorer ces résultats.

Bibliographie

M. Arnaud and X. Emery. *Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatiques*. Hermès Science, 2000. ISBN 9782746201385.

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2 ; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

N. A. C. Cressie. *Statistics for Spatial Data.* Wiley-Interscience, revised Edition edition, January 1993. ISBN 0471002550.

Brian Ferris, Dirk Hähnel, and Dieter Fox. Gaussian Processes for Signal Strength-Based Location Estimation. In *Robotics : Science and Systems'06*, pages –1–1, 2006.

X. Guyon. Statistique spatiale. In *Conférence S.A.D.A. ' 07*, 2007.

D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society*, 52 :119–139, 1951.

G. Matheron. La théorie des variables régionalisées at ses appplications. Cahier du Centre de Morphologie Mathématique de fontainebleau, École des Mines, Paris, 1970.

Troisième partie

Détermination des cartes de propagation sans calibration préalable

L'objectif de cette partie est d'étudier la possibilité de se passer de la campagne de mesures préalable dans la construction d'un estimateur des cartes de propagations $f_\star$. En effet, cette campagne de mesures est contraignante, puisque, selon la taille de l'environnement à géolocaliser, elle peut prendre plusieurs heures. De plus, la fonction $f_\star$ peut être amenée à être modifiée dans le temps, par exemple lorsque des obstacles influant la propagation des ondes sont ajoutés ou lorsque leur emplacement est modifié. Ainsi, lorsqu'un estimateur de $f_\star$ est construit grâce à une campagne préliminaire de mesures, la prise en considération de ces changements ne peut se faire qu'en réitérant cette campagne régulièrement. L'enjeu ici est de construire cet estimateur sur la base des observations $\{Y_t\}_{t\in\mathbb{N}}$, en effet, toute modification affectant $f_\star$ affectera dans les mêmes proportions les mesures $\{Y_t\}_{t\in\mathbb{N}}$. Dans cette partie, nous nous attarderons sur l'estimation de $f_\star$ au sens du maximum de vraisemblance.

Chapitre 6

Introductions des chapitres 7 et 8

Sommaire

Deux approches seront étudiées dans les chapitres 7 et 8. La première, celle du chapitre 7 porte sur l'estimation de $f_\star$ dans le modèle du chapitre 5. La deuxième approche, celle du chapitre 8 porte sur la construction et la consistance d'un estimateur de $f_\star$ dans le cadre général où l'espace d'états n'est pas discrétisé.

6.1 Introduction du chapitre 7

Le chapitre 7 est une retranscription de l'article *Simultanous localisation and mapping problem in wireless sensor networks* (Dumont and Le Corff [2012b]), soumis à la revue *IEEE Signal Processing*. Ce travail a donné lieu à un d'un dépôt de brevet par l'entreprise Id Services (Dumont and Gassiat [2012]).

6.1.1 Position du problème

L'enjeu du chapitre 7 est l'estimation de la fonction $f_\star$ dans le modèle semi-paramétrique décrit dans le chapitre 5. Une discrétisation fine de l'environnement à géolocaliser est effectuée, de sorte que l'espace d'états K que l'on considère est fini. Nous supposons que la fonction $f_\star$, définie sur K, vérifie l'équation

$$f_\star = \mu_\star + \delta_\star \ ,$$

où $\mu_\star$ représente la propagation moyenne des ondes et vérifie, pour tout $j = 1, \ldots, \ell$ et tout x de K,

$$\mu_{\star,j}(x) \stackrel{\text{def}}{=} c^\star_{1,j} + c^\star_{2,j} \log \|O_j - x\|_{\mathbb{R}^2} \ ,$$

avec O_j, position du point d'accès j, et où $\delta_\star = \{\delta_{\star,j}\}_{j=1}^{\ell}$, représentant les perturbations des ondes dues aux obstacles, est une collection de processus Gaussiens sur K, indépendants , de moyennes nulles, et de matrices de covariances $\{\Sigma_j\}_{j=1}^{\ell}$ supposées ici connues. Notons $c^\star_1 = \{c^\star_{1,j}\}_{j=1}^{\ell}$, $c^\star_2 = \{c^\star_{2,j}\}_{j=1}^{\ell}$ et $\theta_\star = (c^\star_1, c^\star_2, \delta_\star, \sigma^2_\star)$ (on rappelle que $\sigma^2_\star$ représente la variance du bruit gaussien d'observation).

Nous optons pour une estimation *en ligne* des paramètres, c'est à dire en utilisant les observations $\{Y_t\}_{t\geq 0}$ de manière séquentielle. Cette démarche s'inspire des techniques de localisation et de cartographie simultanées (ou SLAM pour *Simultaneous Localization and Mapping*), notamment utilisées en robotique (Durrant-Whyte and Bailey [2006a,b]). La méthode d'inférence, décrite dans le chapitre 7, est basée sur l'algorithme EM en ligne, dont le principe est décrit dans la section suivante.

6.1.2 Algorithme EM en ligne

L'algorithme EM, décrit dans le chapitre 3, permet le calcul du prédicteur de maximum *a posteriori* $\widehat{\theta}_n$ construit sur la base des observations jusqu'au pas de temps n, $Y_{0:n}$. L'étape E de l'algorithme EM est effectuée grâce à l'algorithme *Forward-Backward*, parcourant l'ensemble des mesures $Y_{0:n}$. La mise en oeuvre de cet algorithme nécessite donc l'enregistrement préalable des mesures $Y_{0:n}$ qui seront alors parcourues à chaque itération de l'algorithme EM. Lorsque le nombre de mesures n'est pas *a priori* fini, les besoins, en terme de place mémoire, nécessaires pour mettre en place l'algorithme EM grandissent alors avec le nombre d'observations. La méthode d'estimation présentée dans cette section est issue des travaux de Cappé [2011], elle permet l'estimation séquentielle du paramètre $\theta_\star$, par la construction d'une suite d'estimateurs $\{\widehat{\theta}_t\}_{t\geq 0}$, de sorte que, pour tout $t \geq 0$, $\widehat{\theta}_t$ est calculé sur la base des observations jusqu'au temps t, notées $Y_{0:t}$. Au lieu d'enregistrer toutes les mesures $\{Y_t\}_{t\geq 0}$, cette méthode passe par le calcul récursif de ce que l'on appellera les statistiques exhaustives (ou suffisantes), occupant une place mémoire fixe, contrairement à l'algorithme EM classique, et résumant l'information contenue dans les données. Cette procédure d'estimation baptisée EM en ligne a été introduite dans Cappé and Moulines [2009] pour le cas des variables i.i.d.. Le cadre d'application de l'EM en ligne décrit par Cappé [2011] est le suivant. $(X_t, Y_t)_{t\in\mathbb{Z}}$ est généré selon un modèle de Markov caché stationnaire sur l'espace $\mathbb{X} \times \mathbb{Y}$, où $\mathbb{X}$ est supposé fini, et décrit par un paramètre inconnu $\theta_\star$ appartenant à un espace de paramètres Θ. Nous supposons que le modèle appartient à une famille exponentielle courbe, définie de la manière suivante,

Famille exponentielle. Il existe une fonction $s : \mathbb{X}^2 \times \mathbb{Y} \to \mathcal{S}$, appelée vecteur de statistiques exhaustives sur les données complètes, une fonction $h : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^+$ et deux fonctions différentiables

ψ et A sur Θ, telles que, pour tous x_{t-1}, x_t, y_t et θ,

$$p_\theta(x_t, y_t|x_{t-1}) = h(x_t, y_t) \exp\left(\langle \psi(\theta)\ ,\ s(x_{t-1}, x_t, y_t)\rangle\right) - A(\theta)\ , \tag{6.1}$$

où $\langle \cdot\ , \cdot \rangle$ désigne le produit scalaire, $p_\theta(\cdot, \cdot|x_{t-1})$ représente la densité de probabilité du couple (X_t, Y_t) conditionnellement à $X_{t-1} = x_{t-1}$, et $\mathcal{S}$ est l'espace des statistiques exhaustives.

Nous supposons de plus que l'étape M de l'algorithme EM est explicite, c'est à dire que,

Étape M explicite. pour toute statistique exhaustive $S \in \mathcal{S}$, l'équation sur θ définie par,

$$\nabla_\theta \psi(\theta) S - \nabla_\theta A(\theta) = 0\ , \tag{6.2}$$

où ∇_θ désigne le gradient, possède une unique solution notée $\overline{\theta}(S)$.

Dans cette classe de modèles, la k-ième récursion de l'algorithme EM peut alors être réécrite de la manière suivante.

Étape E. Calculer

$$S_{k+1} = \frac{1}{n} \mathbb{E}_{\theta_k} \left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n} \right]\ . \tag{6.3}$$

Étape M. Définir $\theta_{k+1} = \overline{\theta}(S_{k+1})$.

L'idée sous-jacente à l'algorithme EM en ligne décrit par Cappé [2011] est que la quantité définie par (6.3) peut être calculée récursivement, pour cela, introduisons la quantité intermédiaire suivante : pour tout $n \geq 0$, pour tout $\theta \in \Theta$ et pour tout $x \in \mathbb{X}$,

$$\rho_{n,\theta}(x) = \frac{1}{n} \mathbb{E}_\theta \left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n},\ X_n = x \right]\ .$$

Si $\phi_{n,\theta}$ désigne la densité de filtrage sous le paramètre θ, $(\forall x \in \mathbb{X},\ \phi_n(x) = \mathbb{P}_\theta(X_n = x|Y_{0:n}))$, alors on peut facilement vérifier que,

$$\frac{1}{n} \sum_{x \in \mathbb{X}} \phi_{n,\theta}(x) \rho_{n,\theta}(x) = \frac{1}{n} \mathbb{E}_\theta \left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n} \right]\ . \tag{6.4}$$

La proposition suivante explicite le calcul récursif des quantités $\{\rho_{n,\theta}\}_{n\geq 0}$ et $\{\phi_{n,\theta}\}_{n\geq 0}$.

Proposition 6.1.1 (Proposition 1 de Cappé [2011]). ***Initialisation.*** *Pour tout x de $\mathbb{X}$, poser*

$$\begin{aligned} \phi_{0,\theta}(x) &= \frac{\nu(x) g_\theta(x, Y_0)}{\sum_{x' \in \mathbb{X}} \nu(x') g_\theta(x', Y_0)}\ , \\ \rho_{0,\theta}(x) &= 0\ . \end{aligned}$$

Récursion. *Pour tout x de $\mathbb{X}$,*

$$\begin{aligned} \phi_{n+1,\theta}(x) &= \frac{\sum_{x' \in \mathbb{X}} \phi_{n,\theta}(x') q_\theta(x', x) g_\theta(x, Y_{n+1})}{\sum_{(x', x'') \in \mathbb{X}^2} \phi_{n,\theta}(x') q_\theta(x', x'') g_\theta(x'', Y_{n+1})}\ , \\ \rho_{n+1,\theta}(x) &= \sum_{x' \in \mathbb{X}} \left\{ \frac{1}{n+1} s(x', x, Y_{n+1}) + \left(1 - \frac{1}{n+1}\right) \rho_{n,\theta}(x') \right\} \frac{\phi_{n,\theta}(x') q_\theta(x', x)}{\sum_{x'' \in \mathbb{X}} \phi_{n,\theta}(x'') q_\theta(x'', x)}\ . \end{aligned}$$

Algorithm 3 Algorithme EM en ligne

1: Choisir une suite décroissante $\{\gamma_n\}_{n\geq 1}$, satisfaisant $\sum_{n\geq 1}\gamma_n = \infty$ et $\sum_{n\geq 1}\gamma_n^2 < \infty$. Choisir un paramètre d'initialisation $\widehat{\theta}_0$ et un nombre $n_{\min}$ d'observations avant la première mise à jour de $\widehat{\theta}$.

2: **Initialisation** Calculer, pour tout x de $\mathbb{X}$,

$$\widehat{\phi}_0(x) = \frac{\nu(x) g_{\widehat{\theta}_0}(x, Y_0)}{\sum_{x'\in\mathbb{X}} \nu(x') g_{\widehat{\theta}_0}(x', Y_0)} ,$$
$$\widehat{\rho}_0(x) = 0 .$$

3: **Récursion** Pour tout $n \geq 0$, calculer, pour tout $x \in \mathbb{X}$,

$$\widehat{\phi}_{n+1}(x) = \frac{\sum_{x'\in\mathbb{X}} \widehat{\phi}_n(x') q_{\widehat{\theta}_n}(x', x) g_{\widehat{\theta}_n}(x, Y_{n+1})}{\sum_{(x',x'')\in\mathbb{X}^2} \widehat{\phi}_n(x') q_{\widehat{\theta}_n}(x', x'') g_{\widehat{\theta}_n}(x'', Y_{n+1})} ,$$
$$\widehat{\rho}_{n+1}(x) = \sum_{x'\in\mathbb{X}} \left\{\gamma_{n+1} s(x', x, Y_{n+1}) + (1-\gamma_{n+1})\widehat{\rho}_{n+1}(x')\right\} \frac{\widehat{\phi}_n(x') q_{\widehat{\theta}_n}(x', x)}{\sum_{x''\in\mathbb{X}} \widehat{\phi}_n(x'') q_{\widehat{\theta}_n}(x'', x)}$$

4: Si $n \geq n_{\min}$, mettre à jours le paramètre selon

$$\widehat{\theta}_{n+1} = \overline{\theta}\left(\sum_{x\in\mathbb{X}} \widehat{\phi}_{n+1}(x)\widehat{\rho}_{n+1}(x)\right) ,$$

5: sinon, poser $\widehat{\theta}_{n+1} = \widehat{\theta}_n$.

Cette méthode de calcul récursif, à θ fixé, permet d'effectuer l'étape E de l'algorithme EM classique de manière *Forward*. L'algorithme 3 permet la mise à jours récursive de $\widehat{\theta}_n$ selon le principe décrit par la Proposition 6.1.1, il reprend l'algorithme EM en ligne de Cappé [2011].

Cet algorithme nécessite le calcul des distributions de filtrage approchées $\{\widehat{\phi}_n\}_{n\geq 0}$, ceci est rendu possible ici car l'espace d'états $\mathbb{X}$ est supposé fini. Cet algorithme a été étendu dans le cas d'espaces d'états généraux par l'utilisation d'algorithmes de Monte Carlo séquentiels (comme le filtre *bootstrap* présenté dans la section 5.2.2) permettant l'approximation de ces lois de filtrage (voir Cappé [2009], Del Moral et al. [2010]). Dans ce cas, la loi de filtrage $\widehat{\phi}_n$, $n \geq 0$, est approximée par un jeu de particules $\{\widehat{\xi}_n^p, \widehat{\omega}_n^p\}_{p=1}^N$, simulé selon la valeur $\widehat{\theta}_n$ du paramètre courant, grâce à l'algorithme SISR décrit dans la section 5.2.2,

$$\widehat{\phi}_n = \sum_{p=1}^N \widehat{\omega}_n^p \delta_{\widehat{\xi}_n^p} ,$$

où δ_x désigne la mesure de Dirac au point x. De même, $\widehat{\rho}_n$, $n \geq 0$, est approximé par l'ensemble de

quantités intermédiaires relatives à chaque particule $\{\widehat{\rho}_n^p\}_{p=1}^N$, où

$$\widehat{\rho}_n^p = \sum_{p'=1}^{N} \left\{ \gamma_n s(\widehat{\xi}_{n-1}^{p'}, \widehat{\xi}_n^p, Y_n) + (1-\gamma_n)\, \widehat{\rho}_{n-1}^{p'} \right\} \frac{\widehat{\omega}_{n-1}^{p'} q_{\widehat{\theta}_n}\left(\widehat{\xi}_{n-1}^{p'}, \widehat{\xi}_n^p\right)}{\sum_{p'=1}^{N} \widehat{\omega}_{n-1}^{p'} q_{\widehat{\theta}_n}\left(\widehat{\xi}_{n-1}^{p'}, \widehat{\xi}_n^p\right)} ,$$

de sorte que l'étape 4 de l'algorithme 3 peut être remplacée par,

$$\widehat{\theta}_n = \overline{\theta}\left(\sum_{p=1}^{N} \widehat{\omega}_n^p \widehat{\rho}_n^p \right) . \tag{6.5}$$

6.1.3 Contribution du chapitre 7

Dans le chapitre 7 nous décrirons un nouvel algorithme inspiré de l'algorithme EM en ligne, il est d'ailleurs facile de vérifier que notre modèle appartient bien à une famille exponentielle et que l'étape M de l'algorithme EM est explicite (*c.f.* calculs de la section 5.1.1). L'inconvénient avec l'utilisation de l'algorithme EM en ligne, comme décrit par l'algorithme 3, est qu'il est très instable lorsqu'il est utilisé sur des modèles complexes, dépendants d'un grand nombre de paramètres. Cette instabilité peut être due à des problème de convergence de $\{\widehat{\theta}_n\}_{n\geq 0}$ vers des minima locaux de ce qu'on appelle le contraste limite $c_{\theta_\star}$ défini, pour tout θ, par

$$c_{\theta_\star}(\theta) = \mathbb{E}_{\theta_\star}\left[\log L_\theta(Y_0|Y_{-\infty:-1})\right], \tag{6.6}$$

et limite $\mathbb{P}^\star$-presque sûre (sous certaines conditions) de la vraisemblance $\frac{1}{n}\log L_\theta(Y_0,\ldots,Y_n)$, que l'on cherche à maximiser avec l'algorithme EM (voir Baum et al. [1970], Douc et al. [2004]).

L'algorithme BOEM introduit dans Le Corff and Fort [2011], est à mi-chemin entre l'algorithme EM classique et l'algorithme EM en ligne. Cet algorithme est construit sur le même principe que l'algorithme EM en ligne, à la différence près que les mises à jour des paramètres ne sont pas effectuées à chaque pas de temps mais à des instants déterministes que l'on note $\{T_k\}_{k\geq 0}$. De plus, les quantités intermédiaires $\{\widehat{\rho}_n\}_{n\geq 0}$ sont ré-initialisées à 0 après chaque mise à jour du paramètre, de sorte que pour tout $k \geq 0$, et pour tout $t \in \{0,\ldots,T_{k+1}-T_k\}$, si $n = T_k + t$,

$$\forall x \in \mathbb{X},\ \widehat{\rho}_n(x) = \begin{cases} 0 & \text{si t=0} \\ \sum_{x'\in\mathbb{X}} \left\{ \frac{1}{t} s(x', x, Y_n) + (1-\frac{1}{t})\widehat{\rho}_{n-1}(x') \right\} \frac{\widehat{\phi}_n(x') q_{\widehat{\theta}_k}(x',x)}{\sum_{x''\in\mathbb{X}} \widehat{\phi}_n(x'') q_{\widehat{\theta}_k}(x'',x)} & \text{sinon} \end{cases} ,$$

où les densités de filtrages $\{\widehat{\phi}_n(x')\}_{x'\in\mathbb{X}}$ sont calculées sous le paramètre courant $\widehat{\theta}_k$. Ainsi, à chaque instant $t = T_k$, $k \geq 0$, la statistique exhaustive $\mathrm{S}_k = \sum_{x\in\mathbb{X}} \widehat{\phi}_{T_k}(x)\widehat{\rho}_{T_k}(x)$, construite sur la base du k^{ieme} block d'observations $Y_{T_{k-1}+1:T_k}$, est calculée et $\widehat{\theta}_k$ est défini par $\widehat{\theta}_k = \overline{\theta}(\mathrm{S}_k)$. Si l'on définit $\tau_0 = T_0$ et pour tout $k \geq 1$, $\tau_k = T_k - T_{k-1}$, Le Corff and Fort [2011] introduit un estimateur, dit moyennisé, de $\theta_\star$ en définissant, pour tout $k \geq 0$, $\widetilde{\mathrm{S}}_k = \frac{1}{T_k}\sum_{l=0}^{k} \tau_l \mathrm{S}_l$, et $\widetilde{\theta}_k = \overline{\theta}(\widetilde{\mathrm{S}}_k)$. Comme pour l'EM en ligne, une approximation de Monte Carlo peut être utilisée pour permettre l'approximation stochastique des statistiques exhaustives lorsque l'espace d'état n'est pas fini.

Dans le chapitre 7, nous étudions les performances de l'algorithme BOEM sur des simulations construites à partir du modèle de la section 5. L'algorithme *Bootstrap* est alors utilisé pour l'approximation stochastique des lois de filtrage. Nous construisons la suite d'estimateurs $\{\widehat{f}_k\}_{k\geq 0}$, suivant

l'algorithme BOEM et dans l'esprit des calculs de la section 5.1.1, ainsi que leur version moyennisée $\{\widetilde{f}_k\}_{k\geq 0}$. Sur ces simulations (*c.f.* figure 7.2(b)) nous observons une divergence flagrante de ces estimateurs avec cependant un meilleur comportement pour la suite d'estimateurs moyennisés $\{\widetilde{f}_k\}_{k\geq 0}$.

L'algorithme 5 du chapitre 7 introduit alors une étape supplémentaire, dite de stabilisation, afin d'exploiter le bon comportement de la suite d'estimateurs moyennisés $\{\widetilde{f}_k\}_{k\geq 0}$ pour éviter que les paramètres $\{\widehat{f}_k\}_{k\geq 0}$ ne finissent par diverger. Ainsi, régulièrement, nous substituons la valeur du paramètre moyennisé $\widetilde{f}_k$ à celle du paramètre courant $\widehat{f}_k$ (tous les N_b blocs d'observations). Nous observons alors sur les simulations (*c.f.* figure 7.3(b)) que l'étape de stabilisation permet d'éviter ces phénomènes de divergence (pour notre modèle) et que la suite d'estimateurs moyennisés $\{\widetilde{f}_k\}_{k\geq 0}$ semble converger vers le vrai paramètre $f_\star$.

Nous introduisons alors un deuxième système de particules $\{\widetilde{\xi}_t^p, \widetilde{\omega}_t^p\}_{p=1}^N,\ t\geq 0$, construit à partir de l'algorithme *bootstrap*, et utilisant la valeur du paramètre moyennisé courant $\widetilde{f}_k$. Nous définissons alors la suite de prédicteurs de maximum *a posteriori* $\{\widetilde{X}_t\}_{t\geq 0}$ tels que, pour tout $t\geq 0$,

$$\begin{aligned} \widetilde{X}_t &= \widetilde{\xi}_t^{p_t^{\max}}\ , \\ \text{où}\quad p_t^{\max} &= \underset{p=1,\dots,N}{\operatorname{argmax}}\, \widetilde{\omega}_t^p\ . \end{aligned}$$

Nous observons alors sur les figures 7.3 et 7.4, obtenues à partir de données simulées, que l'erreur en positionnement $\{\|\widetilde{X}_t - X_t\|_{\mathbb{R}^2}\}_{t\in\mathbb{N}}$ décroît elle aussi en fonction du nombre de mises à jour des paramètres. Nous avons alors soumis notre algorithme aux données réelles (*c.f.* figure 7.7). Le nombre de mesures à l'issue de la phase de collection de données (d'une durée de 7h environ à raison de 2 mesures par secondes) est d'environ 20000. Il faut savoir qu'en pratique, tous les points d'accès ne sont pas mesurés à chaque instant, le nombre de mesures par points d'accès varie de 5500 à 19000 environ dans les données que nous avons collectées. Notre algorithme BOEM avec étape de stabilisation doit donc être adapté de manière à estimer les paramètres $\{f_{\star,j}\}_{j=1}^\ell$ séparément. Le nombre de réestimation des cartes de propagations $\{\widetilde{f}_j\}_{j=1}^\ell$, varie alors de 2 à 7 selon le point d'accès considéré. La précision obtenue à l'issue de ces tests est cependant comparable à la précision que l'on obtient dans le même environnement, avec le même nombre de points d'accès, et en estimant $f_\star$ grâce à une campagne de mesures préalable.

6.1.4 Perspectives

Le Corff and Fort [2011] prouve, sous certaines conditions sur le modèle, la convergence des paramètres $\{\widehat{\theta}_k\}_{k\geq 0}$ produits par l'algorithmes BOEM, ou Monte Carlo BOEM, vers les points stationnaires de la fonction contraste limite (6.6), ainsi que la convergence des versions moyennisées $\{\widetilde{\theta}_k\}_{k\geq 0}$, lorsque la taille des blocs augmente. La convergence de l'algorithme BOEM avec étape de stabilisation reste à démontrer, mais il semblerait, sur nos simulations, que cet algorithme permette de limiter les phénomènes de dérives (convergence des estimateurs non moyennisés, $\{\widehat{\theta}\}_{n\geq 0}$, vers des minimas locaux, non globaux, de la fonction contraste) que l'on peut rencontrer, notamment lorsque l'espace de paramètres est de grande dimension. Les phénomènes de *label switching* ont, quant à eux, été traités, comme dans le chapitre 5, grâce au prior Gaussien sur la fonction $f_\star$. L'utilisation d'autres priors pourraient être étudiée sur données réelles. Nous pouvons d'ailleurs mentionner les travaux de

Gabriel et al. [2011], en statistiques spatiales, sur la détection de rupture pour les processus Gaussiens dont nous pourrions nous inspirer afin de détecter les ruptures dans la propagation des ondes lors de la traversée d'obstacles importants (murs épais, quantité élevée de métaux,...), provoquant de brusques baisses des puissances des signaux, qu'un prior gaussien, comme celui utilisé ici, aurait pour effet de lisser.

6.2 Introduction du chapitre 8

6.2.1 Position du problème

Le chapitre 8 est une retranscription de l'article *Nonparametric estimation in hidden Markov models* (Dumont and Le Corff [2012c]), soumis à la revue *Annals of Statistics*. Dans ce chapitre nous nous intéresserons à une méthode d'estimation non paramétrique dans une classe de modèles de Markov cachés. Nous considérons un processus aléatoire bivarié $\{X_k, Y_k\}_{k\in\mathbb{Z}}$ à valeurs dans $\mathbb{R}^m \times \mathbb{R}^\ell$, où $\{X_k\}_{k\in\mathbb{Z}}$ est une chaîne de Markov non observée sur un compact K d'intérieur non vide de $\mathbb{R}^m$. Le processus $\{Y_k\}_{k\in\mathbb{Z}}$, appelé processus d'observations, est supposé indépendant conditionnellement à $\{X_k\}_{k\in\mathbb{Z}}$, vérifiant que, pour tout $k_0 \geq 0$, la distribution de Y_{k_0} conditionnellement à $\{X_k\}_{k\in\mathbb{Z}}$ est une loi normale multivariée sur $\mathbb{R}^\ell$, de moyenne $f_\star(X_{k_0}) \in \mathbb{R}^\ell$ et de matrice de covariance $\sigma^2 I_\ell$ où I_ℓ représente la matrice identité de taille ℓ et σ^2 est un réel positif connu. La fonction $f_\star$ est une fonction de K à valeurs dans $\mathbb{R}^\ell$. L'enjeu principal du chapitre 8 est la construction et l'étude d'un estimateur non paramétrique $\widehat{f}_n$ de cette fonction $f_\star$.

6.2.2 Cadre et notations

Le cadre fixé pour cette étude est le suivant. Nous supposons que la chaîne de Markov $\{X_k\}_{k\in\mathbb{Z}}$ est stationnaire, homogène, isotrope et à valeurs dans un compact K de $\mathbb{R}^m$. Nous supposons que son noyau de transition est connu à facteur d'échelle, $a_\star > 0$, près et que ce noyau est à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^m$. Nous supposons l'existence d'une fonction positive connue q telle que cette densité soit donnée par $q_{a_\star}$ où, pour tout $a > 0$, q_a est défini, pour tous x et x' de K, par

$$q_a(x, x') \propto q\left(\frac{\|x - x'\|_{\mathbb{R}^2}}{a}\right) .$$

Nous supposons la connaissance d'un réel strictement positif a_- tel que $a_\star \geq a_-$. La fonction q est supposée strictement monotone. La stricte monotonie de q implique nécessairement la récurrence et la phi-irréductibilité de tous les noyaux de transitions Q_a de densité q_a par rapport à la mesure de Lebesgue λ sur K (*i.e.* pour tout borélien A de K d'intérieur non vide, et pour tout x de K, presque surement et conditionnellement à $X_0 = x$, la chaîne de Markov $\{X_k\}_{k\geq 1}$ retournera dans A en un temps fini). La récurrence et la phi-irréductibilité impliquent l'existence et l'unicité d'une loi invariante par Q_a (voir [Cappé et al., 2005, Théorème 14.2.25]), notée ν_a, pour tout $a > 0$. Nous pouvons d'ailleurs vérifier que ν_a possède une densité par rapport à la mesure de Lebesgue sur K, toujours notée ν_a, donnée par :

$$\forall x \in K,\ \nu_a(x) = \frac{\int_K q\left(\frac{\|x'-x\|_{\mathbb{R}^2}}{a}\right) \mathrm{d}x'}{\int_{K^2} q\left(\frac{\|x'-x''\|_{\mathbb{R}^2}}{a}\right) \mathrm{d}x'\mathrm{d}x''} .$$

Parmi les hypothèses de base sur le modèle, nous supposons que $f_\star$ est une fonction régulière, appartenant à un espace de Sobolev $W^{s,p}$, avec $s \in \mathbb{N}$ et $p \geq 1$. Si l'on définit, pour toute fonction $f : K \to \mathbb{R}^\ell$,

$$\|f\|_{W^{s,p}}^p \stackrel{\text{def}}{=} \sum_{j=1}^{\ell} \sum_{0 \leq |\alpha| \leq s} \|D^\alpha f_j\|_{\mathrm{L}_p}^p ,$$

où les α sont pris dans $\mathbb{N}^m$ et $|\alpha|$ désigne la somme des composantes de α, et où $D^\alpha f$ désigne la dérivée partielle d'ordre α de f. L'espace $W^{s,p}$ est alors défini par

$$W^{s,p} \stackrel{\text{def}}{=} \left\{ f : K \to \mathbb{R}^\ell \; ; \; \|f\|_{W^{s,p}}^p < \infty \right\} .$$

Soit $\left(\widehat{f}_n, \widehat{a}_n\right)$, estimateur de $(f_\star, a_\star)$, construit sur la base des mesures $Y_0, \ldots, Y_{2n-1}$, défini par :

$$\left(\widehat{f}_n, \widehat{a}_n\right) \stackrel{\text{def}}{=} \underset{f \in W^{s,p},\ a \geq a_-}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{k=0}^{n-1} \ln p_{f,a}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 I^2(f) \right\} , \tag{6.7}$$

où $\{\lambda_n\}_{n \geq 1}$ est une suite de réels positifs à déterminer, $I^2(f) \stackrel{\text{def}}{=} \|f\|_{W^{s,p}}^{\upsilon+1}$, avec $\upsilon > 0$ et où $p_{f,a}$ désigne la densité de probabilité du couple (Y_0, Y_1) défini, pour tous y_0 et y_1 de $\mathbb{R}^\ell$, par :

$$p_{f,a}(y_0, y_1) \stackrel{\text{def}}{=} \int_{K^2} \varphi(y_0 - f(x_0)) \varphi(y_1 - f(x_1)) \nu_a(x_0) q_a(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 ,$$

φ désignant la densité de la loi normale $\mathcal{N}(0, \sigma^2 I_\ell)$. Finalement, nous introduisons $\widehat{p}_n$, que l'on appellera estimateur du maximum de vraisemblance pénalisé (ou MLE pour *Maximum Likelihood Estimator*) que nous définissons par

$$\widehat{p}_n \stackrel{\text{def}}{=} p_{\widehat{f}_n, \widehat{a}_n} .$$

6.2.3 Hypothèses supplémentaires et résultats principaux

Nous nous intéresserons tout d'abord à l'identifiabilité de notre modèle, que l'on prouve sur l'espace de paramètres $\mathcal{C}^1 \times [a_-, \infty]$, nous verrons pourquoi dans cette section. Puis nous verrons un résultat de convergences en distance de Hellinger du MLE $\widehat{p}_n$, finalement nous établirons un résultat de consistance en probabilité de notre suite d'estimateurs $(\widehat{f}_n, \widehat{a}_n)$. Tout d'abord, introduisons les hypothèses supplémentaires utilisées pour établir ces résultats.

Hyp1 (i) K est homéomorphe à un sous ensemble convexe de $\mathbb{R}^m$.

(ii) K possède un frontière localement lipschitzienne.

K possède un frontière localement lipschitzienne si, quelque soit x sur la frontière ∂K de K, il existe un voisinage V de x dans ∂K qui est le graphe d'une fonction lipschitzienne.

Hyp2 $f_\star : K \to \mathrm{Im}(f_\star)$ est un diffeomorphisme.

On dit qu'une fonction $f : K \to \mathrm{Im}(f)$ est un diffeomorphisme s'il existe un voisinage ouvert V de K dans $\mathbb{R}^m$ et un difféomorphisme $\overline{f} : V \to \mathrm{Im}(\overline{f})$ tel que $\overline{f}|_V = f$.

Hyp3 $s > m/p + 1$.

Hyp4 $\upsilon > 2\ell$.

Lorsque l'on observe la définition de l'estimateur $\widehat{f}_n$, nous sommes en droit de nous demander à quel espace $\widehat{f}_n$ appartient. L'argmax de la définition (6.7) est pris sur toutes les fonctions de $W_{s,p}$, $\widehat{f}_n$ appartient alors à l'adhérence de cet espace. Cependant, sous l'hypothèse Hyp1 sur le compact K, et sous Hyp3, par un théorème de plongement des espaces de Sobolev ([Adams and Fournier, 2003, Theorem 6.3]), si $\mathcal{C}^1$ désigne l'ensemble des fonctions $\mathcal{C}^1$ sur K à valeurs dans $\mathbb{R}^\ell$, et si $\|\cdot\|_{\mathcal{C}^1}$ désigne la norme sur $\mathcal{C}^1$, définie, pour toute fonction $f \in \mathcal{C}^1$, par

$$\|f\|_{\mathcal{C}^1} = \sum_{j=1}^{\ell} \sup_{|\alpha|\leq 1} \sup_{x\in K} |D^\alpha f(x)| \ ,$$

alors $W^{s,p}$ est plongé de manière compacte dans $(\mathcal{C}^1, \|\cdot\|_{\mathcal{C}^1})$. Ainsi, $W^{s,p}$ peut être vu comme sous-ensemble de $\mathcal{C}^1$, et toute boule fermée de $W^{s,p}$ est un compact de $\mathcal{C}^1$. De plus, l'application identité étant linéaire, il existe un coefficient κ tel que pour tout f de $W_{s,p}$, $\|f\|_{\mathcal{C}^1} \leq \kappa \|f\|_{W^{s,p}}$. Par la définition (6.7) de $\widehat{f}_n$, nécessairement, $\|\widehat{f}_n\|_{W^{s,p}} < \infty$, et finalement, $\widehat{f}_n$ appartient à $\mathcal{C}^1$.

Dans le même esprit, l'estimateur $\widehat{a}_n$, défini par l'équation (6.7) peut valoir ∞. Nous devons alors étendre les définitions des quantités de la section 6.2.2 afin de définir correctement le MLE $\widehat{p}_n$. Par le Théorème de convergence dominé, quels que soient x_0, x_1 dans K, y_0, y_1 dans $\mathbb{R}^\ell$, et f fonction mesurable, $q_a(x_0, x_1)$, $\nu_a(x_0)$ et $p_{f,a}(y_0, y_1)$ convergent, lorsque a tend vers ∞, vers $q_\infty(x_0, x_1)$, $\nu_\infty(x_0)$ et $p_{f,\infty}(y_0, y_1)$, définis par :

$$\nu_\infty(x_0) \stackrel{\text{def}}{=} \mu(K)^{-1} \ , \ q_\infty(x_0, x_1) \stackrel{\text{def}}{=} \mu(K)^{-1} \ ,$$
$$p_{f,\infty}(y_0, y_1) \stackrel{\text{def}}{=} \mu(K)^{-2} \int \varphi(y_0 - f(x_0))\mathrm{d}x_0 \int \varphi(y_1 - f(x_1))\mathrm{d}x_1 \ .$$

Sous le noyau Q_∞, le processus $\{X_k\}_{k\in\mathbb{Z}}$ est alors i.i.d. de loi uniforme sur K. Les propriétés asymptotiques de $\widehat{p}_n$ seront étudiées en terme de distance de Hellinger entre $\widehat{p}_n$ et $p_{f_\star,a_\star}$. La distance de Hellinger est définie, pour tout couple de densités de probabilités (p_1, p_2) sur $\mathbb{R}^{2\ell}$, par

$$h(p_1, p_2) \stackrel{\text{def}}{=} \left[\frac{1}{2}\int_{\mathbb{R}^{2\ell}} \left(p_1^{1/2}(y, y') - p_2^{1/2}(y, y')\right)^2 \mathrm{d}y\mathrm{d}y'\right]^{1/2} \ .$$

Identifiabilité du modèle

Le premier résultat important que nous proposons dans le chapitre 8 est le Théorème 8.3.1, qui montre que, sous certaines hypothèses décrites dans la section précédente et dans une certaine mesure, il ne peut exister de paramètres $(f, b) \in \mathcal{C}^1 \times]0, \infty]$, autres que $(f_\star, a_\star)$, pouvant décrire la distribution du processus $\{Y_k\}_{k\in\mathbb{Z}}$. Introduisons tout d'abord la relation d'équivalence suivante. On

dit que deux fonctions f_1 et f_2 de K sont équivalentes s'il existe une isométrie ϕ de K telle que $f_1 = f_2 \circ \phi$, cette relation d'équivalence sera notée $f_1 \sim f_2$. Le théorème suivant reprend le Théorème 8.3.1 du chapitre 8,

Theorem 6.2.1. *Soit $f : K \to \mathbb{R}^\ell$ dans $\mathcal{C}^1$ et $0 < b \leq \infty$. Sous les hypothèses sur K et $\{X_k\}_{k\in\mathbb{Z}}$ de la section 6.2.2 et les hypothèses Hyp1 et Hyp2,*

$$\text{si } h(p_{f,b}, p_{f_\star,a_\star}) = 0 \text{ alors, } b = a_\star \text{ et } f \overset{\mathcal{I}}{\sim} f_\star.$$

Ce théorème est un résultat clé permettant de prouver la consistance de $(\widehat{f}_n, \widehat{a}_n)$, en effet, il nous permet dès lors de travailler sur le MLE $\widehat{p}_n$ car la consistance du MLE entrainera, par le Théorème 6.2.1, la consistance de notre estimateur.

Pour démontrer ce théorème, nous montrons tout d'abord que, pour tout couple (f, b) de paramètres vérifiant $h(p_{f,b}, p_{f_\star,a_\star}) = 0$, la fonction $\phi = f_\star^{-1} \circ f$ est bijective. La démonstration de ce résultat se fait en deux étapes, la première étape consiste à montrer que le jacobien de la fonction ϕ, qui est $\mathcal{C}^1$ par hypothèse, ne peux s'annuler (*c.f.* Lemme 8.5.1), la preuve de ce résultat repose sur la formule de changement de variable généralisée (*c.f.* [Evans and Gariepy, 1992, Theorem 2, p.99]) appliquée à la densité de probabilité du couple $(\phi(X_0), \phi(X_1))$, ainsi que sur le Théorème de Sard. Nous abordons la deuxième étape de la démonstration d'un point de vue topologique. L'hypothèse $h(p_{f,b}, p_{f_\star,a_\star}) = 0$ implique que ϕ, qui est donc un difféomorphisme local sur K, est nécessairement surjective, ϕ est alors un *revêtement* de K (*c.f.* [Lee, 2000, Chapitre 11]). Nous montrons alors dans le lemme 8.5.1 que, sous l'hypothèse de convexité de K, tout revêtement de K est nécessairement injectif, ceci achevant la démonstration que ϕ est bijective.

Nous terminons la démonstration du théorème par un argument de point fixe sur ϕ. L'existence d'un point fixe pour la fonction ϕ et la définition des noyaux $Q_a,\ a \geq a^-$, permettent de manipuler le noyau de transition de la chaîne de Markov $\{\phi(X_t)\}_{t\in\mathbb{N}}$ afin de prouver que pour tous x et x' de K,

$$\|x - x'\|_{\mathbb{R}^m} = \|\phi(x) - \phi(x')\|_{\mathbb{R}^m} ,$$

ceci achevant la preuve que ϕ est une isométrie de K.

Résultats de consistance

Nous nous intéressons désormais aux propriétés asymptotiques de $\widehat{p}_n$. Le Théorème 8.3.5 assure la convergence en distance de Hellinger de $\widehat{p}_n$ vers la vraie densité de probabilité, $p_{f_\star,a_\star}$, du couple (Y_0, Y_1). Il assure aussi que la norme de Sobolev de $\widehat{f}_n$ reste bornée lorsque n tend vers l'infini. Le Théorème 6.2.2 reprend ce résultat.

Theorem 6.2.2. *Sous les hypothèses sur K et $\{X_k\}_{k\in\mathbb{Z}}$ de la section 6.2.2 et sous les hypothèses Hyp1 et Hyp3-4, si la suite $\{\lambda_n\}_{n\in\mathbb{N}}$ vérifie les conditions suivantes,*

$$\lambda_n \underset{n\to+\infty}{\longrightarrow} 0 \text{ and } \lambda_n^2 n^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty , \tag{6.8}$$

alors,

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) = O_{\mathbb{P}_\star}(\lambda_n^2) \quad \text{and} \quad I^2(\widehat{f}_n) = O_{\mathbb{P}_\star}(1) . \tag{6.9}$$

Où l'on utilise la notation de Landau, $Z_n = O_{\mathbb{P}_\star}(\alpha_n)$, pour signifier que

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbb{P}_\star(Z_n \geq T\alpha_n) = 0 \ .$$

La démonstration de ce théorème utilise un schéma de preuve classique pour démontrer la consistance du MLE qui passe par un contrôle du processus empirique indexé par une classe de fonction $\mathcal{G}$. Nous définissons la classe de fonction $\mathcal{G}$ comme l'ensemble des fonctions $g_{p_{f,a}}$, $f \in W^{s,p}$, $a \geq a^-$, où, pour toute densité de probabilité p sur $\mathbb{R}^{2\ell}$,

$$g_p \stackrel{\text{def}}{=} \frac{1}{2}\log\frac{p + p_{f_\star,a_\star}}{2p_{f_\star,a_\star}} \ .$$

Le contrôle de l'erreur en distance d'Hellinger du MLE se fait grâce à l'inégalité basique (*c.f.* [Van De Geer, 2009, Lemme 10.5]) suivante :

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) + 4\lambda_n^2 I^2(\widehat{f}_n) \leq 16 \int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) + 4\lambda_n^2 I^2(f_\star) \ , \tag{6.10}$$

où $\mathbb{P}_n$ désigne la distribution empirique basée sur les observations $\{(Y_{2k}, Y_{2k+1})\}_{k=0}^{n-1}$. Notons, pour toute fonction g de $\mathcal{G}$, $\nu_n(g) = \sqrt{n}\int g\mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)$. Tout l'enjeu de la démonstration réside alors dans le contrôle des déviations du processus empirique $\{\nu_n(g)\}_{g\in\mathcal{G}}$. Le contrôle que nous avons obtenu apparaît dans la Proposition 8.3.6 du chapitre 8, réécrite ci-après,

Proposition 6.2.3. *Sous les mêmes hypothèses que le Théorème 6.2.2, il existe des constantes Σ, K et T telles que, pour tout $x > 0$,*

$$\mathbb{P}_\star\left\{\sup_{f\in W^{s,p},\ a\geq a_-} \frac{|\nu_n(g_{p_{f,a}})|}{I^2(f)\vee 1} \geq T + x\right\} \leq Ke^{-\Sigma x} \ . \tag{6.11}$$

Le Théorème 6.2.2 est alors une conséquence directe de la Proposition 6.2.3 et de l'équation (6.10). Pour démontrer la Proposition 6.2.3, nous avons d'abord établi une inégalité de déviation sur les processus empiriques indexés par les classes de fonctions $\mathcal{G}_M$, $M \geq 1$ définis, pour tout $M \geq 1$, par

$$\mathcal{G}_M = \left\{g_{p_{f,a}} \ ; \ \|f\|_{W^{s,p}} \geq M \ , \ a \geq a_-\right\} \ .$$

Pour ce faire, les outils propres au cas i.i.d. (comme l'inégalité de Hoeffding ou de Bernstein pour le contrôle de la concentration du processus empirique) n'ont pu être utilisés, en effet le processus $\{(Y_{2k}, Y_{2k+1})\}_{k\geq 0}$ n'est pas indépendant. Cependant nous montrons que ce processus est β-mélangeant (cette notion est définie dans Doukhan et al. [1995]), nous permettant d'établir une inégalité maximale pour ν_n (contrôle de $\mathbb{E}\left(\sup_{g\in\mathcal{G}_M}|\nu_n(g)|\right)$), ceci en appliquant [Doukhan et al., 1995, Théorème 3]. Pour établir cette inégalité maximale, une analyse préalable de l'entropie à crochet de la classe de fonction $\mathcal{G}_M$ est requise. L'inégalité de déviation est finalement obtenue grâce à un contrôle de la concentration de $\sup_{g\in\mathcal{G}_M}|\nu_n(g)|$, utilisant les récents travaux de Adamczak and Bednorz [2012], qui nous permettent d'établir des inégalités de concentration pour les fonctionnelles additives de chaînes de Markov. L'inégalité de déviation sur le processus empirique, indexé par $\mathcal{G}_M$,

que nous obtenons est la suivante, il existe quatre constantes positives K_1, K_2, K_3 et C, telles que, pour tout $M \geq 1$, tout $n \geq 1$ et tout $t \geq C/\sqrt{n}$,

$$\mathbb{P}_\star \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq K_3 M^{\upsilon+1} + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{6.12}$$

La démonstration de l'inégalité (6.12) est détaillée dans le papier supplémentaire Dumont and Le Corff [2012a], retranscrit dans le chapitre 9. L'équation (6.11) est alors obtenue en appliquant une procédure classique de *peeling*, consistant au découpage de la classe de fonctions $\mathcal{G}$ en anneaux disjoints $\{\mathcal{G}_M \setminus \mathcal{G}_{M-1}\}_{M \geq 1}$, et en appliquant l'inégalité (6.12) à chacun de ces anneaux.

Finalement, la consistance en probabilité de l'estimateur $(\widehat{f}_n, \widehat{a}_n)$ peut être démontrée en ayant une nouvelle fois recours à l'inégalité basique (6.10) et grâce aux théorèmes 6.2.2 et 6.2.1. Ce résultat de consistance est établi par le Théorème 8.3.7 du chapitre 8 retranscrit ci-après. Notons tout d'abord $\mathcal{F}_\star$ la classe d'équivalence de $f_\star$ sous la relation $\sim$. Définissons de plus $d_{\mathcal{C}^1}$, la distance associée à la norme $\|\cdot\|_{\mathcal{C}^1}$. Alors,

Theorem 6.2.4. *Sous les hypothèses sur K et $\{X_k\}_{k \in \mathbb{Z}}$ de la section 6.2.2, et en supposant Hyp1-4, si la suite $\{\lambda_n\}_{n \in \mathbb{N}}$ vérifie*

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

Alors,

$$d_{\mathcal{C}^1}(\hat{f}_n, \mathcal{F}_\star) \underset{n \to +\infty}{\longrightarrow} 0 \quad et \quad \widehat{a}_n \underset{n \to +\infty}{\longrightarrow} a_\star \text{ en } \mathbb{P}_\star - \text{probabilité} , \tag{6.13}$$

Résultats de simulations

Dans la section 8.4 nous nous intéresserons à la construction de l'estimateur $\widehat{f}_n$ dans le cas où le paramètre $a_\star$ est supposé connu et où $p = 2$. Le Théorème 6.2.4 impose la condition $\upsilon > 2\ell$, cependant comme nous en discuterons dans la section suivante, cette condition, dépendant de notre contrôle d'entropie a de grande chance d'être sous-optimale. Nous choisirons donc, dans la section 8.4, $\upsilon = 1$. Dans cas, $\widehat{f}_n$ est défini comme maximisant, sur $W^{s,2}$, la fonction

$$T : \ f \mapsto \frac{1}{n} \sum_{k=0}^{n-1} \ln p_{f,a_\star}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 \|f\|_{W^{s,2}}^2 .$$

Nous choisissons de construire $\widehat{f}_n$ à l'aide de l'algorithme EM, construisant ainsi une suite d'estimateurs $\left\{ \widehat{f}_n^p \right\}_{p \geq 0}$, où, pour tout $p \geq 0$, $\widehat{f}_n^{p+1}$ est défini comme maximisant la quantité intermédiaire :

$$f \mapsto Q(f; \widehat{f}_n^p) \overset{\text{def}}{=} \frac{1}{n} \sum_{k=0}^{n} \mathbb{E}_{\widehat{f}_n^p} \left[\ln p_f \left(X_{2k}, Y_{2k}, X_{2k+1}, Y_{2k+1} \right) \middle| Y_{2k}, Y_{2k+1} \right] - \lambda_n^2 \|f\|_{W^{s,2}}^2 ,$$

où p_f est donnée, pour tous x_0, y_0, x_1, y_1, par

$$p_f(x_0, y_0, x_1, y_1) = \nu_{a_\star}(x_0) \varphi(y_0 - f(x_0)) q_{a_\star}(x_0, x_1) \varphi(y_1 - f(x_1)) .$$

Le choix $p = 2$ et $\upsilon = 1$ a pour conséquence que la fonction $f \mapsto Q(f; \widehat{f}_n^p)$ est facilement différentiable et, comme nous le verrons dans la section 8.4, $\widehat{f}_n^p$ peut être défini comme solution d'une équation aux dérivées partielle (EDP) exprimée sous forme variationnelle. Dans le cas où $K = [0, 1]$, la résolution de cette EDP revient à résoudre une équation différentielle avec conditions aux bords . Nous avons résolu cette équation différentielle dans le cas $s = 2$ (voir le détail des calculs dans le papier supplémentaire chapitre 9), les résultats obtenus sur données simulées (présentées sur les figures 8.1 et 8.2) montrent la décroissance des erreurs d'estimations $\{\|f_\star - \widehat{f}_n\|_{\mathrm{L}_2}\}_{n\geq 0}$ et $\{\|f_\star - \widehat{f}_n\|_{\mathrm{L}_\infty}\}_{n\geq 0}$, cependant, la figure 8.2 montre que ces erreurs ne tendent pas vers 0 lorsque le nombre d'observations tend vers l'infini. Ce phénomène semble être du à des problèmes d'approximation au bord du domaine $K = [0, 1]$ lors de notre résolution de l'équation différentielle. Outre ces problèmes de bords, les simulations semblent indiquer que notre estimateur converges vers la vrai fonction $f_\star$, bien que le choix de υ ne corresponde pas à l'hypothèse Hyp4.

6.2.4 Discussion

Discussion sur les hypothèses

Commençons tout d'abord par l'hypothèse Hyp4. Comme nous venons d'en discuter, les résultats de simulations semblent indiquer que l'hypothèse Hyp4 est trop forte, puisque $\upsilon = 1$ semble convenir. Cette hypothèse provient du contrôle de l'entropie à crochet pour les classes de fonctions $\mathcal{G}_M$ (*c.f.* définition 8.6.2). Pour cela, nous utilisons un résultat de Nickel and Potscher [2001] permettant le contrôle de l'entropie à crochets pour les classes de fonctions, à valeurs réelles, de type Besov ou Sobolev. Cependant, nous appliquons ce résultat composante par composante sur l'espace $\mathbb{R}^\ell$, rendant irrémédiablement l'entropie de la classe de fonctions $\mathcal{G}_M$ dépendante de ℓ. Une approche différente de ce problème de contrôle d'entropie à crochets pourrait permettre d'alléger les hypothèses sur υ.

L'hypothèse Hyp1(ii) permet, par le Théorème de Stein [Adams and Fournier, 2003, Theorem 5.24], le prolongement de toute fonction f de $\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$, bornées ($f$ bornée ainsi que ses dérivées partielles), en une fonction $\overline{f}$ de $\mathcal{C}^1(V, \mathbb{R}^\ell)$ où V est un voisinage ouvert de K dans $\mathbb{R}^\ell$.

L'hypothèse Hyp1(i) est nécessaire pour appliquer un théorème de point fixe (Théorème de Schauder) dans la preuve d'identifiabilité. Si l'on arrive à prouver l'identifiabilité sans argument de point fixe, cette hypothèse peut être allégée en "K simplement connexe" (condition nécessaire et suffisante pour la démonstration du lemme 8.3.4 démontrant que les revêtements de K sont nécessairement injectifs).

L'hypothèse Hyp2 suppose que $f_\star$ est un difféomorphisme sur son image ou, dit autrement, que $f_\star$ est un plongement de K dans $\mathbb{R}^\ell$. Cette hypothèse impose $m \leq \ell$. De plus, lorsque les dimensions des espaces de départ et d'arrivée de $f_\star$ vérifient $\ell \geq 2m+1$, le Théorème de Whitney (*c.f.* [Laudenbach, 1996, Section 5.5.2]), assure que l'espace des plongements de K dans $\mathbb{R}^\ell$ est un ouvert dense de $\mathcal{C}^1(K, \mathbb{R}^\ell)$, en revenant à la problématique de la géolocalisation sur $K \subset \mathbb{R}^2$, ceci entraîne qu'il suffit que le nombre de points d'accès soit plus grand que 5 pour être "certain" que la fonction de propagation $f_\star$ soit un plongement.

Conclusion et perspectives

Le chapitre 8 traite de l'estimation non-paramétrique de la fonction $f_\star$ dans le modèle spatio-temporel à états latents

$$Y_t = f_\star(X_t) + \epsilon_t \ ,$$

lorsque $\{X_t\}_{t\in\mathbb{N}}$ est une chaîne de Markov non-observée, homogène, stationnaire et isotrope de noyau de transition connu à facteur d'échelle près $a_\star$. Nous y construisons un estimateur $(\widehat{f}_n, \widehat{a}_n)$ de $(f_\star, a_\star)$, maximisant un critère de pseudo-vraisemblance pénalisé, dont nous prouvons la consistance en probabilité sous certaines conditions sur le modèle. Nous avons aussi étudié les propriétés asymptotiques de ce que nous avons appelé le MLE, $\widehat{p}_n$, estimateur de la densité de probabilité de (Y_0, Y_1). Nous avons démontré que la distance d'Hellinger entre $\widehat{p}_n$ et la vraie densité de (Y_0, Y_1) tend vers 0 en probabilité à une vitesse λ_n^2, plus lente que $n^{-1/2}$ mais aussi proche de $n^{-1/2}$ que désiré. Un des axes de recherche serait donc de montrer que le choix $\lambda_n^2 = n^{-1/2}$ dans la définition 6.7 permet d'obtenir la consistance du MLE, qui convergerait alors vers la vraie densité à une vitesse en $n^{-1/2}$.

Nous pourrions aussi envisager d'établir une inégalité oracle pour le MLE. L'établissement d'une telle inégalité reposerait sur un contrôle plus fin de la concentration du processus empirique, passant par une analyse au niveau local de ce processus (comme le contrôle des variances de $\nu_n(g_{p_{f,a}})$, $f \in W^{s,p}$, $a \geq a_-$).

Une autre perspective de recherche serait d'étendre les résultats obtenus, notamment en terme d'identifiabilité, à une plus grande classe de chaînes de Markov, par exemple, en tentant de généraliser l'hypothèse d'isotropie du noyau de transition de la chaîne de Markov $\{X_t\}_{t\in\mathbb{N}}$.

Finalement, il serait intéressant d'étudier les propriétés asymptotiques de l'estimateur du maximum de vraisemblance pénalisé :

$$(\widetilde{f}_n, \widetilde{a}_n) \stackrel{\text{def}}{=} \operatorname*{argmax}_{f,a} \left\{ \log p_{f,a}(Y_{0:n}) - \mathrm{pen}(n, f) \right\} \ ,$$

où $p_{f,a}(Y_{0:n})$ représenterait ici la vraisemblance des observations sous les paramètres f et a, définie, pour tout $n \geq 0$, et tout $y_{0:n} \in \mathbb{Y}$, par

$$p_{f,a}(y_{0:n}) \stackrel{\text{def}}{=} \int_{x_{0:n}} \nu_a(x_0) \prod_{k=0}^{n-1} q_a(x_k, x_{k+1}) \prod_{k=0}^{n} \varphi(y_k - f(x_k)) \mathrm{d}x_{0:n} \ .$$

L'analyse de l'estimateur $(\widetilde{f}_n, \widetilde{a}_n)$ nécessiterait cependant un contrôle des fluctuations de cette vraisemblance qui semble difficile à réaliser dans notre cadre non paramétrique.

Bibliographie

R. Adamczak and W. Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. arXiv :1201.3569v1, Jan 2012.

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Number vol. 140 in Pure and Applied Mathematics. Academic Press, 2003. ISBN 9780120441433.

L.E. Baum, T.Petrie, G.Soules, and N.Weiss. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41 (1) :164–171, 1970. ISSN 00034851. doi : 10.2307/2239727.

O. Cappé. Online sequential Monte Carlo EM algorithm. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

O. Cappé. Online EM algorithm for Hidden Markov Models. *To appear in J. Comput. Graph. Statist.*, 2011.

O. Cappé and E. Moulines. Online Expectation Maximization Algorithm for Latent Data Models. *J. Roy. Statist. Soc. B*, 71(3) :593–613, 2009. doi : 10.1111/j.1467-9868.2009.00698.x.

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2 ; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

M. Del Moral, A. Doucet, and S.S Singh. Forward smoothing using sequential Monte Carlo. Preprint, Dec 2010.

R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Stat.*, 32(5) :2254–2304, 2004. doi : 10.1214/009053604000000021.

P. Doukhan, P. Massart, and E. Rio. Invariance principle for absolutely regular processes. *Annales de l'Institut Henri Poincaré*, 31 :393–427, 1995.

T. Dumont and E. Gassiat. Procédé de mise à jour continu d'un paramètre représentatif d'une grandeur physique dépendant de sa localisation, et dispositif associé. Patent Provis. app. num. : 1000146721, 2012.

T. Dumont and S. Le Corff. Supplement paper to nonparametric estimation in hidden Markov models. Technical report, 2012a.

T. Dumont and S. Le Corff. Simultaneous localization and mapping problem in wireless sensor networks. Technical report, 2012b.

T. Dumont and S. Le Corff. Nonparametric estimation in hidden Markov models. Technical report, 2012c.

H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping : part I. *Robotics Automation Magazine, IEEE*, 13(2) :99 –110, june 2006a. ISSN 1070-9932. doi : 10.1109/MRA.2006.1638022.

H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping (SLAM) : part II. *Robotics Automation Magazine, IEEE*, 13(3) :108 –117, sept. 2006b. ISSN 1070-9932. doi : 10.1109/MRA.2006.1678144.

L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, 1992. ISBN 9780849371578.

Edith Gabriel, Denis Allard, and Jean-Noël Bacro. Estimating and testing zones of abrupt change for spatial data. *Statistics and Computing*, 21(1) :107–120, January 2011. ISSN 0960-3174. doi : 10.1007/s11222-009-9151-x. URL `http://dx.doi.org/10.1007/s11222-009-9151-x`.

F Laudenbach. Topologie différentielle, 1996. Notes de cours, École polytechnique (Palaiseau).

S. Le Corff and G. Fort. Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Technical report, arXiv, 2011.

J.M. Lee. *Introduction to Topological Manifolds*. Graduate Texts in Mathematics. Springer, 2000. ISBN 9780387950266.

R. Nickel and B.M. Potscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov and Sobolev type. *J. Theor. Probab.*, 20 :177–199, 2001.

S.A. Van De Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2009. ISBN 9780521123259.

Chapter 7

Simultaneous localisation and mapping problem in wireless sensor networks

Thierry Dumont and Sylvain Le Corff

Sommaire

Abstract

Mobile device localization in wireless sensor networks is a challenging task. It has already been addressed when the WiFI propagation maps of the access points are modeled deterministically. However, this procedure does not take into account the environmental dynamics and also assumes an offline human training calibration. In this paper, we propose a semiparametric model: these maps are made of a parametric model combined with a non parametric perturbation field which represents the influence of the environment. The device localization is dealt with using Sequential Monte Carlo methods and relies on the estimation of the propagation maps. This inference tasks is performed online, i.e. using the observations sequentially, with a recently proposed online Expectation Maximization based algorithm. The performance of the algorithm are illustrated through Monte Carlo experiments.

Keywords: Indoor localization, Received Signal Strength Indicator, WiFi, Signal Propagation, Simultaneous localization and Mapping.

7.1 Introduction

Wireless sensor networks (Gaura et al. [2010]) generally consist of a data acquisition network and a data distribution network, monitored and controlled by a management center. These networks have many applications such as environmental monitoring (Barrenetxea et al. [2008]) or target tracking (Lau et al. [2009], Bahl and Padmanabhan [2000], Chen et al. [2005]). In this paper, we consider a WiFi communication network made up of a mobile device (such as a hand held mobile computer or a smartphone), a server and WiFi access points (APs). We are interested in the estimation of the localization of the mobile device in the environment using the signal strength of the surrounding APs. The mobile device collects the power of the signals and sends the data to the server which uses them to build an estimator of the device's position. The key step to provide such an estimator is to understand the behaviour of the WiFi signal strength for different positions in the environment. However, predicting the propagation of WiFi signals in an indoor environment is challenging since they are subject to many perturbations (*e.g.* shadowing, reflection...).

Two main techniques exist to approximate the WiFi signal propagation map of each AP: the first ones use deterministic models based on the localization and characteristics of the surrounding APs as well as the localization of the obstacles involved in the environment, see for instance Gorce et al. [2007]. Other famous techniques are based on a previous hand made offline training phase in which a human operator performs a site survey by measuring the received signal strength indicator (RSSI) from different APs at some fixed sampled points, see Bahl and Padmanabhan [2000], Evennou and Marx [2006]. However, representing the RF indoor propagation map using a deterministic model is challenging since several obstacles cannot be taken into account. On the contrary, the site survey method allows to build an accurate estimation of the signal strength, but only for a finite number of sampled points. Nevertheless, Ferris et al. [2006] provides a method to extend these measures to the entire map using Gaussian processes techniques.

In this paper, we propose an estimation method that does not require any calibration procedure. The propagation maps are estimated online (*i.e.* without storing the observations) using the data sent by the mobile device. Any modification in the way the WiFi signals propagate inside the environment (due to new obstacles for instance) affects the data sent by the mobile device. Then, while these changes deteriorate the accuracy of localization systems using fixed estimators of the propagation maps, our system learns these changes by taking them into account in the construction of our map estimators. Thus, as illustrated Section 7.5.2, the accuracy of our localization method improves with time instead of degrading.

A semiparametric statistical model is used: the propagation maps are made of a parametric average indoor model in addition of a non parametric perturbation field. This model combines a prior knowledge on the signal propagation with random perturbations due to the obstacles. Based on the data collected by the mobile device, parameters and perturbation field estimators can be defined. We simultaneously provide an estimator for the device position. The procedure relies on an online Expectation-Maximisation (EM) based algorithm for the estimation of the propagation maps and on particle filtering for the estimation of the device position.

The structure of this paper is the following. Section 7.2 describes the model and defines the notations. Section 7.3 presents the online EM algorithm and Section 7.4 gives a general algorithm for online inference in our Simultaneous localization and Mapping (SLAM) problem. Section 7.5 illustrates this algorithm with numerical experiments.

7.2 Model and assumption

Let $\{X_t\}_{t\geq 1}$ be the cartesian coordinates of the mobile device in a two-dimensional compact space. This continuous environment is discretized into a grid map, denoted by C, for purposes of numerical computation. It is assumed that $\{X_t\}_{t\geq 1}$ is a Markov chain taking values in C with initial distribution ν and Markov transition probability density function given, for all $(x, x') \in \mathcal{C}^2$, by

$$q(x, x') \propto \mathrm{e}^{-\|x-x'\|^2/a} , \tag{7.1}$$

where $a \in \mathbb{R}_+^\star$ depends on the average speed of the mobile and is assumed to be known and $\|\cdot\|$ denotes the usual euclidean norm in $\mathbb{R}^2$. Let $|\mathcal{C}|$ be the cardinality of $\mathcal{C}$ and $F^\star$ the $B \times |\mathcal{C}|$ matrix where $F^\star_{j,x}$ is the j-th AP expected signal strength at position x. At each time step t, the mobile device measures and sends to the server the observation Y_t taking values in $\mathbb{R}^B$. For all $t \geq 0$, the observation Y_t is given by

$$Y_t \stackrel{\text{def}}{=} F^\star_{\cdot,X_t} + \varepsilon_t , \tag{7.2}$$

with $\{\varepsilon_t\}_{t\geq 0}$ a sequence of i.i.d Gaussian random vectors with mean 0 and covariance matrix $\Sigma \stackrel{\text{def}}{=} \sigma^2 I_B$ (where I_B is the identity matrix of size $B \times B$).

The position of the B APs are assumed to be known and denoted by $\{O_j\}_{j=1}^B$. In order to take into account the perturbations in the signal propagation (due to the fact that radio waves are prone to shadowing, reflections and so on), we propose the following decomposition of $F^\star$: for all $x \in \mathcal{C}$ and all $j \in \{1, \cdots, B\}$,

$$F^\star_{j,x} \stackrel{\text{def}}{=} \mu^\star_{j,x} + \delta^\star_{j,x} . \tag{7.3}$$

For any $j \in \{1, \cdots, B\}$, $\mu^\star_{j,\cdot}$ is the average indoor propagation and is such that for all $x \in \mathcal{C}$, $\mu^\star_{j,x}$ only depends on the distance between x and O_j. In the sequel, we use the so-called Friis transmission equation, see Friis [1946], given by,

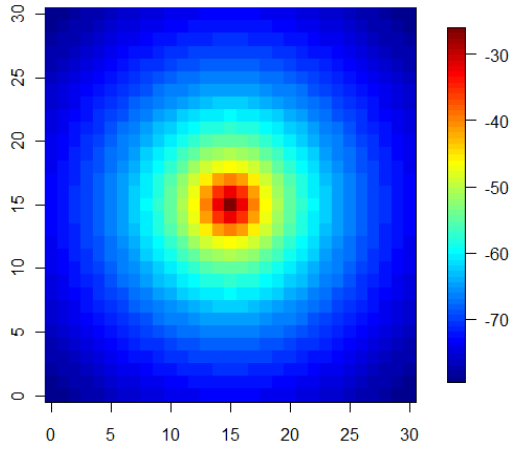
$$\mu^\star_{j,x} \stackrel{\text{def}}{=} c^\star_{1,j} + c^\star_{2,j} \log \|x - O_j\| , \tag{7.4}$$

where $c^\star_{1,j}$ and $c^\star_{2,j}$ are parameters depending on the environment and log is the logarithm to the base e.

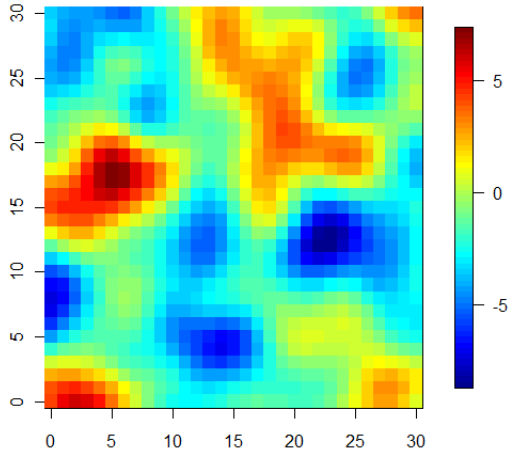
$\delta^\star_j$ is an additive term due to random perturbations such as walls effects (a similar model of WiFi propagation maps using Gaussian processes can be found in Ferris et al. [2006]). It is assumed that $\{\delta^\star_j\}_{j=1}^B$ are independent Gaussian vectors with mean 0 and known covariance matrix Σ_j. In the sequel, for any matrix A, A^T denotes the transpose of A. The probability density function of $\delta^\star$ is denoted by π and is given, for all $\delta \in \mathbb{R}^{B|\mathcal{C}|}$, by

$$\pi(\delta) \propto \exp\left\{ -\frac{1}{2} \sum_{j=1}^{B} \delta_j^T \Sigma_j^{-1} \delta_j \right\} .$$

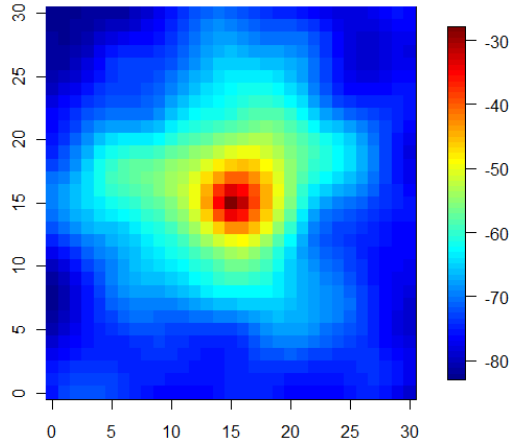
For any $B \times \mathbb{R}^{|\mathcal{C}|}$ matrix A, we use the shorthand notation A_j for the vector $\{A_{j,x}\}_{x\in\mathcal{C}}$. Figure (7.1) represents the functions $\mu^\star_j$, $\delta^\star_j$ and $F^\star_j$ defined on the grid $\mathcal{C} = \{0, \ldots, 30\} \times \{0, \ldots, 30\}$. The parameters used in this figure are $O_j = (15, 15)$, and $c^\star_{1,j}$, $c^\star_{2,j}$ and Σ_j are given in Section 7.5, their values were calibrated after a measurement campaign in an office environment.



(a) $\mu_j^\star$.



(b) $\delta_j^\star$.



(c) $F_j^\star = \mu_j^\star + \delta_j^\star$.

Figure 7.1: Example of spatial representations of the functions $\mu_j^\star$, $\delta_j^\star$ and $F_j^\star = \mu_j^\star + \delta_j^\star$ (in dBm)

In the sequel, we write $\theta^\star \stackrel{\text{def}}{=} (c_1^\star,\ c_2^\star,\ \delta^\star,\ \sigma^{\star,2})$, where $c_1^\star \stackrel{\text{def}}{=} \{c_{1,j}^\star\}_{j=1}^B$, $c_2^\star \stackrel{\text{def}}{=} \{c_{2,j}^\star\}_{j=1}^B$ and $\delta^\star \stackrel{\text{def}}{=} \{\delta_j^\star\}_{j=1}^B$. For any $x \in \mathcal{C}$, the distribution of Y_t conditionally to $X_t = x$ has a density with respect to the Lebesgue measure on $\mathbb{R}^B$ given, for all $y \stackrel{\text{def}}{=} (y_1, \cdots, y_B) \in \mathbb{R}^B$, by

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

Therefore, $\{X_t\}_{t\geq 0}$ is the hidden process of a hidden Markov model observed through the process $\{Y_t\}_{t\geq 0}$. The estimation of the mobile device's position X_t relies on the knowledge of the map $F^\star$. Unfortunately, as mentioned above, the indoor wave propagation is too complex to have an explicit $F^\star$. The observations $\{Y_t\}_{t\geq 1}$ are used to estimate simultaneously the mobile device's position and the map $F^\star$. This simultaneous localization and mapping problem may be seen as an instance of inference in hidden Markov models. For any positive integer n, any observation set $(y_1, \ldots, y_n)$, shortly denoted by $y_{1:n}$ and any parameter $\theta = (c_1, c_2, \delta, \sigma^2)$, the likelihood of the observations $L_\theta(y_{1:n})$ is given by:

$$L_\theta(y_{1:n}) \stackrel{\text{def}}{=} \sum_{x_{1:n}\in\mathcal{C}^n} \nu(x_1) g_\theta(x_1, y_1) \prod_{t=2}^n q(x_{t-1}, x_t) g_\theta(x_t, y_t) , \tag{7.5}$$

where $\nu = \{\nu(x)\}_{x\in\mathcal{C}}$ is an initial distribution on $\mathcal{C}$. Let n be a positive integer and $Y_{1:n}$ be a set of observations, we set as the estimator of $\theta^\star$, the maximum a posteriori estimator defined as $\operatorname{argmax}_\theta n^{-1}\ell_\theta(Y_{1:n})$, where:

$$\ell_\theta(Y_{1:n}) \stackrel{\text{def}}{=} \log L_\theta(Y_{1:n}) + \log \pi(\delta) . \tag{7.6}$$

The next section provides a description of the EM algorithm and of online EM algorithms for the computation of maximum likelihood estimators (without the penalty term). In Section 7.4, we explain how such techniques can be used in our framework.

7.3 Online EM

The EM algorithm is a well-known iterative algorithm to perform maximum likelihood estimation in hidden Markov models Dempster et al. [1977]. Each iteration of this algorithm consists in a E-step where the expectation of the complete data log-likelihood (log of the joint distribution of the states and the observations) conditionally to the observations is computed; and a M-step, which updates the parameter estimate. Except for simple models the E-step is intractable and has to be approximated e.g. by Monte Carlo methods (see e.g. Cappé et al. [2005], Fort and Moulines [2003]).

Let $Y_{1:n}$ be a fixed set of observations and $\hat{\theta}$ be the current parameter estimate.

i) The E-step consists in evaluating the conditional expectation

$$Q_{\hat{\theta}}(Y_{1:n};\theta) = \mathbb{E}_{\hat{\theta}}\left[\frac{1}{n}\log p_\theta(X_{1:n}, Y_{1:n})\middle| Y_{1:n}\right] , \tag{7.7}$$

where $\log p_\theta(X_{1:n}, Y_{1:n})$ is the complete data log-likelihood and $\mathbb{E}_{\hat{\theta}}[\cdot|Y_{1:n}]$ is the conditional expectation given $Y_{1:n}$ when the parameter's value is $\hat{\theta}$.

ii) The M-step updates the current value $\hat{\theta}$ taking the parameter θ maximizing (7.7).

The EM algorithm is of practical interest when the model belongs to the curved exponential family which states that there exist functions $S : \mathbb{X}^2 \times \mathbb{Y} \to \mathcal{S} \subset \mathbb{R}^d$, $\phi : \Theta \to \mathbb{R}$ and $\psi : \Theta \to \mathbb{R}^d$ such that

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

where $\langle \cdot, \cdot \rangle$ denotes the scalar product on $\mathbb{R}^d$. Moreover, it is assumed that there exists a continuous function $\bar{\theta} : \mathcal{S} \to \Theta$ s.t. for any $s \in \mathcal{S}$,

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

In this case the intermediate quantity defined by (7.7) can be written

$$Q_{\hat{\theta}}(Y_{1:n}; \theta) = \phi(\theta) + \left\langle \mathbb{E}_{\hat{\theta}} \left[\frac{1}{n} \sum_{t=1}^{n} S(X_{t-1}, X_t, Y_t) \middle| Y_{1:T} \right], \psi(\theta) \right\rangle \ . \tag{7.8}$$

Therefore, the E-steps amounts to computing only one conditional expectation $\mathbb{E}_{\hat{\theta}} \left[\frac{1}{n} \sum_{t=1}^{n} S(X_{t-1}, X_t, Y_t) \middle| Y_{1:n} \right]$ when the current parameter's value is $\hat{\theta}$. The M-step relies simply on the evaluation of $\bar{\theta}$ at this conditional expectation. This two steps process is repeated till convergence. However, when the observations are obtained sequentially or when the E-step relies on a large data set, the EM algorithm might become impractical. *Online* variants of the EM algorithm have been proposed to obtain parameter estimates each time a new observation is available. In the case of independent and identically distributed (i.i.d.) observations, Cappé and Moulines [2009] proposed the first EM based online algorithm. The E-step amounts to computing intermediate quantities known as *sufficient statistics* (see below for a explicit definition) and Cappé and Moulines [2009] proposed to replace these computations by a stochastic approximation step. When both the observations and the states take a finite number of values (resp. when the state-space is finite) an online EM-based algorithm was proposed by Mongillo and Denève [2008] (resp. by Cappé [2011]). These algorithms combine an online approximation of the filtering distributions of the hidden states and a stochastic approximation step to compute an online approximation of the sufficient statistics. This has been extended to the case of general state-space models with Sequential Monte Carlo algorithms (see Cappé [2009], Del Moral et al. [2010] and Le Corff et al. [2011]). More recently, Le Corff and Fort [2011] proposed a block online algorithm in which the parameter estimate is kept fixed on block of observations. The parameter's update then occurs at the end of each block.

In this paper, we use the online variant of the EM algorithm introduced in Le Corff and Fort [2011] to perform the parameter estimation and to solve the localization problem presented above. This algorithm, called the *Block Online EM* (BOEM) algorithm relies on the ability to compute sequentially quantities of the form:

$$\mathsf{S}_{n,\tau}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta \left[\frac{1}{\tau} \sum_{t=1}^{\tau} S(X_{t+n-1}, X_{t+n}, Y_{t+n}) \middle| Y_{n:n+\tau} \right] \ .$$

Such quantities are called sufficient statistics. The BOEM algorithm uses a sequence of block-sizes $\{\tau_k\}_{k \geq 0}$. Define $T_0 \stackrel{\text{def}}{=} 0$ and, for any $k \geq 0$, $T_k \stackrel{\text{def}}{=} \sum_{i=1}^{k} \tau_i$. Let k be a positive integer,

within each block of observations $Y_{T_k+1:T_{k+1}}$, the parameter's value $\widehat{\theta}_k$ is kept fixed and the sufficient statistic $\mathsf{S}_{T_k,\tau_{k+1}}(\widehat{\theta}_k)$ is computed sequentially. The estimate $\widehat{\theta}_{k+1}$ is computed at the end of the block $Y_{T_k+1:T_{k+1}}$ through the evaluation of the function $\bar{\theta}$.

Unlike the traditional use of the EM algorithm where the conditional expectations are computed using forward-backward techniques, Cappé [2011], Del Moral et al. [2010] and Le Corff and Fort [2011] rely on recursive computations of the conditional expectations. Indeed, let $\phi_{n,\theta}$ denotes the filtering distribution of X_n given the observations $Y_{1:n}$ when the parameter's value is θ:

$$\forall x \in \mathbb{X} \ , \ \phi_{n,\theta}(x) = P_\theta \left(X_n = x | Y_{1:n} \right) \ .$$

Following Cappé [2011], Del Moral et al. [2010], defining for all $x \in \mathbb{X}$ and all $\theta \in \Theta$,

$$\rho_{n,\theta}(x) = \mathbb{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} S(X_{t-1}, X_t, Y_t) \middle| Y_{1:n}, \ X_n = x \right] \ ,$$

we have,

$$\mathbb{E}_\theta \left[\frac{1}{n} \sum_{t=1}^{n} S(X_{t-1}, X_t, Y_t) \middle| Y_{1:n} \right] = \sum_{x \in C} \phi_{n,\theta}(x) \rho_{n,\theta}(x) \ .$$

Proposition 1 of Cappé [2011] illustrates the usefulness of this decomposition. Let ν be an initial distribution for the Markov chain $\{X_t\}_{t \geq 0}$ on $\mathcal{C}$.

Proposition 1 (of Cappé [2011]).
Initialisation:
For all $x \in \mathcal{C}$ and all $\theta \in \Theta$,

$$\begin{aligned} \phi_{1,\theta}(x) &= \frac{\nu(x) g_\theta(x, Y_1)}{\sum_{x' \in \mathcal{C}} \nu(x') g_\theta(x', Y_1)} \ , \\ \rho_{1,\theta}(x) &= 0 \ . \end{aligned}$$

Recursion:
For all $t \geq 2$ and all $x \in \mathcal{C}$,

$$\phi_{t,\theta}(x) = \frac{\sum_{x' \in \mathcal{C}} \phi_{t-1,\theta}(x') q(x'x) g_\theta(x, Y_t)}{\sum_{(x',x'') \in \mathcal{C}^2} \phi_{t-1,\theta}(x') q(x'x'') g_\theta(x'', Y_t)} \ , \tag{7.9}$$

$$\rho_{t,\theta}(x) = \sum_{x' \in \mathcal{C}} \left\{ \frac{1}{t} s(x', x, Y_t) + \left(1 - \frac{1}{t} \right) \rho_{t-1,\theta}(x') \right\} . \frac{\phi_{t-1,\theta}(x') q(x', x)}{\sum_{x'' \in \mathcal{C}} \phi_{t-1,\theta}(x'') q(x'', x)} \ . \tag{7.10}$$

Except in simple models (linear Gaussian models and finite state-space HMM), this algorithm requires forward computations which are not available in closed form and which have to be approximated, e.g. using sequential Monte Carlo methods (see Cappé [2009], Del Moral et al. [2010]). In this case, $\phi_{t,\theta}$ is approximated by weighted samples $\{\widehat{\xi}_t^p, \widehat{\omega}_t^p\}_{p=1}^N$ such that $\widehat{\phi}_{t,\theta}(x) = \sum_{p=1}^{N} \widehat{\omega}_t^p \delta_{\xi_t^p}(x)$. In

the sequel, $\{\widehat{\xi}_t^p\}_{p=1}^N$ will be referred to as the particle set at time step t. Plugging this approximation in (7.10) yields:

$$\rho_t^p = \sum_{\ell=1}^{N} \widehat{\omega}_{t-1}^{\ell} q(\widehat{\xi}_{t-1}^{\ell}, \widehat{\xi}_t^p) \times \frac{\frac{1}{t} s(\widehat{\xi}_{t-1}^{\ell}, \widehat{\xi}_t^p, Y_t) + (1 - \frac{1}{t})\rho_{t-1}^{\ell}}{\sum_{\ell=1}^{N} \widehat{\omega}_{t-1}^{\ell} q(\widehat{\xi}_{t-1}^{\ell}, \widehat{\xi}_t^p)} , \tag{7.11}$$

where ρ_t^p is the approximation of ρ_t evaluated at $\widehat{\xi}_t^p$. At each time step, the new population of particles is built from the previous population using Algorithm 4 referred to as the *bootstrap filter*, see e.g. Cappé et al. [2005]. The Bootstrap filter combines sequential importance sampling and sampling importance resampling steps to produce a set of random particles with associated importance weights. Implementations of such procedures are detailed in Cappé et al. [2005], Cappé [2001], Del Moral [2004], Doucet and Johansen [2009].

Algorithm 4 Bootstrap_filter_recursion (BFR)

Require: $\{\xi_{t-1}^{\ell}, \omega_{t-1}^{\ell}\}_{\ell=1}^N$, Y_t, θ.
1: **for** $p = 1$ to N **do**
2: Draw I in $1, \ldots, N$ with probabilities proportional to $\{\omega_{t-1}^{\ell}\}_{\ell=1}^N$.
3: Sample $\xi_t^p \sim q(\xi_{t-1}^I, \cdot)$.
4: Set $\omega_t^p \propto g_\theta(\xi_t^p, Y_t)$.
5: **end for**
6: **return** $\{\xi_t^p, \omega_t^p\}_{p=1}^N$

This leads to the Algorithm 5 presented below. Algorithm 5 is the adaptation of the BOEM to our model. It recursively updates the parameter θ at the end of each block. The BOEM proposed in Le Corff and Fort [2011] also introduced an averaged estimate based on a weighted mean of all the sufficient statistics computed in the past. It is proved in Le Corff and Fort [2011] that this averaged estimator has an optimal rate of convergence. Lines 21 to 25 of Algorithm 5 computes this averaged sufficient statistics and line 26 computes the sequence $\{\widetilde{\theta}_k\}_{k\geq 0}$ of map estimates based on the averaged statistics. The BOEM algorithm is adapted by introducing a second particle system $\{\widetilde{\xi}_t^p, \widetilde{\omega}_t^p\}_{p=1}^N$. This additional particle system is generated using the averaged parameter estimate. As this estimate is supposed to be more accurate than the original estimator computed on each block, we use the second system of particles to build a better estimator of the device's position. At each time step, we then compute two estimators of the device's position, one for each particle system. Both of them are set as the particle with the greatest importance weight. Line 18 performs the update of the sufficient statistics and line 19 the parameter's update at the end of the block.

Finally, in lines 28 to 30 of Algorithm 5, we add a stabilization step (which is not in the original BOEM) which only consists in regularly replace the original map estimate by the averaged one. This step is needed to ensure the convergence as detailed in Section 7.5.

7.4 Application of the algorithms to the SLAM in wireless networks

In our framework, the objective is the maximisation of the penalized loglikelihood (7.6). This task can be performed using a similar technique as the one described in Section 7.3 since the additional

penalty term only appears in the definition of the function $\bar{\theta}$. Define, for any $(x, y) \in \mathcal{C} \times \mathbb{R}^B$ and any $j \in \{1, \ldots, B\}$,

$$\begin{aligned} s_1(x) &\stackrel{\text{def}}{=} \{1_{x'}(x)\}_{x' \in \mathcal{C}} \ , \\ s_{2,j}(x, y) &\stackrel{\text{def}}{=} \{1_{x'}(x) y_j\}_{x' \in \mathcal{C}} \ , \\ s_{3,j}(y) &\stackrel{\text{def}}{=} y_j^2 \ . \end{aligned}$$

The constant a being known, and since our model belongs to the curved exponential family, the penalized intermediate quantity can be written, up to an additive constant, as:

$$Q_{\hat{\theta}}(Y_{1:n}; \theta) = -\frac{1}{2n} \sum_{j=1}^{B} \delta_j^T \Sigma_j^{-1} \delta_j - \frac{B}{2} \log \sigma^2 - \sum_{j=1}^{B} \frac{\left\{ \mathsf{S}_{3,j} - 2 \left\langle \mathsf{S}_{2,j}, F_j \right\rangle + \left\langle \mathsf{S}_1, F_j^2 \right\rangle \right\}}{2\sigma^2} \ , \tag{7.12}$$

where, $F_j^2 \stackrel{\text{def}}{=} \{F_{j,x}^2\}_{x \in \mathcal{C}}$ and

$$\mathsf{S}_1 \stackrel{\text{def}}{=} \frac{1}{n} \mathbb{E}_{\widehat{\theta}} \left[\sum_{t=1}^{n} s_1(X_t) \middle| Y_{1:n} \right] \ ,$$

and, for all $j \in \{1, \ldots, B\}$,

$$\begin{aligned} \mathsf{S}_{2,j} &\stackrel{\text{def}}{=} \mathbb{E}_{\widehat{\theta}} \left[\frac{1}{n} \sum_{t=1}^{n} s_{2,j}(X_t, Y_t) \middle| Y_{1:n} \right] \ , \\ \mathsf{S}_{3,j} &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} s_{3,j}(Y_t) \ . \end{aligned}$$

For any $\mathsf{S} = \left(\mathsf{S}_1, \ \{\mathsf{S}_{2,j}\}_{j=1}^B, \ \{\mathsf{S}_{3,j}\}_{j=1}^B \right) \in [0,1]^{|\mathcal{C}|} \times \mathbb{R}^{|\mathcal{C}| \times B} \times \left(\mathbb{R}_+^\star\right)^{|\mathcal{C}| \times B}$, we denote by $\bar{\theta}(\mathsf{S}, n)$ one of the parameter $\theta = (c_1, c_2, \delta, \sigma^2)$ maximizing the expression:

$$-\sum_{j=1}^{B} \frac{\left\{ \mathsf{S}_{3,j} - 2 \left\langle \mathsf{S}_{2,j}, F_j \right\rangle + \left\langle \mathsf{S}_1, F_j^2 \right\rangle \right\}}{2\sigma^2} + \frac{1}{n} \log \pi(\delta) - \frac{B}{2} \log \sigma^2 \ .$$

As mentioned in Section 7.2, for any $j \in \{1, \ldots, B\}$, F_j is written as $F_j \stackrel{\text{def}}{=} \mu_j + \delta_j$. In these experiments, the whole set of parameters ($c_1^\star$, $c_2^\star$ and $\sigma^{\star,2}$) and the unknown perturbation Gaussian fields $\{\delta_j^\star\}_{j=1}^B$ are estimated using Algorithm 5. For all $j \in \{1, \ldots, B\}$, we write $D_j \stackrel{\text{def}}{=} \{\log \|x -$

$O_j\|\}_{x\in\mathcal{C}}$ and

$$\begin{aligned}
M_{0,j} &\stackrel{\text{def}}{=} \left[\operatorname{diag}(\mathsf{S}_1) + \frac{\sigma^2}{n+1}\Sigma_j^{-1}\right] , \\
M_{1,j} &\stackrel{\text{def}}{=} \operatorname{diag}(\mathsf{S}_1)\left[I - M_{0,j}^{-1}\operatorname{diag}(\mathsf{S}_1)\right] , \\
M_{2,j} &\stackrel{\text{def}}{=} I - \operatorname{diag}(\mathsf{S}_1)M_{0,j}^{-1} , \\
W_{1,j} &\stackrel{\text{def}}{=} \mathbb{1}^T M_{1,j}\mathbb{1} , \\
W_{2,j} &\stackrel{\text{def}}{=} \mathbb{1}^T M_{1,j} D_j , \\
W_{3,j} &\stackrel{\text{def}}{=} D_j^T M_{1,j} D_j , \\
d_j &\stackrel{\text{def}}{=} W_{1,j}W_{4,j} - W_{2,j}^2 .
\end{aligned}$$

Thus, $\theta = \bar{\theta}(\mathsf{S}, n)$ is given, by $\theta = (c_1, c_2, \delta, \sigma^2)$ where, for all $j \in \{1, \ldots, B\}$,

$$\begin{aligned}
c_{1,j} &= d_j^{-1}\left[W_{3,j}\mathbb{1}^T - W_{2,j}D_j^T\right] M_{2,j}\mathsf{S}_{2,j} , \\
c_{2,j} &= d_j^{-1}\left[-W_{2,j}\mathbb{1}^T + W_{1,j}D_j^T\right] M_2\mathsf{S}_{2,j} , \\
\delta_j &= M_{0,j}\left[\mathsf{S}_{2,j} - \operatorname{diag}(\mathsf{S}_1)(c_{1,j}\mathbb{1} + c_{2,j}D_j)\right] , \\
F_j &= c_{1,j}\mathbb{1} + c_{2,j}D_j + \delta_j
\end{aligned}$$

and

$$\sigma^2 = \frac{1}{B}\sum_{j=1}^{B}\left\{F_j^T \operatorname{diag}(\mathsf{S}_1)F_j - 2\mathsf{S}_{2,j}^T F_j + \mathsf{S}_{3,j}\right\} .$$

7.5 Experiments

7.5.1 Simulated data

In this section, the performance of the proposed BOEM algorithm is illustrated with simulated data. All experiments are performed on the grid $\mathcal{C} = \{0, \ldots, 30\} \times \{0, \ldots, 30\}$. We use $B = 17$ APs, each AP being modelled by the same coefficients $c_1^\star$ and $c_2^\star$, see (7.4),

$$\forall j \in \{1, \ldots, B\},\ c_{1,j}^\star = -26 \quad \text{and} \quad c_{2,j}^\star = -17.5 .$$

For all $j \in \{1, \ldots, B\}$, Σ_j is a Gaussian covariance function defined by $\Sigma_j(x, x') \stackrel{\text{def}}{=} v_1 * \exp(-|x - x'|^2/v_2)$ with $v_1 = 10$ and $v_2 = 18$. The variance of the observation noise is $\sigma^{\star,2} = 25$. The variance of the transition kernel defined in (7.1) is chosen such that $a = 6$.

All runs are started with the same initial estimates $\theta^0 = \left(c_1^0, c_2^0, \delta^0, \sigma^{0,2}\right)$ where, $\delta^0 = 0$

$$\forall j \in \{1, \ldots, B\},\ c_{1,j}^0 = -10 ,\ c_{2,j}^0 = -30 \quad \text{and} \quad \sigma^{0,2} = 30 .$$

The number of particles $N = 25$ is kept fixed and the initial position of each particle is chosen randomly and uniformly in $\mathcal{C}$. For each map $F_j^\star$, the estimation error is set as the normalized L_1 error, such that the distance of a given map F_j from the true map $F_j^\star$ is

$$\epsilon_j \stackrel{\text{def}}{=} \frac{1}{|\mathcal{C}|} \sum_{x \in \mathcal{C}} \left| F_{j,x} - F_{j,x}^\star \right| \ ,$$

and the error displayed is the mean over all maps:

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

The block sizes are given by

$$\forall k \in \mathbb{N},\ \tau_k = 10k + 500 \ .$$

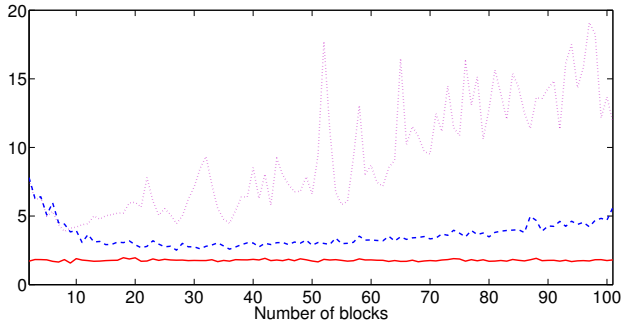
On each block, the localization error is set as the 0.8-quantile of the distance between the true localization and the estimated position. Figure 7.2 displays the error on the estimation of the maps and on the localization when the stabilization step in Algorithm 5 is omitted (lines 29 to 31). This case corresponds to the BOEM algorithm. The localization part is dealt with using two different procedures.

- *Nonaveraged estimate*: the estimate is given with the original particle system (see line 10 of Algorithm 5).
- *Averaged estimate*: the estimate is given with a second particle system run with the average estimation of the map (see line 11 of Algorithm 5).

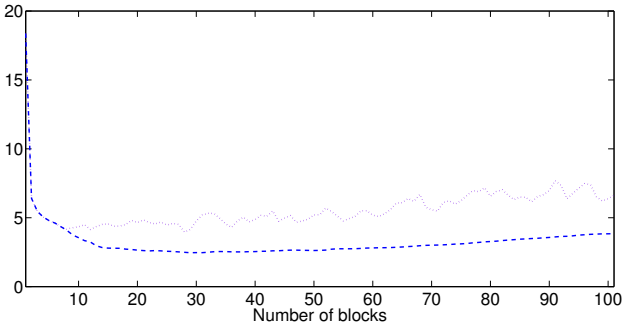
In order to give fair results, the optimal estimate is shown, i.e. the estimated position given with a particle system run with the true maps $F_j^\star$, $j \in \{1, \ldots, B\}$.

As shown in Figure 7.2(a) the estimated position does not converge as the number of blocks (i.e. as the number of estimations) increases. After 50 blocks (about 40000 observations) the position, which is badly estimated, does not provide good map estimates which increases the error on the averaged map estimate. Figure 7.2(b) displays the error on the map estimate. It is clear that both the estimate and its averaged version do not converge. This convergence problem of the BOEM algorithm can be due to the curse of dimensionality that can occur when the number of parameters to estimate is high. Moreover, the higher the parameter space dimension is, the more likely EM based algorithms are prone to converge towards local minima (see Cappé et al. [2005]). To overcome this difficulty, we propose to use the good behaviour of the averaged map estimate during the first 50 blocks. The map estimate is regularly replaced by its averaged version, see lines 28 to 30 of Algorithm 5. This will prevent the map estimate from diverging and thus, this will reduce the error on the estimated position. In Figure 7.3, this stabilization process is performed each time $N_b = 5$ blocks have been used. As shown by Figure 7.3(a) and Figure 7.3(b), this greatly improves the performance of the estimation of both the maps and the localization. Hence, the proposed algorithm is based on this stabilization procedure and uses the averaged position estimate to perform the localization part.

Figures 7.4 and 7.5 illustrate the performance of the algorithm for the localization and for the estimation of the maps over 50 independent Monte Carlo runs. In Figure 7.4, the optimal localization error (i.e. when the maps are known) is also displayed. The convergence of the localization error to the optimal error is almost reached after 100 blocks (about 100000 observations). Similarly, the

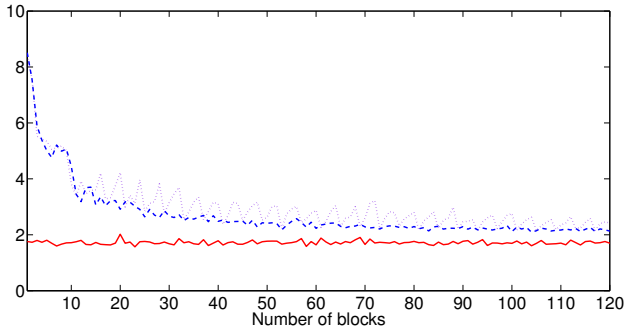


(a) 0.8-quantile of the distance between the true localization and the estimated position. The localization error is given with the nonaveraged estimate (dotted line), the averaged estimate (dashed line) and the optimal estimate (bold line).

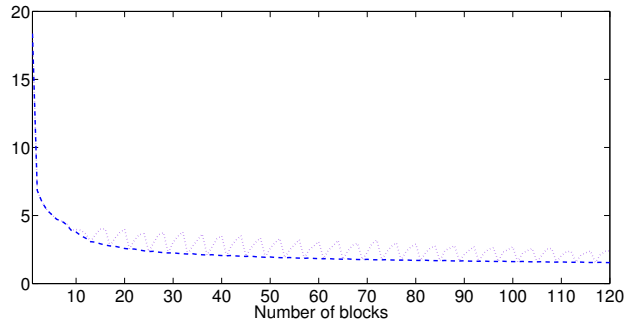


(b) Mean L_1 error on the map estimate with the initial estimate (dotted line) and the averaged estimate (dashed line).

Figure 7.2: Errors on the map estimation and localization processes with the original algorithm.



(a) 0.8-quantile of the distance between the true localization and the estimated position with the stabilization process. The localization error is given with the nonaveraged estimate (dotted line), the averaged estimate (dashed line) and the optimal estimate (bold line).



(b) Mean L_1 error on the map estimate with the initial estimate (dotted line) and the averaged estimate (dashed line) with the stabilization process.

Figure 7.3: Errors on the map estimation and localization procedure with the stabilized algorithm.

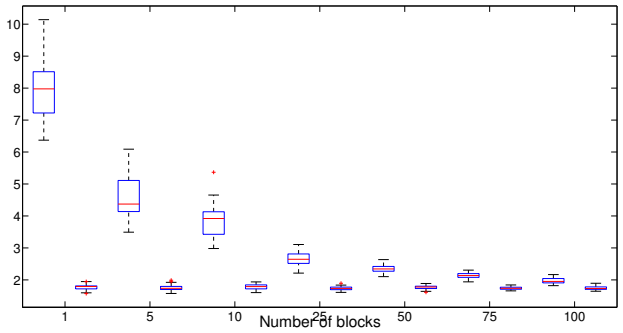


Figure 7.4: Boxplots of the localization error given by the stabilized algorithm with the averaged estimate (left) and the optimal estimate (right) as a function of the number of blocks.

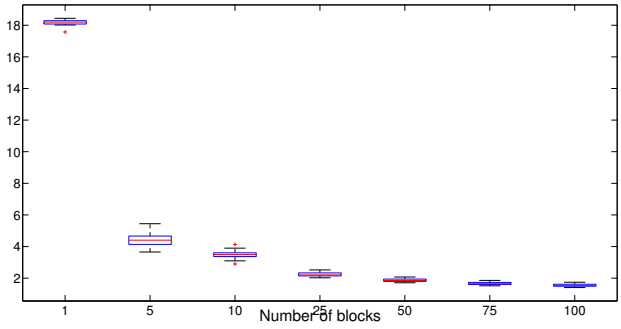


Figure 7.5: Boxplots of the mean L_1 error on the map estimate with the stabilized algorithm and the averaged estimate as a function of the number of blocks.

error for the estimation of the maps given by the averaged algorithm goes on decreasing after 100 blocks (the decrease is slower after 75 blocks).

7.5.2 True data

In this section, the behaviour of our SLAM algorithm is illustrated in a real situation. 10 access points are set up in an office environment (Figure 7.6 represents a map of this environment as well as the position of the access points). The map is discretized using a grid $\mathcal{C} \subset [0,30] \times [0,30]$. The variance $\sigma^{\star,2}$ is assumed to be known and its value ($\sigma^{\star,2} = 25dBm^2$) is calibrated using a measurement campaign at a fixed position. Around $T = 20000$ measures of the RSSI have been made on the map using a WiFi device. Algorithm 5 produces position estimates but we do not have a direct access to the real position and thus cannot observe the localization error. To overcome this difficulty, a test data sample is built by producing measures along 12 paths in the environment such that, for each measure, the associated position is registered. The test data sample is made of $T_{\text{test}} = 1100$ measures: $\{X_t^{\text{test}}, Y_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$. The test data sample is used to compare the localization accuracy provided by different values of the parameter m. Note a major difference between the model given in Section 7.2 and the real data situation. For any measure Y sent by the device, only several

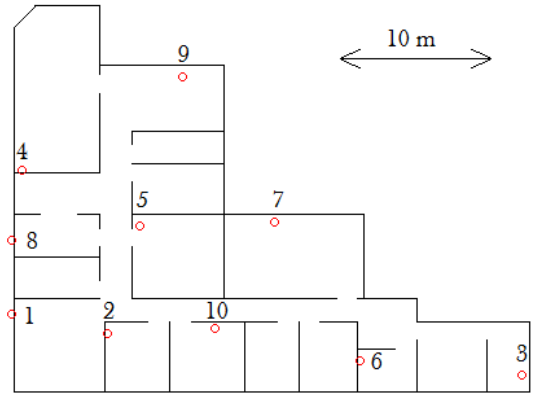


Figure 7.6: Map of the indoor environment used for the test with the position the access points (red circles) and their associated identification numbers.

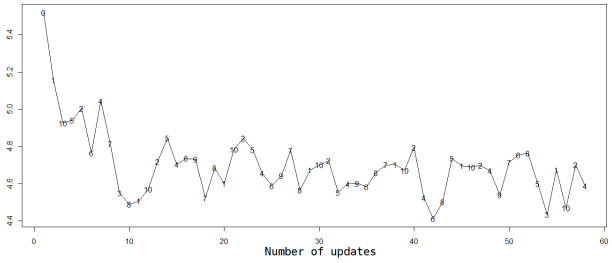


Figure 7.7: 0.8-quantile of the distance (in meter) between the true localization and the averaged estimate obtained with the stabilized algorithm. The localization error is computed on the test data sample each time we update one of the estimated map.

APs are represented in Y. Therefore, the maps $\widetilde{F}_j$, $j \in \{1, \dots, B\}$ are not estimated simultaneously as, for any time step t, two APs might appear a different number of times in $Y_{1:t}$. We thus slightly modify Algorithm 5 by introducing specific blocks and measure counters relatively to each AP. Each time the value of $\widetilde{F}_j$ for any $j \in \{1, \dots, B\}$ is updated using a block of the 20000 measures, we submit the new estimator $\widetilde{F}$ to the test data sample: Algorithm 5 is run on the test data sample $\{Y_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$. Only the averaged particle system is computed and no parameter update is performed. We can then compute the localization error relatively to the test data sample as the 0.8-quantile of the error between $\{X_t^{\text{test}}\}_{t=1}^{T_{\text{test}}}$ and the averaged position estimate. Figure 7.7 displays the results of this experiment by representing the 0.8-quantile of the error as a function of the number of updates. The numbers on the graph in Figure 7.7 indicate which AP were updated for each update. The initial map estimates are given, for any $j \in \{1, \dots, 10\}$ by $c_{1,j}^0 = -26$ and $c_{2,j}^0 = -17.5$ and $\delta_0 = 0$. Despite the relatively small test sample size, Figure 7.7 shows that the localization error seems to adopt the same behaviour as the localization error for the simulated data. The parameter F_j was updated a maximum of 7 times (for AP $j = 10$ for instance) and a minimum of 2 times (for AP

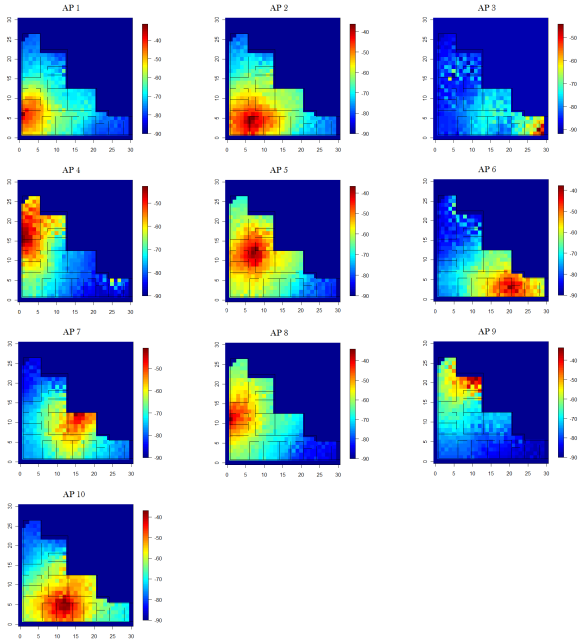


Figure 7.8: Graphical representation of the final propagation maps estimations: $\{\widetilde{F}_j\}_{j=1}^{10}$ (in dBm).

$j = 3$). Figure 7.8 represents the final estimate of the propagation maps $\widehat{F}_j$, $j \in \{1, \ldots, 10\}$. For some map estimates (see for instance the access points 1, 4 and 7), the signal strength drops when passing walls while the walls are responsible for a part of the indoor waves propagation disturbances.

7.6 Conclusion

In this paper we propose a stabilized version of the BOEM algorithm to estimate the signal propagation maps needed in any WiFi based localization system. The main difference with the existing solutions is that these propagation maps are estimated using the data sent by the mobile device originally used for localization purposes. On the contrary, the existing WiFi based localization systems establish these propagation maps either in a deterministic way or by running a previous hand made survey. In case of environmental modifications, the propagation maps are thereby changed. Our technique can easily be adapted to these changes by regularly reinitializing the sufficient statistics while hand made survey based systems can not take into account these modifications without renewing the survey. However, further tests are needed to evaluate the accuracy provided by our method and to compare it with other methods. Many elements should be analyzed such as the number and the position of the access points, the size of the environment or the materials constituting the obstacles in the environment.

Bibliography

P. Bahl and V.N. Padmanabhan. RADAR: An In-Building RF-Based User Location and Tracking System. In *INFOCOM*, pages 775–784, 2000.

G. Barrenetxea, F. Ingelrest, G. Schaefer, and M. Vetterli. Wireless Sensor Networks for Environmental Monitoring: The SensorScope Experience. In *The 20th IEEE International Zurich Seminar on Communications (IZS 2008)*, 2008. Invited paper.

O. Cappé. Recursive Computation of Smoothed Functionals of Hidden Markovian Processes using a Particle Approximation. *Monte Carlo Methods Appl.*, 7(1–2):81–92, 2001.

O. Cappé. Online sequential Monte Carlo EM algorithm. In *IEEE Workshop on Statistical Signal Processing (SSP)*, 2009.

O. Cappé. Online EM algorithm for Hidden Markov Models. *To appear in J. Comput. Graph. Statist.*, 2011.

O. Cappé and E. Moulines. Online Expectation Maximization Algorithm for Latent Data Models. *J. Roy. Statist. Soc. B*, 71(3):593–613, 2009. doi: 10.1111/j.1467-9868.2009.00698.x.

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

Y.C. Chen, J.R. Chiang, H.H Chu, P Huang, and A.W. Tsui. Sensor-assisted wi-fi indoor location system for adapting to environmental dynamics. In *Proceedings of the 8th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, MSWiM '05, pages 118–125, New York, NY, USA, 2005. ACM. ISBN 1-59593-188-0. doi: 10.1145/1089444.1089466.

M. Del Moral, A. Doucet, and S.S Singh. Forward smoothing using sequential Monte Carlo. Preprint, Dec 2010.

P. Del Moral. *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications.* Springer, 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

A. Doucet and A.M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. *Oxford handbook of nonlinear filtering*, 2009.

F. Evennou and F. Marx. Advanced integration of WIFI and inertial navigation systems for indoor mobile positioning. *EURASIP J. Appl. Signal Process.*, 2006:164–164, January 2006. ISSN 1110-8657. doi: 10.1155/ASP/2006/86706.

Brian Ferris, Dirk Hähnel, and Dieter Fox. Gaussian Processes for Signal Strength-Based Location Estimation. In *Robotics: Science and Systems'06*, pages –1–1, 2006.

G. Fort and E. Moulines. Convergence of the Monte Carlo Expectation Maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259, 2003.

H. T. Friis. A Note on a Simple Transmission Formula. *Proceedings of the IRE*, 34(5):254–256, September 1946.

E. Gaura, L. Girod, J. Brusey, M. Allen, and G. Challen. *Wireless Sensor Networks: Deployments and Design Frameworks*. Springer, 2010. ISBN 9781441958334.

J M Gorce, K Jaffres-Runser, and G De La Roche. Deterministic Approach for Fast Simulations of Indoor Radio Wave Propagation, 2007.

S.Y. Lau, T.H. Lin, T.Y. Huang, I.H. Ng, and P. Huang. A measurement study of zigbee-based indoor localization systems under RF interference. In *Proceedings of the 4th ACM international workshop on Experimental evaluation and characterization*, WINTECH '09, pages 35–42, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-740-0. doi: 10.1145/1614293.1614300.

S. Le Corff and G. Fort. Online Expectation Maximization based algorithms for inference in Hidden Markov Models. Technical report, arXiv, 2011.

S. Le Corff, G. Fort, and E. Moulines. Online EM algorithm to solve the SLAM problem. In *IEEE Workshop on Statistical Signal Processing (SPP)*, 2011.

G. Mongillo and S. Denève. Online Learning with Hidden Markov Models. *Neural Computation*, 20 (7):1706–1716, 2008. doi: 10.1162/neco.2008.10-06-351.

Algorithm 5 BOEM_SLAM_indoor

Require: θ^0, $\{\tau_k\}_{k\geq 1}$, $\{Y_t\}_{t\geq 0}$, N, N_b .

Set $\widehat{\theta} = \widetilde{\theta} = \theta^0$.
Sample $\{\widehat{\xi}_0^p\}_{p=1}^N$ and $\{\widetilde{\xi}_0^p\}_{p=1}^N$ independently and uniformly in $\mathcal{C}$.
Set $\widehat{\omega}_0^p = \widetilde{\omega}_0^p = \frac{1}{N}$ for all $p \in \{1, \ldots, N\}$.
Set $\rho_0^p = 0$ for all $p \in \{1, \ldots, N\}$, $k = 1$, $T_0 = 0$, $T_1 = \tau_1$.
for all $t \geq 1$ **do**
  Selection and propagation step.
  Set $\{\widehat{\xi}_t^p, \widehat{\omega}_t^p\}_{p=1}^N = BFR\left(\{\widehat{\xi}_{t-1}^\ell, \widehat{\omega}_{t-1}^\ell\}_{\ell=1}^N, Y_t, \widehat{\theta}\right)$
  Set $\{\widetilde{\xi}_t^p, \widetilde{\omega}_t^p\}_{p=1}^N = BFR\left(\{\widetilde{\xi}_{t-1}^\ell, \widetilde{\omega}_{t-1}^\ell\}_{\ell=1}^N, Y_t, \widetilde{\theta}\right)$
  Position estimations.
  Set $\widehat{p} = \underset{p\in\{1,\ldots,N\}}{\operatorname{argmax}}\, \widehat{\omega}_t^p$ and $\widehat{X}_t = \widehat{\xi}_t^{\widehat{p}}$.
  Set $\widetilde{p} = \underset{p\in\{1,\ldots,N\}}{\operatorname{argmax}}\, \widetilde{\omega}_t^p$ and $\widetilde{X}_t = \widetilde{\xi}_t^{\widetilde{p}}$.
  Forward computation of the intermediate quantity.
  for $p = 1$ to N **do**
    Compute $\{\rho_t^p\}_{p=1}^N$ following (7.11).
  end for
  Map estimation.
  if $t = T_k$ **then**
    Set

$$\widehat{\mathcal{S}}_k = \sum_{p=1}^{N} \widehat{\omega}_t^p \rho_t^p \ .$$

    $\widehat{\theta} = \bar{\theta}(\widehat{\mathcal{S}}_k, \tau_k)$.
    Set $\rho_t^p = 0$ for all $p \in \{1, \ldots, N\}$.
    if $k = 1$ **then**
      Set $\widetilde{\mathcal{S}}_k = \widehat{\mathcal{S}}_k$.
    else
      Set

$$\widetilde{\mathcal{S}}_k = \left(T_{k-1}\widetilde{\mathcal{S}}_{k-1} + \tau_k \widehat{\mathcal{S}}_k\right)/T_k \ .$$

    end if
    $\widetilde{\theta} = \bar{\theta}(\widetilde{\mathcal{S}}_k, T_k)$.
    Stabilization step.
    if $k = 0 \bmod N_b$ **then**
      Set $\widehat{\theta} = \widetilde{\theta}$
    end if
    $k = k + 1$ and $T_k = T_{k-1} + \tau_k$.
  end if
end for

Chapter 8

Nonparametric estimation in hidden Markov models

Thierry Dumont and Sylvain Le Corff

Sommaire

Abstract

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 $f_\star(X_k)$. We propose a maximum likelihood based procedure to estimate the function $f_\star$ using a block of observations $Y_{0:2n-1}$. This paper shows the identifiability of the model under several assumptions on the Markov chain and on the function $f_\star$. We also provide a proof of the consistency of the estimator of $f_\star$ as the number of observations grows to infinity. This consistency result relies on the Hellinger consistency of an estimator of the likelihood of the observations. Finally, we provide numerical experiments to highlight the performance of the estimator.

Keywords : Markov chains, hidden Markov models, nonparametric estimation, Maximum likelihood.

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 and Elliott [2007]), biology (Churchill [1992]) or speech recognition (Juang and Rabiner [1991]).

In this paper, the state-space of the Markov chain $\{X_k\}_{k\geq 0}$ is assumed to be a compact subset of $\mathbb{R}^m$ homeomorphic to a convex subset of $\mathbb{R}^m$ with a Lipschitz boundary. 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 $f_\star$ and the parameter $a_\star$ using only the observations $\{Y_k\}_{k\geq 0}$.

In 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 and Hall [1988], Carroll and Stefanski [1990] and Fan and Truong [1993] for an estimation based on kernel density estimators; see also Comte and 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 [2008a] 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 [2008b] 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 of 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 and 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 bijective, 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 (8.12)). 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 and 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 displayed 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 the appendices and in the supplement paper Dumont and Le Corff [2012a].

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 known 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 euclidean 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, positive, 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 a compact subset of $\mathbb{R}^m$.

(ii) K is homeomorphic to a convex subset of $\mathbb{R}^m$.

(iii) K has a local Lipschitz boundary.

K has a local Lipschitz boundary if, for any x in the boundary ∂K of K, there exists a neighbourhood V of x in ∂K which is the graph of a Lipschitz function. As an immediate consequence of the compactness of K and of the positivity of q, there exists $0 < \sigma_-(a) < \sigma_+(a) < +\infty$ such that, for all $x, x' \in K$,

$$\sigma_-(a) \le q_a(x,x') \le \sigma_+(a) . \tag{8.3}$$

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 and Tweedie, 2009, 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

$$\forall x \in K,\ \nu_a(x) \stackrel{\text{def}}{=} \frac{\int_K q\left(\frac{\|x'-x\|}{a}\right) \mathrm{d}x'}{\int_{K^2} q\left(\frac{\|x'-x''\|}{a}\right) \mathrm{d}x'\mathrm{d}x''} . \tag{8.4}$$

Let $p \ge 1$, define

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

For any m-tuple $\alpha \stackrel{\text{def}}{=} \{\alpha_i\}_{i=1}^m$ of non-negative integers, we write $|\alpha| \stackrel{\text{def}}{=} \sum_{i=1}^m \alpha_i$. 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. Let $s \in \mathbb{N}$, define $W^{s,p}$ be the Sobolev space on K with parameters s and p, *i.e.*,

$$W^{s,p} \stackrel{\text{def}}{=} \{ f \in \mathrm{L}^p;\ D^\alpha f \in \mathrm{L}^p, \alpha \in \mathbb{N}^m \text{ and } |\alpha| \le s \} , \tag{8.5}$$

where $D^\alpha f : K \to \mathbb{R}^\ell$ represents here the vector of partial derivatives of order α, in the sense of distributions, 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

$$\|f\|_{W^{s,p}} \stackrel{\text{def}}{=} \left(\sum_{0 \le |\alpha| \le s} \|D^\alpha f\|_{\mathrm{L}^p}^p \right)^{1/p} . \tag{8.6}$$

For any subset Ω_0 of $\mathbb{R}^m$, and any $k \ge 0$, let $\mathcal{C}^k(\Omega_0)$ be the vector space of all the functions $f : \Omega_0 \to \mathbb{R}$ such that there exists an open neighbourhood Ω of Ω_0 (if Ω_0 is open we can take $\Omega = \Omega_0$) in $\mathbb{R}^m$ and a function $\overline{f} : \Omega \to \mathbb{R}$ such that the restriction $\overline{f}|_{\Omega_0}$ of $\overline{f}$ on Ω_0 satisfies $\overline{f}|_{\Omega_0} = f$ and $\overline{f}$ is $\mathcal{C}^k$-regular on Ω, which means that $\overline{f}$ and all its partial derivatives $D^\alpha \overline{f}$ are continuous on Ω. Define, for any x in Ω_0, $D^\alpha f(x) = D^\alpha \overline{f}(x)$. Let $\|\cdot\|_{\mathcal{C}^k(\Omega_0)}$ be the norm on $\mathcal{C}^k(\Omega_0)$ defined by $\|f\|_{\mathcal{C}^k(\Omega_0)} \stackrel{\text{def}}{=} \sup_{|\alpha| \le k} \|D^\alpha f\|_\infty$. We also define $\mathcal{C}^k(\Omega_0, \mathbb{R}^\ell)$ by $\mathcal{C}^k(\Omega_0, \mathbb{R}^\ell) = \mathcal{C}^k(\Omega_0)^\ell$.

Remark 8.2.1.

i) By H2(iii) and the Stein Theorem [Adams and Fournier, 2003, Theorem 5.24], there exists a positive constant C such that any bounded function f in $\mathcal{C}^1(\overset{\circ}{K})$ can be extended by a function $\overline{f}$ in $\mathcal{C}^1(\mathbb{R}^m)$, with $\|\overline{f}\|_{\mathcal{C}^1(\mathbb{R}^m)} \le C \|f\|_{\mathcal{C}^1(\overset{\circ}{K})}$.

ii) 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 $k \ge 0$, by [Adams and 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 the subspace of bounded functions in $\left(\mathcal{C}^k(\overset{\circ}{K}), \|\cdot\|_{\mathcal{C}^k(\overset{\circ}{K})}\right)$. Provided that $s > m/p + 1$, and arguing component by component, $W^{s,p}$ is compactly embedded into the subspace of bounded functions $\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$. Moreover, the identity function $id : W^{s,p} \to \mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$ being linear and continuous, there exists a positive coefficient κ such that, for any $f \in W^{s,p}$,

$$\|f\|_{\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)} \le \kappa \|f\|_{W^{s,p}} , \tag{8.7}$$

thus f is a bounded function in $\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$ and, by i), can be extended by a function in $\mathcal{C}^1(K, \mathbb{R}^\ell)$ shortly denoted by $\mathcal{C}^1$, and

$$\|f\|_{\mathcal{C}^1} \le \kappa \|f\|_{W^{s,p}} . \tag{8.8}$$

H3 $s > m/p + 1$.

For any $f \in \mathcal{C}^1$ and any $x \in K$, the Jacobian of f at x, is defined by

$$J_f^2(x) \stackrel{\text{def}}{=} \mathrm{Det}\left[D_f(x)^T \, D_f(x) \right] ,$$

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

$$D_f(x)_{j,i} \stackrel{\text{def}}{=} \frac{\partial f_j}{\partial x_i}(x) .$$

For any sets E and F, $f : E \mapsto F$, denote by $\mathrm{Im}(f)$ the image in F of f, $\mathrm{Im}(f) \stackrel{\text{def}}{=} f(E)$.

H4 (i) $f_\star \in W^{s,p}$.

(ii) $f_\star : K \to \mathrm{Im}(f_\star)$ is a diffeomorphism.

Remark 8.2.2. i) We say that a function $f : K \to \mathrm{Im}(f)$ is a diffeomorphism if there exists an open neighbourhood V of K in $\mathbb{R}^m$ and a diffeomorphism $\overline{f} : V \to \mathrm{Im}(\overline{f})$ such that $\overline{f}|_V = f$.

ii) By H4(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 (see proof of Proposition 8.3.6), 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

$$\left(\widehat{f}_n, \widehat{a}_n\right) \stackrel{\mathrm{def}}{=} \underset{f \in W^{s,p}, a \geq a_-}{\mathrm{argmax}} \left\{ \frac{1}{n} \sum_{k=0}^{n-1} \ln p_{f,a}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 I^2(f) \right\} , \tag{8.9}$$

where, for all y_0, y_1 in $\mathbb{R}^\ell$,

$$p_{f,a}(y_0, y_1) \stackrel{\mathrm{def}}{=} \int \varphi(y_0 - f(x_0))\varphi(y_1 - f(x_1))\nu_a(x_0)q_a(x_0, x_1)\mathrm{d}x_0\mathrm{d}x_1 \tag{8.10}$$

and, for some positive υ,

$$I^2(f) \stackrel{\mathrm{def}}{=} ||f||^{\upsilon+1}_{W^{s,p}} . \tag{8.11}$$

Remark 8.2.3. 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 (8.9), (8.11) and Remark 8.2.1, $\widehat{f}_n$ 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 (8.9), $\widehat{a}_n$ 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:

$$\nu_\infty(x_0) \stackrel{\mathrm{def}}{=} \mu(K)^{-1} , \; q_\infty(x_0, x_1) \stackrel{\mathrm{def}}{=} \mu(K)^{-1} ,$$
$$p_{f,\infty}(y_0, y_1) \stackrel{\mathrm{def}}{=} \mu(K)^{-2} \int \varphi(y_0 - f(x_0))\mathrm{d}x_0 \int \varphi(y_1 - f(x_1))\mathrm{d}x_1 .$$

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

$$\widehat{p}_n \stackrel{\mathrm{def}}{=} p_{\widehat{f}_n, \widehat{a}_n} . \tag{8.12}$$

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

$$h(p_1,p_2) \stackrel{\text{def}}{=} \left[\frac{1}{2}\int \left(p_1^{1/2}(y)-p_2^{1/2}(y)\right)^2 \mathrm{d}y\right]^{1/2} . \tag{8.13}$$

Remark 8.2.4. 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 (8.11) can be easily manipulated compared with the $\mathcal{C}^1$ norm. Moreover, Theorem 8.3.5 ensures that $\|\widehat{f}_n\|_{W^{s,p}}$ stays bounded and thus, by Remark 8.2.1, 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 8.3.7.

Section 8.3 provides the main results of the paper. Theorem 8.3.1 establishes the identifiability of our model. Then, the Hellinger consistency of the MLE (8.12) is shown in Theorem 8.3.5. This result does not imply, *a priori*, the consistency of the estimators $\left(\widehat{f}_n,\widehat{a}_n\right)$ defined by (8.9). However, by Theorem 8.3.1, 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 8.3.7.

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 $f \stackrel{\mathcal{I}}{\sim} h$ 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 ϕ on K such that $f = h\circ\phi$. In the sequel, for any random variables X and Y, we write $X \stackrel{\mathcal{D}}{=} Y$ if X and Y have the same distribution.

Theorem 8.3.1. *Assume H1-2 and H4. 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* (8.10)*. Then, $b = a_\star$ and $f \stackrel{\mathcal{I}}{\sim} f_\star$.*

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 $\{X'_k\}_{k\geq 0}$. 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 $k \geq 0$, $(Y_k, Y_{k+1}) \stackrel{\mathcal{D}}{=} (Y'_k, Y'_{k+1})$. The density φ being known, this yields

$$(f(X'_k), f(X'_{k+1})) \stackrel{\mathcal{D}}{=} (f_\star(X_k), f_\star(X_{k+1})) . \tag{8.14}$$

(8.14) 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 H4, $f_\star$ is a diffeomorphism. Let $(f_\star)^{-1}$ denotes its inverse function and define

$$\phi \stackrel{\text{def}}{=} (f_\star)^{-1}\circ f . \tag{8.15}$$

Since $f_\star$ is a diffeomorphism and $f \in \mathcal{C}^1$, $\phi \in \mathcal{C}^1$. The purposes of the following lemmas is to prove that ϕ is bijective on K and that, for any x in K, $J_\phi(x) > 0$ which is showed in Lemma 8.3.2.

Lemma 8.3.2. *Assume H2*(i) *and H4. For all* $x \in K$, $J_\phi(x) > 0$, *where* ϕ *is defined by* (8.15).

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 8.3.3 and Lemma 8.3.4.

$\phi : K \to K$ 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 8.3.3. *Assume H2*(i) *and H4. Then, the function* ϕ *defined by* (8.15) *is a covering map.*

Lemma 8.3.4. *Assume H2*(ii). *Then, every covering map* $\phi : K \to K$ *is a one to one function.*

By Lemma 8.3.3 and Lemma 8.3.4, ϕ, defined by (8.15) is bijective, denote by ϕ^{-1} the inverse function of ϕ. By Lemma 8.3.2, $J_\phi > 0$ on K and thus $\phi^{-1} \in \mathcal{C}^1$.

By (8.14), for all $x \in K$ and all positive measurable function h on K,

$$\begin{aligned} Q_{a_\star}(x,h) = \mathbb{E}\left[h(X_k)|X_{k-1}=x\right] &= \mathbb{E}\left[h\circ\phi(X'_k)|X'_{k-1}=\phi^{-1}(x)\right] , \\ &= Q_b(\phi^{-1}(x), h\circ\phi) . \end{aligned}$$

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, by continuity, 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),

$$C_{a_\star}(x) q\left(\frac{\|x'-x\|}{a_\star}\right) = C_b(\phi^{-1}(x)) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x)\|}{b}\right)|J_{\phi^{-1}}(x')| . \tag{8.16}$$

Therefore, for all $x \in K$, applying (8.16) with $x' = x$ yields

$$|J_{\phi^{-1}}(x)| = \frac{C_{a_\star}(x)}{C_b(\phi^{-1}(x))} . \tag{8.17}$$

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 (8.16), 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 (8.17) in (8.16) yields

$$q\left(\frac{\|x-x_0\|}{a_\star}\right) q\left(\frac{\|x'-x\|}{a_\star}\right) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x_0)\|}{b}\right) = q\left(\frac{\|\phi^{-1}(x)-\phi^{-1}(x_0)\|}{b}\right) q\left(\frac{\|\phi^{-1}(x')-\phi^{-1}(x)\|}{b}\right) q\left(\frac{\|x'-x_0\|}{a_\star}\right) . \quad (8.18)$$

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, (8.18) 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 8.3.5 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 8.3.1 and Theorem 8.3.5 lead to the second main result, Theorem 8.3.7, which guarantees that $(\widehat{f}_n, \widehat{a}_n)$ is also consistent. The proof of Theorem 8.3.5 uses the same classical proof scheme as in the independent case, see [Van De Geer, 2009, 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 and Bednorz [2012]. Denote by $\mathbb{P}_\star$ 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 $Z_n = O_{\mathbb{P}_\star}(\alpha_n)$ if

$$\lim_{T\to+\infty} \limsup_{n\to+\infty} \mathbb{P}_\star\{|Z_n| > T\alpha_n\} = 0 .$$

Theorem 8.3.5. *Assume H1-3, H4*(i) *and H5. Let* $\left(\widehat{f}_n, \widehat{a}_n\right)$ *be defined by* (8.9) *and* $I(f)$ *by* (8.11). *Then, provided that*

$$\lambda_n \underset{n\to+\infty}{\longrightarrow} 0 \text{ and } \lambda_n^2 n^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty , \quad (8.19)$$

we have

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) = O_{\mathbb{P}_\star}(\lambda_n^2) \quad \text{and} \quad I^2(\widehat{f}_n) = O_{\mathbb{P}_\star}(1) . \quad (8.20)$$

Sketch of 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 [2009]. The control of this empirical process will be done in Proposition 8.3.6. We set, for any density function p on $\mathbb{R}^{2\ell}$,

$$g_p \stackrel{\text{def}}{=} \frac{1}{2} \ln \frac{p + p_{f_\star,a_\star}}{2 p_{f_\star,a_\star}} \; . \tag{8.21}$$

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 (8.9) and (8.12), the basic inequality of [Van De Geer, 2009, Lemma 10.5], states that:

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) + 4\lambda_n^2 I^2(\widehat{f}_n) \leq 16 \int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) + 4\lambda_n^2 I^2(f_\star) \; . \tag{8.22}$$

Therefore, a control of the term $\int g_{\widehat{p}_n} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)$ in the right hand side of (8.22) 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

$$\nu_n(g_{p_{f,a}}) \stackrel{\text{def}}{=} \sqrt{n} \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \; ,$$

where $g_{p_{f,a}}$ is defined by (8.21). Proposition 8.3.6 provides a deviation inequality for the supremum of the normalized empirical process.

Proposition 8.3.6. *Assume H1-3, H4*(i) *and H5. There exist some positive constants K, Σ and T such that, for any $x > 0$,*

$$\mathbb{P}_\star \left\{ \sup_{f \in W^{s,p},\ a \geq a_-} \frac{|\nu_n(g_{p_{f,a}})|}{I^2(f) \vee 1} \geq T + x \right\} \leq K e^{-\Sigma x} \; . \tag{8.23}$$

Proposition 8.3.6 is proved in Section 8.5.2 below. It ensures that

$$\sup_{f \in W^{s,p},\ a \geq a-} \frac{|\int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)|}{I^2(f) \vee 1} = O_{\mathbb{P}_\star}(n^{-1/2}) \; .$$

Plugging this bound into (8.22) gives

$$(4 + O_{\mathbb{P}_\star}(n^{-1/2}\lambda_n^{-2})) I^2(\widehat{f}_n) \leq 4 I^2(f_\star) + O_{\mathbb{P}_\star}(n^{-1/2}\lambda_n^{-2}) \; .$$

By (8.19), this establishes the second statement of (8.20). Combining this result with (8.22) gives:

$$h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq O_{\mathbb{P}_\star}(n^{-1/2}) + O_{\mathbb{P}_\star}(\lambda_n^2)$$

which proves the first statement of (8.20) and concludes the proof of Theorem 8.3.5. □

Equations (8.19) and (8.20) 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, $I^2(\widehat{f}_n) = O_{\mathbb{P}_\star}(1)$ and the Sobolev embedding described in Remark 8.2.1 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.*

$$\mathcal{F}_\star \stackrel{\text{def}}{=} \{f;\ f \stackrel{\mathcal{I}}{\sim} f_\star\}\ . \tag{8.24}$$

Theorem 8.3.7. *Assume H1-5. Let* $\left(\widehat{f}_n, \widehat{a}_n\right)$ *be defined by* (8.9) *and* $I(f)$ *by* (8.11). *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,

$$d_{\mathcal{C}^1}(\hat{f}_n, \mathcal{F}_\star) \underset{n\to+\infty}{\longrightarrow} 0 \quad \text{and} \quad \widehat{a}_n \underset{n\to+\infty}{\longrightarrow} a_\star \text{ in } \mathbb{P}_\star - \text{probability}\ , \tag{8.25}$$

where $\mathcal{F}_\star$ *is defined by* (8.24).

Proof. We prove (8.25) 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 $\{f \in W^{s,p};\ I(f) \le I\}$, 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 (8.25) will result from Theorem 8.3.5 and continuity arguments on the function $(f, a) \mapsto h^2(p_{f,a}, p_{f_\star,a_\star})$.

By Theorem 8.3.5, for any $\gamma > 0$, there exist $\epsilon > 0$ and $I > 0$ such that:

$$\limsup_{n\to+\infty} \mathbb{P}_\star \left\{h^2(\widehat{p}_n, p_{f_\star,a_\star}) > \epsilon\lambda_n^2\right\} \le \frac{\gamma}{2}\ , \tag{8.26}$$

$$\limsup_{n\to+\infty} \mathbb{P}_\star \left\{I(\widehat{f}_n) > I\right\} \le \frac{\gamma}{2}\ . \tag{8.27}$$

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 8.3.1 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,

$$\begin{aligned}\mathbb{P}_\star\{d((\widehat{f}_n, \widehat{a}_n), (\mathcal{F}_\star, a_\star)) > \eta\} \leq \mathbb{P}_\star\{I(\widehat{f}_n) > I\} + \mathbb{P}_\star\left\{h^2(\widehat{p}_n, p_{f_\star,a_\star}) > \epsilon\lambda_n^2\right\} \\ + \mathbb{P}_\star\left\{I(\widehat{f}_n) \leq I,\, h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq \epsilon\lambda_n^2, d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\} .\end{aligned}$$

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 $\epsilon\lambda_n^2 < \epsilon_\eta$,

$$\mathbb{P}_\star\left\{I(\widehat{f}_n) \leq I,\, h^2(\widehat{p}_n, p_{f_\star,a_\star}) \leq \epsilon\lambda_n^2,\, d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\} = 0 .$$

and, by (8.26) and (8.27),

$$\limsup_{n\to+\infty} \mathbb{P}_\star\left\{d\left(\left(\widehat{f}_n, \widehat{a}_n\right), (\mathcal{F}_\star, a_\star)\right) > \eta\right\} \leq \gamma .$$

Since γ can be chosen arbitrarily small, for any $\eta > 0$,

$$\limsup_{n\to+\infty} \mathbb{P}_\star\left\{d_{\mathcal{C}_1}(\widehat{f}_n, \mathcal{F}_\star) + |\arctan(\widehat{a}_n) - \arctan(a_\star)| > \eta\right\} = 0 ,$$

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 (8.9). For practical considerations, we choose $\upsilon = 1$ in (8.11) 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 (8.9) that maximizes the function T defined by

$$\begin{array}{rccl} T & : & W^{s,2} & \to & \mathbb{R} \\ & & f & \mapsto & \frac{1}{n}\sum_{k=0}^{n-1} \ln p_{f,a_\star}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 ||f||^2_{W^{s,2}} . \end{array}$$

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 $\widehat{f}^{p+1}$ 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

$$S_{n,1}(\widehat{f}^p, f, h) \stackrel{\text{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p} \left[\langle h(X_{2k}), f(X_{2k}) - Y_{2k} \rangle | Y_{2k:2k+1} \right] \ ,$$

$$S_{n,2}(\widehat{f}^p, f, h) \stackrel{\text{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p} \left[\langle h(X_{2k+1}), f(X_{2k+1}) - Y_{2k+1} \rangle | Y_{2k:2k+1} \right] \ .$$

$\widehat{f}^{p+1}$ 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 $h \in W^{2,2}([0,1], \mathbb{R})$,

$$S_{n,1}(\widehat{f}^p, f, h) + S_{n,2}(\widehat{f}^p, f, h) - 2\lambda_n^2 \sum_{\alpha=0}^{2} \left\langle f^{(\alpha)}, h^{(\alpha)} \right\rangle_{\mathrm{L}_2} = 0 \ . \tag{8.28}$$

This equation can be applied to any function h in $W_0^{2,2} \stackrel{\text{def}}{=} \{h \in W([0,1], \mathbb{R});\ h(0) = h(1) = 0\}$. Using integration by parts, this yields, for any component f_j and any $x \in [0, 1]$,

$$\left(1 + \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \right) f_j(x)$$

$$- f_j^{(2)}(x) + f_j^{(4)}(x) = \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ Y_{2k} \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + Y_{2k+1} \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \ , \tag{8.29}$$

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

$$\phi^{\widehat{f}^p,a_\star}_{2k|2k:2k+1}(x) \stackrel{\text{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x, x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x'}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \ ,$$

$$\phi^{\widehat{f}^p,a_\star}_{2k+1|2k:2k+1}(x') \stackrel{\text{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x, x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \ .$$

8.4.1 Numerical approximations

Let $N \geq 1$ be an integer. The differential system (8.29) 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 (8.29) 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 (8.28) 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 $\widehat{f}_j^{p+1}$. The procedure is displayed in Algorithm 6. The numerical approximations and the computations of all the constants are detailed in the supplement paper [Dumont and Le Corff, 2012a, Section 3]

Algorithm 6 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 (8.29).
    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 6 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

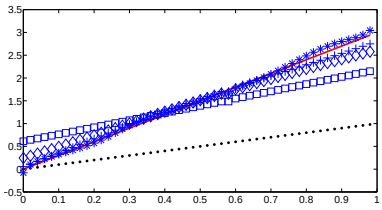
$$\begin{array}{rccc} \widehat{f}^0 & : & [0,1] & \to & \mathbb{R}^3 \\ & & x & \mapsto & (x, 0, 0) \ . \end{array}$$

The Algorithm 6 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 and the L_∞-error respectively defined, for $h_1, h_2 : [0,1] \to \mathbb{R}$, by

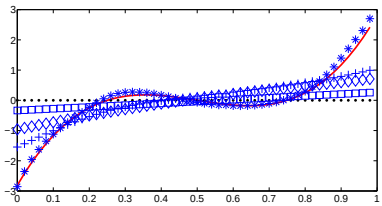
$$\|h_1 - h_2\|_2 \stackrel{\text{def}}{=} \left(\frac{1}{N} \sum_{i=1}^{N} \left| h_1\left(\frac{i}{N}\right) - h_2\left(\frac{i}{N}\right) \right|^2 \right)^{1/2} ,$$

$$\|h_1 - h_2\|_\infty \stackrel{\text{def}}{=} \sup_{1 \le i \le N} \left| h_1\left(\frac{i}{N}\right) - h_2\left(\frac{i}{N}\right) \right| .$$

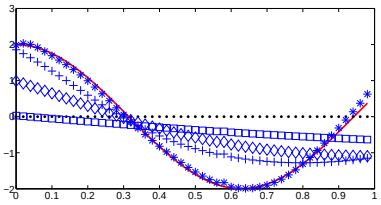
For each number of observations 50 independent Monte Carlo runs are used to compute the L_2-error after 25 iterations of the algorithm. Figure 8.2 shows the median and the lower and upper quartiles over the 50 independent Monte Carlo runs.



(a) f_1 and its estimates.

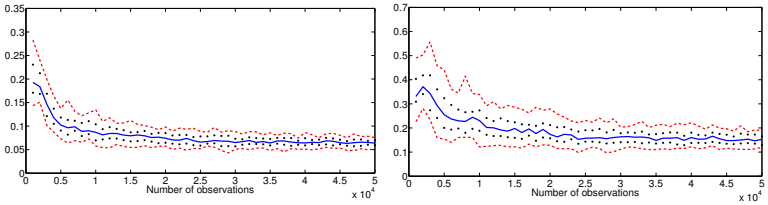


(b) f_2 and its estimates.

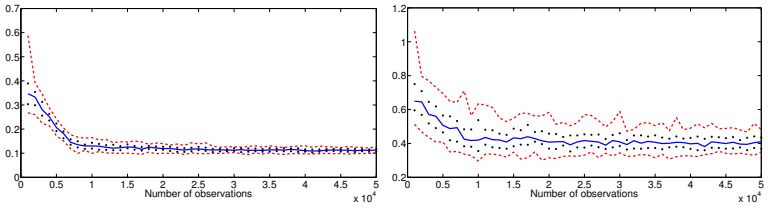


(c) f_3 and its estimates.

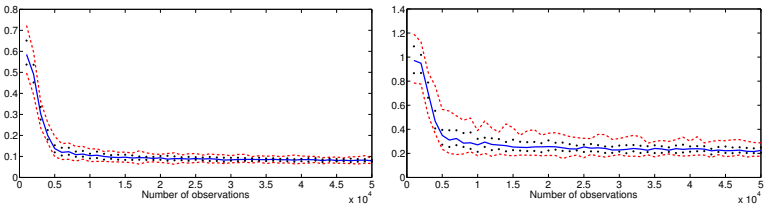
Figure 8.1: Estimation of f_1, f_2 and f_3 after 25 iterations of the algorithm. The true function (bold line) and the initial estimate (dots) are displayed along with the estimates after 1 (squares), 2 (diamonds), 3 (crosses) and 25 (stars) iterations.



(a) L_2-error for f_1 (b) L_∞-error for f_1



(c) L_2-error for f_2 (d) L_∞-error for f_2



(e) L_2-error for f_3 (f) L_∞-error for f_3

Figure 8.2: L_2 (left) and L_∞ (right) errors 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

Lemma (Lemma 8.3.2). *Assume H2*(i) *and H4. For all* $x \in K$, $J_\phi(x) > 0$, *where* ϕ *is defined by* (8.15).

Proof. By (8.15), (8.14) 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

$$\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] = \int h(\phi(x_0), \phi(x_1)) \nu_b(x_0) q_b(x_0, x_1) \mathrm{d}x_0 \mathrm{d}x_1 \; . \tag{8.30}$$

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

$$h(x_0, x_1) \stackrel{\text{def}}{=} h_2(x_0, x_1)\mathbf{1}_A(x_0)\mathbf{1}_A(x_1) \;, \tag{8.31}$$

where h_2 is any bounded measurable function. We have

$$\begin{aligned}\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] &= \int h_2(\phi(x_0), \phi(x_1))\nu_b(x_0)q_b(x_0, x_1)\mathbf{1}_A(\phi(x_0))\mathbf{1}_A(\phi(x_1))\mathrm{d}x_0\mathrm{d}x_1 \\ &= \int h_2(\phi(x_0), \phi(x_1))\frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)}\mathbf{1}_A(\phi(x_0))\mathbf{1}_A(\phi(x_1))J_\phi(x_0)J_\phi(x_1)\mathrm{d}x_0\mathrm{d}x_1 \;.\end{aligned}$$

By [Evans and 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 $z_0 = \phi(x_0)$, $z_1 = \phi(x_1)$.

$$\mathbb{E}\left[h(\phi(X_0'), \phi(X_1'))\right] = \int h_2(z_0, z_1)\mathbf{1}_A(z_0)\mathbf{1}_A(z_1) \times \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)}\mathrm{d}z_0\mathrm{d}z_1 \;.$$

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 $(z_0, z_1) \in K \times K$,

$$\nu_{a_\star}(z_0)q_{a_\star}(z_0, z_1)\mathbf{1}_A(z_0)\mathbf{1}_A(z_1) = \mathbf{1}_A(z_0)\mathbf{1}_A(z_1) \times \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)} \;.$$

By Sard Theorem (see Bröcker and 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 $(z_0, z_1) \in K \times K$,

$$\nu_{a_\star}(z_0)q_{a_\star}(z_0, z_1) = \sum_{\substack{x_0 \in \phi^{-1}(\{z_0\}) \\ x_1 \in \phi^{-1}(\{z_1\})}} \frac{\nu_b(x_0)q_b(x_0, x_1)}{J_\phi(x_0)J_\phi(x_1)} \;. \tag{8.32}$$

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 (8.32) 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. □

Lemma (Lemma 8.3.3). *Assume H2*(i) *and H4. Then, the function ϕ defined by* (8.15) *is a covering map.*

Proof. (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 8.3.2, 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 $V \stackrel{\text{def}}{=} \cap_{i=1}^n \phi(V_i)$, $O_i = \phi_{|V_i}^{-1}(V)$. Then, V is an open neighborhood of z which concludes the proof. □

Lemma (Lemma 8.3.4). *Assume H2*(ii). *Then, every covering map $\phi : K \to K$ is a one to one function.*

Proof. 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 8.3.6

Proposition 8.3.6 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 $\mathbb{R}^{2\ell}$:

$$\mathcal{G}_M \stackrel{\text{def}}{=} \left\{ g_{p_{f,a}};\ f \in W_M^{s,p},\ a > 0 \right\} \text{ and } \overline{\mathcal{G}}_M \stackrel{\text{def}}{=} \left\{ g - \mathbb{E}_\star \left[g(Y_0, Y_1) \right];\ g \in \mathcal{G}_M \right\} ,$$

where $\mathbb{E}_\star$ is the expectation under the distribution $\mathbb{P}_\star$.

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 (8.11) sufficient to establish Proposition 8.3.6. 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 8.5.1 gives a concentration inequality on the restricted empirical processes.

Proposition 8.5.1. *Assume H1, H2*(i), *H4*(i) *and H2*(iii), *H3. 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 $t \geq Cn^{-1/2}$,*

$$\mathbb{P}_\star \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{8.33}$$

The proof of Proposition 8.5.1 is given in the supplement paper [Dumont and Le Corff, 2012a, Section 2] and relies on the concentration results of Adamczak and Bednorz [2012]. It remains to control $\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right]$ for any $M \geq 1$.

Proposition 8.5.2. *Assume H1, H2*(i), *H2*(iii), *H4*(i) *and H3-5. There exists a positive constant K depending on υ, such that, for any $M \geq 1$,*

$$\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] \leq K M^{\upsilon+1} . \tag{8.34}$$

The proof of Proposition 8.5.2 is given in Appendix 8.6.2. We now combine Proposition 8.5.1 and Proposition 8.5.2 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 $t \geq \frac{C}{\sqrt{n}}$,

$$\mathbb{P}_\star \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq K_3 M^{\upsilon+1} + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{8.35}$$

Proposition 8.3.6 is obtained applying the peeling device as in [Van De Geer, 2009, 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 $x_k \leq 2^{kv}$, we have $e^{-K_2 t^2} \leq e^{-K_2 t}$ and $t \leq 2^{kv}(x+1)$. Plugging these relations into (8.35) leads to

$$\mathbb{P}_\star \left\{ \sup_{f \in W_{2^k}^{s,p},\ a \geq a_-} \left| \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \right| \geq \frac{2^{k(\upsilon+1)}}{\sqrt{n}} (K_3' + x) \right\} \leq K_1' e^{-K_2(x+x_k)} , \tag{8.36}$$

where $K_1' \stackrel{\text{def}}{=} 2K_1$ and $K_3' \stackrel{\text{def}}{=} K_3 + 1$. If $T \stackrel{\text{def}}{=} 2^{\upsilon+1} K_3'$,

$$\begin{aligned}
\mathbb{P}_\star \left\{ \sup_{f \in W^{s,p},\ a \geq a_-} \frac{|\int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)|}{I^2(f) \vee 1} \geq \frac{T+x}{\sqrt{n}} \right\} \\
\leq \mathbb{P}_\star \left\{ \sup_{f \in W_1^{s,p},\ a \geq a_-} \left| \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \right| \geq \frac{T+x}{\sqrt{n}} \right\} \\
+ \sum_{k=0}^{\infty} \mathbb{P}_\star \left\{ \sup_{f \in W_{2^{k+1}}^{s,p},\ a \geq a_-} \left| \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \right| \geq 2^{k(\upsilon+1)} \frac{T+x}{\sqrt{n}} \right\} .
\end{aligned}$$

However, since $T \geq K_3'$ and $x_1 \geq 0$, by (8.36) applied with $k = 0$,

$$\mathbb{P}_\star \left\{ \sup_{f \in W_1^{s,p},\ a \geq a_-} \left| \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \right| \geq \frac{T+x}{\sqrt{n}} \right\} \leq K_1' e^{-K_2 x} .$$

Therefore, by the definition of T and by (8.36),

$$\begin{aligned}
\mathbb{P}_\star \left\{ \sup_{f \in W^{s,p},\ a \geq a_-} \frac{|\int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)|}{I^2(f) \vee 1} \geq \frac{T+x}{\sqrt{n}} \right\} \\
\leq K_1' e^{-K_2 x} + \sum_{k=0}^{\infty} \mathbb{P}_\star \left\{ \sup_{f \in W_{2^{k+1}}^{s,p},\ a \geq a_-} \left| \int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star) \right| \right. \\
\left. \geq (2^{k+1})^{\upsilon+1} \frac{K_3' + x/2^{\upsilon+1}}{\sqrt{n}} \right\} \\
\leq K_1' e^{-K_2 x} + \sum_{k=0}^{\infty} K_1' e^{-K_2\left(x/2^{\upsilon+1} + x_{k+1}\right)} .
\end{aligned}$$

This last equation ensures the existence of some positive constants K and Σ such that

$$\mathbb{P}_\star \left\{ \sup_{f \in W^{s,p},\ a \geq a_-} \frac{|\int g_{p_{f,a}} \mathrm{d}(\mathbb{P}_n - \mathbb{P}_\star)|}{I^2(f) \vee 1} \geq \frac{T+x}{\sqrt{n}} \right\} \leq K e^{-\Sigma x} .$$

8.6 Appendices

8.6.1 Appendix A

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 $a > 0$,

$$\boldsymbol{f}(\mathbf{x}) \stackrel{\text{def}}{=} (f(x_0), f(x_1)) \in \mathbb{R}^{2\ell} \quad \text{and} \quad \boldsymbol{\nu}_a(\mathbf{x}) \stackrel{\text{def}}{=} \nu_a(x_0) q_a(x_0, x_1) \; . \tag{8.37}$$

This appendix is devoted to the proof of 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 8.6.1. *Assume H2*(i), *H2*(iii), *H4*(i) *and H3. 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 (8.37), 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 H3 and H4(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 (8.8). □

Lemma 8.6.1 implies that there exists a constant $C > 0$ such that, for any $\mathbf{y} \in \mathbb{R}^{2\ell}$,

$$\overline{\mathrm{G}}_M(\mathbf{y}) \leq C \left(1 + M\|\mathbf{y}\|\right) \; . \tag{8.38}$$

8.6.2 Appendix B

We prove Proposition 8.5.2 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 8.6.2. *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 $n \geq 1$,

$$\beta_n \stackrel{\text{def}}{=} \sup_{u>0} \sup_{A \in \mathcal{G}_{u+n}^{\mathbf{Y}}} \left| \mathbb{P}_\star \left(A \middle| \mathcal{H}_u^{\mathbf{Y}} \right) - \Pr_\star (A) \right| , \tag{8.39}$$

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 (8.39), 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. [2004], it can be proved that there exist β in $(0,1)$ and $C > 0$ such that, for any $n \geq 1$,

$$\beta_n \leq C \beta^n . \tag{8.40}$$

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 $\mathcal{L}_{2,\beta}(\mathbb{P}_\star)$ the class of numerical functions g such that $\|g\|_{2,\beta} < \infty$.

Proposition 8.6.3. *Assume H2(i), H2(iii) and H3. 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 8.6.3 is given in Appendix 8.6.3. Proposition 8.6.3 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 $r > 1$,

$$\|g\|_{2,\beta} \leq \|g\|_{\mathrm{L}_{2r}(\mathbb{P}_\star)} \sqrt{\int_0^1 u^{-1/r}\beta^{-1}(u)\mathrm{d}u} \, . \tag{8.41}$$

Moreover, by [Massart and 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 $a > 0$,

$$\|g_{p_{f,a}}\|^{2r}_{\mathrm{L}_{2r}(\mathbb{P}_\star)} \leq Ch(p_{f,a}, p_{f^\star,a_\star}) \, . \tag{8.42}$$

The Hellinger distance being bounded, (8.42) and (8.41) 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

$$\mathbb{E}_\star\left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)|\right] \leq C\varphi_M\left(1 + \frac{\delta_M(1 \wedge \epsilon_{n,M})}{d}\right), \tag{8.43}$$

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 (8.43). By Proposition 8.6.3 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

$$\varphi_M \leq C\left(M^{s'+b+\frac{2}{p'}\ell}\right)^{\ell/s'} \int_0^d u^{-2r\ell/s'}\mathrm{d}u \, , \tag{8.44}$$

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 (8.44), then $\left(M^{s'+b+\frac{2}{p'}\ell}\right)^{\ell/s'} \underset{s' \to \infty}{\longrightarrow} M^{2\ell}$ and it follows that,

$$\forall \eta > 0, \ \exists C > 0, \ \forall M \geq 1, \ \varphi_M \leq CM^{2\ell+\eta} \, .$$

By H5, there exists a constant C such that

$$\varphi_M \leq CM^{\upsilon} \, . \tag{8.45}$$

Lemma 8.6.4. *Assume H4*(i) *and H3. 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 8.6.1,

$$\begin{aligned}
\mathbb{P}_\star\{G_M(\mathbf{Y}_0) \geq u\} &\leq \mathbb{P}_\star\{C_G(1 + M\|\mathbf{Y}_0\|) \geq u\} \\
&\leq \mathbb{P}_\star\left\{\|\mathbf{Y}_0\| \geq \frac{u/C_G - 1}{M}\right\} \\
&\leq \mathbb{P}_\star\left\{\|\boldsymbol{f}^\star(\mathbf{X}_0)\| + \|\boldsymbol{\epsilon}_0\| \geq \frac{u/C_G - 1}{M}\right\} \\
&\leq \mathbb{P}_\star\left\{\|\boldsymbol{\epsilon}_0\| \geq \frac{u/C_G - 1}{M} - c_\infty\right\} ,
\end{aligned}$$

where $\|\boldsymbol{f}^\star(\mathbf{x})\| \leq c_\infty$ for all $\mathbf{x}$ in K^2 ($f_\star$ is bounded by H3 and H4(i)). Using Cirelson-Ibragimov-Sudakov inequality, see [Massart and Picard, 2007, Section 1.2.1], for any $x > 0$

$$\mathbb{P}_\star\left\{\frac{1}{\sigma}\left(\|\boldsymbol{\epsilon}_0\| - \mathbb{E}(\|\boldsymbol{\epsilon}_0\|)\right) \geq x\right\} \leq \mathrm{e}^{-\frac{x^2}{2}} .$$

Hence,

$$\begin{aligned}
\mathbb{P}_\star\{G_M(\mathbf{Y}_0) \geq u\} &\leq \exp\left(-\frac{\left(\frac{u/C_G - 1}{M} - c_\infty - \mathbb{E}(\|\boldsymbol{\epsilon}_0\|)\right)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{\left(\frac{c_1 u - 1}{M} - c_2\right)^2}{2\sigma^2}\right) ,
\end{aligned}$$

where $c_1 \stackrel{\text{def}}{=} \frac{1}{C_G}$ and $c_2 \stackrel{\text{def}}{=} c_\infty + \mathbb{E}_\star[\|\boldsymbol{\epsilon}_0\|]$. Setting $1 \geq t > 0$, let u be such that $t = \mathbb{P}_\star\{G_M(\mathbf{Y}_0) \geq u\}$, 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 (8.40), there exists a constant $C > 0$ such that,

$$\forall x \in (0,1),\ B(x) \leq Cx\left(1 + \ln\left(\frac{1}{x}\right)\right) . \tag{8.46}$$

Lemma 8.6.5. *Assume H4*(i) *and H3. 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 8.6.4,

$$\mathcal{Q}_{G_M}(t) \leq CM \left(1 + \ln^{1/2}\left(\frac{1}{t}\right) \right) \ .$$

Therefore, by (8.46)

$$\mathcal{Q}_{G_M}(t)\sqrt{B(t)} \leq CM \left(1 + \ln^{1/2}\left(\frac{1}{t}\right) \right) \sqrt{t\left\{1 + \ln\left(\frac{1}{t}\right)\right\}} \ .$$

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 8.6.5 ensures that $\delta_M(\epsilon) \underset{\epsilon \to 0}{\longrightarrow} 0$ for any $M \geq 1$, and Proposition 8.5.2 results from (8.43), Lemma 8.6.5, and (8.45).

8.6.3 Appendix C

The aim of this appendix is to prove Proposition 8.6.3. 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 8.6.6. *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 and Picard, 2007, Lemma 7.26], there exists a positive constant C such that

$$\|g_{P_U} - g_{P_L}\|^{2r}_{\mathrm{L}_{2r}(\mathbb{P}_\star)} \leq C \|\sqrt{P_U} - \sqrt{P_L}\|^2_{\mathrm{L}_2(\mathbb{R}^{2\ell})} \ .$$

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 (8.41). □

Nickel and 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 8.6.7. *Assume H2*(i), *H2*(iii) *and H3. 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 $\alpha \in \mathbb{N}^{2\ell}$,

$$D^\alpha \left(p_{f,a} \cdot \langle \mathbf{y} \rangle^b\right) = \sum_{\alpha' \leq \alpha} \binom{\alpha}{\alpha'} D^{\alpha'}(\langle \mathbf{y} \rangle^b) D^{\alpha - \alpha'}(p_{f,a}) , \tag{8.47}$$

where $\binom{\alpha}{\alpha'} \stackrel{\text{def}}{=} \prod_{j=1}^{2\ell} \binom{\alpha_j}{\alpha'_j}$. Thus, Lemma 8.6.7 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 $\mathbf{y} \in \mathbb{R}^{2\ell}$,

$$D^\alpha p_{f,a}(\mathbf{y}) = \int_{\mathbf{x} \in K^2} P_\alpha(\boldsymbol{f}(\mathbf{x}) - \mathbf{y}) \exp\left\{ -\frac{\|\boldsymbol{f}(\mathbf{x}) - \mathbf{y}\|^2}{2\sigma^2} \right\} \boldsymbol{\nu}_a(\mathbf{x}) \mathrm{d}\mathbf{x} . \tag{8.48}$$

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 (8.8) 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,

$$\left| P_{b,\alpha^{(1)}}(\mathbf{y}) P_{\alpha^{(2)}}(\boldsymbol{f}(\mathbf{x}) - \mathbf{y}) \right| \leq C(1 + \|\mathbf{y}\|)^{b - |\alpha^{(1)}|} \mathbf{1}_{|\alpha^{(1)}| \leq b} \times \left(\sqrt{2}\kappa \left(1 \vee \|f\|_{W^{s,p}}\right) + \|\mathbf{y}\| \right)^{|\alpha^{(2)}|} .$$

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 $\mathbb{R}^{2\ell}$,

$$\left|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^b)D^{\alpha^{(2)}}(p_{f,a})(\mathbf{y})\right| \leq C(1\vee\|f\|_{W^{s,p}})^{\alpha^{(2)}}\cdot(1+\|\mathbf{y}\|)^{b-|\alpha^{(1)}|}\left(1+\frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})}\right)^{|\alpha^{(2)}|} \times\left[\mathbf{1}_{A_f}(\mathbf{y})+\mathbf{1}_{A_f^c}(\mathbf{y})e^{-\frac{1}{2\sigma^2}\left(\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})-\|y\|\right)^2}\right].$$

Therefore, for any $p' \geq 1$,

$$\|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^b)D^{\alpha^{(2)}}(p_{f,a})\|^{p'}_{\mathrm{L}_{p'}} \leq C(1\vee\|f\|_{W^{s,p}})^{p'\alpha^{(2)}}\,(I_1+I_2)\ ,$$

where,

$$I_1 \stackrel{\text{def}}{=} \int_{A_f}(1+\|\mathbf{y}\|)^{p'(b-|\alpha^{(1)}|)}\left(1+\frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})}\right)^{p'|\alpha^{(2)}|}\mathrm{d}\mathbf{y}\ ,$$

$$I_2 \stackrel{\text{def}}{=} \int_{A_f^c}(1+\|\mathbf{y}\|)^{p'(b-|\alpha^{(1)}|)}\left(1+\frac{\|\mathbf{y}\|}{\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})}\right)^{p'|\alpha^{(2)}|} \times \mathrm{e}^{-\frac{p'}{2\sigma^2}\left(\sqrt{2}\kappa(1\vee\|f\|_{W^{s,p}})-\|y\|\right)^2}\mathrm{d}\mathbf{y}\ .$$

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

$$\|D^{\alpha^{(1)}}(\langle\mathbf{y}\rangle^b)D^{\alpha^{(2)}}(p_{f,a})\|^{p'}_{\mathrm{L}_{p'}} \leq C(1\vee\|f\|_{W^{s,p}})^{p'(|\alpha^{(2)}|-|\alpha^{(1)}|+b)+2\ell}\ . \tag{8.49}$$

Using (8.49) in (8.47) with $\alpha^{(1)} = \alpha'$ and $\alpha^{(2)} = \alpha - \alpha'$ for any $|\alpha| \leq s'$ and $\alpha' \leq \alpha$ concludes the proof. □

Hence Lemma 8.6.7 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 and 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

$$\forall\epsilon>0,\ H_{[]}\left(\epsilon,\mathcal{P}_M/M^{s'+b+\frac{2}{p'}\ell},\|\cdot\|_{\mathrm{L}_1(\mathbb{R}^{2\ell})}\right) \leq C\epsilon^{-2\ell/s'}\ . \tag{8.50}$$

Lemma 8.6.6 and (8.50) conclude the proof of Proposition 8.6.3.

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.

Bibliography

R. Adamczak and W. Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. arXiv:1201.3569v1, Jan 2012.

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Number vol. 140 in Pure and Applied Mathematics. Academic Press, 2003. ISBN 9780120441433.

T. Bröcker and L. Lander. *Differentiable Germs and Catastrophes*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1975. ISBN 9780521206815.

R.J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, pages 1184–1186, 1988.

R.J. Carroll and L.A. Stefanski. Deconvolving kernel density estimators. *Statistics*, 21:169–184, 1990.

G. Churchill. Hidden Markov Chains and the Analysis of Genome Structure. *Computers & Chemistry*, 16(2):107–115, 1992.

F. Comte and M.-L. Taupin. Nonparametric estimation of the regression function in an errors-in-variables model. *Statistica sinica*, 17(3):1065–1090, 2007.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

R. Douc, E. Moulines, and J. Rosenthal. Quantitative Bounds for Geometric Convergence Rates of Markov Chains. *Ann. Appl. Probab.*, 14(4):1643–1665, 2004.

P. Doukhan, P. Massart, and E. Rio. Invariance principle for absolutely regular processes. *Annales de l'Institut Henri Poincaré*, 31:393–427, 1995.

T. Dumont and S. Le Corff. Supplement paper to nonparametric estimation in hidden Markov models. Technical report, 2012a.

T. Dumont and S. Le Corff. Simultaneous localization and mapping problem in wireless sensor networks. Technical report, 2012b.

L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, 1992. ISBN 9780849371578.

J. Fan and Y.K. Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21: 1900–1925, 1993.

B. Juang and L. Rabiner. Hidden Markov Models for Speech Recognition. *Technometrics*, 33:251–272, 1991.

C. Lacour. Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Processes and their Applications*, 118(2):232 – 260, 2008a. ISSN 0304-4149.

C. Lacour. Adaptive estimation of the transition density of a particular hidden Markov chain. *Journal of Multivariate Analysis*, 99(5):787–814, 2008b.

J.M. Lee. *Introduction to Topological Manifolds.* Graduate Texts in Mathematics. Springer, 2000. ISBN 9780387950266.

R.S. Mamon and R.J. Elliott. *Hidden Markov Models in Finance*, volume 104 of *International Series in Operations Research & Management Science.* Springer, Berlin, 2007.

P. Massart and J. Picard. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003.* Number vol. 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer-Verlag, 2007. ISBN 9783540484974.

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Communications and control engineering. Cambridge University Press, 2009. ISBN 9780521731829.

R. Nickel and B.M. Potscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov and Sobolev type. *J. Theor. Probab.*, 20:177–199, 2001.

E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants.* Springer, 1990.

D.R. Smart. *Fixed point theorems.* Cambridge University Press, 1980.

S.A. Van De Geer. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2009. ISBN 9780521123259.

Chapter 9

Supplement paper to "Nonparametric estimation in hidden Markov models"

Thierry Dumont and Sylvain Le Corff

Sommaire

Abstract

This document is a supplementary material to the article "Nonparametric estimation in hidden Markov models". It provides additional proofs of some technical results given in the original paper. Section 9.1 recalls the model, the definitions and the assumptions used in the paper. Section 9.2 provides proofs of some results stated in the paper and Section 9.3 gives details on the algorithm used to perform the Expectation-Maximization based estimation.

Keywords : Markov chains, hidden Markov models, nonparametric estimation, Maximum likelihood.

9.1 Model and definitions

In this section, we recall the model and the assumptions given in Dumont and Le Corff [2012]. Comments on these assumptions can be found in [Dumont and Le Corff, 2012, Section 2]. 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 $\mathbb{R}^m$. $\{X_k\}_{k\in\mathbb{N}}$ 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

$$q_a(x, x') \stackrel{\text{def}}{=} C_a(x) q\left(\frac{\|x' - x\|}{a}\right) \ , \tag{9.1}$$

where q is a known, positive, continuous and strictly monotone function on $\mathbb{R}_+$ and where

$$C_a(x) \stackrel{\text{def}}{=} \left(\int_K q\left(\frac{\|x' - x\|}{a}\right) \mathrm{d}x'\right)^{-1} \ , \tag{9.2}$$

where $\mathrm{d}x'$ is a shorthand notation for $\mu(\mathrm{d}x')$. 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

H6 $\{X_k\}_{k\in\mathbb{Z}}$ is a stationary Markov chain with transition kernel $Q_{a_\star}$.

Assume the following statement on the set K:

H7 (i) K is compact.

(ii) K is homeomorphic to a convex subset of $\mathbb{R}^m$.

(iii) K has a local Lipschitz boundary.

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 $x, x' \in K$,

$$\sigma_-(a) \leq q_a(x, x') \leq \sigma_+(a) \ . \tag{9.3}$$

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 and Tweedie, 2009, 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

$$\forall x \in K,\ \nu_a(x) \stackrel{\text{def}}{=} \frac{\int_K q\left(\frac{\|x'-x\|}{a}\right) \mathrm{d}x'}{\int_{K^2} q\left(\frac{\|x'-x''\|}{a}\right) \mathrm{d}x' \mathrm{d}x''} \ . \tag{9.4}$$

Let $s \in \mathbb{N}$ and $p \geq 1$.

Remark 9.1.1. i) By H7(iii) and the Stein Theorem [Adams and Fournier, 2003, Theorem 5.24], there exists a positive constant C such that any bounded function f in $\mathcal{C}^1(\overset{\circ}{K})$ can be extended by a function $\overline{f}$ in $\mathcal{C}^1(\mathbb{R}^m)$, with $\|\overline{f}\|_{\mathcal{C}^1(\mathbb{R}^m)} \leq C\|f\|_{\mathcal{C}^1(\overset{\circ}{K})}$.

ii) 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 and Fournier, 2003, Theorem 6.3], assuming that K satisfies H7(i) and H7(iii) and $s > m/p + k$, $W^{s,p}(K, \mathbb{R})$ is compactly embedded into the subspace of bounded functions in $\left(\mathcal{C}^k(\overset{\circ}{K}), \|\cdot\|_{\mathcal{C}^k(\overset{\circ}{K})}\right)$. Provided that $s > m/p + 1$, and arguing component by component, $W^{s,p}$ is compactly embedded into the subspace of bounded functions $\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$. Moreover, the identity function $id : W^{s,p} \to \mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$ being linear and continuous, there exists a positive coefficient κ such that, for any $f \in W^{s,p}$,

$$\|f\|_{\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)} \leq \kappa \|f\|_{W^{s,p}} , \tag{9.5}$$

thus f is a bounded function in $\mathcal{C}^1(\overset{\circ}{K}, \mathbb{R}^\ell)$ and, by i), can be extended by a function in $\mathcal{C}^1(K, \mathbb{R}^\ell)$ shortly denoted by $\mathcal{C}^1$, and

$$\|f\|_{\mathcal{C}^1} \leq \kappa \|f\|_{W^{s,p}} . \tag{9.6}$$

H8 $s > m/p + 1$.

For any $f \in \mathcal{C}^1$ and any $x \in K$, the Jacobian of f at x, is defined by

$$J_f^2(x) \overset{\text{def}}{=} \text{Det}\left[D_f(x)^T \; D_f(x)\right] ,$$

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

$$D_f(x)_{j,i} \overset{\text{def}}{=} \frac{\partial f_j}{\partial x_i}(x) .$$

H9 (i) $f_\star \in W^{s,p}$.

(ii) $f_\star : K \to \text{Im}(f_\star)$ is a diffeomorphism.

Consider the following assumption on υ.

H10 $\upsilon > 2\ell$.

9.2 Additional proofs

This section is devoted to the proof of [Dumont and Le Corff, 2012, Propositon 5.1]. First of all, for any $M \geq 1$ the Sobolev ball of radius M centred in 0 is denoted by $W_M^{s,p}$. Recall that, for any probability density function p on $\mathbb{R}^{2\ell}$,

$$g_p \overset{\text{def}}{=} \frac{1}{2} \ln \frac{p + p_{f_\star, a_\star}}{2 p_{f_\star, a_\star}} .$$

Define the following collections of functions on $\mathbb{R}^{2\ell}$:

$$\mathcal{G}_M \stackrel{\text{def}}{=} \left\{ g_{p_{f,a}};\ f \in W_M^{s,p},\ a > 0 \right\} \text{ and } \overline{\mathcal{G}}_M \stackrel{\text{def}}{=} \left\{ g - \mathbb{E}_\star \left[g(Y_0, Y_1) \right];\ g \in \mathcal{G}_M \right\} ,$$

where $\mathbb{E}_\star$ is the expectation under the distribution $\mathbb{P}_\star$. Under H6, H7(i), H7(iii), H9(i) and H8, [Dumont and Le Corff, 2012, Propositon 5.1] states that 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 $t \geq Cn^{-1/2}$,

$$\mathbb{P}_\star \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{9.7}$$

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 $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ in $K^2 \times (\mathbb{R}^\ell)^2$, 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 $\mathbf{z}$.

The proof of (9.7) is obtained by integration of the following result with respect to the invariant distribution of the Markov chain $\boldsymbol{Z}$.

Proposition 9.2.1. *Assume that H7*(i)*, H9*(i) *and H8 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 $t \geq Cn^{-1/2}$,*

$$\mathbb{P}_{\mathbf{z}} \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} \leq K_1 \left(e^{-K_2 t^2} + e^{-K_2 t} \right) . \tag{9.8}$$

Proposition 9.2.1 is an application of [Adamczak and Bednorz, 2012, Theorem 7] 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 9.2.2 is proved in [Dumont and Le Corff, 2012, Lemma *A*.1].

Lemma 9.2.2. *Assume H7*(i)*, H7*(iii) *H8 and H9*(i)*.There exists a constant $C > 0$ such that, for any $\mathbf{y} \in \mathbb{R}^{2\ell}$,*

$$\mathrm{G}_M(\mathbf{y}) \leq C_G \left(1 + M\|\mathbf{y}\| \right) \quad \textit{and} \quad \overline{\mathrm{G}}_M(\mathbf{y}) \leq C \left(1 + M\|\mathbf{y}\| \right) .$$

Proposition 9.2.1 is then an application of [Adamczak and Bednorz, 2012, Theorem 7] to the class $\{\bar{g}/M;\ \bar{g} \in \overline{\mathcal{G}}_M\}$. Indeed, Lemma 9.2.2 gives an upper bound for $\overline{\mathrm{G}}_M/M$ which is independent from M. This allows us to apply [Adamczak and Bednorz, 2012, Theorem 7] where all the constants in the upper bound of

$$\mathbb{P}_{\mathbf{z}} \left\{ \sup_{g \in \mathcal{G}_M} |\nu_n(g)| \geq c\mathbb{E}_\star \left[\sup_{g \in \mathcal{G}_M} |\nu_n(g)| \right] + Mt \right\} ,$$

do not depend on M.

By [Adamczak and Bednorz, 2012, Section 3.2], it is sufficient to prove that there exists a small set D, see [Meyn and Tweedie, 2009, 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 and Tweedie, 2009, Theorem 14.2.3 and Theorem 14.2.4], ii) is satisfied if

$$\sup_{\mathbf{z}\in D} \mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D - 1} \exp\left\{\frac{\overline{\mathrm{G}}_M}{M}(\mathbf{Y}_k)\right\}\right] < +\infty \ , \tag{9.9}$$

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 9.2.2 there exists $K > 0$ (independent from M) such that the function V is upper bounded by K on D. Therefore, [Adamczak and 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 9.2.1.

We now turn to the proof of i) and (9.9). By (9.3), 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 (9.9) can be established by a proper choice of D. By H9(i), there exists $M_\star < +\infty$ such that $\|f_\star\|_{W^{s,p}} = M_\star$. Furthermore, by Remark 9.1.1, 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 9.2.3. *For all $k \geq 0$ and all $\mathbf{z} \in D$,*

$$\mathbb{P}_{\mathbf{z}}\left\{\tau_C > k\right\} \leq \exp\left\{-\lambda(\rho)k\right\} \ ,$$

where

$$\lambda(\rho) \stackrel{\text{def}}{=} \left(\frac{t^2 - \ell}{2}\right) - \ln 2 - \frac{\ell}{2}\ln\left(\frac{t^2}{\ell}\right) \quad \textit{and} \quad t^2 \stackrel{\text{def}}{=} \frac{\rho^2}{\sigma^2} \ . \tag{9.10}$$

Proof. For all $i \geq 1$,

$$\mathbb{P}_{X_{2i}}\left\{Y_{2i} \notin \mathrm{B}(0, \sqrt{\ell}\kappa M_\star + \rho)\right\} = \mathbb{P}_{X_{2i}}\left\{\|f_\star(X_{2i}) + \epsilon_{2i}\| \geq \sqrt{\ell}\kappa M_\star + \rho\right\}$$
$$\leq \mathbb{P}_{X_{2i}}\left\{\|\epsilon_{2i}\| \geq \rho\right\} .$$

Since ϵ_{2i} is $\mathcal{N}_\ell(0, \sigma^2 I_\ell)$, if $t^2 > \ell$,

$$\mathbb{P}_{X_{2i}}\left\{\|\epsilon_{2i}\| \geq \rho\right\} \leq \exp\left\{\frac{\ell}{2}\ln\left(\frac{t^2}{\ell}\right) - \left(\frac{t^2-\ell}{2}\right)\right\} .$$

This concludes the proof. □

Then, for $\kappa > 1$ and $\mathbf{z} \in D$,

$$\mathbb{E}_{\mathbf{z}}\left[\kappa^{\tau_D}\right] = \sum_{k\geq 1} \mathbb{P}_{\mathbf{z}}\left\{\tau_D = k\right\}\kappa^k$$
$$\leq \mathrm{e}^{\lambda(\rho)} \sum_{k\geq 1} \mathrm{e}^{-(\lambda(\rho)-\ln\kappa)k} .$$

The right hand side of the last equation is finite if ρ is chosen sufficiently large. This concludes the proof of i).

Proof of (9.9). Let $\mathbf{z} \in D$. By Lemma 9.2.2, there exists a constant C such that

$$\mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D-1} \exp\left\{\frac{\overline{\mathrm{G}}_M}{M}(\mathbf{Y}_k)\right\}\right] \leq \mathrm{e}^C \mathbb{E}_{\mathbf{z}}\left[\sum_{k=0}^{\tau_D-1} \exp\left\{C\|\mathbf{Y}_k\|\right\}\right]$$
$$\leq \mathrm{e}^C \mathbb{E}_{\mathbf{z}}\left[\tau_D \exp\left\{C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right]$$
$$\leq \mathrm{e}^C \mathbb{E}_{\mathbf{z}}\left[\tau_D^2\right]^{1/2} \mathbb{E}_{\mathbf{z}}\left[\exp\left\{2C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right]^{1/2} .$$

By Lemma 9.2.3, $\sup_{\mathbf{z}\in D}\mathbb{E}_{\mathbf{z}}\left[\tau_D^2\right] < +\infty$. For the second term we write

$$\mathbb{E}_{\mathbf{z}}\left[\exp\left\{2C\sum_{k=0}^{\tau_D-1}\|\mathbf{Y}_k\|\right\}\right] = \sum_{p\geq 1}\mathbb{E}_{\mathbf{z}}\left[\mathbf{1}_{\tau_D=p}\exp\left\{2C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right]$$
$$\leq \sum_{p\geq 1}\mathbb{P}_{\mathbf{z}}\left\{\tau_D = p\right\}^{1/2}\mathbb{E}_{\mathbf{z}}\left[\exp\left\{4C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right]^{1/2} , \tag{9.11}$$

where

$$\mathbb{E}_z\left[\exp\left\{4C\sum_{k=0}^{p-1}\|\mathbf{Y}_k\|\right\}\right] \leq \exp\left\{4C\|\mathbf{y}\|\right\}$$
$$\times\left(\left(2\pi\sigma^2\right)^{-\ell}\int\exp\left\{-\frac{\|\mathbf{y}\|^2}{2\sigma^2} + \|\mathbf{y}\|\left(\frac{\sqrt{2\ell}\kappa M_\star}{\sigma^2} + 4C\right)\right\}\mathrm{d}\mathbf{y}\right)^{p-1} .$$

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,

$$
\begin{aligned}
\mathbb{E}_z\left[\exp\left\{4C\sum_{i=0}^{p-1}\|\mathbf{Y}_i\|\right\}\right] &\leq \exp\left\{4C\|\mathbf{y}\|\right\}\left[B(2\ell-1)\mathcal{H}(S^{2\ell-1})\right]^{p-1} \\
&\quad \times \exp\left\{\frac{(\sqrt{2\ell}\kappa M_\star + 4C\sigma^2)^2}{2\sigma^2} p\right\} (1+\sqrt{2\ell}\kappa M_\star + 4C\sigma^2)^{(2\ell-1)p} \, .
\end{aligned}
$$

Finally, for all $p \geq 1$ and all $z \stackrel{\text{def}}{=} (\mathbf{x}, \mathbf{y}) \in C$,

$$\mathbb{E}_z \left[\exp \left\{ 4C \sum_{i=0}^{p-1} \|\mathbf{Y}_i\| \right\} \right] \leq \exp\{4C\|\mathbf{y}\|\} \exp\{\eta(4C)(p-1)\} \ ,$$

where

$$\eta(4C) \stackrel{\text{def}}{=} \ln(\kappa(\ell)) + \frac{\left(\sqrt{2\ell}\kappa M_\star + 4C\sigma^2\right)^2}{2\sigma^2} + (2\ell - 1) \ln\left(1 + \sqrt{2\ell}\kappa M_\star + 4C\sigma^2\right) \tag{9.12}$$

and where $\kappa(\ell)$ is a constant depending only on ℓ. Therefore, by (9.11) and Lemma 9.2.3, this concludes the proof for a sufficiently large ρ. □

9.3 Numerical experiments

Let n be a positive integer, in this section, we denote by $\widehat{f}$ the estimator defined as a maximizer of the function T defined by

$$\begin{array}{rccl} T & : & W^{s,2} & \to & \mathbb{R} \\ & & f & \mapsto & \frac{1}{n}\sum_{k=0}^{n-1} \log p_{f,a_\star}(Y_{2k}, Y_{2k+1}) - \lambda_n^2 ||f||^2_{W^{s,2}} \ . \end{array}$$

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 $\widehat{f}^{p+1}$ is defined as one of the maximizer of the function Q defined by

$$f \mapsto Q(f, \widehat{f}^p) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p} \left[\log p_{f,a_\star}\left(X_{2k}, Y_{2k}, X_{2k+1}, Y_{2k}\right) \middle| Y_{2k},\ Y_{2k+1} \right] - \lambda_n^2 ||f||^2_{W^{s,2}} \ ,$$

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

$$\mathrm{d}_f Q(\cdot, \widehat{f}^p)(h) = S_{n,1}(\widehat{f}^p, f, h) + S_{n,2}(\widehat{f}^p, f, h) - 2\lambda_n^2 \sum_{0 \leq |\alpha| \leq s} \langle D^\alpha f, D^\alpha h \rangle_{\mathrm{L}_2} \ ,$$

where

$$S_{n,1}(\widehat{f}^p, f, h) \stackrel{\text{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p} \left[\langle h(X_{2k}), f(X_{2k}) - Y_{2k} \rangle | Y_{2k:2k+1} \right] \ ,$$

$$S_{n,2}(\widehat{f}^p, f, h) \stackrel{\text{def}}{=} \frac{1}{n\sigma^2} \sum_{k=0}^{n-1} \mathbb{E}_{\widehat{f}^p} \left[\langle h(X_{2k+1}), f(X_{2k+1}) - Y_{2k+1} \rangle | Y_{2k:2k+1} \right] \ .$$

$\widehat{f}^{p+1}$ 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 $h \in W([0,1], \mathbb{R})$,

$$S_{n,1}(\widehat{f}^p, f, h) + S_{n,2}(\widehat{f}^p, f, h) - 2\lambda_n^2 \sum_{\alpha=0}^{2} \left\langle f^{(\alpha)}, h^{(\alpha)} \right\rangle_{\mathrm{L}_2} = 0 \,. \tag{9.13}$$

This equation can be applied to any function h in $W_0^{2,2} \stackrel{\text{def}}{=} \{h \in W([0,1], \mathbb{R});\ h(0) = h(1) = 0\}$. Using integration by parts, this yields, for any component f_j and any $x \in [0,1]$,

$$\left(1 + \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \right) f_j(x)$$
$$- f_j^{(2)}(x) + f_j^{(4)}(x) = \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ Y_{2k} \phi^{\widehat{f}^p,a}_{2k|2k:2k+1}(x) + Y_{2k+1} \phi^{\widehat{f}^p,a}_{2k+1|2k:2k+1}(x) \right\} \,, \tag{9.14}$$

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

$$\phi^{\widehat{f}^p,a_\star}_{2k|2k:2k+1}(x) \stackrel{\text{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x,x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x'}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \,,$$
$$\phi^{\widehat{f}^p,a_\star}_{2k+1|2k:2k+1}(x') \stackrel{\text{def}}{=} \frac{\int \nu_{a_\star}(x) q_{a_\star}(x,x') \varphi(Y_{2k} - \widehat{f}^p(x)) \varphi(Y_{2k+1} - \widehat{f}^p(x')) \mathrm{d}x}{p_{\widehat{f}^p,a_\star}(Y_{2k}, Y_{2k+1})} \,.$$

Numerical approximations Let $N \geq 1$ be an integer. The differential system (9.14) is solved using a discretization of the state space $[0,1]$ by $\left\{\frac{i}{N}\right\}_{i=0}^{N}$. Let $\widehat{q}_{a^\star}$ be the transition probability associated to this discretization: for any $i, j \in \{0, \cdots, N\}$,

$$\widehat{q}_{a_\star}(i,j) \stackrel{\text{def}}{=} \frac{q_{a_\star}(\frac{i}{N}, \frac{j}{N})}{\sum_{j=0}^{N} q_{a_\star}(\frac{i}{N}, \frac{j}{N})}$$

and $\widehat{\nu}_{a_\star}$ the invariant distribution of $\widehat{q}_{a_\star}$ on $\{0, \ldots, N\}$. Define $\widehat{\nu}_{a_\star}$ as the invariant distribution of $\widehat{q}_{a_\star}$ on $\{0, \ldots, N\}$. The filtering distributions $\phi^{\widehat{f}^t,a_\star}_{2k|2k:2k+1}$ and $\phi^{\widehat{f}^p,a_\star}_{2k+1|2k:2k+1}$ 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

$$\underline{\phi}_k^{\widehat{f}^p,a_\star}(x) \stackrel{\text{def}}{=} \sum_{i=0}^{N-1} \mathbf{1}_{\left[\frac{i}{N}, \frac{i+1}{N}\right[}(x)\, \underline{\varphi}^{\widehat{f}^p}_{i,k} \quad \text{and} \quad \overline{\phi}_k^{\widehat{f}^p,a_\star}(x) \stackrel{\text{def}}{=} \sum_{i=0}^{N-1} \mathbf{1}_{\left[\frac{i}{N}, \frac{i+1}{N}\right[}(x)\, \overline{\varphi}^{\widehat{f}^t}_{i,k} \,,$$

where, for any $i, j \in \{0, \ldots, N-1\}$,

$$\underline{\varphi}^{\widehat{f}^p}_{i,k} \stackrel{\text{def}}{=} \frac{\sum_{j=0}^{N-1} \widehat{\nu}_{a_\star}(i) \widehat{q}_{a_\star}(i,j) \varphi\left(\widehat{f}^p\left(\frac{i}{N}\right) - Y_{2k}\right) \varphi\left(\widehat{f}^p\left(\frac{j}{N}\right) - Y_{2k+1}\right)}{\sum_{i',j'=0}^{N-1} \widehat{\nu}_{a_\star}(i') \widehat{q}_{a_\star}(i',j') \varphi\left(\widehat{f}^p\left(\frac{i'}{N}\right) - Y_{2k}\right) \varphi\left(\widehat{f}^p\left(\frac{j'}{N}\right) - Y_{2k+1}\right)} \,, \tag{9.15}$$

$$\overline{\varphi}^{\widehat{f}^p}_{j,k} \stackrel{\text{def}}{=} \frac{\sum_{i=0}^{N-1} \widehat{\nu}_{a_\star}(i) \widehat{q}_{a_\star}(i,j) \varphi\left(\widehat{f}^p\left(\frac{i}{N}\right) - Y_{2k}\right) \varphi\left(\widehat{f}^p\left(\frac{j}{N}\right) - Y_{2k+1}\right)}{\sum_{i',j'=0}^{N-1} \widehat{\nu}_{a_\star}(i') \widehat{q}_{a_\star}(i',j') \varphi\left(\widehat{f}^p\left(\frac{i'}{N}\right) - Y_{2k}\right) \varphi\left(\widehat{f}^p\left(\frac{j'}{N}\right) - Y_{2k+1}\right)} \,. \tag{9.16}$$

The equation (9.14) 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.

Computation of $\widehat{f}^{p+1}$

Let $i \in \{0, \cdots, N-1\}$ and $j \in \{1, \cdots, \ell\}$. We denote by f_i the solution of (9.14) on $\left]\frac{i}{N}, \frac{i+1}{N}\right[$. We have, for each component f_j, $j \in \{1, \cdots, \ell\}$,

$$f_{j,i}^{(4)}(x) - f_{j,i}^{(2)}(x) + (1+\alpha_i)\, f_{j,i}(x) = \beta_{j,i} \ ,$$

where

$$\alpha_i \stackrel{\text{def}}{=} \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ \underline{\varphi}_{i,k}^{\widehat{f}_n} + \overline{\varphi}_{i,k}^{\widehat{f}_n} \right\} \ ,$$
$$\beta_{j,i} \stackrel{\text{def}}{=} \frac{1}{2n\lambda_n^2\sigma^2} \sum_{k=0}^{n-1} \left\{ Y_{2k,j} \underline{\varphi}_{i,k}^{\widehat{f}_n} + Y_{2k+1,j} \overline{\varphi}_{j,k}^{\widehat{f}_n} \right\} \ .$$

Therefore, there exist $c_{i,1}$, $c_{i,2}$, $s_{i,1}$ and $s_{i,2}$ such that, for any $x \in \left]\frac{i}{N}, \frac{i+1}{N}\right[$,

$$f_{j,i}(x) = \mathrm{e}^{\eta_i x} \left[c_{1,i} \cos(\gamma_i x) + s_{1,i} \sin(\gamma_i x)\right] + \mathrm{e}^{-\eta_i x} \left[c_{2,i} \cos(\gamma_i x) + s_{2,i} \sin(\gamma_i x)\right] + \frac{\beta_{j,i}}{1+\alpha_i} \ ,$$

where, if $r_i \stackrel{\text{def}}{=} \sqrt{1+\alpha_i}$

$$\eta_i \stackrel{\text{def}}{=} \frac{\sqrt{1+2r_i}}{2} \quad \text{and} \quad \gamma_i \stackrel{\text{def}}{=} \frac{\sqrt{2r_i - 1}}{2} \ .$$

Therefore, $4N$ 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: for any $i \in \{0, \cdots, N-2\}$,

$$f_{j,i}\left(\frac{i+1}{N}\right) = f_{j,i+1}\left(\frac{i+1}{N}\right) \ , \qquad f'_{j,i}\left(\frac{i+1}{N}\right) = f'_{j,i+1}\left(\frac{i+1}{N}\right) \ ,$$
$$f_{j,i}^{(2)}\left(\frac{i+1}{N}\right) = f_{j,i+1}^{(2)}\left(\frac{i+1}{N}\right) \ , \quad f_{j,i}^{(3)}\left(\frac{i+1}{N}\right) = f_{j,i+1}^{(3)}\left(\frac{i+1}{N}\right) \ , \tag{9.17}$$

where, for any $x \in \left]\frac{i}{N}, \frac{i+1}{N}\right[$,

$$
\begin{aligned}
f'_{j,i}(x) &= \mathrm{e}^{\eta_i x}\left[(\eta_i c_{1,i} + \gamma_i s_{1,i})\cos(\gamma_i x) + (\eta_i s_{1,i} - \gamma_i c_{1,i})\sin(\gamma_i x)\right] \\
&\quad + \mathrm{e}^{-\eta_i x}\left[(\gamma_i s_{2,i} - \eta_i c_{2,i})\cos(\gamma_i x) - (\gamma_i c_{2,i} + \eta_i s_{2,i})\sin(\gamma_i x)\right] , \\
f^{(2)}_{j,i}(x) &= \mathrm{e}^{\eta_i x}\left\{\eta_i(\eta_i c_{1,i} + \gamma_i s_{1,i}) + \gamma_i(\eta_i s_{1,i} - \gamma_i c_{1,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{\eta_i x}\left\{\eta_i(\eta_i s_{1,i} - \gamma_i c_{1,i}) - \gamma_i(\eta_i c_{1,i} + \gamma_i s_{1,i})\right\}\sin(\gamma_i x) \\
&\quad + \mathrm{e}^{-\eta_i x}\left\{-\eta_i(\gamma_i s_{2,i} - \eta_i c_{2,i}) - \gamma_i(\gamma_i c_{2,i} + \eta_i s_{2,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{-\eta_i x}\left\{\eta_i(\gamma_i c_{2,i} + \eta_i s_{2,i}) - \gamma_i(\gamma_i s_{2,i} - \eta_i c_{2,i})\right\}\sin(\gamma_i x) , \\
f^{(3)}_{j,i}(x) &= \mathrm{e}^{\eta_i x}\eta_i\left\{\eta_i(\eta_i c_{1,i} + \gamma_i s_{1,i}) + \gamma_i(\eta_i s_{1,i} - \gamma_i c_{1,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{-\eta_i x}\eta_i\left\{\eta_i(\gamma_i s_{2,i} - \eta_i c_{2,i}) + \gamma_i(\gamma_i c_{2,i} + \eta_i s_{2,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{\eta_i x}\gamma_i\left\{\eta_i(\eta_i s_{1,i} - \gamma_i c_{1,i}) - \gamma_i(\eta_i c_{1,i} + \gamma_i s_{1,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{-\eta_i x}\gamma_i\left\{\eta_i(\gamma_i c_{2,i} + \eta_i s_{2,i}) - \gamma_i(\gamma_i s_{2,i} - \eta_i c_{2,i})\right\}\cos(\gamma_i x) \\
&\quad + \mathrm{e}^{\eta_i x}\eta_i\left\{\eta_i(\eta_i s_{1,i} - \gamma_i c_{1,i}) - \gamma_i(\eta_i c_{1,i} + \gamma_i s_{1,i})\right\}\sin(\gamma_i x) \\
&\quad - \mathrm{e}^{-\eta_i x}\eta_i\left\{\eta_i(\gamma_i c_{2,i} + \eta_i s_{2,i}) - \gamma_i(\gamma_i s_{2,i} - \eta_i c_{2,i})\right\}\sin(\gamma_i x) \\
&\quad - \mathrm{e}^{\eta_i x}\gamma_i\left\{\eta_i(\eta_i c_{1,i} + \gamma_i s_{1,i}) + \gamma_i(\eta_i s_{1,i} - \gamma_i c_{1,i})\right\}\sin(\gamma_i x) \\
&\quad + \mathrm{e}^{-\eta_i x}\gamma_i\left\{\eta_i(\gamma_i s_{2,i} - \eta_i c_{2,i}) + \gamma_i(\gamma_i c_{2,i} + \eta_i s_{2,i})\right\}\cos(\gamma_i x) .
\end{aligned}
$$

Solving (9.13) 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 $\widehat{f}_j^{p+1}$.

Bibliography

R. Adamczak and W. Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. arXiv:1201.3569v1, Jan 2012.

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces*. Number vol. 140 in Pure and Applied Mathematics. Academic Press, 2003. ISBN 9780120441433.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion), 1977.

T. Dumont and S. Le Corff. Nonparametric estimation in hidden Markov models. Technical report, 2012.

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and control engineering. Cambridge University Press, 2009. ISBN 9780521731829.

Quatrième partie

Modèles de Markov cachés à longueur variable

Chapitre 10

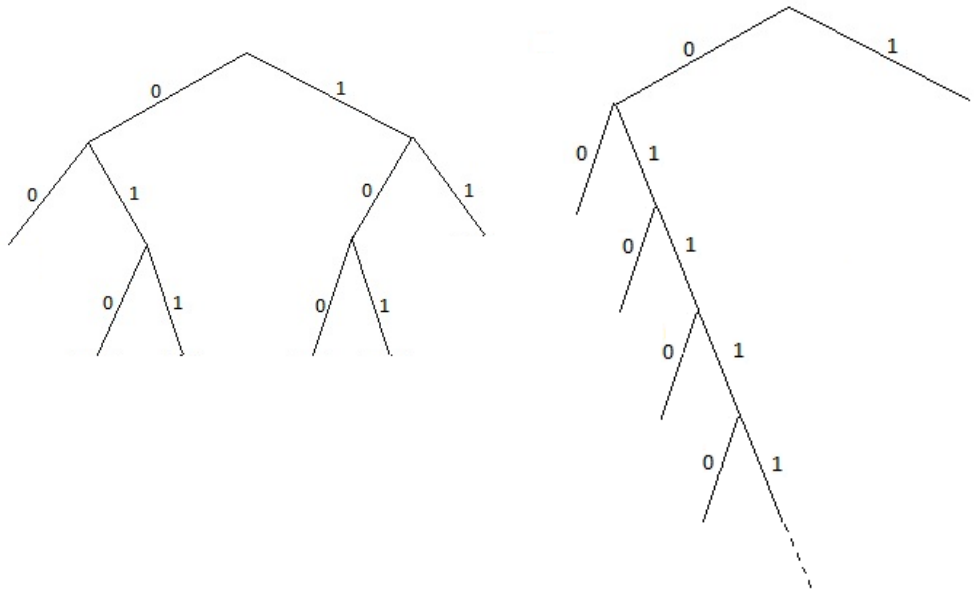
Généralités et introduction du chapitre 11

Sommaire

Dans ce chapitre, la modélisation du processus $\{X_t\}_{t\in\mathbb{N}}$ est modifiée par rapport au modèle général présenté dans la section 3.5. Dans le cas où un découpage grossier de l'environnement est effectué, nous cherchons à optimiser la précision obtenue en affinant notre compréhension de la manière dont le terminal mobile évolue dans l'environnement. Dans la section précédente, le processus $\{X_t\}_{t\in\mathbb{N}}$ était considéré comme une marche aléatoire. Cependant, si le terminal se trouve dans un couloir à l'instant t, par exemple, alors la prise en considération des positions du terminal aux instants $t-1$, $t-2$, $t-3$,... dans le modèle de transition permettrait d'évaluer la vitesse et/ou l'accélération du mobile dans le couloir. Par contre, lorsque le terminal se trouve dans une salle, la prise en compte des positions antérieures peut ne pas être nécessaire dans la description du processus, le processus $\{X_t\}_{t\in\mathbb{N}}$ peut alors être considéré comme étant à mémoire variable. Dans cette section nous étudions un modèle peu étudié jusqu'à présent : les chaînes de Markov cachées à longueur variable (ou VLHMM pour *variable length hidden Markov models*). Nous aborderons tout d'abord les chaînes de Markov à longueur variable (ou VLMC pour *variable length Markov chains*) introduites dans les années 1980 par Rissanen dans le cadre de la compression universelle de données. Nous discuterons notamment la notion d'arbre de contextes associé à une chaîne de Markov cachée à longueur variable ainsi que des résultats d'estimation de cet arbre. Nous introduirons ensuite les VLHMM, équivalent "caché" des chaînes de Markov à longueur variable. Après ces définitions, nous introduirons l'article *Context tree estimation in variable length hidden Markov models* à ce jour en révision pour la revue *IEEE Transactions on information theory*.

10.1 Chaînes de Markov à longueur variable

Soit $\mathbb{X}$ un ensemble fini. Pour $k \leq l$ entiers naturels fixés, un élément $x_k x_{k+1}, \dots x_l$ de $\mathbb{X}^{l-k+1}$ est appelé *mot* et est noté de manière plus concise $x_{k:l}$. Nous notons $l(x_{k:l}) = l - k + 1$ la longueur



(a) Représentation de τ_1 (b) Représentation de τ_2

FIGURE 10.1 – Exemple de représentation de deux arbres de contextes

du mot $x_{k:l}$. La concaténation d'un mot u et d'un mot v est notée uv et on dit qu'un mot s a pour suffixe v (ou v est suffixe de s) s'il existe un mot u tel que $s = uv$, u pouvant être vide.

Définition 7 (Arbres de contextes)**.** *Si τ est un ensemble de mots (fini ou infini), τ est appelé arbre de contextes si, pour tout élément s de τ, il n'existe pas d'élément s' dans τ tel que s' soit un suffixe de s. Cette propriété sera désignée comme étant la propriété d'arbre. De plus, l'arbre τ est dit irréductible si aucun élément s de τ ne peut être remplacé par un de ses suffixes s' sans violer la propriété d'arbre.*

Si τ est un arbre de contexte, et $s = x_{-l(s)+1:0}$ un élément de τ, alors s est appelé contexte ou feuille de l'arbre τ et, pour tout $0 \leq i \leq l(s) - 2$, $x_{-i:0}$ est appelé noeud de τ. τ possède alors une représentation d'arbre, les éléments de τ correspondant alors aux feuilles de l'arbre. La figure 10.1 illustre cette représentation sur deux exemples dans le cas $\mathbb{X} = \{0, 1\}$. La figure 10.1 (a) représente l'arbre de contextes fini τ_1 donné par $\tau_1 \stackrel{\text{def}}{=} \{00, 010, 110, 001, 101, 11\}$ et la figure 10.1 (b) représente l'arbre de contextes infini τ_2 donné par $\tau_2 \stackrel{\text{def}}{=} \{1, 00, 010, 0110, 01110, \ldots\}$.

Définition 8. *Un arbre de contextes irréductible est dit complet si, toute suite infinie $x_{-\infty:0}$ possède un unique suffixe dans τ.*

Soit s un noeud de l'arbre de contexte τ, soit $x \in \mathbb{X}$, le mot xs est appelé "enfant du noeud s" s'il existe un élément s' de τ tel que xs soit suffixe de s'. Un arbre irréductible est alors complet si et seulement si tout noeud de τ a exactement $|\mathbb{X}|$ enfants, où $|\mathbb{X}|$ désigne le cardinal de $\mathbb{X}$. Si τ est un arbre complet, on appelle fonction contexte associée à τ la fonction f_τ qui, à toute suite semi-infinie à gauche $x_{-\infty:0}$ de $\mathbb{X}^\infty$ associe l'unique élément de τ $f_\tau(x_{-\infty:0})$ suffixe de $x_{-\infty:0}$. On

note alors $d(\tau) \stackrel{\text{def}}{=} \max\{l(f_\tau(x_{-\infty:0})) \mid x_{-\infty:0} \in \mathbb{X}^\infty\}$ la *profondeur* de l'arbre τ. Si $d(\tau) < \infty$ alors τ est dit fini, sinon τ est dit infini.

Définition 9 (Chaînes de Markov à longueur variable)**.** *Un processus $\{X_t\}_{t\in\mathbb{Z}}$ sur $\mathbb{X}$ est à arbre de contextes τ (arbre complet) si il est stationnaire et si, pour toute suite semi-infinie à gauche $x_{-\infty:n}$ de $\mathbb{X}^\infty$, et pour tout $x \in \mathbb{X}$, si $s = f_\tau(x_{-\infty:n})$, alors*

$$\mathbb{P}\left(X_{n+1} = x | X_{-\infty:n} = x_{-\infty:n}\right) = \mathbb{P}\left(X_{n+1} = x | X_{-l(s)+n+1:n} = s\right)$$

Le processus $\{X_t\}_{t\in\mathbb{Z}}$ est alors appelé chaîne de Markov à longueur variable.

Remarque 5.
- *Si τ est une arbre fini, et si $\{X_t\}_{t\in\mathbb{Z}}$ est à contexte τ, alors $\{X_{-d(\tau)+t+1:t}\}_{t\in\mathbb{Z}}$ est une chaîne de Markov sur $\mathbb{X}^{d(\tau)}$.*
- *La distribution du processus $\{X_t\}_{t\in\mathbb{Z}}$ est entièrement déterminée par les probabilités de passage $\{P_{s,x}\}_{s\in\tau,\ x\in\mathbb{X}}$ où*

$$P_{s,x} \stackrel{\text{def}}{=} \mathbb{P}\left(X_1 = x | X_{-l(s)+1:0} = s\right) .$$

  Dans le cas où τ est fini, le nombre de paramètres de transitions déterminant la loi de $\{X_t\}_{t\in\mathbb{Z}}$ est donc égal à $|\tau| \times (|\mathbb{X}| - 1)$.
- *Un regroupement des facteurs par contexte d'apparition permet d'exprimer de manière concise la vraisemblance :*

$$\begin{aligned}\mathbb{P}(X_{1:n} = x_{1:n} | X_{-\infty:0} = x_{-\infty:0}) &= \prod_{t=1}^{n} P_{f_\tau(x_{-\infty:t-1}),x_t} \\ &= \prod_{s\in\tau} P_s(S^\star(s, x_{1:n}; x_{-\infty:0})) ,\end{aligned}$$

  où $S^\star(s, x_{1:n}; x_{-\infty:0})$ est la concaténation des éléments x_i, $1 \le i \le n$ tels que $x_{-l(s)+i:i-1} = s$, et $P_s(S^\star(s, x_{1:n}; x_{-\infty:0})) = \prod_{x\in S^\star(s,x_{1:n};x_{-\infty:0})} P_{s,x}$.

Les chaînes de Markov à longueur variable peuvent être utilisées de manière très efficace dans le domaine de la compression de données sans perte. En particulier, après les travaux précurseurs de Rissanen (Rissanen [1983]) et son algorithme *context*, le codage par double mélange dit *Context Tree Weighting* ou CTW (voir Willems et al. [1995], Willems [1994]) permet la construction d'une loi de codage mélange remarquablement efficace pour le codage des processus à mémoire finie. Son implémentation est d'ailleurs rendue tout aussi efficace grâce à l'algorithme CTW qui permet le codage de sources (processus) à mémoire finie de type VLMC sans connaissance *a priori* de l'arbre de contexte régissant le processus.

Dans la prochaine section, nous nous intéresserons à l'estimation de l'arbre de contexte associé à une VLMC non-observée, il convient d'ailleurs de souligner que si un processus $\{X_t\}_{t\in\mathbb{Z}}$ est à contexte τ, alors, pour tout arbre τ' couvrant τ (*i.e.* tout contexte s de τ est suffixe d'un contexte s' de τ') vérifie que $\{X_t\}_{t\in\mathbb{Z}}$ est à contexte τ'. L'objectif alors est l'estimation du plus "petit" arbre τ tel que $\{X_t\}_{t\in\mathbb{Z}}$ soit à contexte τ. Lorsque la VLMC est observée, Csiszar and Talata [2006] montre la consistance forte de l'estimateur MDL (pour *Minimum Description Length*) et de l'estimateur BIC. Cependant, Csiszar and Talata [2006] utilise une hypothèse de bornitude sur la profondeur de l'arbre à estimer, Garivier [2006] montre que cette hypothèse n'est pas nécessaire lorsque l'arbre à estimer est fini.

10.2 Modèles de Markov cachés à longueur variable

Ici, nous nous intéressons au cas où la VLMC $\{X_t\}_{t\in\mathbb{Z}}$ n'est pas observée directement mais à travers un processus dit d'observations $\{Y_t\}_{t\in\mathbb{Z}}$, ce que nous appellerons les modèles de Markov cachés à longueur variable (ou VLHMM pour *variable length hidden Markov models*) :

Définition 10 (Modèles de Markov cachés à longueur variable)**.** *Un processus $\{X_t, Y_t\}_{t\in\mathbb{Z}}$ est appelé modèle de Markov caché à longueur variable (ou VLHMM) si $\{X_t\}_{t\in\mathbb{Z}}$ est une chaîne de Markov cachée à longueur variable et si les variables aléatoires Y_t , $t \in \mathbb{Z}$, sont indépendantes conditionnellement à $\{X_t\}_{t\in\mathbb{Z}}$.*

Remarque 6. *Dans le cas où la chaîne de Markov cachée est à arbre de contextes τ fini (i.e. lorsque le processus $\{X_t\}_{t\in\mathbb{Z}}$ est markovien), les modèles de Markov cachés à longueur variable peuvent être considérés comme un cas particulier de modèles de Markov cachés en considérant le processus étendu $\{X_{-d(\tau)+t+1:t}, Y_t\}_{t\in\mathbb{Z}}$. Les distributions conditionnelles de $Y_t|X_{-d(\tau)+t+1:t}$ ne dépendent alors que de la valeur prise par la variable X_t.*

Les modèles VLHMM sont particulièrement intéressants pour leurs applications, comme dans les travaux de Wang [2005] et Wang and Liu [2005] sur l'analyse des mouvements humains où la dynamique des mouvements est prise en compte grâce à la modélisation de la suite de positions prises par le corps humain, observée à travers les réponses de capteurs sensoriels, par une VLHMM. Collet et al. [2008] prouve la consistance en probabilité d'un estimateur inspiré de l'estimateur de l'arbre de contexte de Rissanen dans le cas où $\mathbb{X} = \{0, 1\}$ et où les distributions d'émission des variables Y_t correspondent à des sauts de type Bernouilli des variables X_t.

Nous considérons $\{X_t, Y_t\}_{t\in\mathbb{Z}}$ une VLHMM sur $\mathbb{X} \times \mathbb{Y}$ où $\mathbb{X}$ est un ensemble fini muni de la mesure de comptage et $\mathbb{Y}$ est un espace polonais muni d'une σ-algèbre $\mathcal{F}_{\mathbb{Y}}$ et d'une mesure λ sur $(\mathbb{Y}, \mathcal{F}_{\mathbb{Y}})$. Nous supposons tout d'abord que la VLMC $\{X_t\}_{t\in\mathbb{Z}}$ est stationnaire et à arbre de contextes $\tau_\star$ inconnu. Nous supposons néanmoins que $\tau_\star$ est fini et complet. Nous supposons que le noyau de transition G sur $\mathbb{X} \times \mathbb{Y}$ définissant la distribution de passage d'un état $x \in \mathbb{X}$ à Y_t est à densité par rapport à λ et que cette densité appartient à un ensemble paramétrique $\{g_{\theta_e}\}_{\theta_e\in\Theta_e}$ où Θ_e est l'espace de paramètres d'émission. Pour tout arbre de contextes τ fini complet, considérons $\Theta_{t,\tau}$, appelé espace de paramètres de transitions, et défini par

$$\Theta_{t,\tau} \stackrel{\text{def}}{=} \left\{ \theta_t = \{p_{s,x}\}_{s\in\tau,\ x\in\mathbb{X}} \in \mathbb{R}_+^{|\tau|\cdot|\mathbb{X}|} \;\middle|\; \forall s \in \tau,\ \sum_{x\in\mathbb{X}} p_{s,x} = 1 \right\} .$$

Définissons alors pour tout arbre complet et fini τ, l'espace de paramètres associé à τ par

$$\Theta_\tau \stackrel{\text{def}}{=} \Theta_{t,\tau} \times \Theta_e .$$

Sans perte de généralité, nous supposons que l'espace de paramètres Θ_e est de la forme suivante,

$$\Theta_e = \left\{\theta_e = \left(\theta_{e,1}, \ldots, \theta_{e,|\mathbb{X}|}, \eta\right)\right\} \in \left(\mathbb{R}^{d_e}\right)^{|\mathbb{X}|} \times \mathbb{R}^{m_e} ,$$

et que, pour tout (x, y) de $\mathbb{X} \times \mathbb{Y}$, pour tout $\theta = \left(\theta_{e,1}, \ldots, \theta_{e,|\mathbb{X}|}, \eta\right)$, les densités d'émission sont de la forme

$$g_{\theta_e}(x, y) = g_{\theta_{e,x},\eta}(y) .$$

Pour un arbre de contexte τ donné, et pour tout $\theta = (\theta_t, \theta_e) \in \Theta_\tau$, avec $\theta_t = \{p_{s,x}\}_{s\in\tau,\ x\in\mathbb{X}}$, notons L_θ que nous appellerons vraisemblance des observations sous θ, définie, quelque soit $y_{1:n}$ vecteur d'observations de $\mathbb{Y}^n$, par

$$L_\theta(y_{1:n}) \stackrel{\text{def}}{=} \sum_{x_{1:n}\in\mathbb{X}^n} \left[\prod_{t=1}^{n} g_{\theta_{e,x_i},\eta}(y_i)\right] g_{\theta_t}(x_{1:n}) \ ,$$

où

$$g_{\theta_t}(x_{1:n}) \stackrel{\text{def}}{=} \sum_{x_{-d(\tau)+1:0}\in\mathbb{X}^{d(\tau)}} \nu_{d(\tau),\theta_t}(x_{-d(\tau)+1:0}) \prod_{t=1}^{n} p_{f_\tau(x_{-d(\tau)+1:t-1}),x_t} \ ,$$

avec $\nu_{d(\tau),\theta_t}$ étant une densité de probabilité par rapport à la mesure de comptage sur $\mathbb{X}^{d(\tau)}$. Nous définissons alors l'estimateur du maximum de vraisemblance pénalisé $\widehat{\tau}_n$ par

$$\widehat{\tau}_n \stackrel{\text{def}}{=} \underset{\tau \text{ fini complet}}{\operatorname{argmin}} \left\{ - \sup_{\theta\in\Theta_\tau} \log L_\theta(y_{1:n}) + \operatorname{pen}(n,\tau) \right\} \ .$$

Les résultats présentés dans la section 11.3 démontrent la consistance forte de $\widehat{\tau}_n$ lorsque la pénalité $\operatorname{pen}(n,\tau)$ est de la forme $C(\tau)\log n$. L'outil fondamental pour obtenir notre résultat est une inégalité empruntée à la théorie de l'information qui permet de contrôler la vraisemblance $L_\theta(y_{1:n})$ par une densité de mélange $\mathbb{KT}^n_\tau$ définie sur $\mathbb{Y}^n$. Cette densité mélange est uniquement utilisée comme outil de preuve et son calcul n'est pas nécessaire dans la construction de $\widehat{\tau}_n$. $\mathbb{KT}^n_\tau$ est défini, pour tout $y_{1:n} \in \mathbb{Y}^n$ par

$$\mathbb{KT}^n_\tau(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n} \mathbb{KT}_{\tau,t}(x_{1:n}) \mathbb{KT}^n_e(y_{1:n}|X_{1:n}) \ .$$

avec

$$\mathbb{KT}^n_e(y_{1:n}|x_{1:n}) = \int_{\Theta_e} \left[\prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(y_i)\right] \pi^n_e(\mathrm{d}\theta_e) \ ,$$

où π^n_e est appelée loi mélange d'émission, le choix de cette loi d'émission dépend du modèle considéré et peut changer avec n. $\mathbb{KT}_{\tau,t}$ est le mélange de Krichevski-Trofimov associé aux priors de Dirichlet $\{\pi_s\}_{s\in\tau}$, distributions de Dirichlet $\mathcal{D}\left(\frac{1}{2},\ldots,\frac{1}{2}\right)$ sur $[0,1]^{|\mathbb{X}|}$ et défini, pour tout $x_{1:n} \in \mathbb{X}^n$ et tout $\theta_t = \{P_{s,x}\}_{s\in\tau,\ x\in\mathbb{X}}$, par

$$\mathbb{KT}_{\tau,t}(x_{1:n}) = \left(\frac{1}{|\mathbb{X}|}\right)^{d(\tau)} \int_{\Theta_t} \prod_{s\in\tau} \int_{\{P_{s,x}\}_{x\in\mathbb{X}}\in[0,1]^{|\mathbb{X}|}} \prod_{x\in\mathbb{X}} P_{s,x}^{a_s^x(x_{1:n})} \pi_s\left(\mathrm{d}\left\{P_{s,x}\right\}_{x\in\mathbb{X}}\right) \ ,$$

où $a_s^x(x_{1:n})$ représente le nombre de fois où x apparaît dans le contexte s. Le résultat principal de la section 11.3 démontre la consistance forte de $\widehat{\tau}_n$, sous des conditions classiques d'irréductibilité de la VLMC $\{X_t\}_{t\in\mathbb{Z}}$ et d'identifiabilité du modèle lorsque la pénalité $\operatorname{pen}(n,\tau)$ est de la forme

$$\operatorname{pen}(n,\tau) = \left[\sum_{t=1}^{|\tau|} \frac{(k-1)t+\alpha}{2}\right] \log n \ , \tag{10.1}$$

avec α réel positif. Notons $\tau^\star$ l'arbre de contexte associé à la VLMC $\{X_t\}_{t\in\mathbb{Z}}$, le théorème 1 de la section 11.3 assure alors que, si l'on peut trouver une suite de distributions prior $\{\pi_e\}_{n\geq 1}$ telle que,

$$\sup_{\theta_e\in\Theta_e}\sup_{x_{1:n}}\left[\log\prod_{t=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq b\log n\ , \tag{10.2}$$

avec $b>0$, $\mathbb{P}$-éventuellement presque sûrement, alors, si $\alpha > 2(b+1)$ dans (10.1), $\widehat{\tau}_n = \tau^\star$, $\mathbb{P}$-éventuellement presque sûrement, à permutation des éléments de $\mathbb{X}$ près.

La démonstration de ce résultat repose principalement sur une inégalité empruntée à la théorie de l'information permettant le contrôles des densités g_{θ_t} par la loi mélange $\mathbb{KT}_{\tau,t}$:

$$0 \leq \sup_{\theta_t\in\Theta_{t,\tau}}\max_{x_{1:n}}\{\log g_{\theta_t}(x_{1:n}) - \log \mathbb{KT}_{\tau,t}(x_{1:n})\} \leq |\tau|\gamma\left(\frac{n}{|\tau|}\right) + d(\tau)\log|\mathbb{X}|\ ,$$

avec $\gamma(x) = \frac{|\mathbb{X}|-1}{2}\log x + \log|\mathbb{X}|$. Ainsi, lorsque l'équation (10.2) est vérifiée, les fluctuations de la vraisemblance L_θ peuvent être contrôlées par la loi mélange $\mathbb{KT}_\tau^n$. Ce résultat s'applique facilement pour certains modèles d'émission comme le cas gaussien à variance connue, ou le cas poissonien.

L'autre résultat important, démontré dans la section 11.4, est la démonstration de l'inégalité (10.2) dans le cas Gaussien avec variance inconnue. Dans ce cas, $\mathbb{Y}=\mathbb{R}$, l'espace de paramètres d'émission est de la forme $\Theta_e = \left\{(m_1,\ldots,m_{|\mathbb{X}|},\sigma^2)\ ;\ m_i\in\mathbb{R},\ \sigma^2>0\right\}$ et les densités d'émission sont données, pour tout $x\in\mathbb{X}$ et $y\in\mathbb{R}$, par :

$$g_{m_x,\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(m_x-y)^2}{2\sigma^2}\right)\ .$$

La difficulté de cette démonstration réside en grande partie dans le choix de la suite de priors $\{\pi_e^n\}_{n\geq 1}$ permettant d'établir l'inégalité (10.2).

Un algorithme permettant le calcul de $\widehat{\tau}_n$ dans le cas général, basé sur l'algorithme EM et utilisant des techniques d'élagage similaires à celle utilisée par Rissanen et son algorithme *context* sera aussi présenté dans la section 11.5. Cet algorithme a pour objectif de trouver l'arbre de contexte τ minimisant $-\sup_{\theta\in\Theta_\tau} L_\theta(Y_{1:n}) + \mathrm{pen}(n,\tau)$ (que l'on appellera score de tau) sans pour autant avoir à calculer cette quantité pour tous les arbres τ possibles. Nous verrons dans la section 11.5.2 une illustration de cet algorithme sur données simulées. La convergence presque sûre de l'estimateur $\widehat{\tau}_n$ vers le vrai arbre de contexte $\tau^\star$ peut être observée sur les résultats (Voir tableaux 11.1 à 11.4 de la section 11.5.2), cependant, notre algorithme semble parfois sélectionner le mauvais arbre : dans certains résultats de la simulation, l'algorithme sélectionne un arbre τ alors que le score de τ est supérieur à celui de $\tau^\star$. Ceci peut être du à la méthode d'estimation des quantités $\sup_{\theta\in\Theta_\tau} L_\theta(Y_{1:n})$ qui font appel à un algorithme récursif de type EM pouvant conduire à une convergence vers un maximum local et non global de la vraisemblance limite.

Bien que la convergence de $\widehat{\tau}_n$ semble se confirmer sur les résultats de simulation (sauf dans les cas mentionnés précédemment), on s'aperçoit que notre estimateur a tendance à sous estimer $\tau^\star$, symptôme d'un choix de pénalité trop grand retardant (en terme de taille d'échantillon) la convergence. Les résultats des simulations de la section 11.5.2 comparent les performances de notre estimateur avec les performance de BIC (l'estimateur construit en choisissant $\mathrm{pen}(n,\tau) = \frac{|\mathbb{X}|-1}{2}|\tau|\log n$, soit une pénalité plus petite que (10.1)). La convergence presque sûre de BIC semble apparaître sur les

simulations, avec quelquefois des surestimations pouvant être causées par les phénomènes de minima locaux, mais la convergence est atteinte plus tôt que celle de notre estimateur.

Application de l'algorithme à la géolocalisation :

Cet algorithme possède une complexité bien plus élevée que les algorithmes développés pour les VLMC puisqu'il nécessite le calcul de la vraisemblance (effectué grâce à l'algorithme EM) pour un grand nombre d'arbres de contextes. L'estimation de l'arbre de contextes dans le cadre de la géolocalisation (où le nombre d'états à considérer est élevé) afin de prendre en compte la dynamique du mobile dans le modèle, nécessiterait donc une optimisation de l'algorithme. De plus, il faudrait donner une structure de dépendance spatiale (comme au chapitre 5) afin d'atténuer les effets du *label switching*, présenté dans la section 4.1.

Bibliographie

P. Collet, A. Galves, and F. Leonardi. Random perturbations of stochastic processes with unbounded variable length memory. *Electron. J. Probab.*, 13 :no. 48, 1345–1361, 2008. ISSN 1083-6489. doi : 10.1214/EJP.v13-538.

I. Csiszar and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inf. Theor.*, 52(3) :1007–1016, March 2006. ISSN 0018-9448. doi : 10.1109/TIT.2005.864431. URL `http://dx.doi.org/10.1109/TIT.2005.864431`.

A. Garivier. Consistency of the unlimited BIC Context Tree estimator. *IEEE Trans. Inform. Theory*, 52 :4630–4635, 2006.

J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29 :656 – 664, 1983.

Y. Wang. The variable-length hidden Markov Model and its applications on sequential data mining. Technical report, Departement of computer science, 2005.

Zhou L. Wang J. Wang, Y. and Z.Q. Liu. Mining complex time-series by learning Markovian models. In *Proceedings ICDM'06, sixth international conference on data mining*, China, 2005.

F. M. J. Willems. The Context-Tree Weighting Method : Extensions. *IEEE Transactions on Information Theory*, 44 :792–798, 1994.

F. M. J. Willems, Y.i M. Shtarkov, and T. J. Tjalkens. The Context Tree Weighting Method : Basic Properties. *IEEE Transactions on Information Theory*, 41 :653–664, 1995.

Chapter 11

Context tree estimation in variable length hidden Markov models

Thierry Dumont

Sommaire

Abstract

We address the issue of context tree estimation in variable length hidden Markov models. We propose an estimator of the context tree of the hidden Markov process which needs no prior upper bound on the depth of the context tree. We prove that the estimator is strongly consistent. This uses information-theoretic mixture inequalities in the spirit of Finesso [1990], Gassiat and Boucheron [2003]. We propose an algorithm to efficiently compute the estimator and provide simulation studies to support our result.

Keywords: Variable length, hidden Markov models, context tree, consistent estimator, mixture inequalities.

11.1 Introduction

A variable length hidden Markov model (VLHMM) is a bivariate stochastic process $(X_n, Y_n)_{n\geq 0}$ where $(X_n)_{n\geq 0}$ (the state sequence) is a variable length Markov chain (VLMC) in a state space $\mathbb{X}$ and, conditionally on $(X_n)_{n\geq 0}$, $(Y_n)_{n\geq 0}$ is a sequence of independent variables in a state space $\mathbb{Y}$ such that the conditional distribution of Y_n given the state sequence (called the emission distribution) depends on X_n only. Such processes fall into the general framework of latent variable processes, and reduce to hidden Markov models (HMM) in case the state sequence is a Markov chain. Latent variable processes are used as a flexible tool to model dependent non-Markovian time series, and the statistical problem is to estimate the parameters of the distribution when only $(Y_n)_{n\geq 0}$ is observed. We will consider in this paper the case where the hidden process may take only a fixed and known number of values, that is the case where the state space $\mathbb{X}$ is finite with known cardinality k.

The dependence structure of a latent variable process is driven by that of the hidden process $(X_n)_{n\geq 0}$, which is assumed here to be a variable length Markov chain (VLMC). Such processes were first introduced by Rissanen in Rissanen [1983] as a flexible and parsimonious modelization tool for data compression, approximating Markov chains of finite orders. Recall that a Markov process of order d is such that the conditional distribution of X_n given all past values depends only on the d previous ones $X_{n-1}, \ldots, X_{n-d}$. But different past values may lead to identical conditional distributions, so that all k^d possible past values are not needed to describe the distribution of the process. A VLMC is such that the probability of the present state depends only on a finite part of the past, and the length of this relevant portion, called context, is a function of the past itself. No context may be a proper postfix of any other context, so that the set of all contexts may be represented as a rooted labelled tree. This set is called the context tree of the VLMC.

Variable length hidden Markov models appear for the first time, to our knowledge, in movement analysis Wang and Liu [2005], Wang [2005]. Human movement analysis is the interpretation of movements as sequences of poses. Wang [2005] analyses the movement through 3D rotations of 19 major joints of human body. Wang and al. then use a VLHMM representation where X_n is the pose at time n and Y_n is the body position given by the 3D rotations of the 19 major points. They argue that "VLHMM is superior in its efficiency and accuracy of modeling multivariate time-series data with highly-varied dynamics".

VLHMM could also be used in WIFI based indoor positioning systems (see Evennou [2007]). Here X_n is a mobile device position at time n and Y_n is the received signal strength (RSS) vector at time n. Each component of the RSS vector represents the strength of a signal sent by a WIFI access point. In practice, the aim is to estimate the positions of the device $(X_n)_{n\geq 0}$ on the basis of the observations $(Y_n)_{n\geq 0}$. The distribution of Y_n given $X_n = x$ for any location x is beforehand calibrated for a finite number of locations $(L_1, ..., L_k)$. A Markov chain on the finite set $(L_1, ..., L_k)$ is then used to model the sequence of positions $(X_n)_{n\geq 0}$. Again VLHMM model would lead to efficient and accurate estimation of the device position.

The aim of this paper is to provide a statistical analysis of variable length hidden Markov models and, in particular, to propose a consistent estimator of the context tree of the hidden VLMC on the basis of the observations $(Y_n)_{n\geq 0}$ only. We consider a parametrized family of VLHMM, and we use a penalized likelihood method to estimate the context tree of the hidden VLMC. To each possible

context tree τ, if Θ_τ is the set of possible parameters, we define

$$\hat{\tau}_n = \underset{\tau}{\operatorname{argmin}} \left\{ - \sup_{\theta \in \Theta_\tau} \log g_\theta(Y_{1:n}) + pen(n, \tau) \right\} ,$$

where $g_\theta(y_{1:n})$ is the density of the distribution of the observation $Y_{1:n} = (Y_1, \ldots, Y_n)$ under the parameter θ with respect to some dominating positive measure, and $pen(n, \tau)$ is a penalty that depends on the number n of observations and the context tree τ. Our aim is to find penalties for which the estimator is strongly consistent without any prior upper bound on the depth of the context tree, and to provide a practical algorithm to compute the estimator.

Context tree estimation for a VLHMM is similar to order estimation for a HMM in which the order is defined as the unknown cardinality of the state space $\mathbb{X}$. The main difficulty lies in the calibration of the penalty, which requires some understanding of the growth of the likelihood ratios (with respect to orders and to the number of observations). In particular cases, the fluctuations of the likelihood ratios may be understood via empirical process theory, see the recent works van Handel [2011] for finite state Markov chains and Gassiat and van Handel [2010] for independent identically distributed observations. Latent variable models are much more complicated, see for instance Gassiat and Keribin [2000] where it is proved in the HMM situation that the likelihood ratio statistics converges to infinity for overestimated order. We thus use an approach based on information theory tools to understand the behavior of likelihood ratios. Such tools have been successfull for HMM order estimation problems and were used in Gassiat and Boucheron [2003], Finesso [1990] for discrete observations and in Chambaz et al. [2009] for Poisson emission distributions or Gaussian emission distributions with known variance. Our main result shows that for a penalty of form $C(\tau) \log n$, $\hat{\tau}_n$ is strongly consistent, that is converges almost surely to the true unknown context tree. Here, $C(\tau)$ has an explicit formulation but is slightly bigger than $(k-1)|\tau|/2$ which gives the popular BIC penalty. We study the important situation of Gaussian emissions with unknown variance, and prove that our consistency theorem holds in this case.

Computation of the estimator requires computation of the maximum likelihood for all possible context trees. As usual, the EM algorithm may be used to compute the maximum likelihood estimator for the parameters when the context tree is fixed. We then propose an algorithm to compute the estimator, which prevents the exploration of a too large number of context trees. In general the EM algorithm needs to be run several times with different initial values to avoid local extrema traps. In the important situation of Gaussian emissions, we propose a way to choose the initial parameters so that only one run of the EM algorithm is needed. Simulations compare penalized maximum likelihood estimators of the context tree τ of the hidden VLMC using our penalty and using BIC penalty.

The structure of this paper is the following. Section 11.2 describes the model and gives the notations. Section 11.3 presents the information theory tools we use, states the main consistency result and applies it to Poisson emission distributions and Gaussian emission distributions with known variance. Section 11.4 proves the result for Gaussian emission distributions with unknown variance. In section 11.5, we describe the algorithm to compute the estimator and we give the simulation results. The proofs that are not essential at first reading are detailed in the Appendix.

11.2 Basic settings and notations

Let $\mathbb{X}$ be a finite set whose cardinality is denoted by $|\mathbb{X}| = k$, that we identify with $\{1, \dots, k\}$. Let $\mathcal{F}_{\mathbb{X}}$ be the finite collection of subsets of $\mathbb{X}$. Let $\mathbb{Y}$ be a Polish space endowed with its Borel sigma-field $\mathcal{F}_{\mathbb{Y}}$. We will work on the measurable space $(\Omega, \mathcal{F})$ with $\Omega = (\mathbb{X} \times \mathbb{Y})^{\mathbb{N}}$ and $\mathcal{F} = (\mathcal{F}_{\mathbb{X}} \otimes \mathcal{F}_{\mathbb{Y}})^{\otimes \mathbb{N}}$.

11.2.1 Context trees and variable length Markov chains

A string $s = x_k x_{k+1} ... x_l \in \mathbb{X}^{l-k+1}$ is denoted by $x_{k:l}$ and its length is then $l(s) = l - k + 1$. We call letters of s its components $x_i,\ i = k, \dots, l$. The concatenation of the strings u and v is denoted by uv. A string v is a *postfix* of a string s if there exists a string u such that $s = uv$.
A set τ of strings and possibly semi-infinite sequences is called a *tree* if the following *tree property* holds : no $s \in \tau$ is postfix of any other $s' \in \tau$. A tree τ is *irreducible* if no element $s \in \tau$ can be replaced by a postfix without violating the tree property. It is *complete* if each node except the leaves has $|\mathbb{X}|$ children exactly. We denote by $d(\tau)$ the depth of τ : $d(\tau) = \max\left\{l(s) \;\middle|\; s \in \tau\right\}$.

Let now Q be the distribution of an ergodic stationary process $(X_n)_{n\in\mathbb{Z}}$ on $(\mathbb{X}^{\mathbb{Z}}, \mathcal{F}_{\mathbb{X}}^{\otimes\mathbb{Z}})$, and for any $m \leq n$ and any $x_{m:n}$ in $\mathbb{X}^{n-m+1}$, write $Q(x_{m:n})$ for $Q(X_{0:n-m} = x_{m:n})$.

Definition 1. *Let τ be a tree. τ is called a Q-adapted context tree if for any string s in τ such that $Q(s) > 0$:*

$$\forall x_0 \in \mathbb{X}, Q(X_0 = x_0 \big| X_{-\infty:-1} = x_{-\infty:-1}) = Q(X_0 = x_0 \big| X_{-l(s):-1} = s)\ , \tag{11.1}$$

whenever s is postfix of the semi infinite sequence $x_{-\infty:-1}$. Moreover, if for any $s \in \tau$, $Q(s) > 0$ and no proper postfix of s has the property (11.1)*, then τ is called the minimal context tree of the distribution Q, and $(X_n)_{n\in\mathbb{Z}}$ is called a variable length Markov chain (VLMC).*

If a tree τ is Q-adapted, then for all sequences $x_{-\infty:-1}$ such that for any $M \geq 1$, $Q(x_{-M:-1}) > 0$, there exists a unique string in τ which is postfix of $x_{-\infty:-1}$. We denote this postfix by $\tau(x_{-\infty:-1})$.
A tree τ is said to be a *subtree* of τ' if for each string s' in τ' there exists a string s in τ which is postfix of s'. Then if τ is a Q-adapted tree, any tree τ' such that τ is a subtree of τ' will be Q-adapted.

Definition 2. *Let Q be the distribution of a VLMC $(X_n)_{n\in\mathbb{Z}}$. Let τ_0 be its minimal context tree. There exists a unique complete tree $\tau^\star$ such that τ_0 is a subtree of $\tau^\star$ and*

$$|\tau^\star| = \min\left\{|\tau| \ :\ \tau \text{ is a complete tree and } \tau_0 \text{ is a subtree of } \tau\right\}\ .$$

$\tau^\star$ is called the minimal complete context tree of the distribution Q of the VLMC $(X_n)_{n\in\mathbb{Z}}$.

Let us define, for any complete tree τ, the set of transition parameters:

$$\Theta_{t,\tau} = \left\{ (P_{s,i})_{s\in\tau, i\in\mathbb{X}} \ :\ \forall s \in \tau, \forall i \in \mathbb{X},\ P_{s,i} \geq 0 \text{ and } \sum_{i=1}^{k} P_{s,i} = 1 \right\}\ .$$

If $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with minimal complete context tree $\tau^\star$ and transition parameters $\theta_t^\star = \left(P^\star_{s,i}\right)_{s\in\tau^\star, i\in\mathbb{X}} \in \Theta_{t,\tau^\star}$, for any complete tree τ such that $\tau^\star$ is a subtree of τ, there exists a unique

$\theta_t = (P_{s,i})_{s\in\tau, i\in\mathbb{X}} \in \Theta_{t,\tau}$ that defines the same VLMC transition probabilities, namely: for any $s \in \tau$, there exists a unique $u \in \tau^\star$ which is a postfix of s, and for all $i \in \mathbb{X}$, $P_{s,i} = P^\star_{u,i}$. Of course, a parameter in $\Theta_{t,\tau}$ might be not sufficient to define a unique distribution of a VLMC (if there is no unique stationary distribution). But the parameter defines a unique distribution of VLMC if, for instance, the Markov chain $([X_{n-d(\tau)+1}, \ldots, X_n])_{n\in\mathbb{Z}}$ it defines is irreducible.

11.2.2 Variable length hidden Markov models

A variable length hidden Markov model (VLHMM) is a bivariate stochastic process $(X_n, Y_n)_{n\geq 0}$ where $(X_n)_{n\geq 0}$ (the state sequence) is a (non observed) stochastic process which is the restriction to non negative indices of a VLMC $(X_n)_{n\in\mathbb{Z}}$ with values in $\mathbb{X}$ and, conditionally on $(X_n)_{n\geq 0}$, $(Y_n)_{n\geq 0}$ is a sequence of independent variables in the state space $\mathbb{Y}$ such that for any integer n, the conditional distribution of Y_n given the state sequence (called the emission distribution) depends on X_n only.

We assume that the emission distributions are absolutely continuous with respect to some positive measure μ on $(\mathbb{Y}, \mathcal{F}_\mathbb{Y})$ and are parametrized by a set of parameters $\Theta_e \subset (\mathbb{R}^{d_e})^k \times \mathbb{R}^{m_e}$, so that the set of emission densities (the possible densities of the distribution of Y_n conditional to $X_n = x$) is $\{(g_{\theta_{e,x},\eta}(.))_{x\in\mathbb{X}},\ \theta_e = (\theta_{e,1}, \ldots, \theta_{e,k}, \eta) \in \Theta_e\}$. For any complete tree τ, we define now the parameter set :

$$\Theta_\tau = \Theta_{t,\tau} \times \Theta_e,$$

and define, for $\theta = (\theta_t, \theta_e) \in \Theta_\tau$, $\mathbb{P}_\theta$ the probability of the VLHMM $(X_n, Y_n)_{n\geq 0}$ such that $(X_n)_{n\in\mathbb{Z}}$ is the VLMC with complete context tree τ, transition parameter θ_t, and for any $(u_1, u_2) \in \mathbb{N}^2$, $u_1 \leq u_2$, any sets $A_{u_1}, \ldots, A_{u_2}$ in $\mathcal{F}_Y$, any $x_{u_1:u_2} \in \mathbb{X}^{u_2-u_1+1}$,

$$\mathbb{P}_\theta\left(Y_{u_1} \in A_{u_1}, \ldots, Y_{u_2} \in A_{u_2} \middle| X_{u_1} = x_{u_1}, \ldots, X_{u_2} = x_{u_2}\right) = \prod_{u=u_1}^{u_2} \left[\int_{A_u} g_{\theta_{e,x_u},\eta}(y) d\mu(y)\right] .$$

Of course, as noted before, it can happen that θ_t does not define a unique VLHMM. We shall however do not consider this question since we shall assume that the true parameter defines an irreducible hidden VLMC, and we shall introduce initial distributions to define a computable likelihood: throughout the paper we shall assume that the observations $(Y_1, ..., Y_n) = Y_{1:n}$ come from a VLHMM with parameter $\theta^\star$ such that $\tau^\star$ is the minimal *complete* context tree of the hidden VLMC, and such that $([X_{n-d(\tau^\star)+1}, \ldots, X_n])_{n\in\mathbb{Z}}$ is a stationary and irreducible Markov chain. And to define a computable likelihood, we introduce, for any positive integer d, a probability distribution ν_d on $\mathbb{X}^d$ so that, for any complete tree τ and any $\theta = (\theta_t, \theta_e) \in \Theta_\tau$, we set what will be called the likelihood:

$$\forall y_{1:n} \in \mathbb{Y}^n,\ g_\theta(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n} \left[\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)\right] g_{\theta_t}(x_{1:n}) , \tag{11.2}$$

where, if $\theta_t = (P_{s,x})_{s\in\tau, x\in\mathbb{X}}$:

$$g_{\theta_t}(x_{1:n}) = \sum_{x_{-d(\tau)+1:0}\in\mathbb{X}^{d(\tau)}} \left[\nu_{d(\tau)}(x_{-d(\tau)+1:0}) \prod_{i=1}^n P_{\tau(x_{-d(\tau)+i:i-1}),x_i}\right] . \tag{11.3}$$

We are concerned with the statistical estimation of the tree $\tau^\star$ using a method that involves no prior upper bound on the depth of $\tau^\star$. Define the following estimator of the minimal complete context tree $\tau^\star$:

$$\hat{\tau}_n = \underset{\tau \text{ complete tree}}{\operatorname{argmin}} \Big[- \sup_{\theta \in \Theta_\tau} \log g_\theta(Y_{1:n}) + pen(n, \tau) \Big] , \tag{11.4}$$

where $pen(n, \tau)$ is a penalty term depending on the number of observations n and the complete tree τ.

The label switching phenomenon occurs in statistical inference of VLHMM as it occurs in statistical inference of HMM and of population mixtures. That is: applying a label permutation on $\mathbb{X}$ does not change the distribution of $(Y_n)_{n\geq 0}$. Thus, if σ is a permutation of $\{1, ..., k\}$ and τ is a complete tree, we define the complete tree $\sigma(\tau)$ by

$$\sigma(\tau) = \left\{ \sigma(x_1)...\sigma(x_l) \middle| \; x_{1:l} \in \tau \right\} .$$

Definition 3. *If τ and τ' are two complete trees, we say that τ and τ' are equivalent, and denote it by $\tau \sim \tau'$, if there exists a permutation σ of $\mathbb{X}$ such that $\sigma(\tau) = \tau'$.*

We then choose $pen(n, \tau)$ to be invariant by permutation, that is: for any permutation σ of $\mathbb{X}$, $pen(n, \sigma(\tau)) = pen(n, \tau)$. In this case, for any complete tree τ,

$$- \sup_{\theta \in \Theta_{\hat{\tau}_n}} \log g_\theta(Y_{1:n}) + pen(n, \tau) = - \sup_{\theta \in \Theta_{\sigma(\hat{\tau}_n)}} \log g_\theta(Y_{1:n}) + pen(n, \sigma(\tau))$$

so that the definition of $\hat{\tau}_n$ requires a choice in the set of minimizers of (11.4).

Our aim is now to find penalties allowing to prove the strong consistency of $\hat{\tau}_n$, that is such that $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$- eventually almost surely as $n \to \infty$.

11.3 The general strong consistency theorem

In this section, we first recall the tools borrowed from information theory, and set the result that we use in order to find a penalty insuring the strong consistency of $\hat{\tau}_n$. Then we give our general strong consistency theorem, and straightforward applications. Application to Gaussian emissions with unknown variance, which is more involved, is deferred to the next section.

11.3.1 An information theoretic inequality

We shall introduce mixture probability distributions on $\mathbb{Y}^n$ and compare them to the maximum likelihood, in the same way as Krichevsky and Trofimov [1981] first did; see also Catoni and Picard [2004] and Gassiat [2011] for tutorials and use of such ideas in statistical methods. For any complete tree τ, we define, for all positive integer n, the mixture measure $\mathbb{KT}^n_\tau$ on $\mathbb{Y}^n$ using a prior π^n on Θ_τ:

$$\pi^n(d\theta) = \pi_t(d\theta_t) \otimes \pi^n_e(d\theta_e)$$

where π^n_e is a prior on Θ_e that may change with n, and π_t the prior on Θ_t such that, if $\theta_t = (P_{s,i})_{s\in\tau, i\in\mathbb{X}}$,

$$\pi_t(d\theta_t) = \otimes_{s\in\tau} \pi_s(d(P_{s,i})_{i\in\mathbb{X}}) ,$$

where $(\pi_s)_{s\in\tau}$ are Dirichlet $\mathcal{D}(\frac{1}{2},...,\frac{1}{2})$ distributions on $[0,1]^{|\mathbb{X}|}$. Then $\mathbb{KT}^n_\tau$ is defined on $\mathbb{Y}^n$ by

$$\mathbb{KT}^n_\tau(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n} \mathbb{KT}_{\tau,t}(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n}) \ ,$$

where

$$\mathbb{KT}^n_e(y_{1:n}|x_{1:n}) = \int_{\Theta_e} \left[\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)\right] \pi^n_e(d\theta_e) \ ,$$

and

$$\begin{aligned}\mathbb{KT}_{\tau,t}(x_{1:n}) &= \left(\frac{1}{k}\right)^{d(\tau)} \int_{\Theta_t} \mathbb{P}_{\theta_t}\left(x_{d(\tau)+1:n}|x_{1:d(\tau)}\right) \pi_t(d\theta_t) \\ &= \left(\frac{1}{k}\right)^{d(\tau)} \prod_{s\in\tau} \int_{[0,1]^{|\mathbb{X}|}} \prod_{i=1}^k P_{s,i}^{a^x_s(x_{1:n})} \pi_s(d(P_{s,i})_{i\in\mathbb{X}}) \ ,\end{aligned}$$

where $a^x_s(x_{1:n})$ is the number of times that x appears in context s, that is

$$a^x_s(x_{1:n}) = \sum_{i=d(\tau)+1}^n \mathbf{1}_{x_i=x, x_{i-l(s),i-1}=s} \ .$$

The following inequality will be a key tool to control the fluctuations of the likelihood.

Proposition 2. *There exists a finite constant D depending only on k such that for any complete tree τ, and any $y_{1:n}\in\mathbb{Y}^n$:*

$$\begin{aligned}0 \le \sup_{\theta\in\Theta_\tau} \log g_\theta(y_{1:n}) - \log \mathbb{KT}^n_\tau(y_{1:n}) \le \sup_{x_{1:n}} &\left[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n})\right] \\ &+ \frac{k-1}{2}|\tau|\log n + D \ .\end{aligned}$$

Proof. Let τ be a complete tree. For any $\theta\in\Theta_\tau$,

$$\begin{aligned}\frac{g_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} &= \frac{\sum_{x_{1:n}} g_{\theta_t}(x_{1:n}) \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)}{\sum_{x_{1:n}} \mathbb{KT}_\tau(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n})} \\ &\le \max_{x_{1:n}} \frac{g_{\theta_t}(x_{1:n}) \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i)}{\mathbb{KT}_\tau(x_{1:n})\mathbb{KT}^n_e(y_{1:n}|x_{1:n})} \ .\end{aligned}$$

Thus,

$$\log \frac{g_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} \le \sup_{x_{1:n}} \left[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n}) + |\tau|\gamma\left(\frac{n}{|\tau|}\right) + d(\tau)\log k\right] ,$$

where $\gamma(x) = \frac{k-1}{2}\log x + \log k$, using Gassiat [2011]. Then

$$\log \frac{g_\theta(y_{1:n})}{\mathbb{KT}^n_\tau(y_{1:n})} \leq \sup_{x_{1:n}} \left[\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(y_i) - \log \mathbb{KT}^n_e(y_{1:n}|x_{1:n}) \right] + \frac{k-1}{2}|\tau|\log n + D(\tau) \ ,$$

where $D(\tau) = -\frac{k-1}{2}|\tau|\log|\tau| + |\tau|\log k + d(\tau)\log k$. Now, since τ is complete, $d(\tau) \leq \frac{|\tau|-k}{k-1}$, so that

$$D(\tau) \leq |\tau|\Big(\log k - \frac{k-1}{2}\log|\tau|\Big) + \frac{|\tau|-k}{k-1}\log k \ .$$

But the upper bound in the inequality tends to $-\infty$ when $|\tau|$ tends to ∞, so that there exists a constant D depending only on k such that for any complete tree τ, $D(\tau) \leq D$. □

11.3.2 Strong consistency theorem

Let $\theta^\star = (\theta^\star_t, \theta^\star_e)$ with $\theta^\star_t = (P^\star_{s,i})_{s\in\tau^\star, i\in\mathbb{X}}$, and $\theta^\star_e = (\theta^\star_{e,1}, ..., \theta^\star_{e,k}, \eta^\star)$ be the true parameters of the VLHMM.
Let us now define for any positive α, the penalty:

$$pen_\alpha(n,\tau) = \left[\sum_{t=1}^{|\tau|} \frac{(k-1)t+\alpha}{2} \right] \log n \ . \tag{11.5}$$

Notice that the complexity of the model is taken into account through the cardinality of the tree τ. We need to introduce further assumptions.

- **(A1)**. The Markov chain $((X_{n-d(\tau^\star)+1}, \ldots, X_n))_{n\geq d(\tau^\star)}$ is irreducible.
- **(A2)**. For any complete tree τ such that $|\tau| \leq |\tau^\star|$ and which is not equivalent to $\tau^\star$, for any $\theta \in \Theta_\tau$, the random sequence $(\theta_{e,X_n})_{n\in\mathbb{Z}}$ where $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with transition probabilities θ_t, has a different distribution than $(\theta^\star_{e,X_n})_{n\in\mathbb{Z}}$ where $(X_n)_{n\in\mathbb{Z}}$ is a VLMC with transition probabilities $\theta^\star_t$.
- **(A3)**. The family $\{g_{\theta_e}, \theta_e \in \Theta_e, \}$ is such that for any probability distributions $(\alpha_i)_{i=1,\ldots,k}$ and $(\alpha'_i)_{i=1,\ldots,k}$ on $\{1, ..., k\}$, any $(\theta_1, \ldots, \theta_k, \eta) \in \Theta_e$ and $(\theta'_1, \ldots, \theta'_k, \eta') \in \Theta_e$, if

$$\sum_{i=1}^k \alpha_i g_{\theta_i,\eta} = \sum_{i=1}^k \alpha'_i g_{\theta'_i,\eta'}$$

  then,

$$\sum_{i=1}^k \alpha_i \delta_{\theta_i} = \sum_{i=1}^k \alpha'_i \delta_{\theta'_i} \text{ and } \eta = \eta' \ .$$

- **(A4)**. For any $y \in \mathbb{Y}$, $\theta_e \longmapsto g_{\theta_e}(y) = (g_{\theta_{e,i},\eta}(y))_{i\in\mathbb{X}}$ is continuous and tends to zero when $||\theta_e||$ tends to infinity.
- **(A5)**. For any $i \in \mathbb{X}$, $E_{\theta^\star}\Big[|\log g_{\theta^\star_{e,i},\eta^\star}(Y_1)|\Big] < \infty$.

– **(A6)**. For any $\theta_e \in \Theta_e$, there exists $\delta > 0$ such that : $E_{\theta^\star}\left[\sup_{||\theta'_e - \theta_e|| < \delta} (\log g_{\theta'_e}(Y_1))^+\right] < \infty$.

Theorem 1. *Assume that* **(A1)** *to* **(A6)** *hold, and that moreover there exists a positive real number* b *such that*

$$\sup_{\theta_e \in \Theta_e} \sup_{x_{1:n}} \left[\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq b \log n \quad , \tag{11.6}$$

$\mathbb{P}_{\theta^\star}$ - *eventually almost surely. If one chooses* $\alpha > 2(b+1)$ *in the penalty (11.5), then* $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ - *eventually almost surely.*

Notice that, to apply this theorem, one has to find a sequence of priors π_e^n on Θ_e such that (11.6) holds. The remaining of the section will prove that it is possible for situations in which priors may be defined as in previous works about HMM order estimation, while in the next section, we will prove that it is possible to find a prior in the important case of Gaussian emissions with unknown variance. In the following proof, the assumption (11.6) insures that $|\hat{\tau}_n| \leq |\tau^\star|$ eventually almost surely, while assumptions **(A1-6)** insure that for any complete tree τ such that $|\tau| < |\tau^\star|$ or $|\tau| = |\tau^\star|$ and $\tau \nsim \tau^\star$, $\hat{\tau}_n \neq \tau^\star$ $\mathbb{P}_{\theta^\star}$ - eventually almost surely. In particular **(A2)** holds whenever $\theta^\star_{e,x} \neq \theta^\star_{e,y}$ if $(x,y) \in \mathbb{X}^2$ and $x \neq y$.

Proof. The proof will be structured as follow : we first prove that $\mathbb{P}_{\theta^\star}$ - eventually almost surely, $|\hat{\tau}_n| \leq |\tau^\star|$. We then prove that for any complete tree τ such that $|\tau| \leq |\tau^\star|$ and $\tau \nsim \tau^\star$, $\hat{\tau}_n \nsim \tau$ $\mathbb{P}_{\theta^\star}$ - eventually almost surely. This will end the proof since there is a finite number of such trees. For any $n \in \mathbb{N}$, we denote by E_n the event

$$E_n \ : \ \left[\sup_{\theta_e \in \Theta_e} \sup_{x_{1:n}} \left(\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right) \leq b \log n \quad \right] .$$

By using (11.6) and Borel-Cantelli Lemma, to get that $\mathbb{P}_{\theta^\star}$ - eventually almost surely, $|\hat{\tau}_n| \leq |\tau^\star|$, it is enough to show that

$$\sum_{n=1}^{\infty} \mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|) \bigcap E_n\right\} < \infty .$$

Let τ be a complete tree such that $|\tau| > |\tau^\star|$. Using Proposition 2,

$$\begin{aligned}
&\mathbb{P}_{\theta^\star}\left\{(\hat{\tau}_n = \tau)\bigcap E_n\right\} \\
&\quad\leq \mathbb{P}_{\theta^\star}\left\{\left(\sup_{\theta\in\Theta_\tau}\log g_\theta(Y_{1:n}) - pen_\alpha(n,\tau) \geq \log g_{\theta^\star}(Y_{1:n}) - pen_\alpha(n,\tau^\star)\right)\bigcap E_n\right\}, \\
&\quad\leq \mathbb{P}_{\theta^\star}\left\{\left(\log \mathbb{KT}^n_\tau(y_{1:n}) + \sup_{\theta_e\in\Theta_e}\sup_{x_{1:n}}\left[\log\prod_{i=1}^n g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}^n_e(Y_{1:n}|x_{1:n})\right]\right.\right. \\
&\qquad\left.\left. + \frac{k-1}{2}|\tau|\log n + D - \log g_{\theta^\star}(Y_{1:n}) + pen_\alpha(n,\tau^\star) - pen_\alpha(n,\tau) \geq 0\right)\bigcap E_n\right\}, \\
&\quad\leq \mathbb{P}_{\theta^\star}\left\{g_{\theta^\star}(Y_{1:n}) \leq \mathbb{KT}^n_\tau(Y_{1:n})\right\}\exp\left(e_{\tau,n}\right) \ ,
\end{aligned}$$

with

$$e_{\tau,n,} = \frac{k-1}{2}|\tau|\log n + b\log n + D + pen_\alpha(n,\tau^\star) - pen_\alpha(n,\tau) \ .$$

But

$$\begin{aligned}
e_{\tau,n} &= \frac{k-1}{2}|\tau|\log n + b\log n + D + \sum_{t=1}^{|\tau^\star|}\frac{(k-1)t+\alpha}{2}\log n - \sum_{t=1}^{|\tau|}\frac{(k-1)t+\alpha}{2}\log n \\
&= \frac{k-1}{2}|\tau|\log n + b\log n + D - \sum_{t=|\tau^\star|+1}^{|\tau|}\frac{(k-1)t+\alpha}{2}\log n \\
&\leq -\frac{\alpha}{2}\left(|\tau| - |\tau^\star|\right)\log n + b\log n + D \ ,
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{P}_{\theta^\star}\left\{(\hat{\tau}_n = \tau)\bigcap E_n\right\} &\leq e^{-\frac{\alpha}{2}\left(|\tau|-|\tau^\star|\right)\log n + b\log n + D} \\
&= C.n^{-\frac{\alpha}{2}(|\tau|-|\tau^\star|)+b} \ ,
\end{aligned}$$

for some constant C. Thus

$$\mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|)\bigcap E_n\right\} \leq C\sum_{t=|\tau^\star|+1}^{\infty} CT(t)n^{-\frac{\alpha}{2}(t-|\tau^\star|)+b} \ ,$$

where $CT(t)$ is the number of complete trees with t leaves. But using Lemma 2 in Garivier [2006], $CT(t) \leq 16^t$ so that

$$\begin{aligned}
\mathbb{P}_{\theta^\star}\left\{(|\hat{\tau}_n| > |\tau^\star|)\bigcap E_n\right\} &\leq Cn^b 16^{|\tau^\star|}\sum_{t=1}^{\infty}\left[16n^{-\alpha/2}\right]^t \\
&= O(n^{-\alpha/2+b}) \ ,
\end{aligned}$$

which is summable if $\alpha > 2(b+1)$.

Let now τ be a tree such that $|\tau| \leq |\tau^\star|$ and $\tau \nsim \tau^\star$. Let τ_M be a complete tree such that τ and $\tau^\star$ are both a subtree of τ_M. Then, by setting for any integer $n \geq d(\tau_M) - 1$, $W_n = [X_{n-d(\tau_M)+1:n}]$, for any $\theta \in \Theta_\tau \cup \Theta_{\tau^\star}$, $(W_n, Y_n)_{n\in\mathbb{Z}}$ is a HMM under $\mathbb{P}_\theta$. Following the proof of Theorem 3 of Leroux [1992], we obtain that there exists $K > 0$ such that $\mathbb{P}_{\theta^\star}$-eventually a.s.,

$$\frac{1}{n}\log g_{\theta^\star}(Y_{1:n}) - \sup_{\theta\in\Theta_\tau} \frac{1}{n}\log g_\theta(Y_{1:n}) \geq K \ ,$$

so that

$$\log g_{\theta^\star}(Y_{1:n}) - pen(n,\tau^\star) - \sup_{\theta\in\Theta_\tau} \log g_\theta(Y_{1:n}) + pen(n,\tau) > 0, \mathbb{P}_{\theta^\star}\text{-eventually a.s.},$$

which finishes the proof of Theorem 1. □

11.3.3 Gaussian emissions with known variance

Here, we do not need the parameter η so we omit it. Then $\Theta_e = \{\theta_e = (m_1, \ldots, m_k) \in \mathbb{R}^k\}$. The conditional likelihood is given, for any $\theta_e = (m_x)_{x\in\mathbb{X}}$ by

$$g_{\theta_{e,x}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-m_x)^2}{2\sigma^2}) \ .$$

Proposition 3. *Assume* **(A1-2)**. *If one chooses $\alpha > k+2$ in the penalty (11.5), $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ - eventually a.s.*

Proof. The identifiability of the Gaussian model **(A3)** has been proved by Yakowitz and Spragins in Yakowitz and Spragins [1968], it is easy to see that Assumptions **(A4)** to **(A6)** hold. Now, we define the prior measure π_e^n on Θ_e as the probability distribution under which $\theta_e = (m_1, ..., m_k)$ is a vector of k independent random variables with centered Gaussian distribution with variance τ_n^2. Then, using Chambaz et al. [2009], $\mathbb{P}_{\theta^\star}$ -eventually a.s.,

$$\sup_{\theta_e\in\Theta_e} \max_{x_{1:n}\in\mathbb{X}^n} \left[\log \prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq \frac{k}{2}\log(1 + \frac{n\tau_n^2}{k\sigma^2}) + \frac{k}{2\tau_n^2}5\sigma^2 \log n \ .$$

Thus, by choosing $\tau_n^2 = \dfrac{5\sigma^2 k \log(n)}{2}$, we get that for any $\epsilon > 0$,

$$\sup_{\theta_e\in\Theta_e} \max_{x_{1:n}\in\mathbb{X}^n} \left[\log \prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n})\right] \leq \frac{k+\epsilon}{2}\log n \ ,$$

$\mathbb{P}_{\theta^\star}$ -eventually almost surely, and (11.6) holds for any $b > \dfrac{k}{2}$. □

11.3.4 Poisson emissions

Now the conditional distribution of Y given $X = x$ is Poisson with mean m_x and

$$\Theta_e = \left\{\theta_e = (m_1, ..., m_k) \middle| \ \forall j \in \mathbb{X},\ m_j > 0\right\} \ .$$

Proposition 4. *Assume* **(A1-2)**. *If one chooses* $\alpha > k + 2$ *in in the penalty (11.5),* $\hat{\tau}_n \sim \tau^\star$ $\mathbb{P}$ *-eventually a.s.*

Proof. The identifiability of the Gaussian model **(A3)** has been proved by Teicher in Teicher [1961], it is easy to see that Assumptions **(A4)** to **(A6)** hold. The prior π_e^n on Θ_e is now defined such that $m_1, ..., m_k$ are independent identically distributed with distribution Gamma$(t,\ 1/2)$. Then, using Chambaz et al. [2009]:

$$\sup_{\theta_e \in \Theta_e} \max_{x_{1:n} \in \mathbb{X}^n} \left\{ \log \prod_{i=1}^n g_{\theta_{e,x_i}}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \right\} \leq \frac{k}{2} \log \frac{n}{k} + kt \frac{\log n}{\sqrt{\log \log n}} + \frac{k}{2}(1 + t \log t) \ ,$$

$\mathbb{P}_{\theta^\star}$ -eventually a.s.. Then, for any fixed $t > 0$, for any $\epsilon > 0$, eventually almost surely :

$$\sup_{\theta_e = (m_1,...,m_k) \in \Theta_e} \max_{x_{1:n} \in \mathbb{X}^n} \left\{ \log g_{\theta_e}(Y_{1:n}|x_{1:n}) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \right\} \leq \left(\frac{k}{2} + \epsilon \right) \log n \ ,$$

$\mathbb{P}_{\theta^\star}$ -eventually almost surely, and (11.6) holds for any $b > \dfrac{k}{2}$. □

11.4 Gaussian emissions with unknown variance

We consider the situation where the emission distributions are Gaussian with the same, but unknown, variance $\sigma_\star^2$ and with a mean depending on the hidden state x. Let $\eta = -\dfrac{1}{2\sigma^2}$ and $\theta_{e,j} = \dfrac{m_j}{\sigma^2}$ for all $j \in \mathbb{X} = \{1, .., k\}$. Here

$$\Theta_e = \left\{ \left(\eta, \left(\theta_{e,j} \right)_{j=1,...,k} \right) \ \middle| \ \theta_{e,j} \in \mathbb{R}, \eta < 0 \right\} \ .$$

If $x_{1:n} \in \mathbb{X}^n$, for any $j \in \mathbb{X}$, we set $I_j = \{i | x_i = j\}$ and $n_j = |I_j|$. For sake of simplicity we omit $x_{1:n}$ in the notation though I_j and n_j depend on $x_{1:n}$. The conditional likelihood is given, for any $x_{1:n}$ in $\mathbb{X}^n$, for any $y_{1:n}$ in $\mathbb{Y}^n$, by

$$\prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(y_i\right) = \frac{1}{\sqrt{2\pi}^n} \prod_{j=1}^k \exp\left[\eta \sum_{i \in I_j} y_i^2 + \theta_{e,j} \sum_{i \in I_j} y_i - n_j A\left(\eta, \theta_{e,j}\right) \right] \ ,$$

where

$$A\left(\eta, \theta_{e,j}\right) = -\frac{\theta_{e,j}^2}{4\eta} - \frac{1}{2} \log(-2\eta) \ .$$

Theorem 2. *Assume* **(A1-2)**. *If one chooses* $\alpha > k+3$ *in the penalty (11.5), then* $\hat{\tau}_n \sim \tau^\star$, $\mathbb{P}_{\theta^\star}$ - *eventually a.s.*

Proof. We shall prove that Theorem 1 applies. First, it is easy to see that Assumptions **(A4)** to **(A6)** hold and the proof of **(A3)** can be found in Yakowitz and Spragins [1968].
Define now the conjugate exponential prior on Θ_e :

$$\pi_e^n(d\theta_e) = \exp\Bigg[\alpha_1^n \eta + \sum_{j=1}^k \alpha_{2,j}^n \theta_{e,j} - \sum_{j=1}^k \beta_j^n A\left(\eta, \theta_{e,j}\right) - B\left(\alpha_1^n, \alpha_{2,1}^n, \ldots, \alpha_{2,k}^n, \beta_1^n, \ldots, \beta_k^n\right)\Bigg] d\eta d\theta_{e,1} \cdots d\theta_{e,k} \ ,$$

where the parameters α_1^n, $(\alpha_{2,j}^n)_{j=1,\ldots,k}$ and $(\beta_j^n)_{j=1,\ldots,k}$ will be chosen later, and the normalizing constant may be computed as

$$\exp\left\{B\left(\alpha_1^n, \alpha_{2,1}^n, \ldots, \alpha_{2,k}^n, \beta_1^n, \ldots, \beta_k^n\right)\right\} = \frac{2^{k+\frac{\sum_{j=1}^k \beta_j^n}{2}} \pi^{\frac{k}{2}} \Gamma\left(\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}\right)}{\left(\prod_{j=1}^k \sqrt{\beta_j^n}\right)\left(\alpha_1^n - \sum_{j=1}^k \frac{(\alpha_{2,j}^n)^2}{\beta_j^n}\right)^{\frac{\sum_{j=1}^k \beta_j^n + k + 2}{2}}} \ ,$$

where we recall the Gamma function: $\Gamma(z) = \int_0^{+\infty} u^{z-1} e^{-u} du$ for any complex number z. Theorem 2 follows now from Theorem 1 and the proposition below. □

Proposition 5. *If* **(A1)** *holds, it is possible to choose the parameters* α_1^n, $(\alpha_{2,j}^n)_{j=1,\ldots,k}$ *and* $(\beta_j^n)_{j=1,\ldots,k}$ *such that for any* $\epsilon > 0$,

$$\max_{x_{1:n}} \left\{ \sup_{\theta_e \in \Theta_e} \log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \right\} \leq \frac{k+1+\epsilon}{2} \log n \ ,$$

$\mathbb{P}_{\theta^\star}$ - *eventually a.s.*.

Proof. For any $x_{1:n} \in \mathbb{X}^n$, the parameters $\left(\widehat{\eta}, (\widehat{\theta}_{e,j})_j\right)$ maximizing the conditional likelihood are given by

$$\widehat{\eta} = -\frac{1}{2\widehat{\sigma}_{x_{1:n}}^2}, \ \widehat{\theta}_{e,j} = \frac{\widehat{m}_{x_{1:n},j}}{\widehat{\sigma}_{x_{1:n}}^2} \ ,$$

with

$$\widehat{m}_{x_{1:n},j} = \frac{\sum_{i \in I_j} Y_i}{n_j}, \ \widehat{\sigma}_{x_{1:n}}^2 = \frac{1}{n} \sum_{j=1}^k \sum_{i \in I_j} \left(Y_i - \hat{m}_{x_{1:n},j}\right)^2 \ .$$

so that

$$\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}\left(Y_i\right) \leq -n \log \widehat{\sigma}_{x_{1:n}} - \frac{n}{2} \log 2\pi - \frac{n}{2} \ .$$

Also,

$$\mathbb{KT}_e^n(y_{1:n}|x_{1:n}) = \frac{1}{\sqrt{2\pi}^n} \exp\left[B\left(\alpha_1^n + \sum_{i=1}^n Y_i^2, (\alpha_{2,j}^n + \sum_{i\in I_j} Y_i)_{1\le j\le k}, (\beta_j^n + n_j)_{1\le j=1\le k}\right) - B\left(\alpha_1^n, (\alpha_{2,j}^n)_{1\le j\le k}, (\beta_j^n)_{1\le j\le k}\right)\right].$$

Recall that for all $z > 0$ (see for instance Whittaker and Watson [1996])

$$\sqrt{2\pi}e^{-z}z^{z-\frac{1}{2}} \le \Gamma(z) \le \sqrt{2\pi}e^{-z+\frac{1}{12z}}z^{z-\frac{1}{2}}$$

so that one gets that, for any $x_{1:n} \in \mathbb{X}^n$ and any $\theta_e \in \Theta_e$,

$$\begin{aligned}
&\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(y_{1:n}|x_{1:n}) \\
&\quad \le o(\log n) - \frac{n}{2}\log \widehat{\sigma}_{x_{1:n}}^2 - \frac{n}{2}(1+\log 2) + \frac{k}{2}\log\left(\frac{n+\sum_{j=1}^k \beta_j^n}{k}\right) \\
&\quad\quad - \left[-\frac{n+\sum_{j=1}^k \beta_j^n + k + 2}{2} + \left(\frac{n+\sum_{j=1}^k \beta_j^n + k + 1}{2}\right)\log\frac{n+\sum_{j=1}^k \beta_j^n + k + 2}{2}\right] \\
&\quad\quad\quad + \frac{n+\sum_{j=1}^k \beta_j^n + k + 2}{2}\log\left(\alpha_1^n + \sum_{i=1}^n Y_i^2 - \sum_{j=1}^k \frac{\left(\alpha_{2,j}^n + \sum_{i\in I_j} Y_i\right)^2}{n_j + \beta_j^n}\right).
\end{aligned}$$

Choose now

$$\beta_j^n = \alpha_{2,j}^n = \frac{1}{n},\ \alpha_{2,j}^n = \sqrt{\beta_j^n},\ j = 1,\ldots,k,\ \alpha_1^n = k+1\,. \tag{11.7}$$

Then one easily gets that for any $x_{1:n} \in \mathbb{X}^n$ and any $\theta_e \in \Theta_e$,

$$\begin{aligned}
&\log \prod_{i=1}^n g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}_e^n(Y_{1:n}|x_{1:n}) \\
&\le o(\log n) + \frac{n+\sum_{j=1}^k \beta_j^n + k + 2}{2}\log\left(1 + \frac{1}{n\widehat{\sigma}_{x_{1:n}}^2}\left[k+1 \quad + \sum_{j=1}^k \left\{ \widehat{m}_{x_{1:n},j}^2\left(n_j - \frac{n_j^2}{n_j + 1/n}\right)\right.\right.\right. \\
&\quad\quad \left.\left.\left. - 2\frac{n_j}{n.n_j+1}\widehat{m}_{x_{1:n},j} - \frac{1}{n^2 n_j + n}\right\}\right]\right) + \frac{k+1}{2}\log n + \frac{k/n+k+2}{2}\log\widehat{\sigma}_{x_{1:n}}^2\,.
\end{aligned}$$

Let now $|Y|_{(n)} = \max_{1\le i\le n}|Y_i|$. Then for any $x_{1:n} \in \mathbb{X}^n$,

$$\widehat{\sigma}_{x_{1:n}}^2 \le |Y|_{(n)}^2 \text{ and } |\widehat{m}_{x_{1:n},j}| \le |Y|_{(n)},\ j = 1,\ldots,k.$$

Also, for any partition $(I_i, \ldots, I_k)$ of $\mathbb{R}$ in k intervals, define :

$$\widehat{\sigma}^2_{I_i,\ldots,I_k} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} \mathbf{1}_{Y_i \in I_j}\left(Y_i - \frac{\sum_{i'=1}^{n} \mathbf{1}_{Y_{i'}\in I_j} Y_{i'}}{\sum_{i'=1}^{n} \mathbf{1}_{Y_{i'}\in I_j}}\right)^2 ,$$

and

$$\sigma^2_{I_i,\ldots,I_k} = \sum_{j=1}^{k} \mathbb{P}_{\theta^\star}(Y_1 \in I_k) Var_{\theta^\star}(Y_1 | Y_1 \in I_k) ,$$

where $Var_{\theta^\star}(Y_1|Y_1 \in I_k)$ is the conditional variance of Y_1 given that $Y_1 \in I_k$. The *k-means* algorithm, see McQueen [1967], Inaba and Imai [1994], allows to find a local minimum of the function $x_{1:n} \longrightarrow \widehat{\sigma}^2_{x_{1:n}}$ starting with any initial configuration $x_{1:n}$. Each step of the algorithm produces an assignment of the values $Y_{1:n}$ in k clusters (by partitioning the observations according to the Voronoï diagram generated by the means of each cluster). Here, the values $Y_{1:n}$ being real numbers, a Voronoï diagram clustering on $\mathbb{R}$ is nothing else than a clustering by intervals. Because the *k-means* algorithm converges, in a finite time, to a local minimum of the quantity $x_{1:n} \longrightarrow \widehat{\sigma}^2_{x_{1:n}}$, if the initial configuration is the $x^0_{1:n}$ that minimizes $\widehat{\sigma}^2_{x_{1:n}}$, the $k-means$ algorithm will lead to the same configuration $x^0_{1:n}$. Thus, the minimum of $\widehat{\sigma}^2_{x_{1:n}}$ is a clustering by intervals, that is

$$\inf_{x_{1:n}\in\mathbb{X}^n} \widehat{\sigma}^2_{x_{1:n}} = \inf_{I_i,\ldots,I_k} \widehat{\sigma}^2_{I_i,\ldots,I_k} ,$$

where the infimum is over all partitions of $\mathbb{R}$ in k intervals.
We now get:

$$\begin{aligned}
&\log \prod_{i=1}^{n} g_{\theta_{e,x_i},\eta}(Y_i) - \log \mathbb{KT}^n_e(Y_{1:n}|x_{1:n}) \\
&\qquad \leq o(\log n) + \frac{n + \sum_{j=1}^{k}\beta^n_j + k + 2}{2} \log\left(1 + \frac{1}{n \inf_{I_i,\ldots,I_k} \widehat{\sigma}^2_{I_i,\ldots,I_k}}\left[k+1\right.\right. \\
&\qquad\qquad \left.\left. + \sum_{j=1}^{k}\left\{|Y|^2_{(n)}\left(n_j - \frac{n_j^2}{n_j + 1/n}\right) + 2\frac{n_j}{n.n_j+1}|Y|(n)\right\}\right]\right) \\
&\qquad\qquad + \frac{k+1}{2}\log n + \frac{k/n + k + 2}{2}\log |Y|^2_{(n)} ,
\end{aligned}$$

and Proposition 5 follows from the choice (11.7) and the lemmas below, whose proofs are given in the Appendix. □

Lemma 1. *If* **(A1)** *holds,*
$\sup_{I_i,\ldots,I_k} \left|\widehat{\sigma}^2_{I_i,\ldots,I_k} - \sigma^2_{I_i,\ldots,I_k}\right|$ *converges to* 0 *as* n *tends to infinity* $\mathbb{P}_{\theta^\star}$ *- a.s. (Here the supremum is over all partitions of* $\mathbb{R}$ *in* k *intervals). Also, the infimum* $s_{\inf}$ *of* $\sigma^2_{I_i,\ldots,I_k}$ *over all partitions of* $\mathbb{R}$ *in* k *intervals satisfies* $s_{\inf} > 0$.

Lemma 2. *If* **(A1)** *holds,* $\mathbb{P}_{\theta^\star}$ *- eventually a.s. ,* $|Y|^2_{(n)} \leq 5\sigma_\star^2 \log n$.

11.5 Algorithm and simulations

In this section we first present our practical algorithm. We then apply it in the case of Gaussian emissions with unknown common variance and compare our estimator with the BIC estimator that is when we choose in (11.4) the BIC penalty $pen(n,\tau) = \frac{k-1}{2}|\tau| \log n$.

11.5.1 Algorithm

We start this section with the definition of the terms used below :

- A maximal node of a complete tree τ is a string u such that, for any x in $\mathbb{X}$, ux belongs to τ. We denote by $N(\tau)$ the set of maximal nodes in the tree τ.
- The score of a complete tree τ on the basis of the observation $(Y_1, \ldots, Y_n)$ is the penalized maximum likelihood associated with τ :

$$sc(\tau) = -\sup_{\theta \in \Theta_\tau} \log g_\theta(Y_{1:n}) + pen(n,\tau) \; . \tag{11.8}$$

We also require that the emission model belongs to an exponential family such that :

(i) There exists $D \in \mathbb{N}^\star$, a function $s : \mathbb{X} \times \mathbb{Y} \longrightarrow \mathbb{R}^D$ of sufficient statistic and functions $h : \mathbb{X} \times \mathbb{Y} \longrightarrow \mathbb{R}$, $\psi : \Theta_e \longrightarrow \mathbb{R}^D$, and $A : \Theta_e \longrightarrow \mathbb{R}$, such that the emission density can be written as :

$$g_{\theta_e,x,\eta}(y) = h(x,y) \exp\left[\langle \psi(\theta_e), s(x,y) \rangle - A(\theta_e)\right] \; ,$$

where $\langle ., . \rangle$ denotes the scalar product in $\mathbb{R}^D$.

(ii) For all $S \in \mathbb{R}^D$, the equation :

$$\nabla_{\theta_e} \psi(\theta_e) S - \nabla_{\theta_e} A(\theta_e) = 0 \; ,$$

where ∇_{θ_e} denotes the gradient, has a unique solution denoted by $\bar{\theta}_e(S)$.

Assumption (ii) states that the function $\bar{\theta}_e : S \in \mathbb{R}^D \to \bar{\theta}_e(S) \in \Theta_e$ that returns the complete data maximum likelihood estimator corresponding to any feasible value of the sufficient statistics is available in closed-form.

The key idea of our algorithm is a "bottom to the top" pruning technique. Starting from the maximal complete tree of depth $M = \lfloor \log n \rfloor$, denoted by τ_M, we change each maximal node into a leaf whenever the resulting tree decreases the score.

We then need to compute the maximum likelihood of any complete tree subtree of τ_M. We start the algorithm by running several iterations of the EM algorithm. During this preliminary step we build estimators of sufficient statistics. These statistics will be used later in the computation of the maximum likelihood estimator $\hat{\theta}_\tau \in \Theta_\tau$ which realizes the supremum in (11.8) for any complete context tree τ subtree of τ_M.

For any $n \geq 0$, we denote by W_n the vectorial random sequence $W_n = (X_{n-M+1}, \ldots, X_n)$. For n big enough, $M \geq d(\tau^\star)$ and $(W_n)_n$ is a Markov chain. The intermediate quantity (see Cappé et al. [2005]) needed in the EM algorithm for the HMM (W_n, Y_n) can be written as:
for any (θ, θ') in Θ_{τ_M} :

$$\begin{aligned} Q_{\theta,\theta'} &= E_{\theta'}(\log(g_\theta(W_{1:n}, Y_{1:n}))\big|Y_{1:n}) \\ &= E_{\theta'}(\nu(W_1)\big|Y_{1:n}) + \sum_{i=1}^{n-1} E_{\theta'}(\log P_{\theta_t}(W_i, W_{i+1})\big|Y_{1:n}) \\ &+ \sum_{i=1}^{n} E_{\theta'}(\log g_{\theta_{e,W_{i,M}},\eta}(Y_i)\big|Y_{1:n}) \ . \end{aligned}$$

Notice, for any $\theta \in \Theta_{\tau_M}$, if $(w, w') \in (\mathbb{X}^M)^2$ are such that $w_{2:M} \neq w'_{1:M-1}$, then $P_{\theta_t}(w, w') = 0$.
For any $w \in \mathbb{X}^M$ and any $w' \in \mathbb{X}^M$ if we denote by

$$\begin{aligned} &\forall\ i = 1, \ldots, n,\ \Phi^{\theta'}_{i|n}(w) = P_{\theta'}(W_i = w | Y_{1:n}) \ , \\ &\forall\ i = 1, \ldots, n-1,\ \Phi^{\theta'}_{i:i+1|n}(w, w') = P_{\theta'}(W_i = w, W_{i+1} = w' | Y_{1:n}) \ , \end{aligned}$$

and

$$S^{\theta'}_{t,n} = \left(\frac{\left(\sum_{i=1}^{n-1} \Phi^{\theta'}_{i:i+1|n}(w, w') \right)}{n} \right)_{(w,w') \in \mathbb{X}^M} ,$$

$$S^{\theta'}_{e,n} = \frac{1}{n} \sum_{x \in \mathbb{X}} \sum_{i=1}^{n} \left(\sum_{w \in \mathbb{X}^M | w_M = x} \Phi^{\theta'}_{i|n}(w) \right) s(x, Y_i) \ ,$$

then there exists a function C such that :

$$\frac{1}{n} Q_{\theta,\theta'} = \frac{1}{n} C(\theta', Y_{1:n}) + \left\langle S^{\theta'}_{t,n}, \log P_{\theta_t} \right\rangle + \left\langle S^{\theta'}_{e,n}, \psi(\theta_e) \right\rangle - A(\theta_e) \ . \tag{11.9}$$

If, for some complete tree τ, we restrict θ_t in $\Theta_{t,\tau}$, then for any s in τ, for any w in $\mathbb{X}^M$ such that s is postfix of w, for any x in $\mathbb{X}$, we have $P_{\theta_t}(w, (w_{2:M} x)) = P_{s,x}(\theta_t)$.

Thus, the vector $P_{s,\cdot}$ maximising this equation is solution of the Lagrangian,

$$\begin{cases} \dfrac{\delta}{\delta P_{s,x}} & \left[\dfrac{1}{n} Q_{\theta,\theta'} + \Lambda(\sum\limits_{x' \in \mathbb{X}} P_{s,x'} - 1) \right] = 0 \ , \ \forall x \in \mathbb{X} \ , \\ \dfrac{\delta}{\delta \Lambda} & \left[\dfrac{1}{n} Q_{\theta,\theta'} + \Lambda(\sum\limits_{x' \in \mathbb{X}} P_{s,x'} - 1) \right] = 0 \ . \end{cases}$$

and, finally, the estimator of $\theta_t \in \Theta_{t,\tau}$ maximising the quantity $Q(\theta', .)$ only depends on the sufficient statistic $S^{\theta'}_{t,n}$ and is given by :

$$\bar{P}_{s,x}(S^{\theta'}_{t,n}) = \frac{\sum_{w \in \mathbb{X}^M |\ s \text{ postfix } of\ w} S^{\theta'}_{t,n}(w,(w_{2:M}x))}{\sum_{x' \in \mathbb{X}} \sum_{w \in \mathbb{X}^M |\ s \text{ postfix } of\ w} S^{\theta'}_{t,n}(w,(w_{2:M}x'))} . \tag{11.10}$$

Algorithm 7 Preliminary computation of the sufficient statistics

Require: $\theta_0 = (\theta_{t,0}, \theta_{e,0}) \in \Theta_{\tau_M}$ be an initial value for the parameter θ.
Require: Let t_{EM} be a threshold.
1: $stop = 0$
2: $i = 0$
3: **while** $(stop = 0)$ **do**
4: $i = i + 1$
5: M step : compute the quantities $S^{\theta_{i-1}}_{t,n}$ and $S^{\theta_{i-1}}_{e,n}$
6: E step : set

$$\theta_i = \left(\left(\bar{P}_{w,x}(S^{\theta_{i-1}}_{t,n}) \right)_{w,x} , \bar{\theta}_e(S^{\theta_{i-1}}_{e,n}) \right)$$

7: **if** $(||\theta_i - \theta_{i-1}|| < t_{Em})$ **then**
8: $stop = 1$
9: **end if**
10: **end while**
11: M step : compute the quantities $S^{\theta_i}_{t,n}$ and $S^{\theta_i}_{e,n}$
12: $S_t = S^{\theta_i}_{t,n}$ and $S_e = S^{\theta_i}_{e,n}$
13: **return** (S_t, S_e)

While Algorithm 7 computes the sufficient statistics S_t and S_e on the basis of the observations $(Y_k)_{k\in\{1,\dots,n\}}$, Algorithm 8 is our pruning Algorithm. This algorithm begins with the estimation of the exhaustive statistics calling Algorithm 7. As Algorithm 7 is prone to the convergence towards local maxima, we set our initial parameter value θ_0 after running a preliminary *k-means* algorithm (see McQueen [1967], Inaba and Imai [1994]): we assign the values $Y_{1:n}$ into k clusters which produces a sequence of "clusters" $\tilde{X}_{1:n}$. A first estimation of the emission parameters is then possible using this clustering, the initial transition parameter $\theta_{0,t} = \left(P^0_{w,i}\right)_{w\in\mathbb{X}^M,\ i\in\mathbb{X}}$ is also computed on the basis of the sequence $\tilde{X}_{1:n}$ using the relation :

$$\forall w \in \mathbb{X}^M,\ \forall x \in \mathbb{X}, P^0_{w,x} = \frac{\sum_{i=1}^{n-M} \mathbf{1}_{\tilde{X}_{i:i+M-1}=w} \mathbf{1}_{\tilde{X}_{i+M}=x}}{\sum_{i=1}^{n-M} \mathbf{1}_{\tilde{X}_{i:i+M-1}=w}} .$$

Then, starting with the initialisation $\tau = \tau_M$, we consider, one after the other, the maximal nodes u of τ. We build a new tree τ_{test} by taking out of τ all the contexts s having u as postfix and adding u as a

new context: $\tau_{test} = \tau \setminus \{ux | ux \in \tau,\ x \in \mathbb{X}\} \bigcup \{u\}$. Let $\hat{\theta}_{\text{test}} = \left(((\bar{P}_{s,x}(S_t))_{s\in\tau_{\text{test}},x\in\mathbb{X}}, \bar{\theta}_e(S_e)\right)$ which, hopefully, becomes an acceptable proxy for $\underset{\theta\in\Theta_{\tau_{test}}}{\text{argmax}} \log g_\theta(Y_{1:n})$. Let $-\log g_{\hat{\theta}_{\text{test}}}(Y_{1:n}) + pen(n, \tau_{\text{test}})$ be an approximation of the score of the context tree τ_{test} still denoted by $sc(\tau_{\text{test}})$, then, if $sc(\tau_{\text{test}}) < sc(\tau)$, we set $\tau = \tau_{\text{test}}$. In Algorithm 8, the role of τ_2 is to insure that all the branches of τ are tested before shortening again a branch already tested.

Algorithm 8 Bottom to the top pruning algorithm

Require: Let t_{EM} a threshold.
Compute (S_t, S_e) with Algorithm 7 with the t_{EM} threshold.
$\hat{\theta} = \left(\left(\bar{P}_{w,x}(S_t)\right)_{w\in\tau_M,x\in\mathbb{X}}, \bar{\theta}_e(S_e)\right)$
Pruning procedure :
$\tau = \tau_2 = \tau_M$
$change = YES$
while ($change = YES$ AND $|\tau| \geq 1$) **do**
  $change = NO$
  for ($u \in N(\tau)$) **do**
    if ($u \in N(\tau_2)$) **then**
      $L_u(\tau_2) = \{s \in \tau_2 | u\ postfix\ of\ s\}$
      $\tau_{\text{test}} = [\tau_2 \setminus L_u(\tau_2)] \bigcup \{u\}$
      $\hat{\theta}_{\text{test}} = \left(\left(\bar{P}_{s,x}(S_t)\right)_{s\in\tau,x\in\mathbb{X}}, \bar{\theta}_e(S_e)\right)$
      if ($sc(\tau_{\text{test}}) < sc(\tau_2)$) **then**
        $\tau_2 = \tau_{\text{test}}$
        $\hat{\theta} = \hat{\theta}_{\text{test}}$
        $change = YES$
      end if
    end if
  end for
  $\tau = \tau_2$
end while
return τ

11.5.2 Simulations

We propose to illustrate the a.s convergence of $\hat{\tau}_n$ using Algorithm 8 in the case of Gaussian emission with unknown variance. We set $k = 2$, and use as minimal complete context tree one of the two complete trees represented in Figure 11.1 and Figure 11.2. The true transitions probabilities associated with each trees are indicated in boxes under each context.

For each tree $\tau_1^\star$ and $\tau_2^\star$, we will simulate 3 samples of the VLHMM, choosing as true emission parameters $m_0^\star = 0$, $\sigma^{2,\star} = 1$ and $m_1^\star$ varying in $\{2, 3, 4\}$. In the preliminary EM steps, we use as threshold $t_{EM} = 0.001$

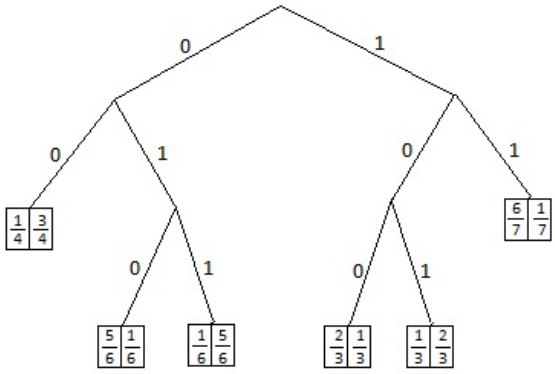


Figure 11.1: Graphic representation of the complete context tree $\tau_1^\star$ with transition probabilities indicated in the box under each leaf s: $P^\star_{s,0} \mid P^\star_{s,1}$

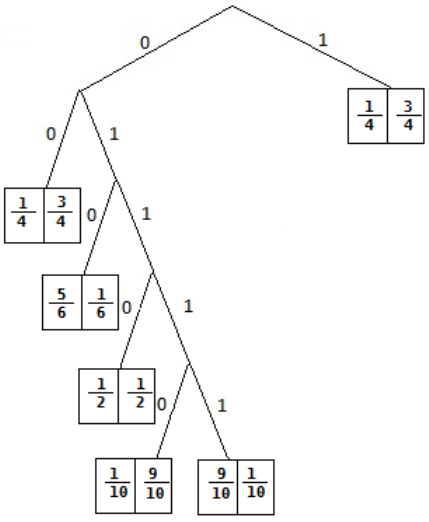


Figure 11.2: Graphic representation of the complete context tree $\tau_2^\star$ with transition probabilities indicated in the box under each leaf s: $P^\star_{s,0} \mid P^\star_{s,1}$.

The results of our simulations are summarized in Tables 11.1 to 11.4. The size of the estimated tree $|\hat{\tau}_n|$ for different values of n and $m_1^\star$ are noticed in Table 11.1 when $\tau^\star = \tau_1^\star$ (resp. in the table

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (11.5)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	2	2	2	2	3	3
1000	2	2	2	7	**6**	**6**
2000	2	2	4	**6**	**6**	**6**
5000	2	4	4	7	**6**	**6**
10000	4	**6**	**6**	7	**6**	**6**
20000	5	**6**	**6**	**6**	**6**	**6**
30000	5	**6**	**6**	**6**	**6**	**6**
40000	**6**	**6**	**6**	7	**6**	**6**
50000	**6**	**6**	**6**	7	**6**	**6**

Table 11.1: Case $\tau^\star = \tau_1^\star$. Comparison of $|\hat{\tau}_n|$ between our estimator and the BIC estimator for different values of n and $m_1^\star$.

$\tau^\star = \tau_1^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (11.5)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	-202	-202	-190	-6	-6	2
1000	-235	-213	-155	4	-2	25
2000	-221	-129	-88	8	-4	4
5000	-144	-36	-20	5	-4	-5
10000	-75	-5	-4	4	-5	-4
20000	-6	-4	-4	10	-4	-4
30000	21	-5	-4	10	-5	-4
40000	12	-4	-3	10	-4	-3
50000	12	-7	-4	10	-4	-4

Table 11.2: Case $\tau^\star = \tau_1^\star$. Score difference $sc(\hat{\tau}_n) - sc(\tau^\star)$.

Figure 11.3 when $\tau^\star = \tau_2^\star$) for the two choices of penalties $pen_\alpha(n,\tau) = \sum_{t=1}^{|\tau|} \frac{(k-1)t+\alpha}{2} \log n$ with $\alpha = 5.1$ and $pen(n,\tau) = \frac{k-1}{2}|\tau| \log n$. The first important remark we make regarding Tables 11.1 and 11.3 is that, on each simulation and whatever the penalty we used, when $|\hat{\tau}_n| = |\tau^\star|$ we also had $\hat{\tau}_n = \tau^\star$, in the same way, each time $|\hat{\tau}_n| < |\tau^\star|$ (resp. $|\hat{\tau}_n| > |\tau^\star|$), $\hat{\tau}_n$ was a subtree of $\tau^\star$ (resp. $\tau^\star$ was a subtree of $\hat{\tau}_n$). For any combination of $\tau^\star$ and $m_1^\star$, both estimators seem to converge, except our estimator in the case $\tau^\star = \tau_2^\star$ and $m_1^\star = 2$, where 50 000 measures is not enough to reach the convergence. However, for small samples, smaller models are systematically chosen with our estimator, while the BIC estimator is reaching the right model for relatively small samples. This

$\tau^\star = \tau_2^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (11.5)			BIC penalty		
$n/m_1^\star$	2	3	4	2	3	4
100	2	2	2	2	2	2
1000	2	2	2	3	**6**	**6**
2000	2	2	2	**6**	**6**	**6**
5000	2	3	3	**6**	**6**	**6**
10000	3	3	3	**6**	**6**	**6**
20000	3	3	**6**	**6**	**6**	**6**
30000	3	3	**6**	**6**	**6**	**6**
40000	3	**6**	**6**	**6**	**6**	**6**
50000	3	**6**	**6**	**6**	**6**	**6**

Table 11.3: Case $\tau^\star = \tau_2^\star$. Comparison of $|\hat{\tau}_n|$ between our estimator and the BIC estimator for different values of n and $m_1^\star$.

$\tau^\star = \tau_2^\star$, $\lvert\tau^\star\rvert = 6$						
	Penalty (11.5)			BIC penalty		
n/m_1	2	3	4	2	3	4
100	-201	-202	-195	-10	-6	1
1000	-266	-246	-229	5	-1	-2
2000	-272	-239	67	4	-1	324
5000	-272	-200	-151	2	-2	-5
10000	-242	-128	-52	6	-2	-4
20000	-227	12	-6	6	-6	-6
30000	-191	141	-6	7	-5	-6
40000	-159	-6	-8	8	-6	-8
50000	-136	-6	-9	7	-6	-8

Table 11.4: Case $\tau^\star = \tau_2^\star$. Score difference $sc(\hat{\tau}_n) - sc(\tau^\star)$.

behaviour of our estimator shows that our penalty is too heavy.

The score differences $sc(\hat{\tau}_n) - sc(\tau^\star)$ Table 11.2 when $\tau^\star = \tau_1^\star$ and Table 11.4 when $\tau^\star = \tau_2^\star$ are the differences between the score of $\hat{\tau}_n$ computed with the estimated parameter $\hat{\theta}_n$ and the score of $\tau^\star$ computed with the the real parameters. These informations allow us to know when the estimators $\hat{\tau}_n, \hat{\theta}_n$ are well estimated by Algorithm 8. Indeed, when $\hat{\tau}_n \neq \tau^\star$, if the score of $\tau^\star$ computed with the real transition and emission parameters is smaller than the score of our estimator with estimated parameters (non negative score difference), then the estimator given by Algorithm 8 is not the expected estimator defined by (11.4). In particular, Table 11.2 shows that the over estimation of the BIC estimator in the case $m_1^\star = 2$ (Table 11.2) can be due to a local minima problem: Algorithm 8 selected a tree τ such that $|\tau| > |\tau^\star|$ whereas $\tau^\star$ had a smaller score. This problem might occur

because we use an EM type algorithm which often leads to local minima. Although we try to take an initial value of the parameters in a neighbourhood of the real ones using the preliminary k-means algorithm, this problem persists. Extra EM loops for each tested tree in Algorithm 8 could also provide a better estimation of the parameters and then improve the score estimation for each tested tree, but it would also increase the complexity of the algorithm.

Finally, we observe that bigger the quantity $|m_0^\star - m_1^\star|$ is, quicker the convergence of our estimator or BIC estimator occurs. This phenomenon can be easily understood as very different emission distributions for different states leads to an easier estimation of the underlying state sequence on the basis of the observations and allows us to build a more precise description of the VLMC behaviour.

11.6 Conclusion

In this paper, we were interested in the statistical analysis of Variable Length Hidden Markov Models (VLHMM). We have presented such models then we estimated the context tree of the hidden process using penalized maximum likelihood. We have shown how to choose the penalty so that the estimator is strongly consistent without any prior upper bound on the depth or on the size of the context tree of the hidden process. We have proved that our general consistency theorem applies when the emission distributions are Gaussian with unknown means and the same unknown variance. We have proposed a pruning algorithm and have applied it to simulated data sets. This illustrates the consistency of our estimator, but also suggests that smaller penalty could lead to consistent estimation.

Finding the minimal penalty insuring the strong consistency of the estimator with no prior upper bound remains unsolved. A similar problem has been solved by R. van Handel van Handel [2011] to estimate the order of finite state Markov chains, and by E. Gassiat and R. van Handel Gassiat and van Handel [2010] to estimate the number of populations in a mixture with i.i.d. observations. The basic idea is that the maximum likelihood behaves as the maximum of approximate chi-square variables, and that the behavior of the maximum likelihood statistic may be investigated using empirical process theory tools to obtain a $\log\log n$ rate of growth. However, it is known for HMM that the maximum likelihood does not behave this way and converges weakly to infinity, see Gassiat and Keribin [2000]. We did by-pass the problem by using information theoretic inequalities, but understanding the pathwise fluctuations of the likelihood in HMM models remains a difficult problem to be solved.

11.7 Appendices

11.7.1 Proof of Lemma 1

For any partition $(I_i, \ldots, I_k)$ of $\mathbb{R}$ in k intervals,

$$\begin{aligned}\sigma^2_{I_i,\ldots,I_k} &= \sum_{j=1}^{k} \mathbb{P}_{\theta^\star}(Y_1 \in I_k) Var_{\theta^\star}(Y_1 | Y_1 \in I_k) \ , \\ &\geq \frac{1}{k} \inf_{I : \mathbb{P}_{\theta^\star}(Y \in I) \geq \frac{1}{k}} Var_{\theta^\star}(Y_1 | Y_1 \in I) \ .\end{aligned}$$

where the infimum is over all intervals I of $\mathbb{R}$. The distribution of Y_1 is the Gaussian mixture with density $g^\star = \sum_{x\in\mathbb{X}} \pi^\star(x)\phi_{m_x^\star,\sigma_\star^2}$, where $\pi^\star$ is the stationary distribution of $(X_n)_{n\geq 0}$ and $\phi_{m_x^\star,\sigma_\star^2}$ is the density of the normal distribution with mean $m_x^\star$ and variance $\sigma_\star^2$. The repartition function $F^\star$ of the distribution of Y_1 is continuous and increasing, with continuous and increasing inverse quantile function. Thus,

$$\inf_{I_i,\ldots,I_k} \sigma^2_{I_i,\ldots,I_k} \geq \inf_{\substack{-\infty\leq a<b\leq+\infty:\\ F^\star(a)+\frac{1}{k}\leq F^\star(b)}} Var_{\theta^\star}(Y_1|Y_1\in]a,b[) \ .$$

But $Var_{\theta^\star}(Y_1|Y_1\in]a,b[)$ is a continuous function of (a,b), and the infimum at the righ-hand side of the inequality is attained at some $(\overline{a},\overline{b})$ (eventually infinite) such that $F^\star(\overline{a})+\frac{1}{k}\leq F^\star(\overline{b})$. Thus $Var_{\theta^\star}(Y_1|Y_1\in]\overline{a},\overline{b}[)>0$, and $s_{inf}>0$.
For any partition $(I_i,\ldots,I_k)$ of $\mathbb{R}$ in k intervals,

$$\hat{\sigma}^2_{I_i,\ldots,I_k}(Y_{1:n}) - \sigma^2_{I_i,\ldots,I_k} = \frac{1}{n}\sum_{i=1}^n Y_i^2 - E(Y_1^2) - \sum_{j=1}^k \left(\frac{(\sum_{i=1}^n Y_i \mathbf{1}_{I_j}(Y_i))^2}{n^2} \frac{n}{\sum_{i=1}^n \mathbf{1}_{I_j}(Y_i)} - \frac{E(Y\mathbf{1}_{I_j}(Y_1))^2}{E(\mathbf{1}_{I_j}(Y_1))} \right) ,$$

so that

$$\sup_{I_1,\ldots,I_k} \left|\hat{\sigma}^2_{I_i,\ldots,I_k}(Y_{1:n}) - \sigma^2_{I_i,\ldots,I_k}\right|$$
$$\leq \frac{1}{n}\left|\sum_{i=1}^n Y_i^2 - E(Y_1^2)\right| + k \sup_{I \text{ interval of } \mathbb{R}} \left| \frac{(\sum_{i=1}^n Y_i \mathbf{1}_I(Y_i))^2}{n^2} \frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)} - \frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))} \right| .$$

Using Leroux [1992], $(Y_n)_{n\geq 0}$ is a stationary ergodic process, so that $\frac{1}{n}\sum_{i=1}^n Y_i^2 - E(Y_1^2)$ tends to 0 $\mathbb{P}_{\theta^\star}$ a.s. Let $\epsilon>0$. We now consider separately the intervals I such that $E(\mathbf{1}_I(Y))\leq\epsilon$ or $E(\mathbf{1}_I(Y))>\epsilon$.
• Let I be such that $E(\mathbf{1}_I(Y_1))\leq\epsilon$.
Using Cauchy Schwarz inequality,

$$\begin{aligned} \left(\frac{1}{n}\sum Y_i\mathbf{1}_I(Y_i)\right)^2 &\leq \left(\frac{1}{n}\sum Y_i^2\mathbf{1}_I(Y_i)\right)\times\left(\frac{1}{n}\sum \mathbf{1}_I(Y_i)\right) , \\ E\left(Y_1\mathbf{1}_I(Y_1)\right)^2 &\leq E\left(Y_1^2\mathbf{1}_I(Y_1)\right) E\left(\mathbf{1}_I(Y_1)\right) , \end{aligned}$$

and,

$$E\left(Y_1^2\mathbf{1}_I(Y_1)\right) \leq \sqrt{E(Y_1^4)}\sqrt{E(\mathbf{1}_I(Y_1))} \leq M\sqrt{\epsilon} ,$$

for some fixed positive constant M. Thus,

$$\begin{aligned}\left|\frac{(\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)} - \frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right| &\leq \frac{1}{n}\sum_{i=1}^n Y_i^2\mathbf{1}_I(Y_i) + E(Y_1^2\mathbf{1}(Y_1)) \ , \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n Y_i^2\mathbf{1}_I(Y_i) - E(Y_1^2\mathbf{1}(Y_1))\right| + 2E(Y_1^2\mathbf{1}(Y_1)) \ , \\ &\leq \left|\frac{1}{n}\sum_{i=1}^n Y_i^2\mathbf{1}_I(Y_i) - E(Y_1^2\mathbf{1}(Y_1))\right| + 2M\sqrt{\epsilon} \ .\end{aligned}$$

• Let now I be such that $E(\mathbf{1}_I(Y_1)) > \epsilon$.

$$\begin{aligned}&\left|\frac{(\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)} - \frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right| \\ &\qquad = \left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n}\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} - \frac{E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right| \\ &\qquad \times \left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n}\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} + \frac{E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right| , \\ &\qquad \leq \left[\left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n}\right|\left|\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} - \frac{1}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right. \\ &\qquad \left. + \left|\frac{\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right] \\ &\qquad \times \left[\left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n}\right|\left|\frac{1}{\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}}} + \frac{1}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right. \\ &\qquad \left. + \left|\frac{\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n} - E(Y\mathbf{1}_I(Y_1))}{\sqrt{E(\mathbf{1}_I(Y_1))}}\right|\right] \ ,\end{aligned}$$

and, finally,

$$\left|\frac{(\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)} - \frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))}\right|$$
$$\leq \left[\left(\frac{\sum_{i=1}^n |Y_i|}{n}\right)\frac{\left|\sqrt{\frac{\sum_{i=1}^n \mathbf{1}_I(Y_i)}{n}} - \sqrt{E(\mathbf{1}_I(Y_1))}\right|}{\epsilon}\right.$$
$$\left. + \frac{\left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))\right|}{\sqrt{\epsilon}}\right]$$
$$\times \left[\frac{2}{\epsilon}\frac{\sum_{i=1}^n |Y_i|}{n} + \frac{\left|\frac{\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i)}{n} - E(Y_1\mathbf{1}_I(Y_1))\right|}{\sqrt{\epsilon}}\right] .$$

Now, using Lemma 3 below, one gets that, for all positive ϵ,

$$\limsup_{n\to\infty} \sup_{I \text{ interval of } \mathbb{R}} \left|\frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))} - \frac{(\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)}\right| \leq 2M\sqrt{\epsilon} ,$$

$\mathbb{P}_{\theta^\star}$-a.s. so that,

$$\lim_{n\to\infty} \sup_{I \text{ interval of } \mathbb{R}} \left|\frac{E(Y_1\mathbf{1}_I(Y_1))^2}{E(\mathbf{1}_I(Y_1))} - \frac{(\sum_{i=1}^n Y_i\mathbf{1}_I(Y_i))^2}{n^2}\frac{n}{\sum_{i=1}^n \mathbf{1}_I(Y_i)}\right| = 0 ,$$

$\mathbb{P}_{\theta^\star}$-a.s. and the Lemma follows.

Lemma 3. $\sup_I \left|\frac{1}{n}\sum Y_i^2\mathbf{1}_I(Y_i) - E\left(Y_1^2\mathbf{1}_I(Y)\right)\right|$, $\sup_I \left|\frac{1}{n}\sum Y_i\mathbf{1}_I(Y_i) - E\left(Y_1\mathbf{1}_I(Y_1)\right)\right|$ *and* $\sup_I \left|\frac{1}{n}\sum \mathbf{1}_I(Y_i) - E\left(\mathbf{1}_I(Y_1)\right)\right|$ *(where the supremum is over all intervals I in $\mathbb{R}$) tend to* 0 *as n tends to infinity, $\mathbb{P}_{\theta^\star}$ a.s.*

Proof. Let us note $\mathcal{F}_a = \{x \to x^a\mathbf{1}_I(x) : I \text{ interval of } \mathbb{R}\}$ for $a = 0, 1, 2$. Since the sequence of random variables $(Y_n)_{n\geq 0}$ is stationary and ergodic, it is enough to prove that, for $a = 0, 1, 2$, for any positive ϵ, there exists a finite set of functions $\tilde{\mathcal{F}}_a$ such that for any $f \in \mathcal{F}_a$, there exists l, u in $\tilde{\mathcal{F}}_a$ such that $l \leq f \leq u$ and $E(u(Y_1) - l(Y_1)) \leq \epsilon$.
For the cases a=0 or 2 and for any positive ϵ, there exist real numbers : $L^1_{a,\epsilon}$ and $L^2_{a,\epsilon}$ such that $\int_{-\infty}^{L^1_{a,\epsilon}} x^a g^\star(x)dx \leq \epsilon$ and $\int_{L^2_{a,\epsilon}}^{+\infty} x^a g^\star(x)dx \leq \epsilon$, and there exists real numbers $x_{a,1} = L^1_\epsilon < x_{a,2} < .... < x_{a,N_{a,\epsilon}-2} < L^2_{a,\epsilon} = x_{a,N_{a,\epsilon}-1}$ such that $\int_{x_{a,i}}^{x_{a,i+1}} x^a g^\star(x)dx < \epsilon/2$, $i = 1, \ldots, N_{a,\epsilon} - 2$. Then we define

- $I^1_{N_\epsilon} = \mathbb{R}$,
- for any $i = 1, ..., N_{a,\epsilon}$, $I^1_{a,i} = [-\infty\ ,\ x_{a,i}]$
- and for any $i = 1, ..., N_{a,\epsilon}$, $I^2_{a,i} = [x_{a,i}\ ,\ \infty]$

so that if $\mathcal{I}_a$ is the set $\mathcal{I}_a = \left\{I^j_{a,i} | i = 1, ..., N_{a,\epsilon},\ j = 1, 2\right\} \bigcup \{[x_{a,i_1}, x_{a,i_2}]\}_{i_1<i_2}$ the set $\tilde{\mathcal{F}}_a = \{x^a\mathbf{1}_I | I \in \mathcal{I}_a\}$ verifies the above conditions.
For the case $a = 1$ the construction of the sequence $x_{a,1} = L^1_\epsilon < x_{a,2} < .... < x_{N_{a,\epsilon}-2} < L^2_{a,\epsilon} = x_{N_{a,\epsilon}-1}$ is such that $\int_{x_i}^{x_{i+1}} |x| g^\star(x)dx < \epsilon/2$ is similar except that we introduce 0 in the sequence : $x_{1:N_{a,\epsilon}}$. □

11.7.2 Proof of Lemma 2

Let $t_n = 5\sigma_\star^2 \log n$. One has

$$\begin{aligned}
\mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n\right) &\leq \max_{x_{1:n}\in\mathbb{X}^n} \mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n | X_{1:n} = x_{1:n}\right) , \\
&= \max_{x_{1:n}\in\mathbb{X}^n} \left\{1 - \prod_{i=1}^{n} \mathbb{P}_{\theta^\star}\left(Y_i^2 \leq t_n | X_i = x_i\right)\right\} , \\
&\leq 1 - \left[\mathbb{P}\left(U^2 \leq \frac{t_n - M}{\sigma_\star}\right)\right]^n ,
\end{aligned}$$

where $M = \max_{i=1,\ldots,k} m_i^\star$ and U is a Gaussian random variable with distribution $\mathcal{N}(0,1)$. Then, for large enough n :

$$\mathbb{P}_{\theta^\star}\left(|Y|^2_{(n)} \geq t_n\right) \leq \frac{1}{n^{3/2}} ,$$

and the result follows from Borel Cantelli Lemma.

Bibliography

O. Cappé, E. Moulines, and T.Rydén. *Inference in hidden Markov models.* Springer Series in Statistics. Springer, New York, 2005. ISBN 978-0387-40264-2; 0-387-40264-0. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

O. Catoni and J. Picard. *Statistical learning theory and stochastic optimization: Ecole D'été de Probabilités de Saint-Flour XXXI-2001*, volume 1851 of *Ecole d'Eté de Probabilités de Saint-Flour*. Springer-Verlag, Berlin, 2004. ISBN 9783540225720.

A. Chambaz, A. Garivier, and E. Gassiat. A MDL approach to HMM with Poisson and Gaussian emissions. Application to order indentification. *Journal of Stat. Planning and Inf.*, 139:962–977, 2009.

F. Evennou. *Techniques et technologies de localisation avancées pour terminaux mobiles dans les environnements indoor.* 2007. Ph.D. Thesis, Univ. Joseph Fourier, Grenoble, France.

L. Finesso. Consistent estimation of the order for Markov and hidden Markov chains, 1990. Ph.D. Thesis, Univ. of Maryland.

A. Garivier. Consistency of the unlimited BIC Context Tree estimator. *IEEE Trans. Inform. Theory*, 52:4630–4635, 2006.

E Gassiat. Codage universel et sélection de modèles emboités, 2011. Notes de cours, M2 Orsay.

E. Gassiat and S. Boucheron. Optimal error exponents in hidden Markov model order estimation. *IEEE Trans. Info. Theory*, 48:964–980, 2003.

E. Gassiat and C. Keribin. The likelihood ratio tets for the number of components in a mixture with Markov regime. *ESAIM P&S*, 2000.

E. Gassiat and R. van Handel. Pathwise fluctuations of likelihood ratios and consistent order estimation, 2010.

Katoh N. Inaba, M. and H. Imai. Applications of weighted voronoi diagrams and randomization to variance based k-clustering. In *Proceedings of the tenth annual symposium computational geometry*, pages 332–339, 1994.

R.E. Krichevsky and V.K. Trofimov. The performance of universal encoding. *IEEE Trans. Inform. Theory*, 27:199–207, 1981.

B. Leroux. Maximum-likelihood estimator for hidden Markov models. *Stoch. Proc. Appl.*, 40:127–143, 1992.

J. McQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Sympos. Math. Statist. and Probability*, pages 281–297, Berkeley, California, 1967. University of California Press.

J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29:656 – 664, 1983.

H. Teicher. Identifiability of mixtures. *Ann. Math. Stat.*, 32(1):244–248, 1961.

R. van Handel. On the minimal penalty for Markov order estimation. *Probab. Th. Rel. Fields*, 2011. to appear.

Y. Wang. The variable-length hidden Markov Model and its applications on sequential data mining. Technical report, Departement of computer science, 2005.

Zhou L. Wang J. Wang, Y. and Z.Q. Liu. Mining complex time-series by learning Markovian models. In *Proceedings ICDM'06, sixth international conference on data mining*, China, 2005.

E.T. Whittaker and G.N. Watson. *A Course of Modern Analysis: An Introduction to the General Theory of Infinite Processes and of Analytic Functions, with an Account of the Principal Transcendental Functions*. Cambridge Mathematical Library. Cambridge University Press, 1996. ISBN 9780521588072.

S.Y. Yakowitz and J.D. Spragins. On the identifiability of finite mixtures. *Ann. Math. Stat.*, 39(1): 209–214, 1968.