Thèse de doctorat de
l'Université Pierre et Marie Curie

Spécialité
Mathématiques

Option
Statistique Mathématique

Présentée par
M. Julien CORNEBISE

Pour obtenir le grade de
Docteur de l'Université Pierre et Marie Curie

Effectuée sous la direction de
M. Eric MOULINES, M. Paul DEHEUVELS

Méthodes de Monte Carlo Séquentielles Adaptatives

Adaptive Sequential Monte Carlo Methods

Soutenue le 25 Juin 2009

Devant le jury composé de :

M. Paul DEHEUVELS	(Directeur)
M. Eric MOULINES	(Directeur)
M. Fabien CAMPILLO	(Rapporteur)
M. Paul FEARNHEAD	(Rapporteur)
M. Christophe ANDRIEU	(Examinateur)
M. Gérard BIAU	(Président)
M. Arnaud DOUCET	(Examinateur)
M. Christian ROBERT	(Examinateur)

À Robert Erra,
professeur d'abord,
devenu mentor,
et enfin ami.

Remerciements / Greetings

Alors que se tourne pour moi la dernière page de cette thèse et qu'un lecteur peut désormais en tourner la première, le temps est enfin venu de remercier tous ceux grâce à qui j'ai pu traverser ces années de doctorat et celles qui y ont mené, aussi bien l'an passé qu'il y a vingt ans. Ces remerciements sont nettement plus nombreux que ne le veut l'usage : j'ai eu la chance exceptionnelle de rencontrer beaucoup de personnes qui ne le sont pas moins. Loin de voir là une accumulation insignifiante où la gratitude serait diluée, je les prie d'y voir à quel point chacune d'elles m'a permis de m'ouvrir à toutes les rencontres qui ont suivi. La reconnaissance, comme la responsabilité, ne se partage pas : elle se duplique.

Tout d'abord, je suis particulièrement reconnaissant à mon directeur, **Eric Moulines**, pour la formation scientifique qu'il m'a prodiguée, ainsi que pour les moyens aussi bien matériels qu'intellectuels qu'il m'a consacrés. Si l'éternité est longue, surtout vers la fin, une thèse l'est aussi, surtout au début : sans sa façon parfois très énergique ("le management à l'école du rugby", disait un autre de ses doctorants) d'atteler et ré-atteler ses thésards à l'ouvrage, et sa patience devant le fol éparpillement d'une jeunesse émerveillée par toutes les opportunités mathématiques, j'ai hautement conscience que ces travaux n'existeraient pas. J'ose espérer qu'il ne me tiendra pas grief d'un enthousiasme qui ne demandait qu'à être cadré, et avoir le plaisir de bénéficier longtemps de l'étendue incroyablement vaste de ses connaissances et de son sens scientifique aigü.

Je suis également particulièrement reconnaissant à mon co-directeur, **Paul Deheuvels**. Son accueil dans sa formation de DEA pendant que j'effectuais en parallèle ma dernière année d'école d'ingénieur dans un domaine relativement différent restera l'une des plus portes les plus significatives qui m'aient été ouvertes. "Je vous accepte : sachez que vous prenez un risque, mais si vous êtes prêt à le courir, je vous donne votre chance" tranchait parmi plusieurs autres réponses moins enthousiastes ou plus protectrices, positivement négatives, données à un jeune ingénieur formé à l'informatique et souhaitant se tourner vers les mathématiques. Cette ouverture d'esprit est pour moi un exemple.

De pair avec avec mes directeurs de thèse, il me faut mentionner – et bien que cela ne soit pas coutumier à cet endroit des remerciements – un autre professeur, **Robert Erra**,

qui fut mon directeur avant l'heure, tout au long de mes cinq années à l'Ecole Supérieur d'Informatique Électronique Automatique, et continue de l'être à ce jour. J'ai désormais la fierté de le compter comme ami. Il m'a transmis la passion pour l'algorithmique, pour le raisonnement formel allié à l'intuition mathématique, la joie absolue d'un éclair de "Ah-ah !" lorsque toutes les pièces tombent en place d'un coup : une jubilation absolue, sans comparaison, et qui a encore plus de goût lorsqu'elle est partagée. Je ne compte plus les conversations téléphoniques à minuit, et les innombrables heures consacrées à ce trublion qui déboulait à tout instant dans son bureau en s'exclamant "Robert, il faut que je te montre un truc !" et sa variante "j'ai une grave question existentiello-algorithmique !" occuperaient sans défaut le service cumulé de plusieurs professeurs d'université à temps complet ! Pour tous tes conseils, pour cette passion, pour cette formation, Robert, je te dédie cette thèse.

C'est avec gratitude que je remercie les rapporteurs de ce manuscrit, **Fabien Campillo** et **Paul Fearnhead**, qui ont accepté de prendre de leur temps particulièrement précieux pour se pencher sur mes travaux. Leurs remarques me sont précieuses pour étendre mes recherches et les affiner avant publication finale. J'ai également une dette envers les réputés membres de mon jury, **Christophe Andrieu**, **Gérard Biau**, **Arnaud Doucet**, et **Christian Robert**, qui, en acceptant aujourd'hui de critiquer mon travail et d'analyser par leurs questions mes compétences mathématiques, me permettent d'entrer dans la communauté scientifique sous la vigilance de prestigieux aînés. Rarement doctorant a eu le redoutable privilège de défendre ses résultats devant une réunion de tels experts du domaine. Si votre présence est un honneur, elle ne m'en motive que davantage à la mériter.

I cannot even think of this thesis without paying due tribute to my co-author, **Jimmy Olsson**. His infinite enthusiasm, combined with his impressive mastering of mathematical techniques and his pedagogical gift, has been worthy beyond words. From our first work on the mystic "grape effect" of refueling up to our most recents developments on optimal weights, through his frequent trips to Paris (sorry for the torture of staying at Maison des Eleves on your thirtieth birthday !) and the exceptional welcome I received in Lund early 2008, this is a collaboration that went far beyond anything I could expect. Nonetheless did you shape my mathematical rigor (any persisting lack is my entire fault) but also displayed a brilliant example of what a young statistician/probabilist/mathematician can be. Nobody knowing you doubts that you are a great researcher ; I can now testify first-hand that you will also be an awesome Ph.D. advisor.

Par ailleurs, une thèse, bien qu'immatérielle, prend corps dans un laboratoire, au sein d'une équipe et d'un entourage scientifique. J'ai eu la chance d'avoir deux laboratoires, le LSTA à Chevaleret et l'équipe STA de TSI rue Dareau. Les membres permanents y sont autant de sources de conseils extérieurs, d'articles, d'expérience, ou d'occasions d'exposer dans différentes conférences. Il me faut particulièrement remercier ici **Gersende Fort** pour ses conseils toujours judicieux, **Randal Douc** pour le temps qu'il a accordé à mes balbutiements particulaires quand bien même il brûlait d'aller à la vitesse de sa propre connaissance du domaine, ainsi que toute l'équipe STA fréquentée ces années, **Olivier Cappé**, **François Roueff**, **Céline Levy-Leduc**, **Cédric Fevotte**, **Jamal Najim**, pour leurs encouragements. Je n'oublierai pas l'un des trésors cachés de Telecom, **Sophie-Charlotte Barrière**, une sysadmin en or massif, c'est vital, qu'elle soit remerciée pour sa patience et la souplesse avec laquelle elle gère les machines Unix !

Au LSTA, il m'est impossible de ne pas mentionner, outre **Gérard Biau** (déjà cité), source de conseil depuis ma première conférence avant même le DEA (AMSDA 2005),

Philippe Saint-Pierre, ainsi que l'ensemble des professeurs et maîtres de conférence. Merci aussi à **Louise Lamart** et **Anne Durrande** pour toutes les difficultés administratives qu'elles aplanissent sans relâche.

The final scientific developments of this thesis took place during my first stay at the Statistical and Applied Mathematical Sciences Institute (SAMSI), in North Carolina, whose extremely stimulating environment has been a tremendous boost. It is therefore a great pleasure to thank here its **whole directorate** and in particular its director, **Jim Berger**, for his great benevolence and flexibility. The interactions with the numerous researchers meeting there, staying, leaving, coming back, were fruitful occasions to get an incredibly wider view of my field and to spot exciting new developments. Last but not least, the very helpful staff, especially **Denise Auger**, **Rita Fortune**, and **Terri Nida**, has been precious in making my mind free of any material contingencies, allowing me to focus solely on my research.

Regardant vers le futur au SAMSI, je n'oublie pas mon passé à l'ESIEA, qui m'a permis d'arriver en DEA, et tout spécialement l'équipe de choc de mes professeurs d'algorithmique et d'informatique, **Laurent Beaudoin**, **Stéphane Duval**, et **Sophie Maucorps**, ainsi que **Nicole Viaud** et **Dominique Rivolier** : j'ai trouvé à l'ESIEA bien plus que je ne pensais y trouver en entrant, dans des domaines certes scientifiques mais également (ou plus encore) humains, dont je n'imaginais pas alors l'existence. Je tiens également à remercier le directeur de l'ESIEA, **Pierre Aliphat**, pour le soutien qu'il m'a apporté, notamment en m'autorisant à poursuivre en parallèle ma dernière année de l'ESIEA et mon DEA. A tous, ainsi qu'à mes autres professeurs, je tiens à dire que la place qu'accordent leurs enseignements et notre école aux profils les plus divers a fait de l'ESIEA un lieu où je me suis épanoui comme jamais je n'aurais pu le faire ailleurs. Et tant que j'en suis à regarder dans le passé de mon parcours, qu'il me soit permis d'aller chercher plus loin encore, et de remercier **Jean-Marie B.** qui m'a apporté une confiance cruciale et prouvé que les épreuves les plus effrayantes peuvent n'être que des pages d'écritures, et **Martine** pour m'avoir offert, par le paradis des rayons de sa bibliothèque, des mondes et des univers d'une richesse inimaginable, qui ne cessent de m'accompagner depuis l'enfance.

Au-delà des conseils des permanents, il y a aussi (voire surtout !) dans les laboratoires ceux avec qui j'ai partagé (de nombreux !) bureaux, galères, bruits de couloirs, tuyaux sur les conférences, astuces LaTeX, lemmes utiles, soirées world-food, doutes, exultations, calembours brameux, ordinateurs chantant l'internationale, bref, la joyeuse équipe de thésards, postdocs et tous jeunes maîtres de conférences. Aux troisième et quatrième étages de Darreau, **Alexandre L.** (enfin quelqu'un qui fait du Python !), **Anne-Laure B.** (mais si, mais si, ça va marcher ton algo), **Aurélia F.** (signal processeuse et matheuse, quel combo !), **Christophe T.** (aux disques improbables et à la mauvaise foi réthorique aussi légendaire qu'indispensable), **Cyril C.** (et tes commentaires sur mon utilisation très personnelle de `diff@`), **Jean-François G.** (passé du côté "costume" de la force), **Jean-Louis D.** (qui n'a jamais hésité à descendre au troisième pour appeller au FIAP), **Jean L.** (encore un écran ?), **Jérôme G.**(Heavy Metal is the only law), **Loïs R.** (tu devrais passer plus souvent !), **Malika K.** (prend soin de ce bureau, et remets vite des affiches !), **Marine D.** (sainte relecture), **Nancy B.** (la projection L^2 de Cécilia de PhDcomics sur l'espace engendré par Telecom), **Natalyia S.** ("Ah mais non !" – ou la Sainte Russie et son rideau de fer faits femme), **Olaf K.** (et tes délicieux mets camerounais – un peu de steak sous ton poivre ?), **Sarah F.** (et ses principes, et les fork bombs permises, et `xhost +` c'est mal mais c'est rigolo), **Steffen B.** (ça fait deux ans, Steffen, tu devrais avoir fini – et ferme ce site web, je te vois), **Tabea R.** (merci d'avoir mis tout ton calme germanique à tolérer mon hideux poster hard-rock, je

t'assure qu'il était esthétique !), **Teodora P.** ("Djoudjou, arlrlête tes bêtises !"), **Thomas T.** (mon PhDrérot ! glorieux aîné, si tu nous lis. . .), **Zaïd H.** (frangin de thèse, tant de choses à raconter. . . pizzas, starbucks, ELLE magazine, bouclages nocturnes d'articles, savons reçus, . . . mon accès JSTOR est à ta disposition). Ayant deux laboratoires, j'ai eu deux fois plus de personnes, d'amis, avec qui partager tout cela. Je remercie donc pour le support – et l'endurance devant les calembours brameux – ma Chevaleret-sque famille : **Aurélie F.** (un sérieux immuable dans ce bureau, j'admire !), **Cécile A.** (une geekette fan de Cthulu arrêtant les réacteurs nucléaires avec des modules de Drinfeld, quel bonheur), **Claire C.** (tu suis les traces de tes aînés, bientôt tu t'attaqueras aux boîtes mails ouvertes), **Clara Z.** (et les discussions passionnantes sur l'éducation en milieu urbain), **Esterina M.** (ma che sono belle le italiene. . .), **Gwladys T.** (là aussi, 3 ans de thèse de concert, tant à dire . . . C'est un homme qui rentre dans un café et plouf), **Ismaël S.** (et les rencontres à l'Arobase en sortant de TD), **Jeanne C.** (que vivent les ID roses bizarroïdes en plastique et les bons de deux mètres), **Jérôme G.** (courage, si les voyages forment la jeunesse, les jeunes pères sont excusés), **Jean-Baptiste A.** (prince du calembour, roi du carambar), **Nathalie K.** (l'Île de Pâques et leurs statues), **Olivier B.** (grand successeur dans l'organisation du GTT LSTA, et un tueur à Geo Challenge), **Olivier F.** (preuve que l'amour vache existe), **Omar E.-D.** (l'expert des astuces mathématiques et du café si serré que la cuiller tient à la verticale), **Pierre R.** (fournisseur officiel du 8A27 en Kinder Surprise, gardien d'enfant devant l'éternel, et qui ne joue jamais à Kdo-kdo – si si, Esterina, il ne joue jamais, jamais, jamais, jamais), **Rosalba I.** (troveremo la cura !), **Samuela L.-A.** (ciao la Mama[3] !), **Segolen G.** (les crêpes, c'est quand tu veux !), **Sophie D.** et **Vincent B.** (longue vie au GTT des collègues du LPMA), ainsi que les doctorants et ex-doctorants du plateau B du 8ème, **Boris L.**, **Lahcen D.**, **Salim B.**, **Véronique V.**, et tous ceux que dans le feu de l'action je suis coupable d'oublier.

Il y a également les amis encore plus proches, ceux que l'on croise au détour d'une école, d'un labo, et avec qui on fera toujours le 401$^{\text{ème}}$ coup des fameux quatre-cents. Pas forcément directement impliqués dans la thèse, et pourtant tellement indispensables : **Benjamin C.** (c'est le Bien !), **Benoît G.** (18 ans ne se résument pas en une phrase – Coin _x<), **François B.** (mon Biii !), **Manuel P.-G.** (jamais vu apprenant plus rapide. . .).

Sarah, tu m'as accompagné dans ce chemin, supportant l'humeur massacrante d'un algorithmicien frustré, d'un codeur buggé, d'un matheux éprouvé et d'un statisticien plus limite que central. Tu m'as apporté beaume, apaisement, calme, et tant de joie de tant de moments partagés. Comment peux-tu être encore à mes côtés ? ! Permets moi désormais d'être aux tiens.

Enfin, je tiens à remercier ma famille : mes parents **Didier** et **Marie-Françoise Cornebise**, et ma soeur **Marion**. Si la plupart des jeunes docteurs mentionnent leurs proches, il m'est impossible de savoir si tous ont autant à coeur ce qu'ici je souhaiterais dire – et devant mon impuissance à trouver les mots justes, j'en doute. Écrire à tous trois ce que j'ai à l'esprit dépasserait de loin les bornes de ces remerciements déjà très personnels. Ils savent. Je leur dois plus encore qu'ils ne le pensent.

Résumé

Cette thèse est consacrée à l'étude et au développement des algorithmes de Monte Carlo séquentiels (*SMC*), aussi connus sous le nom de *méthodes particulaires*, tels que définis par Gordon et al. (1993) avec le *bootstrap filter*, étendus par Pitt and Shephard (1999) avec le filtre particulaire auxiliaire (*APF*), résumées par Doucet et al. (2001), et analysées notamment par Del Moral (2004) et Cappé et al. (2005); Douc and Moulines (2008).

Nous nous penchons sur la conception et l'analyse d'algorithmes adaptatifs, c'est à dire capables de choisir automatiquement les paramètres tels que les *poids d'ajustement multiplicatifs* (ou "poids de première étape") du APF, ou le noyau de proposition. Le but est d'effectuer les choix optimaux en termes d'efficacité calculatoire et de précision des estimateurs obtenus, cette optimalité étant formalisée mathématiquement par des critères dont l'étude est elle-même un sujet de recherche actif abordé dans ce travail.

Nous apportons tout d'abord une étude théorique des pratiques existantes, notamment au travers d'une analyse asymptotique des critères tels que le *coefficient de variation* et l'*entropie* des poids d'importance, actuellement utilisés sur des bases empiriques. Nous les relions aux divergences du χ^2 et de Kullback-Leibler (*KLD*) entre deux distributions que nous explicitons. Nous développons également de nouveaux critères ayant des propriétés plus adaptées à l'usage souhaité, permettant tout particulièrement de découpler le problème d'adaptation des poids d'ajustement et celui de l'adaptation du noyau de proposition.

Nous établissons de nouveaux algorithmes susceptibles d'être utilisés sur les modèles actuels les plus complexes, tout en gardant à l'esprit la simplicité calculatoire requise dans le cadre fondamentalement itératif des méthodes SMC. Pour l'adaptation des poids d'ajustement, nous proposons et analysons des algorithmes de *rechargement* de l'échantillon combinés à un *échantillon exploratoire* (*pilot sample* en anglais). Pour celà, nous établissons au préalable la convergence de l'APF avec poids d'ajustement aléatoires.

Enfin, nous adaptons le noyau de proposition en ajustant un mélange de noyaux paramétrisé, de façon similaire aux *mélange d'experts* en apprentissage statistique. Les noyaux des composantes appartiennent à la vaste famille des distributions exponentielles courbes *intégrées*, plus riche que la classique famille exponentielle. Les poids du mélange proviennent d'une régression logistique, qui partitionne l'espace des particules d'origine en sous-régions, auxquelles sont allouées un petit nombre de noyaux spécialisés. L'algorithme d'optimisation présenté est basé sur les algorithmes Stochastic Approximation EM, Monte Carlo EM, et la méthode de la Cross-Entropy. Nous illustrons ses performances en terme de réduction de la KLD sur plusieurs exemples numériques étudiés en profondeur.

Abstract

This dissertation focuses on the study and development of sequential Monte Carlo algorithms (*SMC*), also known as *particle algorithms*, as defined in Gordon et al. (1993) with the celebrated *bootstrap filter*, extended by Pitt and Shephard (1999) with the auxiliary particle filter (*APF*), summarized in Doucet et al. (2001), and analyzed in Del Moral (2004) and Cappé et al. (2005); Douc and Moulines (2008), among others.

We focus on the conception and design of adaptive algorithms that are able to automatically tune the optimal parameters such as the APF's *adjustment multiplier weights* (or "first stage weights") or the *proposal kernel*. We aim for the optimal choice in terms of computational efficiency and accuracy of the resulting estimates; this optimality is mathematically formalized by critera whose study is itself an active research topic discussed in this work.

We first bring a theoretical study of existing practices, most notably through an asymptotic analysis of criteria such as the *coefficient of variation* and the *entropy* of the importance weights, which are currently used on an empirical basis. We link them to χ^2 and Kullback-Leibler divergences, respectively, between two distributions that we explicit. We also develop new critera with properties better suited to the purpose, specifically allowing for decoupling of the adaptation of the adjustment multiplier weights and the adaptation of the proposal kernel.

We establish new algorithms likely to be used on today's most intricate models, keeping in mind the mandatory computational efficiency of the fundamentally iterative SMC methods. Concerning the adaptation of the adjustment multiplier weights, we propose and analyze *refueling* algorithms combined with a *pilot sample*. To achieve this, we first state the convergence of the APF with random adjustment multiplier weights.

Finally, we adapt the proposal kernel by fitting a parametric mixture of kernels, closely connected to the *mixture of experts* from machine-learning. The components' kernels belong to the broad family of *integrated* curved exponential distributions, richer than the more classical exponential family. The weights of the mixture stem from a logistic regression, that partitions the space of the original particles into subregions to which are assigned a few specialized kernels. The flexibility of this family allows for fitting highly nonlinear and multi-modal distributions. The optimization algorithm presented is inspired by Stochastic Approximation EM, Monte Carlo EM, and Cross-Entropy method. We illustrate its performance, in terms of KLD reduction and impact on the importance weights, on several thoroughly examined numerical examples.

Sommaire / Contents

Introduction et survol de la thèse

Les méthodes de Monte Carlo séquentielles (*MCS*), ou filtres particulaires, ont connu des avancées majeurs dans la dernière décennie, amenées en premier lieu par les besoins du filtrage stochastique, et, par là-même, poussant à une interaction sans cesse accrue entre communauté appliquée (traitement du signal, ingénierie, et désormais biologistes, physiciens, chimistes) et communauté théorique (statisticiens, probabilistes). L'article novateur de Gordon et al. (1993), donne jour au *filtre bootstrap* qui utilise l'idée clé de rééchantillonner parmi plusieurs trajectoires issues d'un échantillonnage préférentiel séquentiel classique (EPS, voir le Chapitre 1 de la présente thèse), a été rapidement suivi par les articles Kong et al. (1994); Liu and Chen (1995, 1998), qui développèrent cette méthodologie avec innovation, en privilégiant dans ces premières années les utilisations pratiques sur les justfications théoriques. Une autre avancée majeure, qui, aussi reconnue soit-elle, n'est pas encore exploitée à sa pleine valeur, a été faite par Pitt and Shephard (1999), introduisant la notion de filtre particulaire auxiliaire (*FPA*, voir le Chapitre 2) qui fournit un degré de liberté supplémentaire au moyen de variables auxiliaires d'indice permettant, via des poids d'ajustement calculés préalablement à l'étape de rééchantillonnage du filtre bootstrap, d'approcher au mieux les distributions d'intérêt.

Vint ensuite l'ouvrage collectif Doucet et al. (2001), qui, avec l'article Doucet et al. (2000), a attiré une grande attention des champs d'application les plus variés : ingénierie financière, ingénierie biologique (voir par exemple Liu (2001)), traitement du signal et de l'image, simulation des évènements rares, optimisation combinatoire (Rubinstein and Kroese, 2004), suivi de cible (Ristic et al., 2004), et même robotique (Fox, 2003; Thrun et al., 2005). Cette liste s'allonge toujours, de nouvelles recherches permettant de relever des défis toujours plus grands, que ce soit en terme de complexité des modèles (estimation du terme source d'une diffusion atmosphérique en milieu urbain, (Johannesson et al., 2004, Section 5.2) et Septier et al. (2009) – suivi de plusieurs cibles en parallèle, Pang et al. (2009); F. Septier and Carmi (2009) – suivi de l'évolution et du lignage d'un groupe de cellules, Wang et al. (2009)) ou en terme de contraintes temporelles fortes (systèmes robotiques embarqués, Montemerlo (2003)).

Du point de vue théorique, de nombreux articles ont analysé les détails de ces algorithmes, amenant au fil des années de puissants outils et une compréhension profonde des mécanismes mathématiques mis en oeuvre, montrant le chemin pour une nouvelle classe d'algorithmes toujours plus performants. Plusieurs articles de Pierre Del Moral à la fin des années 1990 (dont notamment Del Moral (1996); Del Moral and Miclo (2001); Del Moral and Jacod (2001)) ont culminé en son livre Del Moral (2004), qui

présente en profondeur une théorie générale de la plupart des algorithmes MCS sur les fluxs de Feynman-Kac, pour des fonctions d'intérêt bornées. Ces fondations remarquables ont ensuite permis d'établir de nombreux résultats, aussi fins, par exemple, que ceux concernant l'étude de la coalescence de l'arbre ancestral des particules (Del Moral et al., 2006). En parallèle, une autre approche théorique se fonde sur une décomposition des algorithmes en étapes élémentaires analysées séparament, par opposition à la considération du flux de Feynman-Kac dans son ensemble. Cette technique est exposée dans Cappé et al. (2005), qui se concentre sur les modèles de Markov cachés (MMC, cf. Section 1.2, et développée plus avant par Douc and Moulines (2008) (voir Annexe A), apportant des théorèmes de convergence (loi des grands nombres, théorème limite central) pour des fonctions d'intérêt non-bornées. Dans cette lignée, plusieurs travaux ont permis d'éclairer de multiples variantes, comme Olsson et al. (2008) sur le lissage à délai fixé, ou Douc et al. (2008) sur le FPA (voir de nouveau le Chapitre 2). Des approches alternatives sur ce dernier algorithme ont été mises au point indépendamment dans la communauté, notamment par Johansen and Doucet (2008), qui considère un espace étendu permettant de traiter le FPA comme un algorithme MCS classique.

Ces méthodes sont désormais si bien établies qu'elles donnent le jour à des algorithmes de plus en plus génériques. Les derniers exemples en sont les échantillonneurs MCS (Del Moral et al., 2006), qui considèrent des séquences arbitraire de distributions sur un seul et même espace à l'aide de noyaux "backward", ou les méthodes de Population Monte-Carlo (*PMC*) telles qu'étudiées dans Cappé et al. (2008). Des hybrides entre les méthodes de Monte Carlo à chaînes de Markov (*MCMC*, voir Robert and Casella (2004)) et les filtres particulaires sont également développées dans Andrieu et al. (2009). Ceci ouvre des perspectives de plus en plus concrètes pour une utilisation de variantes de MCS pour des problèmes atteignant les limites des capacités des méthodes MCMC.

Tous ces algorithmes, du filtre bootstrap de base aux derniers échantillonneurs MCS, partagent une architecture commune : un échantillon pondéré (composé de *particules*) approchand une distribution cible, facultativement repondéré par des *poids d'ajustement multiplicatifs* (cas du FPA), est rééchantillonné puis propagé au moyen d'un noyau de transition Markovien appelé *noyau de proposition*, après quoi les poids sont mis à jours afin d'approcher la nouvelle distribution cible. Le grand défi actuel dans la communauté MCS est la conception d'algorithmes adaptatifs, c'est à dire à même d'ajuster automatiquement les paramètres tels que le nombre de particules (Legland and Oudjane, 2006; Fox, 2003), l'agenda de rééchantillonnage (Chen et al., 2005), les poids d'ajustement (ou "poids de première étape") du FPA (Douc et al. (2008) dans le cas particulier d'une fonction d'intérêt fixée), et le noyau de proposition (Pitt and Shephard, 1999; Doucet et al., 2001; Van der Merwe et al., 2000). Le but est d'effectuer automatiquement les choix permettant une qualité optimale des estimateurs résultants, cette qualité étant formalisée mathématiquement par des critères dont l'étude est en elle-même un sujet de recherche actif – la variance étant le critère canonique des méthodes de Monte Carlo mais n'étant pas suffisante dans le cadre des méthodes séquentielles (voir le Chapitre 3). Ce besoin est d'autant plus net que les algorithmes récents évoqués plus haut, ou des variantes d'algorithmes existants telles que l'échantillonnage par bloc de Doucet et al. (2006), impliquent des quantités de plus en plus élaborées pour lesquelles une intervention de l'utilisateur à chaque étape n'est pas envisageable.

L'objectif de cette thèse est de contribuer aux progrès de la communauté MCS sur ces challenges,

– en apportant tout d'abord une étude théorique des pratiques existantes, notamment au travers d'une analyse mathématique des critères actuellement utilisés

sur des bases empiriques (coefficient de variation et entropie des poids), tout en développant de nouveaux critères ayant des propriétés plus adaptées à l'usage souhaité,
– puis en mettant cette étude à profit pour établir de nouveaux algorithmes à même d'être utilisés sur les modèles actuels les plus complexes, tout en gardant à l'esprit la simplicité calculatoire requise dans le cadre fondamentalement itératif des méthodes MCS.

Après cette courte introduction volontairement exempte de toute notation mathématique, nous présentons dans le Chapitre 1 une introduction aux méthodes MCS. Basée sur (Cappé et al., 2005, Chapitre 7), avec les éléments nécessaires de (Cappé et al., 2005, Chapitres 1, 3, 6) pour se suffir à lui-même, et sur un léger changement d'approche des récursions fondamentales – fruit de trois années d'expériences depuis publication –, elle est conçue pour être accessible à un doctorant débutant une thèse sur le domaine, voire à un bon étudiant de Master 2, et fournir le socle nécessaire à la compréhension des résultats de la présente thèse – à défaut de leur élaboration. Ce chapitre est complété d'une façon plus théorique par l'Annexe A qui rappelle les théorèmes fondamentaux de Douc and Moulines (2008). Il annonce également dans leur contexte les travaux propres des Chapitres 3 et 5, dont les buts sont compréhensibles avec les éléments standards de MCS, sans recourir à la notion du filtre particulaire auxiliaire.

Ce dernier est présenté en détail dans le Chapitre 2, moins introductif que le précédent. Nous y décrivons le FPA, nécessaire à la compréhension des détails des travaux propres de cette thèse, ainsi que les théorèmes de convergence établis dans Olsson et al. (2008) Ceci permet alors de revenir sur l'introduction des travaux des Chapitres 3 et 5 et d'annoncer ceux du Chapitre 4. Viennent ensuite les contributions originales de cette thèse, en anglais puisqu'il s'agit de trois articles, l'un déjà publié et les deux autres en cours de soumission.

Le Chapitre 3 correspond à l'article Cornebise et al. (2008) publié dans le courant de cette thèse. Seules les notations en ont été changées, par souci d'uniformisation avec les autres chapitres. Sa portée est triple. Tout d'abord, nous y étudions mathématiquement la convergence (lorsque la taille de l'échantillon pondéré – le nombre de particules – augmente) et la signification des critères de qualité omniprésents en MCS que sont le coefficient de variation et l'entropie des poids d'importance. Nous démontrons qu'ils convergent vers, respectivement, la divergence du χ^2 et celle de Kullback-Leibler entre deux distributions pouvant être interprétées comme les distributions asymptotiques mises en jeu par le FPA. A notre connaissance, la seule analyse du coefficient de variation proposé par Kong et al. (1994) était jusqu'alors l'article Liu and Chen (1995), qui précise procéder de façon empirique ("rule of thumb" dans le texte original). Ensuite, nous proposons deux critères alternatifs exprimés en termes de divergence du χ^2 et de Kullback-Leibler entre la distribution de proposition et la distribution cible à taille d'échantillon fixe. Ils sont asymptotiquement équivalents aux deux critères précédents puisqu'ils convergent vers les mêmes quantités, mais entrent par ailleurs plus aisément dans une approche d'analyse du risque indépendant des fonctions d'intérêts pour une taille d'échantillon fixe. Enfin, nous remarquons que ces deux critères permettent une optimisation bien plus aisée. En particulier, le critère basé sur la divergence de Kullback-Leibler sépare l'adaptation de la loi de proposition en deux sous-problèmes distincts et indépendants : l'adaptation des poids d'ajustement du filtre particulaire auxiliaire, et l'adaptation du noyau de proposition. Ces deux sous-problèmes constituent le coeur du reste de cette thèse, respectivement les Chapitres 4 et 5, en alliant étude théorique et développement pratiques. Auparavant, dans la fin de ce chapitre, nous posons les premiers éléments permettant l'adaptation des noyaux en attirant

l'attention sur ses ressemblances avec les problèmes usuellement résolus par l'algorithme Expectation-Maximisation (*EM*, Dempster et al. (1977), sa variante Monte Carlo (*MCEM*, Fort and Moulines (2003)), et la Cross-Entropy (*CE*, Rubinstein and Kroese (2004)). Cette approche n'est alors qu'esquissée, son traitement en profondeur faisant l'objet du Chapitre 5.

Le Chapitre 4, s'appuyant sur les critères proposés dans le Chapitre 3, aborde l'adaptation des poids d'ajustement du filtre particulaire auxiliaire, élément trop souvent négligé qui ajoute à la distribution de proposition un degré de liberté nécessaire pour ajuster au mieux la distribution cible – et donc pour augmenter la performance des méthodes MCS. Nous commençons par donner une analyse précise de l'algorithme de *pilot-sampling* (Zhang and Liu, 2002) en terme de consistence et de normalité asymptotique. Cet algorithme, construit sur la base d'heuristiques empiriques, n'a à notre connaissance jamais été étudié théoriquement. Cette première étude permet de proposer un nouvel algorithme qui inclut une étape d'exploration pilote, mais également un "rechargement" de l'échantillon en nouvelles particules pour garantir que les critères de qualités proposés dans le Chapitre 3 soient maintenus sous un certain seuil. Les preuves de convergence de cet algorithme sont également données.

Le Chapitre 5 propose un nouvel algorithme pour adapter le noyau de proposition des méthodes MCS. Nous le présentons dans le cadre du FPA pour permettre la plus grande généralité, mais il est également directement applicable au filtre bootstrap. L'accent est mis sur la capacité de ces algorithmes à approcher les noyaux optimaux bien plus compliqués que ne le permettent les méthodes existantes comme l'approximation Laplacienne de Pitt and Shephard (1999), le filtre particulaire de Kalman étendu de Doucet et al. (2001) ou le filtre particulaire du Kalman sans parfum de Van der Merwe et al. (2000), tout en conservant un coût informatique raisonnable. Les remarques finales du Chapitre 3 y sont alors pleinement développées, en ajustant au noyau optimal du filtre particulaire auxiliaire une famille de *mélange d'experts* très proche des travaux de Jordan and Jacobs (1994); Jordan and Xu (1995). Cette famille consiste en un mélange de noyaux appartenant à la très large famille exponentielle courbe *intégrée*, classe générale dont nous illustrons les deux membres les plus connus que sont les Gaussiennes et les t-Student multivariées. Les poids de ce mélange constituent une régression logistique permettant une partition douce de l'espace des particules à l'instant précédent, spécialisant chaque noyau pour une ou plusieurs sous-régions. Les algorithmes de ce chapitre sont inspirés de l'algorithme Stochastic Approximation EM (*SAEM*, Delyon et al. (1999); Kuhn and Lavielle (2004); Andrieu et al. (2005)) ainsi que de MCEM et CE déjà cités.

Introduction and outline of the work

Sequential Monte-Carlo (*SMC*) methods, also known as particle filters, have known major breakthroughs in the last, brought at first by the needs of stochastic filtering, and from there, leading to an ever-increasing interaction between the applied community (signal processing, engineering, and more recently biology, physics, chemistry) and the theoretical community (statisticians, probabilists). The seminal article Gordon et al. (1993), introducing the *bootstrap filter* which uses the key idea of resampling amongst several paths of classical Sequential Importance Sampling (*SIS*, see Chapter 1 of the present thesis), has been quickly followed by Kong et al. (1994); Liu and Chen (1995, 1998), who developed this methodology in innovating ways, focussing more, in these early days, on the practical uses than on the theoretical justifications. Another major step has been achieved by Pitt and Shephard (1999) who introduced the notion of Auxiliary Particle Filter (*APF*, see Chapter 2), bringing another degree of freedom by introducing auxiliary index variable, thus allowing for adjustment multiplier weights before the resampling step.

Then came the collection Doucet et al. (2001), which, along with Doucet et al. (2000), brought a great deal of attention on these methods from the most diverse application fields: financial engineering, biological engineering (see e.g. Liu (2001)), signal and image processing, rare-event simulations, combinatorial optimization (Rubinstein and Kroese, 2004), tracking (Ristic et al., 2004), and even robotics (Fox, 2003; Thrun et al., 2005). This list is groing each day, with new researches allowing to take up higher and higher challenges, either in terms of model complexity (source term estimation for atmospheric release, (Johannesson et al., 2004, Section 5.2) and Septier et al. (2009) – multiple objects tracking, Pang et al. (2009); F. Septier and Carmi (2009)) – lineage estimation in celltracking, Wang et al. (2009)) or in terms of strong temporal constraints (on-board robotic systems, Montemerlo (2003)).

On the theoretical side, many articles have analyzed the details of the algorithms, bringing over the years powerful tools and deep understanding of the underlying mathematical mechansisms, paving way for a whole class of refined algorithms. Several articles from Del Moral since the late 1990's (including Del Moral (1996); Del Moral and Miclo (2001); Del Moral and Jacod (2001)) culminated in his book Del Moral (2004), which exposes an in-depth general theory of most aspects of SMC algorithms on generic Feynman-Kac flows, for bounded functions of interest. These remarkable foundations then permitted to establish such subtle results as, for example, the study of the coalescence tree of the particles (Del Moral et al., 2006). In parallel, another theoretical approach relies on a step-by-step study, decomposing SMC algorithms into elemen-

tary steps analyzed separately, rather than considering the whole Feynman-Kac flow. This theory is exposed in Cappé et al. (2005), which focuses on Hidden Markov Models (*HMM*, see Section 1.2), and is further developed in Douc and Moulines (2008) (see Appendix A), bringing convergence theorems (law of large numbers, central limit theorems) for unbounded functions of interest. Along these lines, many works have then shed brighter light on several variants, such as Olsson et al. (2008) on the fixed-lag smoothing, or Douc et al. (2008) on the APF (see once again Chapter 2). Alternative approaches for this latter algorithm have stemmed in the community, such as Johansen and Doucet (2008), who use extended spaces to fit the APF in the more classical SMC algorithms previously studied.

These methods are now so well established that they give birth to more and more generic algorithms. The latest additions are the SMC samplers (Del Moral et al., 2006) that consider arbitrary sequences of distributions on a single space by means of "backward" kernels, or Population Monte-Carlo (*PMC*) as studied in Cappé et al. (2008). Hybrids between Markov Chain Monte Carlo (*MCMC*, see Robert and Casella (2004)) methods and particle filters are also under study (Andrieu et al., 2009). This open tremendous perspectives for a possible use of SMC on problems reaching the limits of MCMC possibilities.

All of these algorithms, from the most basic bootstrap filter up to the latest SMC sampler, share a common skeleton: a cloud of weighted samples (named *particles*) targeting a distribution, possibly reweighted by so-called *adjustment multiplier weights* (in APF), is then resampled and propagated by means of a Markov transition kernel called *proposal kernel*; the weights are finally updated to approximate the new target distribution. The current grand challenge in the SMC community is the design of adaptive algorithms, i.e. self-tuning parameters such as the number of particles (Legland and Oudjane, 2006; Fox, 2003), the resampling schedule (Chen et al., 2005), the adjustment multiplier (or "first-stage") weights of the APF ((Douc et al., 2008) for the particular case of a fixed function of interest), and the proposal (or mutation) kernel. (Pitt and Shephard, 1999; Doucet et al., 2001; Van der Merwe et al., 2000). The aim is to to automatically pick the choices that lead to an optimal quality of the resulting estimators, this quality being mathematically formalized in terms of criteria whose study is an active research field in its own right – the variance, although the canonical criterion in Monte Carlo methods, is not sufficient for sequential methods (see Chapter 3). This need is all the clearer as the recent algorithms mentionned above, or variants of existing algorithms such as block-sampling of Doucet et al. (2006), rely on more and more intricate quantities which can hardly be user defined at each step.

The aim of this thesis is to contribue to the SMC community's progresses on these challenges,

- first, by leading a theoretical study of existing practices, most noticeably through a mathematical analysis of criteria which are widely used on an empirical basis only (coefficient of variation and entropy of the importance weights), along with developing new criteria enjoying properties more adapted to SMC needs,
- then, by making the most of this study by establishing new algorithms, which can be used on today's most intricate models, though keeping in mind the mandatory low computational overhead required by (fundamentally iterative) SMC methods.

After this short introduction, purposefuly free of any mathematical notation, we expose in Chapter 1 a crash course on SMC methods, in french. Based on (Cappé et al., 2005, Chapter 7) with the material from (Cappé et al., 2005, Chapters 1, 3, 6) required to make it largely self-contained, and including a slight change in the decomposition of the key recursions – fruit of three years of experience since original publication –, it is

meant to be within reach of a beginning graduate student starting in the field, and to provide him/her with the basis required for understanding the results of the present thesis – in default of their construction. It also announces in context the genuine work exposed in Chapters 3 and 5, whose goal can be understood without mastering the APF. This chapter is completed by the more theoretical Appendix A, which recapitulates the fundamental theorems of Douc and Moulines (2008).

The said APF is presented in full details in Chapter 2, less introductory than the previous chapter. We describe there the APF which is mandatory to understand the details of the genuine work of this thesis, as well as some convergence theorems established in Olsson et al. (2008). This allows to give slightly more introductory details on the findings exposed in Chapters 3 and 5 and to announce those of Chapter 4. Then come the real focus of this thesis, which correspond to three journal articles, one already published and to with pending submission.

Chapter 3 corresponds to the article Cornebise et al. (2008) published last year. The only change lies in the notations, which have been uniformized with the other chapters. Its aim is threefold. Firstly, we study mathematically the convergence (as the size of the weighted sample – i.e. the number of particles – grows) and the meaning of the ubiquitous quality criteria in SMC, namely the coefficient of variation and the entropy of the importance weights. We prove that they converge to, respectively, the χ^2 and the Kullback-Leibler divergences between two distributions that can be seen as the asymptotic distributions put into play by the APF. To the best of our knowledge, the only analysis of the coefficient of variation proposed by Kong et al. (1994) was the article Liu and Chen (1995) which specifies that its results are established by means of "rules of thumb". Secondly, we propose two alternative criteria expressed in terms of χ^2 and Kullback-Leibler divergence between the proposal distribution and the target distribution for fixed sample size. Not only are they asymptotically equivalent to the two former criteria (as they converge to the same quantities), they also fit in a function-free risk analysis for a fixed sample size. Finally, we outline that these two criterion allow for much easier optimization. Specifically, the criterion based on the Kullback-Leibler divergence slits the adaptation of the proposal distribution in two distinct and independent subproblems: adaptation of the adjustment multiplier weights of the APF, and adaptation of the proposal kernel – the treatment of these two subproblems is the core of the remaining of this thesis, Chapters 4 and 5 respectively, which ally theoretical study and practical algorithm design. Before that, in the end of this chapter, we lay the first blocks of an algorithm to solve the latter, outlining the resemblance with problems routinely solved by means of Expectation-Maximization algorithm (*EM*, Dempster et al. (1977)), its Monte Carlo (*MCEM*, Fort and Moulines (2003)) and Stochastic Approximation (*SAEM*, Delyon et al. (1999); Kuhn and Lavielle (2004); Andrieu et al. (2005)) variants, and the Cross Entropy (*CE*, Rubinstein and Kroese (2004)). This approach is only a sketch, its full explanation being the core of Chapter 5.

Chapter 4, relying on the criteria proposed in Chapter 3, treats of adapting the multiplier adjustment weights of the FPA, which are too often negelected even though they add to the proposal distribution a degree of freedom which is required to hope fitting exactly the target distribution – and hence to increase the performance of SMC methods. We start by giving a precise analysis of the *pilot-sampling* algorithm (Zhang and Liu, 2002) in terms of concistency and asymptotic normality. This algorithm, built upon empirical heuristics, has never been studied theoretically, to the best of our knowledge. This first study then makes room for a new algorithm which includes a pilot exploration step followed by a "refueling" of the sample by new particles to ensure that the quality criteria proposed in Chapter 3 are kept below a chosen threshold. Proofs of convergence

of this algorithm are also given.

Chapter 5 provides a new algorithm to the adapt the proposal kernel of SMC methods. We expose it in the APF framework for greater generality, but its application to bootstrap filter is straightforward. We stress the capacity of this algorithm to fit optimal kernels far more complicated than those allowed by existing methods such as Laplacian approximation ofPitt and Shephard (1999), Extended Kalman particle filter of Doucet et al. (2001) , or Unscented Kalman particle filter of Van der Merwe et al. (2000), while still being of reasonable computational cost. Final remarks of Chapter 3 are then grown up to full maturity as we fit a family of *mixture of experts* very close to the works of Jordan and Jacobs (1994); Jordan and Xu (1995). This family consists in a mixture of kernels belonging to the broad class of *integrated* curved exponential family, from which we exhibit the two most famous examples that are multivariate Gaussian and multivariate t-Student distributions. The weights of this mixture are given by a logistic regression that partitions the space of the original particles with a soft-max function, specializing each kernel for one or several subregions. The algorithms of this chapter are inspired by Stochastic Approximation EM (*SAEM*, Delyon et al. (1999); Kuhn and Lavielle (2004); Andrieu et al. (2005)) as well as the aforementioned MCEM and CE.

Notation

Particles

The following quantities, representing the particles and their weights, are the key notation of this dissertation. We use the generic convention for optional indices and exponents: that $\bullet_{N,i}^{(k),[\ell]}$ is the i^{th} quantity $\bullet$ at time k, iteration ℓ of the algorithm, with an emphazis on the asymptotic analysis in the original number N of particles.

$\boldsymbol{\Xi}$	Space of the current particles
$\tilde{\boldsymbol{\Xi}}$	Space of the intermediate particles
ξ_i	i^{th} particle at a fixed timestep or non-sequentially
$\xi_i^{(k)}$	i^{th} particle at time k
$\xi_{N,i}^{(k)}$	i^{th} particle number at time k, emphasis on the triangular array basing the asymptotic analysis in the original number N of particles
$\xi_i^{(0:k)}$	Trajectory from time 0 to k of particle $\xi_i^{(k)}$
$\xi_i^{(0:k)}(l)$	Component at time l of the trajectory of particle $\xi_i^{(k)}$
$I_i^{(k)}$	Index of the ancestor at time $k-1$ of particle $\xi_i^{(k)}$, that is such that $\xi_{I_i^{(k)}}^{(k-1)} = \xi_i^{(0:k)}(k-1)$
$\tilde{\xi}_i$	i^{th} intermediate particle – often used pairwise with ξ_i when analysing a single step of a SMC algorithm, hence dropping the time index k
$\omega_i^{(k)}$	Importance weight of particle $\xi_i^{(k)}$
$\tilde{\omega}_i$	Importance weight of the intermediate particle $\tilde{\xi}_i$
$\Omega^{(k)}$	$\sum_{i=1}^{N} \omega_i^{(k)}$
$\tilde{\Omega}$	$\sum_{i=1}^{M_N} \tilde{\omega}_i$

The update and proposal kernel, key ingredient of the SMC methods, will be denoted as follows.

L_k	Unnormalized update kernel at time k
l_k	Density function of L_k
L_k^*	Normalized optimal kernel at time k

l_k^*	Density function of L^*
$L_{\mathrm{p},k}$	Pathwise version of L_k
L_{p}^*	Density function of $L_{\mathrm{p},k}$
R_k	Proposal (hence Markovian) kernel at timestep k
r_k	Density function of R_k
$R_{[}^{\mathrm{p}}k]$	Pathwise proposal kernel (hence transition kernel) at timestep k

The auxiliary particle filter needs the following specific notation.

$\psi_i^{(k)}$	Adjustment multiplier weight associated to particle $\xi_i^{(k)}$
$\Psi^{(k)}$	Adjustment function, such that $\psi_i^{(k)} = \Psi^{(k)}(\xi_i^{(k)})$
$\psi^{*}{}_i^{(k)}$	Optimal adustment multiplier weight
$\Psi^{*(k)}$	Optimal adustment function
$\Phi(\xi, \tilde{\xi})$	Weighting function $\frac{1}{\Psi(\xi)} \frac{dL(\xi,\cdot)}{dR(\xi,\cdot)}(\tilde{\xi})$
$\hat{\xi}_i^{(0:k)}$	i^{th} intermediate particle in APF asymptotic analysis
$J_i^{(k)}$	Same as $I_i^{(k)}$, for the optional second resampling stage of Algorithm 2.2.1
$\mu_{\text{aux}}(i, \tilde{\xi})$	Auxiliary target distribution
$\pi_{\text{aux}}(i, \tilde{\xi})$	Auxiliary proposal distribution
$\pi_{\text{aux}}^{\boldsymbol{\theta}}(i, \tilde{\xi})$	Parameterized auxiliary proposal distribution

Hidden Markov models and state space models

$(\mathsf{X}, \mathcal{X})$	Measurable state space at one timestep
$(\mathsf{Y}, \mathcal{Y})$	Measurable space of the observations
$\{X_k\}_{k\geq 0}$	Markov chain of the hidden states
$\{Y_k\}_{k\geq 0}$	Stochastic process of the observations
$\{U_k\}_{k\geq 0}$	Dynamic noise process in the state-space formulation
$\{V_k\}_{k\geq 0}$	Observation noice process in the state-space formulation
χ	Distribution of X_0
$Q^{(k)}(x, dx')$	Prior kernel at time k, i.e. Markovian kernel of $\{X_k\}_{k\geq 0}$
$q^{(k)}(x, x')$	Density transition function of $Q^{(k)}$ w.r.t. measure λ
$G^{(k)}(x, dy)$	Observation kernel at time k
$g^{(k)}(x, y)$	Density transition function of $G^{(k)}$ w.r.t. measure μ
$g_k(x)$	Local likelihood $g^{(k)}(x, Y_k)$ – implicit conditioning
$T^{(k)}\left((x, y), (dx', dy')\right)$	Markovian kernel of the joint Markov Chain $\{(X_k, Y_k)\}_{k\geq 0}$ at time k
$t^{(k)}\left((x, y), (x', y')\right)$	Density transition function of $T^{(k)}$
L_k	The unnormalized update kernel at time k is $g_k Q$
$\phi_{\chi,k:l\|n}$	Distribution of $X_{k:l}$ conditionally on $Y_{0:n}$

$\phi_{\chi,n}$	Filtering distribution, i.e. $\phi_{\chi,n\|n}$

Generic mathematics

$:=$	LHS defined as RHS
$\equiv$	Equality up to an irrelevant (multiplicative or additive) constant – see precise definition p. 150
$\lfloor x \rfloor$	Integer part of x
$\langle x \rangle$	Fractional part of x, i.e. $x - \lfloor x \rfloor$
A^T	Transpose of matrix or vector A
A^{-1}	Inverse of matrix A, if it exists
A^2	$A\,A^T$ for any matrix or vector A
$\operatorname{chol} A$	Upper triangular Cholesky decomposition of matrix A
$x_{j:k}$	$(x_j, x_{j+1}, \dots, x_k)$, for any sequence $(x_n)_{n\in\mathbb{N}}$ and non-negative integers $j \leq k$.
$\overline{x}$	Vector $(x^T, 1)^T$, with vector $x \in \mathbb{R}^d$
$\xrightarrow{\mathbb{P}}$	Convergence in probability
$\xrightarrow{\mathcal{D}}$	Convergence in distribution

Measure-theoretic

We consider a measure space $(\Xi, \mathcal{B}(\Xi), \mu)$, K a kernel from $(\Xi, \mathcal{B}(\Xi))$ to a measurable space $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$, f_1, f_2, g, and h measurable functions on $(\Xi, \mathcal{B}(\Xi))$, $(\Xi, \mathcal{B}(\Xi))$, $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$, and $(\Xi \times \tilde{\Xi}, \mathcal{B}\Xi \otimes \mathcal{B}\tilde{\Xi})$, respectively, such that the integrals below are well defined.

$\mathcal{B}(\Xi)$	σ-algebra of the Borelians of Ξ
$\mathcal{P}(\Xi)$	Set of probability measures on $(\Xi, \mathcal{B}(\Xi))$
$\mathbb{B}(\Xi)$	Set of measurable functions from $(\Xi, \mathcal{B}(\Xi)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$
$\mathcal{F}_\mathrm{b}(\Xi)$	Set of measurable bounded functions from $(\Xi, \mathcal{B}(\Xi)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$
$\mathsf{L}^p(A, \mu)$	Set of function whose p^{th} power is μ-integrable on the set A and its Borelians
$\mu(f)$	$\int_\Xi f(\xi)\,\mu(\mathrm{d}\xi)$
finite kernel	$K(\xi, \tilde{\Xi}) < \infty$ for all $\xi \in \Xi$
transition kernel	$K(\xi, \tilde{\Xi}) = 1$ for all $\xi \in \Xi$
Markovian kernel	Transition kernel from $(\Xi, \mathcal{B}(\Xi))$ to itself
$K(\xi, g)$	$\int_{\tilde{\Xi}} g(\tilde{\xi})\,K(\xi, d\tilde{\xi})$
$\mu K(g)$	$\int_\Xi K(\xi, g)\,\nu(\mathrm{d}\xi)$
$\mu \otimes K(h)$	Outer product, $\iint_{\Xi\times\tilde{\Xi}} \mu(d\xi)\,K(\xi, d\tilde{\xi})h(\xi, \tilde{\xi})$
$\mu[f_1](f_2)$	Modulated measure, $\nu(f_1 \times f_2)$

Note that the product and outer product of a measure and a kernel are trivially generalized as the product and outer product of two kernels, as for any $\xi \in \Xi$, $K(\xi, \cdot)$ is

a measure. For a sequence $\{K_\ell\}_{\ell=m}^n$ of kernels, $\bigotimes_{\ell=m}^n K_\ell := K_m \otimes K_{m+1} \otimes \cdots \otimes K_n$ by recursive application of the outer product.

Divergences and quality criteria

CV^2	Coefficient of variation of the weights
ESS	Effective sample size
d_{χ^2}	Chi-square divergence
$\mathcal{E}$	Negated Shannon entropy of the weights
d_{KL}	Kullback Leibler divergence

Common distributions

Any distribution p with parameter $\theta_1, \ldots, \theta_n$ admits two writings: $p(\theta_1, \ldots, \theta_n)$ denotes the distribution itself, $p(x; \theta_1, \ldots, \theta_n)$ is its mass (or the density, for continuous distributions) evaluated in point x.

Ber	Bernoulli distribution
Bin	Binomial distribution
γ	Gamma distribution
Mult	Multinomial distribution
$\mathcal{N}$	Gaussian distribution
Pois	Poisson distribution
t	t-Student distribution

Abbreviations

AMW	Adjustment Multiplier Weight
AN	Asymptotically Normal
APF	Auxiliary Particle Filter
CE	Cross Entropy
CLD	Chi-Square Divergence
CV	Coefficient of Variation
EKF	Extended Kalman Filter
EM	Expectation Maximization
EPS	Échantillonnage Préférentiel Séquentiel
EPSR	Échantillonnage Préférentiel Séquentiel avec Rééchantillonnage
ESS	Effective Sample Size
FKE	Filtre de Kalman Étendu
FPA	Filtre Particulaire Auxiliaire
HDP	plus Haute Densité à Posteriori

HMM	Hidden Markov Model
KLD	Kullback-Leibler Divergence
MAS	Mutation with Adaptive Selection
MCEM	Monte Carlo Expectation Maximization
MCMC	Markov Chain Monte Carlo
MCS	Monte Carlo Séquentielles
MKF	Mixture Kalman Filter
MMC	Modèle de Markov Caché
MPEAR	Mutation with pilot exploration and adaptive refueling
PMC	Population Monte Carlo
SAEM	Stochastic Approximation Expectation Maximization
SIAS	Sequential Importance Sampling with Adaptive Selection
SIS	Sequential Importance Sampling
SISAR	Sequential Importance Sampling with Adaptive Refueling
SISR	Sequential Importance Sampling with Resampling
SMC	Sequential Monte Carlo
TEE	Taille d'Échantillon Effective
UKF	Unscented Kalman Filter
cf.	*confere*
e.g.	*examplo gratia*
i.e.	*id est*
i.i.d.	independent and identically distributed
w.r.t.	with respect to

CHAPITRE 1

Méthodes de Monte Carlo séquentielles

Sommaire

L'utilisation des méthodes de Monte Carlo séquentielles (MCS) pour le filtrage non-linéaire remonte aux travaux fondateurs de Handschin and Mayne (1969) et Handschin (1970). Ces premières tentatives se basaient sur des versions séquentielles de l'*échantillonnage préférentiel*, qui consiste à simuler des réalisations selon une distribution instrumentale et à approcher ensuite les lois cibles en pondérant ces réalisations par des *poids d'importance* correctement définis. Dans le cadre du filtrage non-linéaire, les algorithmes d'échantillonnage préférentiel peuvent être implémentés séquentiellement au sens où, en définissant avec soin une suite de distributions instrumentales, il n'est pas nécessaire de simuler une nouvelle population de réalisations depuis le début à chaque arrivée d'une nouvelle observation. Cet algorithme est appelé *échantillonnage préférentiel séquentiel* (*EPS – SIS* en anglais). Bien que l'algorithme EPS soit connu depuis le début des années 1970, son utilisation pour des problèmes de filtrage non-linéaire était à l'époque plutôt limité. Vraisemblablement, les capacités de calcul

informatique étaient alors trop limitées pour permettre des applications convaincantes de ces méthodes. Une autre raison moins évidente est que l'algorithme EPS souffre d'un problème majeur qui n'a été clairement identifié et correctement traité qu'à partir de l'article précurseur Gordon et al. (1993). Lorsque le nombre d'itérations croît, les poids d'importance ont tendance à dégénérer, un phénomène connu sous le nom d'*appauvrissement de l'échantillon* ou *dégénérescence des poids*. Sommairement, sur le long terme, la plupart des réalisations ont un poids d'importance normalisé particulièrement faible et ne contribuent donc pas significativement à l'approximation de la distribution cible. La solution proposée par Gordon et al. (1993) est de permettre la régénération de l'échantillon en dupliquant les réalisations ayant un poids d'importance élevé, et, au contraire, en supprimant les réalisations ayant un faible poids.

Le *filtre particulaire* de Gordon et al. (1993) fut la première application couronnée de succès des techniques de Monte Carlo séquentielles au domaine du filtrage non-linéaire. Depuis lors, les méthodes de Monte Carlo séquentielles (*MCS – SMC* en anglais) ont été appliquées dans de nombreux domaines, incluant la vision par ordinateur, le traitement du signal, le contrôle, l'économétrie, la finance, la robotique, et la statistique (Doucet et al., 2001; Ristic et al., 2004). Ce chapitre introductif récapitule les briques fondamentales qui sont nécessaires à l'implémentation d'un algorithme de Monte Carlo séquentiel, en commençant par les concepts de l'échantillonnage préférentiel.

1.1 Échantillonnage préférentiel et rééchantillonnage

1.1.1 Échantillonnage préférentiel

L'échantillonnage préférentiel (*EP – IS* en anglais) remonte au moins à Hammersley and Handscomb (1965), et est couramment utilisé dans plusieurs domaines (pour des références générales sur l'échantillonnage préférentiel, voir Glynn and Iglehart, 1989, Geweke, 1989, Evans and Swartz, 1995, or Robert and Casella, 2004).

Tout au long de cette section, nous noterons μ une mesure de probabilité d'intérêt sur un espace mesurable $(\mathsf{X}, \mathcal{X})$, que nous appellerons la *distribution cible*. Le but est d'approcher des intégrales de la forme $\mu(f) = \int_{\mathsf{X}} f(x)\,\mu(dx)$ pour des fonctions f à valeurs réelles et mesurables. L'approche Monte Carlo élémentaire consiste à simuler un échantillon i.i.d. $\{\xi_1, \ldots, \xi_N\}$ selon la mesure de probabilité μ puis à évaluer la moyenne $N^{-1}\sum_{i=1}^{N} f(\xi_i)$. Bien sûr, cette technique n'est applicable que quand il est possible (et raisonnablement simple) de simuler selon la loi cible μ.

L'échantillonnage préférentiel est basé sur l'idée que, dans certaines situations, il est plus approprié de simuler selon une *distribution instrumentale* ν, puis d'appliquer une formule de changement de mesure afin de prendre en compte le fait que la distribution instrumentale diffère de la distribution cible. Plus rigoureusement, supposons que la mesure de probabilité cible μ est absolument continue par rapport à une *mesure de probabilité instrumentale* ν selon laquelle il est facile de simuler. Notons $d\mu/d\nu$ la dérivée de Radon-Nikodym μ par rapport à ν. Alors pour toute fonction f μ-intégrable,

$$\mu(f) = \int f(x)\,\mu(dx) = \int f(x)\,\frac{d\mu}{d\nu}(x)\,\nu(dx)\;. \tag{1.1.1}$$

En particulier, si $\xi_1, \xi_2, \ldots$ est un échantillon i.i.d. selon ν, (1.1.1) suggère l'estimateur de $\mu(f)$ suivant :

$$\tilde{\mu}^{\mathrm{EP}}_{\nu,N}(f) = N^{-1}\sum_{i=1}^{N} f(\xi_i)\frac{d\mu}{d\nu}(\xi_i)\;. \tag{1.1.2}$$

Cet estimateur étant la moyenne de l'échantillon de variables aléatoires indépendantes, une vaste gamme de résultats est disponible pour évaluer la qualité de $\tilde{\mu}^{\mathrm{EP}}_{\nu,N}(f)$ en tant qu'estimateur de $\mu(f)$. Tout d'abord, la loi forte des grands nombres implique que $\tilde{\mu}^{\mathrm{EP}}_{\nu,N}(f)$ converge vers $\mu(f)$ presque sûrement quand N tend vers l'infini. De plus, le théorème central limite pour les variables i.i.d. (ou les inégalités de déviation) peuvent guider le choix de la distribution de proposition ν, au-delà de la condition évidemment requise qu'elle domine la distribution cible μ.

Dans de nombreuses situations, la mesure de probabilité cible μ ou la mesure de probabilité instrumentale ν est seulement connue à un facteur de normalisation près. Ceci est particulièrement vrai dans les applications de l'échantillonnage préférentiel en statistique Bayésienne. La dérivée de Radon-Nikodym $d\mu/d\nu$ est alors seulement connue à un facteur (constant) multiplicatif près. Il est toutefois toujours possible d'utiliser l'échantillonnage préférentiel dans ce cas, en adoptant la forme auto-normalisée de l'estimateur d'échantillonnage préférentiel

$$\widehat{\mu}^{\mathrm{EP}}_{\nu,N}(f) = \frac{\sum_{i=1}^{N} f(\xi_i)\frac{d\mu}{d\nu}(\xi_i)}{\sum_{i=1}^{N}\frac{d\mu}{d\nu}(\xi_i)} \,. \tag{1.1.3}$$

Cette quantité ne comporte évidemment aucun facteur multiplicatif en $d\mu/d\nu$. L'estimateur d'échantillonnage préférentiel auto-normalisé $\widehat{\mu}^{\mathrm{EP}}_{\nu,N}(f)$ est défini comme le rapport des moyennes sur l'échantillon des fonctions $f_1 = f \times (d\mu/d\nu)$ et $f_2 = d\mu/d\nu$. La loi forte des grands nombres implique donc que $N^{-1}\sum_{i=1}^{N} f_1(\xi_i)$ et $N^{-1}\sum_{i=1}^{N} f_2(\xi_i)$ convergent presque sûrement vers $\mu(f_1)$ et $\nu(d\mu/d\nu) = 1$, respectivement, montrant que $\widehat{\mu}^{\mathrm{EP}}_{\nu,N}(f)$ est un estimateur consistant de $\mu(f)$. Par la suite, le terme *échantillonnage préférentiel* fera référence à la forme auto-normalisée (1.1.3) de l'estimateur d'échantillonnage préférentiel.

1.1.2 Échantillonnage préférentiel avec rééchantillonnage

Bien que l'échantillonnage préférentiel a pour but premier de surmonter la difficulté de simuler sous μ lors de l'approximation d'intégrales de la forme $\mu(f)$, il peut également être utilisé pour simuler (approximativement) selon la distribution μ. Cette dernière opération peut être effectuée par l'*échantillonnage préférentiel avec rééchantillonnage* (*EPR* – SIR en anglais) introduit dans Rubin (1987, 1988). L'échantillonnage préférentiel avec rééchantillonnage est une procédure en deux étapes dans laquelle l'échantillonnage préférentiel décrit ci-dessous est suivi par une étape supplémentaire de tirage aléatoire. Dans la première étape, un échantillon i.i.d. $\{\tilde{\xi}_1, \ldots, \tilde{\xi}_M\}$ est simulé selon la distribution instrumentale ν, et l'on calcule la version normalisée des poids d'importance,

$$\omega_i = \frac{\frac{d\mu}{d\nu}(\tilde{\xi}_i)}{\sum_{i=1}^{M}\frac{d\mu}{d\nu}(\tilde{\xi}_i)} \,, \quad i = 1, \ldots, M \,. \tag{1.1.4}$$

Dans la deuxième étape, l'étape de rééchantillonnage, un échantillon de taille N dénoté $\{\xi_1, \ldots, \xi_N\}$ est de nouveau tiré avec remise parmi l'ensemble de points intermédiaire $\{\tilde{\xi}_1, \ldots, \tilde{\xi}_M\}$, en prenant en compte les poids calculés en (1.1.4). L'idée sous-jacente est que les points $\tilde{\xi}_i$ pour lesquels le poids ω_i dans (1.1.4) est grand sont plus vraisemblables sous la distribution cible μ, et devraient donc être sélectionnés pendant l'étape de rééchantillonnage avec une probabilité plus grande que les points avec un faible poids (normalisé). Le principe est illustré dans la Figure 1.1.

Il y a plusieurs façons d'implémenter cette idée de base, l'approche la plus évidente consistant à tirer avec remise avec une probabilité de tirer chaque ξ_i égale au poids

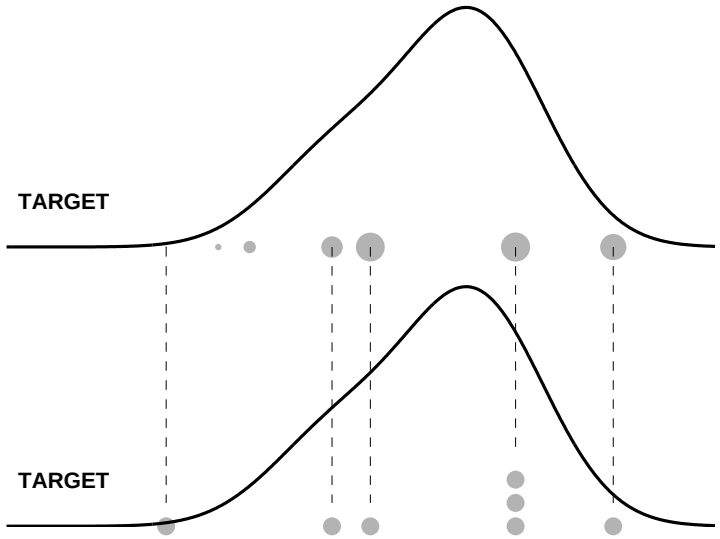


FIGURE 1.1 – Principe du rééchantillonnage. Graphique du haut : l'échantillon tiré selon ν avec les poids d'importance normalisés associés représentés par des disques de rayon proportionnel au poids (la densité cible correspondant à μ est représentée en trait continu). Graphique du bas : après rééchantillonnage, tous les points ont le même poids d'importance, et certains d'entre eux ont été dupliqués ($M = N = 7$).

d'importance ω_i. Ainsi, le nombre de fois N^i dont chaque point particulier $\tilde{\xi}_i$ de l'échantillon de la première étape est sélectionné suit une distribution binomiale $\mathrm{Bin}(N, \omega_i)$. Le vecteur $(N^1, \ldots, N^M)$ est distribué selon $\mathrm{Mult}(N, \omega_1, \ldots, \omega_M)$, la distribution multinomiale avec paramètre N et probabilités de succès $(\omega_1, \ldots, \omega_M)$. Dans cette étape de rééchantillonnage, les poids de la première étape qui sont associés avec de faibles poids d'importance normalisés sont les plus susceptibles d'être défaussés, tandis que les meilleurs points de l'échantillon sont dupliqués proportionnellement à leur poids d'importance. Dans la plupart des applications, M, la taille de l'échantillon de première étape, est typiquement choisi plus grand (et parfois beaucoup plus grand) que N. L'algorithme EPR est résumé ci-dessous.

L'ensemble $(I_1, \ldots, I_N)$ est donc un processus de tirage multinomial et le vecteur des comptes $(N^1, \ldots, N^M)$, où

$$N^i := \sum_{j=1}^{N} \mathbb{1}_{\{I_j = i\}}, \qquad i = 1, \ldots, N \,, \tag{1.1.6}$$

suit la loi multinomiale de paramètre N et de probabilités de succès $(\omega_1, \ldots, \omega_M)$. Ainsi, cette méthode de sélection est appelée schéma de rééchantillonnage *multinomial*. Nous verrons dans la Section 1.5.2 que d'autres schémas de rééchantillonnage existent et présentent certains avantages.

Remarque 1.1.1. Dans la formulation originelle de l'algorithme EPSR, Rubin (1987) suggérait de simuler un échantillon de deuxième étape sans remise, c'est à dire, tirer d'abord un point de $\{\tilde{\xi}_i\}_{1 \leq i \leq M}$ avec probabilités proportionnelles à $\{\frac{d\mu}{d\nu}(\tilde{\xi}_i)\}_{1 \leq i \leq M}$, puis tirer un second point de l'ensemble des $M - 1$ valeurs restantes avec probabilités toujours proportionnelles à $\{\frac{d\mu}{d\nu}(\tilde{\xi}_i)\}_{1 \leq i \leq M}$ et ainsi de suite. C'est bien sûr une idée qui n'a de sens que dans les situations où M est bien plus grand que N. Elle produit toutefois

Algorithme 1.1.1 EPR : Échantillonnage Préférentiel avec Rééchantillonnage

Échantillonnage : Simuler un échantillon i.i.d. $\tilde{\xi}_1, \ldots, \tilde{\xi}_M$ selon la distribution instrumentale ν.

Pondération : Calculer les poids d'importance (normalisés)

$$\omega_i = \frac{\frac{d\mu}{d\nu}(\tilde{\xi}_i)}{\sum_{j=1}^{M} \frac{d\mu}{d\nu}(\tilde{\xi}_j)} \quad \text{for } i = 1, \ldots, M \ .$$

Rééchantillonnage :

- Simuler, indépendamment conditionnellement à $(\tilde{\xi}_1, \ldots, \tilde{\xi}_M)$, N variables aléatoires discrètes $I_1, \ldots, I_N$ selon la loi discrète sur $\{1, \ldots, M\}$ de probabilités $(\omega_1, \ldots, \omega_M)$, i.e.
$$\mathbb{P}(I_1 = j) = \omega_j \ , \quad j = 1, \ldots, M \ . \tag{1.1.5}$$
- Poser, pour $i = 1, \ldots, N$, $\xi_i = \tilde{\xi}_{I_i}$.

un échantillon (approché) d'une distribution qui n'est ni ν ni μ, mais un mélange des deux.

Remarque 1.1.2. Un autre point intéressant à noter est que, à des fins de simulation, même lorsque la dérivée de Radon-Nikodym est connue exactement (c'est à dire avec sa constante de normalisation), les poids renormalisés (1.1.4) sont toujours requis afin que la somme des poids soit égale à 1.

Jusqu'ici, il ne semble peut-être pas évident que l'échantillon $\{\xi_1, \ldots, \xi_N\}$ fourni par l'Algorithme 1.1.1 est effectivement (approximativement) i.i.d. selon μ dans quelque sens utilisable que ce soit. Néanmoins, il peut être prouvé que la moyenne de l'échantillon obtenu avec l'algorithme EPR,

$$\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi_i) \ , \tag{1.1.7}$$

est un estimateur consistante de $\mu(f)$ pour toute fonction f satisfaisant $\mu(|f|) < \infty$. L'étape de rééchantillonnage peut ainsi être vue comme un moyen de transformer l'estimateur d'échantillonnage préférentiel pondéré $\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f)$ défini par (1.1.3) en une moyenne d'un échantillon non pondéré. Rappelons que N^i est le nombre de fois où l'élément $\tilde{\xi}_i$ est resélectionné. En réécrivant

$$\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f) = \frac{1}{N} \sum_{i=1}^{N} f(\xi_i) = \sum_{i=1}^{M} \frac{N^i}{N} f(\tilde{\xi}_i) \ ,$$

on voit facilement que la moyenne $\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)$ de l'échantillon EPR est, conditionnellement à l'échantillon de première étape $\{\tilde{\xi}_1, \ldots, \tilde{\xi}_M\}$, égale à l'estimateur par échantillonnage préférentiel $\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f)$ défini en (1.1.3),

$$\mathbb{E}\left[\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f) \,\middle|\, \tilde{\xi}_1, \ldots, \tilde{\xi}_M\right] = \widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f) \ .$$

Par conséquent, l'estimateur EPR $\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)$ est un estimateur sans biais de $\mu(f)$, mais son erreur quadratique intégrée est toujours plus grande que celle de l'estimateur

d'échantillonnage préférentiel $\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f)$ défini en (1.1.3), suite à la décomposition de la variance bien connue

$$\mathbb{E}\left[\left(\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)-\mu(f)\right)^2\right] = \mathbb{E}\left[\left(\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)-\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f)\right)^2\right]+\mathbb{E}\left[\left(\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f)-\mu(f)\right)^2\right].$$

La variance $\mathbb{E}[(\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)-\widehat{\mu}^{\mathrm{EP}}_{\nu,M}(f))^2]$ peut être interprétée comme le prix à payer pour convertir l'estimateur pondéré de l'échantillonnage préférentiel en une estimation non pondérée.

Montrer que l'estimateur EPR $\hat{\mu}^{\mathrm{EPR}}_{\nu,M,N}(f)$ défini en (1.1.7) est un estimateur consistant et asymptotiquement normal de $\mu(f)$ n'est pas trivial, car $\xi_1, \ldots, \xi_N$ ne sont pas indépendants, à cause de la normalisation des poids suivie du rééchantillonnage. Ainsi, les résultats élémentaires de convergence i.i.d. qui sous-tendent la théorie de l'échantillonnage d'importance ne sont d'aucun secours, et nous renvoyons à l'Annexe A ainsi qu'à l'article Douc and Moulines (2008) pour les preuves correspondantes.

Remarque 1.1.3. Un examen approfondi de la complexité numérique de l'Algorithme 1.1.1 révèle que, bien que toutes les étapes de l'algorithme aient une complexité qui croît proportionnellement à M et N, ceci n'est pas tout à fait vrai pour l'étape de tirage multinomial dont la complexité croît, *a priori*, plus vite que N (environ $N \log_2 M$—voir Section 1.5.1 ci-dessous pour plus de détails). Ceci est particulièrement malheureux, puisque les méthodes de Monte Carlo sont généralement plus utiles quand N est grand (ou plus précisément puisque la qualité de l'approximation s'améliore de plus en plus lentement quand N croît).

Un usage astucieux de résultats probabilistes élémentaires rend toutefois possible la mise au point de méthodes pour simuler N fois selon une même loi discrète avec M issues possibles en utilisant un nombre d'opérations qui ne croît que linéairement avec le maximum de N et M. Afin de ne pas interrompre notre exposé des méthodes de Monte Carlo séquentielles, les algorithmes correspondant sont présentés dans la Section 1.5.1 à la fin de ce chapitre. Notons que nous ne mentionnons ici que des problèmes d'implémentation. Il y a toutefois différentes raisons, abordées dans la Section 1.5.2, pour adopter d'autres schémas de rééchantillonnage que le rééchantillonnage multinomial.

1.2 Problèmes séquentiels et modèles de markov cachés

Nous avons jusqu'à présent exposé l'échantillonnage préférentiel dans un cadre non-séquentiel, c'est à dire où la distribution cible μ comme la distribution de proposition ν sont quelconques et ne présentent pas de structure hiérarchique particulière. De nombreux problèmes actuels exhibent toutefois une structure séquentielle intrinsèque, c'est à dire consistent à approcher non pas une mesure μ mais une suite de mesures $(\mu_k)_{k\in\mathbb{N}}$ décrites par certaines relations de récurrences. De tels problèmes trouvent leur origine dans les questions de filtrage statistique pour des modèles à espace d'état – la littérature sur ces modèles remonte au moins à Kalman and Bucy (1961) qui décrit les premiers algorithmes pour des modèles linéaires gaussiens. Au cours des dernières décennies, la classe des problèmes considérés s'est considérablement élargie, et, devant la complexité de certains modèles parmi lesquels le filtrage de modèles non-linéaires non-gaussiens, l'inférence exacte a cédé le pas aux méthodes de Monte Carlo, notamment aux méthodes de Monte Carlo séquentielles. La performance croissante de ces méthodes permet même désormais de revenir à la simulation dans le cadre non-séquentiel

de mesures particulièrement compliquées, pour lesquelles une suite artificielle de distributions intermédiaires est construite – voir notamment les *SMC samplers* de Del Moral et al. (2006).

Dans cette section, nous introduisons le cadre théorique des *modèles de Markov cachés* (*MMC – HMM* en anglais), qui sont à la base du traitement mathématique des problèmes mentionnés ci-dessus. Après avoir donné les définitions formelles, nous présenterons brièvement les distributions dites de *lissage* et de *filtrage* associées, qui ont motivé le développement de ces modèles, et les relations de récurrence qui les lient. Ceci nous permettra alors, à partir de la Section 1.3, de voir comment généraliser les algorithmes d'échantillonnage préférentiel pour approcher lesdites distributions.

1.2.1 Définitions

Un modèle de Markov caché est souvent décrit comme une chaîne de Markov observée de façon bruitée. Ce modèle comporte en effet une chaîne de Markov, que nous noterons $\{X_k\}_{k\geq 0}$, où k est un indice entier naturel. Cette chaîne de Markov peut être à valeur dans un espace relativement arbitraire, sans restriction quant à sa dénombrabilité. Cette chaîne est *cachée*, en ce sens qu'elle n'est pas observable directement, mais uniquement au travers d'un autre processus stochastique $\{Y_k\}_{k\geq 0}$. Ce dernier est lié à la chaîne de Markov cachée par le fait que X_k gouverne la distribution du Y_k correspondant. La chaîne $\{X_k\}_{k\geq 0}$ est parfois appellée l'*état*.

Un MMC est donc un processus bivarié $\{(X_k, Y_k)\}_{k\geq 0}$ à temps discret, tel que

- le processus $\{X_k\}_{k\geq 0}$ est une chaîne de Markov de noyau de transition Q et de distribution initiale χ ;
- conditionnellement au processus d'état $\{X_k\}_{k\geq 0}$, les variables aléatoires $\{Y_k\}_{k\geq 0}$ sont indépendantes, et pour tout k la distribution conditionnelle de Y_k ne dépend que de X_k.

Nous noterons X l'espace dans lequel les variables aléatoires X_k prennent leurs valeurs. Il est possible d'établir un cadre plus général, dans lequel la variable aléatoire X_k prend ses valeurs dans un espace $\mathsf{X}^{(k)}$, et de noter $\mathsf{X}^{(j:k)} = \mathsf{X}^{(j)} \times \ldots \times \mathsf{X}^{(k)}$ pour tous entiers $j \leq k$. Dans la plupart des cas toutefois, l'espace d'état $\mathsf{X}^{(k)}$ ne dépend pas de l'instant k et l'on peut donc s'en tenir sans perte de généralité au cas $\mathsf{X}^{(k)} = \mathsf{X}$ quel que soit k, et donc où $\mathsf{X}^{(j:k)} := \mathsf{X}^{k-j+1}$ pour tous entiers $j \leq k$. De la même façon, nous noterons Y l'espace dans lequel les variables aléatoires Y_k prennent leurs valeurs, bien qu'il eut été également possible de considérer une suite d'espaces $\mathsf{Y}^{(k)}$. Par ailleurs, outre la simplification des notations, raisonner à espaces homogènes dans le temps permet de respecter la définition classique d'une chaîne de Markov telle que donnée dans (Meyn and Tweedie, 1994, Section 3.1).

Un MMC est donc un processus doublement stochastique, avec un processus stochastique sous-jacent qui n'est pas directement observé (il est "caché") mais peut être observé indirectement à travers un autre processus stochastique qui produit la suite d'observation. Un MMC est souvent défini sous sa forme de *modèle à espace d'état*,

$$\begin{cases} X_{k+1} = a_k(X_k, U_k) \ , \\ Y_k = b_k(X_k, V_k) \ , \end{cases}$$

avec $X_0 \sim \chi$, où $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des suites de variables aléatoires mutuellement indépendantes i.i.d. qui sont indépendentes de X_0, et $\{a_k\}_{k\geq 0}$ et $\{b_k\}_{k\geq 0}$ sont des suites de fonctions mesurables. La première équation est appelée *équation d'état* ou *équation dynamique*, tandis que la seconde est l'*équation d'observation*. Ces deux équations sont

une formulation générative récursive du modèle, contrairement à notre première présentation en termes de distributions jointes des processus $\{X_k\}_{k\geq 0}$ et $\{Y_k\}_{k\geq 0}$. Savoir laquelle de ces deux approches équivalentes est la plus naturelle dépend entièrement de l'application considérée, selon ce que le MMC modélise. La portée de tels modèles est vaste et permet de traiter des problèmes extrêmement variés (voir par exemple Cappé et al. (2005, Section 1.3) pour un bref aperçu). Le cadre formel permettant d'embrasser cette diversité est le suivant.

La notion de dépendance conditionnelle étant délicate à définir mathématiquement dans les espaces les plus généraux (en particulier quand l'espace d'état X n'est pas dénombrable), nous définissons un MMC comme une chaîne de Markov bivariée, observée partiellement, dont le noyau de transition possède une structure particulière. En effet, ce dernier doit être tel que le processus joint $\{(X_k, Y_k)\}_{k\geq 0}$ *et* la chaîne marginale non-observée (ou cachée) $\{X_k\}_{k\geq 0}$ soient Markoviens. Les propriétés d'indépendance conditionnelle esquissées intuitivement ci-dessus découleront de cette définition.

Définition 1.2.1 (Modèle de Markov Caché)**.** Soit $(\mathsf{X}, \mathcal{X})$ et $(\mathsf{Y}, \mathcal{Y})$ deux espaces mesurables, et soit $\{Q^{(k)}\}_{k\geq 0}$ et $\{G^{(k)}\}_{k\geq 0}$, respectivement, une suite de noyaux Markoviens sur $(\mathsf{X}, \mathcal{X})$ et une suite de noyaux de transition de $(\mathsf{X}, \mathcal{X})$ vers $(\mathsf{Y}, \mathcal{Y})$. Considérons la suite de noyaux Markoviens définis sur l'espace produit $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ par

$$T^{(k)}\left((x,y),C\right) = \iint_C Q^{(k)}(x, dx')\, G^{(k+1)}(x', dy')\,, \qquad (x,y) \in \mathsf{X}\times\mathsf{Y},\, C \in \mathcal{X}\otimes\mathcal{Y}\,. \quad (1.2.1)$$

La chaîne de Markov $\{(X_k, Y_k)\}_{k\geq 0}$ de noyaux Markoviens $T^{(k)}$ et de distribution initiale $\chi \otimes G^{(0)}$, où χ est une mesure de probabilité sur $(\mathsf{X}, \mathcal{X})$, est appelée un *modèle de Markov Caché*.

Bien que la définition ci-dessus concerne le processus joint $\{(X_k, Y_k)\}_{k\geq 0}$, le terme *caché* se justifie dans les cas où $\{X_k\}_{k\geq 0}$ n'est pas observable. Cette définition est voulue la plus générique possible. Il y a toutefois plusieurs simplifications couramment rencontrées, notamment :

- modèles homogènes, où les noyaux de transition sont indépendants du temps,
- modèles partiellement dominés, où le noyau de transition liant l'observation courante à l'état caché courant admet une densité par rapport à une mesure de référence,
- modèles complètement dominés, où, en sus de la condition précédente, le noyau de transition Markovien de l'état caché admet lui aussi une densité par rapport à une autre mesure de référence.

De façon formelle, ces trois cas correspondent respectivement aux définitions suivantes.

Définition 1.2.2 (Modèle de Markov Caché Homogène)**.** Le modèle de la Définition 1.2.1 est dit *homogène* s'il existe deux noyaux de transition Q et G, respectivement Markovien sur $(\mathsf{X}, \mathcal{X})$ et de $(\mathsf{X}, \mathcal{X})$ vers $(\mathsf{Y}, \mathcal{Y})$, tels que pour tout $k \geq 0$, $Q^{(k)} = Q$ et $G^{(k)} = G$.

Définition 1.2.3 (Modèle de Markov Caché Partiellement Dominé)**.** Le modèle de la Définition 1.2.1 est dit *partiellement dominé* si il existe une mesure de probabilité μ sur $(\mathsf{Y}, \mathcal{Y})$ telle que pour tout $k \geq 0$, et pour tout $x \in \mathsf{X}$, $G^{(k)}(x, \cdot)$ est absolument continue par rapport à μ, $G^{(k)}(x, \cdot) \ll \mu(\cdot)$, avec pour fonction de densité de transition $g^{(k)}(x, \cdot)$. Alors, pour tout $A \in \mathcal{Y}$, $G^{(k)}(x, A) = \int_A g^{(k)}(x, y')\, \mu(dy')$ et le noyau de transition joint $T^{(k)}$ peut s'écrire comme

$$T^{(k)}\left((x,y),C\right) = \iint_C Q^{(k)}(x, dx') g^{(k+1)}(x', y')\, \mu(dy') \quad C \in \mathcal{X}\otimes\mathcal{Y}\,. \quad (1.2.2)$$

Définition 1.2.4 (Modèle de Markov Caché Entièrement Dominé)**.** Si, en plus des conditions de la Définition 1.2.3, il existe une mesure de probabilité λ sur $(\mathsf{X}, \mathcal{X})$ telle que $\chi \ll \lambda$ et, pour tout $k \geq 0$ et $x \in \mathsf{X}$, $Q^{(k)}Q(x, \cdot) \ll \lambda(\cdot)$ de fonction de densité de transition $q^{(k)}(x, \cdot)$, alors, pour tout $A \in \mathcal{X}$, $Q^{(k)}(x, A) = \int_A q^{(k)}(x, x')\,\lambda(dx')$, et le modèle est dit *entièrement dominé*. Le noyau de transition Markovien $T^{(k)}$ est alors dominé par la mesure produit $\lambda \otimes \mu$ et admet la fonction de densité de transition

$$t^{(k)}\left((x, y), (x', y')\right) := q^{(k)}(x, x')g^{(k)}(x', y') \; . \tag{1.2.3}$$

Notons que pour de tels modèles, nous réutiliserons généralement la notation χ pour représenter la *fonction de densité de probabilité* de l'état initial X_0 (par rapport à λ) plutôt que la distribution elle-même.

Nous présentons maintenant les principes du lissage et du filtrage tels qu'introduits par Baum et al. (1970), dans le cadre général des MMC. Afin d'alléger les notations, nous considérons le cas d'un MMC homogène au sens de la Définition 1.2.2, et partiellement dominé au sens de la Définition 1.2.3.

Remarque 1.2.1 (Notation Abrégée pour les Sous-Suites)*.* Afin d'alléger le propos, nous utiliserons la notation suivante. Pour toute sous-suite $(u_l, \ldots, u_m)$ d'une suite $(u_k)_{k\in\mathbb{N}}$, nous noterons

$$u_{l:m} := (u_l, \ldots, u_m) \; .$$

Par convention, $u_{k:k} := u_k$, bien que dans ce cas la notation u_k soit préférée. De même, lorsque besoin est, $u_{k:l}$ pour $k > l$ est valide et indique un ensemble nul.

1.2.2 Distribution jointe et vraisemblance

La distribution jointe des états non-observables et des observations jusqu'à l'instant n est telle que pour toute fonction f bornée et mesurable par rapport à $(\mathsf{X}^{n+1} \times \mathsf{Y}^{n+1}, \mathcal{X}^{\otimes(n+1)} \otimes \mathcal{Y}^{\otimes(n+1)})$ (ce que nous notons $f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{n+1} \times \mathsf{Y}^{n+1}\right)$),

$$\begin{aligned}\mathbb{E}_\chi[f(X_{0:n}, Y_{0:n})] = \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^{n+1}} & f(x_{0:n}, y_{0:n})\chi(dx_0)g(x_0, y_0) \\ & \times \prod_{k=1}^{n} \{Q(x_{k-1}, dx_k)g(x_k, y_k)\}\, \mu_n(dy_0, \ldots, dy_n) \; ,\end{aligned} \tag{1.2.4}$$

où μ_n est la distribution produit $\mu^{\otimes(n+1)}$ sur $(\mathsf{Y}^{n+1}, \mathcal{Y}^{\otimes(n+1)})$. En marginalisant par rapport aux variables non-observables $X_{0:n}$, nous obtenons la distribution marginale des observations uniquement,

$$\mathbb{E}_\chi[f(Y_{0:n})] = \int_{\mathsf{Y}^{n+1}} f(y_{0:n})\, \mathcal{L}_{\chi,n}(y_{0:n})\, \mu_n(dy_{0:n}) \; , \tag{1.2.5}$$

où $\mathcal{L}_{\chi,n}$ est une quantité importante que nous définissons ci-dessous et qui apparaît naturellement dans l'établissement des récursions.

Définition 1.2.5 (Vraisemblance)**.** La *vraisemblance* des observations est la fonction de densité de probabilité de $Y_{0:n}$ par rapport à μ_n définie, pour tout $y_{0:n} \in \mathsf{Y}^{n+1}$, par

$$\mathcal{L}_{\chi,n}(y_{0:n}) \; = \; \int_{\mathsf{Y}^{n+1}} \chi(dx_0)g(x_0, y_0)Q(x_0, dx_1)g(x_1, y_1)\cdots Q(x_{n-1}, dx_n)g(x_n, y_n) \; . \tag{1.2.6}$$

1.2.3 Filtrage, lissage, prédiction

Nous définissons tout d'abord ce que nous entendons par les termes *lissage*, *filtrage*, *prédiction*, avant de donnner les résultats fondamentaux qui forment le coeur des techniques d'inférence concernées par cette thèse.

Définition 1.2.6 (Lissage, Filtrage, Prédiciton)**.** Pour tous entiers positifs k, l, et n avec $l \geq k$, notons $\phi_{\chi,k:l|n}$ la distribution conditionelle de $X_{k:l}$ sachant $Y_{0:n}$, c'est à dire :

1. $\phi_{\chi,k:l|n}$ est un noyau de transition de $\mathsf{Y}^{(n+1)}$ vers $\mathsf{X}^{(l-k+1)}$:
   - pour tout ensemble $A \in \mathcal{X}^{\otimes(l-k+1)}$, la fonction $y_{0:n} \mapsto \phi_{\chi,k:l|n}(y_{0:n}, A)$ est $\mathcal{Y}^{\otimes(n+1)}$-mesurable,
   - pour toute sous-suite $y_{0:n}$, la distribution $A \mapsto \phi_{\chi,k:l|n}(y_{0:n}, A)$ est une mesure de probabilité sur $(\mathsf{X}^{l-k+1}, \mathcal{X}^{\otimes(l-k+1)})$.
2. le noyau $\phi_{\chi,k:l|n}$ satisfait, pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{l-k+1}\right)$,

$$\mathbb{E}_{\chi}\left[f(X_{k:l}) \mid Y_{0:n}\right] = \int_{\mathsf{X}^{l-k+1}} f(x_{k:l})\, \phi_{\chi,k:l|n}(Y_{0:n}, dx_{k:l}) \;,$$

où l'égalité s'entend $\mathbb{P}_{\chi}$-presque sûrement. Des choix spécifiques de k et l correspondent à différents cas d'intérêt :

Lissage Joint : $\phi_{\chi,0:n|n}$, pour $n \geq 0$, c'est à dire la distribution de la trajectoire de l'état caché jusqu'à l'instant n conditionnellement aux observations jusqu'à ce même instant ;

Lissage Marginal : $\phi_{\chi,k|n}$ pour $n \geq k \geq 0$, c'est à dire la distribution de l'état caché à l'instant k (passé) conditionnellement aux observations jusqu'à l'instant présent n ;

Filtrage : $\phi_{\chi,n|n}$ pour $n \geq 0$, c'est à dire la distribution de l'état caché à l'instant présent n, conditionnellement aux observations jusqu'à ce même instant présent. Le filtrage étant prééminent dans la suite de cette thèse, nous abbrévierons le plus souvent $\phi_{\chi,n|n}$ en $\phi_{\chi,n}$.

Prédiction à p-pas : $\phi_{\chi,n+p|n}$ pour $n, p \geq 0$, c'est à dire la distribution de l'état caché au $p^{\text{ième}}$ instant futur $n+p$, conditionnellement aux observations jusqu'à l'instant présent n ; Par convention, $\phi_{\chi,0|-1}$ dénote χ ;

En toute rigueur, $\phi_{\chi,k:l|n}$ est *une version* de la distribution conditionnelle de $X_{k:l}$ sachant $Y_{0:n}$ (voir par exemple Williams (1991, Chapitre 9)). Puisqu'il n'est pas trivial qu'une telle quantité existe en toute généralité, la proposition ci-dessous complète la Définition 1.2.6 par une approche génerative définissant les quantités de lissage à partir des éléments du MMC.

Proposition 1.2.1. *Soit un MMC partiellement dominé au sens de la Définition 1.2.3, soit n un entier strictement positif et $y_{0:n} \in \mathsf{Y}^{n+1}$ une sous-suite telle que $\mathcal{L}_{\chi,n}(y_{0:n}) > 0$. La distribution de lissage joint $\phi_{\chi,0:n|n}$ satisfait alors*

$$\phi_{\chi,0:n|n}(y_{0:n}, f) = \mathcal{L}_{\chi,n}(y_{0:n})^{-1} \int_{\mathsf{X}^{k+1}} f(x_{0:n}) \times \chi(dx_0) g(x_0, y_0) \prod_{k=1}^{n} Q(x_{k-1}, dx_k) g(x_k, y_k) \quad (1.2.7)$$

pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+1}\right)$. *De même, quel que soit* $p \geq 0$,

$$\phi_{\chi,0:n+p|n}(y_{0:n}, f) = \int_{\mathsf{X}^{n+p+1}} f(x_{0:n+p})\phi_{\chi,0:n|n}(y_{0:n}, dx_{0:n}) \prod_{k=n+1}^{n+p} Q(x_{k-1}, dx_k) \tag{1.2.8}$$

pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+p+1}\right)$.

Démonstration. L'équation (1.2.7) définit $\phi_{\chi,0:n|n}$ d'une façon qui satisfait trivialement la partie (a) de la Définition 1.2.6. Pour prouver la partie (b), considérons une fonction $h \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{Y}^{n+1}\right)$. Par (1.2.4),

$$\begin{aligned}\mathbb{E}_\chi[h(Y_{0:n})f(X_{0:n})] = \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^{n+1}} & h(y_{0:n})f(x_{0:n}) \\ & \times \chi(dx_0)g(x_0, y_0)\left[\prod_{k=1}^{n} Q(x_{k-1}, dx_k)g(x_k, y_k)\right]\mu_n(dy_{0:n}) \,.\end{aligned}$$

La Définition 1.2.5 de la vraisemblance $\mathcal{L}_{\chi,n}$ et (1.2.7) pour $\phi_{\chi,0:n|n}$ entraînent que

$$\begin{aligned}\mathbb{E}_\chi[h(Y_{0:n})f(X_{0:n})] &= \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^{n+1}} h(y_{0:n})\,\phi_{\chi,0:n|n}(y_{0:n}, f)\mathcal{L}_{\chi,n}(y_{0:n})\,\mu_n(dy_{0:n}) \\ &= \mathbb{E}_\chi[h(Y_{0:n})\phi_{\chi,0:n|n}(Y_{0:n}, f)] \,.\end{aligned} \tag{1.2.9}$$

Ainsi $\mathbb{E}_\chi[f(X_{0:n}) \mid Y_{0:n}] = \phi_{\chi,0:n|n}(Y_{0:n}, f)$, $\mathbb{P}_\chi$-presque partout, pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+1}\right)$.

Pour (1.2.8), nous procédons de façon similaire et considérons deux fonctions $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+p+1}\right)$ et $h \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{Y}^{n+1}\right)$. Nous appliquons tout d'abord (1.2.4) pour obtenir

$$\begin{aligned}\mathbb{E}_\chi[h(Y_{0:n})f(X_{0:n+p})] = \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^{n+1}} & f(x_{0:n+p}) \\ & \times \chi(dx_0)g(x_0, y_0)\left[\prod_{k=1}^{n} Q(x_{k-1}, dx_k)g(x_k, y_k)\right] h(y_{0:n}) \\ & \times \left[\prod_{l=n+1}^{n+p} Q(x_{l-1}, dx_l)g(x_l, y_l)\right]\mu_{n+p}(dy_{0:n+p}) \,.\end{aligned}$$

En intégrant par rapport à la sous-suite $y_{n+1:n+p}$, la troisième ligne de l'équation précédente revient à $\prod_{l=n+1}^{n+p} Q(x_{l-1}, dx_l)\mu_n(dy_{0:n})$. Finalement, nous utilisons (1.2.6) et (1.2.7) pour obtenir

$$\begin{aligned}\mathbb{E}_\chi[h(Y_{0:n})f(X_{0:n+p})] = \int_{\mathsf{X}^{n+1}\times\mathsf{Y}^{n+1}} & h(y_{0:n})f(x_{0:n+p}) \\ & \times \phi_{\chi,0:n|n}(y_{0:n}, dx_{0:n})\left[\prod_{k=n+1}^{n+p} Q(x_{k-1}, dx_k)\right]\mathcal{L}_{\chi,n}(y_{0:n})\mu_n(dy_{0:n}) \,,\end{aligned} \tag{1.2.10}$$

ce qui conclue la preuve. □

Remarque 1.2.2. Notons que la vraisemblance $\mathcal{L}_{\chi,n}(y_{0:n})$ n'est autre que la constante de normalisation (ou *fonction de partition*) de la distribution de lissage joint. L'hypothèse qu'elle soit non-nulle est évidemment requise pour que (1.2.7) ait un sens et que (1.2.9) et (1.2.10) soient correctes. Notons que pour tout ensemble S tel que

$\int_{\mathsf{S}} \mathcal{L}_{\chi,n}(y_{0:n})\mu_n(dy_{0:n}) = 0$, $\mathbb{P}_\chi(Y_{0:n} \in \mathsf{S}) = 0$ et la valeur de $\phi_{\chi,0:n|n}(y_{0:n}, \cdot)$ pour $y_{0:n} \in \mathsf{S}$ est sans importance.

Dans la suite, il est implicite que les résultats similaires à ceux de la Proposition 1.2.1 ne sont mentionnés que pour des valeurs de $y_{0:n} \in \mathsf{S}_{\chi,n} \subset \mathsf{Y}^{n+1}$, où l'ensemble $\mathsf{S}_{\chi,n}$ est tel que $\mathbb{P}_\chi(Y_{0:n} \in \mathsf{S}_{\chi,n}) = 1$. Dans la plupart des modèles d'intérêt, cette nuance peut être ignorée car il est possible de poser $\mathsf{S}_{\chi,n} = \mathsf{Y}^{n+1}$. C'est en particulier le cas lorsque $g(x, y)$ est strictement positif pour toute valeur de $(x, y) \in \mathsf{X} \times \mathsf{Y}$. Il y a toutefois des cas plus subtils, que nous n'aborderons pas dans cette thèse, où l'ensemble $\mathsf{S}_{\chi,n}$ dépend de la distribution initiale χ.

La proposition 1.2.1 définit aussi implicitement tous les cas particuliers de noyaux de lissage mentionnés dans la Définition 1.2.6, puisqu'ils sont obtenus par marginalisation. Par exemple, le noyau de lissage marginal $\phi_{\chi,k|n}$ pour $0 \leq k \leq n$ est tel que pour tout $y_{0:n} \in \mathsf{Y}^{n+1}$ et $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$\phi_{\chi,k|n}(y_{0:n}, f) := \int \mathsf{X}^{n+1} f(x_k)\, \phi_{\chi,0:n|n}(y_{0:n}, dx_{0:n}) \,, \tag{1.2.11}$$

où $\phi_{\chi,0:n|n}$ est défini par (1.2.7).

De même, pour tout $y_{0:n} \in \mathsf{Y}^{n+1}$, la distribution de prédiction à p instants $\phi_{\chi,n+p|n}(y_{0:n}, \cdot)$ peut être obtenue en marginalisant la distribution jointe $\phi_{\chi,0:n+p|n}(y_{0:n}, \cdot)$ par rapport à toutes les variables x_k sauf la dernière (correspondant à $k = n + p$). Un examen attentif de (1.2.8), couplée à l'utilisation des équations de Chapman-Kolmogorov servant de base aux itérations de noyaux de transition Markoviens (voir par exemple Meyn and Tweedie (1994, Chapitre 3)), montre directement que $\phi_{\chi,n+p|n}(y_{0:n}, \cdot) = \phi_{\chi,n}(y_{0:n}, \cdot)Q^p$, où $\phi_{\chi,n}$ est la distribution de filtrage (distribution de X_n conditionnellement à $Y_{0:n}$).

Remarque 1.2.3 (Vraisemblance Locale et Simplification des Notations). D'une façon générale, dans cette thèse, nous considèrerons des modèles au moins partiellement dominés au sens de la Définition 1.2.3, faisant intervenir la fonction de densité de transition $g^{(k)}$ de (X, Y) dans $\mathbb{R}$. Par ailleurs, nous aborderons la plupart du temps des problèmes conditionnellement aux variables aléatoires observées $(Y_k)_{k\geq 0}$, tels que les problèmes de filtrage, qui sont, en termes Bayésiens, des problèmes d'inférence *a posteriori*. Typiquement, nous ne mentionnerons pas ici ni les problèmes d'estimation des paramètres du modèle (et non des états cachés) par maximum de vraisemblance, ni ceux d'oubli de la distribution initiale. Par conséquence, et afin d'alléger les notations, nous rendrons implicite la dépendance en l'observation Y_k – sauf cas particulier – en définissant la fonction de *vraisemblance locale*

$$g_k : x \in \mathsf{X} \mapsto g_k(x) := g^{(k)}(x, Y_k) \,. \tag{1.2.12}$$

Avec cette simplification, notamment, l'équation de lissage joint (1.2.7) s'écrit de façon plus concise :

$$\phi_{\chi,0:n|n}(f) = \mathcal{L}_{\chi,n}^{-1} \int_{\mathsf{X}^{n+1}} f(x_{0:n})\, \chi(dx_0) g_0(x_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i) g_i(x_i) \,, \tag{1.2.13}$$

où, encore une fois, la vraisemblance $\mathcal{L}_{\chi,n}$ est simplement la constante de normalisation de la distribution.

1.2.4 Récurrence fondamentale et noyau optimal

Les définitions présentées jusqu'ici nous amènent désormais au coeur de notre propos. L'équation (1.2.13) définissant de façon générale les distributions de lissag est ici

l'élément clé. En l'examinant pour n et $n+1$, on remarque la mise à jour séquentielle suivante de la distribution de lissage joint :

$$\phi_{0:n+1|n+1}(f) = \left(\frac{\mathcal{L}_{n+1}}{\mathcal{L}_n}\right)^{-1} \int_{\mathsf{X}^{n+2}} f(x_{0:n+1}) \phi_{0:n|n}(dx_{0:n})\, Q(x_n, dx_{n+1})\, g_{n+1}(x_{n+1}) \quad (1.2.14)$$

pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathcal{X}^{n+2}\right)$. En explicitant le rapport de vraisemblances correspondant à la renormalisation, on obtient la récurrence suivante :

$$\begin{aligned}\phi_{0:n+1|n+1}(f) &= \frac{\int_{\mathsf{X}^{n+2}} f(x_{0:n+1}) \phi_{0:n|n}(dx_{0:n})\, Q(x_n, dx_{n+1})\, g_{n+1}(x_{n+1})}{\int_{\mathsf{X}^{n+2}} \phi_{0:n|n}(dx_{0:n})\, Q(x_n, dx_{n+1})\, g_{n+1}(x_{n+1})} \\ &= \frac{\int_{\mathsf{X}^{n+2}} f(x_{0:n+1}) \phi_{0:n|n}(dx_{0:n})\, L_n(x_n, dx_{n+1})}{\int_{\mathsf{X}^{n+2}} \phi_{0:n|n}(dx_{0:n})\, L_n(x_n, dx_{n+1})} \end{aligned} \quad (1.2.15)$$

où le noyau L_n de $(\mathsf{X}, \mathcal{X})$ vers $(\mathsf{X}, \mathcal{X})$ est fini – i.e. pour tout $x_n \in \mathsf{X}$, $\int_{\mathsf{X}} L_n(x_n, dx_{n+1}) < \infty$, mais pas nécessairement égale à un – et défini comme

$$L_n(x_n, dx_{n+1}) := Q(x_n, dx_{n+1})\, g_{n+1}(x_{n+1}) \,. \quad (1.2.16)$$

Renormaliser ce noyau fini nous permet de définir le noyau de transition

$$L_n^*(x_n, dx_{n+1}) := \frac{L_n(x_n, dx_{n+1})}{\int_{\mathsf{X}} L_n(x_n, dx_{n+1})} \,, \quad (1.2.17)$$

et nous notons

$$\Psi^{*(n)}(x_n) := \int_{\mathsf{X}} L_n(x_n, dx_{n+1}) \,. \quad (1.2.18)$$

la fonction $\Psi^{*(n)} : \mathsf{X} \to \mathbb{R}^+$ correspondant au facteur de normalisation du noyau L_n. Cette fonction est telle que $\int_{\mathsf{X}^{n+1}} \Psi^{*(n)}(x_n) \phi_{0:n|0:n}(dx_{0:n}) = \mathcal{L}_{n+1}/\mathcal{L}_n$ – comparer pour s'en convaincre (1.2.14) et (1.2.15).

Le noyau de transition (normalisé, donc) L_n^* sera par la suite appelé le *noyau optimal* et L_n le *noyau optimal non-normalisé*. Cette terminologie remonte probablement à Zaritskii et al. (1975) et Akashi and Kumamoto (1977) et est largement adoptée par des auteurs tels que Liu and Chen (1995), Chen and Liu (2000), Doucet et al. (2000), Doucet et al. (2001) et Tanizaki (2003). La fonction de normalisation $\Psi^{*(n)}$ sera, elle, appelée *fonction d'ajustement optimale*, terme qui prendra tout son sens avec l'introduction du filtre particulaire auxiliaire de Pitt and Shephard (1999) dans le Chapitre 2. Ces objets sont optimaux au sens où ils permettent la mise à jour exacte de $\phi_{0:n|n}$ vers $\phi_{0:n+1|n+1}$. L'équation (1.2.14) correspond en effet à une structure simple mais riche dans laquelle la distribution lissage joint est modifiée en appliquant un opérateur qui n'affecte que la dernière coordonnée. Cette propriété a des implications profondes qu'exploitent les approches de Monte Carlo séquentielles, comme nous allons le voir dès la Section 1.3 et tout au long de la présente thèse. Ils sont par ailleurs également "optimaux" au sens de certains critères de qualité que nous traitons dans le Chapitre 3.

Le noyau L_n^* et la fonction $\Psi^{*(n)}$ sont importants au plus haut point et seront au coeur des méthodes adaptatives développées dans les Chapitres 3, 4 et 5, qui toutes cherchent à approcher ces quantités.

1.2.5 Version trajectorielle et cadre théorique général

La version trajectorielle de ces objets est souvent utilisée car elle permet de faire entrer les problèmes de lissage dans le cadre théorique général établi dans Douc and

Moulines (2008) et résumé dans l'Annexe A. Les versions trajectorielles de ces objets de sont rien d'autre que leur extension triviale sur l'espace des trajectoires, laissant les n premières coordonnées inchangées. Formellement, nous définissons le noyau fini $L_{\mathrm{p},n}$ de $(\mathsf{X}^{n+1}, \mathcal{X}^{\otimes n+1})$ vers $(\mathsf{X}^{n+2}, \mathcal{X}^{\otimes n+2})$ comme

$$L_{\mathrm{p},n}(x_{0:n}, dx'_{0:n+1}) := \delta_{x_{0:n}}(dx'_{0:n}) L_n(x_n, dx'_{n+1}) \tag{1.2.19}$$

et sa constante de normalisation

$$\begin{aligned} \Psi^{*\mathrm{p},(n)}(x_{0:n}) &:= \int_{\mathsf{X}^{n+2}} L_{\mathrm{p},n}(x_{0:n}, dx'_{0:n+1}) \\ &= \Psi^{*(n)}(x_n) \ , \end{aligned} \tag{1.2.20}$$

ainsi que la version renormalisée

$$L^*_{\mathrm{p},n}(x_{0:n}, dx'_{0:n+1}) := \frac{L_{\mathrm{p},n}(x_{0:n}, dx'_{0:n+1})}{\Psi^{*\mathrm{p},(n)}(x_{0:n})} \ . \tag{1.2.21}$$

L'indice p dénote la version trajectorielle (*pathwise* en anglais).

La récursion (1.2.15) s'exprime trivialement avec ces versions trajectorielles des noyaux, par

$$\phi_{0:n+1|n+1}(f) = \frac{\iint_{\mathsf{X}^{n+1}\times\mathsf{X}^{n+2}} \phi_{0:n|n}(dx_{0:n}) L_{\mathrm{p}}(x_{0:n}, dx'_{0:n+1}) f(x'_{0:n+1})}{\iint_{\mathsf{X}^{n+1}\times\mathsf{X}^{n+2}} \phi_{0:n|n}(dx_{0:n}) L_{\mathrm{p}}(x_{0:n}, dx'_{0:n+1})}$$

pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathcal{X}^{n+2}\right)$, soit, en utilisant les notations classiques en théorie de la mesure et des chaînes de Markov, la formulation plus concise

$$\phi_{0:n+1|n+1}(f) = \frac{\phi_{0:n|n} L_{\mathrm{p},n}(f)}{\phi_{0:n|n} L_{\mathrm{p},n}(\mathsf{X}^{n+2}}$$

Cette récursion permet d'exprimer la distribution cible $\mu := \phi_{0:n+1|n+1}$ sur u espace $\tilde{\mathbf{\Xi}} := \mathsf{X}^{n+2}$ sous la forme d'une distribution originale $\nu := \phi_{0:n|n}$ sur $\mathbf{\Xi} := \mathsf{X}^n$, mise à jour par un noyau fini $L = L_{\mathrm{p},n}$ de $(\mathbf{\Xi}, \mathcal{B}(\mathbf{\Xi}))$ vers $(\tilde{\mathbf{\Xi}}, \mathcal{B}(\tilde{\mathbf{\Xi}}))$ et renormalisée par $\nu L(\tilde{\mathbf{\Xi}})$. L'apparente complexité de ces expressions permet en fait l'application de la puissante théorie d'analyse résumée dans l'Annexe A, qui s'applique à toute mise à jour de la forme

$$\mu = \frac{\nu L}{\nu L(\tilde{\mathbf{\Xi}})}$$

détaillée en (A.2.1), ce qui englobe bien plus que les MMC et peut être étendue pour envisager les cas plus généraux tels que ceux considérés dans Del Moral et al. (2006). Un autre exemple de problème rentrant dans ce formalisme est l'analyse de la mise à jour des distributions de filtrage – et non de lissage joint comme jusqu'à présent – en considérant le noyau optimal L_n et la récursion facilement vérifiable

$$\phi_{n+1|n+1}(f) = \frac{\phi_{n|n} L_n(f)}{\phi_{n|n} L_n(\mathsf{X})}$$

pour toute fonction $f \in \mathcal{F}_{\mathrm{b}}\left(\mathcal{X}\right)$.

Nous utiliserons ce formalisme théorique dès le Chapitre 2, puisqu'il constitue l'outil principal permettant d'établir les résultats originaux des Chapitres 3, 4, et 5. Nous ne le détaillons toutefois pas ici pour préserver le caractère introductif de ce chapitre – cette sous-section pouvant être considérée comme une bande-annonce.

1.3 Échantillonnage préférentiel séquentiel

Dans cette section, nous nous intéressons désormais à la spécialisation de l'échantillonnage préférentiel de la Section 1.1.1 aux MMC.

1.3.1 Implémentation séquentielle pour les MMC

Nous adoptons un modèle de Markov caché tel que spécifié par la Définition 1.2.3 où le noyau de transition Markovien de la chaîne cachée est noté Q, χ est la distribution de l'état initial X_0, et $g(x, y)$ pour $x \in \mathsf{X}, y \in \mathsf{Y}$) dénote la fonction de densité de transition de l'observation conditionnellement à l'état, par rapport à la mesure μ sur $(\mathsf{Y}, \mathcal{Y})$. Afin de simplifier les expressions mathématiques, nous utiliserons également la notation raccourcie $g_k(\cdot) = g(\cdot, Y_k)$ introduite dans la Remarque 1.2.3. Nous notons $\phi_{0:k|k}$ la distribution de lissage joint , en omettant la dépendance à la distribution initiale χ, qui ne joue pas ici de rôle important. D'après (1.2.15), la distribution de lissage joint peut être mise à jour récursivement dans le temps selon les relations

$$\phi_{0:0|0}(f) = \frac{\int f(x_0)\, g_0(x_0)\, \chi(dx_0)}{\int g_0(x_0)\, \chi(dx_0)}$$

pour tout $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, et

$$\begin{aligned} \phi_{0:k+1|k+1}(f_{k+1}) &= \frac{\int_{\mathsf{X}^{k+2}} f_{k+1}(x_{0:k+1})\, \phi_{0:k|k}(dx_{0:k}) L_k(x_k, dx_{k+1})}{\int_{\mathsf{X}^{k+2}} \phi_{0:k|k}(dx_{0:k}) L_k(x_k, dx_{k+1})} \\ &\propto \int_{\mathsf{X}^{k+2}} f_{k+1}(x_{0:k+1})\, \phi_{0:k|k}(dx_{0:k}) L_k(x_k, dx_{k+1}) \end{aligned} \tag{1.3.1}$$

pour tout $f_{k+1} \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+2}\right)$. Rappelons que pour tout $x \in \mathsf{X}, f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$L_k(f) = \int_{\mathsf{X}} f(x') Q(x, dx') g_{k+1}(x')$$

est le noyau optimal non normalisé défini en (1.2.16), et que le terme de renormalisation au dénominateur de (1.3.1) est égal à

$$\int_{\mathsf{X}^{k+2}} \phi_{0:k|k}(dx_{0:k}) L_k(x_k, dx_{k+1}) = \left(\frac{\mathcal{L}_{k+1}}{\mathcal{L}_k}\right)^{-1}$$

A l'exception de certains cas précis (cas linéaire Gaussien), cette constante de normalisation n'est généralement pas disponible sous forme analytique, rendant impossible l'évaluation analytique de $\phi_{0:k|k}$. Le reste de cette section passe en revue les méthodes d'échantillonnage préférentiel permettant d'approcher $\phi_{0:k|k}$ récursivement en k.

Tout d'abord, puisque l'échantillonnage préférentiel peut être utilisé quand la distribution cible n'est connue qu'à un facteur multiplicatif près, la présence de constantes incalculables telles que $\mathcal{L}_{k+1}/\mathcal{L}_k$ n'empêche pas l'utilisation de l'algorithme. Ensuite, il est pratique de prendre pour distribution instrumentale une mesure de probabilité associée à une chaîne de Markov sur X, possiblement non-homogène. Comme vu ci-dessous, ceci permettra de construire une version séquentielle de l'échantillonnage préférentiel. Notons $\{R_k\}_{k\geq 0}$ une famille de noyaux de transitions Markoviens sur $(\mathsf{X}, \mathcal{X})$, et notons ρ_0 une mesure de probabilité sur $(\mathsf{X}, \mathcal{X})$. Notons également $\{\rho_{0:k}\}_{k\geq 0}$ une famille de mesures de probabilité associées avec la chaîne de Markov non-homogène ayant pour distribution initiale ρ_0 et pour noyaux de transition $\{R_k\}_{k\geq 0}$,

$$\rho_{0:k}(f_k) := \int_{\mathsf{X}^{k+1}} f_k(x_{0:k})\, \rho_0(dx_0) \prod_{l=0}^{k-1} R_l(x_l, dx_{l+1}) \,.$$

Dans ce cas, les noyaux R_k sont appelés *noyaux de propositions* (ou noyaux instrumentaux). Par la suite, nous adoptons les hypothèses suivantes.

Hypothèse 1.3.1 (Échantillonnage Préférentiel Séquentiel). *1. La distribution cible χ est absolument continue par rapport à la distribution instrumentale ρ_0.*

2. Pour tout $k \geq 0$ et tout $x \in \mathsf{X}$, la mesure $L_k(x, \cdot)$ est absolument continue par rapport à $R_k(x, \cdot)$.

Alors pour tout $k \geq 0$ et toute fonction $f_k \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+1}\right)$,

$$\phi_{0:k|k}(f_k) = (\mathcal{L}_k)^{-1} \int_{\mathsf{X}^{k+1}} f_k(x_{0:k})\, g_0(x_0) \frac{d\chi}{d\rho_0}(x_0) \left\{ \prod_{l=0}^{k-1} \frac{dL_l(x_l, \cdot)}{dR_l(x_l, \cdot)}(x_{l+1}) \right\} \rho_{0:k}(dx_{0:k}) \;, \quad (1.3.2)$$

ce qui implique que la distribution cible $\phi_{0:k|k}$ est absolument continue par rapport à la distribution instrumentale $\rho_{0:k}$, et la dérivée de Radon-Nikodym est donnée par

$$\frac{d\phi_{0:k|k}}{d\rho_{0:k}}(x_{0:k}) = (\mathcal{L}_k)^{-1} g_0(x_0) \frac{d\chi}{d\rho_0}(x_0) \prod_{l=0}^{k-1} \frac{dL_l(x_l, \cdot)}{dR_l(x_l, \cdot)}(x_{l+1}) \;. \quad (1.3.3)$$

Il est alors légitime d'utiliser $\rho_{0:k}$ en tant que distribution instrumentale pour calculer les estimations par échantillonnage préférentiel d'intégrales par rapport à $\phi_{0:k|k}$. En notant $\xi_1^{(0:k)}, \ldots, \xi_N^{(0:k)}$ N suites aléatoires i.i.d. ayant pour distribution commune $\rho_{0:k}$, l'estimateur d'échantillonnage préférentiel de $\phi_{0:k|k}(f_k)$ pour $f_k \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+1}\right)$ est défini comme

$$\hat{\phi}^{\mathrm{EP}}_{0:k|k}(f_k) = \frac{\sum_{i=1}^N \omega_i^{(k)} f_k(\xi_i^{(0:k)})}{\sum_{i=1}^N \omega_i^{(k)}} \;, \quad (1.3.4)$$

où $\omega_i^{(k)}$ sont les poids d'importance non-normalisés définis récursivement par

$$\omega_i^{(0)} = g_0(\xi_i^{(0)}) \frac{d\chi}{d\rho_0}(\xi_i^{(0)}) \qquad \text{pour } i = 1, \ldots, N \;, \quad (1.3.5)$$

et, pour $k \geq 0$,

$$\omega_i^{(k+1)} = \omega_i^{(k)} \frac{\mathcal{L}_{k+1}}{\mathcal{L}_k} \frac{dL_k(\xi_i^{(k)}, \cdot)}{dR_k(\xi_i^{(k)}, \cdot)}(\xi_i^{(k+1)}) \qquad \text{pour } i = 1, \ldots, N \;. \quad (1.3.6)$$

La décomposition multiplicative des poids d'importance (non-normalisés) dans (1.3.6) implique que ces poids peuvent être calculés récursivement en temps au fur et à mesure de l'arrivée de nouvelles observations. Dans la littérature de Monte Carlo séquentiel, le facteur de mise à jour dL_k/dR_k est souvent appelé le *poids incrémental*. Comme mentionné précédemment dans la Section 1.1.1, l'estimateur auto-normalisé (1.3.4) est inchangé si les poids, ou de façon équivalent les poids incrémentaux, ne sont évalués qu'à une constante près. Ceci permet, en particulier, d'omettre le facteur de renormalisation problématique $\mathcal{L}_k$ rencontré dans la dérivée de Radon-Nyokdym (1.3.3), qui fait apparaître le terme multiplicatif $\mathcal{L}_{k+1}/\mathcal{L}_k$ dans (1.3.3). Il est donc possible de poser

$$\omega_i^{(k+1)} = \omega_i^{(k)} \frac{dL_k(\xi_i^{(k)}, \cdot)}{dR_k(\xi_i^{(k)}, \cdot)}(\xi_i^{(k+1)}) \;, \quad (1.3.7)$$

en lieu et place de (1.3.6). L'échantillonnage préférentiel est donc implémenté en pratique comme décrit dans l'Algorithme 1.3.1.

Algorithme 1.3.1 EPS : Échantillonnage Préférentiel Séquentiel

État initial : Simuler un échantillon i.i.d. $\xi_1^{(0)}, \ldots, \xi_N^{(0)}$ selon ρ_0 et calculer

$$\omega_i^{(0)} = g_0(\xi_i^{(0)}) \frac{d\chi}{d\rho_0}(\xi_i^{(0)}) \qquad \text{for } i = 1, \ldots, N \ .$$

Récursion : Pour $k = 0, 1, \ldots,$

- Simuler $(\xi_1^{(k+1)}, \ldots, \xi_N^{(k+1)})$ indépendamment conditionnellement à $\{\xi_j^{(0:k)}, j = 1, \ldots, N\}$ selon la distribution $\xi_i^{(k+1)} \sim R_k(\xi_i^{(k)}, \cdot)$. Étendre $\xi_i^{(0:k)}$ avec la composante supplémentaire $\xi_i^{(k+1)}$ à $\xi_i^{(0:k)}$ pour former $\xi_i^{(0:k+1)} = (\xi_i^{(0:k)}, \xi_i^{(k+1)})$.
- Calculer les poids d'importance

$$\omega_i^{(k+1)} = \omega_i^{(k)} \times g_{k+1}(\xi_i^{(k+1)}) \frac{dQ(\xi_i^{(k)}, \cdot)}{dR_k(\xi_i^{(k)}, \cdot)}(\xi_i^{(k+1)}), \qquad i = 1, \ldots, N \ .$$

A toute itération k les estimations par échantillonnage préférentiel peuvent être évaluées selon (1.3.4).

Une propriété importante de cet algorithme, qui correspond à la méthode originellement proposée dans Handschin and Mayne (1969) et Handschin (1970), est que les N trajectoires $\xi_1^{(0:k)}, \ldots, \xi_N^{(0:k)}$ sont i.i.d. à tout instant k. En suivant la terminologie en usage dans la communauté du filtrage non-linéaire, nous appellerons l'échantillon $\{\xi_1^{(k)}, \ldots, \xi_N^{(k)}\}$ à l'instant k la population (ou le système) de *particules*, et $\xi_i^{(0:k)}$ pour une valeur spécifique de l'indice de particule i l'historique (ou trajectoire) de la $i^{\text{ème}}$ particule. Le principe de la méthode est illustré Figure 1.2.

1.3.2 Choix du noyau de proposition

Avant de présenter dans la Section 1.4 une série de problèmes de l'Algorithme 1.3.1 qui doivent être corrigés afin d'appliquer la méthode à tout problème d'intérêt, nous examinons des stratégies qui peuvent servir à choisir correctement des noyaux instrumentaux R_k parmi plusieurs modèles (ou familles de modèles) d'intérêt.

Noyau a priori

Le premier choix de noyau de proposition R_k, le plus évident et souvent très simple, consiste à poser $R_k = Q$ (indépendamment de k). Dans ce cas, le noyau de proposition correspond simplement à la distribution a priori du nouvel état en l'absence de l'observation correspondante. Le poids incrémental se simplifie alors en

$$\frac{dL_k(x, \cdot)}{dQ(x, \cdot)}(x') = g_{k+1}(x') \qquad \text{pour tout } (x, x') \in \mathsf{X}^2 \ . \tag{1.3.8}$$

Une propriété distinctive du noyau a priori est que le poids incrémental dans (1.3.8) *ne dépend pas de* x, c'est à dire, de l'état précédent. L'utilisation du noyau a priori $R_k = Q$ est populaire car simuler selon le noyau a priori Q est souvent immédiat, et calculer les poids incrémentaux revient simplement à évaluer la vraisemblance conditionnelle (aussi appelée vraisemblance locale) de la nouvelle observation étant donnée la position de la particule courante. Le noyau a priori satisfait aussi la condition nécessaire minimale de l'échantillonnage préférentiel formulée dans l'Hypothèse 1.3.1. De

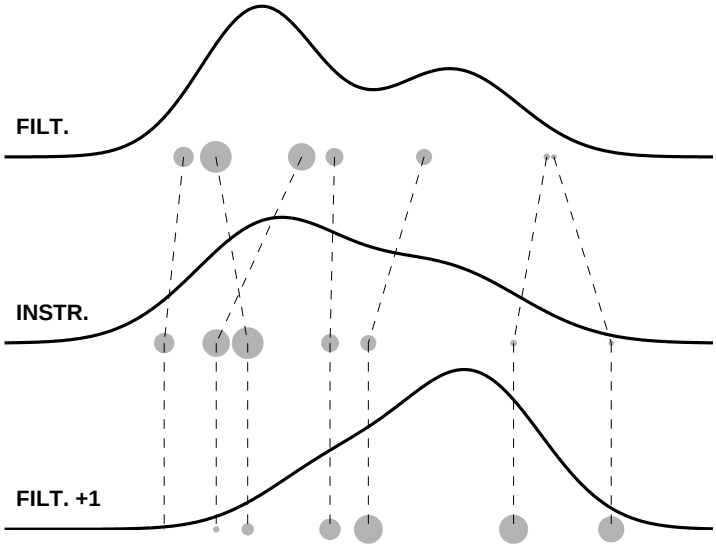


FIGURE 1.2 – Principe de l'échantillonnage préférentiel séquentiel (EPS). Figure du haut : la courbe représente la distribution de filtrage, et les particules pondérées sont représentés le long de l'axe par des disques dont le rayon est proportionnel au poids normalisé de la particule. Figure du milieu : la distribution instrumentale avec les positions des particules rééchantillonnées. Figure du bas : distribution de filtrage à l'instant suivant avec les poids des particules mis à jour. Le cas représenté ici correspond au choix $R_k = Q$.

plus, puisque la fonction d'importance se réduit à g_{k+1}, elle est bornée aussitôt que l'on peut supposer que $\sup_{x\in\mathsf{X},y\in\mathsf{Y}} g(x,y)$ est fini, ce qui est (souvent) une condition très faible. Cependant, en dépit de ces propriétés attirantes, l'utilisation du noyau a priori peut parfois donner de piètres performances, prenant souvent la forme d'un manque de robustesse par rapport aux valeurs prises par la suite observée $\{Y_k\}_{k\geq 0}$. L'exemple suivant illustre ce problème dans une situation très simple.

Exemple 1.3.1 (Modèle AR(1) bruité)**.** Afin d'illustrer les problèmes potentiels de l'utilisation du noyau a priori, Pitt and Shephard (1999) considère le modèle simple où les observations proviennent d'une autorégression linéaire de premier ordre observée en présence de bruit,

$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma_U U_k \ , & U_k &\sim \mathcal{N}(0,1) \ , \\ Y_k &= X_k + \sigma_V V_k \ , & V_k &\sim \mathcal{N}(0,1) \ , \end{aligned}$$

où $\phi = 0.9$, $\sigma_U^2 = 0.01$, $\sigma_V^2 = 1$ et $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des processus de bruits blancs Gaussiens indépendants. La distribution initiale χ est la distribution stationnaire de la chaîne de Markov $\{X_k\}_{k\geq 0}$, c'est à dire, Gaussienne centrée et de variance $\sigma_U^2/(1-\phi^2)$.

Dans la suite, nous supposerons que $n = 5$ et simulerons les cinq premières observations selon le modèle, tandis que la sixième observation sera arbitrairement fixée à la valeur 20. La suite observée est

$$(-0.652,\ -0.345,\ -0.676,\ 1.142,\ 0.721,\ 20) \ .$$

La dernière observation est située à 20 écarts-types de la moyenne (nulle) de la distribution stationnaire, ce qui correspond clairement à une valeur aberrante du point de

vue du modèle. Dans une situation pratique, toutefois, nous serions bien sûrs capables de gérer également des données qui ne proviennent pas nécessairement du modèle considéré. Notons également que dans cet exemple jouet, il est possible d'évaluer la distribution de lissage exacte à l'aide du filtre de Kalman (cf. Kalman and Bucy (1961)).

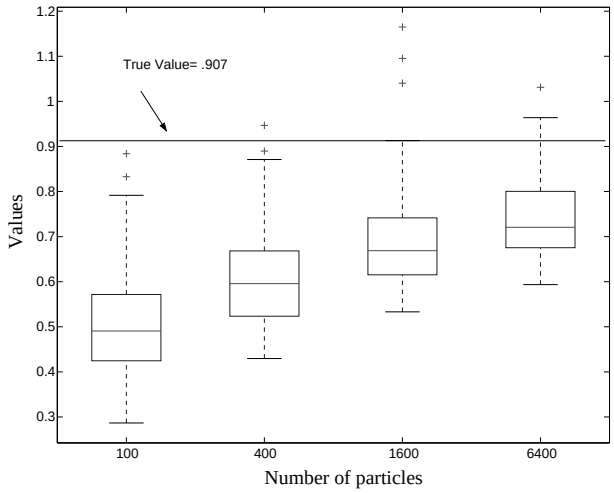


FIGURE 1.3 – Boîtes à moustaches de l'estimation de la moyenne a posteriori de X_5 obtenues sur la base de 125 réplications du filtre EPS en utilisant le noyau a priori et un nombre de particules croissant. La ligne horizontale représente la vraie moyenne a posteriori.

La Figure 1.3 représente un diagramme en boîtes à moustaches des estimations par EPS de la moyenne a posteriori de l'état final X_5 en fonction du nombre N de particules, lors de l'utilisation du noyau a priori. Ces diagrammes ont été obtenus sur la base de 125 réplications indépendantes de l'algorithme EPS. La ligne verticale correspond à la vraie moyenne a posteriori de X_5 sachant $Y_{0:5}$, calculée au moyen du filtre de Kalman. La figure montre que l'algorithme EPS avec le noyau a priori sous-estime grossièrement les valeurs de l'état même lorsque le nombre de particules est élevé. C'est un cas où il y a un conflit entre la distribution a priori et la distribution a posteriori : sous le noyau de proposition, toutes les particules sont proposées dans une région où la fonction de vraisemblance conditionnelle g_5 est extrêmement faible. Dans ce cas, la renormalisation des poids utilisés pour calculer l'estimation de la moyenne filtrée d'après (1.3.4) peut même avoir des conséquences négatives inattendues : un poids proche de 1 ne correspond pas nécessairement à une valeur simulée importante pour la distribution cible. C'est plutôt un poids qui est grand comparativement aux autres poids, encore plus faibles (de particules d'importance encore plus faible pour la distribution de filtrage). C'est une conséquence logique du fait que les poids doivent sommer à 1.

Approximation du noyau de proposition optimal

La disparité entre la distribution instrumentale et la distribution a posteriori dans l'exemple précédent est le type de problème qu'il faudrait éviter à l'aide d'un choix adéquat du noyau de proposition. Un choix intéressant pour traiter ce problème est le noyau optimal

$$L_k^*(x, f) = \frac{\int f(x')\, Q(x, dx') g_{k+1}(x')}{\int Q(x, dx') g_{k+1}(x')} \qquad \text{pour } x \in \mathsf{X}, f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}),$$

précédemment défini en (1.2.17), et qui correspond au noyau L_k renormalisé pour aboutir à un noyau de transition (au sens où pour tout $x \in \mathsf{X}$, $L_k^*(x, \mathsf{X}) = 1$). Le noyau L_k^* peut être interprété comme la distribution conditionnelle de l'état caché X_{k+1} sachant X_k et l'observation courante Y_{k+1}. La principale propriété de L_k^* est que

$$\frac{dL_k(x,\cdot)}{dL_k^*(x,\cdot)}(x') = \Psi^{*(k)}(x) \qquad \text{pour } (x,x') \in \mathsf{X}^2, \tag{1.3.9}$$

où $\Psi^{*(k)}(x)$ a déjà été définie en (1.2.18) et n'est autre que le facteur de renormalisation au dénominateur de (1.2.17), c'est à dire

$$\Psi^{*(k)}(x) := \int Q(x, dx') g_{k+1}(x') = \int L_k(x, dx') \ .$$

L'équation (1.3.9) signifie que le poids incrémental dans (1.3.7) ne dépend désormais plus que de la position précédente de la particule (et non de la nouvelle position proposée à l'instant $k+1$). C'est l'exact opposé de la situation observée précédemment pour le noyau a priori. Le noyau optimal (1.2.17) est attirant car il incorpore l'information tant de la dynamique des états que de l'information courante : avec le noyau a priori, les particules se déplacent en étant aveugles à la nouvelle observation, tandis qu'avec le noyau optimal elles ont tendance à se grouper dans les régions où le produit du noyau a priori et de la vraisemblance locale g_{k+1} est grand. L'utilisation de L_k^* pose toutefois deux problèmes pratiques. Tout d'abord, simuler selon ce noyau n'est en général pas directement possible. Ensuite, le calcul du poids incrémental $\Psi^{*(k)}$ dans (1.2.18) est, de même, souvent impossible analytiquement.

Il apparaît que le noyau optimal peut également être évalué pour une certaine classe de modèles à espace d'état Gaussiens non-linéaires, pourvu que l'équation d'observation soit linéaire (Zaritskii et al., 1975). En effet, considérons le modèle à espace d'état dont l'équation non-linéaire d'évolution de l'état est

$$X_{k+1} = A(X_k) + R(X_k)U_k \ , \qquad U_k \sim \mathcal{N}(0, I) \ , \tag{1.3.10}$$

$$Y_k = BX_k + SV_k \ , \qquad V_k \sim \mathcal{N}(0, I) \ , \tag{1.3.11}$$

où A et R sont des fonctions à valeur dans l'espace des matrices de dimensions appropriées. Un calcul méticuleux mais direct sur le conditionnement de Gaussiennes montre que la distribution conditionnelle du vecteur d'état X_{k+1} sachant $X_k = x$ et Y_{k+1} est une Gaussienne multidimensionnelle de moyenne $m_{k+1}(x)$ et de matrice de covariance $\Sigma_{k+1}(x)$, donnée par

$$\begin{aligned}
K_{k+1}(x) &= R(x)R^t(x)B^t \left[BR(x)R^t(x)B^t + SS^t\right]^{-1} \ , \\
m_{k+1}(x) &= A(x) + K_{k+1}(x)\left[Y_{k+1} - BA(x)\right] \ , \\
\Sigma_{k+1}(x) &= \left[I - K_{k+1}(x)B\right] R(x)R^t(x) \ .
\end{aligned}$$

Ainsi, les nouvelles particules $\xi_i^{(k+1)}$ doivent être simulées selon la distribution

$$\mathcal{N}\left(m_{k+1}(\xi_i^{(k)}), \Sigma_{k+1}(\xi_i^{(k)})\right) \ , \tag{1.3.12}$$

et le poids incrémental correspondant au noyau optimal est proportionnel à

$$\Psi^{*(k)}(x) = \int Q(x, dx') g_{k+1}(x') \propto$$
$$|\Gamma_{k+1}(x)|^{-1/2} \exp\left\{-\frac{1}{2}\left[Y_{k+1} - BA(x)\right]^t \Gamma_{k+1}^{-1}(x)\left[Y_{k+1} - BA(x)\right]\right\}$$

où

$$\Gamma_{k+1}(x) = BR(x)R^t(x)B^t + SS^t \ .$$

Dans d'autres situations, simuler selon le noyau L_k^* et/ou calculer la constante de normalisation $\Psi^{*(k)}$ est une tâche difficile. Il n'y a pas de recette générale pour résoudre ce problème, mais plutôt un ensemble de solutions possibles à prendre en compte.

Exemple 1.3.2 (Modèle AR(1) Bruité, Suite)**.** Nous considérons de nouveau le modèle AR(1) bruité de l'exemple 1.3.1, en utilisant le noyau de proposition optimal, qui correspond au cas particulier où toutes les variables sont réelles et A et R sont constantes dans (1.3.10)–(1.3.11) ci-dessus. Ainsi, la densité de la distribution instrumentale optimale est donnée par

$$l_k^*(x,\cdot) = \mathcal{N}\left(\frac{\sigma_U^2\sigma_V^2}{\sigma_U^2+\sigma_V^2}\left\{\frac{\phi x}{\sigma_U^2}+\frac{Y_k}{\sigma_V^2}\right\}, \frac{\sigma_U^2\sigma_V^2}{\sigma_U^2+\sigma_V^2}\right)$$

et les poids incrémentaux sont proportionnels à

$$\Psi^{*(k)}(x) \propto \exp\left[-\frac{1}{2}\frac{(Y_k-\phi x)^2}{\sigma_U^2+\sigma_V^2}\right] \ .$$

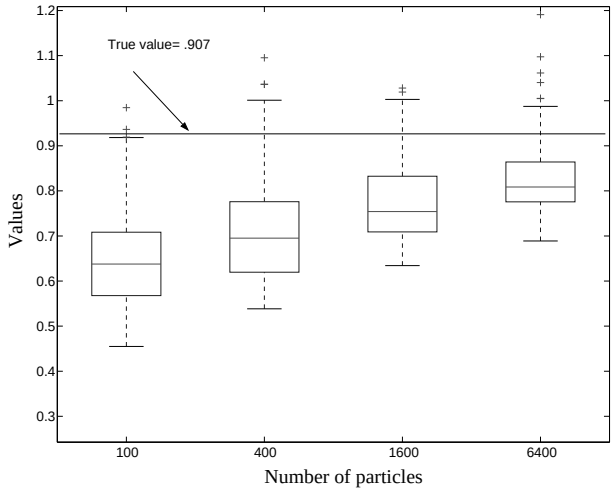


FIGURE 1.4 – Boîtes à moustaches de l'estimation de la moyenne a posteriori de X_5 obtenues sur la base de 125 réplications du filtre EPS en utilisant le noyau optimal et un nombre de particules croissant. Mêmes données et axes que pour la Figure 1.3.

La Figure 1.4 est l'exacte analogue de la Figure 1.3, obtenue également sur la base de 125 exécutions indépendantes de l'algorithme, pour ce nouveau choix du noyau de proposition. La figure montre que, bien que l'estimateur EPS de la moyenne a posteriori soit toujours négativement biaisé, le noyau optimal a tendance à réduire le biais, comparé au noyau a priori. Elle montre également que dès que $N = 400$, il y a au moins plusieurs particules situées autour de la vraie moyenne filtrée de l'état, ce qui signifie que la méthode ne devrait pas se retrouver entièrement perdue lors de l'arrivée de nouvelles observations.

Afin d'illustrer graphiquement les avantages du noyau optimal sur le noyau a priori, nous considérons de nouveau le modèle (1.3.10)–(1.3.11) avec $\phi = 0.9$, $\sigma_u^2 = 0.4$, $\sigma_v^2 = 0.6$, et $(0, 2.6, 0.6)$ en tant que suite d'observations (de longueur 3). La distribution initiale est un mélange $0.6\mathcal{N}(-1, 0.3) + 0.4\mathcal{N}(1, 0.4)$ de deux Gaussiennes, pour lequel est il est

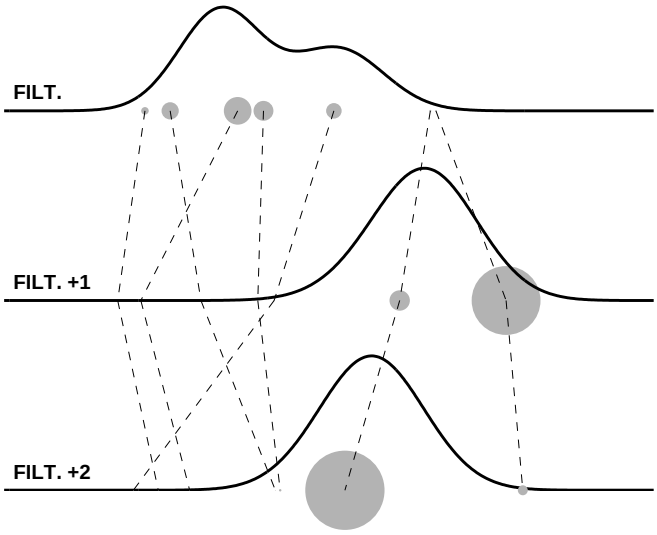


FIGURE 1.5 – EPS utilisant le noyau a priori. Les positions des particules sont indiquées par des disques dont les rayons sont proportionnels aux poids d'importance normalisés. La ligne continue montre les distributions de filtrage pour trois instants consécutifs.

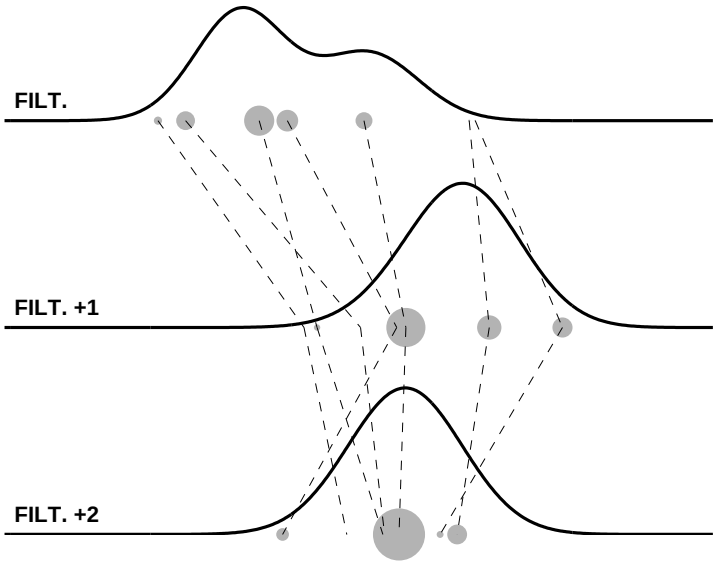


FIGURE 1.6 – EPS utilisant le noyau optimal (mêmes données et affichage que dans la Figure 1.5).

encore possible d'évaluer la distribution de filtrage en tant que mélange de deux filtres de Kalman utilisant, respectivement, $\mathcal{N}(-1, 0.3)$ et $\mathcal{N}(1, 0.4)$ pour distributions initiales de X_0. Nous n'utilisons que sept particules afin de permettre une interprétation graphique. Les Figures 1.5 et 1.6 montrent les positions des particules, propagées à l'aide du noyau a priori et du noyau optimal, respectivement. A l'instant 1, il y a conflit entre la loi a priori et la loi a posteriori, car l'observation n'est pas en accord avec l'approximation particulaire de la distribution prédictive. Avec le noyau a priori (Figure 1.5), la masse se retrouve concentrée sur une seule particule, et plusieurs particules sont perdues dans la queue gauche de la distribution avec des poids négligeables. A l'opposé, dans la Figure 1.6 la plupart des particules restent dans des régions de haute probabilité tout au long des itérations, avec plusieurs particules distinctes ayant des poids non négligeables. La raison en est précisément que le noyau optimal "tire" les particules vers des régions où la vraisemblance locale $g_{k+1}(x) = g^{(k+1)}(x, Y_k)$ est grande, ce que ne fait pas le noyau a priori.

Algorithme d'acceptation-rejet

Simuler selon le noyau optimal L_k^* n'étant pas toujours directement possible, une première idée consiste naturellement à essayer la méthode d'acceptation-rejet (voir Devroye (1986, Chapitre 2)), approche versatile pour simuler selon une distribution choisie. A cause de l'impossibilité d'évaluer la constante de normalisation $\Psi^{*(k)}$ de L_k^*, il nous faut recourir à la version non-normalisée de l'algorithme d'acceptation-rejet, aussi simple à implémenter mais dont le nombre moyen d'itérations avant acceptation dépend de ces constantes inconnues. Afin de simuler selon le noyau de proposition optimal $L_k^*(x, \cdot)$ défini par (1.2.17), il faut un noyau de proposition $R_k(x, \cdot)$ selon lequel il soit facile de simuler et tel qu'il existe M tel que $\frac{dL_k(x,\cdot)}{dR_k(x,\cdot)}(x') \leq M$ (quel que soit $x \in \mathsf{X}$). L'algorithme consiste alors à simuler des paires (ξ, U) de variables aléatoires indépendantes avec $\xi \sim R_k(x, \cdot)$ et U uniformément distribuée sur $[0, 1]$, et à accepter ξ si

$$U \leq \frac{1}{M} \frac{dQ(x, \cdot)}{dR_k(x, \cdot)}(\xi) g_{k+1}(\xi) \; .$$

Rappelons que la distribution du nombre de réalisations requises est géométrique de paramètre

$$p(x) = \frac{\int Q(x, dx') g_{k+1}(x')}{M} = \frac{\Psi^*(x)}{M} \; .$$

La force de la technique d'acceptation-rejet est que, à l'aide de tout noyau de proposition R_k satisfaisant la condition de domination, il est possible d'obtenir une réalisation selon le noyau de proposition optimal L_k^*. Par exemple, quand la vraisemblance locale $g_{k+1}(x)$ de l'observation —vue comme une fonction de x— est bornée, le noyau a priori Q peut être utilisé comme distribution instrumentale pour l'algorithme d'acceptation-rejet. Dans ce cas,

$$\frac{dL_k^*(x, \cdot)}{dQ(x, \cdot)}(x') = \frac{g_{k+1}(x')}{\int g_{k+1}(u) \, Q(x, du)} \leq \frac{\sup_{x' \in \mathsf{X}} g_{k+1}(x')}{\int g_{k+1}(u) \, Q(x, du)} \; .$$

L'algorithme consiste alors à simuler ξ selon le noyau a priori $Q(x, \cdot)$, U uniformément sur $[0, 1]$, et accepter la réalisation si $U \leq g_{k+1}(\xi) / \sup_{x \in \mathsf{X}} g_{k+1}(x)$. Le taux d'acceptation de cet algorithme est alors

$$p(x) = \frac{\int_{\mathsf{X}} Q(x, dx') g_{k+1}(x')}{\sup_{x' \in \mathsf{X}} g_{k+1}(x')} \; .$$

Malheureusement, il n'est pas toujours possible de construire un noyau de proposition $R_k(x,\cdot)$ selon lequel il soit facile de simuler, pour lequel la borne M soit finie, *et* tel que le taux d'acceptation $p(x)$ soit raisonnablement grand.

Approximation locale du noyau de proposition optimal

Une autre option est d'essayer d'approcher le noyau optimal L_k^* par un noyau de proposition R_k plus simple qui rende la simulation facile. Idéalement, R_k devrait être tel que $R_k(x,\cdot)$, à la fois ait des queues plus lourdes que $L_k^*(x,\cdot)$ et soit proche de $L_k^*(x,\cdot)$ autour de ses modes, avec l'objectif de maintenir le rapport $\frac{dL_k^*(x,\cdot)}{dR_k(x,\cdot)}(x')$ aussi petit que possible. A cette fin, des auteurs tels que Pitt and Shephard (1999) et Doucet et al. (2000) suggèrent de commencer par situer les régions de haute densité de la distribution optimale $L_k^*(x,\cdot)$, puis d'utiliser une approximation sur-dispersée (c'est à dire ayant des queues suffisamment lourdes) de $L_k^*(x,\cdot)$. La première partie de cette procédure est principalement pertinente lorsqu'il est connu que la distribution $L_k^*(x,\cdot)$ est unimodale et que son mode peut être localisé d'une façon ou d'une autre. Cette méthode doit toutefois être répétée N fois pour x correspondant tour à tour à chacune des particules actuelles. Ainsi, l'approche utilisée pour construire l'approximation doit être raisonnablement simple si l'on souhaite que les avantages potentiels de l'utilisation d'un "bon" noyau de proposition ne soient pas ruinés par une augmentation intolérable des coûts de calcul.

Une première remarque intéressante est qu'il y a une large classe de modèles à espace d'états pour lesquels il peut être prouvé, à l'aide d'arguments de convexité, que la distribution $L_k^*(x,\cdot)$ est unimodale. Dans la suite de cette section, nous supposerons que $\mathsf{X} = \mathbb{R}^d$ et que le modèle de Markov caché est entièrement dominé (au sens de la Définition 1.2.4), en notant q la densité de transition associée avec la chaîne cachée. Rappelons que pour une certaine forme de modèles non-linéaires à espace d'états donnée par (1.3.10)–(1.3.11), nous avons pu calculer explicitement le noyau optimal L^* – et donc la constante de normalisation Ψ^*. Considérons maintenant le cas où l'état évolue selon (1.3.10), de telle sorte que

$$q(x,x') \propto \exp\left[-\frac{1}{2}(x'-A(x))^t \left\{R(c)R^t(x)\right\}^{-1}(x'-A(x))\right] ,$$

et (x,y) est simplement contrainte d'être une fonction log-concave de son argument x. Ceci inclut bien sûr le modèle à observations linéaires Gaussiennes considéré précédemment dans (1.3.11), mais également de nombreux autres cas tels que celui à observations non-linéaires de l'Exemple 1.3.3 ci-dessous. La densité du noyau de transition optimal $l_k^*(x,x') = (\mathcal{L}_{k+1}/\mathcal{L}_k)^{-1} q(x,x') g_k(x')$ est alors également une fonction log-concave en son argument x', car son logarithme est une somme de fonctions concaves (et d'un terme constant). Ceci implique en particulier que $x' \mapsto l_k^*(x,x')$ est unimodale et que ses modes peuvent être localisés à l'aide de méthodes numériques efficaces telles que des itérations de Newton.

La densité du noyau de proposition est habituellement choisie au sein d'une famille paramétrique $\{r_\theta\}_{\theta\in\Theta}$ indicée par un paramètre θ. Un choix évident est la distribution Gaussienne multidimensionnelle de moyenne m et de matrice de covariance Γ, auquel cas $\theta = (\mu,\Gamma)$. Un meilleur choix est celui d'une distribution t-Student multidimensionnelle à η degrés de liberté, localisation m, et matrice d'échelle Γ. Rappelons que la densité de cette distribution est proportionnelle à $r_\theta(x') \propto [\eta + (x'-m)^t\Gamma^{-1}(x'-m)]^{(-\eta+d)/2}$. Notons qu'ici l'état précédent x' n'apparait pas explicitement mais sera implicitement présent, le paramètre θ optimal ne l'étant que particule par particule et dépendant donc

de x. Le choix $\eta = 1$ correspond à une distribution de Cauchy. C'est le choix conservateur qui garantit la sur-dispersion, mais si X est de grande dimension la plupart des réalisations selon une Cauchy multidimensionnelle seront probablement trop loin du mode pour approcher raisonnablement la distribution cible. Dans la plupart des situations, des valeurs telles que $\eta = 4$ (trois moments finis) sont plus raisonnables, spécialement si le modèle sous-jacent ne présente pas de distribution à queues lourdes. Rappelons également que la simulation selon la distribution t-Student multidimensionnelle avec η degrés de liberté, localisation m, et échelle Σ peut être facilement obtenue en simulant d'abord selon une distribution Gaussienne multidimensionnelle de moyenne m et de covariance Γ, puis en divisant le résultat par la racine carrée d'une réalisation indépendante selon une distribution du χ^2 avec η degrés de liberté divisée par η.

Afin de choisir le paramètre θ du noyau de proposition r_θ, on peut essayer de minimiser le supremum de la fonction d'importance, i.e. chercher

$$\arg\min_{\theta\in\Theta} \sup_{x'\in\mathsf{X}} \frac{l_k^*(x,x')}{r_\theta(x')} = \arg\min_{\theta\in\Theta} \sup_{x'\in\mathsf{X}} \frac{l_k(x,x')}{r_\theta(x')} \,. \tag{1.3.13}$$

Il s'agit d'une garantie minimax selon laquelle θ est choisi de telle sorte qu'il minimise une borne supérieure sur les poids d'importance. Notons que si r_θ devait être utilisée pour simuler selon $l_k^*(x,\cdot)$ par l'algorithme d'acceptation-rejet, la valeur de θ pour laquelle le minimum est atteint dans (1.3.13) est également celle qui rendrait la probabilité d'acceptation maximale. En pratique, résoudre le problème d'optimisation dans (1.3.13) est souvent trop compliqué, d'autant plus qu'il se pose pour chaque particule. Une stratégie plus générale consiste à localiser le mode de $x' \mapsto l_k^*(x,x')$ à l'aide d'un algorithme itératif puis à évaluer en ce mode la Hessienne de son logarithme. Le paramètre θ est alors choisi de la façon suivante.

Gaussienne multidimensionnelle : fixer la moyenne de la distribution Gaussienne sur le mode de $l_k^*(x,\cdot)$ et fixer la covariance à l'opposé de l'inverse de la Hessienne de $\log l_k^*(x,\cdot)$ évaluée au mode.

t-Student multidimensionnelle : fixer les paramètres de localisation et d'échelle à la moyenne et à la covariance du cas Gaussien ; le nombre de degré de liberté est habituellement choisi arbitrairement (et indépendamment de x) en se basant sur les arguments décrits précédemment.

Nous présentons ci-dessous un modèle important pour lequel une telle stratégie est performante.

Exemple 1.3.3 (Modèle de Volatilité Stochastique)**.** Nous considérons le modèle canonique de volatilité stochastique pour des données à temps discret, tel qu'étudié par (Hull and White, 1987; Jacquier et al., 1994), qui a pour modèle à espace d'état

$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma U_k \,, & U_k &\sim \mathcal{N}(0,1) \,, \\ Y_k &= \beta \exp(X_k/2) V_k \,, & V_k &\sim \mathcal{N}(0,1) \,, \end{aligned}$$

où, en termes financiers, les observations $\{Y_k\}_{k\geq 0}$ sont les log-retours, $\{X_k\}_{k\geq 0}$ est la log-volatilité, qui est supposée suivre une auto-régression stationnaire d'ordre 1, et $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des suites i.i.d. indépendantes. Le paramètre$\beta > 0$ joue le rôle d'un facteur d'échelle constant, $\phi \geq 0$ est la persistance (mémoire) de la volatilité, et σ est la volatilité de la log-volatilité. En dépit d'une représentation simple, ce modèle peut présenter une grande divserité de comportements. Comme les modèles ARCH/GARCH de Engle (1982) et Bollerslev et al. (1994), ce modèle peut donner lieu à une grande persistance de la volatilité. Même avec $\phi = 0$, ce modèle est un mélange à échelle

Gaussienne qui aura toujours un kurtosis excessivement grand pour la distribution marginale des données. Dans les modèles ARCH/GARCH avec erreurs Gaussiennes, le degré de kurtosis est lié aux racines de l'équation de volatilité, et croît avec la corrélation de la volatilité. Dans le modèle de volatilité stochastique, le paramètre σ gouverne le degré de mélange indépendamment du degré de lissage dans l'évolution de la volatilité.

A l'aide des équations d'espace d'état qui définissent le modèle, nous obtenons directement

$$q(x, x') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x' - \phi x)^2}{2\sigma^2}\right] ,$$
$$g_k(x') = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left[-\frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{1}{2}x'\right] .$$

Simuler selon le noyau de transition optimal $l_k^*(x, x')$ est difficile, mais la fonction $x' \mapsto \log(q(x, x')g_k(x'))$ est en revanche (strictement) concave. Le mode $m_k(x)$ de $x' \mapsto l_k^*(x, x')$ est l'unique solution de l'équation non-linéaire

$$-\frac{1}{\sigma^2}(x' - \phi x) + \frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{1}{2} = 0 , \tag{1.3.14}$$

qui peut-être trouvée en utilisant des itérations de Newton. Une fois le mode atteint, l'échelle (au carré) $\sigma_k^2(x)$ est fixée à l'opposé de l'inverse de la dérivée d'ordre deux de $x' \mapsto \log l_k^*(x')$ évaluée au mode $m_k(x)$. Le résultat est

$$\sigma_k^2(x) = \left\{\frac{1}{\sigma^2} + \frac{Y_k^2}{2\beta^2}\exp\left[-m_k(x)\right]\right\}^{-1} . \tag{1.3.15}$$

Dans cet exemple, nous avons utilisé une distribution t-Student avec $\eta = 5$ degrés de liberté, avec paramètres de localisation $m_k(x)$ et d'échelle $\sigma_k(x)$ obtenus comme ci-dessus. Le poids incrémental est alors donné par

$$\frac{\exp\left[-\frac{(x'-\phi x)^2}{2\sigma^2} - \frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{x'}{2}\right]}{\sigma_k^{-1}(x)\left\{\eta + \frac{[x'-m_k(x)]^2}{\sigma_k^2(x)}\right\}^{-(\eta+1)/2}} .$$

Le premier instant ($k = 0$) est particulier, et l'on vérifie facilement que $m_0(x)$ est la solution de

$$-\frac{1-\phi^2}{\sigma^2}x - \frac{1}{2} + \frac{Y_0^2}{2\beta^2}\exp(-x) = 0 ,$$

et que $\sigma_0(x)$ est donnée par

$$\sigma_0^2(x) = \left[\frac{1-\phi^2}{\sigma^2} + \frac{Y_0^2}{2\beta^2}\exp(-m_0)\right]^{-1} .$$

La Figure 1.7 montre un exemple typique des ajustements qui peuvent être obtenus pour le modèle de volatilité stochastique avec cette stratégie en utilisant 1000 particules.

Lorsqu'il n'y a pas de façon simple d'implémenter la technique de linéarisation locale, une idée naturelle explorée par Doucet et al. (2000) et Van der Merwe et al. (2000) consiste à utiliser les procédures classiques de filtrage non-linéaire pour approcher l_k^*. Celles-ci incluent en particulier le filtre de Kalman étendu (*FKE – EKF* en anglais), qui remonte aux années 70 (Anderson and Moore, 1979, Chapitre 10), ainsi que filtre

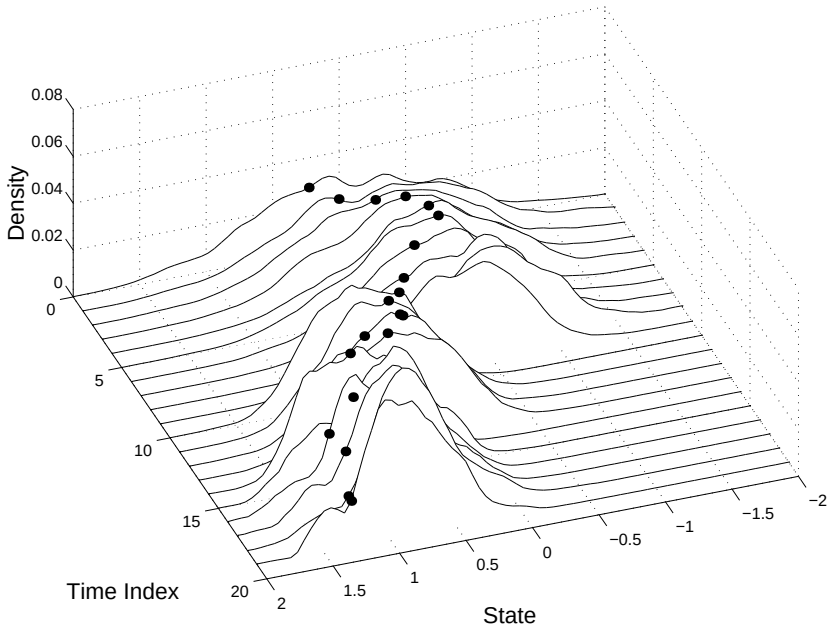


FIGURE 1.7 – Représentation en cascade des distributions de filtrage telles qu'estimées par EPS avec $N = 1000$ particules (densités obtenues avec un noyau d'Epanechnikov, largeur de fenêtre 0.2). Les données correspondent au modèle traité dans Shephard and Pitt (1997), c'est à dire $\phi = 0.98$, $\sigma = 0.14$, et $\beta = 0.66$ pour $n = 20$ instants de données historiques de taux d'échange quotidien.

de Kalman sans parfum (FKU) introduit par Julier and Uhlmann (1997)—voir, par exemple, Ristic et al. (2004, Chapitre 2) pour une revue de ces techniques. Nous illustrons ci-dessous l'emploi du filtre de Kalman étendu dans le cadre de l'EPS.

Nous considérons maintenant la forme la plus générale des modèles à espace d'état à bruits Gaussiens :

$$X_{k+1} = a(X_k, U_k) \,, \qquad U_k \sim \mathcal{N}(0, I) \,, \tag{1.3.16}$$

$$Y_k = b(X_k, V_k) \,, \qquad V_k \sim \mathcal{N}(0, I) \,, \tag{1.3.17}$$

où a, b sont des fonctions mesurables à valeurs dans un espace multidimensionnel. Nous supposons que $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des bruits blancs Gaussiens indépendants. Comme d'habitude, X_0 est supposée suivre une $\mathcal{N}(0, \Sigma_\chi)$ et être indépendante de $\{U_k\}$ et $\{V_k\}$. Le filtre de Kalman étendu consiste à approcher les équations non-linéaires d'espace d'état (1.3.16)–(1.3.17) par un modèle à espace d'état à équation d'observation linéaire. Nous nous ramenons ainsi à un modèle de la forme (1.3.10)–(1.3.11) pour lequel la forme exacte du noyau optimal peut être déterminée en utilisant les formules Gaussiennes. Nous adopterons l'approximation

$$X_k \approx a(X_{k-1}, 0) + R(X_{k-1})U_{k-1} \,, \tag{1.3.18}$$

$$Y_k \approx b\left[a(X_{k-1}, 0), 0\right] + B(X_{k-1})\left[X_k - a(X_{k-1}, 0)\right] + S(X_{k-1})V_k \,, \tag{1.3.19}$$

où

– $R(x)$ est la matrice $d_x \times d_u$ des dérivées partielles de $a(x, u)$ par rapport à u et évaluée en $(x, 0)$,

$$\left[R(x)\right]_{i,j} := \frac{\partial \left[a(x, 0)\right]_i}{\partial u_j} \qquad \text{pour } i = 1, \ldots, d_x \text{ et } j = 1, \ldots, d_u \,;$$

– $B(x)$ et $S(x)$ sont les matrices $d_y \times d_x$ et $d_y \times d_v$ des dérivées partielles de $b(x, v)$ par rapport à x et v, respectivement, et évaluées en $(a(x, 0), 0)$,

$$[B(x)]_{i,j} = \frac{\partial \{b[a(x,0),0]\}_i}{\partial x_j} \qquad \text{pour } i = 1, \dots, d_y \text{ et } j = 1, \dots, d_x ,$$
$$[S(x)]_{i,j} = \frac{\partial \{b[a(x,0),0]\}_i}{\partial v_j} \qquad \text{pour } i = 1, \dots, d_y \text{ et } j = 1, \dots, d_v .$$

Il est important de souligner que l'équation de mesure dans (1.3.19) diffère de (1.3.11) en ce qu'elle dépend à la fois de l'état courant X_k et du précédent X_{k-1}. Le modèle approché spécifié par (1.3.18)–(1.3.19) ne vérifie donc plus les hypothèses MMC. Par ailleurs, *lorsque l'on conditionne* à la valeur de X_{k-1}, les structures des deux modèles (1.3.10)–(1.3.11) et (1.3.18)–(1.3.19) sont exactement similaires. La distribution a posteriori X_k sachant $X_{k-1} = x$ et Y_k est donc une distribution Gaussienne de moyenne $m_k(x)$ et de matrice de covariance $\Gamma_k(x)$, qui peut être évaluée selon

$$\begin{aligned} K_k(x) &= R(x)R^t(x)B^t(x)\left[B(x)R(x)R^t(x)B^t(x) + S(x)S^t(x)\right]^{-1} , \\ m_k(x) &= a(x,0) + K_k(x)\ \{Y_k - b[a(x,0),0]\} , \\ \Gamma(x) &= [I - K_k(x)B(x)]\, R(x)R^t(x) . \end{aligned}$$

La distribution Gaussienne de moyenne $m_k(x)$ et de covariance $\Gamma_k(x)$ peut alors être utilisée comme substitut au noyau de transition optimal $L_k^*(x, \cdot)$. Afin d'améliorer la robustesse de la méthode, il est plus sûr d'augmenter la variance, c'est à dire d'utiliser $c\Gamma_k(x)$ en tant que variance de simulation, où c est un réel plus grand que 1. Une option peut-être plus recommandable consiste à utiliser, comme précédemment, une distribution de proposition avec des queues plus lourdes que la Gaussienne, par exemple une t-Student multidimensionnelle avec localisation $m_k(x)$, échelle $\Gamma_k(x)$, et quatre ou cinq degrés de liberté.

Exemple 1.3.4 (Modèle de croissance)**.** Nous considérons le modèle unidimensionnel de croissance traité par Kitagawa (1987) et Polson et al. (1992) et décrit, sous forme de modèle à espace d'état, par

$$X_k = a_{k-1}(X_{k-1}) + \sigma_u U_{k-1} , \qquad U_k \sim \mathcal{N}(0,1) , \tag{1.3.20}$$
$$Y_k = bX_k^2 + \sigma_v V_k , \qquad V_k \sim \mathcal{N}(0,1) , \tag{1.3.21}$$

où $\{U_k\}_{k\geq 0}$ et $\{V_k\}_{k\geq 0}$ sont des processus de bruit blanc Gaussiens indépendants et

$$a_{k-1}(x) = \alpha_0 x + \alpha_1 \frac{x}{1+x^2} + \alpha_2 \cos[1.2(k-1)] \tag{1.3.22}$$

avec $\alpha_0 = 0.5$, $\alpha_1 = 25$, $\alpha_2 = 8$, $b = 0.05$, et $\sigma_v^2 = 1$ (la valeur de σ_u^2 sera discutée ci-après). L'état initial est connu de façon déterministe et fixé à $X_0 = 0.1$. Ce modèle est non-linéaire tant dans l'équation d'état que dans celle d'observation. Notons que la forme de la vraisemblance locale ajoute une complication supplémentaire au problème : sitôt que $Y_k \leq 0$, la fonction de vraisemblance locale

$$g_k(x) := g(x; Y_k) \propto \exp\left[-\frac{b^2}{2\sigma_v^2}\left(x^2 - Y_k/b\right)^2\right]$$

est unimodale et symétrique par rapport à 0 ; en revanche, quand $Y_k > 0$, la vraisemblance g_k, toujours symétrique par rapport à 0, est cette fois bimodale, avec deux modes situés en $\pm(Y_k/b)^{1/2}$.

L'approximation par FKE du noyau de transition optimal est la distribution Gaussienne de moyenne $m_k(x)$ et de variance $\Gamma_k(x)$ donnée par

$$
\begin{aligned}
K_k(x) &= 2\sigma_u^2 b a_{k-1}(x) \left[4\sigma_u^2 b^2 f_{k-1}^2(x) + \sigma_v^2\right]^{-1} \ , \\
m_k(x) &= a_{k-1}(x) + K_{k-1}(x) \left[Y_k - b a_{k-1}^2(x)\right] \ , \\
\Gamma_k(x) &= \frac{\sigma_v^2 \sigma_u^2}{4\sigma_u^2 b^2 a_{k-1}^2(x) + \sigma_v^2} \ .
\end{aligned}
$$

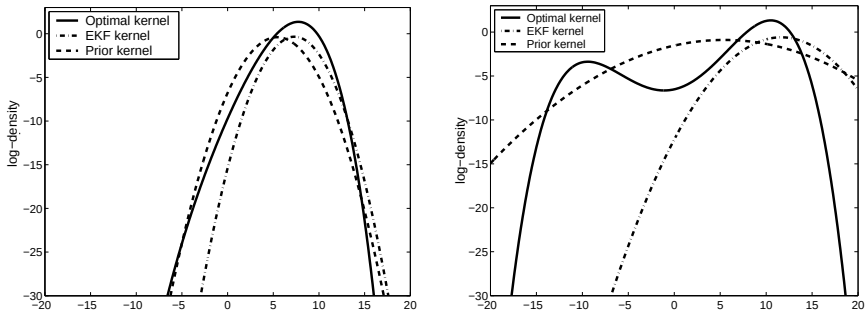


FIGURE 1.8 – Log-densité du noyau optimal (ligne continue), approximation par FKE du noyau optimal (ligne alternée trait-point), et noyau a priori (ligne en traits), évalués en $X_0 = 0.1$, pour deux valeurs différentes de la variance σ_u^2 du bruit d'état : à gauche, $\sigma_u^2 = 1$; à droite, $\sigma_u^2 = 10$.

La Figure 1.8 permet de comparer le noyau optimal, son approximation par FKE, et le noyau a priori, tous évalués en $X_0 = 0.1$, pour deux valeurs différentes de la variance du bruit d'état. Cette figure correspond à l'instant 1, et Y_1 est fixé à 6. Dans le cas où $\sigma_u^2 = 1$ (graphique de gauche de la Figure 1.8), la distribution a priori de l'état, $\mathcal{N}(a_0(X_0), \sigma_u^2)$, s'avère être plus informative (plus pointue, moins diffuse) que la vraisemblance locale g_1. En d'autres termes, l'observation Y_1 n'apporte pas beaucoup d'information sur l'état X_1, comparée à l'information apportée par X_0 ; ceci parce que la variance de mesure σ_v^2 n'est pas faible comparée à σ_u^2. Le noyau de transition optimal, qui tient compte de Y_1, est alors très proche du noyau a priori, et les différences entre les trois noyaux sont mineurs. Dans une telle situation, il ne faut pas attendre une grande amélioration de l'approximation par FKE par rapport au noyau a priori.

Dans le cas du graphique de droite de la Figure 1.8 ($\sigma_u^2 = 10$), la situation est inversée. Maintenant, σ_v^2 est relativement petite comparée à σ_u^2, de telle sorte que l'information sur X_1 contenue dans g_1 est grande comparée à celle fournie par l'information a priori de X_0. C'est le type de situation où nous nous attendons à ce que le noyau optimal améliore considérablement les résultats par rapport au noyau a priori. En effet, puisque $Y_1 > 0$, le noyau optimal est bimodal, et son second mode est bien plus petit que le premier (rappelons que les figures sont à l'échelle logarithmique) ; le noyau FKE choisit correctement le mode dominant. La Figure 1.8 illustre aussi le fait que, contrairement au noyau a priori, le noyau FKE ne domine pas nécessairement le noyau optimal dans les queues, d'où le besoin de simuler selon une version sur-dispersée de l'approximation FKE comme mentionné précédemment.

1.4 Échantillonnage préférentiel séquentiel avec rééchantillonnage

Malgré des résultats plutôt satisfaisant pour des séries de données courtes, comme observé dans l'Exemple 1.3.3, il s'avère que l'approche EPS exposée jusqu'ici est condamnée à l'échec en temps long. Nous donnons d'abord corps à cette affirmation avec un exemple illustratif simple avant d'examiner les solutions à ce travers, basées sur le concept de rééchantillonnage introduit en Section 1.1.2.

1.4.1 Dégénérescence des poids

L'interprétation intuitive des poids d'importance $\omega_i^{(k)}$ consiste à les voir comme mesure de l'adéquation de la distribution instrumentale avec la distribution cible, et donc principalement du noyau de proposition avec le noyau optimal. Deux cas peuvent typiquement expliquer un faible poids $\omega_i^{(k)}$.

i) soit la trajectoire simulée $\xi_i^{(0:k)}$ est en faible adéquation avec la distribution cible $\phi_{0:k|k}$. Un faible poids d'importance implique alors que la trajectoire est simulée loin de la masse principale de la distribution a posteriori $\phi_{0:k|k}$. Plus exactement, elle est située dans les queues de la distribution cible, où la distribution instrumentale propose plus que ne le ferait la distribution cible, d'où une densité de Radon-Nikodym faible. Ceci correspond au cas où le numérateur du poids est faible.

ii) soit la trajectoire $\xi_i^{(0:k)}$ est bel et bien simulée au coeur de la masse principale de la distribution a posteriori $\phi_{0:k|k}$ (ou raisonnablement proche), mais la distribution instrumentale est plus dense que la distribution cible dans cette région, et donc de trop nombreuses particules y sont simulées. Ceci correspond donc au cas où le dénominateur du poids est fort.

De la même façon, un poids anormalement fort peut correspondre soit à une particule en adéquation avec la distribution cible $\phi_{0:k|k}$ (numérateur fort), ou au contraire à une particule située certes en queue de distribution cible mais également encore plus loin dans la queue de la distribution instrumentale (dénominateur très faible). C'est d'ailleurs le contrôle de ce dernier cas (à l'aide de conditions nécessaires d'intégrabilité du carré de la dérivée de Radon-Nikodym) qui permet d'obtenir des variances asymptotiques finies dans les théorèmes limite centraux tels Geweke (1989, Théorème 2) pour l'échantillonnage préférentiel non séquentiel et le Théorème A.2.2 pour l'EPS.

Ces cas typiques illustrent bien la fonction même du poids $\omega_i^{(k)}$, qui consiste à "débiaiser" l'échantillon en corrigeant le fait que la densité de proposition n'est pas la densité cible. Dans tous les cas, un poids faible $\omega_i^{(k)}$ signifie que la particule correspondante $\xi_i^{(0:k)}$ ne contribuera que modérément aux estimations de la forme (1.3.4). En effet, une particule telle que le poids associé soit plusieurs ordres de grandeur plus petit que la somme $\sum_{i=1}^N \omega_i^{(k)}$ est pratiquement inutile. Le problème se pose également avec une particule dont le poids $\omega_i^{(k)}$ est plusieurs ordres de grandeur plus grand que ceux des autres particules : la renormalisation par $\sum_{j=1}^N \omega_j^{(k)}$ lui attribuera alors un poids proche de 1, et les autres particules seront inutilisées. S'il y a trop de particules inutilisées, l'approximation particulaire devient inefficace tant d'un poids de vue statistique qu'informatique : la plupart de l'effort de calcul est gaspillée dans la mise à jour de particules et de poids qui ne contribuent pas significativement à l'estimation. La variance de l'estimateur résultant ne reflètera pas le grand nombre de termes de la

somme, mais seulement la petite partie de particules ayant des poids normalisés non négligeables.

Malheureusement, la situation décrite ci-dessus est la règle plutôt que l'exception, car les poids d'importance vont (presque toujours) dégénérer quand l'instant k augmente, avec la plupart des poids normalisés $\omega_i^{(k)} / \sum_{j=1}^N \omega_j^{(k)}$ proches de 0 à l'exception de quelques uns. Nous considérons ci-dessous le cas de modèles i.i.d. pour lesquels il est possible de montrer, en usant d'arguments simples, que la variance asymptotique (i.e. pour de grands échantillons) de l'estimateur par EPS ne peut que croître avec l'instant k.

Exemple 1.4.1 (Dégénérescence des poids dans le cas I.I.D.)**.** Le cas le plus simple d'application de la technique d'EPS est celui où μ est une mesure de probabilité sur $(\mathsf{X}, \mathcal{X})$ et la suite de distributions cibles correspond aux distributions produits, c'est à dire, à la suite de distributions sur $(\mathsf{X}^{k+1}, \mathcal{X}^{\otimes(k+1)})$ définie récursivement par $\mu_0 = \mu$ et $\mu_k = \mu_{k-1} \otimes \mu$, pour $k \geq 1$. Soit ν une autre mesure de probabilité sur $(\mathsf{X}, \mathcal{X})$, et supposons que μ est absolument continue par rapport à ν et que

$$\int \left[\frac{d\mu}{d\nu}(x)\right]^2 \nu(dx) < \infty \,. \tag{1.4.1}$$

Finalement, soit f une fonction mesurable bornée qui ne soit pas (μ-presque sûrement) constante, telle que sa variance sous μ, $\mu(f^2) - \mu^2(f)$, soit strictement positive.

Considérons l'estimateur par EPS donné par

$$\widehat{\mu}_{k,N}^{\mathrm{EP}}(f) = \sum_{i=1}^N f(\xi_i^{(k)}) \frac{\prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_i^{(k)})}{\sum_{j=1}^N \prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_j^{(k)})} \,, \tag{1.4.2}$$

où les variables aléatoires $\{\xi_j^{(l)}\}$, $l = 1, \ldots, k$, $j = 1, \ldots, N$ sont i.i.d. selon ν. Comme exposé en Section 1.3, les poids d'importance non-normalisés peuvent être calculés récursivement, et ainsi (1.4.2) correspond bien à un estimateur de la forme (1.3.4) dans le cas particulier d'une fonction f_k dépendant uniquement de la dernière composante. Ceci est bien sûr une façon bien compliquée et fort peu performante de construire un estimateur de $\mu(f)$, mais n'en est pas moins un cas valide de l'approche EPS (dans une situation très particulière).

Fixons maintenant k et décomposons

$$N^{1/2}\left\{\widehat{\mu}_{k,N}^{\mathrm{EP}}(f) - \mu(f)\right\} = \frac{N^{-1/2}\sum_{i=1}^N \prod_{l=0}^k \left\{f(\xi_i^{(k)}) - \mu(f)\right\} \frac{d\mu}{d\nu}(\xi_i^{(l)})}{N^{-1}\sum_{i=1}^N \prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_i^{(l)})} \,. \tag{1.4.3}$$

Comme

$$\mathbb{E}\left[\prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_i^{(l)})\right] = 1 \,,$$

la loi faible des grands nombres entraîne que le dénominateur du membre de droite de (1.4.3) converge vers 1 en probabilité lorsque N croît. De même, sous (1.4.1), le théorème limite central montre que le numérateur de la partie droite de (1.4.3) converge en loi vers la distribution Gaussienne $\mathcal{N}(0, \sigma_k^2(f))$, où

$$\begin{aligned} \sigma_k^2(f) &= \mathbb{E}\left(\left\{\prod_{l=0}^k \left[f(\xi_1^{(k)}) - \mu(f)\right]^2 \frac{d\mu}{d\nu}(\xi_1^{(l)})\right\}^2\right) \\ &= \left[\int \left(\frac{d\mu}{d\nu}(x)\right)^2 \nu(dx)\right]^k \int \left[\frac{d\mu}{d\nu}(x)\right]^2 [f(x) - \mu(f)]^2 \nu(dx) \,. \end{aligned} \tag{1.4.4}$$

Le lemme de Slutsky implique alors que (1.4.3) converge également en loi vers la même limite $\mathcal{N}(0, \sigma_k^2(f))$ quand N croît. L'inégalité de Jensen implique maintenant que

$$1 = \left[\int \frac{d\mu}{d\nu}(x)\nu(dx)\right]^2 \leq \int \left[\frac{d\mu}{d\nu}(x)\right]^2 \nu(dx) \,,$$

l'égalité étant atteinte si et seulement si $\mu = \nu$. Ainsi, si $\mu \neq \nu$, la variance asymptotique $\sigma_k^2(f)$ croît exponentiellement avec l'instant k pour toute fonction f telle que

$$\int \left[\frac{d\mu}{d\nu}(x)\right]^2 [f(x) - \mu(f)]^2 \, \nu(dx) = \int \frac{d\mu}{d\nu}(x) \, [f(x) - \mu(f)]^2 \, \mu(dx) \neq 0 \,.$$

Puisque la distribution μ est absolument continue par rapport à ν, $\mu\{x \in \mathsf{X} : d\mu/d\nu(x) = 0\} = 0$ et la dernière intégrale est nulle si et seulement si f est de variance nulle sous μ.

Ainsi, dans le cas i.i.d., la variance asymptotique de l'estimateur par échantillonnage préférentiel (1.4.2) augmente exponentiellement avec l'instant k dès que la loi instrumentale et la loi cible diffèrent (sauf pour les fonctions constantes).

Il est plus difficile de caractériser la dégénérescence des poids pour des distributions cible et instrumentale générales. Il y a eu quelques essais limités pour examiner plus formellement ce phénomène dans des scénarios spécifiques. En particulier, Del Moral and Jacod (2001) ont montré la dégénérescence de l'estimateur par EPS de la moyenne a posteriori dans les modèles linéaires Gaussiens lorsque le noyau de proposition est le noyau a priori. De tels résultats sont en général difficiles à établir (même pour les modèles linéaires Gaussiens où la plupart des calculs peuvent être menés explicitement) et n'apportent guère d'éléments de compréhension supplémentaires. Nul n'est besoin de souligner que, en pratique, la dégénérescence des poids est un problème sérieux et omniprésent, rendant presque inutile la méthode EPS de base telle que discutée jusqu'à présent. La dégénérescence peut se produire après un nombre très limité d'itérations, comme le montre l'exemple suivant.

Exemple 1.4.2 (Modèle de Volatilité Stochastique, Suite)**.** La Figure 1.9 représente l'histogramme du logarithme en base 10 des poids normalisés après 1, 10, et 100 instants pour le modèle de volatilité stochastique considéré dans l'Exemple 1.3.3 (en utilisant le même noyau optimal). Le nombre de particules est fixé à 1000. La Figure 1.9 montre que, malgré le choix d'une approximation raisonnablement bonne du noyau de proposition optimal, les poids normalisés dégénèrent rapidement quand le nombre d'itérations de l'algorithme EPS augmente. Clairement, les résultats présentés dans la Figure 1.7 sont encore raisonnables pour $k = 20$ mais seraient désastreux pour des horizons temporels plus grands, tels $k = 100$.

La dégénérescence des poids étant si problématique, il est très important en pratique de mettre au point des tests pour détecter ce phénomène. Un critère simple pour cela est le coefficient de variation des poids normalisés utilisé par Kong et al. (1994), qui est défini par

$$\mathrm{CV}\left(\{\omega_i\}_{i=1}^N\right) = \left[\frac{1}{N}\sum_{i=1}^N \left(N\frac{\omega_i}{\sum_{j=1}^N \omega_j} - 1\right)^2\right]^{1/2} . \tag{1.4.5}$$

Le coefficient de variation est minimal lorsque les poids normalisés sont tous égaux à$1/N$, auquel cas $\mathrm{CV}\left(\{\omega_i\}_{i=1}^N\right) = 0$. La valeur maximale de $\mathrm{CV}\left(\{\omega_i\}_{i=1}^N\right)$ est $\sqrt{N-1}$,

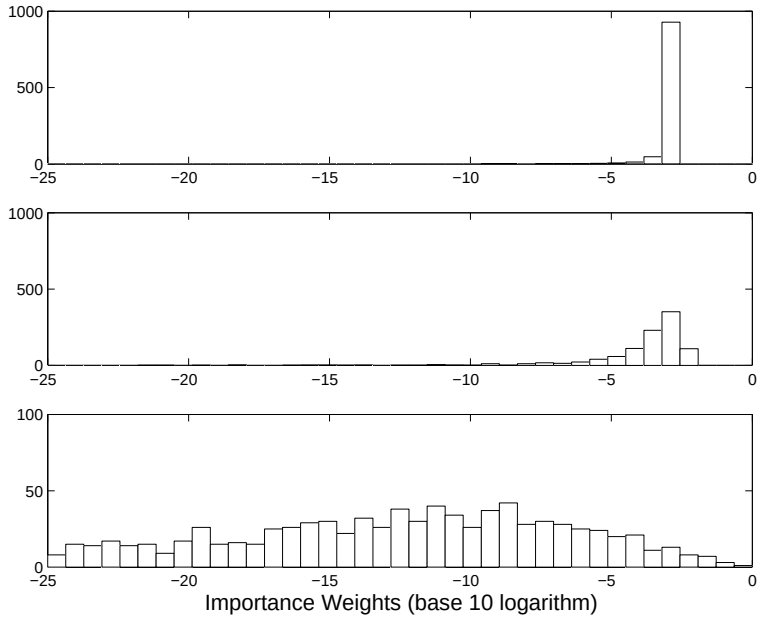


FIGURE 1.9 – Histogrammes des poids normalisés, en logarithme base 10, après (de haut en bas) 1, 10, 100 itérations du modèle de volatilité stochastique de l'Exemple 1.3.3. L'échelle verticale du panneau inférieur a été multipliée par 10.

qui correspond au cas où l'un des poids normalisés vaut 1 et tous les autres sont nuls. Ainsi, le coefficient de variation est souvent interprété comme une mesure du nombre de particules inutiles (celles qui ne contribuent pas significativement à l'estimation). Un critère connexe avec une interprétation plus simple quoique plus approximative est la *taille d'échantillon effective* (***TEE*** – ***ESS*** en anglais) N_{eff} (Liu, 1996), définie comme

$$N_{\text{eff}}\left(\{\omega_i\}_{i=1}^N\right) = \left[\sum_{i=1}^N \left(\frac{\omega_i}{\sum_{j=1}^N \omega_j}\right)^2\right]^{-1}, \tag{1.4.6}$$

qui varie entre 1 (tous les poids nuls sauf un) et N (poids égaux). La vérification de la relation

$$N_{\text{eff}}\left(\{\omega_i\}_{i=1}^N\right) = \frac{N}{1+\text{CV}^2\left(\{\omega_i\}_{i=1}^N\right)}.$$

est immédiate. Quelques éléments supplémentaires ainsi que des heuristiques sur le coefficient de variation sont données dans Liu and Chen (1995), qui est toutefois une analyse plus empirique ("rule of thumb", comme le mentionnent par trois fois les auteurs) que mathématique.

Une autre mesure possible du déséquilibre des poids est encore l'opposé de l'entropie de Shannon des poids d'importance,

$$\mathcal{E}\left(\{\omega_i\}_{i=1}^N\right) = \sum_{i=1}^N \frac{\omega_i}{\sum_{j=1}^N \omega_j} \log\left(\frac{\omega_i}{\sum_{j=1}^N \omega_j}\right). \tag{1.4.7}$$

Quand tous poids normalisés sont nuls sauf un, l'entropie est nulle. Au contraire, si tous les poids sont égaux à $1/N$, alors l'entropie est maximale et égale à $\log N$.

Ces deux critères que sont le coefficient de variation et l'entropie de Shannon sont utilisés extrêmement fréquemment dans la communauté MCS. Toutefois, leur analyse théorique détaillée n'a jamais été effectuée, le seul article approchant étant Liu and

Chen (1995) déjà cité. Le point clé du Chapitre 3 de cette thèse (Cornebise et al., 2008, qui n'est autre que l'article) est précisément une telle analyse. Nous y montrons que ces quantités sont des estimateurs de la divergence du χ^2 et de la divergence de Kullback-Leibler entre deux lois sur l'espace joint des particules à l'instant précédent et des particules proposées, qui sont la loi asymptotique du couple de la particule actuelle et de son successeur selon le modèle (et donc selon la loi cible) d'une part, et selon la distribution de proposition d'autre part. Une telle analyse permet alors la définition de critères de qualité rigoureux, base solide sur laquelle fonder de nombreux développements (algorithmes adaptatifs, notamment). Nous n'entrons volontairement pas plus dans les détails au sein de ce chapitre, et recommandons de se familiariser avec le *filtre particulaire auxiliaire* – éventuellement au moyen du Chapitre 2 – avant d'étudier le Chapitre 3.

Exemple 1.4.3 (Modèle de Volatilité Stochastique, Suite)**.** La Figure 1.10 décrit le coefficient de variation (à gauche) et l'entropie de Shannon (à droite) en fonction de l'instant k, sous les mêmes conditions que pour la Figure 1.9, c'est à dire pour le modèle de volatilité stochastique de l'Exemple 1.3.3. La figure montre que la distribution des poids dégénère régulièrement : le coefficient de variation croît et l'entropie des poids décroît. Après 100 itérations, moins de 50 particules (sur 1000) contribuent significativement à l'estimateur EPS. La plupart des particules a des poids d'importance nuls à la précision machine près, ce qui est bien sûr un gaspillage de ressources de calcul considérable.

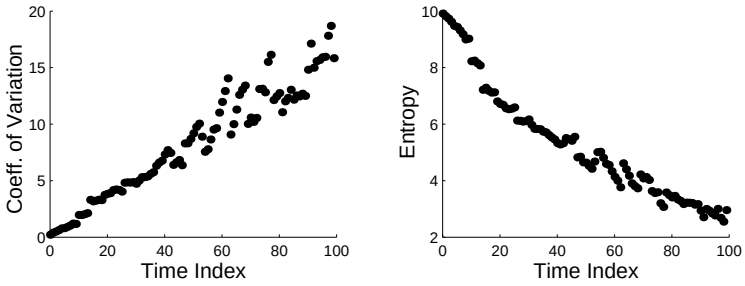


FIGURE 1.10 – Coefficient de variation (gauche) et entropie des poids normalisés en fonction du nombre d'itérations pour le modèle de volatilité stochastique de l'Exemple 1.3.3. Mêmes modèle et données que pour la Figure 1.9.

1.4.2 Rééchantillonnage

La solution proposée par Gordon et al. (1993) pour réduire la dégénérescence des poids d'importance est basée sur le concept de *rééchantillonnage* précédemment abordé dans le contexte de l'échantillonnage préférentiel à la Section 1.1.2. La méthode basique consiste à rééchantillonner parmi la population courante de particules en utilisant les poids normalisés comme probabilités de sélection. Ainsi, les trajectoires ayant un poids d'importance faible sont éliminées, tandis que ceux ayant un grand poids d'importance sont dupliqués. Après rééchantillonnage, les poids d'importance des particules rééchantillonnées sont fixés à 1.

Jusqu'au premier instant où le rééchantillonnage se produit, cette méthode n'est rien d'autre qu'une version de la technique EPS exposée dans la Section 1.1.2. Dans le cadre des méthodes de Monte Carlo séquentielles, cependant, la motivation principale pour rééchantillonner est d'éviter une future dégénérescence des poids en remettant

(périodiquement) tous les poids à une valeur égale. L'étape de rééchantillonnage a toutefois un inconvénient : comme souligné dans la Section 1.1.2, rééchantillonner introduit une variance supplémentaire dans les approximation de Monte Carlo. Dans certaines situations, cette variance supplémentaire peut s'avérer loin d'être négligeable : quand les poids d'importance sont déjà presque égaux, par exemple, rééchantillonner ne peut que réduire le nombre de particules distinctes, et donc dégrader la précision de l'approximation par Monte Carlo. L'effet à un pas du rééchantillonnage est donc négatif mais, sur le long terme, le rééchantillonnage est requis pour garantir la stabilité de l'algorithme. Cette interprétation suggère qu'il peut être avantageux de restreindre le rééchantillonnage aux cas où les poids d'importance deviennent très inégaux. Les critères définis en (1.4.5), (1.4.6), ou (1.4.7) peuvent bien sûr servir ce propos. L'algorithme qui en résulte, généralement connu sous le nom d'*échantillonnage préférentiel séquentiel avec rééchantillonnage* (*EPSR – SISR* en anglais), est résumé ci-dessous.

Algorithme 1.4.1 EPSR : Échantillonnage Préférentiel Séquentiel avec Rééchantillonnage

---

Initialiser les particules comme dans l'Algorithme 1.3.1, avec facultativement l'étape de rééchantillonnage ci-dessous. Pour tous les instants suivants $k \geq 0$, suivre la suite.

Échantillonnage

- Simuler $(\tilde{\xi}_1^{(k+1)}, \ldots, \tilde{\xi}_N^{(k+1)})$ indépendamment conditionnellement à $\{\xi_j^{(0:k)}, j = 1, \ldots, N\}$ selon le noyau de proposition : $\tilde{\xi}_i^{(k+1)} \sim R_k(\xi_i^{(k)}, \cdot)$, $i = 1, \ldots, N$.
- Calculer les poids d'importance mis à jour

$$\omega_i^{(k+1)} = \omega_i^{(k)} g_{k+1}(\tilde{\xi}_i^{(k+1)}) \frac{dQ(\xi_i^{(k)}, \cdot)}{dR_k(\xi_i^{(k)}, \cdot)}(\tilde{\xi}_i^{(k+1)}) \,, \qquad i = 1, \ldots, N \,.$$

Rééchantillonnage (Facultatif) :

- Simuler, indépendamment conditionnellement à $\{(\xi_i^{(0:k)}, \tilde{\xi}_j^{(k+1)}),\ i, j = 1, \ldots, N\}$, les variables $(I_1^{(k+1)}, \ldots I_N^{(k+1)})$ selon la loi discrète sur $\{1, \ldots, N\}$ de probabilités

$$\frac{\omega_1^{(k+1)}}{\sum_j^N \omega_j^{(k+1)}}, \cdots, \frac{\omega_N^{(k+1)}}{\sum_j^N \omega_j^{(k+1)}} \,.$$

- Fixer tous les poids $\omega_i^{(k+1)}$ à une valeur constante pour $i = 1, \ldots, N$.

Si le rééchantillonnage n'est pas appliquer, poser pour $i = 1, \ldots, N$, $I_i^{(k+1)} = i$.

Mise à jour de la trajectoire : pour $i = 1, \ldots, N$,

$$\xi_i^{(0:k+1)} = \left(\xi_{I_i^{(k+1)}}^{(0:k)}, \tilde{\xi}_{I_i^{(k+1)}}^{(k+1)}\right) \,. \tag{1.4.8}$$

Comme précédemment évoqué, l'étape de rééchantillonnage dans l'Algorithme 1.4.1 peut être utilisée systématiquement (pour tous les indices k), mais il est souvent préférable de ne l'effectuer que de temps en temps. Habituellement, le rééchantillonnage est soit utilisé systématiquement avec une fréquence plus faible (agenda déterministe, un instant tous les m, pour un m choisi d'avance) ou à des instants aléatoires basés sur les valeurs des critères que sont le coefficient de variation ou l'entropie des poids définis dans (1.4.5) et (1.4.7), respectivement – nous étudierons cela plus profondément

dans le cadre de méthodes adaptatives au Chapitre 4. Notons qu'en plus des arguments reposant sur la variance de l'approximation par Monte Carlo, il y a généralement un intérêt calculatoire à limiter l'utilisation du rééchantillonnage ; en effet, sauf pour des modèles où l'évaluation des poids incrémentaux est coûteuse (songer aux observations multidimensionnelles en grande dimension, par exemple), le coût de calcul de l'étape de rééchantillonnage peut ne pas être négligeable si elle est implémentée de façon naïve. Les Sections 1.5.1 et 1.5.2 présentent plusieurs implémentations et variantes de l'étape de rééchantillonnage qui rendent ce dernier argument moins problématique.

Le terme *filtre particulaire* est souvent utilisé pour l'Algorithme 1.4.1 bien que la terminologie EPSR soit moins ambigüe, le terme filtrage particulaire étant parfois utilisé de façon générique pour toute méthode de Monte Carlo séquentielle. Gordon et al. (1993) a en effet proposé une version spécifique de l'Algorithme 1.4.1 dans lequel le rééchantillonnage est systématiquement effectué à chaque instant et où le noyau de proposition n'est autre que le noyau a priori $R_k = Q$. Cet algorithme particulier, communément appelé *filtre bootstrap*, est le plus souvent très facile à implémenter car il ne requiert que de simuler sous le noyau de transition Q de la chaîne cachée et d'évaluer la vraisemblance locale g.

Il existe bien sûr toute une gamme de variantes et de raffinements de l'Algorithme 1.4.1. Toutefois une simple remarque serait de faire remarquer que, dans le cas de la méthode EPSR la plus simple présentée en Section 1.1.2, il est possible de rééchantillonner N fois dans une population plus grande de M réalisations intermédiaires, potentiellement selon plusieurs noyaux de proposition. En pratique, cela signifie que l'Algorithme 1.4.1 peut être modifié comme suit aux instants k où le rééchantillonnage est appliqué :

EPS : Pour $i = 1, \dots, N$, simuler α candidats $\tilde{\xi}_{i,1}^{(k+1)}, \dots, \tilde{\xi}_{i,\alpha}^{(k+1)}$ pour chaque noyau de proposition $R_k(\xi_i^{(k)}, \cdot)$.

Rééchantillonnage : Simuler $(N_{k+1}^{1,1}, \dots, N_{k+1}^{1,\alpha}, \dots, N_{k+1}^{N,1}, \dots, N_{k+1}^{N,\alpha})$ selon la distribution multinomiale de paramètre N et de probabilités

$$\frac{\omega_{i,j}^{(k+1)}}{\sum_{l=1}^{N}\sum_{m=1}^{\alpha}\omega_{l,m}^{(k+1)}} \qquad \text{pour } i = 1, \dots, N,\ j = 1, \dots, \alpha \ .$$

Ainsi, tandis que cette forme de rééchantillonnage conserve le nombre de particules constant et égal à N après rééchantillonnage, la population intermédiaire (avant rééchantillonnage) est de taille $M = \alpha \times N$. Bien qu'évidemment plus lourde à implémenter, l'utilisation de α plus grand que 1 diminue la variance de l'approximation de la loi cible obtenue après (cf Théorème A.2.2). Cette amélioration perdure après l'étape de rééchantillonnage (cf Théorème A.3.2).

Remarque 1.4.1 (Interprétation Marginale de l'EPS et de l'EPSR). Les deux Algorithmes 1.3.1 et 1.4.1 ont été introduits en tant que méthodes pour simuler des *trajectoires* $\{\xi_i^{(0:k)}\}_{1 \le i \le N}$ qui approximent la distribution lissage joint $\phi_{0:k|k}$. Ceci se fait assez facilement dans le cas de l'EPS (Algorithme 1.3.1), les trajectoires étant simplement étendues indépendamment les unes des autres lors de nouvelles simulations. Lorsque le rééchantillonnage est utilisé, le procédé devient plus compliqué car il est alors nécessaire de dupliquer ou supprimer certaines trajectoires selon (1.4.8).

Cette présentation des méthodes EPS et EPSR a été adoptée parce qu'elle est la façon la plus naturelle d'introduire les méthodes de Monte Carlo séquentielles. Cela ne signifie pas que, lors de l'implémentation de l'algorithme EPSR, il soit nécessaire de stocker l'intégralité des trajectoires. Nous ne prétendons pas non plus que pour k grand, l'approximation de la distribution jointe $\phi_{0:k|k}$ fournie par les trajectoires des

particules $\{\xi_i^{(0:k)}\}_{1\leq i\leq N}$ soit précise (voir à ce sujet la littérature sur l'estimation de fonctionnelles de la trajectoire, par exemple Olsson et al. (2008) et ses références). Le plus souvent, l'Algorithme 1.4.1 est implémenté en ne sauvant que la génération courante des particules $\{\xi_i^{(k)}\}_{1\leq i\leq N}$, et (1.4.8) se simplifie en

$$\xi_i^{(k+1)} = \tilde{\xi}_{I_i^{(k+1)}}^{(k+1)} \qquad i = 1, \ldots, N \ .$$

Dans ce cas, le système de particules $\{\xi_i^{(k)}\}_{1\leq i\leq N}$, avec ses poids associés $\{\omega_i^{(k)}\}_{1\leq i\leq N}$, fournit une approximation de la distribution de filtrage $\phi_{\chi,k}$, qui est la distribution *marginale* de la distribution lissage joint $\phi_{0:k|k}$.

Remarque 1.4.2 (Arbre généalogique des particules et notation). La notation $\xi_i^{(k)}$ peut être ambigüe en présence de rééchantillonnage, car les $k+1$ premiers éléments de la $i^{\text{ème}}$ trajectoire $\xi_i^{(0:k+1)}$ à l'instant $k+1$ ne coïncident pas forcément avec la $i^{\text{ème}}$ trajectoire $\xi_i^{(0:k)}$ à l'instant k. Par convention, $\xi_i^{(k)}$ *désigne toujours le dernier point de la* i*ème trajectoire, telle que simulée à l'instant* k. De la même façon, $\xi_i^{(l:k)}$ est la partie de la même trajectoire qui débute à l'instant l et s'achève au dernier instant (c'est à dire, k). Lorsque besoin sera, nous noterons $\xi_i^{(0:k)}(l)$ l'élément de l'instant l dans la $i^{\text{ème}}$ trajectoire de particules à l'instant k afin d'éviter toute ambigüité.

Pour conclure cette section sur l'algorithme EPSR, nous revisitons brièvement deux des exemples précédemment considérés, afin de souligner les différences de résultats obtenus par les approches EPS et EPSR.

Exemple 1.4.4 (Modèle de Volatilité Stochastique, Suite)**.** Pour illustrer l'efficacité de la stratégie de rééchantillonnage, nous considérons de nouveau le modèle de volatilité stochastique introduit dans l'Exemple 1.3.3, pour lequel le phénomène de dégénérescence des poids de l'approche EPS de base était flagrant dans les Figures 1.9 et 1.10.

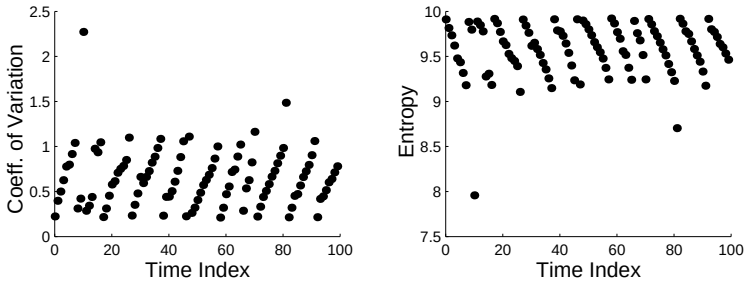


FIGURE 1.11 – Coefficient de variation (gauche) et entropie des poids normalisés en fonction du nombre d'itérations dans le modèle de volatilité stochastique de l'Exemple 1.3.3. Mêmes modèle et données que pour la Figure 1.10. Le rééchantillonnage se produit lorsque le coefficient de variation devient plus grand que 1.

Les Figures 1.11 et 1.12 sont le contrepoint des Figures 1.10 et 1.9, respectivement, lorsque le rééchantillonnage est appliqué dès que le coefficient de variation (1.4.5) des poids normalisés dépasse 1. Notons que la Figure 1.11 représente le coefficient de variation et l'entropie de Shannon calculés pour chaque instant k, *avant rééchantillonnage* aux instants où ce dernier se produit. Contrairement à ce qui se produisait dans l'échantillonnage préférentiel basique, les histogrammes des poids normalisés tracés dans la Figure 1.12 sont remarquablement similaires. Une autre remarque importante dans cet exemple est que les deux critères (le coefficient de variation et l'entropie) sont fortement corrélés. Nous verrons dans le Chapitre 3 que ceci s'explique par le fait que

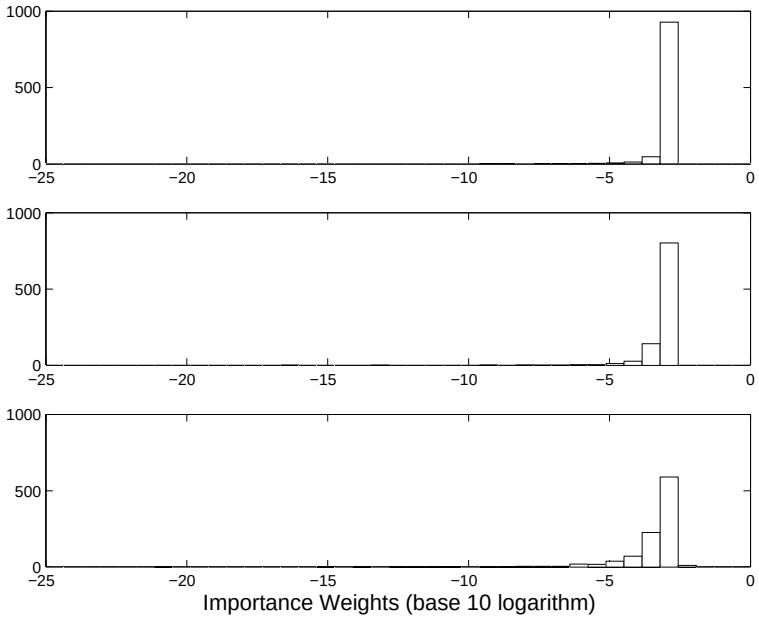


FIGURE 1.12 – Histogrammes des poids normalisés, en logarithme base 10, après (de haut en bas) 1, 10, 100 itérations du modèle de volatilité stochastique de l'Exemple 1.3.3. Mêmes modèles et données que pour la Figure 1.9. Le rééchantillonnage se produit lorsque le coefficient de variation devient plus grand que 1.

ces deux critères approximent deux divergences (certes distinctes, mais pas au point de se contredire systématiquement !) entre les mêmes distributions. Activer le rééchantillonnage dès que l'entropie devient plus faible que, disons, 9.2, serait ainsi presque équivalent au rééchantillonnage se produisant, en moyenne, un instant sur 10.

Exemple 1.4.5 (Modèle de Croissance, Suite)**.** Considérons de nouveau le modèle à espace d'état non-linéaire de l'Exemple 1.3.4, dont la variance σ_u^2 du bruit d'état est fixée à 10 ; ceci rend les observations très informatives comparées à la distribution a priori des états cachés. Les Figures 1.13 et 1.14 représentent les distributions de filtrage estimées pour les 31 premiers instants lorsque la méthode EPS est utilisée avec le noyau a priori Q comme noyau de proposition (Figure 1.13), et la Figure 1.14 en est le pendant pour l'algorithme EPSR correspondant avec rééchantillonnage systématique –autrement dit, le filtre bootstrap. Les deux algorithmes emploient 500 particules.

Pour chaque instant, les graphiques du haut des Figures 1.13 et 1.14 montrent les *régions de plus haute densité a posteriori (HDP)* correspondant aux distributions de filtrage estimées, où les zones gris clair contiennent 95% de la masse de probabilité et les zones plus sombres correspondent à 50% de cette même masse de probabilité. Ces régions HDP sont basées sur une estimation de la densité par noyaux (à l'aide du noyau d'Epanechnikov avec largeur de fenêtre 0.2) calculée sur la base des particules pondérées (c'est à dire, avant le rééchantillonnage dans le cas du filtre bootstrap). Jusqu'à $k = 8$, les deux méthodes donnent des résultats très similaires. Avec l'algorithme EPS, toutefois, le bas de la Figure 1.13 montre que les poids dégénèrent rapidement. Rappelons que la valeur maximale du coefficient de variation (1.4.5) est $\sqrt{N-1}$, qui est environ 22.3 dans le cas de la Figure 1.13. Ainsi, pour $k = 6$ et tous les instants après $k = 12$, le bas de la Figure 1.13 signifie que presque tous les poids normalisés sauf un sont nuls : l'estimation filtrée est concentrée en un seul point, qui est parfois sévèrement loin de la trajectoire du véritable état tel qu'indiquée par les croix – et, on peut donc

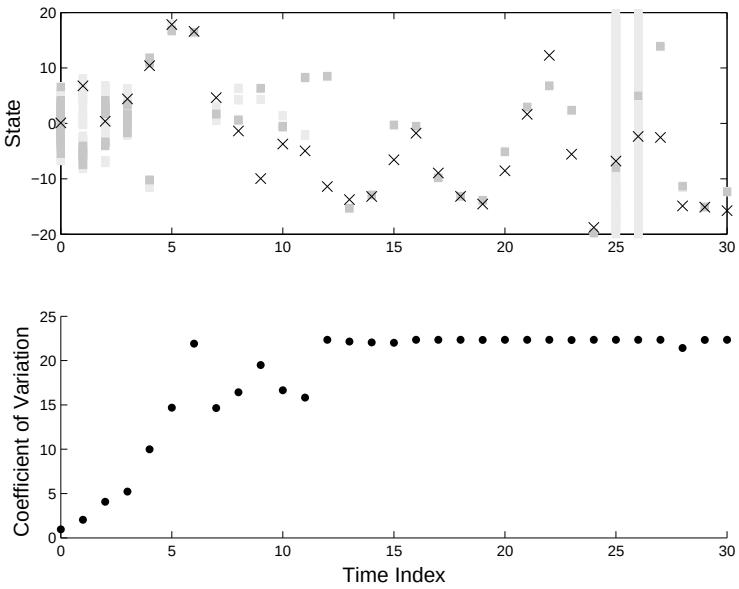
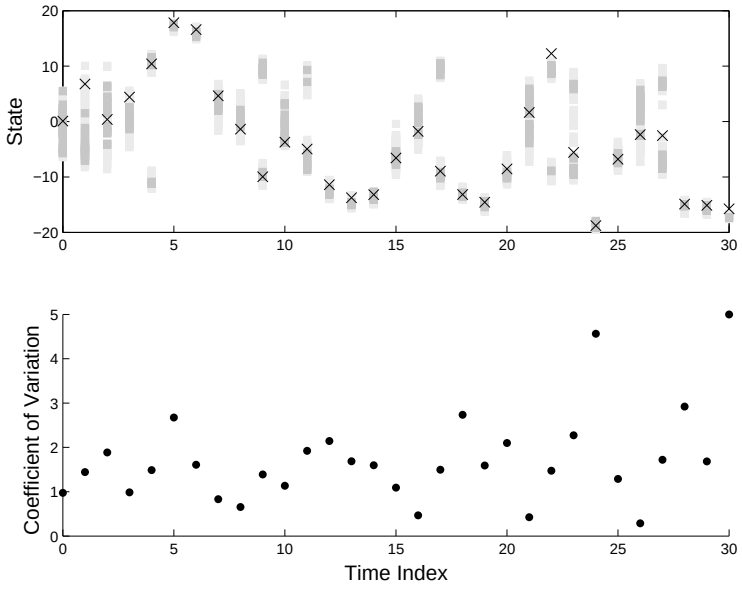


FIGURE 1.13 – Estimations EPS des distributions de filtrage dans le modèle de croissance avec le noyau a priori comme noyau de proposition et 500 particules. En haut : suite des vrais états ($\times$), et régions de plus haute densité posteriori à 95%/50% (gris clair/sombre) des distributions de filtrage estimées. En bas : coefficient de variation des poids normalisés.



p]

FIGURE 1.14 – Même légende que celle de la Figure 1.13, avec les résultats du filtre bootstrap correspondant.

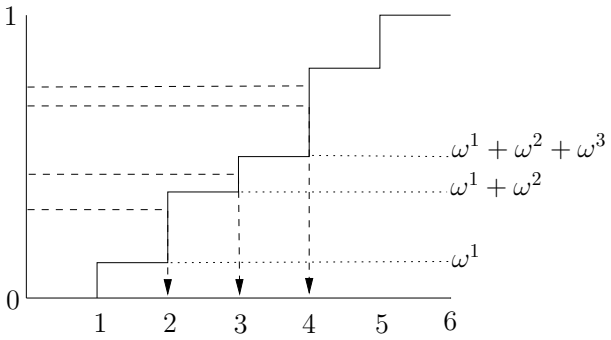


FIGURE 1.15 – Tirage multinomial à partir d'une distribution uniforme par la méthode d'inversion.

le supposer, de la véritable distribution de filtrage, bien qu'en toute rigueur rien ne garantit que la distribution de filtrage soit concentrée autour de l'état caché. Le filtre bootstrap (Figure 1.14), lui, au contraire, semble très stable et fournit des estimations raisonnables de l'état même pour les instants auxquels la distribution de filtrage est fortement bimodale (voir l'Exemple 1.3.4 pour une explication de cette particularité).

1.5 Compléments

Comme expliqué ci-dessus, le rééchantillonnage est un ingrédient clé du succès des méthodes de Monte Carlo séquentielles. Nous présentons ici deux aspects relatifs à cette étape. Tout d'abord, nous montrons qu'il y a plusieurs schémas, basés sur d'astucieux résultats probabilistes, qui peuvent réduire la charge calculatoire associée au rééchantillonnage multinomial. Ensuite, nous examinons des variantes du rééchantillonnage qui atteignent des variances conditionnelles moindres que celle du rééchantillonnage multinomial. Dans ce dernier cas, le but est bien sûr de pouvoir diminuer le nombre de particules sans trop dégrader la qualité de l'approximation – ou encore, à nombre de particules égal, d'améliorer cette qualité.

Tout au long de cette section, nous supposerons que nous devons simuler un échantillon de taille N, $\xi_1, \dots, \xi_N$, parmi un ensemble $\{\tilde{\xi}_1, \dots, \tilde{\xi}_M\}$ généralement plus grand et suivant les poids *normalisés* $\{\omega_1, \dots, \omega_M\}$. Nous noterons $\mathcal{G}$ une tribu (ou σ-algèbre) telle que les variables aléatoires $\omega_1, \dots, \omega_M$ et $\tilde{\xi}_1, \dots, \tilde{\xi}_M$ soient $\mathcal{G}$-mesurables.

1.5.1 Implémentation du rééchantillonnage multinomial

Simuler selon la distribution multinomiale est équivalent à simuler N indices aléatoires $\{I_I, \dots, I_N\}$, indépendants conditionnellement à $\mathcal{G}$, parmi l'ensemble $\{1, \dots, M\}$ et tels que $\mathbb{P}(I_j = i \mid \mathcal{G}) = \omega_i$. Ceci est bien sûr le plus simple exemple de l'utilisation de la *méthode d'inversion* et chaque indice peut être obtenu en simulant d'abord une variable aléatoire U selon la distribution uniforme sur $[0, 1]$ puis en déterminant l'indice I tel que $U \in (\sum_{j=1}^{I-1} \omega_j, \sum_{j=1}^{I} \omega_j]$ (voir la Figure 1.15). Déterminer l'indice I approprié requiert alors en moyenne $\log_2 M$ comparaisons (à l'aide d'une simple recherche par arbre binaire ou une recherche dichotomique, les sommes cumulées des poids normalisés étant, par définition, croissantes). Ainsi, la façon naïve d'implémenter le rééchantillonnage multinomial requiert la simulation de N variables aléatoires uniformes indépendantes et, en moyenne, de l'ordre de $N \log_2 M$ comparaisons – pour être précis, la façon véritablement la plus naïve, par recherche linéaire, requerrait même, en moyenne, de l'ordre de NM comparaisons.

Une solution élégante permettant d'éviter les opérations de recherche répétées consiste à trier au préalable les variables uniformes. Puisque le rééchantillonnage est répété N fois, nous avons besoin de N variable aléatoires uniformes, que nous noterons $U_1, \ldots, U_N$, et nous noterons $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(N)}$ leurs statistiques d'ordre. Il est facilement vérifiable qu'appliquer la méthode d'inversion sur les uniformes *triées* $\{U_{(i)}\}$ ne requiert, dans le pire des cas, que M comparaisons. Le problème est alors que, déterminer les statistiques d'ordre en partant des uniformes non triées $\{U_i\}$ à l'aide d'algorithmes tels que le tri par tas où le quicksort, est une opération qui requiert, au mieux, de l'ordre de $N \log_2 N$ comparaisons (Press et al., 1992, Chapitre 8). Ainsi, sauf dans les cas où $N \ll M$, nous n'avons rien gagné en triant au préalable les variables uniformes avant d'appliquer la méthode d'inversion. Il se trouve néanmoins que deux algorithmes différents permettent de simuler directement les uniformes triées $\{U_{(i)}\}$ avec un nombre d'opérations qui croît linéairement avec N.

Ces deux méthodes sont couvertes en détail dans Devroye (1986, Chapitre 5), et nous ne mentionnons ici que les résultats appropriés, en laissant au lecteur intéressé le soin de consulter Devroye (1986, pp. 207–215) pour les preuves et des références plus approfondies sur le sujet.

Proposition 1.5.1 (Espacements Uniformes)**.** *Soit* $U_{(1)} \leq \ldots \leq U_{(N)}$ *les statistiques d'ordre associées à un échantillon i.i.d. d'une distribution* $\mathrm{U}([0,1])$*. Les incréments*

$$S_i = U_{(i)} - U_{(i-1)}\ , \quad i = 1, \ldots, N\ , \tag{1.5.1}$$

(où $S_1 = U_{(1)}$ *par convention) sont alors appelés espacements uniformes et distribués selon*

$$\frac{E_1}{\sum_{i=1}^{N+1} E_i}, \ldots, \frac{E_N}{\sum_{i=1}^{N+1} E_i}\ ,$$

où $E_1, \ldots, E_{N+1}$ *est une suite de variables aléatoires exponentielles i.i.d..*

Proposition 1.5.2 (Malmquist, 1950)**.** *Soit* $U_{(1)} \leq \ldots \leq U_{(N)}$ *les statistiques d'ordre de* $U_1, U_2, \ldots, U_N$*—une suite de variables aléatoires i.i.d. uniformes* $[0,1]$*. Alors* $U_N^{1/N}$*,* $U_N^{1/N} U_{N-1}^{1/(N-1)}$*, …,* $U_N^{1/N} U_{N-1}^{1/(N-1)} \cdots U_1^{1/1}$ *est distribué comme* $U_{(N)}$*, …,* $U_{(1)}$*.*

Les deux méthodes de simulation associées à ces résultats probabilistes sont résumées dans les Algorithmes 1.5.1 et 1.5.2.

Algorithme 1.5.1 Simulation d'après la Proposition 1.5.1

Pour $i = 1, \ldots, N+1$ **:** Simuler $U_i \sim \mathrm{U}([0,1])$ et poser $E_i = -\log U_i$.

Poser $G = \sum_{i=1}^{N+1} E_i$ **et** $U_{(1)} = E_1/G$**.**

Pour $i = 2, \ldots, n$ **:** $U_{(i)} = U_{(i-1)} + E_i/G$.

Algorithme 1.5.2 Simulation d'après la Proposition 1.5.2

Simuler $V_N \sim \mathrm{U}([0,1])$ et poser $U_{(N)} = V_N^{1/N}$.

Pour $i = N-1$ **à** 1 **:** Simuler $V_i \sim \mathrm{U}([0,1])$ et poser $U_{(i)} = V_i^{1/i} U_{(i+1)}$.

Notons que Devroye (1986) décrit aussi un troisième algorithme légèrement plus compliqué — le tri à seaux de Devroye and Klincsek (1981)— qui a également un coût calculatoire moyen de l'ordre de N. A l'aide de n'importe quelle de ces méthodes, le

coût calculatoire du rééchantillonnage multinomial ne croît plus que linéairement en N et M (au lieu de NM dans une implémentation très naïve), ce qui rend la méthode praticable même lorsqu'un grand nombre de particules est utilisé.

1.5.2 Alternatives au rééchantillonnage multinomial

Au lieu d'utiliser le schéma de rééchantillonnage multinomial, il est aussi possible d'utiliser un schéma de rééchantillonnage (ou de réallocation) différent. Pour $i = 1, \dots, M$, notons N^i le nombre de fois que le $i$ème élément $\tilde{\xi}_i$ est sélectionné. Un schéma de rééchantillonnage sera dit *sans biais par rapport à* $\mathcal{G}$ si

$$\sum_{i=1}^{M} N^i = N \,, \tag{1.5.2}$$

$$\mathbb{E}\left[N^i \middle| \mathcal{G}\right] = N\omega_i, \qquad i = 1, \dots, M \,. \tag{1.5.3}$$

Nous portons ici notre attention sur les techniques de rééchantillonnage qui gardent le nombre de particules constant (voir par exemple Crisan et al., 1999 pour des schémas de rééchantillonnage sans biais avec un nombre aléatoire de particules). Il existe de nombreuses conditions différentes sous lesquelles un schéma de rééchantillonnage est sans biais. Le plus simple schéma sans biais est celui du rééchantillonnage multinomial, pour lequel $(N^1, \dots, N^M)$, conditionnellement à $\mathcal{G}$, suit la distribution multinomiale $\mathrm{Mult}(N, \omega_1, \dots, \omega_N)$. Puisque $I_1, \dots, I_M$ sont i.i.d. conditionnellement à $\mathcal{G}$, il est alors facile d'évaluer la variance conditionnelle de ce schéma :

$$\begin{aligned}\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{\xi}_{I_i}) \middle| \mathcal{G}\right] &= \frac{1}{N}\sum_{i=1}^{M} \omega_i \left[f(\tilde{\xi}_i) - \sum_{j=1}^{M} \omega_j f(\tilde{\xi}_j)\right]^2 \\ &= \frac{1}{N}\left\{\sum_{i=1}^{M} \omega_i f^2(\tilde{\xi}_i) - \left[\sum_{i=1}^{M} \omega_i f(\tilde{\xi}_i)\right]^2\right\} \,. \end{aligned} \tag{1.5.4}$$

Un objectif sensé est d'essayer de construire des schémas de rééchantillonnage pour lesquels la variance conditionnelle $\mathbb{V}(\sum_{i=1}^{N} \frac{N^i}{N} f(\tilde{\xi}_i) \mid \mathcal{G})$ soit aussi petite que possible et, en particulier, plus petite que (1.5.4), de préférence quelque soit la fonction f.

Rééchantillonnage résiduel

Le rééchantillonnage résiduel, ou *rééchantillonnage du reste*, est mentionné par Whitley (1994) (voir aussi Liu and Chen, 1998) comme un moyen simple de diminuer la variance causée par l'étape de rééchantillonnage. Dans ce schéma, pour $i = 1, \dots, M$ nous fixons

$$N^i = \lfloor N\omega_i \rfloor + \bar{N}^i \,, \tag{1.5.5}$$

où $\bar{N}^1, \dots, \bar{N}^M$ sont distribués, conditionnellement à $\mathcal{G}$, selon la distribution multinomiale $\mathrm{Mult}(N - R, \bar{\omega}_1, \dots, \bar{\omega}_N)$ avec $R = \sum_{i=1}^{M} \lfloor N\omega_i \rfloor$ et

$$\bar{\omega}_i = \frac{N\omega_i - \lfloor N\omega_i \rfloor}{N - R}, \quad i = 1, \dots, M \,. \tag{1.5.6}$$

Ce schéma est trivialement sans biais par rapport à $\mathcal{G}$. De façon équivalente, pour toute fonction mesurable f, l'estimateur par échantillonnage résiduel est

$$\frac{1}{N}\sum_{i=1}^{N} f(\xi_i) = \sum_{i=1}^{M} \frac{\lfloor N\omega_i \rfloor}{N} f(\tilde{\xi}_i) + \frac{1}{N}\sum_{i=1}^{N-R} f(\tilde{\xi}_{J^i}) \,, \tag{1.5.7}$$

où $J^1, \dots, J^{N-R}$ sont indépendants conditionnellement à $\mathcal{G}$ avec distribution $\mathbb{P}(J^i = k \mid \mathcal{G}) = \bar{\omega}_k$ pour $i = 1, \dots, N-R$ et $k = 1, \dots, M$. L'estimateur par rééchantillonnage résiduel étant la somme d'un terme qui, sachant $\mathcal{G}$, est déterministe et d'un terme qui implique des étiquettes conditionnellement i.i.d., la variance du rééchantillonnage résiduel est donnée par

$$\begin{aligned} \frac{1}{N^2}\,\mathbb{V}\left[\sum_{i=1}^{N-R} f(\tilde{\xi}_{J^i}) \,\middle|\, \mathcal{G}\right] &= \frac{N-R}{N^2}\,\mathbb{V}\left[f(\tilde{\xi}_{J^1}) \,\middle|\, \mathcal{G}\right] \\ &= \frac{(N-R)}{N^2}\sum_{i=1}^{M}\bar{\omega}_i\left\{f(\tilde{\xi}_i) - \sum_{j=1}^{M}\bar{\omega}_j f(\tilde{\xi}_j)\right\}^2 \\ &= \frac{1}{N}\sum_{i=1}^{M}\omega_i f^2(\tilde{\xi}_i) - \sum_{i=1}^{M}\frac{\lfloor N\omega_i\rfloor}{N^2} f^2(\tilde{\xi}_i) - \frac{N-R}{N^2}\left\{\sum_{i=1}^{M}\bar{\omega}_i f(\tilde{\xi}_i)\right\}^2 . \end{aligned} \tag{1.5.8}$$

L'échantillonnage résiduel domine également l'échantillonnage multinomial en ce sens qu'il a une variance conditionnelle plus faible. En effet, écrivons d'abord

$$\sum_{i=1}^{M}\omega_i f(\tilde{\xi}_i) = \sum_{i=1}^{M}\frac{\lfloor N\omega_i\rfloor}{N} f(\tilde{\xi}_i) + \frac{N-R}{N}\sum_{i=1}^{M}\bar{\omega}_i f(\tilde{\xi}_i) .$$

Notons ensuite que la somme des M nombres $\lfloor N\omega_i\rfloor/N$ plus $(N-R)/N$ égale 1, et que donc cette suite de $M+1$ nombres peut être vue comme une distribution de probabilité. L'inégalité de Jensen appliquée au carré du membre de droite de l'égalité ci-dessus entraîne alors

$$\left\{\sum_{i=1}^{M}\omega_i f(\tilde{\xi}_i)\right\}^2 \le \sum_{i=1}^{M}\frac{\lfloor N\omega_i\rfloor}{N} f^2(\tilde{\xi}_i) + \frac{N-R}{N}\left\{\sum_{i=1}^{M}\bar{\omega}_i f(\tilde{\xi}_i)\right\}^2 .$$

Combiné avec (1.5.8) et (1.5.4), ceci montre que la variance conditionnelle du rééchantillonnage résiduel est toujours plus petite que celle du rééchantillonnage multinomial.

Rééchantillonnage stratifié

La méthode d'inversion pour obtenir une suite de réalisations multinomiales définit une fonction déterministe transformant des variables aléatoires uniformes sur $(0,1)$ $U^1, \dots, U^N$ en indices $I_1, \dots, I_N$. Pour toute fonction f,

$$\sum_{i=1}^{N} f(\tilde{\xi}_{I_i}) = \sum_{i=1}^{N}\Phi_f(U^i) ,$$

où la fonction Φ_f (qui dépend à la fois de f et de $\{\tilde{\xi}_i\}$) est définie, pour tout $u \in (0,1]$, par

$$\Phi_f(u) := f(\tilde{\xi}_{I(u)}), \qquad I(u) = \sum_{i=1}^{M} i\mathbb{1}_{(\sum_{j=1}^{i-1}\omega_j, \sum_{j=1}^{i}\omega_j]}(u) . \tag{1.5.9}$$

Notons que, par construction, $\int_0^1 \Phi_f(u)\,du = \sum_{i=1}^{M}\omega_i f(\tilde{\xi}_i)$. Afin de réduire la variance conditionnelle de $\sum_{i=1}^{N} f(\tilde{\xi}_{I_i})$, nous pouvons changer la façon dont l'échantillon $\{U^1, \dots, U^N\}$ est simulé. Une solution possible, communément utilisée en théorie des sondages, est

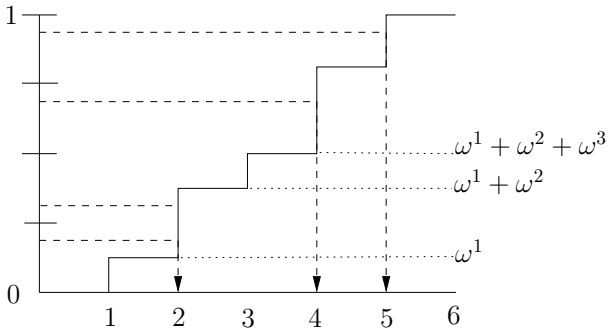


FIGURE 1.16 – Échantillonnage stratifié : l'intervalle $(0, 1]$ est divisé en N intervalles $((i-1)/N, i/N]$. Une réalisation est simulée dans chaque intervalle, indépendamment des réalisations simulées dans les autres intervalles.

basée sur la *stratification* (voir Kitagawa, 1996, et Fearnhead, 1998, Section 5.3, pour un exposé de cette méthode dans le contexte du filtrage particulaire). L'intervalle $(0, 1]$ est partitionné en différentes *strates*, supposées, pour simplifier, être les intervalles $(0, 1] = (0, 1/N] \cup (1/N, 2/N] \cup \cdots \cup (\{N-1\}/N, 1]$. Des partitions plus générales peuvent également être considérées. En particulier, le nombre de partitions n'est pas contraint de sommer à N, et les longueurs des intervalles peut dépendre des ω_i. On simule ensuite un échantillon $\tilde{U}^1, \ldots, \tilde{U}^N$ indépendamment conditionnellement à $\mathcal{G}$ selon la distribution $\tilde{U}^i \sim \mathrm{U}\left(\left(\{i-1\}/N, i/N\right]\right)$ (pour $i = 1, \ldots, N$), et fixons $\tilde{I}^i = I(\tilde{U}^i)$ avec I tel que défini en (1.5.9) (voir Figure 1.16). Par construction, la différence entre $\tilde{N}^i = \sum_{j=1}^N \mathbb{1}_{\{\tilde{I}^j = i\}}$ et la valeur cible (non entière) $N\omega_i$ est plus petite que 1 en valeur absolue. Il s'ensuit également que

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^N f(\tilde{\xi}_{\tilde{I}^i}) \middle| \mathcal{G}\right] &= \mathbb{E}\left[\sum_{i=1}^N \Phi_f(\tilde{U}^i) \middle| \mathcal{G}\right] \\ &= N \sum_{i=1}^N \int_{(i-1)/N}^{i/N} \Phi_f(u)\, du = N \int_0^1 \Phi_f(u)\, du = N \sum_{i=1}^M \omega_i f(\tilde{\xi}_i)\ ,\end{aligned}$$

prouvant que le schéma d'échantillonnage stratifié est sans biais. Puisque $\tilde{U}^1, \ldots, \tilde{U}^N$ sont indépendants conditionnellement à $\mathcal{G}$,

$$\begin{aligned}\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N f(\tilde{\xi}_{\tilde{I}^i}) \middle| \mathcal{G}\right] &= \mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N \Phi_f(\tilde{U}^i) \middle| \mathcal{G}\right] \\ &= \frac{1}{N^2}\sum_{i=1}^N \mathbb{V}\left[\Phi_f(\tilde{U}^i) \middle| \mathcal{G}\right] \\ &= \frac{1}{N}\sum_{i=1}^M \omega_i f^2(\tilde{\xi}_i) - \frac{1}{N}\sum_{i=1}^N \left[N \int_{(i-1)/N}^{i/N} \Phi_f(u) du\right]^2\ ;\end{aligned}$$

ici nous utilisons le fait que $\int_0^1 \Phi_f^2(u)\,du = \int_0^1 \Phi_{f^2}(u)\,du = \sum_{i=1}^M \omega_i f^2(\tilde{\xi}_i)$. Par l'inégalité de Jensen, on a

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \left[N\int_{(i-1)/N}^{i/N} \Phi_f(u)du\right]^2 &\geq \left[\sum_{i=1}^N \int_{(i-1)/N}^{i/N} \Phi_f(u)du\right]^2 \\ &= \left[\sum_{i=1}^M \omega_i f(\tilde{\xi}_i)\right]^2 ,\end{aligned}$$

ce qui prouve que la variance conditionnelle de l'échantillonnage stratifié est toujours inférieure à celle de l'échantillonnage multinomial.

Remarque 1.5.1. Notons que l'échantillonnage stratifié peut être couplé à la méthode d'échantillonnage résiduelle présentée précédemment : la preuve ci-dessus montre qu'utiliser l'échantillonnage stratifié sur les R indices résiduels tirés aléatoirement ne peut que diminuer la variance conditionnelle.

Rééchantillonnage systématique

L'échantillonnage stratifié cherche à réduire la *discrépance*

$$D_N^\star(U^1,\ldots,U^N) := \sup_{a\in(0,1]} \left|\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{(0,a]}(U^i) - a\right|$$

de l'échantillon U selon la distribution uniforme sur $(0,1]$. Ce n'est rien d'autre que la distance de Kolmogorov-Smirnov entre la distribution empirique de l'échantillon et la distribution uniforme. L'inégalité de Koksma-Hlawka (Niederreiter, 1992) montre que quelque soit la fonction f à variation bornée sur $[0,1]$,

$$\left|\frac{1}{N}\sum_{i=1}^N f(u^i) - \int_0^1 f(u)\,du\right| \leq C(f) D_N^\star(u^1,\ldots,u^N) ,$$

où $C(f)$ est la variation de f. Cette inégalité suggère qu'il est souhaitable de construire des suites aléatoires $U^1,\ldots,U^N$ dont la discrépance moyenne soit aussi faible que possible. Ceci fournit une autre explication au gain obtenu par le rééchantillonnage stratifié (comparé à l'échantillonnage multinomial).

Continuer dans cette direction incite à chercher des suites avec une discrépance moyenne encore plus faible. Une telle suite est $U^i = U + (i-1)/N$, où U est simulé selon une distribution $\mathrm{U}((0,1/N])$. En théorie des sondages, cette méthode est connue sous le nom d'*échantillonnage systématique*. Elle fut introduite dans la littérature du filtrage particulaire par Carpenter et al. (1999) mais est mentionnée par Whitley (1994) sous le nom d'échantillonnage *universel*. L'intervalle $(0,1]$ est toujours partitionné en N sous-intervalles $(\{i-1\}/N, i/N]$ et une réalisation est sélectionnée dans chacun d'entre eux, comme dans l'échantillonnage stratifié. Cependant, les réalisations ne sont plus indépendantes, puisqu'elles ont la même position relative au sein de leur propre strate (voir Figure 1.17). Ce schéma d'échantillonnage est évidemment sans biais. Les réalisations n'étant pas sélectionnées indépendamment dans les strates, il n'est toutefois pas possible d'obtenir des formules simples pour la variance conditionnelle (Künsch, 2005). Une conjecture fréquemment faite est que la variance conditionnelle du rééchantillonnage systématique est toujours plus petite que celle du rééchantillonnage multinomial. Ceci est erroné, comme le démontre l'exemple suivant.

Exemple 1.5.1. Considérons le cas où la population initiale de particules $\{\tilde{\xi}_i\}_{1\leq i\leq N}$ est composée de répétitions alternées de seulement deux valeurs distinctes x_0 et x_1, avec même multiplicité (en supposant N pair). En d'autres termes,

$$\{\tilde{\xi}_i\}_{1\leq i\leq N} = \{x_0, x_1, x_0, x_1, \ldots, x_0, x_1\} .$$

Notons $2\omega/N$ la valeur commune des poids normalisés ω_i associés aux $N/2$ particules $\tilde{\xi}_i$ satisfaisant $\tilde{\xi}_i = x_1$, de telle sorte que les particules restantes (qui sont telles que $\tilde{\xi}_i = x_0$) aient un poids commun de $2(1-\omega)/N$. Sans perte de généralité, nous supposons que $1/2 \leq \omega < 1$ et que la fonction d'intérêt f est telle que $f(x_0) = 0$ et $f(x_1) = F$.

Avec le rééchantillonnage multinomial, (1.5.4) montre que la variance conditionnelle de l'estimateur $N^{-1}\sum_{i=1}^N f(\xi_i)$ est donnée par

$$\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N f(\xi_{\text{mult}}i)\middle|\mathcal{G}\right] = \frac{1}{N}(1-\omega)\omega F^2 . \tag{1.5.10}$$

La valeur $2\omega/N$ étant supposée plus grande que $1/N$, il est facile de vérifier que le rééchantillonnage systématique fixe $N/2$ des ξ_i égales à x_1. En fonction de la simulation du décalage initial, ou bien les $N/2$ particules restantes sont *toutes* fixées à x_1, avec probabilité $2\omega - 1$, ou bien sont *toutes* fixées à x_0, avec probabilité $2(1-\omega)$. Ainsi, la variance est celle d'*une seule* simulation selon une Bernoulli multipliée par $N/2$, c'est à dire,

$$\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N f(\xi_{\text{syst}}i)\middle|\mathcal{G}\right] = (\omega - 1/2)(1-\omega)F^2 .$$

Notons que dans ce cas, la variance conditionnelle du rééchantillonnage systématique n'est pas seulement plus grande que (1.5.10) pour la plupart des valeurs de ω (sauf quand ω est très proche de $1/2$), mais elle ne tend même pas vers zéro quand N croît ! Clairement, cette observation dépend fortement de l'ordre dans lequel la population initiale de particules est présentée. Il est intéressant de remarquer que cette propriété est commune aux schémas d'échantillonnage stratifié et systématique, alors que l'approche multinomiale est insensible à cet ordre. Le cas du schéma résiduel, quant à lui, dépend du schéma appliqué aux résidus, selon les mêmes conditions. Dans cet exemple particulier, il est immédiat de vérifier que le rééchantillonnage résiduel (suivi d'une étape multinomiale sur les résidus) et stratifié sont équivalents— ce qui n'est pas le

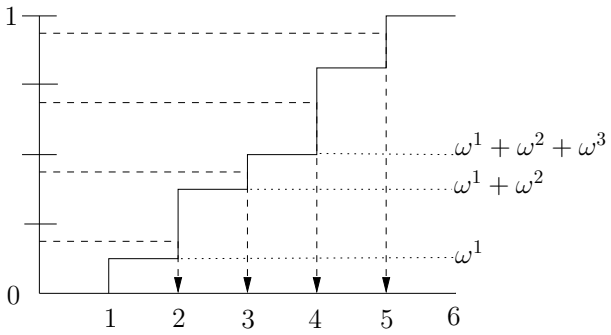


FIGURE 1.17 – Échantillonnage systématique : l'intervalle $(0, 1]$ est divisé en N intervalles $((i-1)/N, i/N]$ et une réalisation est sélectionnée dans chacun d'eux. Contrairement à l'échantillonnage stratifié, chaque réalisation a la même position relative au sein de sa propre strate.

ω	0.51	0.55	0.6	0.65	0.70	0.75
Multinomial	0.050	0.049	0.049	0.048	0.046	0.043
Résiduel, stratifié	0.010	0.021	0.028	0.032	0.035	0.035
Systématique	0.070	0.150	0.200	0.229	0.245	0.250
Systématique avec permutation aléatoire préalable	0.023	0.030	0.029	0.029	0.028	0.025

TABLE 1.1 – Écart-type de différents schémas de rééchantillonnage pour $N = 100$ et $F = 1$. La dernière ligne a été obtenue par simulation, en prenant la moyenne de 100,000 répétitions Monte Carlo.

cas en général — , et reviennent à fixer de façon déterministe $N/2$ particules à la valeur x_1, tandis que les $N/2$ restantes sont simulées par $N/2$ simulations de Bernoulli *conditionnellement indépendantes* dont la probabilité de choisir x_1 est égale à $2\omega - 1$. La variance conditionnelle, tant pour le schéma résiduel que stratifié, est ainsi égale à $N^{-1}(2\omega - 1)(1 - \omega)F^2$. Elle est donc toujours inférieure à (1.5.10), comme prévu par l'étude générale de ces deux méthodes.

Une fois encore, l'échec du rééchantillonnage systématique dans cet exemple est entièrement dû à l'ordre dans lequel les particules sont étiquetées : il est facile de vérifier, du moins empiriquement, que le problème disparaît lorsqu'une permutation aléatoire est appliquée aux particules avant le rééchantillonnage systématique. Le Tableau 1.1 montre également qu'une propriété commune aux schémas résiduel, stratifié, et systématique est d'être très efficace dans des configurations particulières des poids, telles que celle où $\omega = 0.51$, pour laquelle les probabilités de choisir les deux types de particules sont presque égales et où la sélection devient quasi-déterministe. Notons également que la permutation aléatoire préalable compromet quelque peu cette capacité dans le cas du rééchantillonnage systématique.

Dans les applications pratiques des méthodes de Monte Carlo séquentielles, les rééchantillonnages stratifié, résiduel, et systématique ont généralement apporté des résultats comparables. En dépit du manque d'analyse théorique complète de son comportement, le rééchantillonnage systématique est souvent préféré car il présente l'implémentation la plus simple.

CHAPITRE 2

Filtre particulaire auxiliaire

Sommaire

2.1 Introduction

Dans ce chapitre nous analysons le *filtre particulaire auxiliaire* (FPA, APF en anglais), proposé dans Pitt and Shephard (1999), qui s'est avéré être l'une des implémentations les plus utiles et les plus largement adoptées des méthodes MCS. L'analyse en est basée sur les résultats récents sur les échantillons pondérés obtenus par Künsch (2005) et étendus par Chopin (2004) puis Douc and Moulines (2008), ainsi que de la décomposition de l'erreur de Monte Carlo proposée par Del Moral (2004) et affinée ensuite par Olsson et al. (2008). Dans la Section 2.3 nous établissons des théorèmes limite centraux (Théorèmes 2.3.1 et 2.3.2) pour un type de modèle assez général. La convergence est étudiée pour un nombre croissant de particules, et un résultat récent dans le même esprit, a, indépendamment de (Douc et al., 2008), été établi dans (Johansen and Doucet, 2008). Par ailleurs, nous prouvons la convergence uniforme en temps en L^p, sous des conditions plus restrictives (Théorème 2.3.3) d'ergodicité de la chaîne cachée conditionnée aux observations.

Bien que ce chapitre reste comme le Chapitre 1 une introduction à des résultats récents (et non un exposé de résultats originaux), nous y ferons appel à des notions plus avancées que dans le chapitre précédent, et n'hésiterons pas à employer des notations plus précises, concises, et plus théoriques. En particulier, nous faisons désormais appel aux versions trajectorielles des noyaux optimaux et aux récursions correspondantes, tels qu'introduits dans la Section 1.2.5.

2.2 Le filtre particulaire auxiliaire

Rappelons l'algorithme FPA tel que décrit par Pitt and Shephard (1999). Supposons que nous disposons à l'instant k d'un échantillon de particules pondéré $\{(\xi_i^{(0:k)}, \omega_i^{(k)})\}_{i=1}^N$, fournissant une bonne approximation auto-normalisée $(\Omega^{(k)})^{-1} \sum_{i=1}^N \omega_i^{(k)} \delta_{\xi_i^{(0:k)}}(A)$, $A \in \mathcal{X}^{\otimes(k+1)}$, de $\phi_{\chi,0:k|k}$, où $\Omega^{(k)} := \sum_{i=1}^N \omega_i^{(k)}$. Ainsi, lorsque l'observation y_{k+1} devient disponible, une approximation de $\phi_{\chi,0:k+1|k+1}$ est obtenue en injectant la mesure empirique $\phi^N_{\chi,0:k|k}$ dans la récursion (1.2.15), amenant pour tout $A \in \mathcal{X}^{\otimes(k+1)}$ l'approximation

$$\bar{\phi}^N_{\chi,0:k+1|k+1}(A) := \sum_{i=1}^N \frac{\omega_i^{(k)} L_{\mathrm{p},k}(\xi_i^{(0:k)}, \mathsf{X}^{k+2})}{\sum_{j=1}^N \omega_j^{(k)} L_{\mathrm{p},k}(\xi_j^{(0:k)}, \mathsf{X}^{k+2})} L^*_{\mathrm{p},k}(\xi_i^{(0:k)}, A) , \quad A \in \mathcal{X}^{\otimes(n+1)} . \tag{2.2.1}$$

de $\phi_{\chi,0:k+1|k+1}$, qui s'écrit également en faisant figurer les poids d'ajustement optimaux

$$\bar{\phi}^N_{\chi,0:k+1|k+1}(A) := \sum_{i=1}^N \frac{\omega_i^{(k)} \Psi^{*\mathrm{p},(k)}(\xi_i^{(0:k)}),}{\sum_{j=1}^N \omega_j^{(k)} \Psi^{*\mathrm{p},(k)}(\xi_j^{(0:k)}),} L^*_{\mathrm{p},k}(\xi_i^{(0:k)}, A) . \tag{2.2.2}$$

Nous avons ici utilisé les versions trajectorielles du noyau optimal $L^*_{\mathrm{p},k}$ et des poids optimaux tels que présentés dans la Section 1.2.4. Puisque nous voulons établir un échantillon pondéré ciblant $\phi_{\chi,0:k+1|k+1}$ (i.e. consistant et asymptotiquement normal au sens de Douc and Moulines (2008)), il nous faut trouver une façon adéquate de simuler sous $\bar{\phi}^N_{\chi,0:k+1|k+1}$ conditionnellement à $\{(\xi_i^{(0:k)}, \omega_i^{(k)})\}_{i=1}^N$. Dans la plupart des cas, il est possible — mais informatiquement coûteux — de simuler directement selon $\bar{\phi}^N_{\chi,0:k+1|k+1}$ à l'aide de l'algorithme d'acceptation-rejet auxiliaire (voir Hürzeler and Künsch, 1998). Comme indiqué dans la discussion par Künsch (2005, p. 1988), la probabilité moyenne d'acceptation est toutefois inversement proportionnelle à $\|g_{k+1}\|_{\mathsf{X},\infty}$, qui peut-être particulièrement grande si les observations sont informatives. Une solution informatiquement moins coûteuse consiste à simuler un échantillon pondéré ciblant $\bar{\phi}^N_{\chi,0:k+1|k+1}$ en simulant selon la distribution d'échantillonnage préférentiel

$$\rho^N_{0:k+1}(A) := \sum_{i=1}^N \frac{\omega_i^{(k)} \psi_i^{(k)}}{\sum_{j=1}^N \omega_j^{(k)} \psi_j^{(k)}} R_k^{\mathrm{p}}(\xi_i^{(0:k)}, A) , \quad A \in \mathcal{X}^{\otimes(k+2)} . \tag{2.2.3}$$

Ici, $\psi_i^{(k)}$, $1 \leq i \leq N$, sont des nombres positifs appelés *poids multiplicatifs d'ajustement* ou, plus court, poids d'ajustement. (Pitt and Shephard, 1999, utilise le terme *poids de première étape*). Il est important de noter que, sans le degré de liberté supplémentaire (par rapport à l'EPSR présenté en Section 1.4) apporté ces poids d'ajustement, il est vain d'espérer construire une loi de proposition $\rho^N_{0:k+1}$ qui corresponde parfaitement à la loi cible $\bar{\phi}^N_{\chi,0:k+1|k+1}$. En effet, se priver de ces poids d'ajustement, comme le fait l'EPSR standard et de nombreuses variantes, revient à les prendre uniformément égaux à 1, c'est à dire

$$\sum_{i=1}^N \frac{\omega_i^{(k)}}{\sum_{j=1}^N \omega_j^{(k)}} R_k^{\mathrm{p}}(\xi_i^{(0:k)}, A) , \quad A \in \mathcal{X}^{\otimes(k+2)} , \tag{2.2.4}$$

mélange dont les poids sont fixés (conditionnellement à l'échantillon précédent) et ne permettent donc pas d'égaler ceux du mélange cible présenté en (2.2.2).

Dans ce chapitre nous considérons des poids d'ajustement de la forme

$$\psi_i^{(k)} = \Psi^{(k)}(\xi_i^{(0:k)}) \tag{2.2.5}$$

pour une fonction $\Psi^{(k)} : \mathsf{X}^{k+1} \to \mathbb{R}^+$. De plus, le noyau de proposition R_k^{p} est, pour tout $x_{0:k} \in \mathsf{X}^{k+1}$ et $A \in \mathcal{X}^{\otimes(k+2)}$, de la forme

$$R_k^{\mathrm{p}}(x_{0:k}, A) = \int_A \delta_{x_{0:k}}(\mathrm{d}x'_{0:k})\, R_k(x_k, \mathrm{d}x'_{k+1})$$

où R_k est tel que $Q(x, \cdot) \ll R_k(x, \cdot)$ pour tout $x \in \mathsf{X}$. D'autres choix sont possibles, et il est tout à fait envisageable de considérer des noyaux R_k^{p} qui changent des composantes précédentes de la trajectoire. De tels noyaux sont d'ailleurs au coeur des méthodes de *resample and move* de Berzuini and Gilks (2001) ou de *block sampling* de Doucet et al. (2006) – ce dernier englobant ce premier – qui ont pour but de lutter contre la dégénérescence l'échantillon. Nous nous en tiendrons toutefois au cas mentionné précédemment, par simplicité et parcequ'il suffit à traiter la majeure partie des cas. Ainsi, simuler selon $R_k^{\mathrm{p}}(x_{0:k}, \cdot)$ s'obtient en étendant la trajectoire $x_{0:k} \in \mathsf{X}^{k+1}$ avec une composante supplémentaire obtenue en simulant selon $R_k(x_k, \cdot)$. On vérifie facilement (voir Cappé et al., 2005, p. 256, pour de plus amples détails) que pour tout $x_{0:k+1} \in \mathsf{X}^{k+2}$,

$$w_{k+1}(x_{0:k+1}) := \frac{\mathrm{d}\bar{\phi}^N_{\chi,0:k+1|k+1}}{\mathrm{d}\rho^N_{0:k+1|k+1}}(x_{0:k+1}) \propto \sum_{i=1}^N \mathbb{1}_{\xi_i^{(0:k)}}(x_{0:k}) \frac{g_{k+1}(x_{k+1})}{\psi_i^{(k)}} \frac{\mathrm{d}Q(x_k, \cdot)}{\mathrm{d}R_k(x_k, \cdot)}(x_{k+1})\,. \tag{2.2.6}$$

Un échantillon pondéré mis à jour de particules $\{(\tilde{\xi}_i^{(0:k+1)}, \tilde{\omega}_i^{(k+1)})\}_{i=1}^{M_N}$, ciblant la distribution $\bar{\phi}^N_{\chi,0:k+1|k+1}$, est ainsi obtenu en simulant M_N particules $\tilde{\xi}_i^{(0:k+1)}$, $1 \leq i \leq M_N$, selon la distribution instrumentale $\rho^N_{0:k+1}$ et en associant à ces particules les *poids d'importance* $\tilde{\omega}_i^{(k+1)} := w_{k+1}(\tilde{\xi}_i^{(0:k+1)})$, $1 \leq i \leq M_N$. Finalement, dans une seconde étape *facultative* de rééchantillonnage, un échantillon de particules uniformément pondéré $\{(\tilde{\xi}_i^{(0:k+1)}, 1)\}_{i=1}^N$, ciblant toujours la distribution $\bar{\phi}^N_{\chi,0:k+1|k+1}$, est obtenu en rééchantillonnant N particules parmi $\tilde{\xi}_i^{(0:k+1)}$, $1 \leq i \leq M_N$, selon les poids d'importance normalisés. Notons que les nombres de particules M_N et N de ces deux derniers échantillons peuvent être différents. La procédure est ensuite répétée récursivement (avec $\omega_i^{(k+1)} = 1$, $1 \leq i \leq N$) et est initialisée – comme dans l'EPS et l'EPSR mentionnés Sections 1.3 et 1.4, respectivement – par une étape d'échantillonnage préférentiel classique simulant $\{\xi_i^{(0)}\}_{i=1}^{M_N}$ selon $\rho_0^{\otimes M_N}$, avec $\chi \ll \rho_0$, amenant les poids $\omega_i^{(0)} = w_0(\xi_i^{(0)})$ avec $w_0(x) := g_0(x)\,\mathrm{d}\chi/\mathrm{d}\rho_0(x)$, $x \in \mathsf{X}$. Pour résumer, nous obtenons, selon que la deuxième étape de rééchantillonnage est effectuée ou non, les procédures décrites dans les Algorithmes 2.2.1 et 2.2.2. Nous utiliserons le terme FPA comme nom de famille pour ces deux algorithmes, et y ferons séparément référence en tant que *filtre particulaire auxiliaire à deux étapes de rééchantillonnage* (FPA-D) et *filtre particuliaire auxiliaire à simple étape de rééchantillonnage* (FPA-S). Notons que poser $\psi_i^{(k)} \equiv 1$, $1 \leq i \leq N$ dans l'Algorithme 2.2.2 ramène au filtre bootstrap de Gordon et al. (1993) déjà abordé dans la Section 1.4.1. Les étapes de rééchantillonnage du FPA peuvent bien sûr être implémentées à l'aide de techniques différentes du rééchantillonnage multinomial (e.g. rééchantillonnage *résiduel* ou *systématique*) présentées dans la Section 1.5.2, moyennant des adaptations triviales que nous ne présenterons pas ici. Les résultats de l'analyse qui suit sont établis par une approche générique et peuvent donc être étendus à une large classe de schémas de rééchantillonnage.

La question de savoir si la deuxième étape de rééchantillonnage doit ou non avoir lieu (i.e. s'il est préférable d'utiliser l'algorithme FPA-D plutôt que l'algorithme FPA-S) a été posée par plusieurs auteurs. Les résultats théoriques sur la stabilité de l'approximation particulaire et sur la variance asymptotique présentés dans la section suivante indiquent que la deuxième étape de rééchantillonnage doit être évitée, du moins pour

Algorithme 2.2.1 Filtre Particulaire Auxiliaire à Deux Étapes de Rééchantillonnage (FPA-D)

Ensure: $\{(\xi_i^{(0:k)}, \omega_i^{(k)})\}_{i=1}^N$ ciblant $\phi_{\chi,0:k|k}$.

1: **for** $i = 1, \dots, M_N$ **do** ▷ Première étape
2: Simuler l'indice $I_i^{(k)}$ selon la loi discrète sur $\{1, \dots, N\}$ de probabilités $\left\{\omega_j^{(k)}\psi_j^{(k)} / \sum_{\ell=1}^N \omega_\ell^{(k)}\psi_\ell^{(k)}\right\}_{1\leq j\leq N}$;
3: Simuler $\tilde{\xi}_i^{(k+1)} \sim R_k[\xi_{I_i^{(k)}}^{(k)}, \cdot]$, et
4: fixer $\tilde{\xi}_i^{(0:k+1)} := [\xi_{I_i^{(k)}}^{(0:k)}, \tilde{\xi}_i^{(k+1)}]$ et $\tilde{\omega}_i^{(k+1)} := w_{k+1}(\tilde{\xi}_i^{(0:k+1)})$.
5: **end for**
6: **for** $i = 1, \dots, N$ **do** ▷ Deuxième étape
7: Simuler l'indice $J_i^{(k+1)}$ selon la loi discrète sur $\{1, \dots, M_N\}$ de probabilités $\{\tilde{\omega}_j^{(k+1)} / \sum_{\ell=1}^N \tilde{\omega}_\ell^{(k+1)}\}_{1\leq j\leq N}$, et
8: fixer $\xi_i^{(0:k+1)} := \tilde{\xi}_{J_i^{(k+1)}}^{(0:k+1)}$.
9: Finalement, remettre à 1 les poids : $\omega_i^{(k+1)} = 1$.
10: **end for**
11: Utiliser $\{(\xi_i^{(0:k+1)}, 1)\}_{i=1}^N$ comme approximation de $\phi_{\chi,0:k+1|k+1}$.

Algorithme 2.2.2 Filtre Particulaire Auxiliaire à Simple Étape de Rééchantillonnage (FPA-S)

Ensure: $\{(\xi_i^{(0:k)}, \omega_i^{(k)})\}_{i=1}^N$ ciblant $\phi_{\chi,0:k|k}$.

1: **for** $i = 1, \dots, N$ **do**
2: Simuler l'indice $I_i^{(k)}$ selon la loi discrète sur $\{1, \dots, N\}$ et de probabilités $\{\omega_j^{(k)}\psi_j^{(k)} / \sum_{\ell=1}^N \omega_\ell^{(k)}\psi_\ell^{(k)}\}_{1\leq j\leq N}$;
3: Simuler $\tilde{\xi}_i^{(k+1)} \sim R_k[\xi_{I_i^{(k)}}^{(k)}, \cdot]$, et
4: fixer $\tilde{\xi}_i^{(0:k+1)} := [\xi_{I_i^{(k)}}^{(0:k)}, \tilde{\xi}_i^{(k+1)}]$ et $\tilde{\omega}_i^{(k+1)} := w_{k+1}(\tilde{\xi}_i^{(0:k+1)})$.
5: **end for**
6: Utiliser $\{(\tilde{\xi}_i^{(0:k+1)}, \tilde{\omega}_i^{(k+1)})\}_{i=1}^N$ comme approximation de $\phi_{\chi,0:k+1|k+1}$.

le cas $M_N = N$, car elle ne fait qu'augmenter la variance d'échantillonnage. Ainsi, l'idée selon laquelle cette deuxième étape de rééchantillonnage est nécessaire pour éviter la dégénérescence des poids ne tient donc pas. Récemment, (Johansen and Doucet, 2008) ont abouti à une conclusion similaire.

L'avantage que possède le FPA et que n'ont pas les autres méthodes MCS exposées dans le Chapitre 1 est l'apport d'un degré de liberté supplémentaire par le choix des poids d'ajustement$\psi_i^{(k)}$, permettant ainsi de concentrer l'effort calculatoire sur certaines particules plutôt que sur d'autres. Pitt and Shephard (1999) proposent, dans le cas $R_k \equiv Q$ et $\mathsf{X} = \mathbb{R}^d$, d'approcher ce poids d'ajustement optimal par la fonction $\Psi^{(k)}{}_{\mathrm{P\&S}}(x_{0:k}) := g_{k+1}[\int_{\mathsf{X}} x' \, Q(x_k, \mathrm{d}x')]$, $x_{0:k} \in \mathsf{X}^{k+1}$. Il s'agit d'une approximation plutôt rude de l'espérance d'une fonction par la fonction évaluée en l'espérance. L'analyse ci-après montre que ce choix n'est pas toujours bon asymptotiquement.

2.3 Analyse asymptotique

2.3.1 Consistance et normalité asymptotique

Dans cette section, nous établissons la consistance et la normalité asymptotique des échantillons pondérés fournis par les algorithmes FPA-D et FPA-S. Nous utilisons les techniques introduites dans Douc and Moulines (2008) et rappelées dans l'Annexe A.

Pour tout $k \geq 0$ nous définissons la transformation Φ_k sur l'ensemble des fonctions $\phi_{\chi,0:k|k}$-intégrables :

$$\Phi_k[f](x_{0:k}) := f(x_{0:k}) - \phi_{\chi,0:k|k} f \ , \quad x_{0:k} \in \mathsf{X}^{k+1} \ . \tag{2.3.1}$$

De plus, nous imposons les hypothèses suivantes.

(A1) *Pour tout$k \geq 1$, $\Psi^{(k)} \in \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})$ et $w_k \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})$, où $\Psi^{(k)}$ et w_k sont définis dans (2.2.5) et (2.2.6), respectivement.*

(A2) *i) $\mathsf{A}_0 \subseteq \mathsf{L}^1(\mathsf{X}, \phi_{\chi,0|0})$ est un ensemble propre et $\sigma_0 : \mathsf{A}_0 \to \mathbb{R}^+$ est une fonction satisfaisant, pour tout $f \in \mathsf{A}_0$ et $a \in \mathbb{R}$, $\sigma_0(af) = |a|\sigma_0(f)$.*

ii) L'échantillon initial $\{(\xi_{N,i}^{(0)}, 1)\}_{i=1}^N$ est consistant pour $[\mathsf{L}^1(\mathsf{X}, \phi_{\chi,0|0}), \phi_{\chi,0|0}]$ et asymptotiquement normal pour $[\phi_{\chi,0|0}, \mathsf{A}_0, \mathsf{W}_0, \sigma_0, \gamma_0, \{\sqrt{N}\}_{N=1}^\infty]$.

Theorème 2.3.1. *Supposons* **(A1)** *et* **(A2)**, *avec* $(\mathsf{W}_0, \gamma_0) = [\mathsf{L}^1(\mathsf{X}, \phi_{\chi,0|0}), \phi_{\chi,0|0}]$. *Dans le cadre de l'Algorithme 2.2.1, supposons que la limite $\beta := \lim_{N\to\infty} N/M_N$ existe, où $\beta \in [0,1]$. Définissons récursivement la famille $\{\mathsf{A}_k\}_{k=1}^\infty$ par*

$$\begin{aligned}\mathsf{A}_{k+1} := \Big\{ f \in \mathsf{L}^2(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|) L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \\ L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), w_{k+1} f^2 \in \mathsf{L}^1(\mathsf{X}^{k+2}, \ \phi_{\chi,0:k+1|k+1}) \Big\} \ .\end{aligned} \tag{2.3.2}$$

En outre, définissons récursivement $\{\sigma_k\}_{k=1}^\infty$ et les fonctionnelles $\sigma_k : \mathsf{A}_k \to \mathbb{R}^+$ par

$$\begin{aligned}\sigma_{k+1}^2(f) &:= \phi_{\chi,0:k+1|k+1} \Phi_{k+1}^2[f] \\ &+ \frac{\sigma_k^2\{L_{\mathrm{p},k}(\cdot, \Phi_{k+1}[f])\} + \beta \phi_{\chi,0:k|k}(\Psi^{(k)} R_k^{\mathrm{p}}\{\cdot, w_{k+1}^2 (\Phi_{k+1}[f])^2\}) \, \phi_{\chi,0:k|k} \Psi^{(k)}}{[\phi_{\chi,0:k|k} L_{\mathrm{p},k}(\mathsf{X}^{k+2})]^2} \ .\end{aligned} \tag{2.3.3}$$

Alors, pour tout $k \geq 1$, A_k est un ensemble propre. De plus, chaque échantillon $\{(\xi_{N,i}^{(0:k)}, 1)\}_{i=1}^N$ produit par l'Algorithme 2.2.1 est consistant pour $[\mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \phi_{\chi,0:k|k}]$ et asymptotiquement normal pour $[\phi_{\chi,0:k|k}, \mathsf{A}_k, \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \sigma_k, \phi_{\chi,0:k|k}, \{\sqrt{N}\}_{N=1}^\infty]$.

Démonstration. Rappelons le schéma de mise à jour décrit dans l'Algorithme 2.2.1 et décomposons en quatre étapes distinctes :

$$\begin{aligned}\{(\xi_{N,i}^{(0:k)}, 1)\}_{i=1}^N \xrightarrow{\textbf{I} : \text{Pondération}} \{(\xi_{N,i}^{(0:k)}, \psi_{N,i}^{(k)})\}_{i=1}^N \xrightarrow{\textbf{II} : \text{Rééchantillonnage (premier)}} \{(\hat{\xi}_{N,i}^{(0:k)}, 1)\}_{i=1}^{M_N} \to \\ \xrightarrow{\textbf{III} : \text{Mutation}} \{(\tilde{\xi}_{N,i}^{(0:k+1)}, \tilde{\omega}_{N,i}^{(k+1)})\}_{i=1}^{M_N} \xrightarrow{\textbf{IV} : \text{Rééchantillonnage (second)}} \{(\xi_{N,i}^{(0:k+1)}, 1)\}_{i=1}^N \ ,\end{aligned} \tag{2.3.4}$$

où nous avons fixé $\hat{\xi}_{N,i}^{(0:k)} := \xi_{N,I_{N,i}^{(k)}}^{(0:k)}$, $1 \leq i \leq M_N$. Nous établissons maintenant les propriétés asymptotiques décrites dans le Théorème 2.3.1 en construisant une chaîne d'applications de (Douc and Moulines, 2008, Théorèmes 1–4). Nous procédons par induction : supposons que l'échantillon de particules uniformément pondéré $\{(\xi_{N,i}^{(0:k)}, 1)\}_{i=1}^N$ est consistant pour $[\mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \phi_{\chi,0:k|k}]$ et asymptotiquement normal pour $[\phi_{\chi,0:k|k},$

A_k, $\mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \sigma_k, \phi_{\chi,0:k|k}, \{\sqrt{N}\}_{N=1}^{\infty}]$, avec A_k un ensemble propre et σ_k tel que $\sigma_k(af) = |a|\sigma_k(f)$, $f \in \mathsf{A}_k$, $a \in \mathbb{R}$. Nous démontrons, en analysant chacune des étapes **(I–IV)**, qu'une itération de l'algorithme préserve cette propriété. **(I)**. Définissons la mesure

$$\mu_k(A) := \frac{\phi_{\chi,0:k|k}(\Psi^{(k)} \mathbb{1}_A)}{\phi_{\chi,0:k|k} \Psi^{(k)}} , \quad A \in \mathcal{X}^{\otimes(k+1)} .$$

En appliquant (Douc and Moulines, 2008, Théorème 1) pour $R(x_{0:k}, \cdot) = \delta_{x_{0:k}}(\cdot)$, $L(x_{0:k}, \cdot) = \Psi^{(k)}(x_{0:k})\, \delta_{x_{0:k}}(\cdot)$, $\mu = \mu_k$, et $\nu = \phi_{\chi,0:k|k}$, nous concluons que l'échantillon pondéré $\{(\xi_{N,i}^{(0:k)}, \psi_{N,i}^{(k)})\}_{i=1}^N$ est consistant pour $[\{f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k) : \Psi^{(k)}|f| \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})\}, \mu_k] = [\mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k), \mu_k]$. L'égalité est ici basée sur le fait que $\phi_{\chi,0:k|k}(\Psi^{(k)}|f|) = \mu_k|f|\, \phi_{\chi,0:k|k}\Psi^{(k)}$, où le second facteur du membre de droite est borné par l'Hypothèse **(A1)**. De plus, en appliquant (Douc and Moulines, 2008, Théorème 1) nous concluons que$\{(\xi_{N,i}^{(0:k)}, \psi_{N,i}^{(k)})\}_{i=1}^N$ est asymptotiquement normal pour $(\mu_k, \mathsf{A}_{\mathbf{I},k}, \mathsf{W}_{\mathbf{I},k}, \sigma_{\mathbf{I},k}, \gamma_{\mathbf{I},k}, \{\sqrt{N}\}_{N=1}^{\infty})$, où

$$\begin{aligned}
\mathsf{A}_{\mathbf{I},k} &:= \left\{ f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k) : \Psi^{(k)}|f| \in \mathsf{A}_k, \Psi^{(k)} f \in \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \right\} \\
&= \left\{ f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k) : \Psi^{(k)} f \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \right\} , \\
\mathsf{W}_{\mathbf{I},k} &:= \left\{ f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k) : {\Psi^{(k)}}^2 |f| \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \right\}
\end{aligned}$$

sont des ensembles propres, et

$$\begin{aligned}
\sigma^2_{\mathbf{I},k}(f) &:= \sigma_k^2 \left[\frac{\Psi^{(k)}(f - \mu_k f)}{\phi_{\chi,0:k|k}\Psi^{(k)}} \right] = \frac{\sigma_k^2[\Psi^{(k)}(f - \mu_k f)]}{(\phi_{\chi,0:k|k}\Psi^{(k)})^2} , \quad f \in \mathsf{A}_{\mathbf{I},k} , \\
\gamma_{\mathbf{I},k} f &:= \frac{\phi_{\chi,0:k|k}({\Psi^{(k)}}^2 f)}{(\phi_{\chi,0:k|k}\Psi^{(k)})^2} , \quad f \in \mathsf{W}_{\mathbf{I},k} .
\end{aligned}$$

(II). En utilisant (Douc and Moulines, 2008, Théorèmes 3 et 4) nous déduisons que $\{(\hat{\xi}_{N,i}^{(0:k)}, 1)\}_{i=1}^{M_N}$ est consistant pour $[\mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k), \mu_k]$ et asymptotiquement normal pour$[\mu_k, \mathsf{A}_{\mathbf{II},k}, \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k), \sigma_{\mathbf{II},k}, \beta\mu_k, \{\sqrt{N}\}_{N=1}^{\infty}]$, où

$$\mathsf{A}_{\mathbf{II},k} := \left\{ f \in \mathsf{A}_{\mathbf{I},k} : f \in \mathsf{L}^2(\mathsf{X}^{k+1}, \mu_k) \right\} = \left\{ f \in \mathsf{L}^2(\mathsf{X}^{k+1}, \mu_k) : \Psi^{(k)} f \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \right\}$$

est un ensemble propre, et

$$\sigma^2_{\mathbf{II},k}(f) := \beta\mu_k[(f - \mu_k f)^2] + \sigma^2_{\mathbf{I},k}(f) = \beta\mu_k[(f - \mu_k f)^2] + \frac{\sigma_k^2[\Psi^{(k)}(f - \mu_k f)]}{(\phi_{\chi,0:k|k}\Psi^{(k)})^2} , \quad f \in \mathsf{A}_{\mathbf{II},k} .$$

(III). Nous procédons comme pour l'étape **(I)**, mais cette fois avec $\nu = \mu_k$, $R = R_k^{\mathrm{p}}$, et $L(\cdot, A) = R_k^{\mathrm{p}}(\cdot, w_{k+1}\mathbb{1}_A)$, $A \in \mathcal{X}^{\otimes(k+2)}$, obtenant la distribution cible

$$\mu(A) = \frac{\mu_k R_k^{\mathrm{p}}(w_{k+1}\mathbb{1}_A)}{\mu_k R_k^{\mathrm{p}} w_{k+1}} = \frac{\phi_{\chi,0:k|k} L_{\mathrm{p},k}(A)}{\phi_{\chi,0:k|k} L_{\mathrm{p},k}(\mathsf{X}^{k+2})} = \phi_{\chi,0:k+1|k+1}(A) , \quad A \in \mathcal{X}^{\otimes(k+2)} . \tag{2.3.5}$$

Appliquer (Douc and Moulines, 2008, Théorèmes 1 et 2) entraîne que $\{(\tilde{\xi}_{N,i}^{(k+1)}, \tilde{\omega}_{N,i}^{(k+1)})\}_{i=1}^{M_N}$ est consistant pour

$$\begin{aligned}
&\left[\left\{ f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}), R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k) \right\}, \phi_{\chi,0:k+1|k+1} \right] \\
&\qquad = \left[\mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}), \phi_{\chi,0:k+1|k+1} \right] , \quad (2.3.6)
\end{aligned}$$

où (2.3.6) découle de **(A1)**, puisque

$$\mu_k R_k^{\mathrm{p}}(w_{k+1}|f|)\,\phi_{\chi,0:k|k}\Psi^{(k)} = \phi_{\chi,0:k|k}L_{\mathrm{p},k}(\mathsf{X}^{k+2})\,\phi_{\chi,0:k+1|k+1}|f|\ ,$$

et asymptotiquement normal pour $(\phi_{\chi,0:k+1|k+1}, \mathsf{A}_{\mathbf{III},k+1}, \mathsf{W}_{\mathbf{III},k+1}, \sigma_{\mathbf{III},k+1}, \gamma_{\mathbf{III},k+1}, \{\sqrt{N}\}_{N=1}^{\infty})$. Ici

$$\begin{aligned}
&\mathsf{A}_{\mathbf{III},k+1}\\
&:= \left\{f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|) \in \mathsf{A}_{\mathbf{II},k}, R_k^{\mathrm{p}}(\cdot, w_{k+1}^2 f^2) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k)\right\}\\
&= \Big\{f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|) \in \mathsf{L}^2(\mathsf{X}^{k+1}, \mu_k),\\
&\quad \Psi^{(k)}R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|) \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), R_k^{\mathrm{p}}(\cdot, w_{k+1}^2 f^2) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k)\Big\}\\
&= \Big\{f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|)L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}),\\
&\quad L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), w_{k+1}f^2 \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})\Big\}
\end{aligned}$$

et

$$\begin{aligned}
\mathsf{W}_{\mathbf{III},k+1} &:= \left\{f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}^2|f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \mu_k)\right\}\\
&= \left\{f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : w_{k+1}f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})\right\}
\end{aligned}$$

sont des ensembles propres. Par ailleurs, à l'aide de l'identité (2.3.5) nous obtenons que

$$\mu_k R_k^{\mathrm{p}}(w_{k+1}\Phi_{k+1}[f]) = 0\ ,$$

où Φ_{k+1} est défini dans (2.3.1), entraînant

$$\begin{aligned}
&\sigma^2_{\mathbf{III},k+1}(f)\\
&:= \sigma^2_{\mathbf{II},k}\left\{\frac{R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])}{\mu_k R_k^{\mathrm{p}} w_{k+1}}\right\} + \frac{\beta\mu_k R_k^{\mathrm{p}}(\{w_{k+1}\Phi_{k+1}[f] - R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])\}^2)}{(\mu_k R_k^{\mathrm{p}} w_{k+1})^2}\\
&= \frac{\beta\mu_k(\{R_k^{\mathrm{p}}(w_{k+1}\Phi_{k+1}[f])\}^2)}{(\mu_k R_k^{\mathrm{p}} w_{k+1})^2} + \frac{\sigma_k^2\{\Psi^{(k)}R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])\}}{(\phi_{\chi,0:k|k}\Psi^{(k)})^2(\mu_k R_k^{\mathrm{p}} w_{k+1})^2}\\
&\qquad + \frac{\beta\mu_k R_k^{\mathrm{p}}(\{w_{k+1}\Phi_{k+1}[f] - R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])\}^2)}{(\mu_k R_k^{\mathrm{p}} w_{k+1})^2}\ , \quad f \in \mathsf{A}_{\mathbf{III},k+1}\ .
\end{aligned}$$

Maintenant, appliquer l'égalité

$$\begin{aligned}
\{R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])\}^2 + R_k^{\mathrm{p}}(\cdot, \{w_{k+1}\Phi_{k+1}[f] - R_k^{\mathrm{p}}(\cdot, w_{k+1}\Phi_{k+1}[f])\}^2)&\\
= R_k^{\mathrm{p}}(\cdot, w_{k+1}^2\Phi_{k+1}^2[f])&
\end{aligned}$$

aboutit à la variance

$$\sigma^2_{\mathbf{III},k+1}(f) = \frac{\beta\phi_{\chi,0:k|k}\{\Psi^{(k)}R_k^{\mathrm{p}}(\cdot, w_{k+1}^2\Phi_{k+1}^2[f])\}\,\phi_{\chi,0:k|k}\Psi^{(k)} + \sigma_k^2\{L_{\mathrm{p},k}(\cdot, \Phi_{k+1}[f])\}}{[\phi_{\chi,0:k|k}L_{\mathrm{p},k}(\mathsf{X}^{k+2})]^2}\ ,\ . \tag{2.3.7}$$

pour tout $f \in \mathsf{A}_{\mathbf{III},k+1}$. Enfin, pour tout $f \in \mathsf{W}_{\mathbf{III},k+1}$,

$$\gamma_{\mathbf{III},k+1}f := \frac{\beta\mu_k R_k^{\mathrm{p}}(w_{k+1}^2 f)}{(\mu_k R_k^{\mathrm{p}} w_{k+1})^2} = \frac{\beta\phi_{\chi,0:k+1|k+1}(w_{k+1}f)\,\phi_{\chi,0:k|k}\Psi^{(k)}}{\phi_{\chi,0:k|k}L_{\mathrm{p},k}(\mathsf{X}^{k+2})}\ .$$

(IV). La consistance pour $[\mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}), \phi_{\chi,0:k+1|k+1}]$ de l'échantillon de particules uniformément pondéré $\{(\xi_{N,i}^{(0:k+1)}, 1)\}_{i=1}^N$ découle de (Douc and Moulines, 2008, Théorème 3). Qui plus est, appliquer (Douc and Moulines, 2008, Théorème 4) entraîne que ce même échantillon est asymptotiquement normal pour $[\phi_{\chi,0:k+1|k+1}, \mathsf{A}_{\mathbf{IV},k+1}, \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}), \sigma_{\mathbf{IV},k+1}, \phi_{\chi,0:k+1|k+1}, \{\sqrt{N}\}_{N=1}^\infty]$, où

$$\begin{aligned}\mathsf{A}_{\mathbf{IV},k+1} &:= \left\{f \in \mathsf{A}_{\mathbf{III},k+1} : f \in \mathsf{L}^2(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})\right\} \\ &= \Big\{f \in \mathsf{L}^2(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|)L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \\ &\qquad L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{A}_k \cap \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), w_{k+1}f^2 \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})\Big\}\end{aligned}$$

est un ensemble propre, et, pour tout $f \in \mathsf{A}_{\mathbf{IV},k+1}$,

$$\sigma^2_{\mathbf{IV},k+1}(f) := \phi_{\chi,0:k+1|k+1}\Phi^2_{k+1}[f] + \sigma^2_{\mathbf{III},k+1}(f) \,,$$

avec $\sigma^2_{\mathbf{III},k+1}(f)$ défini par (2.3.7). Ceci conclue la démonstration. □

Notons que le résultat similaire suivant a été obtenu pour le FPA-S (Algorithme 2.2.2) en cours de démonstration.

Theorème 2.3.2. *Supposons* **(A1)** *et* **(A2)**. *Définissons les familles* $\{\tilde{\mathsf{W}}_k\}_{k=1}^\infty$ *et* $\{\tilde{\mathsf{A}}_k\}_{k=1}^\infty$ *par*

$$\tilde{\mathsf{W}}_k := \left\{f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) : w_{k+1}f \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})\right\} \,, \quad \tilde{\mathsf{W}}_0 := \mathsf{W}_0 \,,$$

et

$$\begin{aligned}\tilde{\mathsf{A}}_{k+1} := \Big\{&f \in \mathsf{L}^1(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1}) : R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|)L_{\mathrm{p},k}(\cdot, |f|) \in \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \\ &L_{\mathrm{p},k}(\cdot, |f|) \in \tilde{\mathsf{A}}_k, [L_{\mathrm{p},k}(\cdot, |f|)]^2 \in \mathsf{W}_k, w_{k+1}f^2 \in \mathsf{L}^1(\mathsf{X}^{k+2}, \ \phi_{\chi,0:k+1|k+1})\Big\} \,. \quad (2.3.8)\end{aligned}$$

Par ailleurs, définissons récursivement la famille $\{\tilde{\sigma}_k\}_{k=1}^\infty$ *de fonctionnelles* $\tilde{\sigma}_k : \mathsf{A}_k \to \mathbb{R}^+$ *par*

$$\tilde{\sigma}^2_{k+1}(f) := \frac{\tilde{\sigma}^2_k\{L_{\mathrm{p},k}(\cdot, \Phi_{k+1}[f])\} + \phi_{\chi,0:k|k}(\Psi^{(k)}R_k^{\mathrm{p}}\{\cdot, w^2_{k+1}(\Phi_{k+1}[f])^2\})\,\phi_{\chi,0:k|k}\Psi^{(k)}}{[\phi_{\chi,0:k|k}L_{\mathrm{p},k}(\mathsf{X}^{k+2})]^2} \,, \quad (2.3.9)$$

et les mesures $\{\tilde{\gamma}_k\}_{k=1}^\infty$ *par*

$$\tilde{\gamma}_{k+1}f := \frac{\phi_{\chi,0:k+1|k+1}(w_{k+1}f)\,\phi_{\chi,0:k|k}\Psi^{(k)}}{\phi_{\chi,0:k|k}L_{\mathrm{p},k}(\mathsf{X}^{k+2})} \,, \quad f \in \tilde{\mathsf{W}}_{k+1} \,.$$

Alors, pour tout $k \geq 1$, $\tilde{\mathsf{A}}_k$ *est un ensemble propre. De plus, chaque échantillon* $\{(\tilde{\xi}_{N,i}^{(0:k)}, \tilde{\omega}_{N,i}^{(k)})\}_{i=1}^N$ *produit par l'Algorithme 2.2.2 est consistant pour* $[\mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}), \phi_{\chi,0:k|k}]$ *et asymptotiquement normal pour* $[\phi_{\chi,0:k|k}, \tilde{\mathsf{A}}_k, \tilde{\mathsf{W}}_k, \tilde{\sigma}_k, \tilde{\gamma}_k, \{\sqrt{N}\}_{N=1}^\infty]$.

Sous l'hypothèse que la fonction de vraisemblance locale g_k et la fonction de poids d'importance w_k sont bornées, on peut montrer que les TLC établis dans les Théorèmes 2.3.1 et 2.3.2 couvrent toute fonction à moment d'ordre deux, par rapport à la distribution de lissage joint finis, finis ; c'est à dire, sous ces hypothèses, les contraintes supplémentaires sur les ensembles (2.3.2) et (2.3.8) sont automatiquement satisfaites, comme le précise le corollaire ci-dessous.

(A3) *Pour tout* $k \geq 0$, $\|g_k\|_{\mathsf{X},\infty} < \infty$ *et* $\|w_k\|_{\mathsf{X}^{k+1},\infty} < \infty$.

Corollaire 2.3.1. *Supposons* **(A3)** *et soit* $\{\mathsf{A}_k\}_{k\geq 0}$ *et* $\{\tilde{\mathsf{A}}_k\}_{k\geq 0}$ *définis par* (2.3.2) *et* (2.3.8)*, respectivement, avec* $\tilde{\mathsf{A}}_0 = \mathsf{A}_0 := \mathsf{L}^2(\mathsf{X}, \phi_{\chi,0|0})$*. Alors, pour tout* $k \geq 1$*,* $\mathsf{A}_k = \mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})$ *et* $\mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \subseteq \tilde{\mathsf{A}}_k$*.*

Démonstration. Nous choisissons $f \in \mathsf{L}^2(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})$ et montrons que les contraintes de l'ensemble A_{k+1} défini en (2.3.2) sont satisfaites sous l'Hypothèse **(A3)**. Tout d'abord, par l'inégalité de Jensen,

$$\begin{aligned}
&\phi_{\chi,0:k|k}[R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|)L_{\mathrm{p},k}(\cdot, |f|)] \\
&\qquad = \phi_{\chi,0:k|k}\{\Psi^{(k)}[R_k^{\mathrm{p}}(\cdot, w_{k+1}|f|)]^2\} \\
&\qquad \leq \phi_{\chi,0:k|k}[\Psi^{(k)} R_k^{\mathrm{p}}(\cdot, w_{k+1}^2 f^2)] \\
&\qquad = \phi_{\chi,0:k|k} L_{\mathrm{p},k}(w_{k+1} f^2) \\
&\qquad \leq \|w_{k+1}\|_{\mathsf{X}^{k+2},\infty}\, \phi_{\chi,0:k|k} L_{\mathrm{p},k}(\mathsf{X}^{k+2})\, \phi_{\chi,0:k+1|k+1}(f^2) < \infty\,,
\end{aligned}$$

et, de la même façon,

$$\phi_{\chi,0:k|k}\{[L_{\mathrm{p},k}(\cdot, |f|)]^2\} \leq \|g_{k+1}\|_{\mathsf{X},\infty}\, \phi_{\chi,0:k|k} L_{\mathrm{p},k}(\mathsf{X}^{k+2})\, \phi_{\chi,0:k+1|k+1}(f^2) < \infty\,.$$

De ceci, et de la borne

$$\phi_{\chi,0:k+1|k+1}(w_{k+1} f^2) \leq \|w_{k+1}\|_{\mathsf{X}^{k+2},\infty}\, \phi_{\chi,0:k+1|k+1}(f^2) < \infty\,,$$

nous concluons que $\mathsf{A}_{k+1} = \mathsf{L}^2(\mathsf{X}^{k+2}, \phi_{\chi,0:k+1|k+1})$. Pour démontrer que $\mathsf{L}^2(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k}) \subseteq \tilde{\mathsf{A}}_k$, notons que l'Hypothèse **(A3)** implique $\tilde{\mathsf{W}}_k = \mathsf{L}^1(\mathsf{X}^{k+1}, \phi_{\chi,0:k|k})$ et réutilisons les arguments ci-dessus. □

Il est intéressant de noter que les expressions $\tilde{\sigma}_{k+1}^2(f)$ et $\sigma_{k+1}^2(f)$ diffèrent, pour $\beta = 1$, *uniquement en leur terme additif* $\phi_{\chi,0:k+1|k+1}\Phi_{k+1}^2[f]$, c'est à dire, la variance de f sous $\phi_{\chi,0:k+1|k+1}$. Cette quantité représente le coût de l'introduction de la deuxième étape de rééchantillonnage, qui avait pour but d'éviter la dégénérescence de l'approximation particulaire. Dans la section 2.3.2 nous prouverons que les approximations produites par FPA-S sont déjà stables, et que le rééchantillonnage supplémentaire de FPA-D est superflu. Ainsi, il est établi (ce qui est confirmé par les travaux de Johansen and Doucet (2008) menés indépendamment) que la deuxième étape de rééchantillonnage ne doit pas être effectuée.

2.3.2 Bornes L^p et biais

Dans cette section nous examions, sous des conditions de régularité adéquates et pour une population finie de particules, les erreurs des approximations obtenues par le FPA en termes de bornes L^p et de bornes sur le biais. Nous faisons précéder le résultat principal de quelques définitions et hypothèses. Notons $\mathcal{B}_{\mathrm{b}}(\mathcal{X}^m)$ l'espace des fonctions mesurables bornées sur X^m munis de la norme du supremum $\|f\|_{\mathsf{X}^m,\infty} := \sup_{x\in\mathsf{X}^m} |f(x)|$. Soit, pour tout $f \in \mathcal{B}_{\mathrm{b}}(\mathcal{X}^m)$, la *semi-norme d'oscillation* (aussi nommé le *module global de continuité*) définie par $\mathrm{osc}(f) := \sup_{(x,x')\in\mathsf{X}^m\times\mathsf{X}^m} |f(x) - f(x')|$. De plus, notons $\|X\|_p := \mathbb{E}^{1/p}[|X|^p]$ la norme L^p d'une variable aléatoire X. Lorsque nous considérons des sommes, nous utilisons la convention standard $\sum_{k=a}^b c_k = 0$ si $b < a$. Par la suite nous supposerons que toutes les mesures $Q(x,\cdot)$, $x \in \mathsf{X}$, ont pour densité $q(x,\cdot)$ par rapport à une mesure dominante commune μ sur $(\mathsf{X}, \mathcal{X})$. En outre, nous supposerons l'hypothèse suivante vérifiée.

(A4) *i)* $\epsilon_- := \inf_{(x,x')\in\mathsf{X}^2} q(x,x') > 0$, $\epsilon_+ := \sup_{(x,x')\in\mathsf{X}^2} q(x,x') < \infty$.

ii) Pour tout $y \in \mathsf{Y}$, $\int_{\mathsf{X}} g(x,y)\,\mu(\mathrm{d}x) > 0$.

Sous **(A4)**, définissons

$$\rho := 1 - \frac{\epsilon_-}{\epsilon_+}\,. \tag{2.3.10}$$

(A5) *Pour tout* $k \geq 0$, $\left\|\Psi^{(k)}\right\|_{\mathsf{X}^{k+1},\infty} < \infty$.

L'Hypothèse **(A4)** est désormais standard et souvent satisfaite lorsque l'espace d'état X est compact, et implique que la chaîne cachée, lorsqu'elle évolue conditionnellement aux observations, est géométriquement ergodique avec une vitesse de mélange donnée par $\rho < 1$. Pour un traitement complet de telles propriétés de stabilité dans le cadre des modèles à espace d'états, nous renvoyons à Del Moral (2004). Enfin, soit $\mathcal{C}_i(\mathsf{X}^{n+1})$ l'ensemble des fonctions bornées mesurables f sur X^{n+1} de type $f(x_{0:n}) = \bar{f}(x_{i:n})$ pour une quelconque fonction $\bar{f} : \mathsf{X}^{n-i+1} \to \mathbb{R}$. Dans ce cadre, nous avons le résultat suivant.

Theorème 2.3.3. *Supposons* **(A3)**, **(A4)**, **(A5)**, *et soit* $f \in \mathcal{C}_i(\mathsf{X}^{n+1})$ *pour* $0 \leq i \leq n$. *Soit* $\{(\tilde{\xi}^{(0:k)}_{N,i}, \tilde{\omega}^{(k)}_{N,i})\}_{i=1}^{R_N}$ *un échantillon de particules pondéré produit par l'Algorithme* 2.2.r, $r = \{1,2\}$, *avec* $R_N(r) := \mathbb{1}\{r=1\}M_N + \mathbb{1}\{r=2\}N$. *Alors les assertions suivantes sont vraies pour tout* $N \geq 1$ *et* $r = \{1,2\}$.

i) Pour tout $p \geq 2$,

$$\begin{aligned}
&\left\| (\tilde{\Omega}^N_n)^{-1} \sum_{j=1}^{R_N(r)} \tilde{\omega}^{(n)}_{N,j} f_i(\tilde{\xi}^{(0:n)}_{N,j}) - \phi_{\chi,0:n|n} f_i \right\|_p \\
&\qquad \leq B_p \frac{\operatorname{osc}(f_i)}{1-\rho} \left[\frac{1}{\epsilon_- \sqrt{R_N(r)}} \sum_{k=1}^{n} \frac{\|w_k\|_{\mathsf{X}^{k+1},\infty} \left\|\Psi^{(k-1)}\right\|_{\mathsf{X}^k,\infty}}{\mu g_k} \rho^{0 \vee (i-k)} \right. \\
&\qquad\qquad \left. + \frac{\mathbb{1}\{r=1\}}{\sqrt{N}} \left(\frac{\rho}{1-\rho} + n - i \right) + \frac{\|w_0\|_{\mathsf{X},\infty}}{\chi g_0 \sqrt{R_N(r)}} \rho^i \right],
\end{aligned}$$

ii)

$$\begin{aligned}
&\left| \mathbb{E}\left[(\tilde{\Omega}^N_n)^{-1} \sum_{j=1}^{R_N(r)} \tilde{\omega}^{(n)}_{N,j} f_i(\tilde{\xi}^{(0:n)}_{N,j}) \right] - \phi_{\chi,0:n|n} f_i \right| \\
&\qquad \leq B \frac{\operatorname{osc}(f_i)}{(1-\rho)^2} \left[\frac{1}{R_N(r)\epsilon_-^2} \sum_{k=1}^{n} \frac{\|w_k\|^2_{\mathsf{X}^{k+1},\infty} \left\|\Psi^{(k-1)}\right\|^2_{\mathsf{X}^k,\infty}}{(\mu g_k)^2} \rho^{0 \vee (i-k)} \right. \\
&\qquad\qquad \left. + \frac{\mathbb{1}\{r=1\}}{N} \left(\frac{\rho}{1-\rho} + n - i \right) + \frac{\|w_0\|^2_{\mathsf{X},\infty}}{R_N(r)(\chi g_0)^2} \rho^i \right].
\end{aligned}$$

Ici B_p *et* B *sont des constantes universelles telles que* B_p *ne dépend que de* p, *et* ρ *est défini en* (2.3.10).

En particulier, l'utilisation des bornes du Théorème 2.3.3 pour $i = n$, sous l'hypothèse que les fractions $\|w_k\|_{\mathsf{X}^{k+1},\infty} \|\Psi^{(k-1)}\|_{\mathsf{X}^k,\infty} / \mu g_k$ sont toutes uniformément bornées en k, donne des bornes d'erreur de la distribution de filtrage $\phi_{\chi,0:n|n}$ *bornées uniformément* en n. Il en découle que la première étape de rééchantillonnage est suffisante pour préserver la stabilité de l'échantillon. Ainsi avec l'Algorithme 2.2.2 qui évite la deuxième étape de rééchantillonnage, nous pouvons, puisque les termes centraux de la borne ci-dessus s'annulent dans ce cas, obtenir un contrôle encore *plus précis* de l'erreur L^p pour un nombre de particules fixé.

Afin de démontrer le Théorème 2.3.3, nous établissons un lemme de décomposition, qui requiert les notations suivantes. Définissons, pour $r \in \{1,2\}$ et $R_N(r)$ tels que définis dans le Théorème 2.3.3, la mesure empirique des particules

$$\phi^N_{\chi,k|k}(A) := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi^{(0:k)}_{N,i}} \quad \text{et} \quad \tilde{\phi}^N_{\nu,k}(A) := \frac{1}{\tilde{\Omega}^{(k)}_N}\sum_{i=1}^{R_N(r)} \tilde{\omega}^{(k)}_{N,i}\delta_{\tilde{\xi}_{N,0:k}i}(A) \;, \quad A \in \mathcal{X}^{\otimes(k+1)} \;,$$

qui joue le rôle d'approximation de la distribution de lissage $\phi_{\chi,0:k|k}$. Soit $\mathcal{F}_0 := \sigma(\xi^{(0)}_{N,i}; 1 \leq i \leq N)$; alors l'historique des particules jusqu'aux différentes étapes de l'itération $m+1$, $m \geq 0$, de l'Algorithme 2.2.r, $r \in \{1,2\}$, est modélisé par les filtrations $\hat{\mathcal{F}}_m := \mathcal{F}_m \vee \sigma[I^{N,i}_m; 1 \leq i \leq R_N(r)]$, $\tilde{\mathcal{F}}_{m+1} := \mathcal{F}_m \vee \sigma[\tilde{\xi}_{N,0:m+1}i; 1 \leq i \leq R_N(r)]$, et

$$\mathcal{F}_{m+1} := \begin{cases} \tilde{\mathcal{F}}_{m+1} \vee \sigma(J^{N,i}_{m+1}; 1 \leq i \leq N) \;, & \text{pour } r = 1 \;, \\ \tilde{\mathcal{F}}_{m+1} \;, & \text{pour } r = 2 \;. \end{cases}$$

respectivement. Dans la démonstration ci-après nous décrivons une itération de l'algorithme FPA-D comme l'enchaînement des deux opérations suivantes.

$$\{(\xi^{(0:k)}_{N,i}, \omega^{(k)}_{N,i})\}^N_{i=1} \xrightarrow{\text{Simuler selon } \varphi^N_{k+1}} \{(\tilde{\xi}_{N,0:k+1}i, \tilde{\omega}^{(k+1)}_{N,i})\}^{R_N(r)}_{i=1} \rightarrow \xrightarrow{r=1 \,:\, \text{Simuler selon } \tilde{\phi}^N_{\nu,0:k+1}} \{(\xi^{(0:k+1)}_{N,i}, 1)\}^N_{i=1} \;,$$

où, pour tout $A \in \mathcal{X}^{\otimes(k+2)}$,

$$\varphi^N_{k+1}(A) := \mathbb{P}\left(\tilde{\xi}_{N,0:k+1}i_0 \in A \middle| \mathcal{F}_k\right) = \sum_{j=1}^{N} \frac{\omega^{(k)}_{N,j}\psi^{(k)}_{N,j}}{\sum_{\ell=1}^N \omega^{(k)}_{N,\ell}\psi^{(k)}_{N,\ell}} R^{\mathrm{p}}_k(\xi^{(0:k)}_{N,j}, A) = \frac{\phi^N_{\chi,k|k}[\Psi^{(k)} R^{\mathrm{p}}_k(\cdot, A)]}{\phi^N_{\chi,k|k}\Psi^{(k)}} \;, \tag{2.3.11}$$

pour un indice $i_0 \in \{1, \ldots, R_N(r)\}$ (conditionnellemment à $\mathcal{F}_k$, les particules $\tilde{\xi}_{N,0:k+1}i$, $1 \leq i \leq R_N(r)$, sont i.i.d.). Ici les poids initiaux $\{\omega^{(k)}_{N,i}\}^N_{i=1}$ sont tous égaux à un pour $r = 1$. La seconde opération est valide puisque, pour tout $i_0 \in \{1, \ldots, N\}$,

$$\mathbb{P}\left(\xi^{(0:k+1)}_{N,i_0} \in A \middle| \tilde{\mathcal{F}}_{k+1}\right) = \sum_{j=1}^{R_N(r)} \frac{\tilde{\omega}^{(k+1)}_{N,j}}{\tilde{\Omega}^{(k+1)}_N} \delta_{\tilde{\xi}_{N,0:k+1}j}(A) = \tilde{\phi}^N_{\nu,0:k+1}(A) \;, \quad A \in \mathcal{X}^{\otimes(k+2)} \;.$$

Le fait que l'évolution des particules puisse être décrite par deux opérations de simulation i.i.d. rend possible l'analyse de l'erreur à l'aide de l'inégalité de Marcinkiewicz-Zygmund (voir Petrov, 1995, p. 62).

En utilisant cette dernière, posons, pour tout $1 \leq k \leq n$,

$$\alpha^N_k(A) := \int_A \frac{\mathrm{d}\alpha^N_k}{\mathrm{d}\varphi^N_k}(x_{0:k})\, \varphi^N_k(\mathrm{d}x_{0:k}) \;, \quad A \in \mathcal{X}^{\otimes(k+1)} \;, \tag{2.3.12}$$

avec, pour $x_{0:k} \in \mathsf{X}^{k+1}$,

$$\frac{\mathrm{d}\alpha^N_k}{\mathrm{d}\varphi^N_k}(x_{0:k}) := \frac{w_k(x_{0:k}) L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(x_{0:k}, \mathsf{X}^{n+1})\, \phi^N_{\chi,k-1|k-1}\Psi^{(k-1)}}{\phi^N_{\chi,k-1|k-1} L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \;.$$

Nous appliquons ici la convention standard $L_{\mathrm{p},\ell} \cdots L_{\mathrm{p},m} := \mathrm{Id}$ si $m < \ell$. Pour $k = 0$ nous définissons

$$\alpha_0(A) := \int_A \frac{\mathrm{d}\alpha_0}{\mathrm{d}\rho_0}(x_0)\, \rho_0(\mathrm{d}x_0) \;, \quad A \in \mathcal{X} \;,$$

où, pour tout $x_0 \in \mathsf{X}$,

$$\frac{\mathrm{d}\alpha_0}{\mathrm{d}\rho_0}(x_0) := \frac{w_0(x_0) L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1}(x_0, \mathsf{X}^{n+1})}{\chi[g_0 L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1}(\cdot, \mathsf{X}^{n+1})]} \;.$$

De même, pour tout, $0 \leq k \leq n-1$,

$$\beta_k^N(A) := \int_A \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(x_{0:k})\, \tilde{\phi}_{\nu,k}^N(\mathrm{d}x_{0:k}) \;, \quad A \in \mathcal{X}^{\otimes(k+1)} \;, \tag{2.3.13}$$

où, pour $x_{0:k} \in \mathsf{X}^{k+1}$,

$$\frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(x_{0:k}) := \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(x_{0:k}, \mathsf{X}^{n+1})}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \;.$$

Le puissant lemme de décomposition qui suit est une adaptation d'une décomposition similaire établie par Olsson et al. (2008, Lemme 7.2) (le cas EPSR standard), qui est elle même un raffinement d'une décomposition présentée au départ par Del Moral (2004).

Lemme 2.3.1. *Soit $n \geq 0$. Alors, pour tout $f \in \mathcal{B}_{\mathrm{b}}(\mathcal{X}^{n+1})$, $N \geq 1$, et $r \in \{1, 2\}$,*

$$\tilde{\phi}_{\nu,0:n}^N f - \phi_{\chi,0:n|n} f = \sum_{k=1}^{n} A_k^N(f) + I_{N,\{r=1\}} \sum_{k=0}^{n-1} B_k^N(f) + C^N(f) \;, \tag{2.3.14}$$

où

$$A_k^N(f) := \frac{\sum_{i=1}^{R_N(r)} \frac{\mathrm{d}\alpha_k^N}{\mathrm{d}\varphi_k^N}(\tilde{\xi}_{N,0:k} i) \Psi_{k:n}[f](\tilde{\xi}_{N,0:k} i)}{\sum_{j=1}^{R_N(r)} \frac{\mathrm{d}\alpha_k^N}{\mathrm{d}\varphi_k^N}(\tilde{\xi}_{N,0:k} j)} - \alpha_k^N \Psi_{k:n}[f] \;,$$

$$B_k^N(f) := \frac{\sum_{i=1}^{N} \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(\xi_{N,i}^{(0:k)}) \Psi_{k:n}[f](\xi_{N,i}^{(0:k)})}{\sum_{j=1}^{N} \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(\xi_{N,j}^{(0:k)})} - \beta_k^N \Psi_{k:n}[f] \;,$$

$$C^N(f) := \frac{\sum_{i=1}^{N} \frac{\mathrm{d}\beta_{0|n}}{\mathrm{d}\rho_0}(\xi_{N,i}^{(0)}) \Psi_{0:n}[f](\xi_{N,i}^{(0)})}{\sum_{j=1}^{N} \frac{\mathrm{d}\beta_0}{\mathrm{d}\rho_0}(\xi_{N,i}^{(0)})} - \phi_{\chi,0:n|n} \Psi_{0:n}[f] \;,$$

et les opérateurs $\Psi_{k:n} : \mathcal{B}_{\mathrm{b}}(\mathcal{X}^{n+1}) \to \mathcal{B}_{\mathrm{b}}(\mathcal{X}^{n+1})$, $0 \leq k \leq n$, sont, pour des points $\hat{x}_{0:k} \in \mathsf{X}^{k+1}$ fixés, définis par

$$\Psi_{k:n}[f] : x_{0:k} \mapsto \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(x_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(x_{0:k}, \mathsf{X}^{n+1})} - \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \;.$$

Démonstration. Considérons la décomposition

$$\begin{aligned}
\tilde{\phi}_{\nu,0:n}^N f - \phi_{\chi,0:n|n} f = \sum_{k=1}^{n} \left[\frac{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \frac{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1} f}{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \right] \\
+ I_{N,\{r=1\}} \sum_{k=0}^{n-1} \left[\frac{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \frac{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \right] \\
+ \frac{\tilde{\phi}_{\nu,0}^N L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,0}^N L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \phi_{\chi,0:n|n} f \;.
\end{aligned}$$

Nous allons montrer que les trois parties de cette décomposition sont identiques aux trois parties de (2.3.14). Pour $k \geq 1$, à l'aide des définitions (2.3.11) et (2.3.12) de φ_k^N et α_k^N, respectivement, et de la même façon que Olsson et al. (2008, Lemme 7.2), on obtient que

$$
\begin{aligned}
&\frac{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1} L_{\mathrm{p},n-1} f}{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \\
&= \varphi_k^N \left[\frac{w_k(\cdot) L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\cdot) (\phi_{\chi,k-1|k-1}^N \Psi^{(k-1)})}{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \right] \\
&= \varphi_k^N \left[\frac{w_k(\cdot) L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\cdot, \mathsf{X}^{n+1}) (\phi_{\chi,k-1|k-1}^N \Psi^{(k-1)})}{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \left\{ \Psi_{k:n}[f](\cdot) + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \right\} \right] \\
&= \alpha_k^N \left[\Psi_{k:n}[f](\cdot) + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \right] \\
&= \alpha_k^N \Psi_{k:n}[f] + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \,.
\end{aligned}
$$

De plus, par définition,

$$
\frac{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} = \frac{\sum_{i=1}^{R_N(r)} \frac{\mathrm{d}\alpha_k^N}{\mathrm{d}\varphi_k^N}(\tilde{\xi}_{N,0:k} i) \Psi_{k:n}[f](\tilde{\xi}_{N,0:k} i)}{\sum_{j=1}^{R_N(r)} \frac{\mathrm{d}\alpha_k^N}{\mathrm{d}\varphi_k^N}(\tilde{\xi}_{N,0:k} j)} + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \,,
$$

ce qui entraîne

$$
\frac{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \frac{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1} f}{\phi_{\chi,k-1|k-1}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \equiv A_k^N(f) \,.
$$

De même, pour $r = 1$, à l'aide de la définition (2.3.13) de β_k^N,

$$
\begin{aligned}
\frac{\tilde{\phi}_{\nu,0:k}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,0:k}^N L_{\mathrm{p},k-1} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} &= \beta_k^N \left[\frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\cdot)}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \right] \\
&= \beta_k^N \left[\Psi_{k:n}[f](\cdot) + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \right] \\
&= \beta_k^N \Psi_{k:n}[f] + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \,,
\end{aligned}
$$

et en appliquant la relation évidente

$$
\frac{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} = \frac{\sum_{i=1}^{N} \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(\xi_{N,i}^{(0:k)}) \Psi_{k:n}[f](\xi_{N,i}^{(0:k)})}{\sum_{j=1}^{N} \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N}(\xi_{N,j}^{(0:k)})} + \frac{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f(\hat{x}_{0:k})}{L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\hat{x}_{0:k}, \mathsf{X}^{n+1})} \,,
$$

on obtient l'identité

$$
\frac{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\phi_{\chi,k|k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \frac{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,k}^N L_{\mathrm{p},k} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} \equiv B_k^N(f) \,.
$$

L'égalité

$$
\frac{\tilde{\phi}_{\nu,0}^N L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1} f}{\tilde{\phi}_{\nu,0}^N L_{\mathrm{p},0} \cdots L_{\mathrm{p},n-1}(\mathsf{X}^{n+1})} - \phi_{\chi,0:n|n} f \equiv C^N(f)
$$

est obtenue de façon analogue, ce qui conclut la preuve de ce lemme. □

Armés de ce lemme, nous nous tournons maintenant vers la preuve du théorème en lui-même.

Preuve du Théorème 2.3.3. À partir d'ici la preuve est une extension directe de (Olsson et al., 2008, Proposition 7.1). Pour établir la partie (i), on observe que :

– une adaptation triviale de (Olsson et al., 2008, Lemmes 7.3 et 7.4) donne

$$\|\Psi_{k:n}[f_i]\|_{\mathsf{X}^{k+1},\infty} \leq \operatorname{osc}(f_i)\rho^{0\vee(i-k)} , \quad \left\| \frac{\mathrm{d}\alpha_k^N}{\mathrm{d}\varphi_k^N} \right\|_{\mathsf{X}^{k+1},\infty} \leq \frac{\|w_k\|_{\mathsf{X}^{k+1},\infty} \left\|\Psi^{(k-1)}\right\|_{\mathsf{X}^k,\infty}}{\mu g_k(1-\rho)\epsilon_-} . \tag{2.3.15}$$

– En imitant la preuve de (Olsson et al., 2008, Proposition 7.1(i)), c'est à dire, en appliquant l'identité $a/b - c = (a/b)(1-b) + a - c$ à chaque $A_k^N(f_i)$ et en utilisant deux fois l'inégalité de Marcinkiewicz-Zygmund avec les bornes (2.3.15), on obtient la borne

$$\sqrt{R_N(r)} \left\|A_k^N(f_i)\right\|_p \leq B_p \frac{\operatorname{osc}(f_i) \|w_k\|_{\mathsf{X}^{k+1},\infty} \left\|\Psi^{(k-1)}\right\|_{\mathsf{X}^k,\infty}}{\mu g_k(1-\rho)\epsilon_-} \rho^{0\vee(i-k)} ,$$

où B_p est une constante ne dépendant que de p. Nous renvoyons le lecteur intéressé à (Olsson et al., 2008, Proposition 7.1) pour les détails.

– Pour $r = 1$, un examen de la preuve de (Olsson et al., 2008, Lemme 7.4) entraîne immédiatement

$$\left\| \frac{\mathrm{d}\beta_k^N}{\mathrm{d}\tilde{\phi}_{\nu,k}^N} \right\|_{\mathsf{X}^{k+1},\infty} \leq \frac{1}{1-\rho} ,$$

et réutiliser les arguments précédents pour $B_k^N(f_i)$ entraîne

$$\sqrt{N} \left\|B_k^N(f_i)\right\|_p \leq B_p \frac{\operatorname{osc}(f_i)}{1-\rho} \rho^{0\vee(i-k)} .$$

– Les arguments ci-dessus s'appliquent directement à $C^N(f_i)$, amenant

$$\sqrt{N} \left\|C^N(f_i)\right\|_p \leq B_p \frac{\operatorname{osc}(f_i) \|w_0\|_{\mathsf{X},\infty}}{\chi g_0(1-\rho)} \rho^i .$$

Nous concluons la preuve de (i) en additionnant.

La preuve de (ii) (qui imite la preuve de (Olsson et al., 2008, Proposition 7.1(ii)) procède de façon similaire ; en effet, répéter les arguments de (i) ci-dessus pour la décomposition $a/b - c = (a/b)(1-b)^2 + (a-c)(1-b) + c(1-b) + a - c$ nous donne les bornes

$$\begin{aligned}
R_N(r) \left|\mathbb{E}\left[A_k^N(f_i)\right]\right| &\leq B \frac{\operatorname{osc}(f_i) \|w_k\|_{\mathsf{X}^{k+1},\infty}^2 \left\|\Psi^{(k-1)}\right\|_{\mathsf{X}^k,\infty}^2}{(\mu g_k)^2(1-\rho)^2\epsilon_-^2} \rho^{0\vee(i-k)} , \\
N \left|\mathbb{E}\left[B_k^N(f_i)\right]\right| &\leq B \frac{\operatorname{osc}(f_i)}{(1-\rho)^2} \rho^{0\vee(i-k)} , \\
N \left|\mathbb{E}\left[C^N(f_i)\right]\right| &\leq B \frac{\operatorname{osc}(f_i) \|w_0\|_{\mathsf{X},\infty}^2}{(\chi g_0)^2(1-\rho)^2} \rho^i .
\end{aligned}$$

Nous renvoyons de nouveau à (Olsson et al., 2008, Proposition 7.1(ii)) pour les détails, et nous concluons la preuve en additionnant. □

CHAPTER 3

Quality criteria for adaptive sequential Monte Carlo

Contents

This chapter is an article published as Cornebise et al. (2008), to the only difference that, for sake of readability, we included the proofs in the exposure rather than postponing them in an appendix

3.1 Introduction

Easing the role of the user by tuning automatically the key parameters of *sequential Monte Carlo* (SMC) *algorithms* has been a long-standing topic in the community, notably through adaptation of the particle sample size or the way the particles are sampled and weighted. In this paper we focus on the latter issue and develop methods for adjusting adaptively the proposal distribution of the particle filter.

Adaptation of the number of particles has been treated by several authors. In Legland and Oudjane (2006) (and later Hu et al. (2008, Section IV)) the size of the particle sample is increased until the total weight mass reaches a positive threshold, avoiding a situation where all particles are located in regions of the state space having zero posterior probability. Fearnhead and Liu (2007, Section 3.2) adjust the size of the particle cloud in order to control the error introduced by the resampling step. Another approach,

suggested by Fox (2003) and refined in Soto (2005) and Straka and Simandl (2006), consists in increasing the sample size until the *Kullback-Leibler divergence* (KLD) between the true and estimated target distributions is below a given threshold.

Unarguably, setting an appropriate sample size is a key ingredient of any statistical estimation procedure, and there are cases where the methods mentioned above may be used for designing satisfactorily this size; however increasing the sample size only is far from being always sufficient for achieving efficient variance reduction. Indeed, as in any algorithm based on importance sampling, a significant discrepancy between the proposal and target distributions may require an unreasonably large number of samples for decreasing the variance of the estimate under a specified value. For a very simple illustration, consider importance sampling estimation of the mean m of a normal distribution using as importance distribution another normal distribution having zero mean and same variance: in this case, the variance of the estimate grows like $\exp(m^2)/N$, N denoting the number of draws, implying that the sample size required for ensuring a given variance grows exponentially fast with m.

This points to the need for adapting the importance distribution of the particle filter, e.g., by adjusting at each iteration the particle weights and the proposal distributions; see e.g. Doucet et al. (2000), Liu (2001), and Fearnhead (2008) for reviews of various filtering methods. These two quantities are critically important, since the performance of the particle filter is closely related to the ability of proposing particles in state space regions where the posterior is significant. It is well known that sampling using as proposal distribution the mixture composed by the current particle importance weights and the prior kernel (yielding the classical bootstrap particle filter of Gordon et al. (1993)) is usually inefficient when the likelihood is highly peaked or located in the tail of the prior.

In the sequential context, the successive distributions to be approximated (e.g. the successive filtering distributions) are the iterates of a nonlinear random mapping, defined on the space of probability measures; this nonlinear mapping may in general be decomposed into two steps: a prediction step which is linear and a nonlinear correction step which amounts to compute a normalization factor. In this setting, an appealing way to update the current particle approximation consists in sampling new particles from the distribution obtained by propagating the current particle approximation through this mapping; see e.g. Hürzeler and Künsch (1998), Doucet et al. (2000), and Künsch (2005) (and the references therein). This sampling distribution guarantees that the conditional variance of the importance weights is equal to zero. As we shall see below, this proposal distribution enjoys other optimality conditions, and is in the sequel referred to as the *optimal sampling distribution*. However, sampling from the optimal sampling distribution is, except for some specific models, a difficult and time-consuming task (the in general costly auxiliary accept-reject developed and analysed by Künsch (2005) being most often the only available option).

To circumvent this difficulty, several sub-optimal schemes have been proposed. A first type of approaches tries to mimic the behavior of the optimal sampling without suffering the sometimes prohibitive cost of rejection sampling. This typically involves localisation of the modes of the unnormalized optimal sampling distribution by means of some optimisation algorithm, and the fitting of over-dispersed student's t-distributions around these modes; see for example Shephard and Pitt (1997), Doucet et al. (2001), and Liu (2001) (and the references therein). Except in specific cases, locating the modes involves solving an optimization problem for every particle, which is quite time-consuming.

A second class of approaches consists in using some classical approximate non-

linear filtering tools such as the *extended Kalman filter* (EKF) or the *unscented transform Kalman filter* (UT/UKF); see for example Doucet et al. (2001) and the references therein. These techniques assume implicitly that the conditional distribution of the next state given the current state and the observation has a single mode. In the EKF version of the particle filter, the linearisation of the state and observation equations is carried out for each individual particle. Instead of linearising the state and observation dynamics using Jacobian matrices, the UT/UKF particle filter uses a deterministic sampling strategy to capture the mean and covariance with a small set of carefully selected points (*sigma points*), which is also computed for each particle. Since these computations are most often rather involved, a significant computational overhead is introduced.

A third class of techniques is the so-called *auxiliary particle filter* (APF) suggested by Pitt and Shephard (1999), who proposed it as a way to build data-driven proposal distributions (with the initial aim of robustifying standard SMC methods to the presence of outlying observations); see e.g. Fearnhead (2008). The procedure comprises two stages: in the first-stage, the current particle weights are modified in order to select preferentially those particles being most likely proposed in regions where the posterior is significant. Usually this amounts to multiply the weights with so-called *adjustment multiplier weights*, which may depend on the next observation as well as the current position of the particle and (possibly) the proposal transition kernels. Most often, this adjustment weight is chosen to estimate the predictive likelihood of the next observation given the current particle position, but this choice is not necessarily optimal.

In a second stage, a new particle sample from the target distribution is formed using this proposal distribution and associating the proposed particles with weights proportional to the inverse of the adjustment multiplier weight [1]. APF procedures are known to be rather successful when the first-stage distribution is appropriately chosen, which is not always straightforward. The additional computational cost depends mainly on the way the first-stage proposal is designed. The APF method can be mixed with EKF and UKF leading to powerful but computationally involved particle filter; see, e.g., Andrieu et al. (2003).

None of the suboptimal methods mentioned above minimize any sensible risk-theoretic criterion and, more annoyingly, both theoretical and practical evidences show that choices which seem to be intuitively correct may lead to performances even worse than that of the plain bootstrap filter (see for example Douc et al. (2008) for a striking example). The situation is even more unsatisfactory when the particle filter is driven by a state space dynamic different from that generating the observations, as happens frequently when, e.g., the parameters are not known and need to be estimated or when the model is misspecified.

Instead of trying to guess what a good proposal distribution should be, it seems sensible to follow a more risk-theoretically founded approach. The first step in such a construction consists in choosing a sensible risk criterion, which is not a straightforward task in the SMC context. A natural criterion for SMC would be the variance of the estimate of the posterior mean of a target function (or a set of target functions) of interest, but this approach does not lead to a practical implementation for two reasons. Firstly, in SMC methods, though closed-form expression for the variance at any given current timestep of the posterior mean of any function is available, this variance

1. The original APF proposed by Pitt and Shephard (1999) features a second resampling procedure in order to end-up with an equally weighted particle system. This resampling procedure might however severely reduce the accuracy of the filter: Carpenter et al. (1999) give an example where the accuracy is reduced by a factor of 2; see also Douc et al. (2008) for a theoretical proof.

depends explicitly on all the time steps before the current time. Hence, choosing to minimize the variance at a given timestep would require to optimize all the simulations up to that particular time step, which is of course not practical. Because of the recursive form of the variance, the minimization of the conditional variance at each iteration of the algorithm does not necessarily lead to satisfactory performance on the long-run. Secondly, as for the standard importance sampling algorithm, this criterion is not *function-free*, meaning that a choice of a proposal can be appropriate for a given function, but inappropriate for another.

We will focus in the sequel on function-free risk criteria. A first criterion, advocated in Kong et al. (1994) and Liu (2001) is the *chi-square distance* (CSD) between the proposal and the target distributions, which coincides with the *coefficient of variation* (CV^2) of the importance weights. In addition, as heuristically discussed in Kong et al. (1994), the CSD is related to the *effective sample size*, which estimates the number of i.i.d. samples equivalent to the weighted particle system [2]. In practice, the CSD criterion can be estimated, with a complexity that grows linearly with the number of particles, using the empirical CV^2 which can be shown to converge to the CSD as the number of particles tends to infinity. In this paper we show that a similar property still holds in the SMC context, in the sense that the CV^2 still measures a CSD between two distributions μ^* and π^*, which are associated with the proposal and target distributions of the particle filter (see Theorem 3.4.1(ii)). Though this result does not come as a surprise, it provides an additional theoretical footing to an approach which is currently used in practice for triggering resampling steps.

Another function-free risk criterion to assess the performance of importance sampling estimators is the KLD between the proposal and the target distributions; see (Cappé et al., 2005, Chapter 7). The KLD shares some of the attractive properties of the CSD; in particular, the KLD may be estimated using the negated empirical *entropy* $\mathcal{E}$ of the importance weights, whose computational complexity is again linear in the number of particles. In the SMC context, it is shown in Theorem 3.4.1(i) that $\mathcal{E}$ still converges to the KLD between the same two distributions μ^* and π^* associated with the proposal and the target distributions of the particle filter.

Our methodology to design appropriate proposal distributions is based upon the minimization of the CSD and KLD between the proposal and the target distributions. Whereas these quantities (and especially the CSD) have been routinely used to detect sample impoverishment and trigger the resampling step (Kong et al., 1994), they have not been used for adapting the simulation parameters in SMC methods.

We focus here on the auxiliary sampling formulation of the particle filter. In this setting, there are two quantities to optimize: the adjustment multiplier weights (also called *first-stage weights*) and the parameters of the proposal kernel; together these quantites define the mixture used as instrumental distribution in the filter. We first establish a closed-form expression for the limiting value of the CSD and KLD of the auxiliary formulation of the proposal and the target distributions. Using these expressions, we identify a type of auxiliary SMC adjustment multiplier weights which minimize the CSD and the KLD for a given proposal kernel (Proposition 3.4.2). We then propose several optimization techniques for adapting the proposal kernels, always driven by the objective of minimizing the CSD or the KLD, in coherence with what is done to detect sample impoverishment (see Section 3.5). Finally, in the implementation section (Section 3.6), we use the proposed algorithms for approximating the filtering distributions in several state space models, and show that the proposed optimization

2. In some situations, the estimated ESS value can be misleading: see the comments of Stephens and Donnelly (2000) for a further discussion of this.

procedure improves the accuracy of the particle estimates and makes them more robust to outlying observations.

3.2 Informal presentation of the results

3.2.1 Adaptive importance sampling

Before stating and proving rigorously the main results, we discuss informally our findings and introduce the proposed methodology for developing adaptive SMC algorithms. Before entering into the sophistication of sequential methods, we first briefly introduce adaptation of the standard (non-sequential) importance sampling algorithm.

Importance sampling (IS) is a general technique to compute expectations of functions w.r.t. a target distribution with density $p(x)$ while only having samples generated from a different distribution—referred to as the *proposal distribution*—with density $q(x)$ (implicitly, the dominating measure is taken to be the Lebesgue measure on $\mathsf{X} := \mathbb{R}^d$). We sample $\{\xi_i\}_{i=1}^N$ from the proposal distribution q and compute the unnormalized importance weights $\omega_i := W(\xi_i)$, $i = 1, \ldots, N$, where $W(x) := p(x)/q(x)$. For any function f, the self-normalized importance sampling estimator may be expressed as $\mathrm{IS}_N(f) := \Omega^{-1} \sum_{i=1}^N \omega_i f(\xi_i)$, where $\Omega := \sum_{j=1}^N \omega_j$. As usual in applications of the IS methodology to Bayesian inference, the target density p is known only up to a normalization constant; hence we will focus only on a self-normalized version of IS that solely requires the availability of an unnormalized version of p (see Geweke, 1989). Throughout the paper, we call a set $\{\xi_i\}_{i=1}^N$ of random variables, referred to as *particles* and taking values in X, and nonnegative weights $\{\omega_i\}_{i=1}^N$ a *weighted sample* on X. Here N is a (possibly random) integer, though we will take it fixed in the sequel. It is well known (see again Geweke, 1989) that, provided that f is integrable w.r.t. p, i.e. $\int |f(x)|p(x)\,\mathrm{d}x < \infty$, $\mathrm{IS}_N(f)$ converges, as the number of samples tends to infinity, to the target value

$$\mathbb{E}_p[f(X)] := \int f(x)p(x)\,\mathrm{d}x\ ,$$

for any function $f \in \mathsf{C}$, where C is the set of functions which are integrable w.r.t. to the target distribution p. Under some additional technical conditions, th is estimator is also asymptotically normal at rate $\sqrt{N}$; see Geweke (1989).

It is well known that IS estimators are sensitive to the choice of the proposal distribution. A classical approach consists in trying to minimize the asymptotic variance w.r.t. the proposal distribution q. This optimization is in closed form and leads (when f is a non-negative function) to the optimal choice $q^*(x) = f(x)p(x)/\int f(x)p(x)\,\mathrm{d}x$, which is, since the normalization constant is precisely the quantity of interest, rather impractical. Sampling from this distribution can be done by using an accept-reject algorithm, but this does not solve the problem of choosing an appropriate proposal distribution. Note that it is possible to approach this optimal sampling distribution by using the *cross-entropy method*; see Rubinstein and Kroese (2004) and de Boer et al. (2005) and the references therein. We will discuss this point later on.

For reasons that will become clear in the sequel, this type of objective is impractical in the sequential context, since the expression of the asymptotic variance in this case is recursive and the optimization of the variance at a given step is impossible. In addition, in most applications, the proposal density is expected to perform well for a range of typical functions of interest rather than for a specific target function f. We are thus looking for *function-free* criteria. The most often used criterion is the CSD

between the proposal distribution q and the target distribution p, defined as

$$d_{\chi^2}(p||q) = \int \frac{\{p(x) - q(x)\}^2}{q(x)} \, \mathrm{d}x \, , \tag{3.2.1}$$

$$= \int W^2(x) q(x) \, \mathrm{d}x - 1 \, , \tag{3.2.2}$$

$$= \int W(x) p(x) \, \mathrm{d}x - 1 \, . \tag{3.2.3}$$

The CSD between p and q may be expressed as the variance of the importance weight function W under the proposal distribution, i.e.

$$d_{\chi^2}(p||q) = \mathrm{Var}_q[W(X)] \, .$$

This quantity can be estimated by computing the squared coefficient of variation of the unnormalized weights (Evans and Swartz, 1995, Section 4):

$$\mathrm{CV}^2\big(\{\omega_i\}_{i=1}^N \big) := N\Omega^{-2} \sum_{i=1}^N \omega_i^2 - 1 \, . \tag{3.2.4}$$

The CV^2 was suggested by Kong et al. (1994) as a means for detecting weight degeneracy. If all the weights are equal, then CV^2 is equal to zero. On the other hand, if all the weights but one are zero, then the coefficient of variation is equal to $N - 1$ which is its maximum value. From this it follows that using the estimated coefficient of variation for assessing accuracy is equivalent to examining the normalized importance weights to determine if any are relatively large[3]. Kong et al. (1994) showed that the coefficient of variation of the weights $\mathrm{CV}^2\left(\{\omega_i\}_{i=1}^N\right)$ is related to the *effective sample size* (ESS), which is used for measuring the overall efficiency of an IS algorithm:

$$N^{-1}\mathrm{ESS}\big(\{\omega_i\}_{i=1}^N \big) := \frac{1}{1 + \mathrm{CV}^2\big(\{\omega_i\}_{i=1}^N \big)} \to \big\{1 + d_{\chi^2}(p||q)\big\}^{-1} \, .$$

Heuristically, the ESS measures the number of i.i.d. samples (from p) equivalent to the N weighted samples. The smaller the CSD between the proposal and target distributions is, the larger is the ESS. This is why the CSD is of particular interest when measuring efficiency of IS algorithms.

Another possible measure of fit of the proposal distribution is the KLD (also called *relative entropy*) between the proposal and target distributions, defined as

$$d_{\mathrm{KL}}(p||q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) \, \mathrm{d}x \, , \tag{3.2.5}$$

$$= \int p(x) \log W(x) \, \mathrm{d}x \, , \tag{3.2.6}$$

$$= \int W(x) \log W(x) \, q(x) \, \mathrm{d}x \, . \tag{3.2.7}$$

This criterion can be estimated from the importance weights using the negative *Shannon entropy* $\mathcal{E}$ of the importance weights:

$$\mathcal{E}\big(\{\omega_i\}_{i=1}^N \big) := \Omega^{-1} \sum_{i=1}^N \omega_i \log\left(N\Omega^{-1}\omega_i\right) \, . \tag{3.2.8}$$

3. Some care should be taken for small sample sizes N; the CV^2 can be low because q sample only over a subregion where the integrand is nearly constant, which is not always easy to detect.

The Shannon entropy is maximal when all the weights are equal and minimal when all weights are zero but one. In IS (and especially for the estimation of rare events), the KLD between the proposal and target distributions was thoroughly investigated by Rubinstein and Kroese (2004), and is central in the *cross-entropy* (CE) methodology.

Classically, the proposal is chosen from a family of densities q_θ parameterized by θ. Here θ should be thought of as an element of Θ, which is a subset of $\mathbb{R}^k$. The most classical example is the family of student's t-distributions parameterized by mean and covariance. More sophisticated parameterizations, like mixture of multi-dimensional Gaussian or Student's t-distributions, have been proposed; see, e.g., Oh and Berger (1992), Oh and Berger (1993), Evans and Swartz (1995), Givens and Raftery (1996), Liu (2001, Chapter 2, Section 2.6), and, more recently, Cappé et al. (2008) in this issue. In the sequential context, where computational efficiency is a must, we typically use rather simple parameterizations, so that the two criteria above can be (approximatively) solved in a few iterations of a numerical minimization procedure.

The optimal parameters for the CSD and the KLD are those minimizing $\theta \mapsto d_{\chi^2}(p||q_\theta)$ and $\theta \mapsto d_{\mathrm{KL}}(p||q_\theta)$, respectively. In the sequel, we denote by θ^*_{CSD} and θ^*_{KLD} these optimal values. Of course, these quantities cannot be computed in closed form (recall that even the normalization constant of p is most often unknown; even if it is known, the evaluation of these quantities would involve the evaluation of most often high-dimensional integrals). Nevertheless, it is possible to construct consistent estimators of these optimal parameters. There are two classes of methods, detailed below.

The first uses the fact that the the CSD $d_{\chi^2}(p||q_\theta)$ and the KLD $d_{\mathrm{KL}}(p|q_\theta)$ may be approximated by (3.2.4) and (3.2.8), substituting in these expressions the importance weights by $\omega_i = W_\theta(\xi_i^{(\theta)})$, $i = 1, \ldots, N$, where $W_\theta := p/q_\theta$ and $\{\xi_i^{(\theta)}\}_{i=1}^N$ is a sample from q_θ. This optimization problem formally shares some similarities with the classical minimum chi-square or maximum likelihood estimation, but with the following important difference: the integrations in (3.2.1) and (3.2.5) are w.r.t. the proposal distribution q_θ and not the target distribution p. As a consequence, the particles $\{\xi_i^{(\theta)}\}_{i=1}^N$ in the definition of the coefficient of variation (3.2.4) or the entropy (3.2.8) of the weights constitute a sample from q_θ and not from the target distribution p. As the estimation progresses, the samples used to approach the limiting CSD or KLD can, in contrast to standard estimation procedures, be updated (these samples could be kept fixed, but this is of course inefficient).

The computational complexity of these optimization problems depends on the way the proposal is parameterized and how the optimization procedure is implemented. Though the details of the optimization procedure is in general strongly model dependent, some common principles for solving this optimization problem can be outlined. Typically, the optimization is done recursively, i.e. the algorithm defines a sequence θ_ℓ, $\ell = 0, 1, \ldots$, of parameters, where ℓ is the iteration number. At each iteration, the value of θ_ℓ is updated by computing a direction $p_{\ell+1}$ in which to step, a step length $\gamma_{\ell+1}$, and setting

$$\theta_{\ell+1} = \theta_\ell + \gamma_{\ell+1} p_{\ell+1} \, .$$

The search direction is typically computed using either Monte Carlo approximation of the finite-difference or (when the quantities of interest are sufficiently regular) the gradient of the criterion. These quantities are used later in conjunction with classical optimization strategies for computing the step size $\gamma_{\ell+1}$ or normalizing the search direction. These implementation issues, detailed in Section 3.6, are model dependent. We denote by M_ℓ the number of particles used to obtain such an approximation at iteration ℓ. The number of particles may vary with the iteration index; heuristically there is no need for using a large number of simulations during the initial stage of the opti-

mization. Even rather crude estimation of the search direction might suffice to drive the parameters towards the region of interest. However, as the iterations go on, the number of simulations should be increased to avoid "zi g-zagging" when the algorithm approaches convergence. After L iterations, the total number of generated particles is equal to $N = \sum_{\ell=1}^{L} M_\ell$. Another solution, which is not considered in this paper, would be to use a stochastic approximation procedure, which consists in fixing $M_\ell = M$ and letting the stepsize γ_ℓ tend to zero. This appealing solution has been successfully used in Arouna (2004).

The computation of the finite difference or the gradient, being defined as expectations of functions depending on θ, can be performed using two different approaches. Starting from definitions (3.2.3) and (3.2.6), and assuming appropriate regularity conditions, the gradient of $\theta \mapsto d_{\chi^2}(p||q_\theta)$ and $\theta \mapsto d_{\mathrm{KL}}(p||q_\theta)$ may be expressed as

$$G_{\mathrm{CSD}}(\theta) := \nabla_\theta d_{\chi^2}(p||q_\theta) = \int p(x)\nabla_\theta W_\theta(x)\,\mathrm{d}x = \int q_\theta(x)W_\theta(x)\nabla_\theta W_\theta(x)\,\mathrm{d}x\,, \tag{3.2.9}$$

$$G_{\mathrm{KLD}}(\theta) := \nabla_\theta d_{\mathrm{KL}}(p||q_\theta) = \int p(x)\nabla_\theta \log[W_\theta(x)]\,\mathrm{d}x = \int q_\theta(x)\nabla_\theta W_\theta(x)\,\mathrm{d}x\,. \tag{3.2.10}$$

These expressions lead immediately to the following approximations,

$$\hat{G}_{\mathrm{CSD}}(\theta) = M^{-1}\sum_{i=1}^{M} W_\theta(\xi_i^{(\theta)})\nabla_\theta W_\theta(\xi_i^{(\theta)})\,, \tag{3.2.11}$$

$$\hat{G}_{\mathrm{KLD}}(\theta) = M^{-1}\sum_{i=1}^{M} \nabla_\theta W_{\theta_\ell}(\xi_i^{(\theta_\ell)})\,. \tag{3.2.12}$$

There is another way to compute derivatives, which shares some similarities with *pathwise derivative estimates*. Recall that for any $\theta \in \Theta$, one may choose F_θ so that the random variable $F_\theta(\epsilon)$, where ϵ is a vector of independent uniform random variables on $[0,1]^d$, is distributed according to q_θ. Therefore, we may express $\theta \mapsto d_{\chi^2}(p||q_\theta)$ and $\theta \mapsto d_{\mathrm{KL}}(p||q_\theta)$ as the following integrals,

$$d_{\chi^2}(p||q_\theta) = \int_{[0,1]^d} w_\theta(x)\,\mathrm{d}x\,,$$
$$d_{\mathrm{KL}}(p||q_\theta) = \int_{[0,1]^d} w_\theta(x)\log\left[w_\theta(x)\right]\mathrm{d}x\,,$$

where $w_\theta(x) := W_\theta \circ F_\theta(x)$. Assuming appropriate regularity conditions (i.e. that $\theta \mapsto W_\theta \circ F_\theta(x)$ is differentiable and that we can interchange the integration and the differentiation), the differential of these quantities w.r.t. θ may be expressed as

$$G_{\mathrm{CSD}}(\theta) = \int_{[0,1]^d} \nabla_\theta w_\theta(x)\,\mathrm{d}x\,,$$
$$G_{\mathrm{KLD}}(\theta) = \int_{[0,1]^d} \{\nabla_\theta w_\theta(x)\log[w_\theta(x)] + \nabla_\theta w_\theta(x)\}\,\mathrm{d}x\,.$$

For any given x, the quantity $\nabla_\theta w_\theta(x)$ is the pathwise derivative of the function $\theta \mapsto w_\theta(x)$. As a practical matter, we usually think of each x as a realization of of the output of an ideal random generator. Each $w_\theta(x)$ is then the output of the simulation algorithm at parameter θ for the random number x. Each $\nabla_\theta w_\theta(x)$ is the derivative of the simulation output w.r.t. θ with the random numbers held fixed. These two expressions,

which of course coincide with (3.2.9) and (3.2.10), lead to the following estimators,

$$\tilde{G}_{\mathrm{CSD}}(\theta) = M^{-1} \sum_{i=1}^{M} \nabla_\theta w_\theta(\epsilon_i) \ ,$$
$$\tilde{G}_{\mathrm{KLD}}(\theta) = M^{-1} \sum_{i=1}^{M} \{\nabla_\theta w_\theta(\epsilon_i) \log[w_\theta(\epsilon_i)] + \nabla_\theta w_\theta(\epsilon_i)\} \ ,$$

where each element of the sequence $\{\epsilon_i\}_{i=1}^M$ is a vector on $[0,1]^d$ of independent uniform random variables. It is worthwhile to note that if the number $M_\ell = M$ is kept fixed during the iterations and the uniforms $\{\epsilon_i\}_{i=1}^M$ are drawn once and for all (i.e. the same uniforms are used at the different iterations), then the iterative algorithm outlined above solves the following problem:

$$\theta \mapsto \mathrm{CV}^2 \left(\{w_\theta(\epsilon_i)\}_{i=1}^M \right) \ , \tag{3.2.13}$$
$$\theta \mapsto \mathcal{E} \left(\{w_\theta(\epsilon_i)\}_{i=1}^N \right) \ . \tag{3.2.14}$$

From a theoretical standpoint, this optimization problem is very similar to M-estimation, and convergence results for M-estimators can thus be used under rather standard technical assumptions; see for example Van der Vaart (1998). This is the main advantage of fixing the sample $\{\epsilon_i\}_{i=1}^M$. We use this implementation in the simulations.

Under appropriate conditions, the sequence of estimators $\theta^*_{\ell,\mathrm{CSD}}$ or $\theta^*_{\ell,\mathrm{KLD}}$ of these criteria converge, as the number of iterations tends to infinity, to θ^*_{CSD} or θ^*_{KLD} which minimize the criteria $\theta \mapsto d_{\chi^2}(p||q_\theta)$ and $\theta \mapsto d_{\mathrm{KL}}(p||q_\theta)$, respectively; these theoretical issues are considered in a companion paper.

The second class of approaches considered in this paper is used for minimizing the KLD (3.2.14) and is inspired by the cross-entropy method. This algorithm approximates the minimum θ^*_{KLD} of (3.2.14) by a sequence of pairs of steps, where each step of each pair addresses a simpler optimization problem. Compared to the previous method, this algorithm is derivative-free and does not require to select a step size. It is in general simpler to implement and avoid most of the common pitfalls of stochastic approximation. Denote by $\theta_0 \in \Theta$ an initial value. We define recursively the sequence $\{\theta_\ell\}_{\ell \geq 0}$ as follows. In a first step, we draw a sample $\{\xi_i^{(\theta_\ell)}\}_{i=1}^{M_\ell}$ and evaluate the function

$$\theta \mapsto Q_\ell(\theta, \theta_\ell) := \sum_{i=1}^{M_\ell} W_{\theta_\ell}(\xi_i^{(\theta_\ell)}) \log q_\theta(\xi_i^{(\theta_\ell)}) \ . \tag{3.2.15}$$

In a second step, we choose $\theta_{\ell+1}$ to be the (or any, if there are several) value of $\theta \in \Theta$ that maximizes $Q_\ell(\theta, \theta_\ell)$. As above, the number of particles M_ℓ is increased during the successive iterations. This procedure ressembles closely the Monte Carlo EM (Wei and Tanner, 1991) for maximum likelihood in incomplete data models. The advantage of this approach is that the solution of the maximization problem $\theta_{\ell+1} = \mathrm{argmax}_{\theta \in \Theta} \in Q_\ell(\theta, \theta_\ell)$ is often on closed form. In particular, this happens if the distribution q_θ belongs to an *exponential family* (EF) or is a mixture of distributions of NEF; see Cappé et al. (2008) for a discussion. The convergence of this algorithm can be established along the same lines as the convergence of the MCEM algorithm; see Fort and Moulines (2003). As the number of iterations ℓ increases, the sequence of estimators θ_ℓ may be shown to converge to θ^*_{KLD}. These theoretical results are established in a companion paper.

3.2.2 Sequential Monte Carlo methods

In the sequential context, where the problem consists in simulating from a *sequence* $\{p_k\}$ of probability density function, the situation is more difficult. Let X_k be denote the state space of distribution p_k and note that this space may vary with k, e.g. in terms of increasing dimensionality. In many applications, these densities are related to each other by a (possibly random) mapping, i.e. $p_k = \Psi_{k-1}(p_{k-1})$. In the sequel we focus on the case where there exists a non-negative function $l_{k-1} : (\xi, \tilde{\xi}) \mapsto l_{k-1}(\xi, \tilde{\xi})$ such that

$$p_k(\tilde{\xi}) = \frac{\int l_{k-1}(\xi, \tilde{\xi}) p_{k-1}(\xi)\, \mathrm{d}\xi}{\int p_{k-1}(\xi) \int l_{k-1}(\xi, \tilde{\xi})\, \mathrm{d}\tilde{\xi}\, \mathrm{d}\xi} \,. \tag{3.2.16}$$

As an example, consider the following generic nonlinear dynamic system described in state space form:

– *State (system) model*

$$X_k = a(X_{k-1}, U_k) \leftrightarrow \overbrace{q(X_{k-1}, X_k)}^{\text{Transition Density}} \,, \tag{3.2.17}$$

– *Observation (measurement) model*

$$Y_k = b(X_k, V_k) \leftrightarrow \overbrace{g(X_k, Y_k)}^{\text{Observation Density}} \,. \tag{3.2.18}$$

By these equations we mean that each hidden state X_k and data Y_k are assumed to be generated by nonlinear functions $a(\cdot)$ and $b(\cdot)$, respectively, of the state and observation noises U_k and V_k. The state and the observation noises $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are assumed to be mutually independent sequences of i.i.d. random variables. The precise form of the functions and the assumed probability distributions of the state and observation noises U_k and V_k imply, via a change of variables, the transition probability density function $q(x_{k-1}, x_k)$ and the observation probability density function $g(x_k, y_k)$, the latter being referred to as the *likelihood of the observation*. With these definitions, the process $\{X_k\}_{k\geq 0}$ is Markovian, i.e. the conditional probability density of X_k given the past states $X_{0:k-1} := (X_0, \ldots, X_{k-1})$ depends exclusively on X_{k-1}. This distribution is described by the density $q(x_{k-1}, x_k)$. In addition, the conditional probability density of Y_k given the states $X_{0:k}$ and the past observations $Y_{0:k-1}$ depends exclusively on X_k, and this distribution is captured by the likelihood $g(x_k, y_k)$. We assume further that the initial state X_0 is distributed according to a density function $\pi_0(x_0)$. Such nonlinear dynamic systems arise frequently in many areas of science and engineering such as target tracking, computer vision, terrain referenced navigation, finance, pollution monitoring, communications, audio engineering, to list only a few.

Statistical inference for the general nonlinear dynamic system above involves computing the *posterior distribution* of a collection of state variables $X_{s:s'} := (X_s, \ldots, X_{s'})$ conditioned on a batch $Y_{0:k} = (Y_0, \ldots, Y_k)$ of observations. We denote this posterior distribution by $\phi_{s:s'|k}(X_{s:s'}|Y_{0:k})$. Specific problems include *filtering*, corresponding to $s = s' = k$, *fixed lag smoothing*, where $s = s' = k - L$, and *fixed interval smoothing*, with $s = 0$ and $s' = k$. Despite the apparent simplicity of the above problem, the posterior distributions can be computed in closed form only in very specific cases, principally, the linear Gaussian model (where the functions $a(\cdot)$ and $b(\cdot)$ are linear and the state and observation noises $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are Gaussian) and the discrete *hidden Markov model* (where X_k takes its values in a finite alphabet). In the vast majority of cases, nonlinearity or non-Gaussianity render analytic solutions intractable—see Anderson and Moore (1979); Kailath et al. (2000); Ristic et al. (2004); Cappé et al. (2005).

Starting with the initial, or prior, density function $\pi_0(x_0)$, and observations $Y_{0:k} = y_{0:k}$, the posterior density $\phi_{k|k}(x_k|y_{0:k})$ can be obtained using the following *prediction-correction* recursion (Ho and Lee, 1964):

– *Prediction*

$$\phi_{k|k-1}(x_k|y_{0:k-1}) = \phi_{k-1|k-1}(x_{k-1}|y_{0:k-1})q(x_{k-1}, x_k) \ , \tag{3.2.19}$$

– *Correction*

$$\phi_{k|k}(x_k|y_{0:k}) = \frac{g(x_k, y_k)\phi_{k|k-1}(x_k|y_{0:k-1})}{\mathcal{L}_{k|k-1}(y_k|y_{0:k-1})} \ , \tag{3.2.20}$$

where $\mathcal{L}_{k|k-1}$ is the predictive distribution of Y_k given the past observations $Y_{0:k-1}$. For a fixed data realisation, this term is a normalizing constant (independent of the state) and is thus not necessary to compute in standard implementations of SMC methods.

By setting $p_k = \phi_{k|k}$, $p_{k-1} = \phi_{k-1|k-1}$, and

$$l_{k-1}(x, x') = g(x_k, y_k)q(x_{k-1}, x_k) \ ,$$

we conclude that the sequence $\{\phi_{k|k}\}_{k\geq 1}$ of filtering densities can be generated according to (3.2.16).

The case of fixed interval smoothing works entirely analogously: indeed, since

$$\phi_{0:k|k-1}(x_{0:k}|y_{0:k-1}) = \phi_{0:k-1|k-1}(x_{0:k-1}|y_{0:k-1})q(x_{k-1}, x_k)$$

and

$$\phi_{0:k|k}(x_k|y_{0:k}) = \frac{g(x_k, y_k)\phi_{k|k-1}(x_{0:k}|y_{0:k-1})}{\mathcal{L}_{k|k-1}(y_k|y_{0:k-1})} \ ,$$

the flow $\{\phi_{0:k|k}\}_{k\geq 1}$ of smoothing distributions can be generated according to (3.2.16) by letting $p_k = \phi_{0:k|k}$, $p_{k-1} = \phi_{0:k-1|k-1}$, and replacing $l_{k-1}(x_{0:k-1}, x'_{0:k})\,\mathrm{d}x'_{0:k}$ by $g(x'_k, y_k)\,q(x_{k-1}, x'_k)\,\mathrm{d}x'_k$ $\delta_{x_{0:k-1}}(\mathrm{d}x'_{0:k-1})$, where δ_a denotes the Dirac mass located in a. Note that this replacement is done formally since the unnormalized kernel in question lacks a density in the smoothing mode; this is due to the fact that the Dirac measure is singular w.r.t. the Lebesgue measure. This is however handled by the measure theoretic approach in Section 3.4, implying that all theoretical results presented in the following will comprise also fixed interval smoothing.

We now adapt the procedures considered in the previous section to the sampling of densities generated according to (3.2.16). Here we focus on a single timestep, and drop from the notation the dependence on k which is irrelevant at this stage. Moreover, set $p_k = \mu$, $p_{k-1} = \nu$, $l_k = l$, and assume that we have at hand a weighted sample $\{(\xi_i, \omega_i)\}_{i=1}^N$ targeting ν, i.e., for any ν-integrable function f, $\Omega^{-1}\sum_{i=1}^N \omega_i f(\xi_i)$ approximates the corresponding integral $\int f(\xi)\nu(\xi)\,\mathrm{d}\xi$. A natural strategy for sampling from μ is to replace ν in (3.2.16) by its particle approximation, yielding

$$\mu_N(\tilde{\xi}) := \sum_{i=1}^N \frac{\omega_i \int l(\xi_i, \tilde{\xi})\,\mathrm{d}\tilde{\xi}}{\sum_{j=1}^N \omega_j \int l(\xi_j, \tilde{\xi})\,\mathrm{d}\tilde{\xi}} \left[\frac{l(\xi_i, \tilde{\xi})}{\int l(\xi_i, \tilde{\xi})\,\mathrm{d}\tilde{\xi}}\right]$$

as an approximation of μ, and simulate $\tilde{M}_N$ new particles from this distribution; however, in many applications direct simulation from μ_N is infeasible without the application of computationally expensive auxiliary accept-reject techniques introduced by Hürzeler and Künsch (1998) and thoroughly analysed by Künsch (2005). This difficulty can be overcome by simulating new particles $\{\tilde{\xi}_i\}_{i=1}^{\tilde{M}_N}$ from the instrumental mixture distribution with density

$$\pi_N(\tilde{\xi}) := \sum_{i=1}^N \frac{\omega_i \psi_i}{\sum_{j=1}^N \omega_j \psi_j} r(\xi_i, \tilde{\xi}) \ ,$$

where $\{\psi_i\}_{i=1}^N$ are the so-called *adjustment multiplier weights* and r is a transition density function, i.e., $r(\xi, \tilde{\xi})$ is a nonnegative function and, for any $\xi \in \mathsf{X}$, $\int r(\xi, \tilde{\xi}) \, \mathrm{d}\tilde{\xi} = 1$. If one can guess, based on the new observation, which particles are most likely to contribute significantly to the posterior, the resampling stage may be anticipated by increasing (or decreasing) the importance weights. This is the purpose of using the multiplier weights ψ_i. We associate these particles with importance weights $\{\mu_N(\tilde{\xi}_i)/\pi_N(\tilde{\xi}_i)\}_{i=1}^{\tilde{M}_N}$. In this setting, a new particle position is simulated from the transition proposal density $r(\xi_i, \cdot)$ with probability proportional to $\omega_i \psi_i$. Haplessly, the importance weight $\mu_N(\tilde{\xi}_i)/\pi_N(\tilde{\xi}_i)$ is expensive to evaluate since this involves summing over N terms.

We thus introduce, as suggested by Pitt and Shephard (1999), an *auxiliary variable* corresponding to the selected particle, and target instead the probability density

$$\mu_{\mathrm{aux}}(i, \tilde{\xi}) := \frac{\omega_i \int l(\xi_i, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}}{\sum_{j=1}^N \omega_j \int l(\xi_j, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}} \left[\frac{l(\xi_i, \tilde{\xi})}{\int l(\xi_i, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}} \right] \tag{3.2.21}$$

on the product space $\{1, \ldots, N\} \times \mathsf{X}$. Since μ_N is the marginal distribution of μ_{aux} with respect to the particle index i, we may sample from μ_N by simulating instead a set $\{(I_i, \tilde{\xi}_i)\}_{i=1}^{\tilde{M}_N}$ of indices and particle positions from the instrumental distribution

$$\pi_{\mathrm{aux}}(i, \tilde{\xi}) := \frac{\omega_i \psi_i}{\sum_{j=1}^N \omega_j \psi_j} r(\xi_i, \tilde{\xi}) \tag{3.2.22}$$

and assigning each draw $(I_i, \tilde{\xi}_i)$ the weight

$$\tilde{\omega}_i := \frac{\mu_{\mathrm{aux}}(I_i, \tilde{\xi}_i)}{\pi_{\mathrm{aux}}(I_i, \tilde{\xi}_i)} = \psi_{I_i}^{-1} \frac{l(\xi_{I_i}, \tilde{\xi}_i)}{r(\xi_{I_i}, \tilde{\xi}_i)} \, . \tag{3.2.23}$$

Hereafter, we discard the indices and let $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^{\tilde{M}_N}$ approximate the target density μ. Note that setting, for all $i \in \{1, \ldots, N\}$, $\psi_i \equiv 1$ yields the standard bootstrap particle filter presented by Gordon et al. (1993). In the sequel, we assume that each adjustment multiplier weight ψ_i is a function of the particle position $\psi_i = \Psi(\xi_i)$, $i \in \{1, \ldots, N\}$, and define

$$\Phi(\xi, \tilde{\xi}) := \Psi^{-1}(\xi) \frac{l(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})} \, , \tag{3.2.24}$$

so that $\mu_{\mathrm{aux}}(i, \tilde{\xi})/\pi_{\mathrm{aux}}(i, \tilde{\xi})$ is proportional to $\Phi(\xi_i, \tilde{\xi})$. We will refer to the function Ψ as the *adjustment multiplier function*.

3.2.3 Risk minimization for sequential adaptive importance sampling and resampling

We may expect that the efficiency of the algorithm described above depends highly on the choice of adjustment multiplier weights and proposal kernel.

In the context of state space models, Pitt and Shephard (1999) suggested to use an approximation, defined as the value of the likelihood evaluated at the mean of the prior transition, i.e. $\psi_i := g\left(\int x' q(\xi_i, x') \, \mathrm{d}x', y_k\right)$, where y_k is the current observation, of the predictive likelihood as adjustment multiplier weights. Although this choice of the weight outperforms the conventional bootstrap filter in many applications, as pointed out in Andrieu et al. (2003), this approximation of the predictive likelihood could be very poor and lead to performance even worse than that of the conventional approach

if the dynamic model $q(x_{k-1}, x_k)$ is quite scattered and the likelihood $g(x_k, y_k)$ varies significantly over the prior $q(x_{k-1}, x_k)$.

The optimization of the adjustment multiplier weight was also studied by Douc et al. (2008) (see also Olsson et al. (2007)) who identified adjustment multiplier weights for which the increase of asymptotic variance at a single iteration of the algorithm is minimal. Note however that this optimization is done using a *function-specific* criterion, whereas we advocate here the use of *function-free* criteria.

In our risk minimization setting, this means that both the adjustment weights and the proposal kernels need to be adapted. As we will see below, these two problems are in general intertwined; however, in the following it will be clear that the two criteria CSD and KLD behave differently at this point. Because the criteria are rather involved, it is interesting to study their behaviour as the number of particles N grows to infinity. This is done in Theorem 3.4.1, which shows that the CSD $d_{\chi^2}(\mu_{\text{aux}}||\pi_{\text{aux}})$ and KLD $d_{\text{KL}}(\mu_{\text{aux}}||\pi_{\text{aux}})$ converges to $d_{\chi^2}(\mu^*||\pi^*_\Psi)$ and $d_{\text{KL}}(\mu^*||\pi^*_\Psi)$, respectively, where

$$\begin{aligned} \mu^*(\xi, \tilde{\xi}) &:= \frac{\nu(\xi)\, l(\xi, \tilde{\xi})}{\iint \nu(\xi)\, l(\xi, \tilde{\xi})\, \mathrm{d}\xi\, \mathrm{d}\tilde{\xi}}\,, \\ \pi^*_\Psi(\xi, \tilde{\xi}) &:= \frac{\nu(\xi)\Psi(\xi)\, r(\xi, \tilde{\xi})}{\iint \nu(\xi)\Psi(\xi)\, r(\xi, \tilde{\xi})\, \mathrm{d}\xi\, \mathrm{d}\tilde{\xi}}\,. \end{aligned} \tag{3.2.25}$$

The expressions (3.2.25) of the limiting distributions then allow for deriving the adjustment multiplier weight function Ψ and the proposal density l minimizing the corresponding discrepancy measures. In absence of constraints (when Ψ and l can be chosen arbitrarily), the optimal solution for both the CSD and the KLD consists in setting $\Psi = \Psi^*$ and $r = r^*$, where

$$\Psi^*(\xi) := \int l(\xi, \tilde{\xi})\, \mathrm{d}\tilde{\xi} = \int \frac{l(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})} r(\xi, \tilde{\xi})\, \mathrm{d}\tilde{\xi}\,, \tag{3.2.26}$$

$$r^*(\xi, \tilde{\xi}) := l(\xi, \tilde{\xi})/\Psi^*(\xi)\,. \tag{3.2.27}$$

This choice coincides with the so-called *optimal sampling strategy* proposed by Hürzeler and Künsch (1998) and developed further by Künsch (2005), which turns out to be *optimal* (in absence of constraints) in our risk-minimization setting.

Remark 3.2.1. The limiting distributions μ^* and π^*_Ψ have nice interpretations within the framework of state space models (see the previous section). In this setting, the limiting distribution μ^* at time k is the joint distribution $\phi_{k:k+1|k+1}$ of the *filtered* couple $X_{k:k+1}$, that is, the distribution of $X_{k:k+1}$ conditionally on the observation record $Y_{0:k+1}$; this can be seen as the asymptotic target distribution of our particle model. Moreover, the limiting distribution π^* at time k is only slightly more intricate: Its first marginal corresponds to the filtering distribution at time k reweighted by the adjustment function Ψ, which is typically used for incorporating information from the new observation Y_{k+1}. The second marginal of π^* is then obtained by propagating this weighted filtering distribution through the Markovian dynamics of the proposal kernel R; thus, π^*_Ψ describes completely the asymptotic instrumental distribution of the APF, and the two quantities $d_{\text{KL}}(\mu^*||\pi^*)$ and $d_{\chi^2}(\mu^*||\pi^*)$ reflect the asymptotic discrepancy between the true model and the particle model at the given time step.

In presence of constraints on the choice of Ψ and r, the optimization of the adjustment weight function and the proposal kernel density is intertwined. By the so-called *chain rule for entropy* (see Cover and Thomas, 1991, Theorem 2.2.1), we have

$$d_{\text{KL}}(\mu^*||\pi^*_\Psi) = \int \frac{\nu(\xi)}{\nu(\Psi^*)} \Psi^*(\xi) \log\left(\frac{\Psi^*(\xi)/\nu(\Psi^*)}{\Psi(\xi)/\nu(\Psi)}\right) \mathrm{d}\xi + \iint \frac{\nu(\xi)}{\nu(\Psi^*)} l(\xi, \tilde{\xi}) \log\left(\frac{r^*(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})}\right) \mathrm{d}\xi\, \mathrm{d}\tilde{\xi}$$

where $\nu(f) := \int \nu(\xi) f(\xi) \, \mathrm{d}\xi$. Hence, if the optimal adjustment function can be chosen freely, then, whatever the choice of the proposal kernel is, the best choice is still $\Psi^*_{\mathrm{KL},r} = \Psi^*$: the best that we can do is to choose $\Psi^*_{\mathrm{KL},r}$ such that the two marginal distributions $\xi \mapsto \int \mu^*(\xi, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}$ and $\xi \mapsto \int \pi^*(\xi, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}$ are identical. If the choices of the weight adjustment function and the proposal kernels are constrained (if, e.g., the weight should be chosen in a pre-specified family of functions or the proposal kernel belongs to a parametric family), nevertheless, the optimization of Ψ and r decouple asymptotically. The optimization for the CSD does not lead to such a nice decoupling of the adjustment function and the proposal transition; nevertheless, an explicit expression for the adjustment multiplier weights can still be found in this case:

$$\Psi^*_{\chi^2,r}(\xi) := \sqrt{\int \frac{l^2(\xi, \tilde{\xi})}{r(\xi, \tilde{\xi})} \, \mathrm{d}\tilde{\xi}} = \sqrt{\int \frac{l^2(\xi, \tilde{\xi})}{r^2(\xi, \tilde{\xi})} r(\xi, \tilde{\xi}) \, \mathrm{d}\tilde{\xi}} \, . \tag{3.2.28}$$

Compared to (3.2.26), the optimal adjustment function for the CSD is the L^2 (rather than the L^1) norm of $\xi \mapsto l^2(\xi, \tilde{\xi}) / r^2(\xi, \tilde{\xi})$. Since $l(\xi, \tilde{\xi}) = \Psi^*(\xi) r^*(\xi, \tilde{\xi})$ (see definitions (3.2.26) and (3.2.27)), we obtain, not surprisingly, if we set $r = r^*$, $\Psi^*_{\chi^2,r}(\xi) = \Psi^*(\xi)$.

Using this risk minimization formulation, it is possible to select the adjustment weight function as well as the proposal kernel by minimizing either the CSD or the KLD criteria. Of course, compared to the sophisticated adaptation strategies considered for adaptive importance sampling, we focus on elementary schemes, the computational burden being quickly a limiting factor in the SMC context.

To simplify the presentation, we consider in the sequel the adaptation of the proposal kernel; as shown above, it is of course possible and worthwhile to jointly optimize the adjustment weight and the proposal kernel, but for clarity we prefer to postpone the presentation of such a technique to a future work. The optimization of the adjustment weight function is in general rather complex: indeed, as mentioned above, the computation of the optimal adjustment weight function requires the computing of an integral. This integral can be evaluated in closed form only for a rather limited number of models; otherwise, a numerical approximation (based on cubature formulae, Monte Carlo etc) is required, which may therefore incur a quite substantial computational cost. If proper simplifications and approximations are not found (which are, most often, model specific) the gains in efficiency are not necessarily worth the extra cost. In state space (tracking) problems simple and efficient approximations, based either on the EKF or the UKF (see for example Andrieu et al. (2003) or Shen et al. (2004)), have been proposed for several models, but the validity of this sort of approximations cannot necessarily be extended to more general models.

In the light of the discussion above, a natural strategy for adaptive design of π_{aux} is to minimize the empirical estimate $\mathcal{E}$ (or CV^2) of the KLD (or CSD) over all proposal kernels belonging to some parametric family $\{r_\theta\}_{\theta \in \Theta}$. This can be done using straightforward adaptations of the two methods described in Section 3.2.1. We postpone a more precise description of the algorithms and implementation issues to after the next section, where more rigorous measure-theoretic notation is introduced and the main theoretical results are stated.

3.3 Notation and definitions

To state precisely the results, we will now use measure-theoretic notation. In the following we assume that all random variables are defined on a common probability

space $(\Omega, \mathcal{F}, \mathbb{P})$ and let, for any general state space $(\Xi, \mathcal{B}(\Xi))$, $\mathcal{P}(\Xi)$ and $\mathbb{B}(\Xi)$ be the sets of probability measures on $(\Xi, \mathcal{B}(\Xi))$ and measurable functions from Ξ to $\mathbb{R}$, respectively.

A kernel K from $(\Xi, \mathcal{B}(\Xi))$ to some other state space $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ is said *finite* if $K(\xi, \tilde{\Xi}) < \infty$ for all $\xi \in \Xi$ and to be a *transition kernel* if $K(\xi, \tilde{\Xi}) = 1$ for all $\xi \in \Xi$. Additionally, such a transition kernel is called *Markovian* if $\Xi = \tilde{\Xi}$. Moreover, K induces two operators, one transforming a function $f \in \mathbb{B}(\Xi \times \tilde{\Xi})$ satisfying $\int_{\tilde{\Xi}} |f(\xi, \tilde{\xi})| \, K(\xi, \mathrm{d}\tilde{\xi}) < \infty$ into another function

$$\xi \mapsto K(\xi, f) := \int_{\tilde{\Xi}} f(\xi, \tilde{\xi}) \, K(\xi, \mathrm{d}\tilde{\xi})$$

in $\mathbb{B}(\Xi)$; the other transforms a measure $\nu \in \mathcal{P}(\Xi)$ into another measure

$$A \mapsto \nu K(A) := \int_{\Xi} K(\xi, A) \, \nu(\mathrm{d}\xi) \tag{3.3.1}$$

in $\mathcal{P}(\tilde{\Xi})$. Furthermore, for any probability measure $\mu \in \mathcal{P}(\Xi)$ and function $f \in \mathbb{B}(\Xi)$ satisfying $\int_{\Xi} |f(\xi)| \, \mu(\mathrm{d}\xi) < \infty$, we write $\mu(f) := \int_{\Xi} f(\xi) \, \mu(\mathrm{d}\xi)$.

The *outer product* of the measure ν and the kernel K, denoted by $\nu \otimes K$, is defined as the measure on the product space $\Xi \times \tilde{\Xi}$, equipped with the product σ-algebra $\mathcal{B}(\Xi) \otimes \mathcal{B}(\tilde{\Xi})$, satisfying

$$\nu \otimes K(A) := \iint_{\Xi \times \tilde{\Xi}} \nu(\mathrm{d}\xi) \, K(\xi, \mathrm{d}\tilde{\xi}) \mathbb{1}_A(\xi, \xi') \tag{3.3.2}$$

for any $A \in \mathcal{B}(\Xi) \otimes \mathcal{B}(\tilde{\Xi})$. For a non-negative function $f \in \mathbb{B}(\Xi)$, we define the modulated measure $\nu[f]$ on $(\Xi, \mathcal{B}(\Xi))$ by

$$\nu[f](A) := \nu(f \mathbb{1}_A) \, , \tag{3.3.3}$$

for any $A \in \mathcal{B}(\Xi)$.

In the sequel, we will use the following definitions. A set C of real-valued functions on Ξ is said to be *proper* if the following conditions hold: **(i)** C is a linear space; **(ii)** if $g \in \mathsf{C}$ and f is measurable with $|f| \leq |g|$, then $|f| \in \mathsf{C}$; **(iii)** for all $c \in \mathbb{R}$, the constant function $f \equiv c$ belongs to C.

Additionally, in the following definition which regards asymptotic analysis results, we emphasis the dependency in N of the random variables involved by figuring N as a subscript of the particles, (adjustment) weights, and sums of the weights.

Definition 3.3.1. A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be *consistent* for the probability measure $\nu \in \mathcal{P}(\Xi)$ and the set C if, for any $f \in \mathsf{C}$, as $N \to \infty$,

$$\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \nu(f) \, ,$$

$$\Omega_N^{-1} \max_{1 \leq i \leq M_N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0 \, ,$$

where $\Omega_N := \sum_{i=1}^{M_N} \omega_{N,i}$.

Alternatively, we will sometimes say that the weighted sample in Definition 3.3.1 *targets* the measure ν.

Thus, suppose that we are given a weighted sample $\{(\xi_i, \omega_i)\}_{i=1}^{M_N}$ targeting $\nu \in \mathcal{P}(\Xi)$. We wish to transform this sample into a new weighted particle sample approximating the probability measure

$$\mu(\cdot) := \frac{\nu L(\cdot)}{\nu L(\tilde{\Xi})} = \frac{\int_{\Xi} L(\xi, \cdot) \, \nu(\mathrm{d}\xi)}{\int_{\Xi} L(\xi', \tilde{\Xi}) \, \nu(\mathrm{d}\xi')} \tag{3.3.4}$$

on some other state space $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$. Here L is a finite transition kernel from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$. As suggested by Pitt and Shephard (1999), an auxiliary variable corresponding to the selected stratum, and target the measure

$$\mu_{\text{aux}}(\{i\} \times A) := \frac{\omega_i L(\xi_i, \tilde{\Xi})}{\sum_{j=1}^{M_N} \omega_j L(\xi_j, \tilde{\Xi})} \left[\frac{L(\xi_i, A)}{L(\xi_i, \tilde{\Xi})} \right] \tag{3.3.5}$$

on the product space $\{1, \dots, M_N\} \times \Xi$. Since μ_N is the marginal distribution of μ_{aux} with respect to the particle position, we may sample from μ_N by simulating instead a set $\{(I_i, \tilde{\xi}_i)\}_{i=1}^{\tilde{M}_N}$ of indices and particle positions from the instrumental distribution

$$\pi_{\text{aux}}(\{i\} \times A) := \frac{\omega_i \psi_i}{\sum_{j=1}^{M_N} \omega_j \psi_j} R(\xi_i, A) \tag{3.3.6}$$

and assigning each draw $(I_i, \tilde{\xi}_i)$ the weight

$$\tilde{\omega}_i := \psi_{I_i}^{-1} \frac{\mathrm{d}L(\xi_{I_i}, \cdot)}{\mathrm{d}R(\xi_{I_i}, \cdot)}(\tilde{\xi}_i)$$

being proportional to $\mathrm{d}\mu_{\text{aux}}/\mathrm{d}\pi_{\text{aux}}(I_i, \tilde{\xi}_i)$—the formal difference with Equation (3.2.23) lies only in the use of Radon-Nykodym derivatives of the two kernels rather than densities w.r.t. Lebesgue measure. Hereafter, we discard the indices and take $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^{\tilde{M}_N}$ as an approximation of μ. The algorithm is summarised below.

Algorithm 3.3.1 Nonadaptive APF

Require: $\{(\xi_i, \omega_i)\}_{i=1}^{M_N}$ targets ν.

1: Draw $\{I_i\}_{i=1}^{\tilde{M}_N} \sim \mathcal{M}(\tilde{M}_N, \{\omega_j \psi_j / \sum_{\ell=1}^{M_N} \omega_\ell \psi_\ell\}_{j=1}^{M_N})$,

2: simulate $\{\tilde{\xi}_i\}_{i=1}^{\tilde{M}_N} \sim \bigotimes_{i=1}^{\tilde{M}_N} R(\xi_{I_i}, \cdot)$,

3: set, for all $i \in \{1, \dots, \tilde{M}_N\}$,

$$\tilde{\omega}_i \leftarrow \psi_{I_i}^{-1} \mathrm{d}L(\xi_{I_i}, \cdot) / \mathrm{d}R(\xi_{I_i}, \cdot)(\tilde{\xi}_i) \ .$$

4: take $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^{\tilde{M}_N}$ as an approximation of μ.

3.4 Theoretical results

As in Definition 3.3.1, because we state asymptotic results, we emphasize throughout this section the dependency in N of the random variables involved by figuring N as a subscript of the particles, (adjustment) weights, and sums of the weights. Consider the following assumptions.

(A6) *The initial sample* $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ *is consistent for* (ν, C).

(A7) *There exists a function* $\Psi : \Xi \to \mathbb{R}^+$ *such that* $\psi_{N,i} = \Psi(\xi_{N,i})$*; moreover,* $\Psi \in \mathsf{C} \cap \mathsf{L}^1(\Xi, \nu)$ *and* $L(\cdot, \tilde{\Xi}) \in \mathsf{C}$.

Under these assumptions we define for $(\xi, \tilde{\xi}) \in \Xi \times \tilde{\Xi}$ the weight function

$$\Phi(\xi, \tilde{\xi}) := \Psi^{-1}(\xi) \frac{\mathrm{d}L(\xi, \cdot)}{\mathrm{d}R(\xi, \cdot)}(\tilde{\xi}) \ , \tag{3.4.1}$$

so that for every index i, $\tilde{\omega}_{N,i} = \Phi(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})$. The following result describes how the consistency property is passed through one step of the APF algorithm. A somewhat less general version of this result was also proved in Douc et al. (2008) (Theorem 3.1).

Proposition 3.4.1. *Assume* **(A6, A7)**. *Then the weighted sample* $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ *is consistent for* $(\nu, \tilde{\mathsf{C}})$, *where* $\tilde{\mathsf{C}} := \{f \in \mathsf{L}^1(\tilde{\boldsymbol{\Xi}}, \mu), L(\cdot, |f|) \in \mathsf{C}\}$.

Proof. The result above is a direct consequence of Lemma 3.4.2 and the fact that the set $\tilde{\mathsf{C}}$ is proper. □

Let μ and ν be two probability measures in $\mathcal{P}(\boldsymbol{\Lambda})$ such that μ is absolutely continuous with respect to ν. We then recall that the KLD and the CSD are, respectively, given by

$$d_{\mathrm{KL}}(\mu||\nu) := \int_{\boldsymbol{\Lambda}} \log[\mathrm{d}\mu/\mathrm{d}\nu(\lambda)]\, \mu(\mathrm{d}\lambda) \,,$$
$$d_{\chi^2}(\mu||\nu) := \int_{\boldsymbol{\Lambda}} [\mathrm{d}\mu/\mathrm{d}\nu(\lambda) - 1]^2\, \nu(\mathrm{d}\lambda) \,.$$

Define the two probability measures on the product space $(\boldsymbol{\Xi} \times \tilde{\boldsymbol{\Xi}}, \mathcal{B}(\boldsymbol{\Xi}) \otimes \mathcal{B}(\tilde{\boldsymbol{\Xi}}))$:

$$\mu^*(A) := \frac{\nu \otimes L}{\nu L(\tilde{\boldsymbol{\Xi}})}(A) = \frac{\iint_{\boldsymbol{\Xi}\times\tilde{\boldsymbol{\Xi}}} \nu(\mathrm{d}\xi)\, L(\xi, \mathrm{d}\xi') \mathbb{1}_A(\xi, \xi')}{\iint_{\boldsymbol{\Xi}\times\tilde{\boldsymbol{\Xi}}} \nu(\mathrm{d}\xi)\, L(\xi, \mathrm{d}\xi')} \,, \tag{3.4.2}$$

$$\pi^*_\Psi(A) := \frac{\nu[\Psi] \otimes R}{\nu(\Psi)}(A) = \frac{\iint_{\boldsymbol{\Xi}\times\tilde{\boldsymbol{\Xi}}} \nu(\mathrm{d}\xi)\Psi(\xi)\, R(\xi, \mathrm{d}\xi') \mathbb{1}_A(\xi, \xi')}{\iint_{\boldsymbol{\Xi}\times\tilde{\boldsymbol{\Xi}}} \nu(\mathrm{d}\xi)\Psi(\xi)\, R(\xi, \mathrm{d}\xi')} \,, \tag{3.4.3}$$

where $A \in \mathcal{B}(\boldsymbol{\Xi}) \otimes \mathcal{B}(\tilde{\boldsymbol{\Xi}})$ and the outer product $\otimes$ of a measure and a kernel is defined in (3.3.2).

Theorem 3.4.1. *Assume* **(A6, A7)**. *Then the following holds as* $N \to \infty$.

(i) If $L(\cdot, |\log \Phi|) \in \mathsf{C} \cap \mathsf{L}^1(\boldsymbol{\Xi}, \nu)$, *then*

$$d_{\mathrm{KL}}(\mu^N_{\mathrm{aux}}||\pi^N_{\mathrm{aux}}) \xrightarrow{\mathbb{P}} d_{\mathrm{KL}}\left(\mu^* \| \pi^*_\Psi\right) \,, \tag{3.4.4}$$

(ii) If $L(\cdot, \Phi) \in \mathsf{C}$, *then*

$$d_{\chi^2}(\mu^N_{\mathrm{aux}}||\pi^N_{\mathrm{aux}}) \xrightarrow{\mathbb{P}} d_{\chi^2}\left(\mu^* \| \pi^*_\Psi\right) \,, \tag{3.4.5}$$

Additionally, $\mathcal{E}$ and CV^2, defined in (3.2.8) and (3.2.4) respectively, converge to the same limits.

Theorem 3.4.2. *Assume* **(A6, A7)**. *Then the following holds as* $N \to \infty$.

(i) If $L(\cdot, |\log \Phi|) \in \mathsf{C} \cap \mathsf{L}^1(\boldsymbol{\Xi}, \nu)$, *then*

$$\mathcal{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \xrightarrow{\mathbb{P}} d_{\mathrm{KL}}\left(\mu^* \| \pi^*_\Psi\right) \,. \tag{3.4.6}$$

(ii) If $L(\cdot, \Phi) \in \mathsf{C}$, *then*

$$\mathrm{CV}^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) \xrightarrow{\mathbb{P}} d_{\chi^2}\left(\mu^* \| \pi^*_\Psi\right) \,. \tag{3.4.7}$$

We preface the proofs of Theorems 3.4.1 and 3.4.2 with the following two lemma.

Lemma 3.4.1. *Assume* **(A7)**. *Then the following identities hold.*

i) $$d_{\mathrm{KL}}\left(\mu^* \| \pi^*_\Psi\right) = \nu \otimes L\{\log[\Phi\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\}/\nu L(\tilde{\mathsf{X}}) \,,$$

ii) $$d_{\chi^2}\left(\mu^* \| \pi^*_\Psi\right) = \nu(\Psi)\, \nu \otimes L(\Phi)/[\nu L(\tilde{\mathsf{X}})]^2 - 1 \,.$$

Proof. We denote by $q(\xi, \xi')$ the Radon-Nikodym derivative of the probability measure μ^* with respect to $\nu \otimes R$ (where the outer product $\otimes$ of a measure and a kernel is defined in (3.3.2)), that is,

$$q(\xi, \xi') := \frac{\frac{\mathrm{d}L(\xi,\cdot)}{\mathrm{d}R(\xi,\cdot)}(\xi')}{\iint_{\mathsf{X}\times\tilde{\mathsf{X}}} \nu(\mathrm{d}\xi)\, L(\xi, \mathrm{d}\xi')} , \tag{3.4.8}$$

and by $p(\xi)$ the Radon-Nikodym derivative of the probability measure π^* with respect to $\nu \otimes R$:

$$p(\xi) = \frac{\Psi(\xi)}{\nu(\Psi)} . \tag{3.4.9}$$

Using the notation above and definition (3.4.1) of the weight function Φ, we have

$$\frac{\Phi(\xi, \xi')\nu(\Psi)}{\nu L(\tilde{\mathsf{X}})} = \frac{\nu(\Psi)\frac{\mathrm{d}L(\xi,\cdot)}{\mathrm{d}R(\xi,\cdot)}(\xi')}{\Psi(\xi)\nu L(\tilde{\mathsf{X}})} = p^{-1}(\xi)q(\xi, \xi') .$$

This implies that

$$\begin{aligned} d_{\mathrm{KL}}\left(\mu^* \| \pi^*_\Psi\right) &= \iint_{\mathsf{X}\times\tilde{\mathsf{X}}} \nu(\mathrm{d}\xi)\, R(\xi, \mathrm{d}\xi') q(\xi, \xi') \log\left(p^{-1}(\xi)q(\xi, \xi')\right) \\ &= \nu \otimes L\{\log[\Phi\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\}/\nu L(\tilde{\mathsf{X}}) , \end{aligned}$$

which establishes assertion *i)*. Similarly, we may write

$$\begin{aligned} d_{\chi^2}\left(\mu^* \| \pi^*_\Psi\right) &= \iint_{\mathsf{X}\times\tilde{\mathsf{X}}} \nu(\mathrm{d}\xi)\, R(\xi, \mathrm{d}\xi') p^{-1}(\xi) q^2(\xi, \xi') - 1 \\ &= \frac{\iint_{\mathsf{X}\times\tilde{\mathsf{X}}} \nu(\Psi)\, \nu(\mathrm{d}\xi)\, R(\xi, \mathrm{d}\xi') \left[\frac{\mathrm{d}L(\xi,\cdot)}{\mathrm{d}R(\xi,\cdot)}(\xi')\right]^2 \Psi^{-1}(\xi)}{[\nu L(\tilde{\mathsf{X}})]^2} - 1 \\ &= \nu(\Psi)\, \nu \otimes L(\Phi)/[\nu L(\tilde{\mathsf{X}})]^2 - 1 , \end{aligned}$$

showing assertion *ii)*. □

Lemma 3.4.2. *Assume* **(A6, A7)** *and let* $\mathsf{C}^* := \{f \in \mathbb{B}(\mathsf{X} \times \tilde{\mathsf{X}}) : L(\cdot, |f|) \in \mathsf{C} \cap \mathsf{L}^1(\mathsf{X}, \nu)\}$. *Then, for all* $f \in \mathsf{C}^*$, *as* $N \to \infty$,

$$\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \nu \otimes L(f)/\nu L(\tilde{\mathsf{X}})$$

Proof. It is enough to prove that

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \nu \otimes L(f)/\nu(\Psi) , \tag{3.4.10}$$

for all $f \in \mathsf{C}^*$; indeed, since the function $f \equiv 1$ belongs to C^* under **(A7)**, the result of the lemma will follow from (3.4.10) by Slutsky's theorem. Define the measure $\varphi(A) := \nu(\Psi \mathbb{1}_A)/\nu(\Psi)$, with $A \in \mathcal{B}(\mathsf{X})$. By applying Theorem 1 in Douc and Moulines (2008) we conclude that the weighted sample $\{(\xi_{N,i}, \psi_{N,i})\}_{i=1}^{M_N}$ is consistent for $(\varphi, \{f \in \mathsf{L}^1(\mathsf{X}, \varphi) : \Psi|f| \in \mathsf{C}\})$. Moreover, by Theorem 2 in the same paper this is also true for the uniformly weighted sample $\{(\xi_{N,I_{N,i}}, 1)\}_{i=1}^{\tilde{M}_N}$ (see the proof of Theorem 3.1 in Douc et al. (2008) for details). By definition, for $f \in \mathsf{C}^*$, $\varphi \otimes R(\Phi|f|)\, \nu(\Psi) = \nu \otimes L(|f|) < \infty$ and $\Psi R(\cdot, \Phi|f|) =$

$L(\cdot, |f|) \in \mathsf{C}$. Hence, we conclude that $R(\cdot, \Phi|f|)$ and thus $R(\cdot, \Phi f)$ belong to the proper set $\{f \in \mathsf{L}^1(\mathsf{X}, \varphi) : \Psi|f| \in \mathsf{C}\}$. This implies the convergence

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[\tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \middle| \mathcal{F}_N\right] = \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} R(\xi_{N,I_{N,i}}, \Phi f) \xrightarrow{\mathbb{P}} \varphi \otimes R(\Phi f) = \nu \otimes L(f)/\nu(\Psi) \,, \quad (3.4.11)$$

where $\mathcal{F}_N := \sigma(\{\xi_{N,I_{N,i}}\}_{i=1}^{\tilde{M}_N})$ denotes the σ-algebra generated by the selected particles. It thus suffices to establish that

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \left\{ \mathbb{E}\left[\tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \middle| \mathcal{F}_N\right] - \tilde{\omega}_{N,i} f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i}) \right\} \xrightarrow{\mathbb{P}} 0 \,, \quad (3.4.12)$$

and we do this, following the lines of the proof of Theorem 1 in Douc and Moulines (2008), by verifying the two conditions of Theorem 11 in the same work. The sequence

$$\left\{ \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \middle| \mathcal{F}_N\right] \right\}_N$$

is tight since it tends to $\nu \otimes L(|f|)/\nu(\Psi)$ in probability (cf. (3.4.11)). Thus, the first condition is satisfied. To verify the second condition, take $\epsilon > 0$ and consider, for any $C > 0$, the decomposition

$$\begin{aligned} &\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \mathbb{1}_{\{\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \geq \epsilon\}} \middle| \mathcal{F}_N\right] \\ &\leq \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} R\left(\xi_{N,I_{N,i}}, \Phi|f| \mathbb{1}_{\{\Phi|f| \geq C\}}\right) + \mathbb{1}_{\{\epsilon \tilde{M}_N < C\}} \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \middle| \mathcal{F}_N\right] \,. \end{aligned}$$

Since $R(\cdot, \Phi f)$ belongs to the proper set $\{f \in \mathsf{L}^1(\mathsf{X}, \varphi) : \Psi|f| \in \mathsf{C}\}$, so does the function $R(\cdot, \Phi|f| \mathbb{1}\{\Phi|f| \geq C\})$. Thus, since the indicator $\mathbb{1}\{\epsilon \tilde{M}_N < C\}$ tends to zero, we conclude that the upper bound above has the limit $\varphi \otimes R(\Phi|f| \mathbb{1}\{\Phi|f| \geq C\})$; however, by dominated convergence this limit can be made arbitrarily small by increasing C. Hence

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \mathbb{1}_{\{\tilde{\omega}_{N,i} |f(\xi_{N,I_{N,i}}, \tilde{\xi}_{N,i})| \geq \epsilon\}} \middle| \mathcal{F}_N\right] \xrightarrow{\mathbb{P}} 0 \,,$$

which verifies the second condition of Theorem 11 in Douc and Moulines (2008). Thus, (3.4.12) follows. □

Proof of Theorem 3.4.1. We start with *i)*. In the light of Lemma 3.4.1 we establish the limit

$$d_{\mathrm{KL}}(\mu_{\mathrm{aux}}^N || \pi_{\mathrm{aux}}^N) \xrightarrow{\mathbb{P}} \nu \otimes L\{\log[\Phi\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\}/\nu L(\tilde{\mathsf{X}}) \,, \quad (3.4.13)$$

as $N \to \infty$. Hence, recall the definition (given in Section 3.4) of the KLD and write, for any index $m \in \{1, \ldots, \tilde{M}_N\}$,

$$\begin{aligned} d_{\mathrm{KL}}(\mu_{\mathrm{aux}}^N || \pi_{\mathrm{aux}}^N) = \sum_{i=1}^{M_N} \mathbb{E}_{\mu_{\mathrm{aux}}^N}\left[\log \Phi(\xi_{N,I_{N,m}}, \tilde{\xi}_{N,m}) \middle| I_{N,m} = i\right] \mu_{\mathrm{aux}}^N(\{i\} \times \tilde{\mathsf{X}}) \\ + \log\left[\frac{\sum_{j=1}^{M_N} \omega_{N,j} \psi_{N,j}}{\sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, \tilde{\mathsf{X}})}\right] \,, \quad (3.4.14) \end{aligned}$$

where $\mathbb{E}_{\mu^N_{\text{aux}}}$ denotes the expectation associated with the random measure μ^N_{aux}. For each term of the sum in (3.4.14) we have

$$\mathbb{E}_{\mu^N_{\text{aux}}}\left[\log\Phi(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\middle| I_{N,m}=i\right]\mu^N_{\text{aux}}(\{i\}\times\tilde{\mathsf{X}}) = \frac{\omega_{N,i}L(\xi_{N,i},\log\Phi)}{\sum_{j=1}^{M_N}\omega_{N,i}L(\xi_{N,j},\tilde{\mathsf{X}})}\,,$$

and by using the consistency of $\{(\xi_{N,i},\omega_{N,i})\}_{i=1}^{M_N}$ (under **(A6)**) we obtain the limit

$$\sum_{i=1}^{M_N}\mathbb{E}_{\mu^N_{\text{aux}}}\left[\log\Phi(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\middle| I_{N,m}=i\right]\mu^N_{\text{aux}}(\{i\}\times\tilde{\mathsf{X}}) \xrightarrow{\mathbb{P}} \nu\otimes L(\log\Phi)/\nu L(\tilde{\mathsf{X}})\,,$$

where we used that $L(\cdot,|\log\Phi|)\in\mathsf{C}$ by assumption, implying, since C is proper, $L(\cdot,\log\Phi)\in\mathsf{C}$. Moreover, under **(A7)**, by the continuous mapping theorem,

$$\log\left[\frac{\sum_{j=1}^{M_N}\omega_{N,j}\psi_{N,j}}{\sum_{\ell=1}^{M_N}\omega_{N,\ell}L(\xi_{N,\ell},\tilde{\mathsf{X}})}\right]\xrightarrow{\mathbb{P}}\log[\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\,,$$

yielding

$$\begin{aligned}d_{\text{KL}}(\mu^N_{\text{aux}}||\pi^N_{\text{aux}})\xrightarrow{\mathbb{P}}\nu\otimes L(\log\Phi)/\nu L(\tilde{\mathsf{X}})+\log[\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]&\\ =\nu\otimes L\{\log[\Phi\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\}/\nu L(\tilde{\mathsf{X}})\,,&\end{aligned}$$

which establishes (3.4.13) and, consequently, *i)*.

To prove *ii)* we show that

$$d_{\chi^2}(\mu^N_{\text{aux}}||\pi^N_{\text{aux}})\xrightarrow{\mathbb{P}}\nu(\Psi)\,\nu\otimes L(\Phi)/[\nu L(\tilde{\mathsf{X}})]^2-1 \tag{3.4.15}$$

and apply Lemma 3.4.1. Thus, recall the definition of the CSD and write, for any index $m\in\{1,\ldots,\tilde{M}_N\}$,

$$\begin{aligned}d_{\chi^2}(\mu^N_{\text{aux}}||\pi^N_{\text{aux}}) &= \mathbb{E}_{\mu^N_{\text{aux}}}\left[\frac{\mathrm{d}\mu^N_{\text{aux}}}{\mathrm{d}\pi^N_{\text{aux}}}(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\right]-1\\ &=\sum_{i=1}^{M_N}\mathbb{E}_{\mu^N_{\text{aux}}}\left[\frac{\mathrm{d}\mu^N_{\text{aux}}}{\mathrm{d}\pi^N_{\text{aux}}}(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\middle| I_{N,m}=i\right]\mu^N_{\text{aux}}(\{i\}\times\tilde{\mathsf{X}})-1\,.\end{aligned}$$

Here

$$\begin{aligned}\mathbb{E}_{\mu^N_{\text{aux}}}\left[\frac{\mathrm{d}\mu^N_{\text{aux}}}{\mathrm{d}\pi^N_{\text{aux}}}(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\middle| I_{N,m}=i\right]\mu^N_{\text{aux}}(\{i\}\times\tilde{\mathsf{X}})&\\ =\omega_{N,i}L(\xi_{N,i},\Phi)\left[\sum_{j=1}^{M_N}\omega_{N,i}L(\xi_{N,j},\tilde{\mathsf{X}})\right]^{-2}\sum_{j=1}^{M_N}\omega_{N,i}\psi_{N,i}\,,&\end{aligned}$$

and using the consistency of $\{(\xi_{N,i},\omega_{N,i})\}_{i=1}^{M_N}$ yields the limit

$$\sum_{i=1}^{M_N}\mathbb{E}_{\mu^N_{\text{aux}}}\left[\frac{\mathrm{d}\mu^N_{\text{aux}}}{\mathrm{d}\pi^N_{\text{aux}}}(\xi_{N,I_{N,m}},\tilde{\xi}_{N,m})\middle| I_{N,m}=i\right]\mu^N_{\text{aux}}(\{i\}\times\tilde{\mathsf{X}})\xrightarrow{\mathbb{P}}\nu(\Psi)\nu\otimes L(\Phi)/[\nu L(\tilde{\mathsf{X}})]^2\,.$$

which proves (3.4.15). This completes the proof of *ii)*. □

Proof of Theorem 3.4.2. Applying directly Lemma 3.4.2 for $f = \log \Phi$ (which belongs to C^* by assumption) and the limit (3.4.10) for $f \equiv 1$ yields, by the continuous mapping theorem,

$$\mathcal{E}(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) = \tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} \log \tilde{\omega}_{N,i} + \log(\tilde{M}_N \tilde{\Omega}_N^{-1})$$
$$\xrightarrow{\mathbb{P}} \nu \otimes L(\log \Phi)/\nu L(\tilde{\mathsf{X}}) + \log[\nu(\Psi)/\nu L(\tilde{\mathsf{X}})]$$
$$= \nu \otimes L\{\log[\Phi \nu(\Psi)/\nu L(\tilde{\mathsf{X}})]\}/\nu L(\tilde{\mathsf{X}}) \, .$$

Now, we complete the proof of assertion *i)* by applying Lemma 3.4.1.

We turn to *ii)*. Since Φ belongs to C^* by assumption, we obtain, by applying Lemma 3.4.2 together with (3.4.10),

$$\mathrm{CV}^2(\{\tilde{\omega}_{N,i}\}_{i=1}^{\tilde{M}_N}) = (\tilde{M}_N \tilde{\Omega}_N^{-1}) \tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i}^2 - 1$$
$$\xrightarrow{\mathbb{P}} \nu_{\mathrm{KL}}(\Psi) := \nu(\Psi)\, \nu \otimes L(\Phi)/[\nu L(\tilde{\mathsf{X}})]^2 - 1 \, . \quad (3.4.16)$$

From this *ii)* follows via Lemma 3.4.1. □

Next, it is shown that the adjustment weight function can be chosen so as to minimize the RHS of (3.4.4) and (3.4.5).

Proposition 3.4.2. *Assume* **(A6, A7)**. *Then the following holds.*

(i) If $L(\cdot, |\log \Phi|) \in \mathsf{C} \cap \mathsf{L}^1(\Xi, \nu)$, *then*

$$\arg\min_\Psi d_{\mathrm{KL}}(\mu^* \| \pi_\Psi^*) := \Psi_{\mathrm{KL},R}^* \text{ where } \Psi_{\mathrm{KL},R}^*(\xi) := L(\xi, \tilde{\Xi}) \, .$$

(ii) If $L(\cdot, \Phi) \in \mathsf{C}$, *then*

$$\arg\min_\Psi d_{\chi^2}(\mu^* \| \pi_\Psi^*) := \Psi_{\chi^2,R}^* \text{ where } \Psi_{\chi^2,R}^*(\xi) := \sqrt{\int_{\tilde{\Xi}} \frac{\mathrm{d}L(\xi,\cdot)}{\mathrm{d}R(\xi,\cdot)}(\tilde{\xi})\, L(\xi, \mathrm{d}\tilde{\xi})} \, .$$

Proof. Define by $q(\xi) := \int_{\tilde{\mathsf{X}}} R(\xi, \mathrm{d}\xi') q(\xi, \xi')$ the marginal density of the measure on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, $A \in \mathcal{B}(\mathsf{X}) \mapsto \mu^*(A \times \tilde{\mathsf{X}})$. We denote by $q(\xi'|\xi) = q(\xi, \xi')/q(\xi)$ the conditional distribution. By the chain rule of the entropy, (the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other), we may split the KLD between μ^* and π^* as follows,

$$d_{\mathrm{KL}}(\mu_{\mathrm{aux}}^N || \pi_{\mathrm{aux}}^N) = \int_{\mathsf{X}} \nu(\mathrm{d}\xi) q(\xi) \log(p^{-1}(\xi) q(\xi)) + \iint_{\mathsf{X} \times \tilde{\mathsf{X}}} \nu(\mathrm{d}\xi) R(\xi, \mathrm{d}\xi') q(\xi, \xi') \log q(\xi|\xi') \, .$$

The second term in the RHS of the previous equation does not depend on the adjustment multiplier weight Ψ. The first term is canceled if we set $p = q$, i.e. if

$$\frac{\Psi(\xi)}{\nu(\Psi)} = \int_{\tilde{\mathsf{X}}} R(\xi, \mathrm{d}\xi') q(\xi, \xi') = \frac{L(\xi, \tilde{\mathsf{X}})}{\int_{\mathsf{X}} \nu(\mathrm{d}\xi) L(\xi, \tilde{\mathsf{X}})} \, ,$$

which establishes assertion *i)*.

Consider now assertion *ii)*. Note first that

$$\begin{aligned}
&\iint_{\mathsf{X}\times\tilde{\mathsf{X}}} \nu(\mathrm{d}\xi)\, R(\xi, \mathrm{d}\xi') p^{-1}(\xi) q^2(\xi, \xi') - 1 \\
&\quad = \int_{\mathsf{X}} \nu(\mathrm{d}\xi)\, p^{-1}(\xi) g^2(\xi) - 1 \\
&\quad = \nu^2(g) \left\{ \int_{\mathsf{X}} \nu(\mathrm{d}\xi) \frac{g^2(\xi)}{p(\xi)\nu^2(g)} - 1 \right\} + \nu^2(g) - 1 \,, \qquad (3.4.17)
\end{aligned}$$

where

$$g^2(\xi) = \int_{\tilde{\mathsf{X}}} R(\xi, \mathrm{d}\xi')\, q^2(\xi, \xi') \,.$$

The first term on the RHS of (3.4.17) is the CSD between the probability distributions associated with the densities $g/\nu(g)$ and $\Psi/\nu(\Psi)$ with respect to ν. The second term does not depend on Ψ and the optimal value of the adjustment multiplier weight is obtained by canceling the first term. This establishes assertion *ii)*. □

It is worthwhile to notice that the optimal adjustment weights for the KLD do not depend on the proposal kernel R. The minimal value $d_{\mathrm{KL}}(\mu^* \| \pi^*_{\Psi^*_{\mathrm{KL},R}})$ of the limiting KLD is the conditional relative entropy between μ^* and π^*.

In both cases, letting $R(\cdot, A) = L(\cdot, A)/L(\cdot, \tilde{\boldsymbol{\Xi}})$ yields, as we may expect, the optimal adjustment multiplier weight function $\Psi^*_{\mathrm{KL},R}(\cdot) = \Psi^*_{\chi^2,R}(\cdot) = L(\cdot, \tilde{\boldsymbol{\Xi}})$, resulting in uniform importance weights $\tilde{\omega}_{N,i} \equiv 1$.

It is possible to relate the asymptotic CSD (3.4.5) between μ^N_{aux} and π^N_{aux} to the asymptotic variance of the estimator $\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i})$ of an expectation $\mu(f)$ for a given integrable target function f. More specifically, suppose that $\tilde{M}_N/M_N \to \ell \in [0, \infty]$ as $N \to \infty$ and that the initial sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ satisfies, for all f belonging to a given class A of functions, the central limit theorem

$$a_N \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} [f(\xi_{N,i}) - \mu(f)] \xrightarrow{\mathcal{D}} \mathcal{N}[0, \sigma^2(f)] \,, \qquad (3.4.18)$$

where the sequence $\{a_N\}_N$ is such that $a_N M_N \to \beta \in [0, \infty)$ as $N \to \infty$ and $\sigma : \mathsf{A} \to \mathbb{R}^+$ is a functional. Then the sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$ produced in Algorithm 3.3.1 is, as showed in (Douc et al., 2008, Theorem 3.2), asymptotically normal for a class of functions $\tilde{\mathsf{A}}$ in the sense that, for all $f \in \tilde{\mathsf{A}}$,

$$\tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} [f(\tilde{\xi}_{N,i}) - \mu(f)] \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \tilde{\sigma}^2[\Psi](f)\} \,,$$

where

$$\tilde{\sigma}^2[\Psi](f) = \sigma^2\{L_{\cdot, f-\mu(f)}\}/[\nu L(\cdot, \tilde{\boldsymbol{\Xi}})]^2 + \beta \ell^{-1} \nu(\Psi R\{\cdot, \Phi^2[f - \mu(f)]^2\}) \nu(\Psi)/[\nu L(\tilde{\boldsymbol{\Xi}})]^2$$

and, recalling the definition (3.3.3) of a modulated measure,

$$\begin{aligned}
&\nu(\Psi R\{\cdot, \Phi^2[f - \mu(f)]^2\}) \nu(\Psi)/[\nu L(\tilde{\boldsymbol{\Xi}})]^2 \\
&\qquad = \mu^2(|f|)\, d_{\chi^2}\{\mu^*[|f|]/\mu^*(|f|) \| \pi^*\} \\
&\qquad - 2\mu(f)\mu(f_+^{1/2})\, d_{\chi^2}\{\mu^*[f_+^{1/2}]/\mu^*(f_+^{1/2}) \| \pi^*\} \\
&\qquad + 2\mu(f)\mu(f_-^{1/2})\, d_{\chi^2}\{\mu^*[f_-^{1/2}]/\mu^*(f_-^{1/2}) \| \pi^*\} \\
&\qquad + \mu^2(f)\, d_{\chi^2}(\mu^* \| \pi^*) + \mu^2(|f|) - \mu^2(f) \,. \qquad (3.4.19)
\end{aligned}$$

Here $f_+ := \max(f, 0)$ and $f_- := \max(-f, 0)$ denote the positive and negative parts of f, respectively, and $\mu^*(|f|)$ refers to the expectation of the extended function $|f| : (\xi, \tilde{\xi}) \in \boldsymbol{\Xi} \times \tilde{\boldsymbol{\Xi}} \mapsto |f(\tilde{\xi})| \in \mathbb{R}^+$ under μ^* (and similarly for $\mu^*(f_\pm^{1/2})$). From (3.4.19) we deduce that decreasing $d_{\chi^2}(\mu^*||\pi^*)$ will imply a decrease of asymptotic variance if the discrepancy between μ^* and modulated measure $\mu^*[|f|]/\mu^*(|f|)$ is not too large, that is, we deal with a target function f with a regular behavour in the support of $\mu^*(\boldsymbol{\Xi} \times \cdot)$.

3.5 Adaptive importance sampling

3.5.1 APF adaptation by minimization of estimated KLD and CSD over a parametric family

Assume that there exists a random noise variable ϵ, having distribution λ on some measurable space $(\boldsymbol{\Lambda}, \mathcal{B}(\boldsymbol{\Lambda}))$, and a family $\{F_\theta\}_{\theta \in \Theta}$ of mappings from $\boldsymbol{\Xi} \times \boldsymbol{\Lambda}$ to $\tilde{\boldsymbol{\Xi}}$ such that we are able to simulate $\tilde{\xi} \sim R_\theta(\xi, \cdot)$, for $\xi \in \boldsymbol{\Xi}$, by simulating $\epsilon \sim \lambda$ and letting $\tilde{\xi} = F_\theta(\xi, \epsilon)$. We denote by Φ_θ the importance weight function associated with R_θ, see (3.4.1) and set $\Phi_\theta \circ F_\theta(\xi, \epsilon) := \Phi_\theta(\xi, F_\theta(\xi, \epsilon))$.

Assume that **(A6)** holds and suppose that we have simulated, as in the first step of Algorithm 3.3.1, indices $\{I_i\}_{i=1}^{\tilde{M}_N}$ and noise variables $\{\epsilon_i\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$. Now, keeping these indices and noise variables fixed, we can form an idea of how the KLD varies with θ via the mapping $\theta \mapsto \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_i}, \epsilon_i)\}_{i=1}^{\tilde{M}_N})$. Similarly, the CSD can be studied by using CV^2 instead of $\mathcal{E}$. This suggests an algorithm in which the particles are reproposed using $R_{\theta_N^*}$, with $\theta_N^* = \arg\min_{\theta \in \Theta} \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_i}, \epsilon_i)\}_{i=1}^{\tilde{M}_N})$.

This procedure is summarised in Algorithm 3.5.1, and its modification for minimization of the empirical CSD is straightforward.

Algorithm 3.5.1 Adaptive APF

Require: **(A6)**

1: Draw $\{I_i\}_{i=1}^{\tilde{M}_N} \sim \mathcal{M}(\tilde{M}_N, \{\omega_j \psi_j / \sum_{\ell=1}^{M_N} \omega_\ell \psi_\ell\}_{j=1}^{M_N})$,

2: simulate $\{\epsilon_i\}_{i=1}^{\tilde{M}_N} \sim \lambda^{\otimes \tilde{M}_N}$,

3: $\theta_N^* \leftarrow \arg\min_{\theta \in \Theta} \mathcal{E}(\{\Phi_\theta \circ F_\theta(\xi_{I_i}, \epsilon_i)\}_{i=1}^{\tilde{M}_N})$,

4: set

$$\tilde{\xi}_i \overset{\forall i}{\leftarrow} F_{\theta_N^*}(\xi_{I_i}, \epsilon_i)$$

5: update

$$\tilde{\omega}_i \overset{\forall i}{\leftarrow} \Phi_{\theta_N^*}(\xi_{I_i}, \tilde{\xi}_i) \,,$$

6: let $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^{\tilde{M}_N}$ approximate μ.

Remark 3.5.1. A slight modification of Algorithm 3.5.1, lowering the added computational burden, is to apply the adaptation mechanism only when the estimated KLD (or CSD) is above a chosen threshold.

3.5.2 APF adaptation by cross-entropy methods

Here we construct an algorithm which selects a proposal kernel from a parametric family in a way that minimizes the KLD between the instrumental mixture distribution and the target mixture μ_{aux} (defined in (3.3.5)). Thus, recall that we are given an initial

sample $\{(\xi_i, \omega_i)\}_{i=1}^{M_N}$; we then use IS to approximate the target auxiliary distribution μ_{aux} by sampling from the instrumental auxiliary distribution

$$\pi_{\text{aux}}^{\theta}(\{i\} \times A) := \frac{\omega_i \psi_i}{\sum_{j=1}^{M_N} \omega_j \psi_j} R_\theta(\xi_i, A) \,, \tag{3.5.1}$$

which is a straightforward modification of (3.3.6) where R is replaced by R_θ, that is, a transition kernel from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ belonging to the parametric family $\{R_\theta(\xi, \cdot) : \xi \in \Xi, \quad \theta \in \Theta\}$.

We aim at finding the parameter θ^* which realises the minimum of $\theta \mapsto d_{\text{KL}}(\mu_{\text{aux}} || \pi_{\text{aux}}^{\theta})$ over the parameter space Θ, where

$$d_{\text{KL}}(\mu_{\text{aux}} || \pi_{\text{aux}}^{\theta}) = \sum_{i=1}^{M_N} \int_{\tilde{\Xi}} \log \left(\frac{\mathrm{d}\mu_{\text{aux}}}{\mathrm{d}\pi_{\text{aux}}^{\theta}}(i, \tilde{\xi}) \right) \mu_{\text{aux}}(i, \mathrm{d}\tilde{\xi}) \,. \tag{3.5.2}$$

Since the expectation in (3.5.2) is intractable in most cases, the key idea is to approximate it iteratively using IS from more and more accurate approximations—this idea has been successfully used in CE methods; see e.g. Rubinstein and Kroese (2004). At iteration ℓ, denote by $\theta_N^\ell \in \Theta$ the current fit of the parameter. Each iteration of the algorithm is split into two steps: In the first step we sample, following Algorithm 3.3.1 with $\tilde{M}_N = \tilde{M}_N^\ell$ and $R = R_{\theta_N^\ell}$, M_N^ℓ particles $\{(I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]})\}_{i=1}^{M_N^\ell}$ from $\pi_{\text{aux}}^{\theta_N^\ell}$. Note that the adjustment multiplier weights are kept constant during the iterations, a limitation which is however not necessary. The second step consists in computing the exact solution

$$\theta_N^{\ell+1} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{\tilde{\Omega}^{[\ell]}} \log \left(\frac{\mathrm{d}\mu_{\text{aux}}}{\mathrm{d}\pi_{\text{aux}}^{\theta}}(I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}) \right) \tag{3.5.3}$$

to the problem of minimizing the Monte Carlo approximation of (3.5.2). In the case where the kernels L and R_θ have densities, denoted by l and r_θ, respectively, w.r.t. a common reference measure on $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$, the minimization program (3.5.3) is equivalent to

$$\theta_N^{\ell+1} := \arg\max_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{\tilde{\Omega}^{[\ell]}} \log r_\theta(\xi_{I_i^{[\ell]}}, \tilde{\xi}_i^{[\ell]}) \,. \tag{3.5.4}$$

This algorithm is helpful only in situations where the minimization problem (3.5.3) is sufficiently simple for allowing for closed-form minimization; this happens, for example, if the objective function is a convex combination of concave functions, whose minimum either admits a (simple) closed-form expression or is straightforward to minimize numerically. As mentioned in Section 3.2.1, this is generally the case when the function $r_\theta(\xi, \cdot)$ belongs to an exponential family for any $\xi \in \Xi$.

Since this optimization problem closely ressembles the Monte Carlo EM algorithm, all the implementation details of these algorithms can be readily transposed to that context; see for example Levine and Casella (2001), Eickhoff et al. (2004), and Levine and Fan (2004). Because we use very simple models, convergence occurs, as seen in Section 3.6, within only few iterations. When choosing the successive particle sample sizes $\{\tilde{M}_N^\ell\}_{\ell=1}^L$, we are facing a trade-off between precision of the approximation (3.5.3) of (3.5.2) and computational cost. Numerical evidence typically shows that these sizes may, as high precision is less crucial here than when generating the final population from $\pi_{\text{aux}}^{\theta_N^L}$, be relatively small compared to the final size $\tilde{M}_N$. Besides, it is possible

(and even theoretically recommended) to increase the number of particles with the iteration index, since, heuristically, high accuracy is less required at the first steps. In the current implementation in Section 3.6, we will show that fixing a priori the total number of iterations and using the same number $\tilde{M}_N^\ell = \tilde{M}_N/L$ of particles at each iteration may provide satisfactory results in a first run.

Algorithm 3.5.2 CE-based adaptive APF

Require: $\{(\xi_i, \omega_i)\}_{i=1}^{M_N}$ targets ν.

1: Choose an arbitrary θ_N^0,

2: **for** $\ell = 0, \ldots, L-1$ **do**

3: draw

$$\{I_i^{[\ell]}\}_{i=1}^{\tilde{M}_N^\ell} \sim \mathcal{M}(\tilde{M}_N^\ell, \{\omega_j \psi_j / \sum_{n=1}^{M_N} \omega_n \psi_n\}_{j=1}^{M_N}) \,,$$

4: simulate $\{\tilde{\xi}_i^{[\ell]}\}_{i=1}^{\tilde{M}_N^\ell} \sim \bigotimes_{i=1}^{\tilde{M}_N^\ell} R_{\theta_N^\ell}(\xi_{I_i^{[\ell]}}, \cdot)$,

5: update

$$\tilde{\omega}_i^{[\ell]} \overset{\forall i}{\leftarrow} \Phi_{\theta_N^\ell}(\xi_{I_i^{[\ell]}}, \tilde{\xi}_i^{[\ell]}) \,,$$

6: compute, with available closed-form,

$$\theta_N^{\ell+1} := \arg\min_{\theta \in \Theta} \sum_{i=1}^{\tilde{M}_N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{\tilde{\Omega}^{[\ell]}} \log \left(\frac{\mathrm{d}\mu_{\mathrm{aux}}}{\mathrm{d}\pi_{\mathrm{aux}}^\theta}(I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}) \right) \,,$$

7: **end for**

8: run Algorithm 3.3.1 with $R = R_{\theta_N^L}$.

3.6 Application to state space models

For an illustration of our findings we return to the framework of state space models in Section 3.2.2 and apply the CE-adaptation-based particle method to *filtering* in nonlinear state space models of type

$$\begin{aligned} X_{k+1} &= m(X_k) + \sigma_w(X_k) W_{k+1} \,, \quad k \geq 0 \,, \\ Y_k &= X_k + \sigma_v V_k \,, \quad k \geq 0 \,, \end{aligned} \tag{3.6.1}$$

where the parameter σ_v and the $\mathbb{R}$-valued functions (m, σ_w) are known, and $\{W_k\}_{k=1}^\infty$ and $\{V_k\}_{k=0}^\infty$ are mutually independent sequences of independent standard normal-distributed variables. In this setting, we wish to approximate the filter distributions $\{\phi_{k|k}\}_{k \geq 0}$, defined in Section 3.2.2 as the posterior distributions of X_k given $Y_{0:k}$ (recall that $Y_{0:k} := (Y_0, \ldots, Y_k)$), which in general lack closed-form expressions. For models of this type, the optimal weight and density defined in (3.2.26) and (3.2.27), respectively, can be expressed in closed-form:

$$\Psi_k^*(x) = \mathcal{N}\left(Y_{k+1}; m(x), \sqrt{\sigma_w^2(x) + \sigma_v^2}\right) \,, \tag{3.6.2}$$

where $\mathcal{N}(x; \mu, \sigma) := \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$ and

$$r_k^*(x, x') = \mathcal{N}\left(x'; \tau(x, Y_{k+1}), \eta(x)\right) \,, \tag{3.6.3}$$

with

$$\tau(x, Y_{k+1}) := \frac{\sigma_w^2(x) Y_{k+1} + \sigma_v^2 m(x)}{\sigma_w^2(x) + \sigma_v^2} ,$$
$$\eta^2(x) := \frac{\sigma_w^2(x)\sigma_v^2}{\sigma_w^2(x) + \sigma_v^2} .$$

We may also compute the chi-square optimal adjustment multiplier weight function $\Psi^*_{\chi^2,Q}$ when the prior kernel is used as proposal: at time k,

$$\Psi^*_{\chi^2,Q}(x) \propto \sqrt{\frac{2\sigma_v^2}{2\sigma_w^2(x) + \sigma_v^2}} \exp\left(-\frac{Y_{k+1}^2}{\sigma_v^2} + \frac{m(x)}{2\sigma_w^2(x) + \sigma_v^2}[2Y_{k+1} - m(x)] \right) . \tag{3.6.4}$$

We recall from Proposition 3.4.2 that the optimal adjustment weight function for the KLD is given by $\Psi^*_{KL,Q}(x) = \Psi^*_k(x)$.

In these intentionally chosen simple example we will consider, at each timestep k, adaption over the family

$$\{R_\theta(x, \cdot) := \mathcal{N}(\tau(x, Y_{k+1}), \theta\, \eta(x)) : x \in \mathbb{R}, \theta > 0\} \tag{3.6.5}$$

of proposal kernels. In addition, we keep the adjustment weights constant, that is $\Psi(x) = 1$.

The mode of each proposal kernel is centered at the mode of the optimal kernel, and the variance is proportional to the inverse of the Hessian of the optimal kernel at the mode. Let $r_\theta(x, x') := \mathcal{N}(x'; \tau(x, Y_{k+1}), \theta\, \eta(x))$ denote the density of $R_\theta(x, \cdot)$ w.r.t. the Lebesgue measure. In this setting, at every timestep k, a closed-form expression of the KLD between the target and proposal distributions is available:

$$d_{\mathrm{KL}}(\mu_{\mathrm{aux}} || \pi^\theta_{\mathrm{aux}}) = \sum_{i=1}^{M_N} \frac{\omega_i {\psi^*}_i}{\sum_{j=1}^{M_N} \omega_j {\psi^*}_j} \left[\log\left(\frac{{\psi^*}_i \Omega}{\sum_{j=1}^{M_N} \omega_j {\psi^*}_j} \right) + \log\theta + \frac{1}{2}\left(\frac{1}{\theta^2} - 1 \right) \right] , \tag{3.6.6}$$

where we set ${\psi^*}_i := \Psi^*(\xi_i)$ and $\Omega = \sum_{i=1}^{M_N} \omega_i$.

As we are scaling the optimal standard deviation, it is obvious that

$$\theta^*_N := \arg\min_{\theta > 0} d_{\mathrm{KL}}(\mu_{\mathrm{aux}} || \pi^\theta_{\mathrm{aux}}) = 1 , \tag{3.6.7}$$

which may also be inferred by straightforward derivation of (3.6.6) w.r.t. θ. This provides us with a reference to which the parameter values found by our algorithm can be compared. Note that the instrumental distribution $\pi^{\theta^*_N}_{\mathrm{aux}}$ differs from the target distribution μ_{aux} by the adjustment weights used: recall that every instrumental distribution in the family considered has uniform adjustment weights, $\Psi(x) = 1$, whereas the overall optimal proposal has, since it is equal to the target distribution μ_{aux}, the optimal weights defined in (3.6.2). This entails that

$$d_{\mathrm{KL}}(\mu_{\mathrm{aux}} || \pi^{\theta^*_N}_{\mathrm{aux}}) = \sum_{i=1}^{M_N} \omega_i \frac{{\psi^*}_i}{\sum_{j=1}^{M_N} \omega_j {\psi^*}_j} \log\left(\frac{{\psi^*}_i \Omega}{\sum_{j=1}^{M_N} \omega_j {\psi^*}_j} \right) , \tag{3.6.8}$$

which is zero if all the optimal weights are equal.

The implementation of Algorithm 3.5.2 is straightforward as the optimization program (3.5.4) has the following closed-form solution:

$$\theta^{\ell+1}_N = \left\{ \sum_{i=1}^{M^\ell_N} \frac{\tilde{\omega}_i^{[\ell]}}{\tilde{\Omega}^{[\ell]} \eta^2_{I_i^{[\ell]}}} \left(\tilde{\xi}_i^{[\ell]} - \tau_{I_i^{[\ell]}} \right)^2 \right\}^{1/2} , \tag{3.6.9}$$

where $\tau_i := \tau(\xi_i, Y_{k+1})$ and $\eta_i^2 := \eta^2(\xi_i)$. This is a typical case where the family of proposal kernels allows for efficient minimization. Richer families sharing this property may also be used, but here we are voluntarily willing to keep this toy example as simple as possible.

We will study the following special case of the model (3.6.1):

$$m(x) \equiv 0, \quad \sigma_w(x) = \sqrt{\beta_0 + \beta_1 x^2}\,.$$

This is the classical Gaussian *autoregressive conditional heteroscedasticity* (ARCH) *model* observed in noise (see Bollerslev et al. (1994)). In this case an experiment was conducted where we compared:

(i) a plain nonadaptive particle filter for which $\Psi \equiv 1$, that is, the bootstrap particle filter of Gordon et al. (1993),

(ii) an auxiliary filter based on the prior kernel and chi-square optimal weights $\Psi^*_{\chi^2,Q}$,

(iii) adaptive bootstrap filters with uniform adjustment multiplier weights using numerical minimization of the empirical CSD and

(iv) the empirical KLD (Algorithm 3.5.1),

(v) an adaptive bootstrap filter using direct minimization of $d_{\mathrm{KL}}(\mu_{\mathrm{aux}}||\pi^{\theta}_{\mathrm{aux}})$, see (3.6.7),

(vi) a CE-based adaptive bootstrap filter, and as a reference,

(vi) an optimal auxiliary particle filter, i.e. a filter using the optimal weight and proposal kernel defined in (3.6.2) and (3.6.3), respectively.

This experiment was conducted for the parameter set $(\beta_0, \beta_1, \sigma_v^2) = (1, 0.99, 10)$, yielding (since $\beta_1 < 1$) a geometrically ergodic ARCH(1) model (see Chen and Chen, 2000, Theorem 1); the noise variance σ_v^2 is equal to $1/10$ of the stationary variance, which here is equal to $\sigma_s^2 = \beta_0/(1-\beta_1)$, of the state process.

In order to design a challenging test of the adaptation procedures we set, after having run a hundred burn-in iterations to reach stationarity of the hidden states, the observations to be constantly equal to $Y_k = 6\sigma_s$ for every $k \geq 110$. We expect that the bootstrap filter, having a proposal transition kernel with constant mean $m(x) = 0$, will have a large mean square error (MSE) due a poor number of particles in regions where the likelihood is significant. We aim at illustrating that the adaptive algorithms, whose transition kernels have the same mode as the optimal transition kernel, adjust automatically the variance of the proposals to that of the optimal kernel and reach performances comparable to that of the optimal auxiliary filter.

For these observation records, Figure 3.1 displays MSEs estimates based on 500 filter means. Each filter used $5{,}000$ particles. The reference values used for the MSE estimates were obtained using the optimal auxiliary particle filter with as many as $500{,}000$ particles. This also provided a set from which the initial particles of every filter were drawn, hence allowing for initialisation at the filter distribution a few steps before the outlying observations.

The CE-based filter of algorithm 3.5.2 was implemented in its most simple form, with the inside loop using a constant number of $M_N^{\ell} = N/10 = 500$ particles and only $L = 5$ iterations: a simple prefatory study of the model indicated that the Markov chain $\{\theta_N^{\ell}\}_{l \geq 0}$ stabilised around the value reached in the very first step. We set $\theta_N^0 = 10$ to avoid initialising at the optimal value.

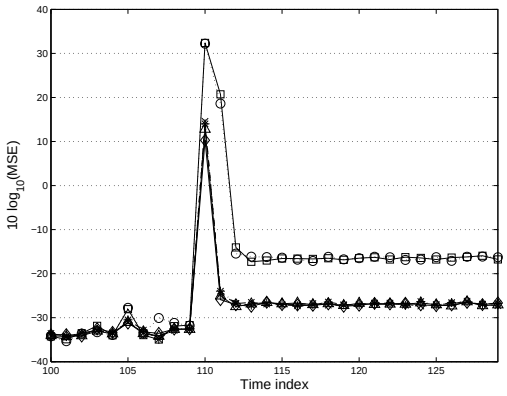
It can be seen in Figure 3.1a that using the CSD optimal weights combined with the prior kernel as proposal does not improve on the plain bootstrap filter, precisely because the observations were chosen in such a way that the prior kernel was helpless. On the contrary, Figures 3.1a and 3.1b show that the adaptive schemes perform

exactly similarly to the optimal filter: they all success in finding the optimal scale of the standard deviation, and using uniform adjustment weights instead of optimal ones does not impact much.

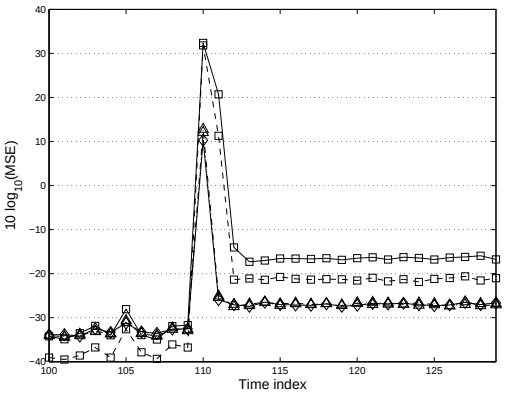
We observe clearly a change of regime, beginning at step 110, corresponding to the outlying constant observations. The adaptive filters recover from the changepoint in one timestep, whereas the bootstrap filter needs several. More important is that the adaptive filters (as well as the optimal one) reduce, in the regime of the outlying observations, the MSE of the bootstrap filter by a factor 10.

Moreover, for a comparison with fixed simulation budget, we ran a bootstrap filter with $3N = 15{,}000$ particles This corresponds to the same simulation budget as the CE-based adaptive scheme with N particles, which is, in this setting, the fastest of our adaptive algorithms. In our setting, the CE-based filter is measured to expand the plain bootstrap runtime by a factor 3, although a basic study of algorithmic complexity shows that this factor should be closer to $\sum_{\ell=1}^{L} M_N^\ell / N = 1.5$—the difference rises from Matlab benefitting from the vectorisation of the plain bootstrap filter, not from the iterative nature of the CE.

The conclusion drawn from Figure 3.1b is that for an equal runtime, the adaptive filter outperforms, by a factor 3.5, the bootstrap filter using even three times more particles.



(a) Auxiliary filter based on chi-square optimal weights $\Psi^*_{\chi^2,Q}$ and prior kernel K ($\circ$), adaptive filters minimizing the empirical KLD ($*$) and CSD ($\times$), and reference filters listed below.



(b) CE-based adaption ($\triangle$, dash-dotted line), bootstrap filter with $3N$ particles ($\square$, dashed line), and reference filters listed below.

Figure 3.1: Plot of MSE performances (on log-scale) on the ARCH model with $(\beta_0, \beta_1, \sigma_v^2) = (1, 0.99, 10)$. Reference filters common to both plots are: the bootstrap filter ($\square$, continuous line), the optimal filter with weights Ψ^* and proposal kernel density r^* ($\diamond$), and a bootstrap filter using a proposal with parameter θ_N^* minimizing the current KLD ($\triangle$, continuous line). The MSE values are computed using $N = 5{,}000$ particles—except for the reference bootstrap using $3N$ particles ($\square$, dashed line)—and $1{,}000$ runs of each algorithm.

Acknowledgements

The authors are grateful to Prof. Paul Fearnhead for encouragements and useful recommandations, and to the anonymous reviewers for insightful comments and suggestions that improved the presentation of the paper.

CHAPTER 4

Adaptation of the adjustment weights by pilot exploration and refueling

Contents

This chapter corresponds to an article which is to be submitted under the (tentative) name *On the use of the coefficient of variation criterion for sequential Monte Carlo adaptation: a statistical perspective*, by J. Cornebise, E. Moulines, J. Olsson, 2009. As we will develop in-depth convergence results, for sake of rigor and unambiguity, we figure the index N for all the random-variables that constitute a triangular array (most noticeably the particles, their weights, and their adjustment weights).

4.1 Introduction

Since the *bootstrap particle filter* was introduced by (Gordon et al., 1993), significant research activity has been devoted to the study of automatic adaptation of the key parameters—such as the particle sample size or the proposal kernel—of *sequential Monte Carlo* (SMC) methods. Such adaptation strategies have a long history, from Gordon et al. (1993) themselves, adjusting heuristically the proposal kernel via so-called *prior editing*, to Cornebise et al. (2008) (see Chapter 3 of the present dissertation) who provide a unified function-free risk-theoretic framework for evaluating the expected quality of the SMC output. As mentioned, the study of SMC adaptation follows historically two main directions: adaptation of the number of particles (Pitt and Shephard, 1999; Doucet and Andrieu, 2001) or of the proposal kernel (Fox, 2003; Legland and

Oudjane, 2006; Soto, 2005). In addition, the problem of designing adaptively the resampling (selection) schedule (i.e. to determine online whether selection should occur or not) has been broadly considered. The first steps in this direction was taken by Liu and Chen (1995), who pointed out that resampling systematically the particles at every time step is suboptimal; instead, the decision to select the particles should be based on criteria describing the quality of the particle sample in terms of weight degeneracy. Rigorous theoretical interpretations of the most frequently used such criteria, i.e., the *coefficient of variation* (CV), the *efficient sample size* (ESS), and the *Shannon entropy* of the importance weights, did not appear until the recent paper by Cornebise et al. (2008) (Chapter 3 of this dissertation) in which it is shown that all these criteria are closely related to the *Kullback-Leibler divergence* (KLD) or the *chi-square distance* (CSD) between well defined instrumental and target distributions associated with a sequence of importance sampling problems solved by the *auxiliary particle filter* APF of Pitt and Shephard (1999) at the different iterations. The moral is simple: large skewness of the particle weights indicates a large distance between the instrumental distribution and the target. This is entirely analogous to what holds for standard importance sampling. The work in question however did not explicitly generalize its results to cases where selection is executed at random timesteps, which, since the criteria in question are usually used for *activating* selection, is a clear limitation. We thus address this issue here and provide a similar theoretical superstructure in the case of adaptive selection.

An additional SMC parameter was introduced by Pitt and Shephard (1999) within the framework of the (APF). In the APF the particle weights are modified by nonnegative multiplicative factors, referred to as *adjustment multiplier weights* (AMWs) (or, following the terminology of the original article, *first stage weights*). The main motivation for introducing the additional degree of freedom imposed by the adjustment weights was to robustify the SMC scheme to outliers in observed state space model data by holding resampling until the subsequent observation becomes available and then incorporating this information into the selection procedure. In this way the survival rate is increased for particles being *expected* to land up in state space regions of high posterior probability (as measured by the likelihood) at the next move. With the exception of Douc et al. (2008), who identified optimal AMWs minimizing the asymptotic variance of the Monte Carlo estimates for a given proposal kernel, very few works deal with improving the generic weights proposed ad hoc by Pitt and Shephard (1999). The purpose of this article is partially to shed new light on these adjustment weights and to subject them to adaptation. In Cornebise et al. (2008) it is showed that no adaptation of the particle filter instrumental distributions can be fully achieved without taking into account these AMWs; this is, as we will see in the forthcoming examples of Chapter 5 (see e.g. Sections 5.5.2 and 5.5.4), specially obvious for state space models with very informative observations or censorship.

The contribution of this article is threefold, since we

1. extend existing convergence results (Douc et al., 2008; Johansen and Doucet, 2008) on the APF by allowing for AMWs which are not necessarily deterministic functions of the ancestor particles. More specifically, we show that each sample in the sequence returned by the APF is, under weak assumptions, *consistent* as well as *asymptotically normal* (these convergence modi being adopted from Douc and Moulines, 2008) also when the AMWs are generated according to Markovian transitions from the ancestors space to the non-negative real numbers half-line. In particular, the results obtained encompasses the *pilot-exploration*-based APF (the so-called SISPER scheme) proposed by Zhang and Liu (2002);

2. extend existing results (Cornebise et al., 2008) characterizing the importance weight CV as a consistent CSD estimate in the case where selection is executed systematically at all time steps, to the more relevant case where selection is performed adaptively. Lead by these theoretical findings, we conclude that the standard way of using the CV for triggering the selection procedure is, since the instrumental distributions of the SMC scheme are not adjusted to the targets until *after* the sample has degenerated in this case, importantly suboptimal due to the lack of foresight;

3. use the results of the two items above for motivating and constructing a novel algorithm, referred to as *sequential importance sampling with adaptive refueling* (SISAR), which is composed of two parts: in a first pilot-exploration step, adopted from the SISPER algorithm of Zhang and Liu (2002), a swarm of particles is sent out to estimate, via the CV, the CSD between the instrumental and target distributions at the *next* time step; in the second *refueling* step the current particles are, if the estimated CSD lies above a pre-specified threshold κ, selected multinomially with respect to the weights of the pilot sample in order to fit the instrumental distribution to the target *before* degeneracy has occurred. In addition, in the refueling step, the number of particles is increased by a factor which is an increasing function φ of the CV.

The chapter is organized as follows. In Section 4.2 we discuss how SMC algorithms are used for approximating sequences of probability measures generated recursively by nonlinear Markovian transitions. The concepts of mutation and selection are introduced, leading to a non-standard description of the APF in which we allow for randomly varying AMWs. The convergence of the random weight APF is stated in Theorems 4.2.1 (consistency) and 4.2.2 (asymptotic normality). The analysis is made under the assumption that selection is carried through systematically at all time steps, an assumption which is lightened in Section 4.3 in which we put the random weight APF into the context of standard CV-triggered adaptation. The convergence of the resulting scheme, referred to as *sequential importance sampling with adaptive selection* (SISAS), is analyzed rigorously (Theorem 4.3.2). In addition, we extend Theorems 4.1–2 in (Cornebise et al., 2008) and show (see Theorem 4.3.2 as well) that the CV of the particle weights and the CSD between specified instrumental and target mixture distributions coincide asymptotically at any time step. The last part, Section 4.3.3, is devoted to the SISAR algorithm and the convergence of the scheme is stated in Corollary 4.3.1.

4.2 The SMC framework

4.2.1 Notation

We preface the precise description of our main SMC algorithm with some measure-theoretic notation. Let $\mathbb{B}(\Xi)$ and $\mathcal{P}(\Xi)$ denote the spaces of measurable functions and probability measures, respectively, on some state space $(\Xi, \mathcal{B}(\Xi))$. For any $\mu \in \mathcal{P}(\Xi)$ and $f \in \mathbb{B}(\Xi)$ satisfying $\int_\Xi |f(\xi)|\,\mu(\mathrm{d}\xi) < \infty$ we let $\mu(f)$ denote $\int_\Xi f(\xi)\,\mu(\mathrm{d}\xi)$. A transition kernel K from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ induces two operations: the first tranforms a function $f \in \mathbb{B}(\Xi \times \tilde{\Xi})$ such that $\int_{\tilde{\Xi}} |f(\xi, \tilde{\xi})|\,K(\xi, \mathrm{d}\tilde{\xi}) < \infty$ into the function $\xi \mapsto K(\xi, f) := \int_{\tilde{\Xi}} f(\xi, \tilde{\xi})\,K(\xi, \mathrm{d}\tilde{\xi})$ in $\mathbb{B}(\Xi)$; the other transforms a measure $\mu \in \mathcal{P}(\Xi)$ into another measure $A \mapsto \mu K(A) := \int_\Xi K(\xi, A)\,\mu(\mathrm{d}\xi)$ in $\mathcal{P}(\tilde{\Xi})$. The *product* of K and another transition kernel T from $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ to $(\bar{\Xi}, \mathcal{B}(\bar{\Xi}))$ is the transition kernel from

$(\Xi, \mathcal{B}(\Xi))$ to $(\bar{\Xi}, \mathcal{B}(\bar{\Xi}))$ given by, for $\xi \in \Xi$ and $A \in \mathcal{B}(\bar{\Xi})$,

$$KT(\xi, A) := \int_{\tilde{\Xi}} K(\xi, \mathrm{d}\tilde{\xi})\, T(\tilde{\xi}, A) \ .$$

The *outer product* $K \otimes T$ of K and T is the transition kernel from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi} \times \bar{\Xi}, \mathcal{B}(\tilde{\Xi} \times \bar{\Xi}))$ given by, for $\xi \in \Xi$ and $A \in \mathcal{B}(\tilde{\Xi} \times \bar{\Xi})$,

$$K \otimes T(\xi, A) := \iint_{\tilde{\Xi} \times \bar{\Xi}} \mathbb{1}_A(\tilde{\xi}, \bar{\xi})\, K(\xi, \mathrm{d}\tilde{\xi})\, T(\tilde{\xi}, \mathrm{d}\bar{\xi}) \ , \tag{4.2.1}$$

and for a sequence $\{K_\ell\}_{\ell=m}^n$ of kernels, $\bigotimes_{\ell=m}^n K_\ell := K_m \otimes K_{m+1} \otimes \cdots \otimes K_n$ is defined by using recursively (4.2.1).

4.2.2 SMC approximation of Feynman-Kac distribution flows

Let $\{(\Xi_n, \mathcal{B}(\Xi_n))\}_{n=0}^\infty$ and $\{L_n\}_{n=0}^\infty$ be sequences of general state spaces and finite transition kernels, respectively, where each L_n describes transitions from $(\Xi_n, \mathcal{B}(\Xi_n))$ to $(\Xi_{n+1}, \mathcal{B}(\Xi_{n+1}))$. In this paper we deal with the problem of approximating efficiently the *Feynman-Kac-flow* $\{\mu_n\}_{n=0}^\infty$ of distributions generated recursively according to

$$\mu_{n+1}(A) := \frac{\mu_n L_n(A)}{\mu_n L_n(\Xi_{n+1})} \ , \quad A \in \mathcal{B}(\Xi_{n+1}) \ , \tag{4.2.2}$$

by a sequence of weighted samples. The recursion is initialized by a measure $\mu_0 \in \mathcal{P}(\Xi_0)$. Though n is not necessarily a temporal index, we will often refer to n as "time". For any $(m, n) \in \mathbb{N}^* \times \mathbb{N}^*$ with $m \leq n$ one easily shows that

$$\mu_n(A) = \frac{\mu_m L_m \cdots L_{n-1}(A)}{\mu_m L_m \cdots L_{n-1}(\Xi_n)} \ , \quad A \in \mathcal{B}(\Xi_n) \ ,$$

under the convention that $L_\ell \cdots L_p := \mathrm{Id}$ if $\ell > p$. Feynman-Kac flows are widely used in many scientific disciplines and a survey of examples from, *e.g.*, financial economics, signal processing, biology, and statistical physics is given by Del Moral (2004, Chapter 1).

In most applications, nonlinear/non-Gaussian model components make closed-form solutions to the recursion (4.2.2) intractable, and the aim of this chapter is thus to develop adaptive SMC methods approximating the distribution flow under consideration. In the *sequential importance sampling* (SIS) approach proposed by Handschin and Mayne (1969) a *weighted sample* approximating μ_n is produced by drawing particle trajectories $\{\xi_{N,i}^{(0:n)}\}_{i=1}^{M_N}$ from an instrumental distribution $\rho_0 \bigotimes_{\ell=0}^{n-1} R_\ell$ in $\mathcal{P}(\Xi_{0:n})$, where each kernel $R_\ell(\xi, \cdot)$ dominates $L_\ell(\xi, \cdot)$ for all $\xi \in \Xi_\ell$. Every particle $\xi_{N,i}^{(0:n)}$ is associated with a nonnegative weight $\omega_{N,i}^{(n)} := \mathrm{d}\mu_0/\mathrm{d}\rho_0(\xi_{N,i}^{(0)}) \Phi_{0,n-1}(\xi_{N,i}^{(0:n)})$, where

$$\Phi_{k,m}(\xi_{k:m+1}) := \prod_{\ell=k}^{m} \frac{\mathrm{d}R_\ell(\xi_\ell, \cdot)}{\mathrm{d}L_\ell(\xi_\ell, \cdot)}(\xi_{\ell+1}) \ , \quad \xi_{k:m+1} \in \Xi_{k:m+1} \ ,$$

which implies that $\omega_{N,i}^{(n)} \propto \mathrm{d}\mu_{0:n}/\mathrm{d}[\rho_0 \bigotimes_{\ell=0}^{n-1} R_\ell](\xi_{N,i}^{(0:n)})$, with

$$\mu_{k:m}(A) := \frac{\mu_k \bigotimes_{\ell=k}^{m-1} L_\ell(A)}{\mu_k L_k \cdots L_{m-1}(\Xi_m)} \ , \quad A \in \mathcal{B}(\Xi_{k:m}) \ .$$

Hence, the self-normalized quantity $\sum_{i=1}^{M_N} \omega_{N,i}^{(n)} f(\xi_{N,i}^{(0:n)})/\Omega_N^{(n)}$, with $\Omega_N^{(n)} := \sum_{\ell=1}^{M_N} \omega_{N,\ell}^{(n)}$, can, for large M_N's, be taken as an estimate of $\mu_{0:n}(f)$ for any f belonging to $\mathsf{L}^1(\mu_{0:n}, \mathbf{\Xi}_{0:n})$. Moreover, since μ_n is the restriction of $\mu_{0:n}$ to $\mathcal{B}(\mathbf{\Xi}_n)$, the marginal particles $\{\xi_{N,i}^{(n)}\}_{i=1}^{M_N}$ can be used for estimating μ_n in the sense that $\sum_{i=1}^{M_N} \omega_{N,i}^{(n)} f'(\xi_{N,i}^{(n)})/\Omega_N^{(n)}$ approximates $\mu_n(f')$ for all f' in $\mathsf{L}^1(\mu_n, \mathbf{\Xi}_n)$. A key observation in this context is that the particular choice of instrumental distribution above allows for a completely sequential implementation of the procedure. More specifically, given particles and weights at time n, a weighted sample approximating $\mu_{0:n+1}$ is obtained by simply extending each particle path $\xi_{N,i}^{(0:n)}$ with an additional component $\xi_{N,i}^{(n+1)}$ simulated according to $R_n(\xi_{N,i}^{(n)}, \cdot)$ and assigning this extended particle the importance weight $\omega_{N,i}^{(n+1)} := \omega_{N,i}^{(n)} \Phi_n(\xi_{N,i}^{(n:n+1)})$. This operation is typically referred to as *mutation* and is described generically in Algorithm 4.2.1 below, where ν and $\mu(\cdot) := \nu L(\cdot)/\nu L(\tilde{\mathbf{\Xi}})$, L being a finite transition kernel, are probability measures on general state spaces $(\mathbf{\Xi}, \mathcal{B}(\mathbf{\Xi}))$ and $(\tilde{\mathbf{\Xi}}, \mathcal{B}(\tilde{\mathbf{\Xi}}))$, respectively, R is a Markovian instrumental kernel (dominating L), and $\Phi(\xi, \tilde{\xi}) := \mathrm{d}L(\xi, \cdot)/\mathrm{d}R(\xi, \cdot)(\tilde{\xi})$ for $(\xi, \tilde{\xi}) \in \mathbf{\Xi} \times \tilde{\mathbf{\Xi}}$.

Algorithm 4.2.1 Mutates a given weighted sample

procedure MUTATION($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, R, Φ)
  for $i \leftarrow 1$ **to** M_N **do**
    simulate, conditionally independently, $\tilde{\xi}_{N,i} \sim R(\xi_{N,i}, \cdot)$;
    $\tilde{\omega}_{N,i} \leftarrow \omega_{N,i} \Phi(\xi_{N,i}, \tilde{\xi}_{N,i})$;
  end for
  return $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{M_N}$
end procedure

Expressed in terms of Algorithm 4.2.1, the weighted sample $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$ is produced as follows: as initialization, draw $\{\xi_{N,i}^{(0)}\}_{i=1}^{M_N}$ from the product measure $\rho_0^{\otimes M_N}$ and set $\omega_{N,0}^{(i)} := \mathrm{d}\mu_0/\mathrm{d}\rho_0(\xi_{N,i}^{(0)})$ for all i. Next, execute Algorithm 4.2.2.

Algorithm 4.2.2 Sequential importance sampling

procedure SIS($\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$, $\{(R_\ell, \Phi_\ell)\}_{\ell=0}^{n-1}$)
  for $\ell \leftarrow 0$ **to** $n-1$ **do**
    $\{(\xi_{N,i}^{(\ell+1)}, \omega_{N,i}^{(\ell+1)})\}_{i=1}^{M_N} \leftarrow$ MUTATION($\{(\xi_{N,i}^{(\ell)}, \omega_{N,i}^{(\ell)})\}_{i=1}^{M_N}, R_\ell, \Phi_\ell$)
  end for
  return $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$
end procedure

In the mutation operation, using the proposal kernel, the particles are scattered randomly in the state space and assigned importance weights reflecting the relevance of the particles as measured by the likelihood ratio Φ. However, piling blindly consecutive mutation steps as in Algorithm 4.2.2 results almost without exception in *weight degeneracy* as n increases. In a situation of a degenerated particle sample, the particle approximation becomes statistically and computationally inefficient, since only a few particles contribute significantly to the Monte Carlo estimation and most computational effort is wasted on updating non-contributing particles and weights. To cope with the problem,

Gordon et al. (1993) combined the mutation operation above with a *selection operation* in which particles having large/small importance weights are duplicated/eliminated by drawing, with replacement, the particles multinomially with respect to the normalized weights.

Algorithm 4.2.3 Selection by multinomial resampling

procedure SELECTION($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, $\tilde{M}_N$)
  for $i \leftarrow 1$ **to** $\tilde{M}_N$ **do**
    simulate $I_{N,i} \sim \mathrm{Mult}(\{\omega_{N,i} / \sum_{\ell=1}^{M_N} \omega_{N,\ell}\}_{i=1}^{M_N})$;
    set $\hat{\xi}_{N,i} \leftarrow \xi_{N,I_{N,i}}$;
  end for
  return $\{(\hat{\xi}_{N,i}, 1), I_{N,i}\}_{i=1}^{\tilde{M}_N}$
end procedure

The selection step is unbiased (in the sense that the number of expected offspring of a certain particle is proportional to its weight) and does not change the target distribution of the particle swarm. In the *auxiliary particle filter* (APF) proposed by Pitt and Shephard (1999) the selection step is prefaced by a first stage *weighting operation* in which the particle weights are multiplied by so-called *adjustment multiplier weights* (AMWs) $\{\psi_{N,i}\}_{i=1}^{M_N}$ (alternatively termed *first stage weights*). Such a weighting operation makes it possible to amplify the weight of particles that are expected to be associated large likelihood ratios Φ (and thus large importance weights) at the subsequent mutation step. To compensate for the first stage weight adjustment, the particles have to be reweighted at an additional weighting step succeeding the mutation operation. In the framework of state space models the adjustment weights incorporate the subsequent observation. The APF with random AMWs is summarized in Algorithm 4.2.4 in which we assume that each weight $\psi_{N,i}$ is a random draw from $\Psi(\xi_{N,i}, \cdot)$, where Ψ is a transition kernel from $\boldsymbol{\Xi}$ to $\mathbb{R}^+$ (hence such that for any $\xi \in \boldsymbol{\Xi}$, $\Psi(\xi, \mathbb{R}^+) = 1$). This yields a significantly more general framework than in most related works (such as Pitt and Shephard, 1999; Douc et al., 2008), since it is standard in the literature to assume that the AMWs are determined by a deterministic function of the ancestor particles, corresponding to $\Psi(\xi, \cdot) = \delta_{h(\xi)}(\cdot)$ for some nonnegative function $h : \boldsymbol{\Xi} \to \mathbb{R}^+$.

Algorithm 4.2.4 One step of the APF with random AMWs

procedure APF($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, R, Φ, Ψ, $\tilde{M}_N$)
  for $i \leftarrow 1$ **to** M_N **do**
    draw, conditionally independently, $\psi_{N,i} \sim \Psi(\xi_{N,i}, \cdot)$;
  end for
  $\{(\hat{\xi}_{N,i}, 1), I_{N,i}\}_{i=1}^{\tilde{M}_N} \leftarrow$SELECTION($\{(\xi_{N,i}, \omega_{N,i}\psi_{N,i})\}_{i=1}^{M_N}, \tilde{M}_N$);
  $\{(\tilde{\xi}_{N,i}, \check{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N} \leftarrow$MUTATION($\{(\hat{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}, R, \Phi$);
  for $i \leftarrow 1$ **to** $\tilde{M}_N$ **do**
    $\tilde{\omega}_{N,i} \leftarrow \check{\omega}_{N,i}\psi_{N,I_{N,i}}^{-1}$;
  end for
  return $\{(\xi_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$
end procedure

From the scheme above it is clear that the choice $\Psi(\xi, \cdot) = \delta_{L(\xi, \tilde{\Xi})}(\cdot)$ results in perfectly uniform second stage weights and consequently a total elimination of the weight degeneracy. In this case the APF is referred to as *fully adapted*. Moreover, Cornebise et al. (2008) showed that this particular first stage weight type minimizes the chi-square divergence between the inherent instrumental and target distributions of the APF when selection is performed systematically at all timesteps. This observation is of key importance also for the development of the present chapter, and we will return to this matter later. Unfortunately, full adaption is achievable only in a few specific models and we are in general referred to (often computationally expensive) approximation-based approaches. On the other hand, it is possible to estimate the optimal optimal weights without bias and with arbitrary accuracy by means of Monte Carlo. The *pilot exploration* approach proposed by Zhang and Liu (2002) is based on the identity

$$\mathbb{E}\left[\Phi(\xi_{N,i}, \tilde{\xi}_{N,i}) \middle| \xi_{N,i}\right] = L(\xi_{N,i}, \tilde{\Xi}) \,, \tag{4.2.3}$$

for all i, where $\{\xi_{N,i}\}_{i=1}^{M_N}$ and $\{\tilde{\xi}_{N,i}\}_{i=1}^{M_N}$ denote, respectively, the input and output particles of Algorithm 4.2.1. Thus, for a given ancestor particle $\xi_{N,i}$, a Monte Carlo estimate of the optimal AMW function $L(\xi_{N,i}, \tilde{\Xi})$ can be computed by drawing, say, α conditionally independent *pilot* particles $\{\bar{\xi}_{N,i}^{[\ell]}\}_{\ell=1}^{\alpha}$ from $R(\xi_{N,i}, \cdot)$ and approximating $L(\xi_{N,i}, \tilde{\Xi})$ by $\alpha^{-1}\sum_{\ell=1}^{\alpha}\Phi(\xi_{N,i}, \bar{\xi}_{N,i}^{[\ell]})$. A crude but computationally efficient estimate is obtained by letting $\alpha = 1$. For this method $\Psi(\xi, \cdot)$ is the convolution of α measures of form $R(\xi, \Phi^{\leftarrow}(\xi, \alpha \cdot A))$, where $\Phi^{\leftarrow}(\xi, S) \in \mathcal{B}(\tilde{\Xi})$ denotes the inverse image of $S \in \mathcal{B}(\mathbb{R}^+)$ under $\Phi(\xi, \cdot)$ and $\alpha \cdot A = \{a\alpha; a \in A\}$. Consequently, a corresponding approximation of $\omega_{N,i}L(\xi_{N,i}, \tilde{\Xi})$ is given by $\alpha^{-1}\sum_{\ell=1}^{\alpha}\bar{\omega}_{N,i}^{(\ell)}$, i.e.by simply averaging over the pilot weights $\bar{\omega}_{N,i}^{(\ell)} = \omega_{N,i}\Phi(\xi_{N,i}, \bar{\xi}_{N,i}^{[\ell]})$. The technique in question shows clearly the importance of allowing for random first stage weights.

4.2.3 Convergence of the random first stage weight APF

We end the current section by stating results describing the convergence of Algorithm 4.2.4. We will consider convergence in the following probabilistic senses.

Definition 4.2.1 (Consistency)**.** A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be *consistent* for the probability measure μ and the set C if, as $N \to \infty$,

$$\Omega_N^{-1}\sum_{i=1}^{M_N}\omega_{N,i}f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \mu(f) \,, \quad f \in \mathsf{C} \,,$$
$$\Omega_N^{-1}\max_{1\le i\le M_N}\omega_{N,i} \xrightarrow{\mathbb{P}} 0 \,.$$

Definition 4.2.2 (Asymptotic normality)**.** A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be *asymptotically normal* (AN) for $(\mu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\}_{N=1}^{\infty})$ if, as $N \to \infty$,

$$a_N\Omega_N^{-1}\sum_{i=1}^{M_N}\omega_{N,i}\{f(\xi_{N,i}) - \mu(f)\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(f)) \,, \quad f \in \mathsf{A} \,,$$
$$a_N^2\Omega_N^{-2}\sum_{i=1}^{M_N}\omega_{N,i}^2 f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \gamma(f) \,, \quad f \in \mathsf{W} \,,$$
$$a_N\Omega_N^{-1}\max_{1\le i\le M_N}\omega_{N,i} \xrightarrow{\mathbb{P}} 0 \,.$$

Now, impose the following assumptions.

(A8) *The sample* $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ *is consistent for* (ν, C)*, where* $\mathsf{C} \subseteq \mathsf{L}^1(\nu, \Xi)$.

(A9) *The sample* $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ *is asymptotically normal for* $(\mu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \alpha, \{a_N\}_{N=1}^{\infty})$*, where* $\mathsf{A} \subseteq \mathsf{L}^1(\nu, \Xi)$.

(A10) *There exists a constant* $|\Phi|_\infty$ *such that, for any* $(\xi, \tilde{\xi}) \in \Xi \times \tilde{\Xi}$, $\Phi(\xi, \tilde{\xi}) \leq |\Phi|_\infty$. *Moreover,* $\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+})$ *and* $\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})$ *belongs to* C.

Here $\mathbf{1}_{\mathbb{R}^+}$ and $\mathbf{1}_{\mathbb{R}^+}^{-1}$ denote, respectively, the identity and inverted identity mappings on $\mathbb{R}^+$, i.e., $\mathbf{1}_{\mathbb{R}^+}(x) = x$ and $\mathbf{1}_{\mathbb{R}^+}^{-1}(x) = 1/x$, for $x \in \mathbb{R}^+$. Under **(A8)** and **(A9)**, define

$$\begin{aligned} \tilde{\mathsf{C}} &:= \left\{ f \in \mathsf{L}^1(\mu, \tilde{\Xi}) : L(\cdot, |f|) \in \mathsf{C} \right\}, \\ \tilde{\mathsf{A}} &:= \left\{ f \in \mathsf{L}^1(\mu, \tilde{\Xi}) : L(\cdot, f) \in \mathsf{A} \cap \mathsf{C}, R(\cdot, \Phi^2 f^2)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C} \right\}, \\ \tilde{\mathsf{W}} &:= \left\{ f : R(\cdot, \Phi^2 |f|)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C}, R(\cdot, \Phi^2 |f|) \in \mathsf{C} \right\}. \end{aligned} \tag{4.2.4}$$

We now have the following results.

Theorem 4.2.1 (Consistency of Algorithm 4.2.4)**.** *Let* $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, Φ, *and* Ψ *satisfy Assumptions* **(A8)** *and* **(A10)**. *Moreover, suppose that* $L(\cdot, \tilde{\Xi})$ *belongs to* C. *Then the set* $\tilde{\mathsf{C}}$ *defined in* (4.2.4) *is proper and weighted particle sample* $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ *obtained in Algorithm 4.2.4 consistent for* $(\mu, \tilde{\mathsf{C}})$.

We preface the proof of Theorem 4.2.1 by a lemma.

Lemma 4.2.1. *Assume* **(A8)** *and let* $h \in \{h' \in \mathbb{B}(\Xi \times \mathbb{R}^+) : \Psi(\cdot, |h'|) \in \mathsf{C} \cap \mathsf{L}^1(\Xi, \nu)\}$. *Then, as* $N \to \infty$,

$$\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} h(\xi_{N,i}, \psi_{N,i}) \xrightarrow{\mathbb{P}} \nu\Psi(h).$$

Proof. Define $U_{N,i} := \omega_{N,i}\Omega_N^{-1} h(\xi_{N,i}, \psi_{N,i})$ and $\mathcal{F}_{N,i} := \sigma(\{(\xi_{N,\ell}, \omega_{N,\ell})\}_{\ell=1}^{M_N}) \vee \sigma(\{U_{N,\ell}\}_{\ell=1}^{i})$. Now, since, as $\Psi(\cdot, h) \in \mathsf{C}$,

$$\sum_{i=1}^{M_N} \mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_{N,i-1}\right] = \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i}\Psi(\xi_{N,i}, h) \xrightarrow{\mathbb{P}} \nu\Psi(h), \tag{4.2.5}$$

it is enough to establish the two conditions of Theorem 11 in (Douc and Moulines, 2008). The first condition follows trivially since the sequence $\{\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i}\Psi(\xi_{N,i}, |h|)\}_{N=1}^{\infty}$ converges (just replace h in (4.2.6) by $|h|$). Thus, we take $\epsilon > 0$ and turn to the second property. Fixing a constant $C > 0$ yields the bound

$$\begin{aligned} \sum_{i=1}^{M_N} \mathbb{E}\left[|U_{N,i}| \mathbb{1}_{\{|U_{N,i}| \geq \epsilon\}} \middle| \mathcal{F}_{N,i-1} \right] \leq \mathbb{1}_{\{\Omega_N^{-1} \max_\ell \omega_{N,\ell} \geq \epsilon C^{-1}\}} \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i}\Psi(\xi_{N,i}, |h|) \\ + \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i}\Psi(\xi_{N,i}, |h| \mathbb{1}_{\{|h| \geq C\}}), \end{aligned}$$

where the right hand side tends, as $\Psi(\cdot, |h| \mathbb{1}_{\{|h| \geq C\}}) \in \mathsf{C}$ since C is proper, to $\nu\Psi(|h| \mathbb{1}_{\{|h| \geq C\}})$ in probability as $N \to \infty$. However, since this limit can, using the dominated convergence theorem, be made arbitrarily small by increasing C, we conclude that the left hand side tends to zero in probability. This completes the proof. □

Proof of Theorem 4.2.1. Properness of $\tilde{\mathsf{C}}$ is checked straightforwardly. Thus, pick $f \in \tilde{\mathsf{C}}$; without loss of generality it is (by Slutsky's theorem) enough to establish the limit, as $N \to \infty$,

$$\tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \frac{\nu L(f)}{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})} \,. \tag{4.2.6}$$

Thus, define, for $i \in \{1, \ldots, \tilde{M}_N\}$, the random variables and σ-algebras $U_{N,i} := \tilde{M}_N^{-1} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i})$ and $\mathcal{F}_{N,i} := \sigma(\{(\xi_{N,\ell}, \omega_{N,\ell}, \psi_{N,\ell})\}_{\ell=1}^{M_N}) \vee \sigma(\{U_{N,\ell}\}_{\ell=1}^{i})$, respectively. Then, for any $i \in \{1, \ldots, \tilde{M}_N\}$,

$$\mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_{N,i-1}\right] = \tilde{M}_N^{-1} \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \psi_{N,\ell} \right)^{-1} \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, f) \,,$$

from which we conclude that, using Lemma 4.2.1 together with **(A8)**,

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_{N,i-1}\right] \xrightarrow{\mathbb{P}} \frac{\nu L(f)}{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})} \,.$$

Thus, we establish the two conditions of Theorem 11 in (Douc and Moulines, 2008). Since, repeating the arguments above, $\{\sum_{i=1}^{M_N} \mathbb{E}[|U_{N,i}||\mathcal{F}_{N,i-1}]\}_{N=1}^{\infty}$ converges (to $\nu L(|f|)$ $/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$) and is thus tight, we set focus on verifying the second condition for a given $\epsilon > 0$. Hence, take a constant $C > 0$ and make the decomposition

$$\begin{aligned} \sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[|U_{N,i}| \mathbb{1}_{\{|U_{N,i}| \geq \epsilon\}} \middle| \mathcal{F}_{N,i-1} \right] \leq \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \psi_{N,\ell} \right)^{-1} \\ \times \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, |f| \mathbb{1}_{\{\Phi|f| \geq C\}}) + \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\tilde{M}_N \psi_{N,\ell} \leq C\epsilon^{-1}\}} L(\xi_{N,\ell}, |f|) \right) \,. \end{aligned} \tag{4.2.7}$$

To establish the limit

$$\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\tilde{M}_N \psi_{N,\ell} \leq C\epsilon^{-1}\}} L(\xi_{N,\ell}, |f|) \xrightarrow{\mathbb{P}} 0 \,, \tag{4.2.8}$$

pick $\delta > 0$ and bound, for all N such that $\tilde{M}_N \geq C(\epsilon\delta)^{-1}$, the quantity of interest by

$$\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\tilde{M}_N \psi_{N,\ell} \leq C\epsilon^{-1}\}} L(\xi_{N,\ell}, |f|) \leq \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell} \leq \delta\}} L(\xi_{N,\ell}, |f|) \,.$$

Now, since $\int_{\psi \leq \delta} \Psi(\cdot, \mathrm{d}\psi) L(\cdot, |f|) \leq L(\cdot, |f|) \in \mathsf{C}$ and C is proper, we obtain, by applying Lemma 4.2.1,

$$\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell} \leq \delta\}} L(\xi_{N,\ell}, |f|) \xrightarrow{\mathbb{P}} \int_{\Xi} \int_{\psi \leq \delta} \Psi(\xi, \mathrm{d}\psi) L(\xi, |f|) \, \nu(\mathrm{d}\xi) \,. \tag{4.2.9}$$

However, the integral on the right hand side of (4.2.9) can, by the dominated convergence theorem, be made arbitrarily small by decreasing δ, which establishes (4.2.8).

Finally, since $L(\cdot, |f|\mathbb{1}_{\{\Phi|f|\geq C\}}) \leq L(\cdot,|f|) \in \mathsf{C}$ and C is proper, we conclude, under (**A8**), that

$$\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} L(\xi_{N,\ell}, |f|\mathbb{1}_{\{\Phi|f|\geq C\}}) \xrightarrow{\mathbb{P}} \nu L(|f|\mathbb{1}_{\{\Phi|f|\geq C\}}) \,. \tag{4.2.10}$$

However, the limit quantity in (4.2.10) can, by increasing C and applying again the dominated convergence theorem, be made arbitrarily small, which shows that

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[|U_{N,i}|\mathbb{1}_{\{|U_{N,i}|\geq\epsilon\}} \middle| \mathcal{F}_{N,i-1} \right] \xrightarrow{\mathbb{P}} 0 \,.$$

Thus, the two conditions of Theorem 11 in (Douc and Moulines, 2008) are satisfied, which concludes the proof of (4.2.6).

It remains to show establish asymptotic smallness of the normalized weights $\{\tilde{\omega}_{N,i}\, \tilde{\Omega}_N^{-1}\}_{i=1}^{\tilde{M}_N}$. By (4.2.6) it is, using Slutsky's theorem, enough to show that, as $N \to \infty$,

$$\tilde{M}_N^{-1} \max_{1\leq i\leq \tilde{M}_N} \tilde{\omega}_{N,i} \xrightarrow{\mathbb{P}} 0 \,. \tag{4.2.11}$$

Thus, write, for any $\delta > 0$,

$$\tilde{M}_N^{-1} \max_{1\leq i\leq \tilde{M}_N} \tilde{\omega}_{N,i} \leq |\Phi|_\infty (\tilde{M}_N \delta)^{-1} + |\Phi|_\infty \tilde{M}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \psi_{N,I_{N,i}}^{-1} \mathbb{1}_{\{\psi_{N,I_{N,i}}\leq\delta\}} \,,$$

where the first term on the right hand side tends to zero. To treat the second term, define, for $i \in \{1,\dots,\tilde{M}_N\}$, $U_{N,i} := \tilde{M}_N^{-1} \psi_{N,I_{N,i}}^{-1} \mathbb{1}_{\{\psi_{N,I_{N,i}}\leq\delta\}}$ and $\mathcal{F}_{N,i}$ as above; now,

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[U_{N,i}\middle|\mathcal{F}_{N,i}\right] = \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell} \right)^{-1} \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell}\leq\delta\}} \,,$$

where, by Lemma 4.2.1, the first factor converges to $\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$, and

$$\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell}\leq\delta\}} \xrightarrow{\mathbb{P}} \int_{\boldsymbol{\Xi}} \int_{\psi\leq\delta} \Psi(\xi, \mathrm{d}\psi)\, \nu(\mathrm{d}\xi) \,. \tag{4.2.12}$$

The limit quantity in (4.2.12) can, by the dominated convergence theorem, be made arbitrarily small by decreasing δ. Thus, since the $U_{N,i}$'s are all positive, it suffices to establish the last condition of Theorem 11 in (Douc and Moulines, 2008). Hence, let $\epsilon > 0$; then, for any $\delta' > 0$, we have the bound

$$\begin{aligned}
&\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[|U_{N,i}|\mathbb{1}_{\{|U_{N,i}|\geq\epsilon\}} \middle| \mathcal{F}_{N,i-1} \right] \leq \\
&\left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell} \right)^{-1} \left(\mathbb{1}_{\{\tilde{M}_N^{-1}\geq\epsilon\delta'\}} \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell}\leq\delta\}} + \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell} \mathbb{1}_{\{\psi_{N,\ell}\leq\delta\wedge\delta'\}} \right) \,.
\end{aligned} \tag{4.2.13}$$

Finally, by applying again the limit (4.2.12) together with Slutsky's theorem we establish that the right hand side of (4.2.13) tends to $\int_{\boldsymbol{\Xi}} \int_{\psi\leq\delta\wedge\delta'} \Psi(\xi,\mathrm{d}\psi)\,\nu(\mathrm{d}\xi)/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$ as $N \to \infty$, a quantity that can be made arbitrarily small by decreasing δ'. This completes the proof. □

Theorem 4.2.2 (Asymptotic normality of Algorithm 4.2.4)**.** *Assume that* $\{(\xi_{N,i},\omega_{N,i})\}_{i=1}^{M_N}$, Φ, *and* Ψ *satisfy Assumptions* **(A8)**, **(A9)**, *and* **(A10)**. *Moreover, suppose that* $L(\cdot,\tilde{\boldsymbol{\Xi}})$ *belongs to* C. *Finally, assume that* $a_N^{-2}M_N \to \beta^{-1}$ *and* $\tilde{M}_N M_N^{-1} \to \rho$, *for* $\beta \in [0,\infty)$ *and* $\rho \in [0,\infty]$. *Then the weighted particle sample* $\{(\tilde{\xi}_{N,i},\tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ *obtained in Algorithm 4.2.4 is asymptotically normal for* $(\mu,\tilde{\mathsf{A}},\tilde{\mathsf{W}},\tilde{\sigma},\{a_N\}_{N=1}^{\infty})$, *where* $\tilde{\mathsf{A}}$ *and* $\tilde{\mathsf{W}}$, *defined in* (4.2.4), *are proper, and*

$$\tilde{\sigma}^2(f) := \frac{\sigma^2\{L[\cdot,f-\mu(f)]\}}{[\nu L(\tilde{\boldsymbol{\Xi}})]^2} + \beta\frac{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})\nu(R\{\cdot,\Phi^2[f-\mu(f)]^2\}\Psi(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1}))}{\rho[\nu L(\tilde{\boldsymbol{\Xi}})]^2}\,, \quad f\in\tilde{\mathsf{A}}\,,$$

$$\tilde{\gamma}(f) := \beta\frac{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})\nu[R(\cdot,\Phi^2 f)\Psi(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{\rho[\nu L(\tilde{\boldsymbol{\Xi}})]^2}\,, \quad f\in\tilde{\mathsf{W}}\,.$$

Proof. We pick $f\in\tilde{\mathsf{A}}$ and assume without loss of generality that $\mu(f)=0$. Define, for $i\in\{1,\ldots,\tilde{M}_N\}$, the random variables $U_{N,i} := a_N\tilde{M}_N^{-1}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})$ and let the σ-algebras $\mathcal{F}_{N,i}$ be defined as in the proof of Theorem 4.2.1. Make the decomposition $a_N\tilde{\Omega}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i}) = (\tilde{M}_N\tilde{\Omega}_N^{-1})(A_N+B_N)$, with

$$A_N := \sum_{i=1}^{\tilde{M}_N}\mathbb{E}\left[U_{N,i}\middle|\,\mathcal{F}_{N,i-1}\right] = \left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\right)^{-1} a_N\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}L(\xi_{N,\ell},f)\,,$$

$$B_N := \sum_{i=1}^{\tilde{M}_N}\left(U_{N,i}-\mathbb{E}\left[U_{N,i}\middle|\,\mathcal{F}_{N,i-1}\right]\right)\,.$$

By **(A9)**, Equation (4.2.6), and Slutsky's theorem it holds that, supposing $L(\cdot,f)\in\mathsf{A}$,

$$A_N \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\sigma^2\left[\frac{L(\cdot,f)}{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})}\right]\right) \tag{4.2.14}$$

as $N\to\infty$. We now establish similar weak convergence of the sequence $\{B_N\}_{N=1}^{\infty}$ by showing that the two conditions of Theorem 13 in (Douc and Moulines, 2008) are satisfied. In order to compute the asymptotic variance, write

$$\begin{aligned}&\sum_{i=1}^{\tilde{M}_N}\mathbb{E}^2\left[U_{N,i}\middle|\,\mathcal{F}_{N,i-1}\right]\\ &\quad= a_N^2\tilde{M}_N^{-1}\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\right)^{-2}\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}L(\xi_{N,\ell},f)\right)^2 \xrightarrow{\mathbb{P}} \beta\frac{\nu^2L(f)}{\rho\nu^2\Psi(\mathbf{1}_{\mathbb{R}^+})}=0\,,\end{aligned}$$

where we used Lemma 4.2.1, Assumption **(A9)**, and the fact that $\nu L(f)=\mu(f)\nu L(\tilde{\boldsymbol{\Xi}})=0$. Moreover, applying again Lemma 4.2.1 yields, since $R(\cdot,\Phi^2f^2)\Psi(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})\in\mathsf{C}$,

$$\begin{aligned}\sum_{i=1}^{\tilde{M}_N}\mathbb{E}\left[U_{N,i}^2\middle|\,\mathcal{F}_{N,i-1}\right] = a_N^2\tilde{M}_N^{-1}\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\right)^{-1}\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}^{-1}R(\xi_{N,\ell},\Phi^2f^2)\\ \xrightarrow{\mathbb{P}} \beta\frac{\nu[R(\cdot,\Phi^2f^2)\Psi(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{\rho\nu\Psi(\mathbf{1}_{\mathbb{R}^+})}\,,\end{aligned}$$

showing that the first condition is satisfied. To show that also the second condition is satisfied, take $\epsilon > 0$ and write

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[U_{N,i}^2 \mathbb{1}_{\{|U_{N,i}|\geq\epsilon\}} \middle| \mathcal{F}_{N,i-1}\right] \leq a_N^2 \tilde{M}_N^{-1} \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell}\right)^{-1} \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell}^{-1} R(\xi_{N,\ell}, \Phi^2 f^2 \mathbb{1}_{\{\Phi|f|\geq C\}}) + \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell}^{-1} \mathbb{1}_{\{a_N^{-1}\tilde{M}_N\psi_{N,\ell}\leq C\epsilon^{-1}\}} R(\xi_{N,\ell}, \Phi^2 f^2)\right) .$$

Now, adapting the arguments in (4.2.8)–(4.2.10) to the two terms on the right hand side of the previous display shows that their sum tends, as $N \to \infty$, to $\nu[R(\cdot, \Phi^2 f^2 \mathbb{1}_{\{\Phi^2 f^2 \geq C\}}) \Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]$, a quantity that can be made arbitrarily small by increasing C. Thus, applying Theorem 13 in (Douc and Moulines, 2008) gives, for any $u \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\mathrm{i}\, u \sum_{i=1}^{\tilde{M}_N} \{U_{N,i} - \mathbb{E}\left[U_{N,i} \middle| \mathcal{F}_{N,i-1}\right]\}\right) \middle| \mathcal{F}_{N,0}\right] \xrightarrow{\mathbb{P}} \exp\left(-u^2 \beta \frac{\nu[R(\cdot, \Phi^2 f^2)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{2\rho\nu\Psi(\mathbf{1}_{\mathbb{R}^+})}\right) , \quad (4.2.15)$$

from which we, via (4.2.14) and the theorems of dominated convergence and Slutsky, draw the conclusion that

$$(\tilde{M}_N \tilde{\Omega}_N^{-1})(A_N + B_N) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tilde{\sigma}^2(f)) .$$

We now aim at establishing the second property of Definition 4.2.2. By (4.2.6), $\tilde{\Omega}_N^2 \tilde{M}_N^{-2} \xrightarrow{\mathbb{P}} [\nu L(\tilde{\Xi})/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})]^2$, and it is hence enough to show that, for any $f \in \tilde{\mathsf{W}}$ and as $N \to \infty$,

$$a_N^2 \tilde{M}_N^{-2} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i}^2 f(\tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \beta \frac{\nu[R(\cdot, \Phi^2 f)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\rho\nu\Psi(\mathbf{1}_{\mathbb{R}^+})} . \quad (4.2.16)$$

Thus, we define, for $i \in \{1, \ldots, \tilde{M}_N\}$, $U'_{N,i} := a_N^2 \tilde{M}_N^{-2} \tilde{\omega}_{N,i}^2 f(\tilde{\xi}_{N,i})$ and $\mathcal{F}_{N,i}$ as above; then, since $R(\cdot, \Phi^2 f) \leq R(\cdot, \Phi^2|f|) \in \mathsf{C}$,

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[U'_{N,i} \middle| \mathcal{F}_{N,i-1}\right] = a_N^2 \tilde{M}_N^{-1} \left(\Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell}\right)^{-1} \Omega_N^{-1} \sum_{\ell=1}^{M_N} \omega_{N,\ell}\psi_{N,\ell}^{-1} R(\xi_{N,\ell}, \Phi^2 f) \xrightarrow{\mathbb{P}} \beta \frac{\nu[R(\cdot, \Phi^2 f)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\rho\nu\Psi(\mathbf{1}_{\mathbb{R}^+})} . \quad (4.2.17)$$

Analogously, the sequence $\{\sum_{i=1}^{\tilde{M}_N} \mathbb{E}[|U'_{N,i}||\mathcal{F}_{N,i-1}]\}_{N=1}^{\infty}$ converges (in probability) to the constant $\beta\rho^{-1}\nu[R(\cdot, \Phi^2|f|)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$ and is thus tight. Moreover, the two terms

in the bound

$$\sum_{i=1}^{\tilde{M}_N} \mathbb{E}\left[|U'_{N,i}|\mathbb{1}_{\{|U'_{N,i}|\geq\epsilon\}}\middle|\mathcal{F}_{N,i-1}\right] \leq a_N^2\tilde{M}_N^{-1}\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\right)^{-1} \times\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}^{-1}R(\xi_{N,\ell},\Phi^2|f|\mathbb{1}_{\{\Phi^2|f|\geq C\}}) +\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}^{-1}\mathbb{1}_{\{a_N^{-2}\tilde{M}_N^2\psi_{N,\ell}^{()}2\leq C\epsilon^{-1}\}}R(\xi_{N,\ell},\Phi^2|f|)\right) ,$$

for fixed $\epsilon > 0$ and $C > 0$, can, since $a_N^{-2}\tilde{M}_N^2 \to \infty$, be treated in analogy with (4.2.8)–(4.2.10), showing that the sequence $\{\sum_{i=1}^{\tilde{M}_N}\mathbb{E}[|U'_{N,i}|\mathbb{1}_{\{|U'_{N,i}|\geq\epsilon\}}|\mathcal{F}_{N,i-1}]\}_{N=1}^{\infty}$ is bounded, asymptotically, by $\beta\rho^{-1}\nu[R(\cdot,\Phi^2|f|\mathbb{1}_{\{\Phi^2|f|\geq C\}})\Psi(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$. The latter quantity can however be made arbitrarily small by increasing C, and appealing to Theorem 11 in (Douc and Moulines, 2008) completes the proof of (4.2.16).

We now establish the third property of Definition 4.2.2, i.e., asymptotic uniform smallness of the weights $\{\tilde{\omega}_{N,i}\tilde{\Omega}_N^{-1}\}_{i=1}^{\tilde{M}_N}$ at the rate a_N. Again, by (4.2.6) and Slutsky's theorem it is sufficient to show that $a_N^2\tilde{M}_N^{-2}\max_{1\leq i\leq\tilde{M}_N}\tilde{\omega}_{N,i}^2$ vanishes in probability as $N \to \infty$. Hence, decompose, for a fixed $\delta > 0$,

$$a_N^2\tilde{M}_N^{-2}\max_{1\leq i\leq\tilde{M}_N}\tilde{\omega}_{N,i}^2 \leq |\Phi|_\infty^2 a_N^2\tilde{M}_N^{-2}\sum_{i=1}^{\tilde{M}_N}\psi_{N,I_{N,i}}^{-2}\mathbb{1}_{\{\psi_{N,I_{N,i}}\leq\delta\}} + \delta^2|\Phi|_\infty^2 a_N^2\tilde{M}_N^{-2} ,$$

where $\delta^2|\Phi|_\infty^2 a_N^2\tilde{M}_N^{-2} \to 0$. For inspecting the first term, define, for $i \in \{1,\ldots,\tilde{M}_N\}$, the random variables $U''_{N,i} := a_N^2\psi_{N,I_{N,i}}^{-2}\mathbb{1}_{\{\psi_{N,I_{N,i}}\leq\delta\}}$ and the σ-algebras $\mathcal{F}_{N,i}$ as previously. Now,

$$\sum_{i=1}^{\tilde{M}_N}\mathbb{E}\left[U''_{N,i}\middle|\mathcal{F}_{N,i-1}\right] = a_N^2\tilde{M}_N^{-1}\left(\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}\right)^{-1}\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}^{-1}\mathbb{1}_{\{\psi_{N,\ell}\leq\delta\}} ,$$

where, on the right hand side, the product of the first three first factors tends to $\beta\rho^{-1}/\nu\Psi(\mathbf{1}_{\mathbb{R}^+})$. By Lemma 4.2.1 and **(A10)**,

$$\Omega_N^{-1}\sum_{\ell=1}^{M_N}\omega_{N,\ell}\psi_{N,\ell}^{-1}\mathbb{1}_{\{\psi_{N,\ell}\leq\delta\}} \xrightarrow{\mathbb{P}} \int_{\Xi}\int_{\psi\leq\delta}\psi^{-1}\,\Psi(\xi,\mathrm{d}\psi)\,\nu(\mathrm{d}\xi) ,$$

where the limit can be made arbitrarily small by increasing C. Finally, arguing along the lines of (4.2.13) gives that, for any $\epsilon > 0$,

$$\sum_{i=1}^{\tilde{M}_N}\mathbb{E}\left[|U''_{N,i}|\mathbb{1}_{\{|U''_{N,i}|\geq\epsilon\}}\middle|\mathcal{F}_{N,i-1}\right] \xrightarrow{\mathbb{P}} 0 ,$$

and the asymptotic smallness follows by applying Theorem 11 in (Douc and Moulines, 2008). This completes the proof of the theorem. □

The results above extend all existing asymptotic convergence results (see Douc et al., 2008) for the APF since the AMWs are allowed to fluctuate randomly as described by the kernel Ψ.

4.3 Adaptation of SMC algorithms

4.3.1 Mutation with adaptive selection (MAS)

Criteria for detecting weight degeneracy

Since, as explained above, the weight degeneracy phenomenon deteriorates drastically the particle approximation, it is of practical importance to set up criteria which detect, online and at a limited computational cost, such degeneracy. In addition, it is of significance that such criteria are not oversensitive inasmuch as selecting well balanced particle weights increases unnecessarily the asymptotic variance of the produced estimates. Denote by, for a set of nonnegative numbers $\{a_i\}_{i=1}^N$,

$$\mathrm{CV}^2(\{a_i\}_{i=1}^N) := N \sum_{i=1}^N \left(\frac{a_\ell}{\sum_{\ell=1}^N a_\ell} \right)^2 - 1$$

the *coefficient of variation* (CV) of $\{a_i\}_{i=1}^N$. This quantity is minimal when all a_i's are equal and maximal when all but one are zero. Thus, Kong et al. (1994) proposed to use $\mathrm{CV}^2(\{\omega_{N,i}\}_{i=1}^{M_N})$ as a means for detecting weight degeneracy of a sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$. Another criterion having similar properties is the *Shannon entropy*-like quantity

$$\mathcal{E}(\{a_i\}_{i=1}^N) := \sum_{i=1}^N \frac{a_i}{\sum_{\ell=1}^N a_\ell} \log \left(\frac{N a_i}{\sum_{\ell=1}^N a_\ell} \right)$$

studied by Cornebise et al. (2008). In Algorithm 4.3.1, the auxiliary selection operation is put on standby and activated only when the CV (or entropy) of the particle weights exceeds a prespecified threshold κ.

Algorithm 4.3.1 Mutation with adaptive selection

procedure MAS($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, R, Φ, Ψ, $\tilde{M}_N$, κ)
  if $\mathrm{CV}^2(\{\omega_{N,i}\}_{i=1}^{M_N}) \geq \kappa$ **then**
    $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N} \leftarrow$ APF($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}, R, \Phi, \Psi, \tilde{M}_N$);
  else
    $\tilde{M}_N \leftarrow M_N$;
    $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N} \leftarrow$ Mutation($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}, R, \Phi$);
  end if

  return $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$
end procedure

Convergence of the MAS algorithm

Define the sets

$$\begin{aligned} \bar{\mathsf{A}} &:= \left\{ f \in \mathsf{L}^1(\mu, \tilde{\Xi}) : L(\cdot, |f|) \in \mathsf{A} \cap \mathsf{C}, R(\cdot, \Phi^2 f^2) \Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C}, R(\cdot, \Phi^2 f^2) \in \mathsf{W} \right\}, \\ \bar{\mathsf{W}} &:= \left\{ f : R(\cdot, \Phi^2 |f|) \Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C}, R(\cdot, \Phi^2 |f|) \in \mathsf{W} \cap \mathsf{C} \right\}. \end{aligned} \tag{4.3.1}$$

Then the following result, whose (short) proof is omitted for brevity, follows straightforwardly from Theorems 4.2.1, 4.2.2, and from the limit theorems for the mutation step from Douc and Moulines (2008) recalled in Appendix A, Theorems A.2.2 and A.2.2.

Theorem 4.3.1 (Consistency and asymptotic normality of Algorithm 4.3.1)**.** *Let the assumptions of Theorem 4.2.2 hold. Then the weighted particle sample* $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ *obtained in Algorithm 4.3.1 is consistent and asymptotically normal for* $(\mu, \tilde{\mathsf{C}})$ *and* $(\mu, \bar{\mathsf{A}}, \bar{\mathsf{W}}, \bar{\sigma}, \bar{\gamma}, \{a_N\}_{N=1}^{\infty})$*, respectively, where* $\bar{\mathsf{A}}$ *and* $\bar{\mathsf{W}}$*, defined in* (4.3.1)*, are proper and*

$$\bar{\sigma}^2(f) := \frac{\sigma^2\{L[\cdot, f-\mu(f)]\} + \varepsilon\gamma R(\{\Phi[f-\mu(f)] - R(\cdot, \Phi[f-\mu(f)])\}^2)}{[\nu L(\tilde{\mathbf{\Xi}})]^2} + \beta(1-\varepsilon)\frac{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})\nu(R\{\cdot, \Phi^2[f-\mu(f)]^2\}\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}))}{\rho[\nu L(\tilde{\mathbf{\Xi}})]^2}, \quad f \in \bar{\mathsf{A}}, \tag{4.3.2}$$

with $\varepsilon := \mathbb{1}_{\{\gamma(\mathbf{\Xi})<\kappa+1\}}$ *and*

$$\bar{\gamma}(f) := \beta(1-\varepsilon)\frac{\nu\Psi(\mathbf{1}_{\mathbb{R}^+})\nu[R(\cdot, \Phi^2 f)\Psi(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\rho[\nu L(\tilde{\mathbf{\Xi}})]^2} + \varepsilon\frac{\gamma R(\Phi^2 f)}{[\nu L(\tilde{\mathbf{\Xi}})]^2}, \quad f \in \bar{\mathsf{W}}. \tag{4.3.3}$$

4.3.2 SIS with adaptive selection (SISAS)

Applying sequentially Algorithm 4.3.1 yields the by far most commonly used technique for preventing adaptively the particle weights from degenerating, and the aim of this section is to investigate this scheme theoretically. Thus, assume that we have produced a weighted sample $\{(\omega_{N,i}^{(0)}, \xi_{N,i}^{(0)})\}_{i=1}^{M_N^0}$ targeting the initial distribution μ_0; then an updated particle sample targeting μ_n is obtained using the recursive procedure

Algorithm 4.3.2 SIS with adaptive selection

procedure SISAS($\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$, $\{(R_\ell, \Phi_\ell, \Psi_\ell, M_N^\ell, \kappa_\ell)\}_{\ell=0}^{n-1}$)
  for $\ell \leftarrow 0$ **to** $n-1$ **do**
    $\{(\xi_{N,i}^{(\ell+1)}, \omega_{N,i}^{(\ell+1)})\}_{i=1}^{M_N^{\ell+1}} \leftarrow$ **MAS**($\{(\xi_{N,i}^{(\ell)}, \omega_{N,i}^{(\ell)})\}_{i=1}^{M_N^\ell}, R_\ell, \Phi_\ell, \Psi_\ell, M_N^{\ell+1}, \kappa_\ell$);
  end for
  return $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N^n}$
end procedure

Setting $\kappa_\ell \equiv 0$ for all ℓ yields an algorithm in which the particles are selected systematically, which eliminates completely all weight degeneracy at the cost of variance added by the multinomial resampling operation. However, as we shall see in the following, the SISAS approach has a significant drawback in the sense that it does not at all adjust adaptively the IS proposal distribution inherent of the particle filter to the corresponding target (by, e.g., minimizing some divergence between these distributions). To describe this formally, let $\{\mu_\ell\}_{\ell=0}^{\infty}$ be the Feynman-Kac flow generated by (4.2.2). Moreover, assume that the initial sample $\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$ satisfies **(A8)** and **(A9)** for some initial parameters (μ_0, C_0) and $(\mu_0, \mathsf{A}_0, \mathsf{W}_0, \sigma_0, \gamma_0, \{a_N\}_{N=1}^{\infty})$, respectively, and define recursively (in accordance with (4.3.1)) the sets

$$\begin{aligned}
\mathsf{C}_{\ell+1} &:= \left\{f \in \mathsf{L}^1(\mu_{\ell+1}, \mathbf{\Xi}_{\ell+1}) : L_\ell(\cdot, |f|) \in \mathsf{C}_\ell\right\}, \\
\mathsf{A}_{\ell+1} &:= \left\{f \in \mathsf{L}^1(\mu_{\ell+1}, \mathbf{\Xi}_{\ell+1}) : L_\ell(\cdot, f) \in \mathsf{A}_\ell \cap \mathsf{C}_\ell, R_\ell(\cdot, \Phi_\ell^2 f^2)\Psi_\ell(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C}_\ell, R_\ell(\cdot, \Phi_\ell^2 f^2) \in \mathsf{W}_\ell\right\}, \\
\mathsf{W}_{\ell+1} &:= \left\{f : R_\ell(\cdot, \Phi_\ell^2|f|)\Psi_\ell(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \in \mathsf{C}_\ell, R_\ell(\cdot, \Phi_\ell^2|f|) \in \mathsf{W}_\ell \cap \mathsf{C}_\ell\right\}.
\end{aligned} \tag{4.3.4}$$

In addition, define recursively

$$\gamma_{\ell+1}(f) := \mathbb{1}_{\{\gamma_\ell(\Xi_\ell)\geq\kappa_\ell+1\}}\frac{\mu_\ell\Psi_\ell(\mathbf{1}_{\mathbb{R}^+})\mu_\ell[R_\ell(\cdot,\Phi_\ell^2 f)\Psi_\ell(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2} + \mathbb{1}_{\{\gamma_\ell(\Xi_\ell)<\kappa_\ell+1\}}\frac{\gamma_\ell R_\ell(\Phi_\ell^2 f)}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2}\,, \quad f\in\mathsf{W}_{\ell+1}\,, \tag{4.3.5}$$

and the asymptotic variances

$$\begin{aligned}\sigma^2_{\ell+1}(f) := {}& \frac{\sigma^2_\ell\{L_\ell[\cdot,f-\mu_{\ell+1}(f)]\}}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2} \\ &+ \mathbb{1}_{\{\gamma_\ell(\Xi_\ell)<\kappa_\ell+1\}}\frac{\gamma_\ell R_\ell(\{\Phi_\ell[f-\mu_{\ell+1}(f)]-R_\ell(\cdot,\Phi_\ell[f-\mu_{\ell+1}(f)])\}^2)}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2} \\ &+ \mathbb{1}_{\{\gamma_\ell(\Xi_\ell)\geq\kappa_\ell+1\}}\frac{\mu_\ell\Psi_\ell(\mathbf{1}_{\mathbb{R}^+})\mu_\ell(R_\ell\{\cdot,\Phi_\ell^2[f-\mu_{\ell+1}(f)]^2\}\Psi_\ell(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1}))}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2}\,, \quad f\in\mathsf{A}_{\ell+1}\,,\end{aligned} \tag{4.3.6}$$

and set $\mathrm{s}(n) := \min\{\ell < n : \gamma_\ell(\Xi_\ell) \geq \kappa_\ell + 1\}$, $n \in \mathbb{N}$. Let $\{(\xi_{N,i}^{(\ell)}, \omega_{N,i}^{(\ell)})\}_{i=1}^{M_N}$, $\ell \geq 0$, be a sequence of weighted samples produced by means of Algorithm 4.3.2 where the particle sample size is the same at all iterations, that is, $M_N^\ell \equiv M_N$ for all ℓ. A framework with varying particle sample sizes is studied in Section 4.3.3. Since, as a consequence of Theorem 4.3.1, $\mathrm{CV}^2(\{\omega_{N,i}^{(\ell)}\}_{i=1}^{M_N})$ tends, for all ℓ, to $\gamma_\ell(\Xi_\ell) - 1$ when N grows to infinity and the initial sample $\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$ is asymptotically normal for μ_0 at the rate $a_N = M_N^{1/2}$ (this result is stated rigorously in Theorem 4.3.2), $\mathrm{s}(n)$ can, for a sufficiently large particle sample size M_N, be interpreted as the last time adaptive selection was performed standing at timestep n. Now consider, for $n \in \mathbb{N}$, the random measures

$$\mu_{N,n}(A) := \sum_{i=1}^{M_N}\frac{\omega_{N,i}^{(\mathrm{s}(n))}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))},\Xi_n)}{\sum_{j=1}^{M_N}\omega_{N,j}^{(\mathrm{s}(n))}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,j}^{(\mathrm{s}(n))},\Xi_n)}\left[\frac{L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))},A)}{L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))},\Xi_n)}\right]\,, \tag{4.3.7}$$

for any $A \in \mathcal{B}(\Xi_n)$, and

$$\pi_{N,n}(A) := \sum_{i=1}^{M_N}\frac{\omega_{N,i}^{(\mathrm{s}(n))}\psi_{N,i}^{(\mathrm{s}(n))}}{\sum_{j=1}^{M_N}\omega_{N,j}^{(\mathrm{s}(n))}\psi_{N,j}^{(\mathrm{s}(n))}}R_{\mathrm{s}(n)}\cdots R_n(\xi_{N,i}^{(\mathrm{s}(n))},A)\,, \quad A\in\mathcal{B}(\Xi_n)\,,$$

where each $\psi_{N,i}^{(\mathrm{s}(n))}$ is a random weight distributed according to $\Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))},\cdot)$. Clearly, $\mu_{N,n}$ is the mixture distribution obtained by simply replacing $\mu_{\mathrm{s}(n)}$ in the multi-step Feynman-Kac transition formula

$$\mu_n(A) = \frac{\mu_{\mathrm{s}(n)}L_{\mathrm{s}(n)}\cdots L_{n-1}(A)}{\mu_{\mathrm{s}(n)}L_{\mathrm{s}(n)}\cdots L_{n-1}(\Xi_n)}\,, \quad A\in\mathcal{B}(\Xi_n)\,,$$

by the weighted empirical measure associated with the weighted particle sample at time $\mathrm{s}(n)$. In the asymptotic (with regard to the particle population size) regime, executing Algorithm 4.3.2 for $\ell \leftarrow \mathrm{s}(n)$ to $n-1$ involves a selection step at time $\mathrm{s}(n)$ followed by a sequence of $n - \mathrm{s}(n)$ mutation operations; this can be expressed equivalently as an IS problem on the product space $\{1,\ldots,M_N\}\times\Xi_{\mathrm{s}(n)+1:n}$ where the target distribution

$$\mu_{\mathrm{aux}}^{N,n}(\{i\}\times A) := \frac{\omega_{N,i}^{(\mathrm{s}(n))}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))},\Xi_n)}{\sum_{j=1}^{M_N}\omega_{N,j}^{(\mathrm{s}(n))}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,j}^{(\mathrm{s}(n))},\Xi_n)}\left[\frac{\bigotimes_{\ell=\mathrm{s}(n)}^{n-1}L_\ell(\xi_{N,i}^{(\mathrm{s}(n))},A)}{L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))},\Xi_n)}\right] \tag{4.3.8}$$

is estimated using the instrumental distribution

$$\pi_{\text{aux}}^{N,n}(\{i\} \times A) := \frac{\omega_{N,i}^{(\mathrm{s}(n))} \psi_{N,i}^{(\mathrm{s}(n))}}{\sum_{j=1}^{M_N} \omega_{N,j}^{(\mathrm{s}(n))} \psi_{N,j}^{(\mathrm{s}(n))}} \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi_{N,i}^{(\mathrm{s}(n))}, A) \, .$$

These measures are closely related to $\mu_{N,n}$ and $\pi_{N,n}$ since the latter are the restrictions (marginal distributions) of the former, respectively, with respect to $\boldsymbol{\Xi}_n$. Thus, a weighted particle sample $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$ targeting $\mu_{N,n}$ can be obtained by simulating pairs $\{(I_{N,i}^{(\mathrm{s}(n))}, \xi_{N,i}^{(\mathrm{s}(n)+1:n)})\}_{i=1}^{M_N}$ of indices and particle trajectories from $\pi_{\text{aux}}^{N,n}$ and associating to each draw the importance weight

$$\omega_{N,i}^{(n)} := \left(\psi_{N,I_{N,i}^{(\mathrm{s}(n))}}^{(\mathrm{s}(n))}\right)^{-1} \Phi_{\mathrm{s}(n),n-1}\left(\xi_{N,I_{N,i}^{(\mathrm{s}(n))}}^{(\mathrm{s}(n))}, \xi_{N,i}^{(\mathrm{s}(n)+1:n)}\right) \propto \frac{\mathrm{d}\mu_{\text{aux}}^{N,n}}{\mathrm{d}\pi_{\text{aux}}^{N,n}}(I_{N,i}^{(\mathrm{s}(n))}, \xi_{N,i}^{(\mathrm{s}(n)+1:n)}) \, .$$

Next, the auxiliary variables $\{(I_{N,i}^{(\mathrm{s}(n))}, \xi_{N,i}^{(\mathrm{s}(n)+1:n-1)})\}_{i=1}^{M_N}$ are discarded, while the weighted sample $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$ is used for estimating $\mu_{N,n}$. The efficiency of this IS approach depends highly on the design of the adjustment weights $\{\psi_{N,i}^{(\mathrm{s}(n))}\}_{i=1}^{M_N}$ and the instrumental kernels $\{R_\ell\}_{\ell=\mathrm{s}(n)}^{n-1}$. In particular, we may expect that the quality of the final sample is high if the discrepancy, as measured by say the CSD, between μ_{aux}^N and π_{aux}^N is low. Define, for $\ell \in \mathbb{N}^*$, the σ-fields $\mathcal{F}_N^\ell := \sigma(\{(\xi_{N,i}^{(\ell)}, \omega_{N,i}^{(\ell)})\}_{i=1}^{M_N})$; then $\mathbb{E}[d_{\chi^2}(\mu_{\text{aux}}^{N,n} || \pi_{\text{aux}}^{N,n}) | \mathcal{F}_N^{\mathrm{s}(n)}]$, where the expectation is taken over the $\psi_{N,i}^{(\mathrm{s}(n))}$'s only, is the *expected* CSD between the proposal and target distributions given the particle system at time $\mathrm{s}(n)$. Impose the following assumption.

(A11) *The functions* $L_{\mathrm{s}(n)} \cdots L_{n-1}(\cdot, \boldsymbol{\Xi}_n)$ *and* $\Psi_{\mathrm{s}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\cdot, \Phi_{\mathrm{s}(n),n-1}^2)$ *belong to* $\mathsf{C}_{\mathrm{s}(n)}$ *for all* $n \in \mathbb{N}$.

Adding **(A11)** to our list of assumptions, the next theorem states that $\mathrm{CV}^2(\{\omega_{N,i}^{(n)}\}_{i=1}^{M_N})$ is a consistent estimate of the expected CSD at any timestep n.

Theorem 4.3.2 (Convergence of Algorithm 4.3.2)**.** *Assume* **(A11)** *and suppose that the sample* $\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$ *satisfies* **(A8)** *and* **(A9)** *for* (μ_0, C_0) *and* $(\mu_0, \mathsf{A}_0, \mathsf{W}_0, \sigma_0, \gamma_0, \{M_N^{1/2}\}_{N=1}^\infty)$*, respectively. In addition, let* $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$*,* $n \in \mathbb{N}$*, the be the sample returned by Algorithm 4.3.2 with input parameters* $M_N^\ell \equiv M_N$ *and* $(\Phi_\ell, \Psi_\ell, \mathsf{C}_\ell)$*,* $\ell \in \{0, \ldots, n-1\}$*, satisfying* **(A10)** *for all* ℓ*. Finally, suppose that* $L_\ell(\cdot, \boldsymbol{\Xi}_{\ell+1}) \in \mathsf{C}_\ell$ *for all* ℓ*. Then, the following holds.*

i) *The output* $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$ *is consistent and asymptotically normal for* (μ_n, C_n) *and* $(\mu_n, \mathsf{A}_n, \mathsf{W}_n, \sigma_n, \gamma_n, \{M_N^{1/2}\}_{N=1}^\infty)$ *where* $(\mathsf{C}_n, \mathsf{A}_n, \mathsf{W}_n)$*,* σ_n*, and* γ_n *are defined via the recursions* (4.3.6) *and* (4.3.5)*, respectively.*

ii)
$$\left| \mathbb{E}\left[d_{\chi^2}(\mu_{\text{aux}}^{N,n} || \pi_{\text{aux}}^{N,n}) \middle| \mathcal{F}_N^{\mathrm{s}(n)} \right] - \mathrm{CV}^2(\{\omega_{N,i}^{(n)}\}_{i=1}^{M_N}) \right| \xrightarrow{\mathbb{P}} 0 \, , \quad N \to \infty \, .$$

Proof. By definition of $\mathrm{s}(n)$ it follows that

$$\begin{aligned} \gamma_n(\boldsymbol{\Xi}_n) &= \frac{\gamma_{n-1} R_{n-1}(\Phi_{n-1}^2)}{[\mu_{n-1} L_{n-1}(\boldsymbol{\Xi}_n)]^2} = \frac{\gamma_{n-2} R_{n-2} \otimes R_{n-1}(\Phi_{n-2,n-1}^2)}{[\mu_{n-2} L_{n-2}(\boldsymbol{\Xi}_{n-1})]^2 [\mu_{n-1} L_{n-1}(\boldsymbol{\Xi}_n)]^2} \\ &= \cdots = \frac{\gamma_{\mathrm{s}(n)+1} \bigotimes_{\ell=\mathrm{s}(n)+1}^{n-1} R_\ell(\Phi_{\mathrm{s}(n)+1,n-1}^2)}{\prod_{\ell=\mathrm{s}(n)+1}^{n-1} [\mu_\ell L_\ell(\boldsymbol{\Xi}_{\ell+1})]^2} = \frac{\gamma_{\mathrm{s}(n)+1} \bigotimes_{\ell=\mathrm{s}(n)+1}^{n-1} R_\ell(\Phi_{\mathrm{s}(n)+1,n-1}^2)}{[\mu_{\mathrm{s}(n)+1} L_{\mathrm{s}(n)+1} \cdots L_{n-1}(\boldsymbol{\Xi}_n)]^2} \, . \end{aligned}$$

Combining this with the identity

$$\gamma_{\mathrm{s}(n)+1}(f) = \frac{\mu_{\mathrm{s}(n)}\Psi_{\mathrm{s}(n)}(\mathbf{1}_{\mathbb{R}^+})\mu_{\mathrm{s}(n)}[R_{\mathrm{s}(n)}(\cdot,\Phi^2_{\mathrm{s}(n)}f)\Psi_{\mathrm{s}(n)}(\cdot,\mathbf{1}^{-1}_{\mathbb{R}^+})]}{[\mu_{\mathrm{s}(n)}L_{\mathrm{s}(n)}(\boldsymbol{\Xi}_{\mathrm{s}(n)+1})]^2}\,, \quad f \in \mathsf{W}_{\mathrm{s}(n)+1}\,,$$

yields

$$\gamma_n(\boldsymbol{\Xi}_n) = \frac{\mu_{\mathrm{s}(n)}\Psi_{\mathrm{s}(n)}(\mathbf{1}_{\mathbb{R}^+})\mu_{\mathrm{s}(n)}[\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\cdot,\Phi^2_{\mathrm{s}(n),n-1})\Psi_{\mathrm{s}(n)}(\cdot,\mathbf{1}^{-1}_{\mathbb{R}^+})]}{[\mu_{\mathrm{s}(n)}L_{\mathrm{s}(n)}\cdots L_{n-1}(\boldsymbol{\Xi}_n)]^2}\,. \tag{4.3.9}$$

On the other hand, the CSD between $\mu^{N,n}_{\mathrm{aux}}$ and $\pi^{N,n}_{\mathrm{aux}}$ can be expressed as

$$\begin{aligned}
\mathbb{E}\left[d_{\chi^2}(\mu^{N,n}_{\mathrm{aux}}||\pi^{N,n}_{\mathrm{aux}})\,\middle|\,\mathcal{F}^{\mathrm{s}(n)}_N\right] &= \sum_{i=1}^{M_N}\mathbb{E}\left[\int \frac{\mathrm{d}\mu^{N,n}_{\mathrm{aux}}}{\mathrm{d}\pi^{N,n}_{\mathrm{aux}}}(i,\xi_{\mathrm{s}(n)+1:n})\,\mu^{N,n}_{\mathrm{aux}}(\{i\}\times \mathrm{d}\xi_{\mathrm{s}(n)+1:n})\,\middle|\,\mathcal{F}^{\mathrm{s}(n)}_N\right] - 1 \\
&= \sum_{i=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,i}\sum_{j=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,j}\mathbb{E}\left[\psi^{(\mathrm{s}(n))}_{N,j}(\psi^{(\mathrm{s}(n))}_{N,i})^{-1}\,\middle|\,\mathcal{F}^{\mathrm{s}(n)}_N\right]\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi^{(\mathrm{s}(n))}_{N,i},\Phi^2_{\mathrm{s}(n),n-1}) \\
&\qquad \times\left(\sum_{l=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,l}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi^{(\mathrm{s}(n))}_{N,l},\boldsymbol{\Xi}_n)\right)^{-2} - 1\,.
\end{aligned}$$

Here, as the $\psi^{(\mathrm{s}(n))}_{N,i}$'s are conditionally independent given $\mathcal{F}^{\mathrm{s}(n)}_N$,

$$\begin{aligned}
\sum_{j=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,j}\mathbb{E}\left[\psi^{(\mathrm{s}(n))}_{N,j}(\psi^{(\mathrm{s}(n))}_{N,i})^{-1}\,\middle|\,\mathcal{F}^{\mathrm{s}(n)}_N\right] &= \Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}^{-1}_{\mathbb{R}^+})\sum_{j=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,j}\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,j},\mathbf{1}_{\mathbb{R}^+}) \\
&\quad - \omega^{(\mathrm{s}(n))}_{N,i}\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}^{-1}_{\mathbb{R}^+})\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}_{\mathbb{R}^+}) + \omega^{(\mathrm{s}(n))}_{N,i}\,,
\end{aligned}$$

yielding

$$\begin{aligned}
&\mathbb{E}\left[d_{\chi^2}(\mu^{N,n}_{\mathrm{aux}}||\pi^{N,n}_{\mathrm{aux}})\,\middle|\,\mathcal{F}^{\mathrm{s}(n)}_N\right] \\
&\qquad = \Bigg((\Omega^{(\mathrm{s}(n))}_N)^{-1}\sum_{j=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,j}\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,j},\mathbf{1}_{\mathbb{R}^+})(\Omega^{(\mathrm{s}(n))}_N)^{-1} \\
&\qquad\quad \times\sum_{i=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,i}\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}^{-1}_{\mathbb{R}^+})\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi^{(\mathrm{s}(n))}_{N,i},\Phi^2_{\mathrm{s}(n),n-1}) \\
&\quad - (\Omega^{(\mathrm{s}(n))}_N)^{-2}\sum_{i=1}^{M_N}(\omega^{(\mathrm{s}(n))}_{N,i})^2\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}^{-1}_{\mathbb{R}^+})\Psi_{\mathrm{s}(n)}(\xi^{(\mathrm{s}(n))}_{N,i},\mathbf{1}_{\mathbb{R}^+})\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi^{(\mathrm{s}(n))}_{N,i},\Phi^2_{\mathrm{s}(n),n-1}) \\
&\qquad\quad +(\Omega^{(\mathrm{s}(n))}_N)^{-2}\sum_{i=1}^{M_N}(\omega^{(\mathrm{s}(n))}_{N,i})^2\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi^{(\mathrm{s}(n))}_{N,i},\Phi^2_{\mathrm{s}(n),n-1})\Bigg) \\
&\qquad\quad \times\left((\Omega^{(\mathrm{s}(n))}_N)^{-1}\sum_{l=1}^{M_N}\omega^{(\mathrm{s}(n))}_{N,l}L_{\mathrm{s}(n)}\cdots L_{n-1}(\xi^{(\mathrm{s}(n))}_{N,l},\boldsymbol{\Xi}_n)\right)^{-2} - 1\,.
\end{aligned} \tag{4.3.10}$$

Now, since $\{(\xi_{N,i}^{(\mathrm{s}(n))}, \omega_{N,i}^{(\mathrm{s}(n))})\}_{i=1}^{M_N}$ is consistent for $(\mu_{\mathrm{s}(n)}, \mathsf{C}_{\mathrm{s}(n)})$ we obtain the limits, as $N \to \infty$,

$$(\Omega_N^{(\mathrm{s}(n))})^{-1} \max_{1 \le i \le M_N} \omega_{N,i}^{(\mathrm{s}(n))} \xrightarrow{\mathbb{P}} 0 \,,$$

$$(\Omega_N^{(\mathrm{s}(n))})^{-1} \sum_{i=1}^{M_N} \omega_{N,i}^{(\mathrm{s}(n))} \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}^{-1}) \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}) \xrightarrow{\mathbb{P}} \mu_{\mathrm{s}(n)}[\Psi_{\mathrm{s}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \Psi_{\mathrm{s}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+})] \,,$$

and, consequently,

$$\begin{aligned} 0 \le (\Omega_N^{(\mathrm{s}(n))})^{-2} \sum_{i=1}^{M_N} (\omega_{N,i}^{(\mathrm{s}(n))})^2 \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi_{N,i}^{(\mathrm{s}(n))}, \Phi_{\mathrm{s}(n),n-1}^2) \\ \le \left((\Omega_N^{(\mathrm{s}(n))})^{-1} \max_{1 \le l \le M_N} \omega_{N,l}^{(\mathrm{s}(n))} \right)^2 \prod_{\ell=\mathrm{s}(n)}^{n-1} |\Phi_\ell|_\infty^2 \xrightarrow{\mathbb{P}} 0 \end{aligned} \tag{4.3.11}$$

and, by Slutsky's theorem,

$$\begin{aligned} 0 \le (\Omega_N^{(\mathrm{s}(n))})^{-2} \sum_{i=1}^{M_N} (\omega_{N,i}^{(\mathrm{s}(n))})^2 \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}^{-1}) \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}) \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi_{N,i}^{(\mathrm{s}(n))}, \Phi_{\mathrm{s}(n),n-1}^2) \\ \le (\Omega_N^{(\mathrm{s}(n))})^{-2} \max_{1 \le l \le M_N} \omega_{N,l}^{(\mathrm{s}(n))} \prod_{\ell=\mathrm{s}(n)}^{n-1} |\Phi_\ell|_\infty^2 \sum_{i=1}^{M_N} \omega_{N,i}^{(\mathrm{s}(n))} \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}^{-1}) \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}) \xrightarrow{\mathbb{P}} 0 \,. \end{aligned} \tag{4.3.12}$$

Applying (4.3.11) and (4.3.12) together with the limits

$$(\Omega_N^{(\mathrm{s}(n))})^{-1} \sum_{i=1}^{M_N} \omega_{N,j}^{(\mathrm{s}(n))} \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}) \xrightarrow{\mathbb{P}} \mu_{\mathrm{s}(n)} \Psi_{\mathrm{s}(n)}(\mathbf{1}_{\mathbb{R}^+}) \,,$$

$$\begin{aligned} (\Omega_N^{(\mathrm{s}(n))})^{-1} \sum_{i=1}^{M_N} \omega_{N,i}^{(\mathrm{s}(n))} \Psi_{\mathrm{s}(n)}(\xi_{N,i}^{(\mathrm{s}(n))}, \mathbf{1}_{\mathbb{R}^+}^{-1}) \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\xi_{N,i}^{(\mathrm{s}(n))}, \Phi_{\mathrm{s}(n),n-1}^2) \xrightarrow{\mathbb{P}} \\ \mu_{\mathrm{s}(n)} \left(\Psi_{\mathrm{s}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}) \bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\cdot, \Phi_{\mathrm{s}(n),n-1}^2) \right) \,, \end{aligned}$$

$$(\Omega_N^{(\mathrm{s}(n))})^{-1} \sum_{i=1}^{M_N} \omega_{N,i}^{(\mathrm{s}(n))} L_{\mathrm{s}(n)} \cdots L_{n-1}(\xi_{N,i}^{(\mathrm{s}(n))}, \Xi_n) \xrightarrow{\mathbb{P}} \mu_{\mathrm{s}(n)} L_{\mathrm{s}(n)} \cdots L_{n-1}(\Xi_n)$$

to (4.3.10) gives (recall (4.3.9))

$$\begin{aligned} \mathbb{E}\left[d_{\chi^2}(\mu_{\mathrm{aux}}^{N,n} || \pi_{\mathrm{aux}}^{N,n}) \,\middle|\, \mathcal{F}_N^{\mathrm{s}(n)} \right] \xrightarrow{\mathbb{P}} \frac{\mu_{\mathrm{s}(n)} \Psi_{\mathrm{s}(n)}(\mathbf{1}_{\mathbb{R}^+}) \mu_{\mathrm{s}(n)}[\bigotimes_{\ell=\mathrm{s}(n)}^{n-1} R_\ell(\cdot, \Phi_{\mathrm{s}(n),n-1}^2) \Psi_{\mathrm{s}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{[\mu_{\mathrm{s}(n)} L_{\mathrm{s}(n)} \cdots L_{n-1}(\Xi_n)]^2} - 1 \\ = \gamma_n(\Xi_n) - 1 \,. \end{aligned} \tag{4.3.13}$$

Finally, since also $\mathrm{CV}^2(\{\omega_{N,i}^{(n)}\}_{i=1}^{M_N}) \xrightarrow{\mathbb{P}} \gamma_n(\Xi_n) - 1$, the proof is completed. □

Theorem 4.3.2 extends similar results obtained by (Cornebise et al., 2008, Theorems 1(i) and 2(i)) under the assumption that selection is performed systematically at all timesteps. Referring to Theorem 4.3.2, selecting, as in the standard SISAS scheme, the particles not until the CV is large does not necessarily improve the quality of the particle sample, since the offspring will be selected from a set of ancestor particles being sampled from a proposal distribution which is badly adjusted to the target $\mu_{\text{aux}}^{N,n}$. Especially, in the extreme situation where the divergence between the instrumental and target distributions in question is so large that *all* particles are proposed in a state space region far from the main mode of $\mu_{\text{aux}}^{N,n}$, trying to improve the particle sample through multinomial selection will be unavailing since the particle cloud has already lost its track.

4.3.3 Mutation with pilot exploration and adaptive refueling (MPEAR)

In the light of the results of the previous section we now propose and justify theoretically a natural modification of the SISAS scheme. In our version, a *pilot sample*, evolving in front of the main particle swarm, detects and reports, via the CV, large CSDs at future timesteps to the main cloud. In addition, since a large CSD indicates a demanding IS problem, we let the occurrence of a large CV activate automatically a *refueling operation* in which the number of particles is increased by a factor determined by an increasing function φ of the CV, as described in Algorithm 4.3.3. Including Algorithm 4.3.3 in the sequential context yields Algorithm 4.3.4.

Algorithm 4.3.3 Mutation with pilot exploration and adaptive refueling

procedure MPEAR($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$, R, Φ, φ, κ)
  $\{(\bar{\xi}_{N,i}, \bar{\omega}_{N,i})\}_{i=1}^{M_N} \leftarrow$MUTATION($\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}, R, \Phi$); ▷ pilot exploration
  if $\mathrm{CV}^2(\{\bar{\omega}_{N,i}\}_{i=1}^{M_N}) \geq \kappa$; **then**
    $\tilde{M}_N \leftarrow \lceil M_N \varphi(\mathrm{CV}^2(\{\bar{\omega}_{N,i}\}_{i=1}^{M_N})) \rceil$; ▷ refueling
    $\{(\hat{\xi}_{N,i}, 1), I_{N,i}\}_{i=1}^{\tilde{M}_N} \leftarrow$SELECTION($\{(\xi_{N,i}, \bar{\omega}_{N,i})\}_{i=1}^{M_N}, \tilde{M}_N$);
    $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N} \leftarrow$MUTATION($\{(\hat{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}, R, \Phi$);
    for $\ell \leftarrow 1$ **to** $\tilde{M}_N$ **do**
      $\tilde{\omega}_{N,\ell} \leftarrow \tilde{\omega}_{N,\ell} \bar{\omega}_{N,I_{N,\ell}}^{-1} \omega_{N,I_{N,\ell}}$;
    end for
    return $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$

  else
    return $\{(\bar{\xi}_{N,i}, \bar{\omega}_{N,i})\}_{i=1}^{M_N}$
  end if
end procedure

Algorithm 4.3.4 SIS with adaptive refueling

procedure SISAR($\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$, $\{(R_\ell, \Phi_\ell, \varphi_\ell, \kappa_\ell)\}_{\ell=0}^{n-1}$)
  for $\ell \leftarrow 0$ **to** $n-1$ **do**
    $\{(\xi_{N,i}^{(\ell+1)}, \omega_{N,i}^{(\ell+1)})\}_{i=1}^{M_N^{\ell+1}} \leftarrow$MPEAR($\{(\xi_{N,i}^{(\ell)}, \omega_{N,i}^{(\ell)})\}_{i=1}^{M_N^{\ell}}, R_\ell, \Phi_\ell, \varphi_\ell, \kappa_\ell$);
  end for
  return $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N^n}$
end procedure

In Step (1) of Algorithm 4.3.3, the pilot sample is produced by simply mutating the ancestor particles one step. By computing the CV of the pilot sample importance weights, an estimate of the CSD between the target and proposal distributions is obtained; if this estimated CSD is small, i.e., below the prespecified threshold κ, the pilot sample itself is returned by the algorithm (Step (9)). On the contrary, if the estimated CSD is large (above κ), the algorithm is reverted one step back, whereupon the ancestor particles are resampled (Step (4)) according to AMWs given by the pilot sample weights. Like in standard SMC schemes, the purpose of the selection step is here to duplicate particles located in the support of the target; however, by executing selection *before* the particle system collapses and including, via the AMWs, information about the subsequent measure in the distribution flow, the instrumental distribution of the filter is *redesigned* adaptively. In addition, when the selection operation in activated, the particle sample size may, in order to parry discrepancies between the location of the particles and the location of the mode of the target, be increased by a factor depending on the estimated CSD (Step (3)). After refueling, the particles are moved (mutated) according to the instrumental kernel as usual and associated with weights being inversely proportional to the pilot exploration-based AMWs (Steps (5–7)).

In the case of the standard SISAS scheme with $\kappa_\ell \equiv 0$, where selection is performed systematically at each timestep, Straka and Simandl (2006) used a refueling-like operation similar to that in step (3) in order to determine the input sample size $M_N^{\ell+1}$ at each iteration: Denote, for any $m \in \mathbb{N}^*$,

$$\mathrm{ESS}_m(\lambda, \eta) := m \frac{1}{1 + d_{\chi^2}(\eta || \lambda)} \,. \tag{4.3.14}$$

the *efficient sample size* (ESS) associated with the problem of sampling η using λ as instrumental distribution (or vice versa). The ESS was proposed by Kong et al. (1994) as a tool for describing the number of draws needed from the target distribution η itself in order to produce an MC estimate with the same quality as an IS-estimate based on m draws from an instrumental distribution λ. Note that $\mathrm{ESS}_m(\lambda, \eta)$ is always smaller than m. Standing at time ℓ, a natural approach for designing the size $M_N^{\ell+1}$ of the particle sample at the next timestep is to assure that $\mathrm{ESS}_{M_N^{\ell+1}}(\pi_{\mathrm{aux}}^{N,\ell+1}, \mu_{\mathrm{aux}}^{N,\ell+1})$ equals, say, the size M_N^ℓ of the previous sample, yielding

$$M_N^{\ell+1} = \left\lceil M_N^\ell \varphi\{ d_{\chi^2}(\pi_{\mathrm{aux}}^{N,\ell+1} || \mu_{\mathrm{aux}}^{N,\ell+1}) \} \right\rceil \,,$$

with $\varphi(x) = (1 + x)$, $x \in \mathbb{R}^+$. A problem with this approach is of course that the CSD lacks closed-form expression. Thus, in the work in question the authors use an MC approach of roughly *quadratic* (in the number of particles) complexity to estimate the desired ESS. However, combining Theorem 4.3.2 with the continuous mapping theorem gives immediately, as $N \to \infty$,

$$\left| \frac{\left\lceil M_N^\ell \varphi\{ d_{\chi^2}(\pi_{\mathrm{aux}}^{N,\ell+1} || \mu_{\mathrm{aux}}^{N,\ell+1}) \} \right\rceil}{\left\lceil M_N^\ell \varphi\{ \mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(\ell+1)}\}_{i=1}^{M_N^\ell}) \} \right\rceil} - 1 \right| \xrightarrow{\mathbb{P}} 0 \,,$$

from which we conclude that the SISAR algorithm can, for large sample sizes, be used for approximating *linearly* and adaptively the necessary sample size with arbitrary accuracy. A full simulation study of the MPEAR algorithm is beyond the scope of this chapter and is left as future work. However, the convergence of the algorithm under consideration is established in the next section.

Convergence of the MPEAR scheme

Theorem 4.3.3 (Consistency and asymptotic normality of Algorithm 4.3.3)**.** *Let the assumptions of Theorem 4.3.1 hold for the adjustment multiplier weight kernel* $\Psi^*(\xi, A) := R(\xi, \Phi^{\leftarrow}(\xi, \alpha \cdot A))$ *with* $\alpha \in \mathbb{N}$*. Then the weighted sample* $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ *returned by Algorithm 4.3.3 is consistent and asymptotically normal for* $(\mu, \tilde{\mathsf{C}})$ *and* $(\mu, \bar{\mathsf{A}}, \bar{\mathsf{W}}, \bar{\sigma}, \bar{\gamma}, \{a_N\}_{N=1}^{\infty})$*, respectively, where* $\bar{\mathsf{A}}$ *and* $\bar{\mathsf{W}}$ *are defined in* (4.3.1)*, and*

$$\breve{\sigma}^2(f) := \frac{\sigma^2\{L[\cdot, f-\mu(f)]\} + \breve{\varepsilon}\gamma R(\{\Phi[f-\mu(f)] - R(\cdot, \Phi[f-\mu(f)])\}^2)}{[\nu L(\tilde{\boldsymbol{\Xi}})]^2} + \beta(1-\breve{\varepsilon})\frac{\nu(R\{\cdot, \Phi^2[f-\mu(f)]^2\}\Psi^*(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1}))}{\varphi(\eta)\nu L(\tilde{\boldsymbol{\Xi}})}, \quad f \in \bar{\mathsf{A}}, \tag{4.3.15}$$

with $\eta := \gamma L(\Phi)/\{\beta[\nu L(\tilde{\boldsymbol{\Xi}})]^2\} - 1$*,* $\breve{\varepsilon} := \mathbb{1}_{\{\eta<\kappa\}}$*, and*

$$\breve{\gamma}(f) := \beta(1-\breve{\varepsilon})\frac{\nu[R(\cdot, \Phi^2 f)\Psi^*(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi(\eta)\nu L(\tilde{\boldsymbol{\Xi}})} + \breve{\varepsilon}\frac{\gamma R(\Phi^2 f)}{[\nu L(\tilde{\boldsymbol{\Xi}})]^2}, \quad f \in \bar{\mathsf{W}}. \tag{4.3.16}$$

Proof. The properness of $\bar{\mathsf{A}}$ and $\bar{\mathsf{W}}$ is checked straightforwardly. We thus turn to the consistency and let η be defined as in the theorem and set, for $N \in \mathbb{N}$, $\beta_N := \lceil M_N \varphi(\eta) \rceil$. Pick $f \in \tilde{\mathsf{C}}$ and express the final estimate of $\mu(f)$ returned by algorithm as

$$\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})<\kappa\}}\bar{\Omega}_N^{-1}\sum_{i=1}^{M_N}\bar{\omega}_{N,i}f(\bar{\xi}_{N,i}) + \mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}\tilde{\Omega}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i}). \tag{4.3.17}$$

Since the pilot sample $\{(\bar{\xi}_{N,i}, \bar{\omega}_{N,i})\}_{i=1}^{M_N}$ is consistent and asymptotically normal under **(A8)** and **(A9)**, we obtain by the limit theorems for the mutation step from Douc and Moulines (2008) recalled in Appendix A, Theorems A.2.2 and A.2.2, as $N \to \infty$,

$$\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N}) \xrightarrow{\mathbb{P}} \eta, \quad \tilde{M}_N \xrightarrow{\mathbb{P}} \beta_N, \quad \bar{\Omega}_N^{-1}\sum_{i=1}^{M_N}\bar{\omega}_{N,i}f(\bar{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \mu(f). \tag{4.3.18}$$

Applying Slutsky's theorem yields immediately

$$\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})<\kappa\}}\bar{\Omega}_N^{-1}\sum_{i=1}^{M_N}\bar{\omega}_{N,i}f(\bar{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \mathbb{1}_{\{\eta<\kappa\}}\mu(f). \tag{4.3.19}$$

To treat the second term of (4.3.17), we write, for a given $\epsilon > 0$,

$$\mathbb{P}\left(\left|\tilde{M}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i}) - \mu(f)\right| \geq \epsilon\right) \leq \mathbb{P}\left(\left|\beta_N^{-1}\sum_{i=1}^{\beta_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i}) - \mu(f)\right| \geq \epsilon\right) + \mathbb{P}\left(\tilde{M}_N \neq \beta_N\right) \to 0, \tag{4.3.20}$$

where the limit follows by (4.2.6) with $\Psi(\xi, \mathbf{1}_{\mathbb{R}^+}) = \Psi^*(\xi, \mathbf{1}_{\mathbb{R}^+}) = L(\xi, \tilde{\boldsymbol{\Xi}})$. Thus, the quantity $\tilde{M}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})$ tends to $\mu(f)$ in probability. Using Slutsky's theorem

and the properness of $\tilde{\mathrm{C}}$ (implying that the constant function $f \equiv 1$ belongs to $\tilde{\mathrm{C}}$), this implies the limit

$$\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}\tilde{\Omega}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})$$
$$= \mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}(\tilde{M}_N^{-1}\tilde{\Omega}_N)^{-1}\tilde{M}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \mathbb{1}_{\{\eta\geq\kappa\}}\mu(f)\ . \quad (4.3.21)$$

Combining (4.3.19) and (4.3.21) shows that (4.3.17) tends to $\mu(f)$ as N tends to infinity, which is property *i)* of definition 4.2.1. We show property *ii)*; using (4.2.11) and repeating the arguments in (4.3.20) yields

$$\tilde{M}_N^{-1}\max_{1\leq i\leq\tilde{M}_N}\tilde{\omega}_{N,i} \xrightarrow{\mathbb{P}} 0\ ,$$

implying that also $\tilde{\Omega}_N^{-1}\max_{1\leq i\leq\tilde{M}_N}\tilde{\omega}_{N,i}$ vanishes as N tends to infinity. In addition, by consistency of the mutation operation (Theorem A.2.1), $\bar{\Omega}_N^{-1}\max_{1\leq i\leq M_N}\bar{\omega}_{N,i}$ tends to zero in probability. Property *ii)* follows.

We turn to asymptotic normality and assume without loss of generality that $\mu(f)=0$. To treat the case $\eta\geq\kappa$, in which the first part of (4.3.17) tends to zero in probability, we recall the σ-field $\mathcal{F}_{N,0}=\sigma(\{(\xi_{N,\ell},\omega_{N,\ell},\bar{\omega}_{N,\ell})\}_{\ell=1}^{M_N})$ and show that, for any $u\in\mathbb{R}$, as $N\to\infty$,

$$a_N\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}\tilde{\Omega}_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})$$
$$\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\beta\frac{\nu[R(\cdot,\Phi^2f^2)\Psi^*(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi(\eta)\nu L(\tilde{\Xi})}+\frac{\sigma^2[L(\cdot,f)]}{[\nu L(\tilde{\Xi})]^2}\right)\ . \quad (4.3.22)$$

Indeed, using the bound $|\exp(\mathrm{i}\,a)-\exp(\mathrm{i}\,b)|\leq|a-b|$, yields

$$\left|\mathbb{E}\left[\exp\left(\mathrm{i}\,ua_N\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}\beta_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})\right)\right.\right.$$
$$\left.\left.-\exp\left(\mathrm{i}\,ua_N\beta_N^{-1}\sum_{i=1}^{\beta_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})\right)\middle|\mathcal{F}_{N,0}\right]\right|$$
$$\leq ua_N\mathbb{1}_{\{\tilde{M}_N\neq\beta_N\}}\mathbb{E}\left[\left|\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})\geq\kappa\}}\beta_N^{-1}\sum_{i=1}^{\tilde{M}_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})-\beta_N^{-1}\sum_{i=1}^{\beta_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})\right|\middle|\mathcal{F}_{N,0}\right]$$
$$+ua_N\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N})<\kappa\}}\mathbb{E}\left[\left|\beta_N^{-1}\sum_{i=1}^{\beta_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})\right|\middle|\mathcal{F}_{N,0}\right]\ ,$$

where the right hand side vanishes in probability as N grows, since the multiplicative indicator functions tend to zero (in probability). Since

$$\mathbb{E}\left[\exp\left(\mathrm{i}\,ua_N\beta_N^{-1}\sum_{i=1}^{\beta_N}\tilde{\omega}_{N,i}f(\tilde{\xi}_{N,i})\right)\middle|\mathcal{F}_{N,0}\right]$$
$$\xrightarrow{\mathbb{P}}\exp\left(-\frac{u^2}{2}\left\{\beta\frac{\nu[R(\cdot,\Phi^2f^2)\Psi^*(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi(\eta)\nu L(\tilde{\Xi})}+\frac{\sigma^2[L(\cdot,f)]}{[\nu L(\tilde{\Xi})]^2}\right\}\right) \quad (4.3.23)$$

and $\tilde{\Omega}_N \beta_N^{-1}$ tends to unity in probability, the weak convergence (4.3.22) follows. We turn to the case $\eta < \kappa$ in which the second part of (4.3.17) tends to zero in probability. By combining Theorem A.2.2 and the limit (4.3.18) we get, since $\nu L(f) = 0$ by assumption,

$$a_N \bar{\Omega}_N^{-1} \sum_{i=1}^{M_N} \bar{\omega}_{N,i} f(\bar{\xi}_{N,i}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2[L(\cdot, f)] + \gamma R\{[\Phi f - R(\cdot, \Phi f)]^2\}}{[\nu L(\tilde{\boldsymbol{\Xi}})]^2}\right) . \tag{4.3.24}$$

Using jointly (4.3.22) and (4.3.24) provides, via the dominated convergence theorem,

$$\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N}) < \kappa\}} \bar{\Omega}_N^{-1} \sum_{i=1}^{M_N} \bar{\omega}_{N,i} f(\bar{\xi}_{N,i}) + \mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N}) \geq \kappa\}} \tilde{\Omega}_N^{-1} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i} f(\tilde{\xi}_{N,i})$$
$$\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \check{\sigma}^2(f)\right) ,$$

where $\check{\sigma}^2$ is defined in (4.3.15). This establishes property *i)* in Definition 4.2.2.

Now, pick $f \in \bar{\mathrm{W}}$; to establish property *ii)* in Definition 4.2.2 we apply argue as in (4.3.20) and apply Theorem 4.2.2, yielding

$$a_N^2 \tilde{\Omega}_N^{-2} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i}^2 f(\tilde{\xi}_{N,i}) \xrightarrow{\mathbb{P}} \beta \frac{\nu[R(\cdot, \Phi^2 f)\Psi^*(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi(\eta)\nu L(\tilde{\boldsymbol{\Xi}})} .$$

By combining this with Theorem A.2.2 we obtain straightforwardly

$$\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N}) < \kappa\}} a_N^2 \bar{\Omega}_N^{-2} \sum_{i=1}^{M_N} \bar{\omega}_{N,i}^2 f(\bar{\xi}_{N,i}) + \mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,\ell}\}_{\ell=1}^{M_N}) \geq \kappa\}} a_N^2 \tilde{\Omega}_N^{-2} \sum_{i=1}^{\tilde{M}_N} \tilde{\omega}_{N,i}^2 f(\tilde{\xi}_{N,i})$$
$$\xrightarrow{\mathbb{P}} \mathbb{1}_{\{\eta < \kappa\}} \frac{\gamma R(\Phi^2 f)}{[\nu L(\tilde{\boldsymbol{\Xi}})]^2} + \mathbb{1}_{\{\eta \geq \kappa\}} \beta \frac{\nu[R(\cdot, \Phi^2 f)\Psi^*(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi(\eta)\nu L(\tilde{\boldsymbol{\Xi}})} = \check{\gamma}(f) ,$$

where $\check{\gamma}$ is defined in (4.3.16).

From the last part of the proof of Theorem 4.2.2 we obtain, repeating the arguments of (4.3.20), the limit

$$a_N \tilde{\Omega}_N^{-1} \max_{1 \leq i \leq \tilde{M}_N} \tilde{\omega}_{N,i} \xrightarrow{\mathbb{P}} 0 ,$$

and combining this with the uniform asymptotic smallness of $\{a_N \bar{\omega}_{N,\ell} \bar{\Omega}_N^{-1}\}_{\ell=1}^{M_N}$ provided by Theorem A.2.2 establishes property *iii)*. This completes the proof. □

In order to illustrate theoretically the advantage of adaptive refueling approach vis-à-vis the standard approach with adaptive selection, we return to the sequential context with a flow $\{\mu_\ell\}_{\ell=0}^{\infty}$ of distributions generated by (4.2.2) and consider the SISAR algorithm. In particular, we wish show that the SISAR scheme is more foresighted than SISAS scheme in the sense that it adjusts adaptively the instrumental distribution to the target distribution *before* the particle sample may have degenerated; recall the discussion in Section 4.3.2. Another point of interest is how that adaptively increased particle sample size effects the asymptotic variance of the produced estimates. Hence, let again $\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$ be an initial sample satisfying **(A8)** and **(A9)** for some initial parameters (μ_0, C_0) and $(\mu_0, \mathrm{A}_0, \mathrm{W}_0, \sigma_0, \gamma_0, \{a_N\}_{N=1}^{\infty})$, respectively. In addition, set $\beta_0 := 1$ and define recursively

$$\beta_{\ell+1} := \frac{\beta_\ell}{\check{\varepsilon}_\ell + (1 - \check{\varepsilon}_\ell)\varphi_\ell(\eta_\ell)} ,$$

with

$$\eta_\ell := \frac{\gamma_\ell L_\ell(\Phi_\ell)}{\beta_\ell[\mu_\ell L_\ell(\Xi_{\ell+1})]^2} - 1 \;,$$
$$\check{\varepsilon}_\ell := \mathbb{1}_{\{\eta_\ell < \kappa_\ell\}} \;,$$

and

$$\check{\gamma}_{\ell+1}(f) := \beta_\ell(1-\check{\varepsilon}_\ell)\frac{\mu_\ell[R_\ell(\cdot,\Phi_\ell^2 f)\Psi_\ell^*(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi_\ell(\eta_\ell)\mu_\ell L_\ell(\Xi_{\ell+1})} + \check{\varepsilon}_\ell \frac{\check{\gamma}_\ell R_\ell(\Phi_\ell^2 f)}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2} \;, \quad f \in \bar{\mathsf{W}}_{\ell+1} \;. \tag{4.3.25}$$

Here the $\bar{\mathsf{W}}_\ell$'s are defined through (4.3.1). Moreover, define the asymptotic variances

$$\check{\sigma}^2_{\ell+1}(f) := \frac{\check{\sigma}^2_\ell\{L_\ell[\cdot, f-\mu_{\ell+1}(f)]\} + \check{\varepsilon}_\ell\gamma_\ell R_\ell(\{\Phi_\ell[f-\mu_{\ell+1}(f)] - R_\ell(\cdot,\Phi_\ell[f-\mu_{\ell+1}(f)])\}^2)}{[\mu_\ell L_\ell(\Xi_{\ell+1})]^2}$$
$$+ \beta_\ell(1-\check{\varepsilon}_\ell)\frac{\mu_\ell(R_\ell\{\cdot,\Phi_\ell^2[f-\mu_{\ell+1}(f)]^2\}\Psi_\ell^*(\cdot,\mathbf{1}_{\mathbb{R}^+}^{-1}))}{\varphi_\ell(\eta_\ell)\mu_\ell L_\ell(\Xi_{\ell+1})} \;, \quad f \in \bar{\mathsf{A}}_{\ell+1} \;. \tag{4.3.26}$$

In addition, let $\check{\mathrm{s}}(n) := \min\{\ell < n : \eta_\ell \geq \kappa_\ell\}$, $n \in \mathbb{N}$.

$$\check{\mu}^{\mathrm{aux}}_{N,n}(\{i\}\times A) := \frac{\omega_{N,i}^{(\check{\mathrm{s}}(n))} L_{\check{\mathrm{s}}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\check{\mathrm{s}}(n))}, \Xi_n)}{\sum_{j=1}^{M_N}\omega_{N,j}^{(\check{\mathrm{s}}(n))} L_{\check{\mathrm{s}}(n)}\cdots L_{n-1}(\xi_{N,j}^{(\check{\mathrm{s}}(n))}, \Xi_n)} \left[\frac{\bigotimes_{\ell=\check{\mathrm{s}}(n)}^{n-1} L_\ell(\xi_{N,i}^{(\check{\mathrm{s}}(n))}, A)}{L_{\check{\mathrm{s}}(n)}\cdots L_{n-1}(\xi_{N,i}^{(\check{\mathrm{s}}(n))}, \Xi_n)}\right] \tag{4.3.27}$$

is estimated using the instrumental distribution

$$\check{\pi}^{\mathrm{aux}}_{N,n}(\{i\}\times A) := \frac{\omega_{N,i}^{(\check{\mathrm{s}}(n))}\bar{\omega}_{N,i}^{(\check{\mathrm{s}}(n))}}{\sum_{j=1}^{M_N}\omega_{N,j}^{(\check{\mathrm{s}}(n))}\bar{\omega}_{N,j}^{(\check{\mathrm{s}}(n))}} \bigotimes_{\ell=\check{\mathrm{s}}(n)}^{n-1} R_\ell(\xi_{N,i}^{(\check{\mathrm{s}}(n))}, A) \;.$$

We then have the following result.

Corollary 4.3.1 (Convergence of Algorithm 4.3.4)**.** *Assume* **(A11)** *and suppose that the sample* $\{(\xi_{N,i}^{(0)}, \omega_{N,i}^{(0)})\}_{i=1}^{M_N}$ *satisfies* **(A8)** *and* **(A9)** *for* (μ_0, C_0) *and* $(\mu_0, \mathsf{A}_0, \mathsf{W}_0, \sigma_0, \gamma_0, \{M_N^{1/2}\}_{N=1}^\infty)$*, respectively. In addition, let* $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$*,* $n \in \mathbb{N}$*, the be the sample returned by Algorithm 4.3.4 with input parameters* $(\check{\Phi}_\ell, \check{\Psi}_\ell)$*,* $\ell \in \{0,\ldots,n-1\}$*, satisfying* **(A10)** *for all* ℓ*. Finally, suppose that* $L_\ell(\cdot, \Xi_{\ell+1}) \in \mathsf{C}_\ell$ *for all* ℓ*. Then the following holds.*

i) *The output* $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N}$ *is consistent and asymptotically normal for* (μ_n, C_n) *and* $(\mu_n, \mathsf{A}_n, \mathsf{W}_n, \check{\sigma}_n, \check{\gamma}_n, \{M_N^{1/2}\}_{N=1}^\infty)$ *where* $(\mathsf{C}_n, \mathsf{A}_n, \mathsf{W}_n)$*,* $\check{\sigma}_n$*, and* $\check{\gamma}_n$ *are defined recursively via* (4.3.26) *and* (4.3.25)*, respectively.*

ii)
$$\mathbb{E}\left[d_{\chi^2}(\check{\mu}^{\mathrm{aux}}_{N,n+1}||\check{\pi}^{\mathrm{aux}}_{N,n+1})\,\middle|\,\mathcal{F}_N^{\check{\mathrm{s}}(n)}\right] \xrightarrow{\mathbb{P}} \eta_n \;, \quad N\to\infty$$

and

$$\mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(n+1)}\}_{i=1}^{M_N^n}) \xrightarrow{\mathbb{P}} \eta_n \;, \quad N\to\infty \;.$$

Proof. We proceed by induction. Thus, suppose that the statement is true for $n \in \mathbb{N}$, i.e., $\{(\xi_{N,i}^{(n)}, \omega_{N,i}^{(n)})\}_{i=1}^{M_N^n}$ is asymptotically normal for (μ_n, C_n) and $(\mu_n, \mathsf{A}_n, \mathsf{W}_n, \check{\sigma}_n, \check{\gamma}_n, \{M_N^{1/2}\}_{N=1}^\infty)$. In addition, suppose that $M_N^n M_N^{-1}$ converges to β_n^{-1} in probability as N tends to infinity. Then $\{(\xi_{N,i}^{(n+1)}, \omega_{N,i}^{(n+1)})\}_{i=1}^{M_N^{n+1}}$ is consistent and asymptotically normal for $(\mu_{n+1}, \mathsf{C}_{n+1})$ and

$(\mu_{n+1}, \mathsf{A}_{n+1}, \mathsf{W}_{n+1}, \check{\sigma}_{n+1}, \check{\gamma}_{n+1}, \{M_N^{1/2}\}_{N=1}^{\infty})$ by Theorem 4.3.3. Moreover,

$$\begin{aligned}
&\mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(n+1)}\}_{i=1}^{M_N^n}) = M_N^n(\bar{\Omega}_N n + 1)^{-2} \sum_{i=1}^{M_N^n} (\bar{\omega}_{N,i}^{(n+1)})^2 - 1 \\
&= (M_N^n M_N^{-1}) M_N (\bar{\Omega}_N n + 1)^{-2} \sum_{i=1}^{M_N^n} (\bar{\omega}_{N,i}^{(n+1)})^2 - 1 \xrightarrow{\mathbb{P}} \frac{\check{\gamma}_n L_n(\Phi_n)}{\beta_n [\mu_n L_n(\boldsymbol{\Xi}_{n+1})]^2} - 1 = \eta_n \,, \quad (4.3.28)
\end{aligned}$$

where the limit follows from the induction hypothesis and the expression of $\tilde{\gamma}$ in Theorem A.2.2. On the other hand, by the definition of $\check{\mathrm{s}}(n)$ it holds that

$$\begin{aligned}
&\frac{\check{\gamma}_n R_n(\Phi_n^2)}{\beta_n [\mu_n L_n(\boldsymbol{\Xi}_{n+1})]^2} = \frac{\check{\gamma}_{n-1} R_{n-1} \otimes R_n(\Phi_{n-1,n}^2)}{\beta_{n-1} [\mu_{n-1} L_{n-1}(\boldsymbol{\Xi}_n)]^2 [\mu_n L_n(\boldsymbol{\Xi}_{n+1})]^2} \\
&= \cdots = \frac{\check{\gamma}_{\check{\mathrm{s}}(n)+1} \bigotimes_{\ell=\check{\mathrm{s}}(n)+1}^n R_\ell(\Phi_{\check{\mathrm{s}}(n)+1,n}^2)}{\beta_{\check{\mathrm{s}}(n)+1} \prod_{\ell=\check{\mathrm{s}}(n)+1}^n [\mu_\ell L_\ell(\boldsymbol{\Xi}_{\ell+1})]^2} = \frac{\check{\gamma}_{\check{\mathrm{s}}(n)+1} \bigotimes_{\ell=\check{\mathrm{s}}(n)+1}^n R_\ell(\Phi_{\check{\mathrm{s}}(n)+1,n}^2)}{\beta_{\check{\mathrm{s}}(n)+1} [\mu_{\check{\mathrm{s}}(n)+1} L_{\check{\mathrm{s}}(n)+1} \cdots L_n(\boldsymbol{\Xi}_{n+1})]^2} \,. \quad (4.3.29)
\end{aligned}$$

Now, plugging the expression

$$\check{\gamma}_{\check{\mathrm{s}}(n)+1}(f) = \frac{\beta_{\check{\mathrm{s}}(n)} \mu_{\check{\mathrm{s}}(n)} [R_{\check{\mathrm{s}}(n)}(\cdot, \Phi_{\check{\mathrm{s}}(n)}^2 f) \Psi_{\check{\mathrm{s}}(n)}(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{\varphi_{\check{\mathrm{s}}(n)}(\eta_{\check{\mathrm{s}}(n)}) \mu_{\check{\mathrm{s}}(n)} L_{\check{\mathrm{s}}(n)}(\boldsymbol{\Xi}_{\check{\mathrm{s}}(n)+1})} \,, \quad f \in \mathsf{W}_{\check{\mathrm{s}}(n)+1} \,,$$

into (4.3.29) and applying the identity $\beta_{\check{\mathrm{s}}(n)} = \beta_{\check{\mathrm{s}}(n+1)} \varphi_{\check{\mathrm{s}}(n)}(\eta_{\check{\mathrm{s}}(n)})$ yields

$$\eta_n = \frac{\check{\gamma}_n R_n(\Phi_n^2)}{\beta_n [\mu_n L_n(\boldsymbol{\Xi}_{n+1})]^2} - 1 = \frac{\mu_{\check{\mathrm{s}}(n)} \Psi_{\check{\mathrm{s}}(n)}^*(\mathbf{1}_{\mathbb{R}^+}) \mu_{\check{\mathrm{s}}(n)} [\bigotimes_{\ell=\check{\mathrm{s}}(n)}^n R_\ell(\cdot, \Phi_{\check{\mathrm{s}}(n),n}^2) \Psi_{\check{\mathrm{s}}(n)}^*(\cdot, \mathbf{1}_{\mathbb{R}^+}^{-1})]}{[\mu_{\check{\mathrm{s}}(n)} L_{\mathrm{s}(n)} \cdots L_n(\boldsymbol{\Xi}_{n+1})]^2} - 1 \,. \quad (4.3.30)$$

However, by simply replacing $\mathrm{s}(n)$ and $\Psi_{\mathrm{s}(n)}$ by $\check{\mathrm{s}}(n)$ and $\Psi_{\check{\mathrm{s}}(n)}^*$, respectively, in the limit (4.3.13) we conclude that

$$\mathbb{E}\left[d_{\chi^2}(\check{\mu}_{N,n+1}^{\mathrm{aux}} || \check{\pi}_{N,n+1}^{\mathrm{aux}}) \middle| \mathcal{F}_N^{\check{\mathrm{s}}(n)} \right] \xrightarrow{\mathbb{P}} \eta_n \,.$$

Finally, since also

$$\begin{aligned}
&M_N^{n+1} M_N^{-1} \\
&= M_N^{-1} M_N^n \left(\mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(n+1)}\}_{i=1}^{M_N^n}) < \kappa_n\}} + \varphi_n \left[\mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(n+1)}\}_{i=1}^{M_N^n}) \right] \mathbb{1}_{\{\mathrm{CV}^2(\{\bar{\omega}_{N,i}^{(n+1)}\}_{i=1}^{M_N^n}) \geq \kappa_n\}} \right) \\
&\qquad \xrightarrow{\mathbb{P}} \beta_n^{-1} \left(\mathbb{1}_{\{\eta_n < \kappa_n\}} + \mathbb{1}_{\{\eta_n \geq \kappa_n\}} \varphi_n(\eta_n) \right) = \beta_{n+1}^{-1} \,,
\end{aligned}$$

we conclude the proof by establishing that the induction hypothesis is true for $n = 0$ and $\beta_0 = 1$. □

We conclude this section with a few remarks on this result. Since η_ℓ equals the value of the pilot sample CV at any time step in the asymptotic regime, $\check{\mathrm{s}}(n)$ is the last time step, standing at time n, selection was occurred when N is large. At all time steps between $\check{\mathrm{s}}(n)$ and n the particle cloud evolves by means of standard subsequent mutation operations. This way of propagating the particles is captured by the mixture $\check{\pi}_{N,n}^{\mathrm{aux}}$, which consequently can be interpreted as the instrumental distribution of the filter for large particle population sizes. Optimally, we would use the target distribution $\check{\mu}_{N,n}^{\mathrm{aux}}$ obtained by simply plugging the particle approximation at time $\check{\mathrm{s}}(n)$ into the multi-step Feynman-Kac transition formula.

In order to investigate the impact of the adaptive refueling operation on the asymptotic variance of the final Monte Carlo estimates, define, for $\xi \in \mathbf{\Xi}_m$ and $f \in \mathbb{B}(\mathbf{\Xi}_n)$,

$$\lambda_{m,n}(\xi, f) := \begin{cases} R_m(\xi, \{\Phi_m L_{m+1} \cdots L_{n-1}[\cdot, f - \mu_n(f)] - R_m(\xi, \Phi_m L_{m+1} \cdots L_{n-1}[\cdot, f - \mu_n(f)])\}^2) , \\ \qquad \text{for } \check{\mathrm{s}}(n) < m \leq n-1 , \\ R_{\check{\mathrm{s}}(n)}(\xi, \Phi^2_{\check{\mathrm{s}}(n)}\{L_{\check{\mathrm{s}}(n)+1} \cdots L_{n-1}[\cdot, f - \mu_n(f)]\}^2) , \qquad \text{for } m = \check{\mathrm{s}}(n) . \end{cases}$$

We now have the following alternative expressions of the asymptotic variances.

Corollary 4.3.2. *Let the assumptions of Corollary 4.3.1 hold and let $\{\check{\sigma}^2_\ell\}_{\ell=0}^\infty$ be the generated recursively through (4.3.26). Then, for any $n \in \mathbb{N}$ and $f \in \mathsf{A}_n$,*

$$\check{\sigma}^2_n(f) = \frac{\check{\sigma}^2_{\check{\mathrm{s}}(n)}\{L_{\check{\mathrm{s}}(n)} \cdots L_{n-1}[\cdot, f - \mu_n(f)]\}}{[\mu_{\check{\mathrm{s}}(n)} L_{\check{\mathrm{s}}(n)} \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} + \frac{\beta_{\check{\mathrm{s}}(n)} \mu_{\check{\mathrm{s}}(n)} \Psi^*_{\check{\mathrm{s}}(n)}(\mathbf{1}_{\mathbb{R}^+}) \sum_{m=\check{\mathrm{s}}(n)}^{n-1} \mu_{\check{\mathrm{s}}(n)}[\bigotimes_{\ell=\check{\mathrm{s}}(n)}^{m-1} R_\ell(\cdot, \Phi^2_{\check{\mathrm{s}}(n),m-1} \lambda_{m,n}(f)) \Psi^*_{\check{\mathrm{s}}(n)}(\cdot, \mathbf{1}^{-1}_{\mathbb{R}^+})]}{\varphi_{\check{\mathrm{s}}(n)}(\eta_{\check{\mathrm{s}}(n)})[\mu_{\check{\mathrm{s}}(n)} L_{\check{\mathrm{s}}(n)} \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} . \tag{4.3.31}$$

Proof. As a direct consequence of (4.3.26), the statement is true for $\check{\mathrm{s}}(n) = n-1$. Thus, we consider the case $\check{\mathrm{s}}(n) < n-1$ and proceed by induction: let $\check{\mathrm{s}}(n) + 1 < \ell \leq n$ and assume that

$$\check{\sigma}^2_n(f) = \frac{\check{\sigma}^2_\ell\{L_\ell \cdots L_{n-1}[\cdot, f - \mu_n(f)]\}}{[\mu_\ell L_\ell \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} + \sum_{m=\ell}^{n-1} \frac{\check{\gamma}_m \lambda_{m,n}(f)}{[\mu_m L_m \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} . \tag{4.3.32}$$

Since $\mu_\ell L_\ell \cdots L_{n-1}[f - \mu_n(f)] = 0$, it holds by (4.3.26) and the definition of $\check{\mathrm{s}}(s)$ that

$$\begin{aligned} \frac{\check{\sigma}^2_\ell\{L_\ell \cdots L_{n-1}[\cdot, f - \mu_n(f)]\}}{[\mu_\ell L_\ell \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} &= \frac{\check{\sigma}^2_{\ell-1}\{L_{\ell-1} \cdots L_{n-1}[f - \mu_n(f)]\}}{[\mu_{\ell-1} L_{\ell-1}(\mathbf{\Xi}_\ell)]^2 [\mu_\ell L_\ell \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} \\ &\quad + \frac{\check{\gamma}_{\ell-1} R_{\ell-1}(\{\Phi_{\ell-1} L_\ell \cdots L_{n-1}[f - \mu_n(f)] - R_\ell(\cdot, \Phi_{\ell-1} L_\ell \cdots L_{n-1}[f - \mu_n(f)])\}^2)}{[\mu_{\ell-1} L_{\ell-1}(\mathbf{\Xi}_\ell)]^2 [\mu_\ell L_\ell \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} \\ &= \frac{\check{\sigma}^2_{\ell-1}\{L_{\ell-1} \cdots L_{n-1}[f - \mu_n(f)]\}}{[\mu_{\ell-1} L_{\ell-1} \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} + \frac{\check{\gamma}_{\ell-1} \lambda_{\ell-1,n}(f)}{[\mu_{\ell-1} L_{\ell-1} \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} , \end{aligned}$$

from which it follows that (4.3.32) still holds when replacing ℓ by $\ell - 1$. As the induction hypothesis is also trivially true for $\ell = n$, we have proved that

$$\check{\sigma}^2_n(f) = \frac{\check{\sigma}^2_{\check{\mathrm{s}}(n)+1}\{L_{\check{\mathrm{s}}(n)+1} \cdots L_{n-1}[\cdot, f - \mu_n(f)]\}}{[\mu_{\check{\mathrm{s}}(n)+1} L_{\check{\mathrm{s}}(n)+1} \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} + \sum_{m=\check{\mathrm{s}}(n)+1}^{n-1} \frac{\check{\gamma}_m \lambda_{m,n}(f)}{[\mu_m L_m \cdots L_{n-1}(\mathbf{\Xi}_n)]^2} . \tag{4.3.33}$$

On the other hand, by equation (4.3.9) it holds, for any $\check{\mathrm{s}}(n) < m \leq n$ and $h \in \mathsf{W}_m$, that

$$\check{\gamma}_m(h) = \frac{\beta_{\check{\mathrm{s}}(n)} \mu_{\check{\mathrm{s}}(n)} \Psi^*_{\check{\mathrm{s}}(n)}(\mathbf{1}_{\mathbb{R}^+}) \mu_{\check{\mathrm{s}}(n)}[\bigotimes_{\ell=\check{\mathrm{s}}(n)}^{m-1} R_\ell(\cdot, \Phi^2_{\check{\mathrm{s}}(n),m-1} h) \Psi^*_{\check{\mathrm{s}}(n)}(\cdot, \mathbf{1}^{-1}_{\mathbb{R}^+})]}{\varphi_{\check{\mathrm{s}}(n)}(\eta_{\check{\mathrm{s}}(n)})[\mu_{\check{\mathrm{s}}(n)} L_{\mathrm{s}(n)} \cdots L_{m-1}(\mathbf{\Xi}_m)]^2} .$$

Inserting these expressions of $\{\check{\gamma}_m\}_{m=\ell}^{n-1}$ into (4.3.33) yields the desired formula (4.3.31). □

The first term in (4.3.31) depends on the behavior, as described by the asymptotic variance $\check{\sigma}^2_{\check{\mathrm{s}}(n)}$, of the particle filter up to time $\check{\mathrm{s}}(n)$ only. The second term is, notably, inversely proportional to $\varphi_{\check{\mathrm{s}}(n)}(\eta_{\check{\mathrm{s}}(n)})$ which grows with the asymptotic CSD between $\check{\mu}^{\mathrm{aux}}_{N,n+1}$ and $\check{\pi}^{\mathrm{aux}}_{N,n+1}$. Thus, incorporating the refueling mechanism in the algorithm will indeed counteract, adaptively to the data, increases of variance caused by badly adjusted proposal distributions.

CHAPTER 5

Adaptation of the proposal kernel by mixture of experts

Contents

This chapter corresponds to an article which is to be submitted under the (tentative) name *Adapting the proposal kernel of sequential importance sampling by means of mixture of experts*, by J. Cornebise, E. Moulines, J. Olsson, 2009. As we focus more on methodology than on asymptotic results, we drop the index N in the random variables, as we will most often display algorithms for a fixed number N of particles.

5.1 Introduction

SMC methods refer to a class of algorithms designed for approximating a *sequence of probability distributions* by updating recursively in time a set of random *particles* with associated nonnegative weights. These algorithms are all based on *selection* and *mutation* operations and can thus be seen as combinations of the *sequential importance sampling* and *sampling importance resampling* methods introduced in Handschin and Mayne (1969) and Rubin (1987), respectively. SMC methods have emerged as a key tool to solve filtering and smoothing problems in general state space models, in rare events simulations, in computational chemistry (Liu, 2001; Doucet et al., 2001; Ristic et al., 2004; Del Moral, 2004; Cappé et al., 2005, and the reference therein).

Suppose that we are given a weighted sample $\{(\xi_i, \omega_i)\}_{i=1}^N$ targeting a probability density function ν. By targeting, we mean that, for any bounded measurable function f, $\Omega^{-1}\sum_{j=1}^N \omega_j f(\xi_j)$ approximates $\int f(\xi)\nu(\xi)\mathrm{d}\xi$. We wish to transform this sample into a new weighted particle sample approximating the probability measure

$$\mu(\xi) := \frac{\int \nu(\xi) l(\xi, \tilde{\xi})\mathrm{d}\xi}{\iint \nu(\xi) l(\xi, \tilde{\xi})\mathrm{d}\xi\mathrm{d}\tilde{\xi}} \,. \tag{5.1.1}$$

where l is a non-normalized transition density kernel. Starting with the sample $\{\xi_i, \omega_i\}_{i=1}^N$ targeting the distribution ν, a natural idea consists in approximating μ with the mixture distribution,

$$\hat{\mu}(\tilde{\xi}) := \frac{\sum_{i=1}^N \omega_i l(\xi_i, \tilde{\xi})}{\sum_{i=1}^N \omega_i l(\xi_i, \tilde{\xi})\mathrm{d}\tilde{\xi}} \,. \tag{5.1.2}$$

As suggested by Pitt and Shephard (1999), $\hat{\mu}$ may be seen as the marginal with respect to the particle index of the *auxiliary distribution* given by

$$\mu_{\mathrm{aux}}(i, \tilde{\xi}) = \frac{\omega_i l(\xi_i, \tilde{\xi})}{\sum_{j=1}^N \omega_j \int l(\xi_j, \tilde{\xi})\mathrm{d}\tilde{\xi}} = \frac{\omega_i \Psi^*(\xi_i)}{\sum_{j=1}^N \omega_j \Psi^*(\xi_j)} l^*(\xi_i, \tilde{\xi}) \tag{5.1.3}$$

where $\Psi^*(x)$ is the partition function of $l(x, \cdot)$ and l^* is the normalized (Markovian) transition density

$$\Psi^*(\xi) := \int l(\xi, \tilde{\xi})\mathrm{d}\tilde{\xi} \,, \quad \text{and} \quad l^*(\xi, \tilde{\xi}) := \frac{l(\xi, \tilde{\xi})}{\int l(\xi, \tilde{\xi})\mathrm{d}\tilde{\xi}} \,. \tag{5.1.4}$$

The decomposition (5.1.3) is largely of theoretical interest because either the computation of Ψ^* is difficult and/or sampling from $l^*(\xi_i, \tilde{\xi})$ is involved. We sample from $\hat{\mu}$ by **(1)** drawing a set $\{I_j\}_{j=1}^M$ of indices from the distribution on $\{1, \dots, N\}$ with weights proportional to $(\omega_1\Psi(\xi_1), \dots, \omega_N\Psi(\xi_N))$, where Ψ is the *adjustment weight function*, **(2)** sampling, conditionally to $\{I_j\}_{j=1}^M$ new particle positions $\{\tilde{\xi}_j\}_{j=1}^M$ using the *proposal density kernel* r from $\boldsymbol{\Xi}$ to $\tilde{\boldsymbol{\Xi}}$, and **(3)** computing their respective importance weights

$$\tilde{\omega}_j := \frac{1}{\Psi(\xi_{i_j})} \frac{l(\xi_{i_j}, \tilde{\xi}_j)}{r(\xi_{i_j}, \tilde{\xi}_j)} \,, \quad j = 1, \dots, N \,. \tag{5.1.5}$$

We finally discard the indices and keep $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^N$ as an approximation of μ. Equivalently, the procedure amounts to sample from the proposal distribution π_{aux} over the product space $\{1, \dots, N\} \times \tilde{\boldsymbol{\Xi}}$

$$\pi_{\mathrm{aux}}(i, \tilde{\xi}) := \frac{\omega_i \Psi(\xi_i)}{\sum_{j=1}^N \omega_j \Psi(\xi_j)} r(\xi_i, \tilde{\xi}) \tag{5.1.6}$$

and discard the indices.

The way to sample in the current set of particles and to propose new particles has a significant impact on the overall performances. This latter concern was already present in the pioneering paper Gordon et al. (1993) with the so-called *prior editing*, which aimed, in the state-space models context, at incorporating the next observation in the proposal by means of accept/reject algorithm. Many techniques to construct better have been investigated so far.

Doucet et al. (2000) suggest to approximate the transition kernel $l(\xi_j, \cdot)$ in the neighborhood of the particle ξ_j by a Gaussian density whose mean $\mu(\xi_j)$ and covariance $\Gamma(\xi_j)$

for each particle are obtained using the Extended Kalman Filter (EKF). This technique is computationally heavy because it requires to compute the gradient of the unnormalized transition kernel at each particle. A somewhat simpler version was later studied by Van der Merwe et al. (2000); Van der Merwe and Wan (2003) consisting in replacing the EKF by the unscented Kalman Filter (UKF). Pitt and Shephard (1999) suggest to approximate $x' \mapsto l(x, x')$ via a form of Laplace approximations, i.e. with Gaussian or t-student distributions centered at its mode. This technique is appropriate when the function $x' \mapsto l(x, x')$ is log-concave (or strongly unimodal). Nevertheless, the determination of the mode requires solving an optimization problem for each particle.

In spite of this intense recent activity in the field, the state-of-the-art algorithms have met mitigated success, as the success of the linearization or the Laplace approximations is heavily model dependent and is effective only under rather restrictive modeling assumptions. In particular, these techniques implicitly assume that the function $x' \mapsto l(x, x')$ has a single mode, the approximation of the optimal kernel being fitted to this mode.

A challenge in the SMC community is hence the design of adaptive methods to derive adjustment weighting functions and proposal kernels; we focus in this paper on the second problem. The idea of adapting the proposal distribution to the target in order to improve sampler performance is classical in the importance sampling literature. A common practice to design a proposal distribution in non-sequential IS (see Rubinstein (1981); Oh and Berger (1993); Rubinstein and Kroese (2008)) is to choose first a family of proposal distributions parameterized by $\boldsymbol{\theta}$, and to try to find a parameter which minimizes a measure of the discrepancy between the target distribution and the proposal distribution. There are many such discrepancy measures, but it has often been found convenient to use the Kullback-Leibler divergence (which is the limit of the entropy of the importance weights). The chi-square divergence (*CSD*) can also be of interest, but will not be considered in this work. Cornebise et al. (2008) have proposed to apply the adaptive IS to the auxiliary particle filter. The idea consists in adapting the parameters $\boldsymbol{\theta}$ by minimizing the d_{KL}between the auxiliary target distribution μ_{aux} and a family of auxiliary proposal distributions $\{\pi_{\mathrm{aux}}^{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathrm{KL}}\left(\mu_{\mathrm{aux}} \| \pi_{\mathrm{aux}}^{\boldsymbol{\theta}}\right) , \tag{5.1.7}$$

where d_{KL} stands for the Kullback-Leibler divergence.

Note that, as pointed out in (Cornebise et al., 2008, Section 2.3), the d_{KL}decouples the adaptation of the adjustment weight function Ψ and of the proposal kernel r, as

$$d_{\mathrm{KL}}\left(\mu_{\mathrm{aux}} \| \pi_{\mathrm{aux}}\right) = \mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log \frac{\mu_{\mathrm{aux}}(I,\tilde{\xi})}{\pi_{\mathrm{aux}}(I,\tilde{\xi})}\right] \tag{5.1.8}$$

$$= \underbrace{\mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log \frac{l^*(I,\tilde{\xi})}{r(I,\tilde{\xi})}\right]}_{\text{Depends only on kernel } r} + \underbrace{\mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log \frac{\omega_I \Psi^*(\xi_I)/\sum_{j=1}^N \omega_j \Psi^*(\xi_j)}{\omega_I \Psi(\xi_I)/\sum_{j=1}^N \omega_j \Psi(\xi_j)}\right]}_{\text{Depends only on function } \Psi} , \tag{5.1.9}$$

where the first term corresponds to the discrepancy induced by the proposal of the particles $\tilde{\xi}$ with a suboptimal proposal kernel, and the second term corresponds to the discrepancy induced by the choice of the ancestors index I with suboptimal adjustment weights. This brings a sound rationale to considering separately adaptation of the proposal kernel and adaptation of the weights: both are optimization problems that do not conflict nor interact. Note that CSD does not lead to such a convenient decoupling.

Moreover, the terms involving the optimal quantities Ψ^* and l^* inside the expectation involved in the KLD can be expressed as irrelevant additive terms, that is,

$$d_{\mathrm{KL}}\left(\mu_{\mathrm{aux}} \| \pi_{\mathrm{aux}}\right) \equiv -\mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log r(I, \tilde{\xi})\right] - \mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log \frac{\Psi(\xi_I)}{\sum_{j=1}^{N} \omega_j \Psi(\xi_j)}\right] \tag{5.1.10}$$

up to an irrelevant additive term (equality up to a constant is denoted by $\equiv$)[1]. When restricting ourselves to the adaptation of the proposal kernel and thus neglecting terms not involved in this optimization, we obtain the extremely simple expression

$$d_{\mathrm{KL}}\left(\mu_{\mathrm{aux}} \| \pi_{\mathrm{aux}}\right) \equiv -\mathbb{E}_{\mu_{\mathrm{aux}}}\left[\log r(I, \tilde{\xi})\right] , \tag{5.1.11}$$

still up to an irrelevant additive term. Therefore, although the $d_{\mathrm{KL}}$is most often not directly computable, such a quantity can be approximated *on-the-fly* using the weighted sample obtained; the resulting algorithm closely resembles the Cross-Entropy (*CE*, Rubinstein and Kroese (2004)) method (with the difference that the parameters can be updated several times).

The chosen parametric family of distributions should offer enough flexibility for allowing to approximate complex transition kernels. On the other hand, it should be simple enough so that sampling from $\pi_{\mathrm{aux}}^{\boldsymbol{\theta}}$ is easy. Finally, the parameterization should be chosen in such a way that the problem of estimating the parameters be as simple as possible. In this paper, we suggest to model the proposal $\pi_{\mathrm{aux}}^{\boldsymbol{\theta}}(i, \tilde{\xi})$ as a mixture of integrated curved exponential distributions. The mixture weights are chosen to depend on the particle ξ_i, thus allowing to partition the input space among regions to which correspond a specialized kernel. Each component of the mixture belongs to a family of integrated curved exponential distributions, whose two most known members are the multivariate Gaussian and t-student distributions. The parameters of the mixture also depend on the particle positions ξ_i. This parameterization of the proposal distribution is closely related to the (hierarchical) mixture of experts from the machine learning community as described in Jordan and Jacobs (1994); Jordan and Xu (1995). The flexibility of this mixture approach allows for fitting of nonlinear and intricate models.

The paper is organized as follows. In Section 5.2, the choice of proposal distributions is described. In Section 5.3, the optimization of the parameters is presented. In Section 5.4, two versions of the adaptive sequential importance sampling algorithms are presented. In Section 5.5, several examples are presented to support our findings.

5.2 Mixture of experts

In this contribution, we let the proposal kernel have density

$$r_{\boldsymbol{\theta}}(\xi, \tilde{\xi}) := \sum_{j=1}^{d} \alpha_j(\xi, \boldsymbol{\beta}) \rho\left(\xi, \tilde{\xi}; \eta_j\right) . \tag{5.2.1}$$

where the functions $\{\alpha_j\}_{j=1}^d$ are the so-called *weighting functions* and $\rho(\cdot, \cdot; \eta_j)$ are Markov transition kernels from $\boldsymbol{\Xi}$ to $\tilde{\boldsymbol{\Xi}}$. The parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$ characterizes the

1. When dealing with an optimization problem, we adopt the following convention in order to get rid of irrelevant terms. For any two functions f and g of two variables x and y, that we wish to optimize in x, we will denote $f(x, y) \equiv g(x, y)$ the *equality up to an irrelevant constant*, that is if and only if there exists a positive function a of y and a function b of y such that $f(x, y) = a(y)g(x, y) + b(y)$. This implies that for any y, $\arg\max_x f(x, y) = \arg\max_x g(x, y)$ and $\arg\min_x f(x, y) = \arg\min_x g(x, y)$. More general definitions could be given but would be unnecessarily intricate.

mixture chosen in the proposal family, where we denote $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)$ the vector of every component's parameters. In the APF framework, the associated proposal distribution $\pi^{\boldsymbol{\theta}}_{\text{aux}}$ is defined by its density

$$p_{\boldsymbol{\theta}}(i, \tilde{\xi}) := \frac{\omega_i \psi_i}{\sum_{k=1}^{N} \omega_k \psi_k} \sum_{j=1}^{d} \alpha_j(\xi_i, \boldsymbol{\beta}) \rho\left(\xi_i, \tilde{\xi}; \eta_j\right) , \tag{5.2.2}$$

where ψ_i, $i = 1, \ldots, N$ are the multiplier adjustment weights. We then assign the importance weight

$$\tilde{\omega}^{\boldsymbol{\theta}}_j = \frac{l(\xi_{I_j}, \tilde{\xi}_j)}{\psi_{I_j} \sum_{\ell=1}^{d} \alpha_\ell(\xi_{I_j}, \boldsymbol{\beta}) \, \rho\left(\xi_{I_j}, \tilde{\xi}_j; \eta_j\right)} \tag{5.2.3}$$

to each draw $(I_j, \tilde{\xi}_j)$ from $\pi^{\boldsymbol{\theta}}_{\text{aux}}$.

The purpose of the weighting function (also referred to as *gating network* in Jordan and Jacobs (1994); Jordan and Xu (1995)), is to partition the input space Ξ into regions which are associated to a single (or perhaps a few) *specialized* transition kernel in the mixture. This is done by assigning a vector of mixture weights to each point of the input space. The weighting function implements a mapping between the input space and the vector of distributions over the set of the mixture indices. As in Jordan and Xu (1995), we consider logistic weight functions, that is

$$\alpha_j(\xi, \boldsymbol{\beta}) := \frac{\exp\left(\beta_j^T \overline{\xi}\right)}{\sum_{u=1}^{d} \exp\left(\beta_u^T \overline{\xi}\right)} , \tag{5.2.4}$$

with $\mathbb{B} = \mathbb{R}^{(p+1)\times(d-1)} \times \{0_{\mathbb{R}^{p+1}}\}$ where, for any vector $x \in \mathbb{R}^p$, $\overline{x}$ denotes

$$\overline{x} := (x, 1) , \tag{5.2.5}$$

the extended vector in $\mathbb{R}^{p+1}$ obtained by appending 1 to x. The boundaries of the input space where $\alpha_j = \alpha_{j'}$ are hyperplanes in $\mathbb{R}^{p+1}$ and the weighting functions induce a smoothed piecewise-planar partitioning of the input space. More complex divisions can be obtained using a hierarchical mixture of experts

In some cases it is of interest to resort to a simpler mixture whose weights do not depend on ξ, by letting $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^d : \sum_{j=1}^{d} \beta_j = 1\}$ and having

$$\alpha_j(\xi, \boldsymbol{\beta}) = \beta_j \tag{5.2.6}$$

be independent of ξ. This model for the transition is then similar to the switching regression model in statistics (Quandt and Ramsey, 1972). The difference between constant and logistic weights is that the constant weights assume that the choice of the proposal transition is independent on the input vector. This latter assumption does not allow for piecewise variations in the choice of the proposal kernels, as all of the proposal kernels then contribute with the same proportions throughout the whole input space.

To ease the implementation – see Section 5.3 – the kernels $\rho\left(\xi, \tilde{\xi}; \eta\right)$ of the family are assumed to have the following *integrated curved exponential* form:

$$\rho\left(\xi, \tilde{\xi}; \eta\right) = \int_{\mathbb{U}} \rho^e\left(\xi, \tilde{\xi}, u; \eta\right) \mathrm{d}u , \tag{5.2.7}$$

where $\rho^e\left(\xi, \tilde{\xi}, u; \eta\right)$ is a *curved exponential distribution*:

$$\rho^e\left(\xi, \tilde{\xi}, u; \eta\right) = \gamma(u) h(\xi, \tilde{\xi}, u) \exp\left(-A(\eta) + \left\langle S\left(\xi, \tilde{\xi}, u\right), B(\eta)\right\rangle\right) . \tag{5.2.8}$$

Here, we have used the following notation: for any two matrices with the same dimensions, A and B, $\langle A, B\rangle = \mathrm{Tr}(A^T B)$. In other words, $\rho\left(\xi, \tilde{\xi}; \eta\right)$ corresponds to the marginal in $(\xi, \tilde{\xi})$ of a curved exponential distribution on $\rho^e\left(\xi, \tilde{\xi}, u; \eta\right)$ defined for some auxiliary variable $u \in \mathbb{U}$ whose own marginal distribution is γ. We denote S the vector of sufficient statistics of this joint distribution.

For reasons that will become clear in Section 5.3, we augment $(I, \tilde{\xi})$ with the index J of the mixture component and with the auxiliary variable U of the curved exponential family to define an *extended* auxiliary variable $(I, J, \tilde{\xi}, U)$, which is distributed according to the extended distribution $\pi_e^{\boldsymbol{\theta}}$ on the product space $\{1, \dots, N\} \times \{1, \dots, d\} \times \tilde{\boldsymbol{\Xi}} \times \mathbb{U}$ with density

$$p_{\boldsymbol{\theta}}^e(i, j, \tilde{\xi}, u) := \frac{\omega_i \psi_i}{\sum_{k=1}^N \omega_k \psi_k} \alpha_j(\xi_i, \boldsymbol{\beta})\, \rho^e\left(\xi_i, \tilde{\xi}, u; \eta_j\right) . \tag{5.2.9}$$

It is easily checked that $\pi_{\mathrm{aux}}^{\boldsymbol{\theta}}$ is the marginal density of $\pi_e^{\boldsymbol{\theta}}$ in I and $\tilde{\xi}$, as

$$\begin{aligned}\sum_{j=1}^d \int_{\mathbb{U}} p_{\boldsymbol{\theta}}^e(i, j, \tilde{\xi}, u)\mathrm{d}u &= \frac{\omega_i \psi_i}{\sum_{k=1}^N \omega_k \psi_k} \sum_{j=1}^d \alpha_j(\xi_i, \boldsymbol{\beta}) \int_{\mathbb{U}} \rho^e\left(\xi_i, \tilde{\xi}, u; \eta_j\right) \mathrm{d}u \\ &= \frac{\omega_i \psi_i}{\sum_{k=1}^N \omega_k \psi_k} \sum_{j=1}^d \alpha_j(\xi_i, \boldsymbol{\beta}) \rho\left(\xi_i, \tilde{\xi}; \eta_j\right) = p_{\boldsymbol{\theta}}\left(i, \tilde{\xi}\right) .\end{aligned}$$

We define the *conditional mixture weights*

$$p_{\boldsymbol{\theta}}\left[j \middle| i, \tilde{\xi}\right] := \mathbb{P}_{\boldsymbol{\theta}}\left[J = j \middle| I = i, \tilde{\xi} = \tilde{\xi}\right] = \frac{\alpha_j(\xi_i, \boldsymbol{\beta})\, \rho\left(\xi_i, \tilde{\xi}; \eta_j\right)}{\sum_{k=1}^d \alpha_k(\xi_i, \boldsymbol{\beta})\, \rho\left(\xi_i, \tilde{\xi}; \eta_k\right)} . \tag{5.2.10}$$

where the probability $\mathbb{P}_{\boldsymbol{\theta}}$ is evaluated w.r.t. the distribution $\pi_e^{\boldsymbol{\theta}}$. In the sequel, it is assumed that for any $\boldsymbol{\eta}$ and any $j \in \{1, \dots, d\}$, the statistics

$$\begin{aligned}S_j\left(i, \tilde{\xi}, \boldsymbol{\theta}\right) &:= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{1}_{\{J=j\}}\, \mathbb{E}_{\eta_J}\left[S\left(\xi, \tilde{\xi}, U\right) \middle| \xi = \xi_i, \tilde{\xi} = \tilde{\xi}\right]\right] \\ &= p_{\boldsymbol{\theta}}\left[j \middle| i, \tilde{\xi}\right] \mathbb{E}_{\eta_j}\left[S\left(\xi, \tilde{\xi}, U\right) \middle| \xi = \xi_i, \tilde{\xi} = \tilde{\xi}\right]\end{aligned} \tag{5.2.11}$$

are available in closed form, the expectation $\mathbb{E}_{\eta_j}$ being w.r.t. the curved exponential distribution $\rho^e\left(\xi, \tilde{\xi}, u; \eta_j\right)$. We will denote

$$S_{j,0}\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\mu_{\mathrm{aux}}}\left[p_{\boldsymbol{\theta}}\left[j \middle| I, \tilde{\xi}\right]\right] \quad \text{and} \quad S_j\left(\boldsymbol{\theta}\right) = \mathbb{E}_{\mu_{\mathrm{aux}}}\left[S_j\left(I, \tilde{\xi}, \boldsymbol{\theta}\right)\right] , 1 \leq j \leq d , \tag{5.2.12}$$

the *expected sufficient statistics of component* j.

This framework now allows for efficient fitting of the optimal parameter $\boldsymbol{\theta}^*$ minimizing the KLD (5.1.7). This is the topic of the Section 5.3. After a generic statement of the algorithm, we will focus in Sections 5.3.2 and 5.3.2 on two families. Following the usage in IS, we consider either multidimensional Gaussian distributions or multidimensional t-distributions. When using multidimensional Gaussian distributions, our proposal kernel coincides with the so-called mixture of experts introduced in machine learning literature to parameterize conditional distributions (see Jordan and Jacobs (1994); Jordan and Xu (1995)). The use of mixtures (but not dynamic mixtures) of multidimensional t-distribution has frequently be used for importance sampling applications: the heavier tails of the t-distribution tend to reduce the variability of the importance weights.

Remark 5.2.1. In addition to the integrated curved exponential families of distributions, these two cases also fit in the commonly used framework of *location and scale* families,

$$\rho^e\left(\xi, \tilde{\xi}, M; \Sigma\right) := |\Sigma|^{-1/2}\, p\left(\Sigma^{-1/2}\left(\tilde{\xi} - M\overline{\xi}\right)\right) \tag{5.2.13}$$

where p is a density w.r.t. to some dominating measure (most often, the Lebesgue measure on $\mathbb{R}^{p'}$). The *location* is a linear function of ξ, i.e. $M\,\overline{\xi}$, where $\overline{\xi} := (\xi^T, 1)^T$, The matrix M (and therefore, all the matrices M_j, $j \in \{1, \ldots, d\}$) hence belong to the space $\mathcal{M}(p, p+1)$ of $p \times (p+1)$ matrices, and will be referred to as *regression matrices*. The matrices Σ is a multidimensional *scaling parameter* which, for simplicity, does not depend on ξ. The matrices Σ (and therefore, all the matrices Σ_j, $j \in \{1, \ldots, d\}$) belong to the space $\mathcal{M}_s^+(p)$ of $p \times p$, symmetric definite positive matrices.

5.3 Parameter estimation techniques

In this section, we discuss the parameter optimization techniques to solve the optimization problem (5.1.7). We will use interchangeably the terms *parameter optimization* and *parameter estimation*, though the problem (5.1.7) involves expectations w.r.t. the whole distribution μ_{aux} rather than w.r.t. the empirical distribution of some observations.

In mixture models, parameter estimation is most often performed using the EM algorithm of Dempster et al. (1977), which allows computationally efficient optimization in latent-data settings (see also Wu (1983) for the theoretical study and McLachlan and Peel (2000) for an up-to-date review). The EM algorithm proceeds iteratively. Denote $\boldsymbol{\theta}^{\ell-1}$ the current fit of the parameter after iteration $\ell-1$. The new fit $\boldsymbol{\theta}^{\ell}$ of the parameter is obtained by maximizing in $\boldsymbol{\theta}$ the *intermediate quantity*

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) := \mathbb{E}_{\mu_{\text{aux}}}\left[\mathbb{E}_{\boldsymbol{\theta}^{\ell-1}}\left[\log p^e_{\boldsymbol{\theta}}(I, J, \tilde{\xi}, U)\middle| I, \tilde{\xi}\right]\right] , \tag{5.3.1}$$

where $p^e_{\boldsymbol{\theta}}$ is the complete likelihood defined in (5.2.9) and $\mathbb{E}_{\boldsymbol{\theta}^{\ell-1}}$ denotes expectation w.r.t. the extended auxiliary distribution $\pi_e^{\boldsymbol{\theta}^{\ell-1}}$. Hence, vanilla EM algorithm lets

$$\boldsymbol{\theta}^{\ell} = (\boldsymbol{\beta}^{\ell}, \boldsymbol{\eta}^{\ell}) = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) . \tag{5.3.2}$$

The intermediate quantity $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1})$, may be decomposed as the following sum – here and in the following, $\equiv$ denotes *equality up to an irrelevant constant* as noted in footnote 1:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\ell-1}) \equiv Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1}) + Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1}) , \tag{5.3.3}$$

where

$$Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1}) := \mathbb{E}_{\mu_{\text{aux}}}\left[\mathbb{E}_{\boldsymbol{\theta}^{\ell-1}}\left[\log \alpha_J(\xi_I, \boldsymbol{\beta})\middle| I, \tilde{\xi}\right]\right] \tag{5.3.4}$$

depends only on the parameter $\boldsymbol{\beta}$ of the mixture weights, and

$$Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1}) := \mathbb{E}_{\mu_{\text{aux}}}\left[\mathbb{E}_{\boldsymbol{\theta}^{\ell-1}}\left[\log \rho^e\left(\xi_I, \tilde{\xi}, U; \eta_J\right)\middle| I, \tilde{\xi}\right]\right] \tag{5.3.5}$$

depends only on the parameter $\boldsymbol{\eta}$ of the kernels. Therefore, at each iteration, the optimization problem (5.3.2) can be split in two subproblems, one dealing with the parameters of the mixture weights, the other with the parameters of the proposal kernels for each component:

$$\boldsymbol{\beta}^{\ell} = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1}) , \quad \text{and} \quad \boldsymbol{\eta}^{\ell} = \arg\max_{\boldsymbol{\eta}\in\mathbb{E}} Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1}) .$$

Additionally, reaching the exact maximum of the intermediate quantity at each iteration is not needed. Under weak additional assumptions – see (Wu, 1983, Section 2) – it suffices to increase the intermediate quantity at each iteration, i.e. ensure $Q(\boldsymbol{\theta}^\ell, \boldsymbol{\theta}^{\ell-1}) > Q(\boldsymbol{\theta}^{\ell-1}, \boldsymbol{\theta}^{\ell-1})$. This class of algorithms is commonly known as *generalized EM* (GEM), and is very helpful when the exact solution of (5.3.2) is not available, as will be the case when optimizing the weights of the mixture in Section 5.3.1.

5.3.1 Optimizing the weighting functions

We now turn to the optimization of the weighting functions (or gating network) α_j, whose design has been discussed in 5.2.

A first natural choice for the mixture weights functions α_j would be to have them constant, independent of the ancestor particle, that is $\alpha_j(\xi, \boldsymbol{\beta}) = \beta_j$, where $\boldsymbol{\beta} \in \mathbb{B} = \{(\beta_1, \dots, \beta_d) \in [0,1]^d : \sum_{j=1}^d \beta_j = 1\}$. This model is commonly known as a *mixture of regressions*. Maximisation of $Q_1(\boldsymbol{\alpha}, \boldsymbol{\theta}^\ell)$, where Q_1 is defined in (5.3.4) , in $\boldsymbol{\alpha}$ under the constraint that $\sum_{j=1}^d \alpha_j = 1$ can here be achieved in closed form and leads to

$$\breve{\alpha_j} = \tilde{S}_{j,0}(\boldsymbol{\theta}) \;, \quad \text{with} \quad \tilde{S}_{j,0}(\boldsymbol{\theta}) := \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I, \tilde{\xi}\right]\right] \;. \tag{5.3.6}$$

For the logistic weights (5.2.4), finding the maximum in $\boldsymbol{\beta}$ of $Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^\ell)$ in closed form is out of reach. However, this exact maximization is not mandatory for the algorithm to converge: increasing the intermediate quantity at each step suffices; see Wu (1983). We propose to use a single step of a simple numerical optimization such as gradient ascent method or Newton-Raphson algorithm. Define an ascent direction

$$\boldsymbol{d}^\ell(\boldsymbol{\beta}^{\ell-1}) := \begin{cases} \left[\nabla^2_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\ell-1}}\right]^{-1} \nabla_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\ell-1}} \;, & \text{(gradient ascent)} \\ \nabla_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\ell-1}} \;, & \text{(Newton-Raphson)} \end{cases} \tag{5.3.7}$$

where, for a mapping $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}'}$ and $\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{x}'}$ its gradient and Hessian matrix, respectively, evaluated in $\boldsymbol{x}'$. A step of optimization then proceeds to

$$\boldsymbol{\beta}^\ell = \boldsymbol{\beta}^{\ell-1} + \tau_\ell \boldsymbol{d}^\ell(\boldsymbol{\beta}^{\ell-1})$$

for some positive step size τ_ℓ either chosen beforehand or, better, determined by a line-search algorithm such as described by, e.g., Fletcher (1987, Section 2.6). We stress here that we are not even looking for the optimal step size, but only ensure that we increase the function Q_1 of interest.

Computation of the gradient $\nabla_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})$ is easily achieved by simplifying (5.3.4) in

$$Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1}) \equiv \sum_{j=1}^d \beta_j^T \, \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^\ell}\left[j \middle| I, \tilde{\xi}\right] \overline{\xi}_I\right] - \mathbb{E}_{\mu_{\text{aux}}}\left[\log\left\{\sum_{j=1}^d \exp\left(\beta_j^T \, \overline{\xi}_I\right)\right\}\right] \;,$$

up to an irrelevant additive constant which does not depend on $\boldsymbol{\beta}$. This immediately leads to the components

$$\frac{\partial Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})}{\partial \beta_k} = \mathbb{E}_{\mu_{\text{aux}}}\left[\left\{p_{\boldsymbol{\theta}^\ell}\left[k \middle| I, \tilde{\xi}\right] - \alpha_k\left(\xi_I, \boldsymbol{\beta}\right)\right\} \overline{\xi}_I\right] \tag{5.3.8}$$

of the gradient. The computation of the Hessian $\nabla^2_{\boldsymbol{\beta}} Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^\ell)$ needed for Newton-Raphson optimization now requires the evaluation of

$$\frac{\partial \alpha_k\left(\overline{\xi}, \boldsymbol{\beta}\right)}{\partial \beta_m} = \alpha_k\left(\xi, \boldsymbol{\beta}\right)\left[\mathbb{1}_{\{k=m\}} - \alpha_m\left(\xi, \boldsymbol{\beta}\right)\right] \overline{\xi} \;. \tag{5.3.9}$$

Hence, for any $k, m \in \{1, \dots, d-1\}$, each of the $(d-1)\times(d-1)$ blocks of size $(p+1)\times(p+1)$ of the Hessian matrix can be expressed as

$$\frac{\partial^2 Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})}{\partial \beta_k \partial \beta_m} = \mathbb{E}_{\mu_{\text{aux}}}\left[\alpha_k\left(\xi_I, \boldsymbol{\beta}\right)\left\{\alpha_m\left(\xi_I, \boldsymbol{\beta}\right) - \mathbb{1}_{\{k=m\}}\right\}\bar{\xi}_I^2\right] , \tag{5.3.10}$$

where, for any vector or matrix A, A^2 is a shorthand notation for $A\,A^T$.

5.3.2 Optimizing the mixture kernels

We first present optimization in the general case of integrated curved exponential kernels defined in (5.2.7), before specializing to the multivariate Gaussian and student-t distribution. Starting from (5.3.5) and using (5.2.12), we have

$$\begin{aligned} Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1}) &= \mathbb{E}_{\mu_{\text{aux}}}\left[\mathbb{E}_{\boldsymbol{\theta}^{\ell-1}}\left[\log \rho^e\left(\xi_I, \tilde{\xi}, U; \eta_J\right)\middle| I, \tilde{\xi}\right]\right] \\ &\equiv \sum_{j=1}^{d} -A\left(\eta_j\right) S_{j,0}\left(\boldsymbol{\theta}^{\ell-1}\right) + \left\langle S_j\left(\boldsymbol{\theta}^{\ell-1}\right), B\left(\eta_j\right)\right\rangle . \end{aligned} \tag{5.3.11}$$

Multivariate Gaussian regression

The Gaussian regression kernels is the simplest family to deal with. To each component j corresponds a linear Gaussian regression, parameterized by $\eta_j = (M_j, \Sigma_j)$. The current particle $\tilde{\xi}$ is distributed according to a p'-dimensional Gaussian with mean $M_j\,\bar{\xi}_I$, where the *regression matrix* M_j belongs to the space $\mathcal{M}(p', p'+1)$ of $p' \times (p+1)$ matrices and the *covariance matrix* Σ_j belongs to the space $\mathcal{M}_s^+(p')$ of $p' \times p'$, symmetric definite positive matrices. The corresponding kernel density

$$\rho\left(\xi, \tilde{\xi}; \eta_j\right) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\left(\tilde{\xi} - M_j\,\xi\right)\Sigma_j^{-1}\left(\tilde{\xi} - M_j\,\xi\right)^T\right\}$$

can be cast into the framework (5.2.7), as it is a curved exponential distribution of the form (5.2.8) in itself; the auxiliary variable U is not needed and we let the functions γ and h be defined as: $\gamma(u) = 1$ and $h(\xi, \tilde{\xi}, u) = (2\pi)^{-p/2}$. The sufficient statistics are given by $S(\xi, \tilde{\xi}, u) = \left(\tilde{\xi}^2, \bar{\xi}^2, \tilde{\xi}\,\bar{\xi}^T\right)$, and lead to the following form of the expected sufficient statistics defined in (5.2.12):

$$\begin{aligned} S_{j,0}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle| I, \tilde{\xi}\right]\right] , & S_{j,1}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle| I, \tilde{\xi}\right]\,\tilde{\xi}^2\right] , \\ S_{j,2}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle| I, \tilde{\xi}\right]\,\bar{\xi}_I^2\right] , & S_{j,3}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle| I, \tilde{\xi}\right]\,\tilde{\xi}\,\bar{\xi}_I^T\right] . \end{aligned} \tag{5.3.12}$$

and the functions A and B are given by

$$A(\eta) = \frac{1}{2}\log|\Sigma| , \quad \text{and} \quad B(\eta) = \left(-\frac{1}{2}\Sigma^{-1}, -\frac{1}{2}M\Sigma^{-1}M^T, -\Sigma^{-1}M\right) .$$

The maximization of $Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1})$ in $\boldsymbol{\eta}$ is given by $M_j^\ell = M\left(S_{j,0:3}\left(\boldsymbol{\theta}^{\ell-1}\right)\right)$ and $\Sigma_j^\ell = \Sigma\left(S_{j,0:3}\left(\boldsymbol{\theta}^{\ell-1}\right)\right)$ where the functions M and Σ are defined as follows

$$M\left(S_{0:3}\right) := S_3\,S_2^{-1} \quad \text{and} \quad \Sigma\left(S_{0:3}\right) := \frac{S_1 - S_3\,S_2^{-1}\,S_3^T}{S_0} . \tag{5.3.13}$$

Multivariate t-Student regression

A common advice in Importance Sampling (see Oh and Berger (1993) for example) is to replace Gaussian by Student's t-distributions with location parameter μ, scale parameter Σ, and a chosen positive number of degrees of freedom $\nu > 0$, resulting in density

$$t_{p'}\left(\tilde{\xi};\mu,\Sigma,\nu\right) = \frac{\Gamma\left(\frac{\nu+p'}{2}\right)}{(\pi\nu)^{\frac{p'}{2}}\Gamma\left(\frac{\nu}{2}\right)\left[1+\delta\left(\tilde{\xi},\mu,\Sigma\right)/\nu\right]^{\frac{1}{2}(\nu+p')}} ,$$

where $\delta\left(\tilde{\xi},\mu,\Sigma\right) = \left|\tilde{\xi}-\mu\right|^2_{\Sigma^{-1}} = \left(\tilde{\xi}-\mu\right)^T \Sigma^{-1} \left(\tilde{\xi}-\mu\right)$ is the Mahalanobis distance with covariance matrix Σ, and $\Gamma(x) := \int_{\mathbb{R}^+} t^{x-1}\exp(-t)\mathrm{d}t$.

The t-distributions have heavier tails than Gaussian distributions, and hence take better account of outlying data or heavy-tailed targets. If $\nu > 1$, the mean of the distribution is 0, and if $\nu > 2$, its covariance matrix is $\nu I_{p'}/(\nu-2)$. Therefore, as ν tends to infinity, the t-distribution converges almost everywhere to the Gaussian density, and so ξ becomes marginally multivariate Gaussian distributed, with mean M and covariance Σ. The family of t distributions hence provides a heavy-tailed alternative to the normal family (if $\nu > 2$), as argued by Liu and Rubin (1995) and Peel and McLachlan (2000).

The use of these t-distribution for regression transition kernels is trivial, replacing location μ by the product $M\overline{\xi}$ of a regression matrix M and an extended regressor ξ. We therefore propose to use

$$\rho\left(\xi,\tilde{\xi};\eta_j\right) = t_{p'}\left(\tilde{\xi};M_j\,\overline{\xi},\Sigma_j,\nu\right) \tag{5.3.14}$$

as a robustified version of the Gaussian kernels exposed in Section 5.3.2.

Remark 5.3.1. For sake of simplicity, the number of degrees of freedom ν of the t-Student distributions is chosen beforehand and kept fixed, typically to $\nu \in \{3,4\}$, and is common to all the components. A similar choice has been adopted, among others, in Cappé et al. (2008), though they involve constant location parameters rather than regressions. Such a choice is also considered in Peel and McLachlan (2000)[Section 7], as it allows for a closed-form M-step. Its only drawback is to weaken the robustness against outlying data, which is not critical in our adaptive design of an importance kernel.

These kernels can be cast into the framework of Section 5.2 thanks to the *Gaussian-Gamma distribution* of multivariate t-Student distributions, introduced in (Liu and Rubin, 1995, Section 2) and (Peel and McLachlan, 2000, Section 3). They show that the multivariate t-Student can be seen as a continuous mixture of scaled Gaussians, that is

$$t\left(\tilde{\xi};\mu,\Sigma,\nu\right) = \int \mathcal{N}\left(\tilde{\xi};\mu,\Sigma/u\right)\gamma\left(u;\frac{\nu}{2},\frac{\nu}{2}\right)\mathrm{d}u , \tag{5.3.15}$$

where

$$\gamma\left(u;a,b\right) := \frac{b^a u^{a-1}}{\Gamma(a)}\exp\left(-bu\right)\mathbb{1}_{\mathbb{R}^+}(u)$$

is the density of a Gamma distribution with shape parameter a and scale parameter b. The multivariate t-distribution kernel (5.3.14) is hence an integrated curved exponential distribution (5.2.7) with $\gamma(u) = \gamma\left(u;\frac{\nu}{2},\frac{\nu}{2}\right)$, $h(\xi,\tilde{\xi},u) = (2\pi)^{-p/2}$, sufficient statistics $S(\xi,\tilde{\xi},u) = \left(u\tilde{\xi}^2, u\overline{\xi}^2, u\tilde{\xi}\,\overline{\xi}^T\right)$, and

$$A(\eta) = \frac{1}{2}\log|\Sigma| \ , \quad \text{and} \quad B(\eta) = \left(-\frac{1}{2}\Sigma^{-1}, -\frac{1}{2}M\Sigma^{-1}M^T, -\Sigma^{-1}M\right) ,$$

To compute the conditional expectation of S, we need to evaluate $\mathbb{E}_\eta\left[U\middle|\xi=\xi_i,\tilde{\xi}\right]$; since the Gamma distribution is the conjugate prior for U, this can be done in closed form, leading to the conditional density:

$$\begin{aligned} p_\eta\left(u\middle|\xi,\tilde{\xi}\right) &= \rho^e\left(\xi,\tilde{\xi},u;\eta\right)/\rho\left(\xi,\tilde{\xi};\eta\right) \\ &= \gamma\left(u;\frac{1}{2}\left(\nu+p'\right),\frac{1}{2}\left(\nu+\delta\left(\tilde{\xi},M\xi,\Sigma\right)\right)\right) . \end{aligned}$$

The conditional expectation of U is hence

$$\mathbb{E}_\eta\left[U\middle|\xi,\tilde{\xi}\right] = \frac{\nu+p'}{\nu+\delta\left(\tilde{\xi},M\bar{\xi},\Sigma\right)} , \tag{5.3.16}$$

which leads to the expected sufficient statistics

$$\begin{aligned} S_{j,0}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle|I,\tilde{\xi}\right]\right] , \\ S_{j,1}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle|I,\tilde{\xi}\right]\frac{\nu+p'}{\nu+\delta\left(\tilde{\xi},M_j\bar{\xi}_I,\Sigma_j\right)}\tilde{\xi}^2\right] , \\ S_{j,2}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle|I,\tilde{\xi}\right]\frac{\nu+p'}{\nu+\delta\left(\tilde{\xi},M_j\bar{\xi}_I,\Sigma_j\right)}\bar{\xi}_I^2\right] , \\ S_{j,3}\left(\boldsymbol{\theta}^{\ell-1}\right) &:= \mathbb{E}_{\mu_{\text{aux}}}\left[p_{\boldsymbol{\theta}^{\ell-1}}\left[j\middle|I,\tilde{\xi}\right]\frac{\nu+p'}{\nu+\delta\left(\tilde{\xi},M_j\bar{\xi}_I,\Sigma_j\right)}\tilde{\xi}\bar{\xi}_I^T\right] . \end{aligned} \tag{5.3.17}$$

Because the functions A and B are unchanged from the Gaussian case, the same equations (5.3.13) can be applied to update the parameter $\boldsymbol{\eta}$ based on these new expected sufficient statistics.

5.4 Stochastic approximation and resulting algorithm

As mentioned above, the algorithms described so far mainly have a theoretical interest, because the computation of the sufficient statistics $\tilde{S}_{j,i}(\boldsymbol{\theta})$, $i=0,\ldots,4$, requires to compute expectations w.r.t. to the target distribution μ_{aux}. In most problems, neither integrating w.r.t., nor sampling from this distribution is possible. Though the expectations cannot therefore be computed exactly or approximated by crude Monte Carlo, they can be approximated by IS. Two different approaches can be considered: the batch and the stochastic approximation approaches.

5.4.1 Batch algorithm

In the batch algorithm, the total number of particles N is split into $L+1$ blocks of sizes $\{N^\ell\}_{\ell=0}^L$. At each iteration of the algorithm, the outer expectation $\mathbb{E}_{\mu_{\text{aux}}}[\cdot]$ in the computation of the intermediate quantities of the EM algorithm $Q_1(\boldsymbol{\beta},\boldsymbol{\theta}^{\ell-1})$ and $Q_2(\boldsymbol{\eta},\boldsymbol{\theta}^{\ell-1})$, defined in (5.3.4) and (5.3.5), is replaced by an importance sampling estimator. More precisely, at iteration ℓ, the current fit of the parameter is $\boldsymbol{\theta}^{\ell-1}$; we draw N^ℓ samples $\{(I_j^{[\ell]},\tilde{\xi}_j^{[\ell]})\}_{j=1}^{N_\ell}$ conditionally independently from the past using the proposal distribution $p_{\boldsymbol{\theta}^{\ell-1}}$, defined in (5.2.2). Every quantity appearing in the definitions of Q_1

and Q_2 which is defined as an expectation $\mathbb{E}_{\mu_{\text{aux}}}\left[F(\boldsymbol{\theta}, I, \tilde{\xi})\right]$ of some measurable function F (see (5.3.6), (5.3.8), (5.3.10), (5.3.12) for the Gaussian distribution and (5.3.17) for the t-distribution) is estimated using

$$\hat{\mathbb{E}}^{[\ell]}_{\mu_{\text{aux}}}\left[F(\boldsymbol{\theta}, I, \tilde{\xi})\right] := \sum_{j=1}^{N^\ell} \frac{\tilde{\omega}_j^{[\ell]}}{\sum_{u=1}^{N^\ell} \tilde{\omega}_u^{[\ell]}} F(\boldsymbol{\theta}; I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]}) \tag{5.4.1}$$

We then update the parameters using the procedures outlined in the previous section – where $\boldsymbol{\theta}^{-1}$ can be chosen according to prior knowledge of the model (see Remark 5.4.1). Such an algorithm is related to the so-called sample average in the sense that, at each step of the algorithm, the objective functions $Q_1(\boldsymbol{\beta}, \boldsymbol{\theta}^{\ell-1})$ and $Q_2(\boldsymbol{\eta}, \boldsymbol{\theta}^{\ell-1})$ are defined as expectations of functions which are approximated by a sample average estimate derived from a random sample. The resulting sample average approximating problem is then solved by deterministic optimization techniques. In the definition of Verweij et al. (2003), such an algorithm is referred to as an interior sampling method, in the sense that the samples used during the optimization are modified by generating new samples obtained using the proposal distribution associated to the current fit of the parameter. Note that it is also possible to consider solutions that use all the samples obtained so far, but the complexity in this case becomes prohibitive. The rationale behind this is that each iteration leads to an *improved* fit $\boldsymbol{\theta}^{\ell-1}$ i.e. a proposal distribution with smaller KLD to the target, and hence makes way for a more reliable approximation of expectations under μ_{aux}.

The theoretical study of the convergence of the algorithm – taking into account the uncertainty added by the random approximations of the solutions – is beyond the scope of this paper and will be considered in a companion paper. The analysis is based on the results obtained on simulation based optimization (see Shapiro (1996) and Shapiro and Homem de Mello (2001)), with the additional difficulty that the optimization steps are themselves included in the iterations of the EM algorithm, much like in the Monte-Carlo version of the EM algorithm (see Fort and Moulines (2003) for an in-depth analysis). In practice, because we do not seek to reach convergence to the optimum value but rather to incrementally improve the proposal distribution, only a fixed number of iterations are performed.

At the end of a whole cycle of optimization, we thus have a set of $L+1$ weighted samples $\{(\tilde{\omega}_j^{[\ell]}, I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]})\}_{j=1}^{N^\ell}$, all targeting the same distribution μ_{aux}. By construction, the weighted sample $\{(\tilde{\omega}_j^{[\ell]}, I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]})\}_{j=1}^{N^\ell}$ is conditionally independent from $\{(\tilde{\omega}_j^{[\ell]}, I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]})\}_{j=1}^{N^l}$ for $l = 1, \dots, \ell-1$. Assume now that we are willing to estimate $\mu_{\text{aux}}(f) := \int f(x)\mu_{\text{aux}}(x)\mathrm{d}x$. The different weighted samples provide us with a sequence of conditionally independent estimates of $\mu_{\text{aux}} f$ given by:

$$\hat{\mu}^{[\ell]}_{\text{aux}}(f) := \sum_{j=1}^{N^\ell} \frac{\tilde{\omega}_j^{[\ell]}}{\sum_{u=1}^{N^\ell} \tilde{\omega}_u^{[\ell]}} f(\tilde{\xi}_j^{[\ell]}) \,.$$

Under appropriate conditions (see for example Douc and Moulines (2008)), we may establish a multidimensional central limit theorem (CLT) proving that the L dimensional vector

$$\sqrt{N}\left(\sqrt{\frac{N^1}{N}}\left(\hat{\mu}^{[0]}_{\text{aux}}(f) - \mu_{\text{aux}}(f)\right), \dots, \sqrt{\frac{N^L}{N}}\left(\hat{\mu}^{[1]}_{\text{aux}}(f) - \mu_{\text{aux}}(f)\right)\right) \tag{5.4.2}$$

is asymptotically normal. This multidimensional CLT suggests to combine these estimators.

5.4.2 Stochastic approximation algorithm

A second approach is more connected to *stochastic approximation* (SA) EM algorithm, as studied by Delyon et al. (1999) and Kuhn and Lavielle (2004). Batch EM can indeed advantageously be replaced by a stochastic approximation scheme, sampling a single particle at each iteration, updating the sufficient statistics according this new particle, and using these updated statistics to get a new fit of the parameters. This may be seen as an extension to the earlier work by Arouna (2004) on adaptive importance sampling. Updating the statistics at each new particle sampled is called *sample adaptation* of the fit, as adaptation takes place for every particle sampled. However, such a frequent adaptation can be cumbersome, due to the computation overhead at each iteration.

It is therefore more relevant to proceed to *block adaptation* of the fit. At iteration ℓ, sample a block of N^ℓ particles, and let each estimate $\hat{S}_{j,i}^\ell$ of an expected sufficient statistic $S_j(i)$ be a convex combination of the former estimate $\hat{S}_{j,i}^{\ell-1}$ and of an IS approximation computed on the new block of particles. The weights of the convex combination are given by a sequence (λ_ℓ) of positive stepsizes, such that $\sum_{\ell=0}^{\infty} \lambda_\ell = \infty$ and $\sum_{\ell=0}^{\infty} \lambda_\ell^2 < \infty$, with $\lambda_0 = 1$ for initialization.

Note that a special scare has to be paid to estimation of the IS normalizing constant, as we cannot use the self-normalized importance sampling estimate anymore: the small number of particles per block would increase the variance of the estimates to an unreasonable point. Recall that in the plain APF setting, we have

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}_i \xrightarrow{\mathbb{P}} \int \nu(\xi) l(\xi, \tilde{\xi}) d\xi d\tilde{\xi} \,, \tag{5.4.3}$$

when N tends to infinity – the proof follows immediatly by the Theorem 2.3.2 from Chapter 2 of this dissertation, which can also be found in Douc et al. (2008). Therefore, relying as in for the batch case on the N^ℓ weighted samples $\{(I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]})\}_{j=1}^{N_\ell}$ of the current iteration, we use as a normalizing constant of the IS estimate the following stochastic approximation of the right hand side of the last display:

$$\mathcal{C}^\ell := \begin{cases} \frac{1}{N^1}\sum_{i=1}^{N^1} \tilde{\omega}_i^{[1]} \,, & \text{for } \ell = 0 \\ (1-\lambda_\ell)\mathcal{C}^{\ell-1} + \lambda_\ell \frac{1}{N^\ell}\sum_{i=1}^{N^\ell} \tilde{\omega}_i^{[\ell]} \,, & \text{for } \ell \geq 1 \,. \end{cases} \tag{5.4.4}$$

We then estimate each expectation $\mathbb{E}_{\mu_{\text{aux}}}\left[F(\boldsymbol{\theta}, I, \tilde{\xi})\right]$ of some measurable function F by

$$\hat{\mathbb{E}}_{\mu_{\text{aux}}}^{[\ell]}\left[F(\boldsymbol{\theta}, I, \tilde{\xi})\right] := (1-\lambda_\ell)\hat{\mathbb{E}}_{\mu_{\text{aux}}}^{[\ell-1]}\left[F(\boldsymbol{\theta}^{\ell-1}, I, \tilde{\xi})\right] + \lambda_\ell \sum_{j=1}^{N^\ell} \frac{\tilde{\omega}_j^{[\ell]}}{N^\ell \mathcal{C}^\ell} F(\boldsymbol{\theta}; I_j^{[\ell]}, \tilde{\xi}_j^{[\ell]}) \,, \tag{5.4.5}$$

to be compared with its batch-EM setting equivalent (5.4.1).

A major difference with the batch-EM algorithm is that, past the initial step $\ell = 0$ which provides the first optimized fit (see Remark 5.4.1 below for insights on the choice of the initial $\boldsymbol{\theta}^{-1}$), the numbers N^ℓ of particles sampled at each iteration $\ell \geq 1$ can be noticeably smaller than in classical EM (potentially by an order of magnitude), as the current fit at iteration ℓ, is based on the whole set of $\sum_{k=0}^{\ell} N^k$ particles generated, by means of the convex combination with the past fit. The algorithm for the Gaussian case is stated in Algorithm 5.4.1, and the t-Student case is obtained by straightforward modifications.

Remark 5.4.1 (Initial fit of the algorithm). Note that the initial parameter fit $\boldsymbol{\theta}^0$ can be automatically chosen based on prior knowledge of the model, e.g. fitting the local likelihood for state space models with very informative observations. Besides, the initial statistics $\mathcal{C}^0$ and $\hat{S}^0_{j,0:3}$ can be set arbitrarily to 0, as $\lambda_1 = 1$.

Remark 5.4.2 (Optimization scheme in the initial iteration). As a subtlety, the maximization of $\boldsymbol{\beta}$ in the first iteration $\ell = 0$ might be achieved by means of a more intricate optimization scheme than a step of gradient: we could use e.g. the *Broyden-Fletcher-Goldfarb-Shanno* (*BFGS*, see (Fletcher, 1987, Chapter 3)) algorithm with permissive stopping conditions. However, we will see in the example of Section 5.5.3 that the added computational cost is not necessarily worth the increased initial accuracy, and that the simplicity of a step of gradient can often be preferred.

Remark 5.4.3 (Proposal kernel in the initial iteration). Finally, we may chose to propose the first step according to another distribution than $\pi^{\boldsymbol{\theta}^0}_{\text{aux}}$ to avoid relying on an arbitrary initial choice which is often quite poor. In the state space model framework, in the examples of Section 5.5, we propose the first sample $\{I^{[0]}_i, \tilde{\xi}^{[0]}_i\}_{i=1}^{N^0}$ according to the prior kernel – the simplest default choice. Of course, the importance weights must be computed accordingly.

5.5 Applications

5.5.1 Non-linear state-spaces model

One of the main achievement of SMC method can be found in *state-space models*. They consist of generic nonlinear dynamic system described in the following form:

– *State (system) equation*

$$X_k = f_{k-1}(X_{k-1}, U_k)$$

– *Observation (measurement) equation*

$$Y_k = h_k(X_k, V_k) \tag{5.5.1}$$

where $\{f_k(\cdot)\}_{k\geq 0}$ and $\{h_k(\cdot)\}_{k\geq 0}$ are sequences of possibly nonlinear, known functions. When these functions do not depend on time k, the state-space model is said *homogeneous* in time. The random variables $\{U_k\}_{k\geq 1}$, $\{V_k\}_{k\geq 0}$ are mutually independent sequences of i.i.d. random variables, known as the *state noise* and *observation noise*, respectively. The precise form of the functions and the assumed probability distributions of the dynamic and observation noises U_k and V_k imply, via a change of variables, the *Hidden Markov Model* (HMM) representation

$$X_k \overset{\cdot|X_{k-1}}{\sim} \overbrace{q_{k-1}(X_{k-1}, \cdot)}^{\text{Prior kernel}} \tag{5.5.2}$$

$$Y_k \overset{\cdot|X_k}{\sim} \underbrace{g_k(X_k, \cdot)}_{\text{Local likelihood}} \tag{5.5.3}$$

With these definitions, the process $\{X_k\}_{k\geq 0}$ is Markovian, i.e. the conditional probability density of X_k given the past states $X_{0:k-1} := (X_0, \ldots, X_{k-1})$ depends exclusively on X_{k-1}, with the random variables X_k being X-valued. The kernel Q is, Markovian on the state space X, admits density q_{k-1}, and is referred to as the *prior kernel*. We assume further that the initial state X_0 is distributed according to a density function $\pi_0(x_0)$. The observation probability density function $g_k(x_k, y_k)$ is referred to as the *local*

Algorithm 5.4.1 Optimization of a mixture of Gaussian experts by SA

Require: Initial parameter fit $\boldsymbol{\theta}^{-1}$. Arbitrary values $\hat{S}^{-1}_{j,0:3}$, $\mathcal{C}^{-1}$. Stepsize $\lambda_0 = 1$.

1: **for** $\ell = 0, 1, \ldots, L$ **do**

2: sample $\{I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\}_{i=1}^{N^\ell}$ from $\pi_{\text{aux}}^{\boldsymbol{\theta}^{\ell-1}}$ conditionally independently from the past,

3: compute their respective importance weights defined in (5.2.3):

$$\tilde{\omega}_i^{[\ell]} := \frac{l(\xi_{I_i^{[\ell]}}, \tilde{\xi}_i^{[\ell]})}{\sum_{j=1}^d \alpha_j^{\ell-1}(\xi_{I_i^{[\ell]}}, \boldsymbol{\beta}^{\ell-1}) \mathcal{N}\left(\tilde{\xi}_i^{[\ell]}; M_j^{\ell-1} \overline{\xi}_{I_i^{[\ell]}}, \Sigma_j^{\ell-1}\right)} , \tag{5.4.6}$$

4: compute their respective conditional mixture weights defined in (5.2.10):

$$p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\right] = \frac{\alpha_j^{\ell-1}(\xi_{I_i^{[\ell]}}, \boldsymbol{\beta}^{\ell-1}) \mathcal{N}\left(\tilde{\xi}_i^{[\ell]}; M_j^{\ell-1} \overline{\xi}_{I_i^{[\ell]}}, \Sigma_j^{\ell-1}\right)}{\sum_{u=1}^d \alpha_u^{\ell-1}(\xi_{I_i^{[\ell]}}, \boldsymbol{\beta}^{\ell-1}) \mathcal{N}\left(\tilde{\xi}_i^{[\ell]}; M_u^{\ell-1} \overline{\xi}_{I_i^{[\ell]}}, \Sigma_u^{\ell-1}\right)} , \tag{5.4.7}$$

reusing quantities already computed for the importance weights,

5: update the approximation of the normalization constant as stated in (5.4.5):

$$\mathcal{C}^\ell := (1 - \lambda_\ell)\mathcal{C}^{\ell-1} + \lambda_\ell \frac{1}{N^\ell} \sum_{i=1}^{N^\ell} \tilde{\omega}_i^{[\ell]} , \tag{5.4.8}$$

6: approximate (5.3.12) by computing the sufficient statistics

$$\hat{S}^\ell_{j,0} := (1 - \lambda_\ell)\hat{S}^{\ell-1}_{j,1} + \lambda_\ell \sum_{i=1}^{N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{N^\ell \mathcal{C}^\ell} p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\right] , \tag{5.4.9}$$

$$\hat{S}^\ell_{j,1} := (1 - \lambda_\ell)\hat{S}^{\ell-1}_{j,1} + \lambda_\ell \sum_{i=1}^{N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{N^\ell \mathcal{C}^\ell} p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\right] \left(\tilde{\xi}_i^{[\ell]}\right)^2 , \tag{5.4.10}$$

$$\hat{S}^\ell_{j,2} := (1 - \lambda_\ell)\hat{S}^{\ell-1}_{j,2} + \lambda_\ell \sum_{i=1}^{N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{N^\ell \mathcal{C}^\ell} p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\right] \left(\overline{\xi}_{I_i^{[\ell]}}\right)^2 , \tag{5.4.11}$$

$$\hat{S}^\ell_{j,3} := (1 - \lambda_\ell)\hat{S}^{\ell-1}_{j,3} + \lambda_\ell \sum_{i=1}^{N^\ell} \frac{\tilde{\omega}_i^{[\ell]}}{N^\ell \mathcal{C}^\ell} p_{\boldsymbol{\theta}^{\ell-1}}\left[j \middle| I_i^{[\ell]}, \tilde{\xi}_i^{[\ell]}\right] \tilde{\xi}_i^{[\ell]} \overline{\xi}^T_{I_i^{[\ell]}} , \tag{5.4.12}$$

7: update the parameters $\boldsymbol{M}^\ell$ and $\boldsymbol{\Sigma}^\ell$ by applying (5.3.13) to the current fit of the sufficient statistics:

$$M_j^\ell := \hat{S}^\ell_{j,3}\left[\hat{S}^\ell_{j,2}\right]^{-1} , \qquad \Sigma_j^\ell := \frac{\hat{S}^\ell_{j,1} - M_j^\ell \left[\hat{S}^\ell_{j,3}\right]^T}{\hat{S}^\ell_{j,0}} , \tag{5.4.13}$$

8: compute the direction $\hat{\boldsymbol{d}}^\ell(\boldsymbol{\beta}^{\ell-1})$ by computing (5.3.7) with the expectations in the gradient (5.3.8) and the Hessian (5.3.10) approximated in a way similar to the sufficient statistics ,

9: update the parameter $\boldsymbol{\beta}^\ell$ by a step of stochastic gradient of stepsize τ_ℓ:

$$\boldsymbol{\beta}^\ell := \boldsymbol{\beta}^{\ell-1} + \tau_\ell \hat{\boldsymbol{d}}^\ell(\boldsymbol{\beta}^{\ell-1}) . \tag{5.4.14}$$

10: **end for**

likelihood. When working conditionally on the observations, the dependence in y_k is often dismissed and the local likelihood is only noted $g_k(x_k)$, as a function of the hidden state. Note that the conditional probability density of Y_k given the states $X_{0:k}$ and the past observations $Y_{0:k-1}$ depends exclusively on X_k, and this distribution is captured by the likelihood $g(x_k, y_k)$. Starting with the initial, or prior, density function $\pi_0(x_0)$, and observations $Y_{0:k} = y_{0:k}$, the posterior density $\phi_{k|k}(x_k|y_{0:k})$ can be obtained using the following *prediction-correction* recursion (Ho and Lee, 1964):

– *Prediction*

$$\phi_{k|k-1}(x_k|y_{0:k-1}) = \phi_{k-1|k-1}(x_{k-1}|y_{0:k-1})q(x_{k-1}, x_k) \ , \tag{5.5.4}$$

– *Correction*

$$\phi_{k|k}(x_k|y_{0:k}) = \frac{g(x_k, y_k)\phi_{k|k-1}(x_k|y_{0:k-1})}{p_{k|k-1}(y_k|y_{0:k-1})} \ , \tag{5.5.5}$$

where $p_{k|k-1}$ is the predictive distribution of Y_k given the past observations $Y_{0:k-1}$. Compared to the classical approaches in the non-linear filtering literature, SMC method construct approximations of the posterior probability distribution of the batch of states, whereas the extended or the unscented Kalman filter estimate the posterior mean and covariance of the state. Most importantly, these approximations are obtained without compromising the state-space model by introducing unrealistic modeling assumptions.

By setting $\Xi = \tilde{\Xi} = \mathsf{X}$, $\nu = \phi_{k|k}$, $\mu = \phi_{k-1|k-1}$ we see that this filtering problem can be cast in the sequential importance sampling framework with the update kernel L having density

$$l(\xi, \tilde{\xi}) = g_k(\tilde{\xi})\, q_{k-1}(\xi, \tilde{\xi}) \ . \tag{5.5.6}$$

This leads to the importance weights

$$\tilde{\omega}_i = \frac{1}{\psi_{I_i}} \frac{g(\tilde{\xi}_i)q(\xi_{I_i}, \tilde{\xi}_i)}{r(\xi_{I_i}, \tilde{\xi}_i)} \ . \tag{5.5.7}$$

5.5.2 Multivariate linear Gaussian model

We will now illustrate our findings on a multivariate state-space model where the optimal kernel and the optimal adjustment, defined in (5.1.4), are available in closed form.

Following the notation from Section 5.5.1, we will consider a single timestep (k) of a non-linear state-space model on $\mathsf{X} = \mathbb{R}^d$ with observations in $\mathsf{Y} = \mathbb{R}^p$ for which

– the prior kernel $Q(X_0, \cdot)$ is a mixture of multivariate Gaussian distributions, i.e. has density

$$q(\xi, \tilde{\xi}) = \sum_{i=1}^{m} \alpha_i \mathcal{N}_d\left(\tilde{\xi}; A_i\, \overline{\xi}, \Sigma_i\right) \tag{5.5.8}$$

where $A_i \in \mathcal{M}(d \times (d+1))$, and $\Sigma_i \in \mathcal{M}_s^+(d)$,

– the local likelihood

$$g(\tilde{\xi}, y) = \sum_{k=1}^{n} \beta_k \mathcal{N}_p\left(y; B_k\, \tilde{\xi}, \Sigma_{Y,k}\right) \ , . \tag{5.5.9}$$

is also a mixture of multivariate Gaussian, with $B_k \in \mathcal{M}(p \times d)$ and $\Sigma_{Y,k} \in \mathcal{M}_s^+(p)$.

In this setting, the optimal adjustment weight function Ψ^* and the optimal kernel L^* are available in closed form. In order to lead a progressive exposure, we will first consider the optimal kernel and adjustment weights in the linear Gaussian case, i.e. when $m = p = 1$. In this case the unnormalized kernel L admits a density

$$l(\xi, \tilde{\xi}) = \mathcal{N}_d\left(\tilde{\xi}; A\, \overline{\xi}, \Sigma\right) \times \mathcal{N}_p\left(y; B\, \tilde{\xi}, \Sigma_Y\right) \ . \tag{5.5.10}$$

Consequently, the optimal adjustment function defined in (5.1.4) is given by

$$\Psi^*(\xi) = \left(\frac{|\Gamma|\,|S|}{|\Sigma|\,|\Sigma_Y|}\right)^{1/2} \exp\left\{-\frac{1}{2}\, y^T\, \left(S^{-1} - \Sigma_Y^{-1}\right)\, y\right\} \mathcal{N}_p\left(y; B\,A\,\overline{\xi}, S\right) \tag{5.5.11}$$

with

$$S := \Sigma_Y + B\,\Sigma\,B^T\ ,$$

and the optimal kernel defined in (5.1.4) has density

$$l^*(\xi, \tilde{\xi}) = \mathcal{N}_d\left(\tilde{\xi}; C_1\,\overline{\xi} + C_2\,y, \Gamma\right)\ , \tag{5.5.12}$$

with $\Gamma := \left(\Sigma^{-1} + B^T\,\Sigma_Y^{-1}\,B\right)^{-1}$, $C_1 := \Gamma\,\Sigma^{-1}\,A$ and $C_2 := \Gamma\,B^T\,\Sigma_Y^{-1}$.

Proof. Recall that $L^*(\xi,\cdot) = L(\xi,\cdot)/\Psi^*(\xi)$. By uniqueness of this decomposition, it is sufficient to exhibit a factor $\mathcal{N}_d\left(\tilde{\xi}; C_1\overline{\xi} + C_2 y, \Gamma\right)$ in $L^*(\xi,\cdot)$ to also obtain the eppression of the optimal adjustment weights. We proceed by identification of the terms of the two quadratic forms

$$\left(\tilde{\xi} - A\,\overline{\xi}\right)^T\, \Sigma^{-1}\, \left(\tilde{\xi} - A\,\overline{\xi}\right) + \left(\tilde{\xi} - B\,y\right)^T\, \Sigma_y^{-1}\, \left(\tilde{\xi} - B\,y\right) \tag{5.5.13}$$

and

$$\left(\tilde{\xi} - C_1\,\overline{\xi} - C_2\,y\right)^T\, \Gamma^{-1}\, \left(\tilde{\xi} - C_1\,\overline{\xi} - C_2\,y\right)\ . \tag{5.5.14}$$

Identifying second order terms leads

$$\begin{aligned} \tilde{\xi}^T\,\Gamma^{-1}\,\tilde{\xi} = \tilde{\xi}^T\left(\Sigma^{-1} + B^T\,\Sigma_Y^{-1}\,B\right)\,\tilde{\xi} \quad &\text{, hence} \\ \Gamma = \left(\Sigma^{-1} + B^T\,\Sigma_Y^{-1}\,B\right)^{-1}\ . \end{aligned}$$

Turning to terms involving $\tilde{\xi}$ and ξ entails

$$\begin{aligned} \tilde{\xi}^T\,\Gamma^{-1}\,C_1\,\overline{\xi} = \tilde{\xi}^T\,\Sigma^{-1}\,A\,\overline{\xi} \quad &\text{, hence} \\ C_1 = \Gamma\,\Sigma^{-1}\,A\ . \end{aligned}$$

From identification of terms involing $\tilde{\xi}$ and y, we obtain

$$\begin{aligned} \tilde{\xi}^T\,\Gamma^{-1}\,C_2\,y = \tilde{\xi}^T\,B^T\,\Sigma_Y^{-1}\,y \quad &\text{, hence} \\ C_2 = \Gamma\,B^T\,\Sigma_Y^{-1}\ . \end{aligned}$$

Similar identifications on the constant terms, i.e. terms not involving $\tilde{\xi}$, and factorization, lead to

$$\begin{aligned} l(\xi, \tilde{\xi}) = \mathcal{N}_d\left(\tilde{\xi}; C_1\,\overline{\xi} + C_2\,y, \Gamma\right) \\ \times \left(\frac{|\Gamma|}{(2\pi)^d\,|\Sigma|\,|\Sigma_Y|}\right)^{1/2} \exp\left\{\frac{1}{2}\left|\Sigma^{-1}\,A\,\overline{\xi} + B^T\,\Sigma_Y^{-1}\,y\right|_\Gamma^2 - \frac{1}{2}\left|A\,\overline{\xi}\right|_{\Sigma^{-1}}^2 - \frac{1}{2}\,|y|_{\Sigma_Y^{-1}}^2\right\} \end{aligned}$$

where $|u|_A := (u^T\,A\,u)^{1/2}$ denotes the norm of vector or matrix u according to metric associated with the square matrix A. This proves (5.5.12). Note that, by Woodbury matrix identity,

$$\Gamma = \Sigma + \Sigma\,B^T\,\left(\Sigma_Y + B\,\Sigma\,B^T\right)^{-1}\,B\,\Sigma\ ,$$

and that

$$\left(\Sigma_Y + B\,\Sigma\,B^T\right)^{-1} B\,\Sigma = \Sigma_Y^{-1}\,B\left(\Sigma^{-1} + B^T\,\Sigma_Y^{-1}\,B\right)^{-1} .$$

With the two latter equalities in mind, Equation (5.5.11) stems from identification of

$$\left|\Sigma^{-1}\,A\,\overline{\xi} + B^T\,\Sigma_Y^{-1}\,y\right|_{\Gamma}^2 - \left|A\,\overline{\xi}\right|_{\Sigma^{-1}}^2 = -\left|B\,A\,\overline{\xi} - y\right|_{S^{-1}}^2 + c(y)$$

in matrix S and function c. □

This result obviously generalizes to arbitrary m and p. The optimal kernel defined in (5.1.4) is itself a mixture of $m \times n$ Gaussian, with density

$$l^*(\xi, \tilde{\xi}) = \sum_{j=1}^{m}\sum_{k=1}^{n} \gamma_{j,k}(\xi)\mathcal{N}_d\left(\tilde{\xi}; C_1^{j,k}\,\overline{\xi} + C_2^{j,k}\,y, \Gamma_{j,k}\right) . \tag{5.5.15}$$

The mixture weights are given by

$$\gamma_{j,k}(\xi) := \frac{\alpha_j\beta_k{\Psi^*}_{j,k}(\xi)}{\Psi^*(\xi)} , \quad \Psi^*(\xi) = \sum_{j=1}^{m}\sum_{k=1}^{n}\alpha_j\beta_k{\Psi^*}_{j,k}(\xi) , \tag{5.5.16}$$

where

$${\Psi^*}_{j,k}(\xi) = \left(\frac{|\Gamma_{j,k}|\,|S_{j,k}|}{|\Sigma_j|\,|\Sigma_{Y,k}|}\right)^{1/2} \exp\left\{-\frac{1}{2}\,y^T\left(S_{j,k}^{-1} - \Sigma_{Y,k}^{-1}\right)\,y\right\}\mathcal{N}_p\left(y; B_k\,A_j\,\overline{\xi}, S_{j,k}\right) \tag{5.5.17}$$

with $S_{j,k} := \Sigma_{Y,k} + B_k\,\Sigma_j\,B_k^T$. The parameters of the proposal kernels are given by

$$C_1^{j,k} := \Gamma_{j,k}\,\Sigma_j^{-1}\,A_j , \quad C_2^{j,k} := \Gamma_{j,k}\,B_k^T\,\Sigma_{Y,k}^{-1} , \quad \Gamma_{j,k} := \left(\Sigma_j^{-1} + B_k^T\,\Sigma_{Y,k}^{-1}\,B_k\right)^{-1} .$$

It is important to note that, with this kind of model,

- although the Gaussian components of the optimal kernel belong to the same family as the components of our mixture of experts,
- the partitions spanned by our logistic weights do not allow for exact matching of the mixture weights $\gamma_{j,k}(\cdot)$.

As a result, we cannot expect to fit exactly the optimal kernel, as it does not belong to our family of mixture of experts. However, as we will see, the optimal adjustment weights $\Psi^*(\cdot)$ are often negligible for large regions of the ancestor's space, i.e. many ancestors only produce sons which are very unlikely in the light of the observation. Those ancestors of low optimal adjustment weights will have very little impact on the KLD that our algorithm is minimizing. Therefore, we can expect that a good fit of the mixture probabilities will be achieved in the regions where the optimal importance weights are large, and neglect the fact that the fit will be less accurate in the regions of the space where optimal importance weights are small.

In order to easily visualize the results of the algorithm in a case where every optimal quantity is known, we consider one step of APF, where the initial particles $\{\xi_i\}_{i=1}^N$ are sampled exactly from the original distribution ν, and hence uniformly weighted, i.e. $\omega_i = 1$ for every $i \in \{1, \ldots, N\}$. The state space is $\mathsf{X} = \mathbb{R}^3$, hence $d = 3$. The distribution ν is chosen to be a mixture of two Gaussians with respective modes $m_1 = (0,0,0)^T$ and $m_2 = (6,6,6)^T$, that is

$$\xi_i \overset{i.i.d.}{\sim} 0.4\mathcal{N}_3\left(0_{\mathbb{R}^3}, I_3\right) + 0.6\mathcal{N}_3\left((6,6,6)^T, 0.5I_3\right) , \tag{5.5.18}$$

where I_3 is the identity matrix of $\mathcal{M}(3 \times 3)$. These original particles are plotted in Figure 5.1. The prior kernel Q is a mixture of two Gaussian distributions, having the same means, but different variance matrices to provide heavier tails than a simple Gaussian, that is of the form (5.5.8) with $m = 2$, $\alpha_1 = 1 - \alpha_2 = 0.25$, regression matrices

$$A_1 = A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

with null intercept, and variance matrices $\Sigma_1 = 3I_3$ and $\Sigma_2 = I_3$. The observation y lies in $\mathbb{R}^2$ with local likelihood g of the form (5.5.8) with dimension $p = 2$, $m = 1$, matrix

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

and $\Sigma_Y = 1.5I_2$, where the index has been dropped as there is only one component in the mixture. Note that the observation only depends on the two first components of the hidden state.

Case of a highly likely observation

In this first example, we chose the observation y is chosen to be $y = B\,A_1\,\overline{m_2} = (18, 12)^T$, that is precisely at the mode of its distribution conditionally on a hidden state itself situated at one of the modes of the prior distribution. This observation is hence not an outlier: it does not lie in the tails of distribution of the observations, conditionally on the hidden state (we will discuss such an outlying observation page 169). The optimal kernel L^* is given in this case by

$$l^*(\xi, \tilde{\xi}) = \sum_{j=1}^{m} \gamma_j(\xi) \mathcal{N}_d\left(\tilde{\xi}; C_1^j\,\overline{\xi} + C_2^j\,y, \Gamma_j\right) .$$

where

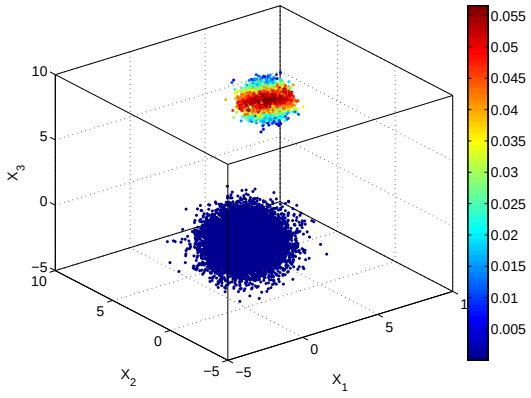
$$C_1^1 = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad C_1^2 = \begin{pmatrix} 0.6 & 0.6 & 0.6 & 0 \\ 0 & 0.6 & 0.6 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$C_2^1 = \begin{pmatrix} 2/3 & 0 \\ 0 & 2/3 \\ 0 & 0 \end{pmatrix}, \qquad C_2^2 = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.4 \\ 0 & 0 \end{pmatrix},$$

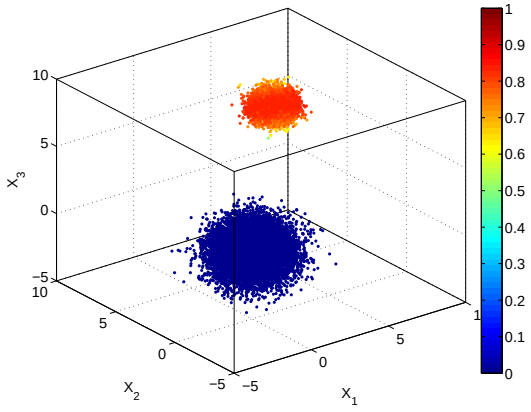
$$\Gamma_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}, \qquad \Gamma_2 = \begin{pmatrix} 0.6 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

These exact computations will allow to check the efficiency of the parameterization of the proposal kernel by a mixture of expert, allowing in particular to judge how closely the optimal KLD can be matched. The mixture weights $\gamma_1(\xi)$ and $\gamma_2(\xi)$ are given by Eq. (5.5.16), and the weights $\gamma_2(\xi_i)$, ranging from 0.85 to 10^{-33} and corresponding to the $20,000$ particles are displayed in Figure 5.1d. Weights $\gamma_1(\xi_i)$ are their complement to unity.

The optimal adjustment weights ${\psi^*}_i = \Psi^*(\xi_i)$, given by (5.5.16), for the same set of $20,000$ original particles ξ_i, are displayed in Figure 5.1c. As expected, the only ancestors to have a non-negligible optimal adjustment weights are those in a single mode of the



(c) Optimal adjustment weights $\Psi^*(\xi_i)$, for a mixture of Gaussians as prior kernel Q and a single Gaussian as local likelihood. The larger weights are associated with the brighter colors.



(d) Weight $\gamma_2(\xi)$ of the second component of the optimal kernel's mixture L^*.

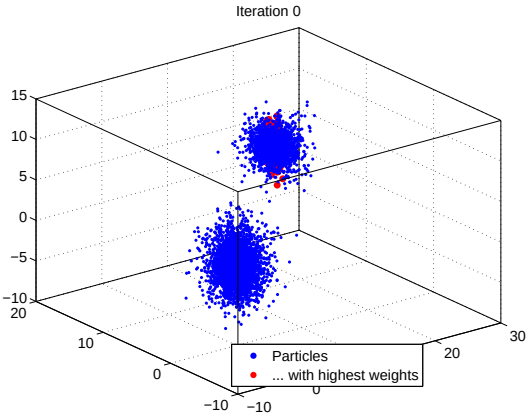
Figure 5.1: Original cloud of $20,000$ particles $\{(\xi_i, 1)\}_{i=1}^{20,000}$ for the bimodal original distribution ν of Equation (5.5.18).

original distribution (the observation is here assumed to be the image by the regression matrix of the mode of this mixture component). As we do not adapt the adjustment weights, all the ancestors are equally sampled and will be equally chosen to have sons. As a result, we can expect that a lot of proposed particles will have a negligible importance weight – as they will in any case have almost null predictive likelihood. This fact will clearly illustrate that our proposed algorithm will be able to adapt the proposal kernel even if the adjustment weights are not selected appropriately; this is in some sense a consequence of the decoupling of the optimization of the adjustment weights and of the proposal kernels evidenced in our companion paper Cornebise et al. (2008).

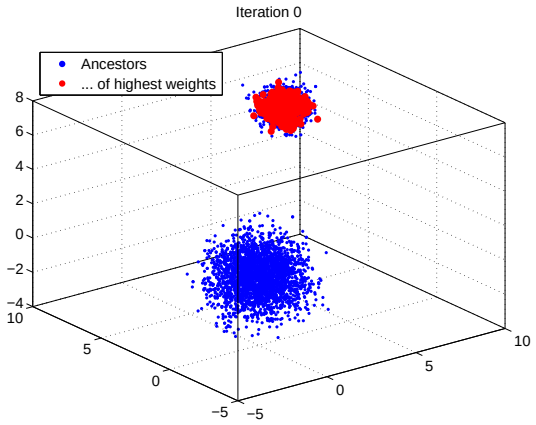
This is confirmed by studying the original sample of the algorithm, which is $1,000$ particles proposed according to the prior kernel, as displayed in Figure 5.2. The 100 particles with the highest weights are all located in the same mode of the proposal distribution, and, similarly, the ancestors of these particles with highest weights belong to a unique mode of the original distribution ν.

Figure 5.3 displays the weights of this first sample obtained from the prior distribution. Many importance weights are nearly equal to zero, which corresponds to particles whose ancestor belongs to the second mode of the prior distribution. The particles with ancestors belonging to the "correct" mode of the prior distributions have, not surprisingly, non-negligible importance weights. However, it is interesting to note that the importance weights of these particles still vary significantly, as can be seen the plot of the sorted importance weights on Figure 5.3a.

This is reflected in the Figure 5.3c, which we call a ***curve of proportions***, an unusual but informative visualization of the weights. We plot the proportion of particles (x-axis), that is $x = x/N^1$ against the proportion of the total mass cumulated by their weights (y-axis). The particles have been sorted by decreasing order of weights, so that the particles with larger weights are considered first. This allows for checking quickly if, for example, 10% of the particles carry 90% of the total mass – which would be a terrible waste of computational effort. In the ideal case where we would be sampling from the target distribution μ_{aux} (rather than from an importance sampling proposal

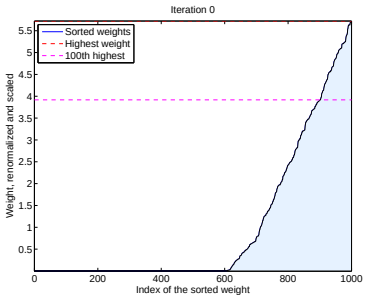


(a) Cloud of $1,000$ particles proposed with the prior kernel. The 100 particles with the highest weights are plotted in red.

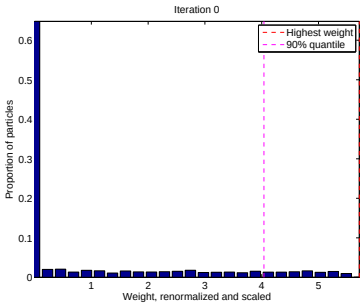


(b) Cloud of the ancestors of the $1,000$ particles proposed with the prior kernel. The ancestors of the 100 particles with the highest weights are plotted in red.

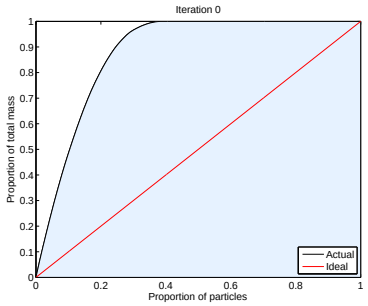
Figure 5.2: Proposing according to the prior kernel.



(a) Sorted weights of the first sample, proposed according to the prior kernel



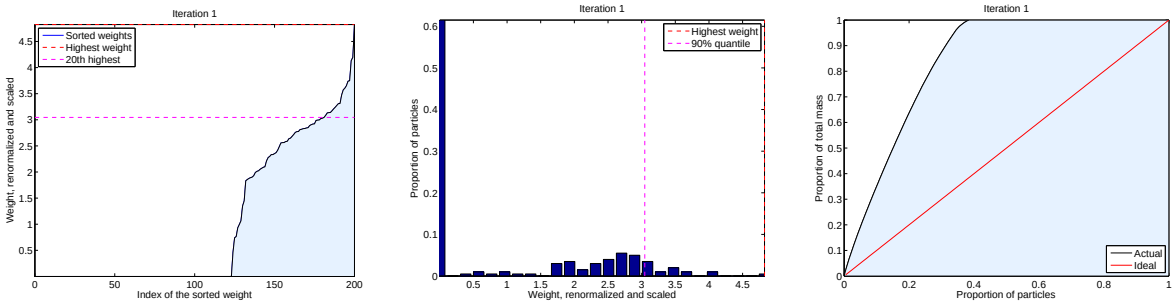
(b) Histogram of these weights.



(c) *Curve of proportions*: proportion of the particles (sorted by decreasing order of importance weight) against proportion of the total mass.

Figure 5.3: First sample, $N^0 = 1,000$ particles from the prior.

π_{aux}), all the weights would be equal, and the resulting plot would be a straight line, displayed here in red as a reference. We observe that using the prior kernel leads to 40% of the particles carrying the whle mass. Otherwise stated, 60% of the computational effort is wasted in particles having a null weight. Besides, the skewed aspect of the curve indicates that the non-null weights vary over a wide range, as can be checked on the histogram of the weights in Figure 5.3b.



(a) Sorted weights of the sample issued after iteration 1.

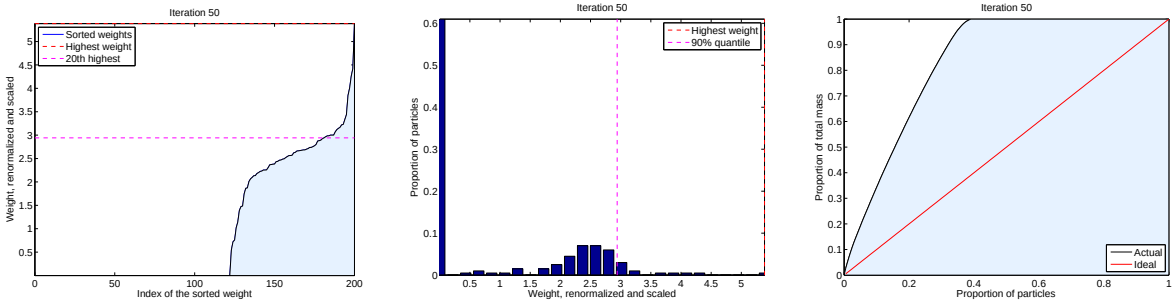
(b) Histogram of these weights.

(c) Curve of proportions.

Figure 5.4: Sample proposed after the first iteration, $N^1 = 100$ particles from the first fit.

We now perform a single iteration of our adaptation, in its stochastic approximation version described in Section 5.4 and summarized in Algorithm 5.4.1. We use the closed-form updates for the regression parameters and use a BFGS algorithm to optimize the parameter of the logistic weights, with a very permissive stopping criterion (i.e. stopping after 2 iterations of the BFGS procedure); see Figure 5.4. It is worthwhile to note that a single iteration is enough to reduce significantly the dispersion of the importance weights, as the curve of proportions is now closer to a straight line – this can be checked on the histogram. Note of course that 60% of particles still have negligible importance weights, which is due to the lack of optimization of the adjustment multiplier weights of the APF: solutions to this problem is the topic of a companion paper – which constitutes Chapter 4 of this dissertation.

Examining the corresponding figures at iteration 50, Figure 5.5 do not show much improvement; as often, the first steps of the optimization procedure bring a lot of improvements, values closed to the optimum are reached very fast.



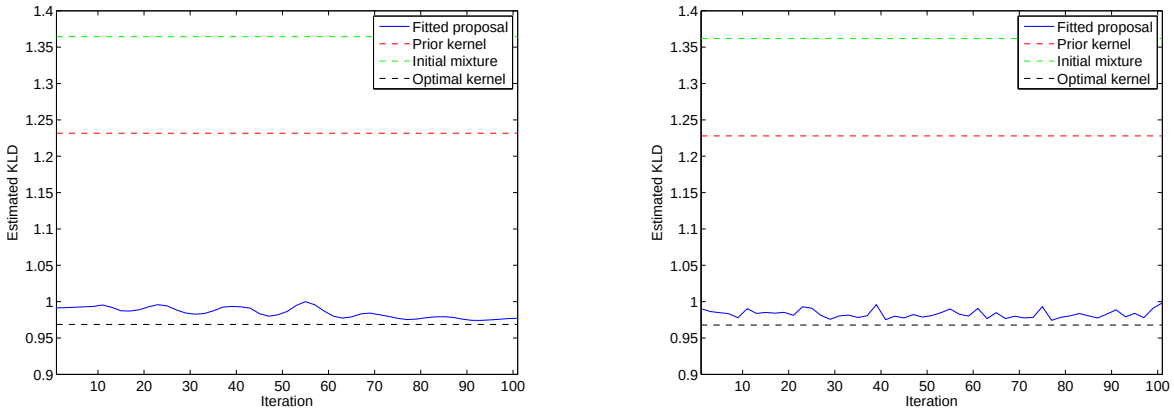
(a) Sorted weights of the sample issued after iteration 50.

(b) Histogram of these weights.

(c) Curve of proportions.

Figure 5.5: Sample proposed after 50 iterations, $N^50 = 100$.

To assess the performance of the algorithm, we estimate the KLD between the fit and the target, over the iterations (and the target distribution). More precisely, we approximate the quantity (5.1.9), where the optimal adjustment weights Ψ^* and optimal kernel l^* are exactly known, and we replace the expectation w.r.t. μ_{aux} by an importance sampling from a reference cloud of $50,000$ particles sampled from the prior kernel.



(a) Step size $0.1/\sqrt{L}$.

(b) Step size $1/\sqrt{L}$.

Figure 5.6: Comparison of the evolution of the KLD for two distinct step sizes, $L = 100$, $N_0 = 1,000$, $N^\ell = 200$.

Figure 5.6a shows the KLD obtained during the successive iterations of the adaptation procedure (the initial fit for the proposal kernel is nothing but the components of the prior kernel, with uniform logistic weights); it is compared to the KLD of the prior kernel and of the optimal kernel – note that in this latter case, we use the optimal kernel but still with uniform adjustment weights, as we are not considering optimization of the adjustment weights. As mentioned earlier, the KLD decreases very fast, most of the improvement being obtained in the first few iterations. Figure 5.6b shows the impact of the choice of the stepsize: unsurprisingly, as the stepsize is increased, the algorithm converges faster but the remaining perturbations are more noticeable – although this is of no importance as we are only looking for a good proposal kernel, not for the exact optimal one.

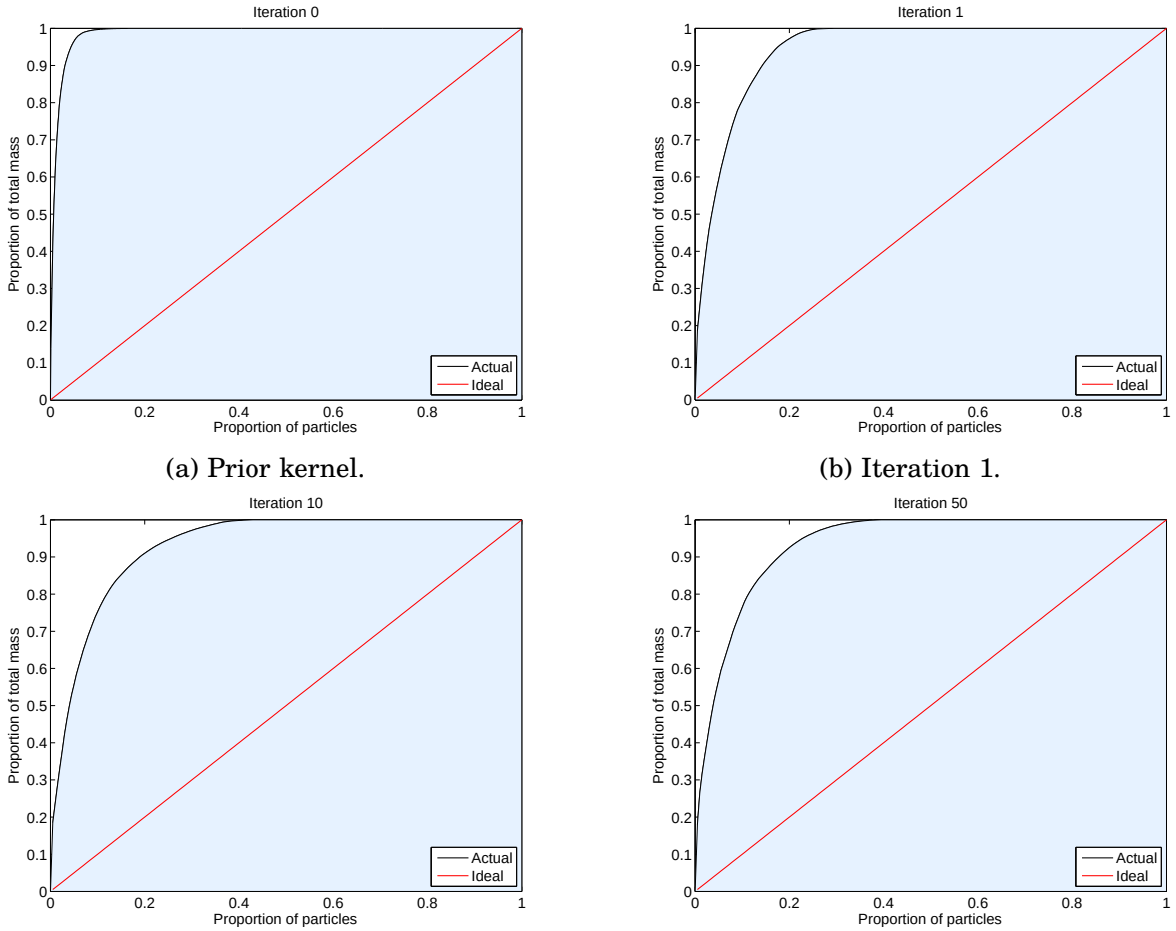
Remark 5.5.1. It would have been possible to neglect the optimal quantities and approximate the KLD only up to an irrelevant constant, by approximating the simple form (5.1.11). This would have been particularly easier for the examples of the next section, whose optimal kernels and optimal weights can only be obtained by numerical integration. However, by approximating the exact KLD (including the constants) and giving the KLD optimal kernel with uniform adjustment weights as a reference, we allow the reader to see which proportion of the KLD is left as a consequence of non-optimal adjustment weights – bringing another rationale for combining these algorithms with those of the companion paper exposed in Chapter 4.

Case of an outlying observation

As a more striking example of the robustness brought to SMC by our algorithm, we now consider a drastically outlying observation

$$y = B\left(A_1\,\overline{m_2} + \frac{4}{\sqrt{3}}\,\mathrm{chol}\left(\Sigma_2\right)^T \begin{pmatrix}1\\1\\1\end{pmatrix}\right) = \begin{pmatrix}22\\16\end{pmatrix}$$

where chol denotes the upper-diagonal Cholesky decomposition. This observation corresponds to an extreme offspring stemming from an ancestor situated in the second mode of the original distribution ν – that is, an offspring four standard deviations away, in the direction defined by vector $(1,1,1)^T$, from the mode of the prior kernel evaluated in $q(m_2,\cdot)$. Therefore, the prior kernel is expected to propose very few particles having a large likelihood, and our algorithm should much improve the efficiency. We keep initalizing our algorithm's components to the prior kernel's components, and the logistic weights' parameters are uniformly set to 0. We set the stepsize $\tau = 1/\sqrt{L}$ to allow for fast updates and speed up the convergence, as we expect that the initialization distribution will be quite far from the optimal. We keep the number of particles as low as in the first setting, with still an initial swarm of $N_0 = 1,000$, and $N^\ell = 200$ particles per iteration.



(a) Prior kernel.

(b) Iteration 1.

(c) Iteration 10.

(d) Iteration 50.

Figure 5.7: Curve of proportions.

As expected, the prior kernel is actually a dramatic proposal, as can be seen on the curve of proportions in Figure 5.7a: only roughly 8% of the particles have non negligible weights. The evolution of the KLD resulting from our algorithm, estimated on a basis of a huge draw of 10^5 particles from the prior kernel, is displayed[2] in Figure 5.8.

2. The attentive reader will have remarked that the KLD estimates for the first few iterations are not displayed. This is due to numerical instabilities in these estimates. This is easily explained by the presence of the logarithm, which is not defined for quantities numerically rounded to zero due to finite

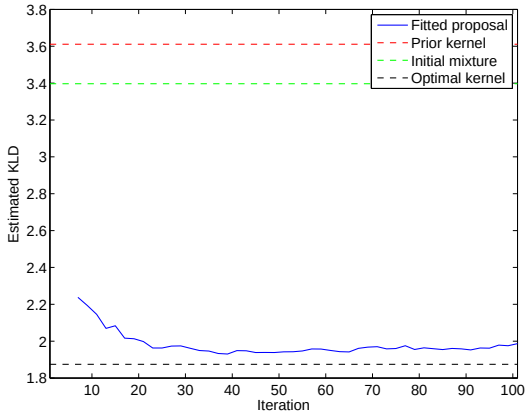


Figure 5.8: Evolution of the KLD in the presence of an outlying observation, $L = 100$, $N_0 = 1,000$, $N^\ell = 200$, stepsize $1/\sqrt{L}$

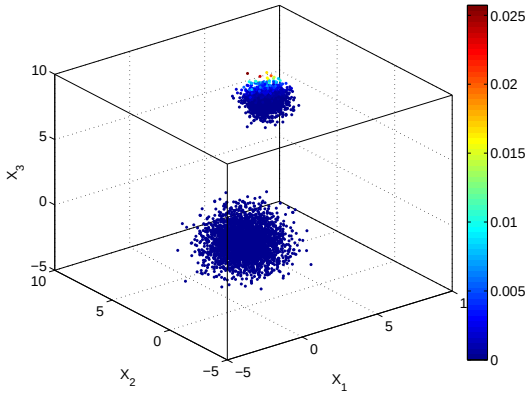
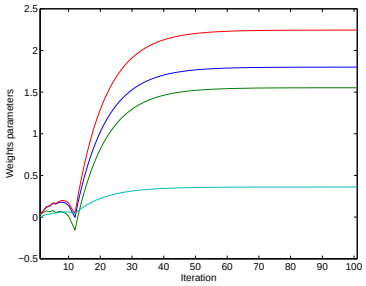
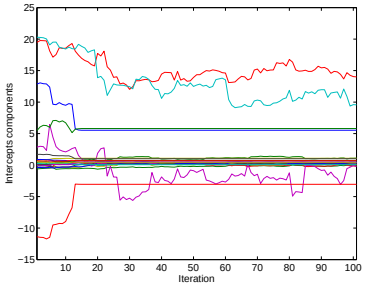


Figure 5.9: Optimal adjustment weights $\Psi^*(\xi_i)$ for the original cloud of $20,000$ particles $\{(\xi_i, 1)\}_{i=1}^{20,000}$ in presence of an outlying observation

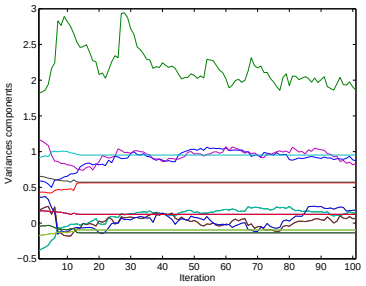
The decrease of the KLD is extremely rapid, and, as in the case of a highly likely observation, a few iterations (here, around 20) suffice to reach a satisfying quality, as is confirmed by the curves of proportions in Figures 5.7b–5.7d show the improvement after iterations 1, 10, and 50. This is a noticeable achievement, given the extremely outlying nature of the observation and the very poor starting fit (prior kernel). Besides, the remaining proportion of particles with null weights seen in Figure 5.7d, as well as the remaining KLD in Figure 5.8, are explained by the highly concentrated optimal adjustment weights displayed in Figure 5.9: only the ancestors in the very extreme region of the original cloud do have non neglectable optimal weights.



(a) Components of the logistic weight parameters β^ℓ.



(b) Components of the intercept matrices M^ℓ.



(c) Components of the covariance matrices Σ^ℓ.

Figure 5.10: Fitted parameters θ^ℓ over 100 iterations of the algorithm.

We also display in Figure 5.10 the evolution of the parameters over the 100 iterations of the algorithm, for the same outlying observation. The evolution of the logistic weights parameters corresponds to a tightening of the transition domain between the components in the ancestor space, that is, achieving a sharper partition. The regression parameters (components of the regression and variance matrices) are also quickly sta-

machine precision – here, the density of the proposal evaluated at some (few) of the (numerous) reference particles sampled from the prior kernel is below the machine precision, though mathematically positive. The algorithm in itself does not suffer from the same instability, as it does not depend on these estimates but on closed form maximization of the intermediate quantities.

bilized. Actually, the remaining variations concern the component to which fewer particles are assigned, thus having a greater variance of its parameter estimates. Moreover, we use a fixed stepsize in the stochastic gradient scheme: using a decreasing stepsize would suppress this behavior. However, again, this subtleties, highly relevant when seeking an accurate optimum, are of absolutely no concern to our importance sampling proposal adaptation setting.

5.5.3 Brownian motion driving a Bessel process observed in noise

As a more intricate example, we consider here filtering the Brownian motion underlying a Bessel process observed in noise, also known as range-only filtering of a Gaussian random walk. The state and measurement equations are given in such case by

$$\begin{cases} X_{k+1} &= X_k + V_k \\ Y_k &= \|X_k\| + W_k \end{cases} \tag{5.5.19}$$

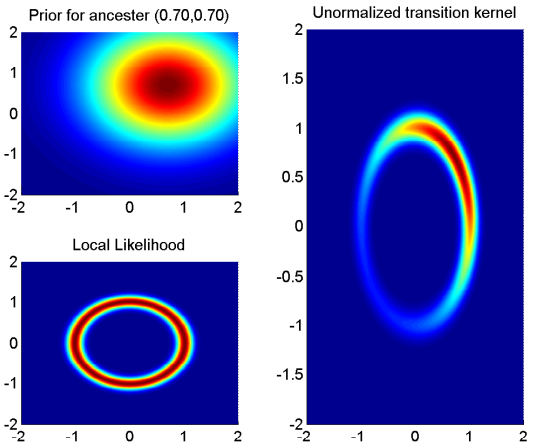
where $V_k \overset{i.i.d.}{\sim} \mathcal{N}_{d_s}(0, \Sigma_x)$, $W_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_x^2)$, and $\|X\| = \sqrt{\sum_{i=1}^{d_s} X_{(i)}^2}$ is the L_2 norm on the state space $\mathsf{X} := \mathbb{R}^{d_s}$. With the notation defined in Section 5.5.1, we consider state-particles, that is $\boldsymbol{\Xi} = \tilde{\boldsymbol{\Xi}} = \mathsf{X} = \mathbb{R}^{d_s}$, and define

$$q(\xi, \tilde{\xi}) := \mathcal{N}_{d_s}\left(\tilde{\xi}; \xi, \Sigma_x\right) \tag{5.5.20}$$

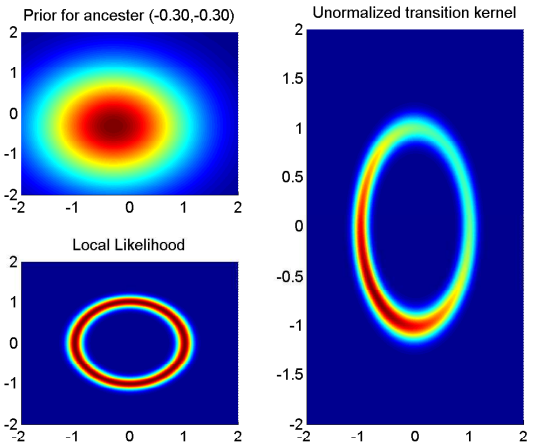
the density of the prior kernel Q, and

$$g(\tilde{\xi}, y) := \mathcal{N}\left(y; \left\|\tilde{\xi}\right\|, \sigma_y^2\right) \tag{5.5.21}$$

the local likelihood.



(a) Densities evaluated for the ancestor located at the mode $(0.7, 0.7)$ of the prior distribution.



(b) Densities evaluated for the ancestor $(-0.3, -0.3)$.

Figure 5.11: Bessel model densities for two particles in the original weighted sample: (a) center of the cloud, and (b) bottom left quadrant. Red is highest, blue is lowest. Note the impact on the unnormalized transition kernel, whose mass shifts consequently.

In our numerical illustration, we let the dimension $d_s = 2$, variance matrix $\Sigma_x = I_{d_s}$ the identity matrix in $\mathbb{R}^{d_s}$, thus having a diffuse prior, compared to the local likelihood

variance $\sigma_y^2 = 0.01$ which corresponds to informative observations. We consider, for illustration purposes, the following setting. The prior distribution of the state at time 0 is Gaussian $\mathcal{N}_{d_s}((0.7, 0.7), 0.5I_2)$ and the observation is set to be $Y_0 = 1.01$. Figure 5.11 displays two examples of the prior kernels $Q(\xi, \cdot)$, local likelihood $g(\cdot, Y_1)$, and the unnormalized transition kernel $L(\xi, \cdot) = g(\cdot, Y_1)Q(\xi, \cdot)$ for $\xi = X_0$ and $\xi = X_0 - (1, 1)$, showing the variety and the nonlinearity of the shapes the optimal kernel can take. Despite the low variance of the measurement noise, as the hidden state is observed range-only, the state equations provides most of the information about the bearings.

We now examine the behavior of the mixture of experts proposal. The original weighted sample $\{(\xi_i, \omega_i)\}_{i=1}^N$ consists of $N = 20,000$ i.i.d. samples drawn from the prior distribution of the state at time 0, hence having uniform weights. They can be seen as the result of a previous iteration of SMC leading up to this timestep.

We will follow the evolution of the adapted mixture by plotting its density for three ancestors, respectively situated in X_0 and at two standard deviations away from it on a line joining the origin, i.e. closer and further from the origin. The unnormalized optimal kernel (i.e. the product of the prior kernel and the local likelihood) for these three ancestors is plotted Figure 5.12.

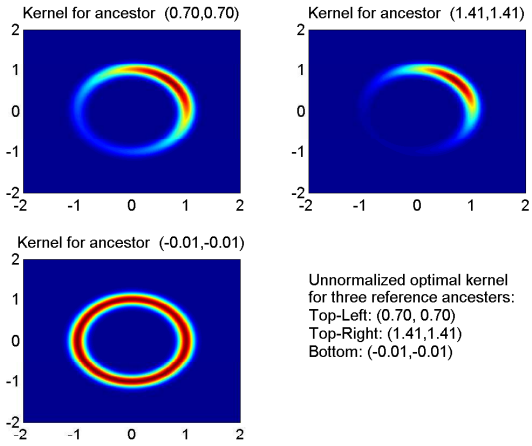


Figure 5.12: Density of the unnormalized optimal kernel evaluated for three distinct ancestors.

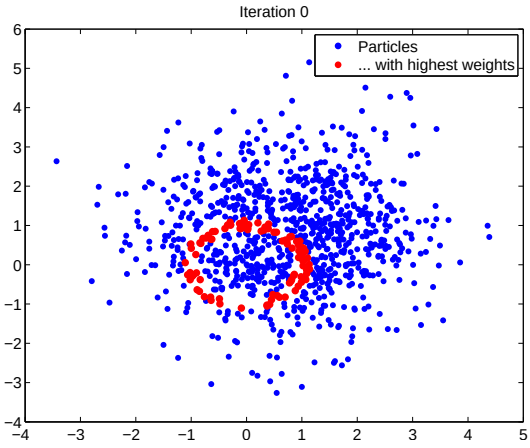
We initialize the algorithm by proposing $N_1 = 1,000$ particles $\{(\tilde{\xi}_i, \tilde{\omega}_i)\}_{i=1}^{N_1}$ according to the prior kernel. The 100 particles with the highest weights are plotted in Figure 5.13a, along with their ancestors in Figure 5.13b. The histogram of the weights, along with their 90% empirical quantile, is available in Figure 5.14b. Most of the particles have negligible weights, and a few particles have comparatively large weights. This is confirmed by Figure 5.14a, which displays the curve of proportions – on the same principle than Figure 5.3c described page 167. It can be seen that only 20% of the proposed particles carry the total mass, which is a dramatic waste of computational time.

This entails two conclusions about the need for adaptation in this model:

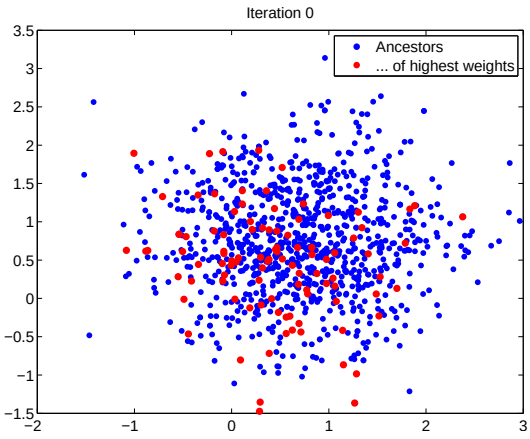
– the support of the proposal distribution is over-spread and does not match the support of the target distribution; it is clear that one may anticipate some benefits from adapting the proposal kernel, as the region of the state space that receives the most relevant particles is not large, and

– adapting the adjustment weights of the original weighted sample would not improve much the performance of the APF, as the ancestors leading to the particles

with significant importance weights are evenly spread in the original weighted sample, and are not located in a specific region of the space.

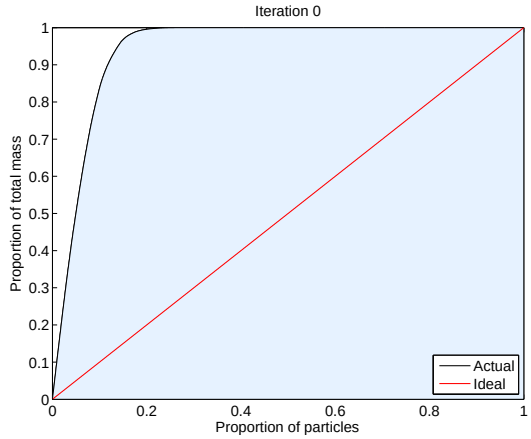


(a) Cloud of $1,000$ particles proposed with the prior kernel. The 100 particles with the highest importance weights are plotted in red.

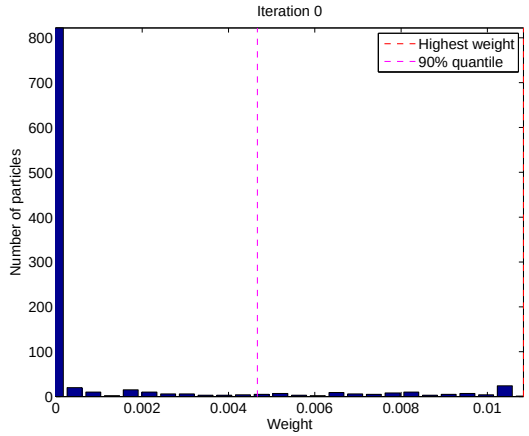


(b) Cloud of the ancestors of the $1,000$ particles proposed with the prior kernel. The ancestors of the 100 particles with the highest importance weights are plotted in red.

Figure 5.13: Particles proposed using the prior kernel.



(a) Proportion mass/weights of a weighted sample proposed using the prior kernel Q



(b) Histogram of these weights

Figure 5.14: Importance weights of $1,000$ particles proposed using the prior kernel Q.

Based on these $1,000$ particles, a preliminary cycle of adaptation of the parameter of the mixture of experts is done by using conditional probabilities from the mixture displayed in Figure 5.15 whose components are chosen to be independent of the ancestor – i.e. only the constant term is non-null.

The kernel obtained after this first adaptation step for the same three reference ancestors, along with the corresponding partition, is plotted in Figure 5.16. It is then used to propose 200 particles, which will serve for the IS approximation of μ_{aux} in iteration 2. The histogram of the weights, along with the sorted weights, 20 particles with the highest weights, and ancestors of these 20 particles, are displayed in Figure 5.17.

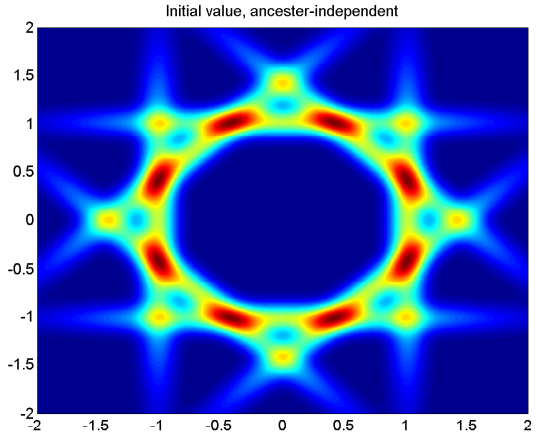
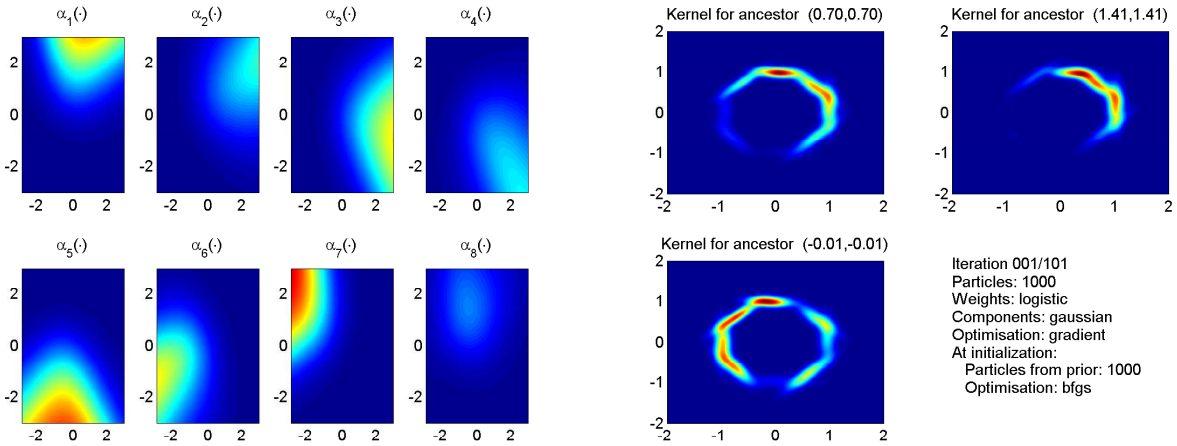


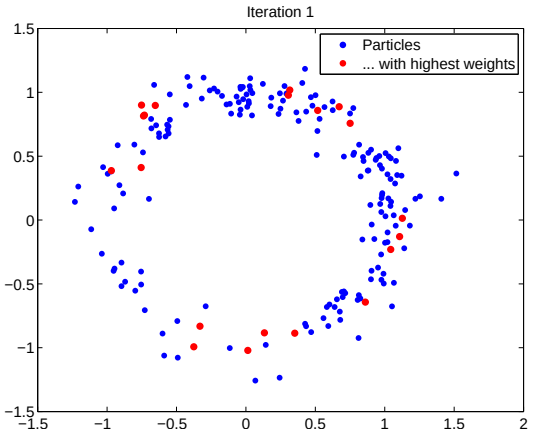
Figure 5.15: Ancestor-independent mixture of regression used for the conditional probabilities to obtain the first fit.



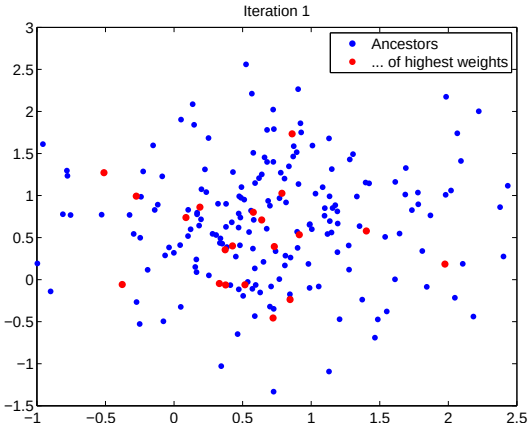
(a) Partition after the initial step

(b) Mixture after the initial step

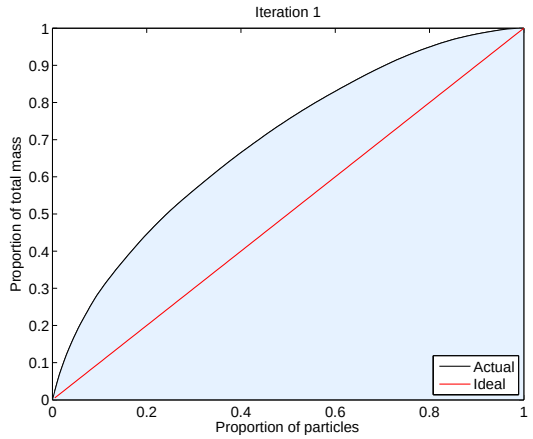
Figure 5.16: Result of the initial step (Iteration 1) of adaptation: partition of the ancestor space, along with mixture kernel evaluated for 3 distinct ancestors.



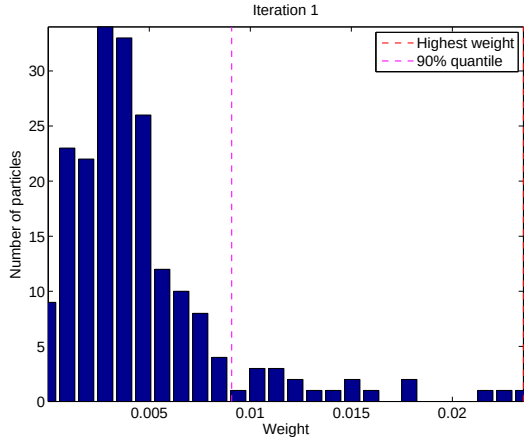
(a) The 20 particles with the highest weights are plotted in red.



(b) Ancestors of these 200 particles. The ancestors of the 20 particles with the highest weights are plotted in red.



(c) Proportion mass/weights

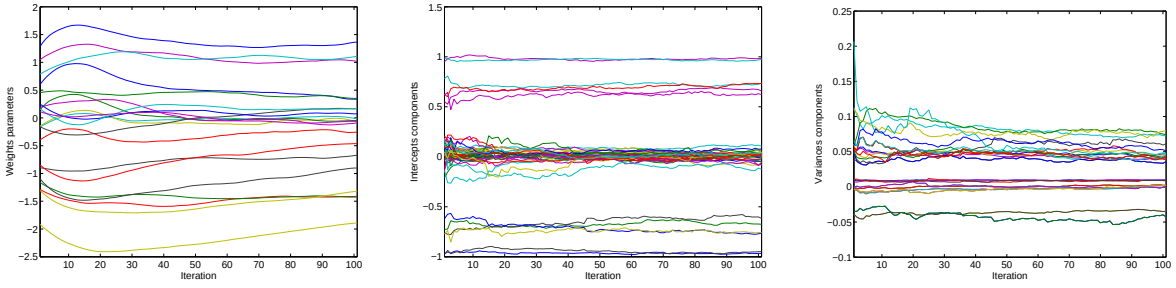


(d) Histogram of the weights.

Figure 5.17: Cloud of 200 particles proposed with the mixture of experts resulting of the first step of adaptation of Gaussian experts.

As can be seen in these two figures, a single step of optimization leads to much more uniformly distributed importance weights, i.e. to a much better proposal. We now proceed to 100 iterations, and see how the weighted sample and the proposal kernel evolve. Figure 5.18 displays the evolution of all the estimated parameters over the 100 iterations, and Figure 5.19 plots the partitions, kernels, sorted weights, histogram of the weights, and particles of highest weights, for iterations 15, 50, and 100.

From this, it can be seen that after very few steps, the fit only varies slightly. Even the very first few iterations could serve as much more efficient proposals than the prior kernel. It appears that the parameters of the logistic weights β^ℓ are those requiring the longest run of iterations before stabilizing. This comes as no surprise, as they do not benefit the closed form updates of the regression parameters: we need to resort stochastic gradient as closed-form updates are not available, as described in Section 5.3.1, and therefore achieve a slower increase.



(a) Components of the logistic weight parameters β^ℓ.

(b) Components of the intercept matrices M^ℓ.

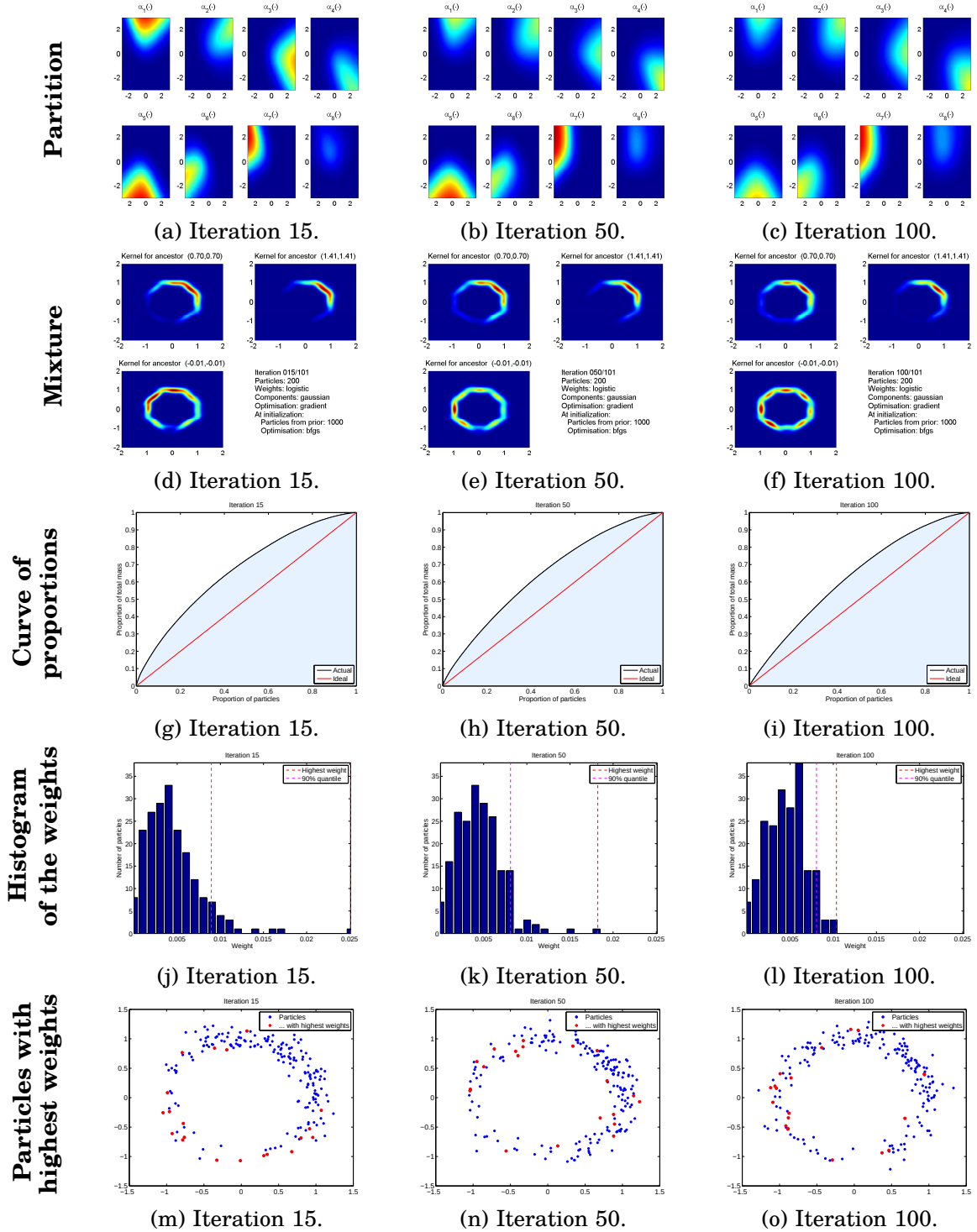
(c) Components of the covariance matrices Σ^ℓ.

Figure 5.18: Fitted parameters θ^ℓ over 100 iterations of the algorithm.

Figure 5.21a displays the decrease of the KLD over the iterations. The estimates of the KLD are obtained as in Section 5.5.2 by estimating the expectation in (5.1.9) with a reference cloud of $50,000$ particles from the prior, and approximating the optimal quantities by numerical integration. As expected, the KLD is decreasing at almost every iteration, and a small number of iterations suffices to dramatically decrease it.

Remark 5.5.2. Note that in this case, the KLD is almost entirely cancelled by adapting the optimal kernel, without optimizing the adjustment weights: the optimal kernel, even with uniform adjustment weights, leads to an almost null KLD. This means that, in contrast to the linear Gaussian model of Section 5.5.2 – as seen for example in Figure 5.8 –, optimum efficiency can be achieved by the sole use of an adapted proposal kernel, without any requirement of adjustment weights adaptation. This is in full concordance with our first thoughts based on Figure 5.13. Actually, an approximation of the optimal adjustment weights, obtained by numerical integration, is displayed in Figure 5.20 for the $20,000$ original particles. Although these weights are not strictly speaking uniform, no region of the input space is critically assigned null weights. This is, once again, the contrary of the linear Gaussian case which was depicted in Figure 5.1c. It outlines the peculiarities of the two sides of adaptation, as depicted in Chapter 4 and in the present chapter.

Finally, as an improvement, we compare the two initialization schemes, BFGS-based and Gradient-based. BFGS, which we have used up to this point, requires several evaluations (up to 176 in our setting !) of the intermediate quantity $Q_2(\beta, \theta^0)$ in the sole line-search – i.e. in the bracketing and sectioning phases as described in (Fletcher,



(a) Iteration 15. (b) Iteration 50. (c) Iteration 100.

(d) Iteration 15. (e) Iteration 50. (f) Iteration 100.

(g) Iteration 15. (h) Iteration 50. (i) Iteration 100.

(j) Iteration 15. (k) Iteration 50. (l) Iteration 100.

(m) Iteration 15. (n) Iteration 50. (o) Iteration 100.

Figure 5.19: Iterations 15, 50, and 100 (final) of the adaptation of Gaussian experts with logistic weights.

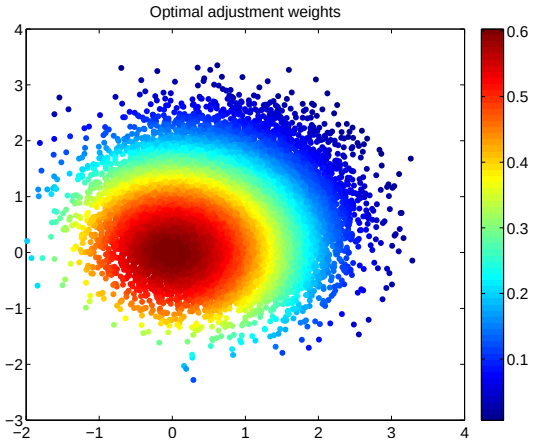
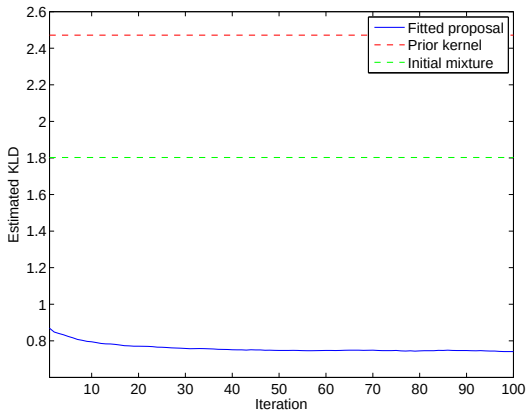
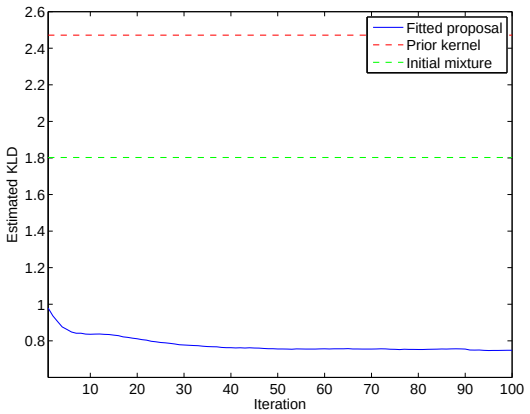


Figure 5.20: Optimal adjustment weights evaluated for the $20,000$ original particles.

1987, Algorithms 2.4 and 2.6) – even with very low precision requirements. Therefore, even though it offers quite an accurate starting point, the use of BFGS for the initial step can be criticized. It is interesting to have a look at the use of gradient descent from the very first iteration, i.e. is replacing the BFGS with one step in the direction of the gradient evaluated at the initialization point. This is signficantly less expensive than the BFGS approach, as a single sum over the particles is needed. Figure 5.21 displays the evaluation of the KLD for the two distinct approaches. Although the gradient approach leads to a slightly worse result after the first iteration, this is compensated in a few steps. Therefore, using the gradient method at the first iteration rather than BFGS indeed starts with a smaller accuracy, but is much faster, and a few iterations compensate for this loss of initial precision for still lower computational cost.



(a) Evolution of the KLD when using BFGS as the optimization method at first step.



(b) Evolution of the KLD when using gradient as the optimization method at first step.

Figure 5.21: Evolution of the KLD for the two possible initialization steps. Note the best result of initialization by BFGS, as well as the small number of iterations needed by the gradient approach to compensate for this worse beginning.

5.5.4 Multivariate tobit model

We now turn to a multivariate dynamic *tobit* model partially observed

$$X_k = A\,X_{k-1} + U_k \tag{5.5.22}$$
$$Y_k = \max\left(B^T\,X_k + V_k, 0\right)\ , \tag{5.5.23}$$

where $\mathsf{X} = \mathbb{R}^2$, A is a 2×2 matrix, and $B \in \mathbb{R}^2$, so that $Y_k \in \mathbb{R}$. The random variables U_k and V_k are independent Gaussian random variables with covariance matrix Σ_U and variance σ_v^2, respectively. The hidden state is partially observed as only the sum of its components is observed. Additionally, the observation suffers from left-censorship. Indeed, if the true state X_1 lies below the line of equation

$$\Delta = \left\{x \in \mathbb{R}^2 \text{ s.t. } B^T\,x = 0\right\}\ ,$$

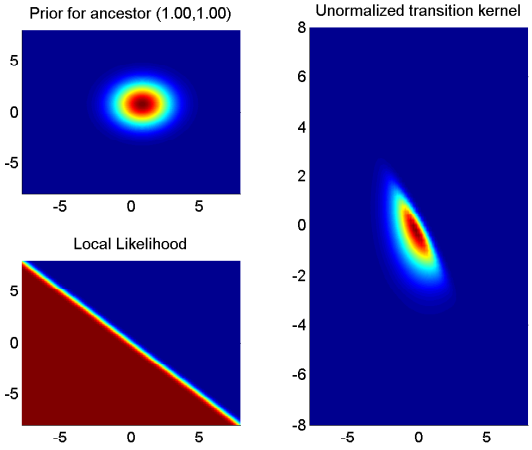
it is highly likely (depending on the variance σ_v^2 of the observation noise) that $y = 0$, and thus that the conditional likelihood is constant positive in the region below Δ and almost null in the region above Δ.

We chose the parameters

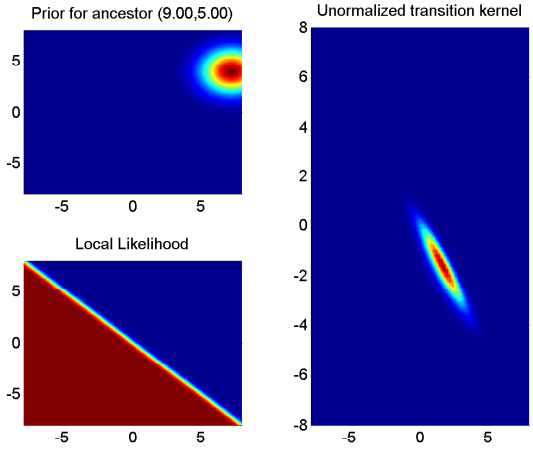
$$A = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.8 \end{pmatrix}\ , \qquad \Sigma_U = 2I_2\ , \qquad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}\ , \qquad \sigma_V^2 = 0.1\ , \tag{5.5.24}$$

so that Δ is here the line of equation $x_2 = -x_1$. We will focus on one filtering step occuring at time 1. Our knowledge on the hidden state at time 0 is $X_0 \sim \mathcal{N}_2\left((1,1)^T, 10I_2\right)$, hence the swarm of $N = 20,000$ ancestors $\{\xi_i, 1\}_{i=1}^N$ where $\xi_i \overset{i.i.d}{\sim} \mathcal{N}_2\left((1,1)^T, 10I_2\right)$. The components of the ancestors roughly ranges from -12 to 13. Note that, in this setting, half of the ancestors are such that the mean $B^T\,A^T\,\xi_i$ of the prior kernel is situated below Δ, i.e. they are quite likely to give birth to offspring corresponding to censored observation. Precisely, at time 1, we observe $y < 0$, i.e. a censored observation. As the observation noise $\sigma_v^2 = 0.1$ is small (compared to the variance terms of the prior kernel given by Σ_U), the local likelihood acts much as an indicator function, taking almost constant maximal value in the region of the state-space below Δ, and almost constant minimal value above Δ, with a narrow transition region around Δ. This is illustrated in the three examples of Figure 5.22; the prior kernel, the local likelihood (independent of the ancestor, hence common to the four subfigures), and the unnormalized optimal kernel, that is, the product of both, of are all plotted for a distinct ancestor in each subfigure.

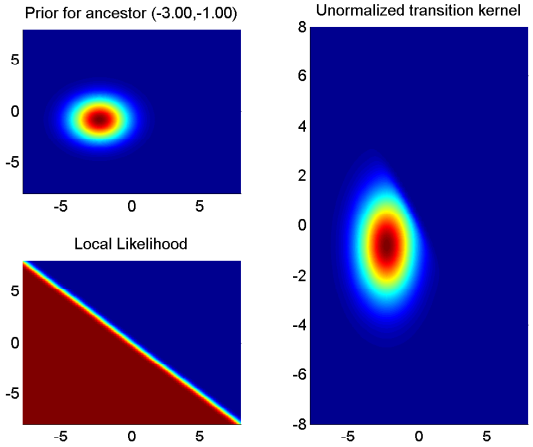
Note that, for ancestors ξ such that $A\,\tilde{\xi}$ is below the line Δ, the prior kernel, though slightly over-dispersed in its upper right tail compared to the optimal kernel, pretty much matches the optimal kernel. However, for the other ancestors, the mismatch between the prior kernel and the optimal kernel can be really significant, leading to very irrelevant offsprings. This is reflected in Figure 5.23a which displays $1,000$ particles from proposed by the APF using the prior kernel and uniform adjustment weights (as, once again, we do not adapt these adjustment weights), highlighting those with highest weights, as well as their respective ancestors in Figure 5.23b. It comes as no surprise that, as outlined above, the prior kernel is only relevant for the lower half of the ancestors, and that many particles are of very low weights. Figure 5.24 displays the proportion of particles (sorted by decreasing order of weights) against the proportion of the total mass, along with the histogram of the weights: only 40% of the particles significantly contribute to the estimator, all of them having almost equal weight. The remaining 60% have practically null weights. Though obviously adaptation



(a) Ancestor (1,1), center of the ancestors sample. Unnormalized optimal kernel (and hence also the normalized optimal kernel) is the prior kernel, truncated in its middle.

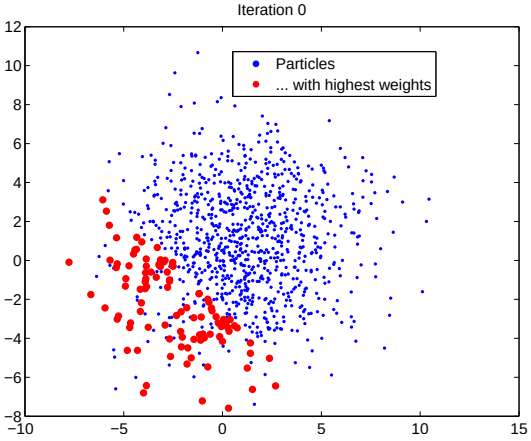


(b) Ancestor (9,5), top right of the ancestors sample. The unnormalized optimal kernel differs widely from the prior kernel, as only the very far tails of the latter match non-null local likelihood.

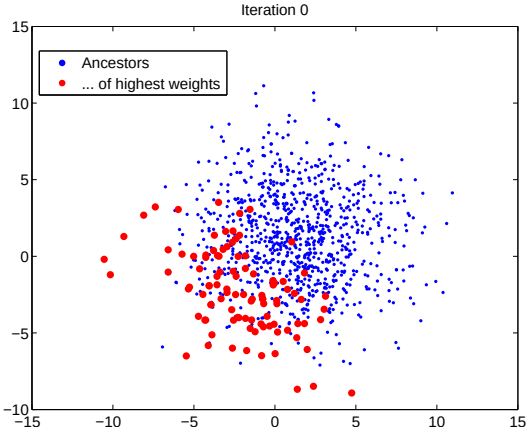


(c) Ancestor (-3,-1), bottom left of the ancestors sample. The unnormalized optimal kernel almost matches the prior, save for a truncation in the upper right tail.

Figure 5.22: Tobit model densities for three particles in the original weighted sample. Red is highest, blue is lowest.

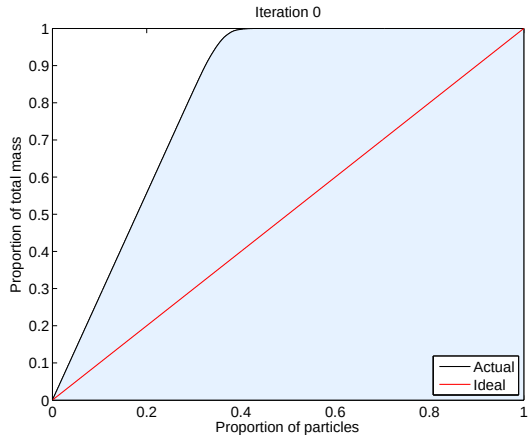


(a) Cloud of $1,000$ particles proposed with the prior kernel. The 100 particles with the highest importance weights are plotted in red.

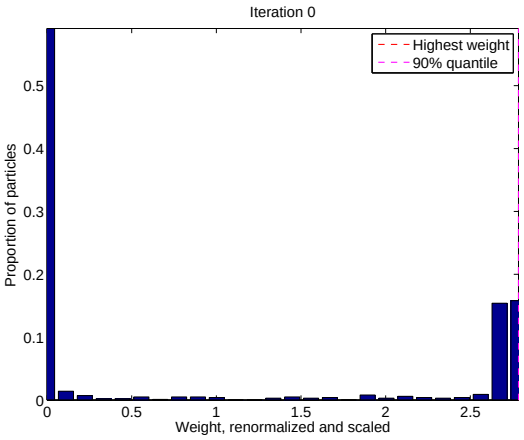


(b) Cloud of the ancestors of the $1,000$ particles proposed with the prior kernel. The ancestors of the 100 particles with the highest importance weights are plotted in red.

Figure 5.23: Particles proposed using the prior kernel for the Dynamic Tobit model.



(a) Curve of proportions.



(b) Histogram of these weights.

Figure 5.24: Importance weights of $1,000$ particles proposed using the prior kernel Q for the dynamic Tobit model.

of the adjustment weights could be used here, as some ancestors are very likely to have low optimal weights (typically those in the far top right tail of the sample), we can still expect that the adaptation will increase the proportion of particles contributing to the estimator.

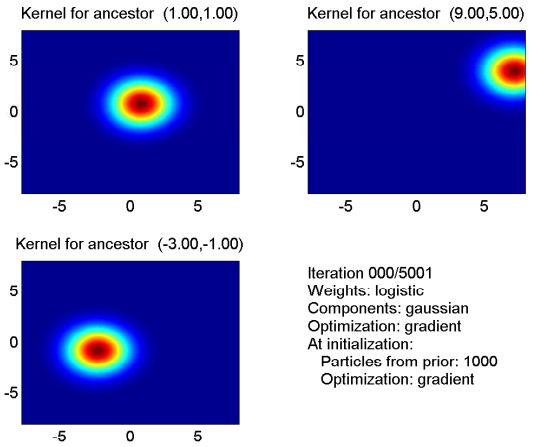
Here, again, we adapt a mixture of Gaussian regression experts, with logistic mixture weights. As the preliminary analysis of the problem indicates that partitioning the ancestor space into two regions could be sufficient, the peculiar shape of the truncated Gaussians lead us to opt for $d = 3$ components to benefit from some added degree of freedom, and better fit the shapes of these target distributions. The initial fit used for evaluating the first conditional probabilities is chosen to be as close to the prior kernel as possible. We choose the logistic weights to be uniform over the whole ancestor space, that is $\beta^0 = 0_{\mathbb{R}^{p+1}(d-1)}$. The variance matrices are all set to the variance Σ_U of the prior kernel, and the regression matrices M_j^0 are chosen so that the components "circle" the prior matrix with an offset of half a standard deviation, that is, more precisely, for any $j \in \{1, \dots, d\}$,

$$M_j^0 := \left(A, \operatorname{chol}(\Sigma_U)\, \mathcal{R}\left(\frac{2j\pi}{d}\right) \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} \right)$$
$$\text{where } \mathcal{R}(t) := \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \text{ is the rotation matrix of angle } t. \quad (5.5.25)$$

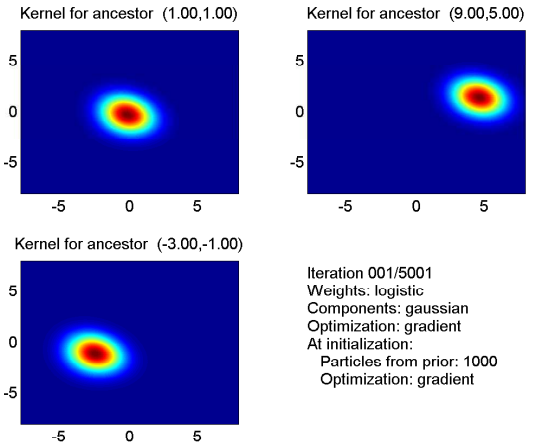
Figure 5.25a plots the initial mixture for the same three reference ancestors as in Figure 5.22. Note that it would make no sense to let all the regression matrices equal to $(A, (0,0)^T)$, because equating the three components would obviously be a degenerate case preventing the algorithm from separating them.

The very first step of the algorithm is based on a sample of $1,000$ particles from the prior kernel, uses the fit described above for the first conditional probabilities, and uses one step of gradient ascent for the parameters of the logistic weights. We stress that, for the latters, we did not use BFGS optimization algorithm, but stuck to the very lightweight step of gradient descent. This leads to the kernel plotted in Figure 5.25b. Note how the mass of the kernels has shifted towards the mass of the optimal kernel. Figure 5.26 displays the logistic partition obtained after this first iteration. Clearly, the partition is suboptimal: it does not correspond to the clear-cut partition that the study of the model lead us to expect. However, studying the proportion graphic of the weights resulting from a sample from this kernel, in Figure 5.27, is very encouraging, especially when compared to Figure 5.24a: almost 60% of the particles now contribute significantly to the estimator, instead of 40% with the prior kernel, and this, only after one single iteration of our algorithm, based on a small sample of $1,000$ particles.

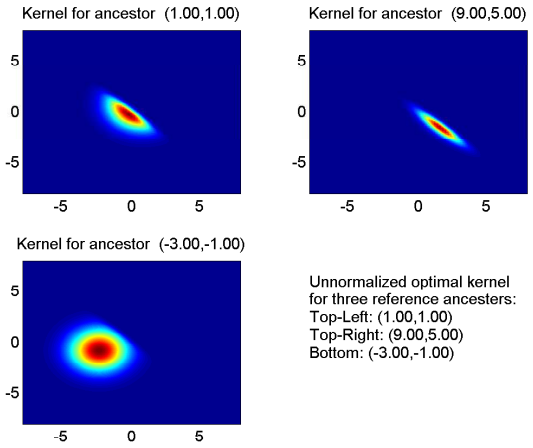
We now consider the behavior of our algorithm on a much longer run, say, $L = 5,000$ iterations. We set the constant stepsize to $\tau_\ell = 1/\sqrt{L}$. In order to provide a sound comparative basis, we chose to approximate the optimal kernel l^* – which is not available in closed form, differing in that of Section 5.5.2 – by obtaining a very close approximation of the optimal adjustment weights $\Psi^*(\xi_i)$ for each particle, as defined in (5.1.4), by a costly numerical integration over a grid. The optimal kernel l^* is then obtained by dividing the unnormalized optimal kernel by this optimal weight. We stress that the reference KLD corresponds to the optimal kernel *with uniform adjustment weights*; in other words, we do not take into account these numerically integrated optimal weights, save for computing the optimal kernel – not when sampling the ancestors indices. Of course, taking these optimal adjustment weights into account when sampling the ancestors indices would lead to a null KLD.



(a) Initial kernel, evaluated for three distinct ancestors, used for computing the first conditional probabilities of the particles generated from the prior. Note that it is chosen to be very similar to the prior kernel.



(b) Kernel obtained after one iteration of the algorithm, based on gradient optimization (not BFGS), and on basis of $1,000$ particles from the prior kernel, for the same three ancestors. Note how the mass has moved towards its optimal setting displayed in Figure 5.25c.



(c) Unnormalized optimal kernel, evaluated for the same three ancestors. This is a duplicate of the right-most subplots of Figure 5.22, presented here for reference purpose.

Figure 5.25: Initial fit and result of the first iteration of the algorithm. Kernels evaluated for the same three ancestors as in Figure 5.22. The rightmost subplots of this latter figure are here displayed in the bottom plot on the same scale as the above plots, for comparison purposes.

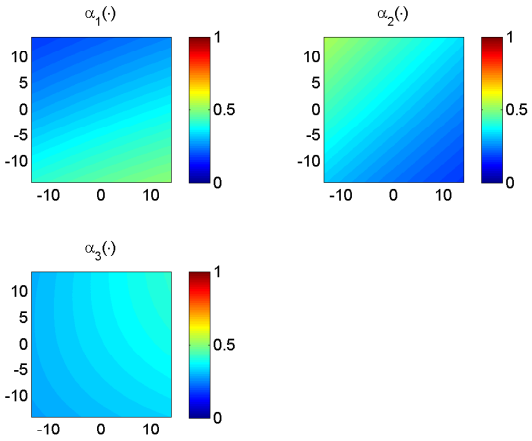


Figure 5.26: Partition of the ancestor space obtained after the first iteration of the adaptation algorithm, corresponding to the parameters β^1.

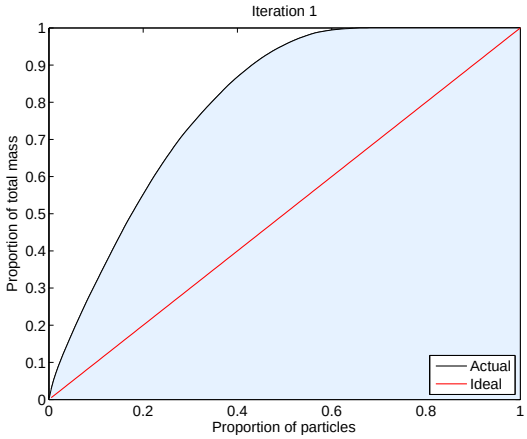


Figure 5.27: Curve of proportions of a weighted sample proposed using the parameters θ^1 resulting from the first iteration of the adaptation algorithm.

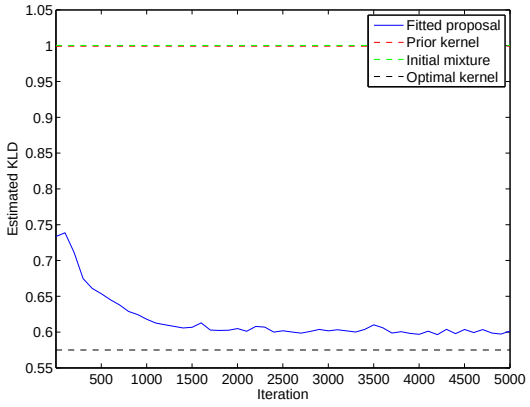
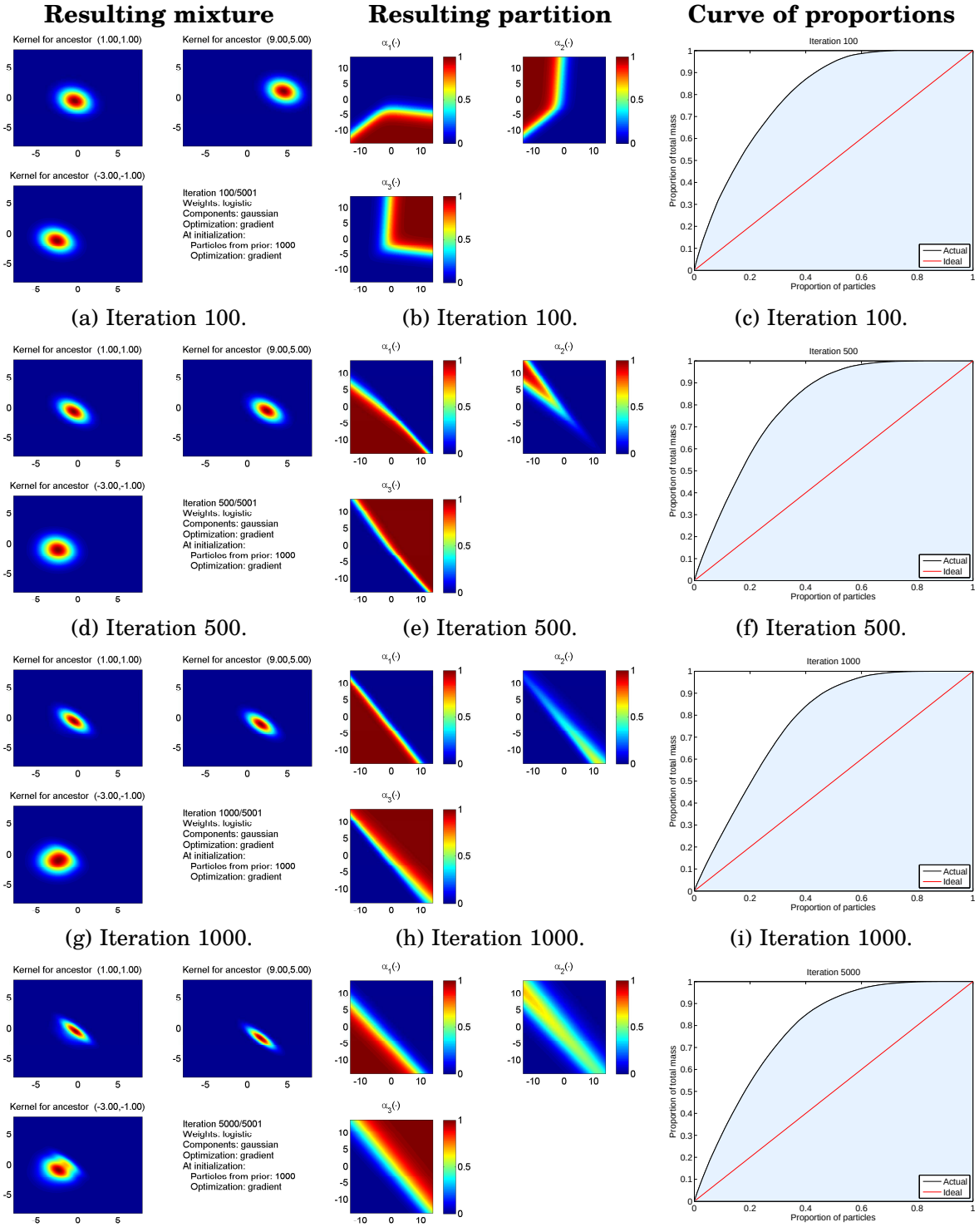


Figure 5.28: Evolution of the KLD over $5,000$ of our algorithm, for the Tobit model.



(a) Iteration 100. (b) Iteration 100. (c) Iteration 100.

(d) Iteration 500. (e) Iteration 500. (f) Iteration 500.

(g) Iteration 1000. (h) Iteration 1000. (i) Iteration 1000.

(j) Iteration 5000. (k) Iteration 5000. (l) Iteration 5000.

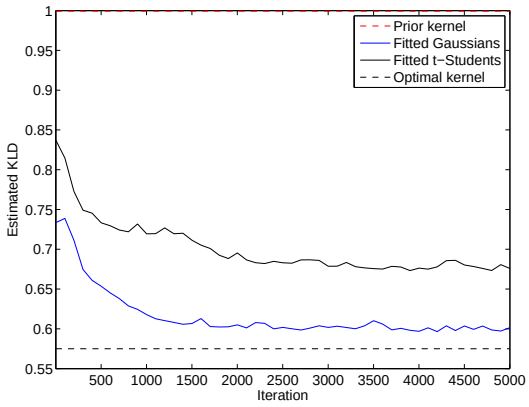
Figure 5.29: Results of our algorithm at iterations 100, 500, $1,000$, $2,000$, and $5,000$. For each iteration ℓ the resulting mixture kernel $R_{\boldsymbol{\theta}^\ell}$ is plotted for the same reference ancestors of Figure 5.25c, along with the partition of the ancestors space corresponding to logistic parameters $\boldsymbol{\beta}^\ell$, and the curve of proportions.

Estimation of the KLD is then achieved, in contrast with the two former examples, by approximating (5.1.9), including the terms that do not depend on the proposal kernel r, making use of the optimal adjustment weights numerically obtained. As usual, though, the expectations w.r.t. μ_{aux} are approximated with a massive sample of $50,000$ particles using the prior kernel. Our reference KLD, 0.57, is hence the exact minimum that can be achieved in this model by only adapting the proposal kernel, without adapting the adjustment weights. Any additional decrease can only be obtained by using adapted adjustment weights, as detailed in Chapter 4. The results displayed in Figure 5.28 shows an impressive decrease of the KLD of the fitted kernel in the first step, then a slower decrease, with convergence seemingly achieved after $1,500$ iteration, where we are extremely close to the optimal KLD achievable. Although this number of iterations is bigger than for the Bessel example of Section 5.5.3, we stress that we do not, by any means, require our algorithm to achieve convergence: it can be stopped at any time, it will in any case be an improvement over the prior kernel, and most of the improvement is achieved in the very first steps. Figure 5.29 displays several iterations of the adaptation algorithm. The lessons from it are the following:

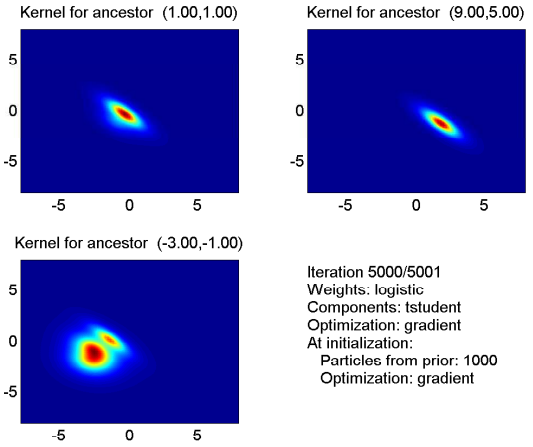
- Few improvement is obtained after $1,000$ iterations, and 500 iterations already provide a very descent fit.
- After a few steps that make the divisions sharper, the partition of the ancestor space converges to a state with only two major regions, with the third region bringing a slight additional flexibility to fit the transition between the two.
- Running the algorithm as far as $5,000$ iterations brings the number of particles with non-null weights up to 75%, though this last result is likely to be negligible compared to the wins that could be obtained by adapting the adjustment weights,

To sum it up, this case again shows that very few iterations (even a *single* iteration) can bring a tremendous improvement over the prior kernel, even with a very naïve initialization scheme, and that, though the first few iterations suffice for practical purposes, pushing the algorithm to convergence actually achieves an extremely close fit of the optimal kernel.

It is also interesting to compare these results, obtained for the mixture of Gaussian experts (described in Section 5.3.2), with the mixture of t-Student experts as described in Section 5.3.2. We consider the exactly same original population, model, observation, stepsize, etc., and set the number of degrees of freedom to $\nu = 3$. The estimation of the resulting evolution of the KLD is plotted in Figure 5.30a, along with the KLD estimation for the Gaussian case already shown in Figure 5.28. Though this latter decreases over time, it stays strictly higher than the KLD obtained in the Gaussian case. It comes as no surprise, as the t-Student family has heavier tails than the Gaussian, and that the Tobit model introduces (loosely speaking) truncated Gaussians. This can be seen on Figure 5.30b, which displays the mixture obtained after as long as $5,000$ iterations of the algorithm, to be compared with its Gaussian equivalent in Figure 5.29j – we do not plot the partition of the ancestor space achieved by the logistic weights in the t-Student case, as it is fundamentally similar to the one achieved in the Gaussian case, as displayed in Figure 5.29k. Because the heavier tails are compensated by a more accentuated mode, the mixture of t-Student experts is visibly less accurate to fit the target distribution. This is the trade-off to benefit from the added robustness of heavier tails – along with slightly increased computational costs, as the conditional probabilities are slightly more costly than in the Gaussian case. A last interesting observation can be made when comparing the target distribution $\hat{\mu}(\tilde{\xi})$ defined in (5.1.2) – which is nothing but the marginalization of the auxiliary target μ_{aux} by summing over the auxiliary variable I – with the marginalization of the proposal $\pi_{\text{aux}}^{\boldsymbol{\theta}}$, i.e. $\sum_{i=1}^{N} \pi_{\text{aux}}^{\boldsymbol{\theta}}(i, \tilde{\xi})$, for a long-



(a) Comparison of the evolution of the KLD for the Gaussian experts and the t-Student expert, over $5,000$ iterations of the algorithm.



(b) Fit achieved after $5,000$ iterations of the algorithm fitting a mixture of t-Student experts, to be compared with Figure 5.29j. We do not display the partition of the ancestor space, as it is very similar to the one obtained in the Gaussian case.

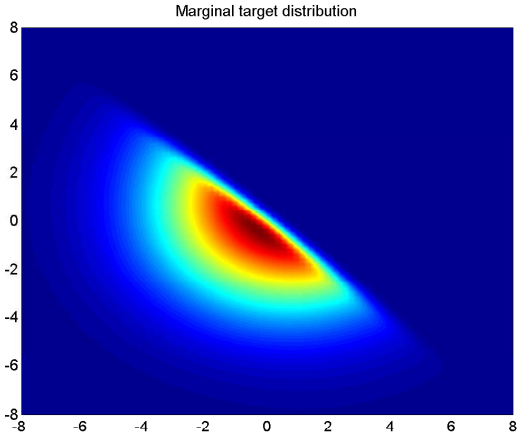
Figure 5.30: Result of $5,000$ iterations of the algorithm, fitting a mixture of t-Student experts on the Tobit model.

term result (at iteration $5,000$) of the Algorithm in both the Gaussian and t-Student cases. Figure 5.31 is achieved by a costly sum over all the $20,000$ ancestor particles and a grid-based evaluation. It shows that, first, the marginal proposal obtained with either the Gaussian or the t-Student experts is very similar to the distribution obtained with the optimal kernel and uniform weights, and that the noticeable difference with the marginal target $\hat{\mu}$ is due to the adjustment weights, which are uniform, instead of taking their optimal form whose numerical approximation (as describded above) is depicted in Figure 5.31f. As discussed at the beginning of the section, the particles in the top-right region should have very small adjustment weights: the fact that they carry, in our approximation, as much mass as those in the bottom-left region explains the high value of the marginalized proposal density in the center of the updated state space.

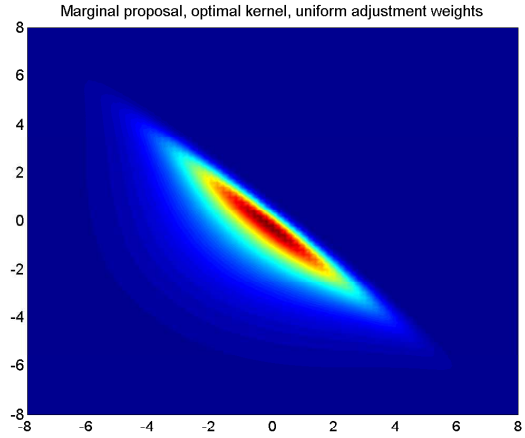
5.6 Future work and conclusion of the dissertation

In the last chapter of this thesis, we have built new algorithms relying on the results from Chapter 3, to approximate the optimal kernel at a given timestep of the APF. After a detailed description of the path leading to the construction of the algorithm, we have shown on several examples how this relatively simple algorithm can improve the criterion considered (decreased the KLD), both in models exhibiting strong non-linearity or distinct behaviors depending on the location of the ancestor particle. Additionally, we have illustrated that very few iterations are enough for achieving a high drop in the KLD, thus minimizing the computational overhead.

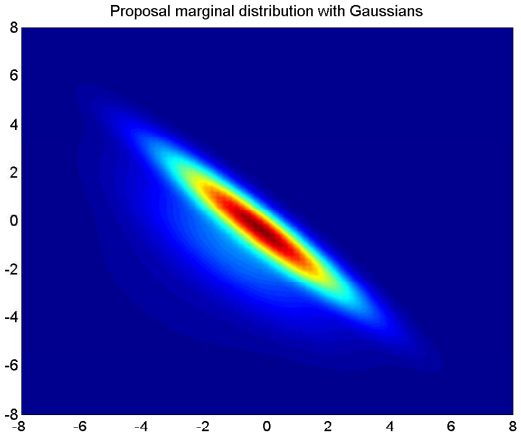
Of course, this APF frameworks also fits without modification the cases where the resampling occurs at random stopping times, e.g. when a quality criterion is above a given threshold. The optimal kernel would then be taken pathwise on the part of the trajectory (or "block", in the terms of Doucet et al. (2006)) between now and the



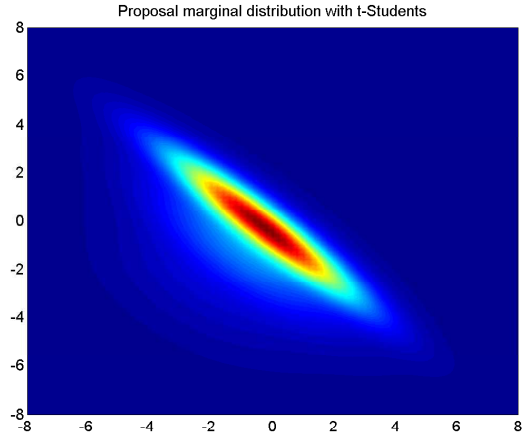
(a) Marginal target distribution $\hat{\mu}$.



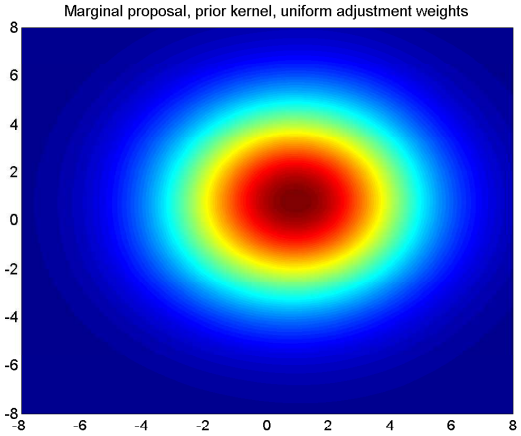
(b) Marginal of the proposal distribution obtained with the optimal kernel and uniform adjustment weights.



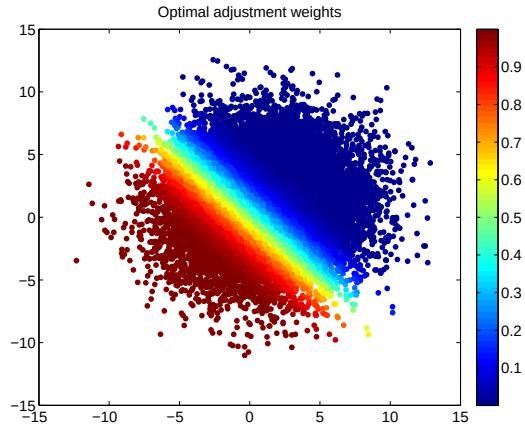
(c) Marginal of the proposal distribution obtained after $5,000$ iterations of the algorithm, fitting a mixture of Gaussian experts.



(d) Marginal of the proposal distribution obtained after $5,000$ iterations of the algorithm, fitting a mixture of t-Student experts.



(e) Marginal of the proposal distribution obtained using the prior kernel, with uniform weights.



(f) Cloud of the original particles, colored proportionally to their optimal adjustment weights as approximated by numerical integration.

Figure 5.31: Marginalized distributions, evaluated on a grid conditionally on the same original sample of $20,000$ particles, and optimal adjustment weights.

last timestep when resampling was triggered. Several variants are possible, depending on how many components of this block would be kept fixed – optimizing the few last components might be sufficient to reduce the KLD.

Considerable work remains to be done along the lines exposed here. First and foremost, proving the convergence of the algorithm is under work and will be published in a companion paper. As stated in Section 5.4, this analysis is based on concepts similar to those used in analyzes of Monte Carlo (*MCEM*, Fort and Moulines (2003)) and Stochastic Approximation (*SAEM*, Delyon et al. (1999); Kuhn and Lavielle (2004); Andrieu et al. (2005)) variants of the EM algorithm. On a more practical side, a comparison of the trade-off performance/cost with the much simpler schemes (particle EKF, particle UKF, Laplace approximation) would be of considerable interest – even though such an analysis, which can only be empirical, cannot escape the bias of implementation details and model-specific considerations. Last, but not least, interesting refinements can be considered, either stemming from the stochastic approximation community or from importance sampling. More precisely, we are currently considering the possible use of so-called *relaxation* (as named in Rubinstein and Kroese (2004)) scheme, also known for a longer time as *chaining* (see Evans and Swartz (1995) and references therein) in classical importance sampling, or as *progressive correction* (Musso et al. (2001)). This consists in targeting a sequence of distributions that progressively evolve to the target distribution, each one getting more intricate than the preceding. A typical example could be to use a decreasing sequence of observation noises or an increasing sequence of exponents (lower than unity) of the local likelihood, thus building a smooth transition between a simple update of the original distribution by the prior kernel (easy to approximate) and the target distribution. This would make perfect sense in our naturally iterative algorithm, and we expect that even a most simplistic scheme fitting within the small number of iterations we do could help tackle the most difficult problems having very narrow posterior distribution.

We believe, and this will conclude the present dissertation, that bringing to SMC methods, as described in Chapters 1 and 2, rigorously and theoretically analyzed quality criterion such as those from Chapter 3, amounts to opening a Pandora box – or a horn of Amalthea – from which the new insights and algorithms exposed in Chapters 4 and 5 are only the first escapees.

APPENDIX A

Elements of asymptotic analysis

Contents

In this appendix, we briefly recapitulate the techniques developed Douc and Moulines (2008).

A.1 Notations

All the random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A state space X is said to be *general* if it is equipped with a countably generated σ-field $\mathcal{X}$. For a general state space X, we denote by $\mathcal{P}(\mathsf{X})$ the set of probability measures on $(\mathsf{X}, \mathcal{X})$ and $\mathbb{B}(\mathsf{X})$ (resp. $\mathbb{B}^+(\mathsf{X})$) the set of all $\mathcal{B}(\mathsf{X})/\mathcal{B}(\mathbb{R})$-measurable (resp. non-negative) functions from X to $\mathbb{R}$ equipped with the Borel σ-field $\mathcal{B}(\mathbb{R})$. A subset $\mathsf{C} \subseteq \mathsf{X}$ is said to be *proper* if the following conditions are satisfied. **(i)** C is a linear space: for any f and g in C and reals α and β, $\alpha f + \beta g \in \mathsf{C}$; **(ii)** if $g \in \mathsf{C}$ and f is measurable with $|f| \leq |g|$, then $|f| \in \mathsf{C}$; **(iii)** for all c, the constant function $f \equiv c$ belongs to C.

For any $\mu \in \mathcal{P}(\mathsf{X})$ and $f \in \mathbb{B}(\mathsf{X})$ satisfying $\int_{\mathsf{X}} \mu(dx)|f(x)| < \infty$, $\mu(f)$ denotes $\int_{\mathsf{X}} f(x)\mu(dx)$. Let X and Y be two general state spaces. A kernel V from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$ is a map from $\mathsf{X} \times \mathcal{Y}$ into $[0, 1]$ such that for each $A \in \mathcal{Y}$, $x \mapsto V(x, A)$ is a nonnegative bounded measurable function on X and for each $x \in \mathsf{X}$, $A \mapsto V(x, A)$ is a measure on $\mathcal{Y}$. We say that V is finite if $V(x, \mathsf{Y}) < \infty$ for any $x \in \mathsf{X}$; it is Markovian if $V(x, \mathsf{X}) \equiv 1$ for any $x \in \mathsf{X}$. For any function $f \in \mathbb{B}(\mathsf{X} \times \mathsf{Y})$ such that $\int_{\mathsf{Y}} V(x, dy)|f(x, y)| < \infty$ we denote by $V(\cdot, f)$ or $Vf(\cdot)$ the function $x \mapsto V(x, f) := \int_{\mathsf{Y}} V(x, dy)f(x, y)$. For ν a measure on $(\mathsf{X}, \mathcal{X})$, we denote by νV the measure on $(\mathsf{Y}, \mathcal{Y})$ defined for any $A \in \mathcal{Y}$ by $\nu V(A) = \int_{\mathsf{X}} \nu(dx)V(x, A)$.

Throughout the paper, we denote by Ξ, μ a probability measure on $(\Xi, \mathcal{B}(\Xi))$, $\{M_N\}_{N\geq 0}$ be a sequence integer-valued random variables, C a proper subsets of Ξ. We approximate the probability measure μ by points $\xi_{N,i} \in \Xi$, $i = 1, \ldots, M_N$ associated to non-negative weights $\omega_{N,i} \geq 0$.

Definition A.1.1. A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be *consistent* for the probability measure μ and the (proper) set C if for any $f \in \mathsf{C}$, as $N \to \infty$, $\Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \mu(f)$ and $\Omega_N^{-1} \max_{i=1}^{M_N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0$ where $\Omega_N = \sum_{i=1}^{M_N} \omega_{N,i}$.

This definition of weighted sample consistency is similar to the notion of *properly weighted sample* introduced in Liu and Chen (1998). The difference stems from the smallness condition which states that the contribution of each individual term in the sum vanishes in the limit as $N \to \infty$.

We denote by γ a finite measure on $(\Xi, \mathcal{B}(\Xi))$, A and W be proper sets of Ξ, and σ a real non negative function on A, and $\{a_N\}$ be a non-decreasing real sequence diverging to infinity.

Definition A.1.2. A weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ on Ξ is said to be *asymptotically normal* for $(\mu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$ if

$$a_N \Omega_N^{-1} \sum_{i=1}^{M_N} \omega_{N,i} \{f(\xi_{N,i}) - \mu(f)\} \xrightarrow{\mathcal{D}} \mathcal{N}\{0, \sigma^2(f)\} \qquad \text{for any } f \in \mathsf{A}\,, \tag{A.1.1}$$

$$a_N^2 \Omega_N^{-2} \sum_{i=1}^{M_N} \omega_{N,i}^2 f(\xi_{N,i}) \xrightarrow{\mathbb{P}} \gamma(f) \qquad \text{for any } f \in \mathsf{W} \tag{A.1.2}$$

$$a_N \Omega_N^{-1} \max_{1 \le i \le M_N} \omega_{N,i} \xrightarrow{\mathbb{P}} 0 \tag{A.1.3}$$

Note that these definitions implicitly implies that the sets C, A and W are proper.

To analyze the sequential Monte Carlo methods discussed in the introduction, we now need to study how the importance sampling and the resampling steps affect the consistent or / and asymptotically normal weighted sample.

A.2 Importance sampling

We will show that the importance sampling step transforms a weighted sample consistent (or asymptotically normal) for a distribution ν on a general state space $(\Xi, \mathcal{B}(\Xi))$ into a weighted sample consistent (or asymptotically normal) for a distribution μ on $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$. Let L be a Markov kernel from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ such that for any $f \in \mathbb{B}(\tilde{\Xi})$,

$$\mu = \frac{\nu L f}{\nu L(\tilde{\Xi})}\,. \tag{A.2.1}$$

We wish to transform a weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ targeting the distribution ν on $(\Xi, \mathcal{B}(\Xi))$ into a weighted sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ targeting μ on $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ where $\tilde{M}_N = \alpha M_N$ (α denoting the number of offsprings of each particle). The use of multiple offsprings has been suggested by Rubin (1987): when the importance sampling step is followed by a resampling step, an increase in the number of distinct particles will increase the number of distinct particles **after** the resampling step. In the sequential context, this operation is a practical mean for contending particle impoverishment. These offsprings are proposed using a Markov kernel denoted R from $(\Xi, \mathcal{B}(\Xi))$ to $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$. We assume that for any $\xi_N \in \Xi$, the probability measure $L(\xi_N, \cdot)$ on $(\tilde{\Xi}, \mathcal{B}(\tilde{\Xi}))$ is absolutely continuous w.r.t. R, which we denote $L(\xi_N, \cdot) \ll R(\xi_N, \cdot)$ and define

$$W(\xi_N, \tilde{\xi}_N) := \frac{dL(\xi_N, \cdot)}{dR(\xi_N, \cdot)}(\tilde{\xi}_N) \tag{A.2.2}$$

The new weighted sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ is constructed as follows. We draw new particle positions $\{\tilde{\xi}_{N,j}\}_{j=1}^{\tilde{M}_N}$ conditionally independently given

$$\mathcal{F}_{N,0} := \sigma\left(M_N, \{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}\right) , \tag{A.2.3}$$

with distribution given for $i = 1, \ldots, M_N$, $k = 1, \ldots, \alpha$ and $A \in \mathcal{B}(\tilde{\Xi})$ by

$$\mathbb{P}\left(\tilde{\xi}_{N,N,\alpha(i-1)+k} \in A \middle| \mathcal{F}_{N,0}\right) = R(\xi_{N,i}, A) \tag{A.2.4}$$

and associate to each new particle positions the importance weight:

$$\tilde{\omega}_{N,\alpha(i-1)+k} = \omega_{N,i} W(\xi_{N,i}, \tilde{\xi}_{N,N,\alpha(i-1)+k}) , \tag{A.2.5}$$

for $i = 1, \ldots, M_N$ and $k = 1, \ldots, \alpha$. The importance sampling step is *unbiased* in the sense that, for any $f \in \mathcal{B}(\tilde{\Xi})$ and $i = 1, \ldots, M_N$,

$$\sum_{j=\alpha(i-1)+1}^{\alpha i} \mathbb{E}\left[\tilde{\omega}_{N,j} f(\tilde{\xi}_{N,j}) \middle| \mathcal{F}_{N,j-1}\right] = \alpha \omega_{N,i} L(\xi_{N,i}, f) , \tag{A.2.6}$$

where for $j = 1, \ldots, \tilde{M}_N$, $\mathcal{F}_{N,j} := \mathcal{F}_{N,0} \vee \sigma(\{\tilde{\xi}_{N,l}\}_{1 \le l \le j})$. The following theorems state conditions under which the importance sampling step described above preserves the weighted sample consistency. Denote by

$$\tilde{\mathsf{C}} := \{f \in \mathsf{L}^1(\tilde{\Xi}, \mu),\ L(\cdot, |f|) \in \mathsf{C}\} . \tag{A.2.7}$$

Theorem A.2.1. *Assume that the weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C) and that $L(\cdot, \tilde{\Xi})$ belongs to C. Then, the set $\tilde{\mathsf{C}}$ defined in (A.2.7) is a proper set and the weighted sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ defined by (A.2.4) and (A.2.5) is consistent for $(\mu, \tilde{\mathsf{C}})$.*

We now turn to prove the asymptotic normality. Define

$$\tilde{\mathsf{A}} := \left\{f \ : L(\cdot, |f|) \in \mathsf{A}, R(\cdot, W^2 f^2) \in \mathsf{W}\right\} , \ \tilde{\mathsf{W}} := \left\{f : R(\cdot, W^2 |f|) \in \mathsf{W}\right\} . \tag{A.2.8}$$

Theorem A.2.2. *Suppose that the assumptions of Theorem A.2.1 hold. Assume in addition that the weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is asymptotically normal for $(\nu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$, and that the function $R(\cdot, W^2)$ belongs to W .*

Then, the sets $\tilde{\mathsf{A}}$ and $\tilde{\mathsf{W}}$ defined in (A.2.8) are proper and the weighted sample $\{(\tilde{\xi}_{N,i}, \tilde{\omega}_{N,i})\}_{i=1}^{\tilde{M}_N}$ is asymptotically normal for $(\mu, \tilde{\mathsf{A}}, \tilde{\mathsf{W}}, \tilde{\sigma}, \tilde{\gamma}, \{a_N\})$ with $\tilde{\gamma}(f) := \alpha^{-1}\gamma R(W^2 f)/(\nu L(\tilde{\Xi}))^2$ and

$$\tilde{\sigma}^2(f) := \sigma^2\{L[f - \mu(f)]\}/(\nu L(\tilde{\Xi}))^2 + \alpha^{-1}\gamma R\{[W[f - \mu(f)] - R(\cdot, W[f - \mu(f)])]^2\}/(\nu L(\tilde{\Xi}))^2 .$$

Remark A.2.1. Reweighting the particles without moving them is a particular case of importance sampling. Thus, Theorem A.2.1 and A.2.2 may also apply in this context.

A.3 Resampling

Resampling converts a weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ targeting a distribution ν $(\Xi, \mathcal{B}(\Xi))$ into an equally weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ targeting the same distribution ν. The resampling step is an essential ingredient in the sequential context because it removes particles with small weights and produces multiple copies of particles with large weights. Denote by $G_{N,i}$ the number of times the i-th particle is replicated. The number of particles after resampling $\tilde{M}_N = \sum_{i=1}^{M_N} G_{N,i}$ is supposed to be an $\mathcal{F}_{N,0}$-measurable integer-valued random variables, where $\mathcal{F}_{N,0}$ is given in (A.2.3); it might differ from the initial number of particles M_N, but will generally be a (deterministic) function of it. There are many different resampling procedures described in the literature. The simplest is the *multinomial* resampling, in which the distribution of $(G_{N,1}, \ldots, G_{N,M_N})$ conditionally to $\mathcal{F}_{N,0}$ is multinomial:

$$(G_{N,1}, \ldots, G_{N,M_N}) | \mathcal{F}_{N,0} \sim \mathsf{Mult}\left(\tilde{M}_N, \{\Omega_N^{-1}\omega_{N,i}\}_{i=1}^{M_N}\right) . \tag{A.3.1}$$

Another possible solution is the *deterministic plus residual multinomial resampling*, introduced in Liu and Chen (1995). Denote by $\lfloor x \rfloor$ the integer part of x and by $\langle x \rangle$ denotes the fractional part of x, $\langle x \rangle := x - \lfloor x \rfloor$. This scheme consists in retaining at least $\lfloor \Omega_N^{-1}\tilde{M}_N\omega_{N,i} \rfloor$, $i = 1, \ldots, M_N$ copies of the particles and then reallocating the remaining particles by applying the multinomial resampling procedure with the residual importance weights defined as $\left\langle \tilde{M}_N\Omega_N^{-1}\omega_{N,i} \right\rangle$, i.e. $G_{N,i} = \lfloor \Omega_N^{-1}\tilde{M}_N\omega_{N,i} \rfloor + H_{N,i}$ where

$$\begin{aligned}(H_{N,1}, \ldots, H_{N,M_N}) | \mathcal{F}_{N,0} \\ \sim \mathsf{Mult}\left(\sum_{i=1}^{M_N} \left\langle \Omega_N^{-1}\tilde{M}_N\omega_{N,i} \right\rangle, \left\{ \frac{\left\langle \Omega_N^{-1}\tilde{M}_N\omega_{N,i} \right\rangle}{\sum_{i=1}^{M_N} \left\langle \Omega_N^{-1}\tilde{M}_N\omega_{N,i} \right\rangle} \right\}_{i=1}^{M_N}\right) .\end{aligned} \tag{A.3.2}$$

If the weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C), where C is a proper subset of $\mathbb{B}(\mathsf{X})$, it is a natural question to ask whether the uniformly weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ is consistent for ν and, if so, what an appropriately defined class of functions on Ξ might be. It happens that a fairly general result can be obtained in this case.

Theorem A.3.1. *Assume that the weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C). Then, the uniformly weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ obtained using either (A.3.1) or (A.3.2) is consistent for (ν, C).*

It is also sensible to strengthen the requirement of consistency into asymptotic normality, and prove that the resampling procedures (A.3.1) and (A.3.2) transform an asymptotically normal weighted sample for ν into an asymptotically normal sample for ν. We consider first the multinomial sampling algorithm. We define

$$\tilde{\mathsf{A}} := \left\{ f \in \mathsf{A},\ f^2 \in \mathsf{C} \right\} , \tag{A.3.3}$$

Theorem A.3.2. *Assume that*

(i) $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C) and asymptotically normal for $(\nu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$; in addition, $a_N^{-2}M_N \xrightarrow{\mathbb{P}} \beta^{-1}$ for some $\beta \in [0, \infty)$.

(ii) $\tilde{M}_N$ *is* $\mathcal{F}_{N,0}$*-measurable, where* $\mathcal{F}_{N,0}$ *is defined in* (A.2.3)*,* $\tilde{M}_N/M_N \xrightarrow{\mathbb{P}} \ell$ *where* $\ell \in [0, \infty]$.

Then $\tilde{\mathsf{A}}$ *is a proper set and the equally weighted particle system* $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ *obtained using* (A.3.1) *is asymptotically normal for* $(\nu, \tilde{\mathsf{A}}, \mathsf{C}, \tilde{\sigma}, \tilde{\gamma}, \{a_N\})$ *with* $\tilde{\sigma}^2(f) = \beta\ell^{-1}\mathbb{V}_\nu(f) + \sigma^2(f)$ *and* $\tilde{\gamma} = \beta\ell^{-1}\nu$.

The analysis of the deterministic plus multinomial residual sampling is more involved. To carry out the analysis, it is required to consider situations where the importance weights are a function of the particle position, i.e. $\omega_{N,i} = \Phi(\xi_{N,i})$, where $\Phi \in \mathbb{B}^+(\mathbf{\Xi})$. This condition is fulfilled in most applications of sequential Monte-Carlo methods and should therefore not be considered as a stringent limitation. For $\ell \in \mathbb{R}^+$, and ν a probability measure on $\mathbf{\Xi}$, define $\nu_{\ell,\Phi}$ the measure $\nu_{\ell,\Phi}(f) = \nu\left(\frac{\langle \ell\nu(\Phi^{-1})\Phi\rangle}{\ell\nu(\Phi^{-1})\Phi} f\right)$ for $f \in \mathbb{B}^+(\mathbf{\Xi})$.

Theorem A.3.3. *Assume that*

(i) $\{(\xi_{N,i}, \Phi(\xi_{N,i}))\}_{i=1}^{M_N}$ *is consistent for* (ν, C) *and asymptotically normal for* $(\nu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$*; in addition,* $a_N^{-2} M_N \xrightarrow{\mathbb{P}} \beta^{-1}$ *for some* $\beta \in [0, \infty)$.

(ii) $\tilde{M}_N$ *is* $\mathcal{F}_{N,0}$*-measurable, where* $\mathcal{F}_{N,0}$ *is defined in* (A.2.3)*,* $\tilde{M}_N/M_N \xrightarrow{\mathbb{P}} \ell$ *where* $\ell \in [0, \infty]$.

(iii) $\Phi^{-1} \in \mathsf{C}$*, and* $\nu\left(\ell\nu(\Phi^{-1})\Phi \in \mathbb{N} \cup \{\infty\}\right) = 0$.

Then, the uniformly weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ *obtained using* (A.3.2) *is asymptotically normal for* $(\nu, \tilde{\mathsf{A}}, \mathsf{C}, \tilde{\sigma}, \tilde{\gamma}, \{a_N\})$ *where* $\tilde{\mathsf{A}}$ *is given by* (A.3.3)*,* $\tilde{\gamma} := \beta\ell^{-1}\nu$*, and*

$$\tilde{\sigma}^2(f) := \beta\ell^{-1}\nu_{\ell,\Phi}\left\{(f - \nu_{\ell,\Phi}(f)/\nu_{\ell,\Phi}(\mathbb{1}))^2\right\} + \sigma^2(f) \quad \textit{for } f \in \tilde{\mathsf{A}}\,.$$

Remark A.3.1. Because $\left\langle \ell\nu(\Phi^{-1})\Phi \right\rangle / \ell\nu(\Phi^{-1})\Phi \leq 1$, for any $f \in \tilde{\mathsf{A}}$,

$$\nu_{\ell,\Phi}\left\{(f - \nu_{\ell,\Phi}(f)/\nu_{\ell,\Phi}(\mathbb{1}))^2\right\} = \inf_{c\in\mathbb{R}} \nu_{\ell,\Phi}\left\{(f-c)^2\right\} \leq \inf_{c\in\mathbb{R}} \nu\left\{(f-c)^2\right\} = \mathbb{V}_\nu(f)\,,$$

showing that the variance of the residual plus deterministic sampling is always lower than that of the multinomial sampling. These results extend (Chopin, 2004, Theorem 2) that derive an expression of the variance of the residual sampling in a specific case. Note however the assumption Theorem A.3.3-(iii) missing in the statement of (Chopin, 2004, Theorem 2). This assumption cannot be relaxed, as shown in (Douc and Moulines, 2008, Appdenix D)

A.4 Branching

Branching procedures have been considered as an alternative to resampling procedure (see Crisan et al. (1998), Del Moral and Miclo (2000), Crisan (2003) and (Del Moral, 2004, Chapter 11)); these procedures are easier to implement than resampling and are popular among practitioners. In the branching procedures, the number of times each particle is replicated $(G_{N,1}, \ldots, G_{N,M_N})$ are independent conditionally to $\mathcal{F}_{N,0}$ and are distributed in such a way that $\mathbb{E}\left[G_{N,i}|\mathcal{F}_{N,0}\right] = \tilde{m}_N \Omega_N^{-1} \omega_{N,i}$, $i = 1, \ldots, M_N$, where $\tilde{m}_N$ is the targeted number of particle, assumed to be a $\mathcal{F}_{N,0}$ random variables. Most

often, $\tilde{m}_N$ is chosen to be a deterministic function of the current number of particle M_N, e.g. $\tilde{m}_N = M_N$ or $\tilde{m}_N = N$ (in which case we target a "deterministic" number of particles). Contrary to the resampling procedures, the number of particles $\tilde{M}_N$ after branching is no longer $\mathcal{F}_{N,0}$-measurable, i.e. the actual number of particles $\tilde{M}_N$ is different from the targeted number $\tilde{m}_N$ and cannot be predicted before the branching numbers $\{G_{N,i}\}_{i=1}^{M_N}$ are drawn. There are of course many different ways to select the branching numbers. In the Poisson branching, the branching numbers $\{G_{N,i}\}_{i=0}^{M_N}$ are conditionally independent given $\mathcal{F}_{N,0}$ with Poisson distribution with parameters $\{\tilde{m}_N \Omega_N^{-1} \omega_{N,i}\}_{i=1}^{M_N}$,

$$\{G_{N,i}\}_{i=1}^{M_N} | \mathcal{F}_{N,0} \sim \bigotimes_{i=1}^{M_N} \mathrm{Pois}\left(\tilde{m}_N \Omega_N^{-1} \omega_{N,i}\right) , \tag{A.4.1}$$

where $\otimes$ denotes the tensor product of measures. Similarly, in the binomial branching, the branching numbers $\{G_{N,i}\}_{i=0}^{M_N}$ are conditionally independent given $\mathcal{F}_{N,0}$ with binomial distribution of parameters $\{(\tilde{m}_N, \Omega_N^{-1}\omega_{N,i})\}_{i=1}^{M_N}$

$$\{G_{N,i}\}_{i=1}^{M_N} | \mathcal{F}_{N,0} \sim \bigotimes_{i=1}^{M_N} \mathrm{Bin}\left(\tilde{m}_N, \Omega_N^{-1} \omega_{N,i}\right) , \tag{A.4.2}$$

The third branching algorithm, referred to as the Bernoulli branching algorithm, shares similarities with the deterministic plus residual multinomial sampling. In this case, for each i-th, $\lfloor \tilde{m}_N \Omega_N^{-1} \omega_{N,i} \rfloor$ are retained; to correct for the truncation, an additional particle is eventually added, i.e. $G_{N,i} = \lfloor \tilde{m}_N \Omega_N^{-1} \omega_{N,i} \rfloor + H_{N,i}$ where $\{H_{N,i}\}_{i=1}^{M_N}$ are conditionally independent given $\mathcal{F}_{N,0}$ with Bernoulli distribution of parameter $\{\langle \tilde{m}_N \Omega_N^{-1} \omega_{N,i} \rangle\}_{i=1}^{M_N}$,

$$G_{N,i} = \lfloor \tilde{m}_N \Omega_N^{-1} \omega_{N,i} \rfloor + H_{N,i} , \quad \{H_{N,i}\}_{i=1}^{M_N} | \mathcal{F}_{N,0} \sim \bigotimes_{i=1}^{M_N} \mathrm{Ber}\left(\langle \tilde{m}_N \Omega_N^{-1} \omega_{N,i} \rangle\right) . \tag{A.4.3}$$

As above, it may be shown that these branching algorithms preserve consistency.

Theorem A.4.1. *Assume that the weighted sample $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C). Then, $\tilde{M}_N / \tilde{m}_N \xrightarrow{\mathbb{P}} 1$ and the uniformly weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ obtained using either (A.4.1), (A.4.2) and (A.4.3) is consistent for (ν, C).*

We may also strengthen the conditions to establish the asymptotic normality. For the Poisson and the binomial branching, the asymptotic normality is satisfied under almost the same conditions than for the multinomial sampling (see Theorem A.3.2); in addition, the asymptotic variance of these procedures are equal.

Theorem A.4.2. *Assume that*

(i) $\{(\xi_{N,i}, \omega_{N,i})\}_{i=1}^{M_N}$ is consistent for (ν, C) and asymptotically normal for $(\nu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$; in addition, $a_N^{-2} M_N \xrightarrow{\mathbb{P}} \beta^{-1}$ for some $\beta \in [0, \infty)$.

(ii) $\tilde{m}_N$ is $\mathcal{F}_{N,0}$-measurable, where $\mathcal{F}_{N,0}$ is defined in (A.2.3), $M_N^{-1} \tilde{m}_N \xrightarrow{\mathbb{P}} \ell$ where $\ell \in [0, \infty]$.

Then the equally weighted particle system $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ obtained using either (A.4.1) or (A.4.2) is asymptotically normal for $(\nu, \tilde{\mathsf{A}}, \mathsf{C}, \tilde{\sigma}, \tilde{\gamma}, \{a_N\})$, with $\tilde{\mathsf{A}} := \{f, f^2 \in \mathsf{C} \cap \mathsf{W}\}$, $\tilde{\sigma}^2(f) = \beta \ell^{-1} \mathbb{V}_\nu(f) + \sigma^2(f)$, and $\tilde{\gamma} = \beta \ell^{-1} \nu$.

We now consider the case of the Bernoulli branching. As for the deterministic plus residual sampling, it is here required to assume that the weights are a function of the particle positions, i.e. $\omega_{N,i} = \Phi(\xi_{N,i})$.

Theorem A.4.3. *Assume that*

(i) $\{(\xi_{N,i}, \Phi(\xi_{N,i}))\}_{i=1}^{M_N}$ *is consistent for* (ν, C) *and asymptotically normal for* $(\nu, \mathsf{A}, \mathsf{W}, \sigma, \gamma, \{a_N\})$*; in addition,* $a_N^{-2} M_N \xrightarrow{\mathbb{P}} \beta^{-1}$ *for some* $\beta \in [0, \infty)$.

(ii) $\tilde{m}_N$ *is* $\mathcal{F}_{N,0}$*-measurable, where* $\mathcal{F}_{N,0}$ *is defined in* (A.2.3) *and* $\tilde{m}_N / M_N \xrightarrow{\mathbb{P}} \ell$ *where* $\ell \in [0, \infty]$,

(iii) $\Phi^{-1} \in \mathsf{C}$, *and* $\nu\left(\ell\nu(\Phi^{-1})\Phi \in \mathbb{N} \cup \{\infty\}\right) = 0$.

Then, the uniformly weighted sample $\{(\tilde{\xi}_{N,i}, 1)\}_{i=1}^{\tilde{M}_N}$ *defined by* (A.4.3)*) is asymptotically normal for* $(\nu, \tilde{\mathsf{A}}, \mathsf{C}, \tilde{\sigma}, \tilde{\gamma}, \{a_N\})$ *where* $\tilde{\mathsf{A}} := \{f \in \mathsf{A}, (1+\Phi) f^2 \in \mathsf{C}\}$, $\tilde{\gamma} := \beta \ell^{-1} \nu$ *and*

$$\tilde{\sigma}^2(f) := \beta \ell^{-1} \nu \left(\frac{\left\langle \ell\nu(\Phi^{-1})\Phi \right\rangle \left(1 - \left\langle \ell\nu(\Phi^{-1})\Phi \right\rangle\right)}{\ell\nu(\Phi^{-1})\Phi} (f - \nu f)^2 \right) + \sigma^2(f) \ , \quad f \in \tilde{\mathsf{A}} \ .$$

Remark A.4.1. Since $\frac{\left\langle \ell\nu(\Phi^{-1})\Phi \right\rangle \left(1 - \left\langle \ell\nu(\Phi^{-1})\Phi \right\rangle\right)}{\ell\nu(\Phi^{-1})\Phi} \leq 1$, the asymptotic variance of the Bernoulli branching is always lower than the asymptotic variance of the multinomial resampling. Compared with the deterministic-plus-residual sampling, the two quantities are not ordered uniformly w.r.t f.

Bibliographie / Bibliography

Akashi, H. and Kumamoto, H. (1977). Random sampling approach to state estimation in switching environment. *Automatica*, **13**, 429–434. 41

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall. 54, 100

Andrieu, C., Davy, M., and Doucet, A. (2003). Efficient particle filtering for jump Markov systems. Application to time-varying autoregressions. *IEEE Trans. Signal Process.*, **51**(7), 1762–1770. 93, 102, 104

Andrieu, C., Doucet, A., and Holenstein, R. (2009). Particle Markov chain Monte Carlo. Submitted. 16, 20

Andrieu, C., Moulines, ., and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, **44**(1), 283–312 (electronic). 18, 21, 22, 190

Arouna, B. (2004). Robbins-monro algorithms and variance reduction in finance. *Journal of Computational Finance*, **7**(2). 98, 159

Baum, L. E., Petrie, T. P., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**(1), 164–171. 37

Berzuini, C. and Gilks, W. R. (2001). Resample-move filtering with cross-model jumps. *In* A. Doucet, N. De Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, Springer. pp. 117–138. 79

Bollerslev, T., Engle, R. F., and Nelson, D. (1994). ARCH models. *In* R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, North-Holland. pp. 2959–3038. 53, 117

Cappé, O., Douc, R., Guillin, A., Marin, J. M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, **18**(4), 447–459. 16, 20, 97, 99, 156

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer. 9, 11, 16, 17, 20, 36, 79, 94, 100, 147

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). An improved particle filter for non-linear problems. *IEE Proc. Radar Sonar Navig.*, **146**, 2–7. 73, 93

Chen, M. and Chen, G. (2000). Geometric ergodicity of nonlinear autoregressive models withchanging conditional variances. *The Canadian Journal of Statistics*, **28**(3), 605–613. 117

Chen, R. and Liu, J. S. (2000). Mixture Kalman filter. *J. Roy. Statist. Soc. Ser. B*, **62**(3), 493–508. 41

Chen, Y., Xie, J., and Liu, J. (2005). Stopping-time resampling for sequential Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B*, **67**(2), 199–217. 16, 20

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Ann. Statist.*, **32**(6), 2385–2411. 77, 195

Cornebise, J., Moulines, E., and Olsson, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Stat. Comput.*, **18**(4), 461–480. 17, 21, 62, 91, 119, 120, 121, 125, 132, 138, 149, 166

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley. 103

Crisan, D. (2003). Exact rates of convergence for a branching particle approximation to the solution of the Zakai equation. *Ann. Probab.*, **31**(2), 693–718. 195

Crisan, D., Del Moral, P., and Lyons, T. (1999). Discrete filtering using branching and interacting particle systems. *Markov Process. Related Fields*, **5**(3), 293–318. 70

Crisan, D., Gaines, J., and Lyons, T. (1998). Convergence of a branching particle method to the solution of the Zakai equation. *SIAM J. Appl. Math.*, **58**(5), 1568–1590 (electronic). 195

de Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.*, **134**, 19–67. 95

Del Moral, P. (1996). Nonlinear filtering : interacting particle solution. *Markov Process. Related Fields*, **2**, 555–579. 15, 19

—— (2004). *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer. 9, 11, 15, 19, 77, 86, 88, 122, 147, 195

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *J. Roy. Statist. Soc. Ser. B*, **68**(3), 411. 16, 20, 35, 42

Del Moral, P. and Jacod, J. (2001). Interacting particle filtering with discrete observations. *In* A. Doucet, N. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, Springer. pp. 43–77. 15, 19, 60

Del Moral, P. and Miclo, L. (2000). Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. *In Séminaire de Probabilités, XXXIV*, vol. 1729 of *Lecture Notes in Math.*, Springer, Berlin. pp. 1–145. 195

Del Moral, P. and Miclo, L. (2001). Genealogies and increasing propagation of chaos for feynman-kac and genetic models. *Ann. Appl. Probab.*, **11**, 1166–1198. 15, 19

Del Moral, P., Patras, F., and Rubenthaler, S. (2006). Coalescent tree based functional representations for some Feynman-Kac particle models. Preprint. 16, 19

Delyon, B., Lavielle, M., and Moulines, E. (1999). On a stochastic approximation version of the EM algorithm. *Ann. Statist.*, **27**(1). 18, 21, 22, 159, 190

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**(1), 1–38 (with discussion). 18, 21, 153

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer. 51, 69

Devroye, L. and Klincsek, T. (1981). Average time behavior of distributive sorting algorithms. *Computing*, **26**, 1–7. 69

Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequentialMonte Carlo. *Ann. Statist.*, **36**. To appear. 9, 11, 16, 17, 20, 21, 34, 41, 77, 78, 81, 82, 84, 108, 109, 120, 126, 127, 128, 129, 130, 131, 132, 140, 158, 191, 195

Douc, R., Moulines, E., and Olsson, J. (2008). Optimality of the auxiliary particle filter. *Probab. Math. Statist.*, **28**(2). To appear. 16, 20, 77, 93, 103, 106, 108, 112, 120, 124, 131, 159

Doucet, A. and Andrieu, C. (2001). Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Trans. Signal Process.*, **49**(6), 1216–1227. 119

Doucet, A., Briers, M., and Senecal, S. (2006). Efficient block sampling strategies for sequential Monte Carlo methods. *J. Comput. Graph. Statist.*, **15**(3), 693. 16, 20, 79, 188

Doucet, A., De Freitas, N., and Gordon, N. (Eds.) (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York. 9, 11, 15, 16, 18, 19, 20, 22, 30, 41, 92, 93, 147

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, **10**, 197–208. 15, 19, 41, 52, 54, 92, 148

Eickhoff, J. C., Zhu, J., and Amemiya, Y. (2004). On the simulation size and the convergence of the Monte Carlo EM algorithm via likelihood-based distances. *Statist. Probab. Lett.*, **67**(2), 161–171. 114

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, **50**, 987–1007. 53

Evans, M. and Swartz, T. (1995). Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems. *Statist. Sci.*, **10**, 254–272. 30, 96, 97, 190

F. Septier, S. K. Pang, S. J. G. and Carmi, A. (2009). Tracking of coordinated groups using marginalised MCMC-based particle algorithm. *In Proc. of IEEE Aerospace Conference*, vol. to appear. Big Sky, Montana, vol. to appear. 15, 19

Fearnhead, P. (1998). *Sequential Monte Carlo methods in filter theory*. Ph.D. thesis, University of Oxford. 72

—— (2008). Computational methods for complex stochastic systems : a review of some alternatives to MCMC. *Stat. Comput.*, **18**, 151–171. 92, 93

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *J. Roy. Statist. Soc. Ser. B*, **69**(4), 590–605. 91

Fletcher, R. (1987). *Practical Methods of Optimization, second edition*. Wiley. 154, 160, 177

Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.*, **31**(4), 1220–1259. 18, 21, 99, 158, 190

Fox, D. (2003). Adapting the sample size in particle filters through KLD-sampling. *Int. J. Rob. Res.*, **22**(11), 985–1004. 15, 16, 19, 20, 92, 119

Geweke, J. (1989). Bayesian inference in econometric models using Monte-Carlo integration. *Econometrica*, **57**(6), 1317–1339. 30, 58, 95

Givens, G. and Raftery, A. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Amer. Statist. Assoc.*, **91**(433), 132–141. 97

Glynn, P. W. and Iglehart, D. (1989). Importance sampling for stochastic simulations. *Management Science*, **35**(11), 1367–1392. 30

Gordon, N., Salmond, D., and Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, **140**, 107–113. 9, 11, 15, 19, 30, 62, 64, 79, 92, 102, 117, 119, 124, 148

Hammersley, J. M. and Handscomb, D. C. (1965). *Monte Carlo Methods*. Methuen & Co., London. 30

Handschin, J. (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, **6**, 555–563. 29, 45

Handschin, J. and Mayne, D. (1969). Monte Carlo techniques to estimate the conditionnal expectation in multi-stage non-linear filtering. *In Int. J. Control*, vol. 9. vol. 9, pp. 547–559. 29, 45, 122, 147

Ho, Y. C. and Lee, R. C. K. (1964). A Bayesian approach to problems in stochastic estimation and control. *IEEE Trans. Automat. Control*, **9**(4), 333–339. 101, 162

Hu, X.-L., Schon, T. B., and Ljung, L. (2008). A basic convergence result for particle filtering. *IEEE Trans. Signal Process.*, **56**(4), 1337–1348. 91

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *J. Finance*, **42**, 281–300. 53

Hürzeler, M. and Künsch, H. R. (1998). Monte Carlo approximations for general state-space models. *J. Comput. Graph. Statist.*, **7**, 175–193. 78, 92, 101, 103

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models (with discussion). *J. Bus. Econom. Statist.*, **12**, 371–417. 53

Johannesson, G., Hanley, B., and Nitao, J. (2004). Dynamic Bayesian models via Monte Carlo-an introduction with examples. *Tech. Rep.*, UCRL-TR-207173, Lawrence Livermore National Lab., Livermore, CA (US). 15, 19

Johansen, A. and Doucet, A. (2008). A note on auxiliary particle filter. *Statistics and Probability Letters*, **78**(12), 1498–1504. 16, 20, 77, 80, 85, 120

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, **6**, 181–214. 18, 22, 150, 151, 152

Jordan, M. I. and Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural Networks*, **8**(9), 1409–1431. 18, 22, 150, 151, 152

Julier, S. J. and Uhlmann, J. K. (1997). A new extension of the Kalman filter to nonlinear systems. *In AeroSense : The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*. 55

Kailath, T., Sayed, A., and Hassibi, B. (2000). *Linear Estimation*. Prentice-Hall. 100

Kalman, R. E. and Bucy, R. (1961). New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME, Series D*, **83**(3), 95–108. 34, 47

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series. *J. Am. Statist. Assoc.*, **82**(400), 1023–1063. 56

—— (1996). Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, **1**, 1–25. 72

Kong, A., Liu, J. S., and Wong, W. (1994). Sequential imputation and Bayesian missing data problems. *J. Am. Statist. Assoc.*, **89**(278-288), 590–599. 15, 17, 19, 21, 60, 94, 96, 132, 139

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Statist.*, **8**, 115–131. 18, 21, 22, 159, 190

Künsch, H. R. (2005). Recursive Monte-Carlo filters : algorithms and theoretical analysis. *Ann. Statist.*, **33**(5), 1983–2021. 73, 77, 78, 92, 101, 103

Legland, F. and Oudjane, N. (2006). A sequential algorithm that keeps the particle system alive. *Tech. Rep.*, Rapport de recherche 5826, INRIA. 16, 20, 91, 119

Levine, R. A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *J. Comput. Graph. Statist.*, **10**(3), 422–439. 114

Levine, R. A. and Fan, J. (2004). An automated (Markov chain) Monte Carlo EM algorithm. *J. Stat. Comput. Simul.*, **74**(5), 349–359. 114

Liu, C. and Rubin, D. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, **5**(1), 19–39. 156

Liu, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. 15, 19, 92, 94, 97, 147

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputations. *J. Am. Statist. Assoc.*, **90**(420), 567–576. 15, 17, 19, 21, 41, 61, 120, 194

—— (1998). Sequential Monte-Carlo methods for dynamic systems. *J. Am. Statist. Assoc.*, **93**(443), 1032–1044. 15, 19, 70, 192

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.*, **6**, 113–119. 61

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons. 153

Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for convergence rates of Markov chains. *Annals of Applied Probability*, **4**, 981–1011. 35, 40

Montemerlo, M. (2003). *FastSLAM : A Factored Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association*. Ph.D. thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. 15, 19

Musso, C., Oudjane, N., and Le Gland, F. (2001). Improving regularized particle filters. *In* A. Doucet, N. De Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, Springer. 190

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM, Philadelphia. 73

Oh, M.-S. and Berger, J. O. (1992). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation*, **41**(3-4), 143–168. 97

—— (1993). Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.*, **88**(422), 450–456. 97, 149, 156

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in non-linear state space model. *Bernoulli*, **14**(1), 155–179. 16, 17, 20, 21, 65, 77, 88, 89, 90

Olsson, J., Moulines, E., and Douc, R. (2007). Improving the performance of the two-stage sampling particle filter : a statistical perspective. *In Proceedings of the IEEE/SP 14th Workshop on Statistical Signal Processing*. Madison, USA, pp. 284–288. 103

Pang, S., Li, J., and Godsill, S. (2009). Detection and Tracking of Coordinated Groups. *IEEE Trans. Aerospace and Electronic Systems*, **in revision**. 15, 19

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348. 156

Petrov, V. V. (1995). *Limit Theorems of Probability Theory*. Oxford University Press. 87

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation : Auxiliary particle filters. *J. Am. Statist. Assoc.*, **94**(446), 590–599. 9, 11, 15, 16, 18, 19, 20, 22, 41, 46, 52, 77, 78, 80, 93, 102, 106, 119, 120, 124, 148, 149

Polson, N. G., Carlin, B. P., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Assoc.*, **87**(418), 493–500. 56

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, 2nd edn.. 69

Quandt, R. and Ramsey, J. (1972). A new approach to estimating switching regressions. *J. Am. Statist. Assoc.*, **67**, 306–310. 151

Ristic, B., Arulampalam, M., and Gordon, A. (2004). *Beyond Kalman Filters : Particle Filters for Target Tracking*. Artech House. 15, 19, 30, 55, 100, 147

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, 2nd edn.. 16, 20, 30

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fraction of missing information is modest : the SIR algorithm (discussion of Tanner and Wong). *J. Am. Statist. Assoc.*, **82**, 543–546. 31, 32, 147, 192

—— (1988). Using the SIR algorithm to simulate posterior distribution. *In* J. M. Bernardo, M. DeGroot, D. Lindley, and A. Smith (Eds.), *Bayesian Statistics 3*, Clarendon Press. pp. 395–402. 31

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. John Wiley and Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics. 149

Rubinstein, R. Y. and Kroese, D. P. (2004). *The Cross-Entropy Method*. Springer. 15, 18, 19, 21, 95, 97, 114, 150, 190

—— (2008). *Simulation and the Monte Carlo method*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley and Sons], Hoboken, NJ, second edn.. 149

Septier, F., Carmi, A., and Godsill, S. (2009). Tracking of multiple contaminant clouds. *In Submitted to Proc. of the 12th International Conf. on Information Fusion*. Seattle, Washington. 15, 19

Shapiro, A. (1996). Simulation-based optimization : Convergence analysis and statistical inference. *Stochastic Models*, **12**, 425–454. 158

Shapiro, A. and Homem de Mello, T. (2001). On rate of convergence of monte carlo approximations of stochastic programs. *SIAM Journal on Optimization*, **11**, 70–86. 158

Shen, C., van den Hengel, A., Dick, A., and Brooks, M. J. (2004). Enhanced importance sampling : unscented auxiliary particle filtering for visual tracking. *In AI 2004 : Advances in artificial intelligence*, vol. 3339 of *Lecture Notes in Comput. Sci.*, Springer, Berlin. pp. 180–191. 104

Shephard, N. and Pitt, M. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**(3), 653–667. Erratum in volume 91, 249–250, 2004. 55, 92

Soto, A. (2005). Self adaptive particle filter. *In* L. P. Kaelbling and A. Saffiotti (Eds.), *Proceedings of the 19th International Joint Conferences on Artificial Intelligence (IJCAI)*. Edinburgh, Scotland, pp. 1398–1406. 92, 120

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62**(4), 605–655. With discussion and a reply by the authors. 94

Straka, O. and Simandl, M. (2006). Particle filter adaptation based on efficient sample size. *In Proceedings of the 14th IFAC Symposium on System Identification*. Newcastle, Australia, pp. 991–996. 92, 139

Tanizaki, H. (2003). Nonlinear and non-Gaussian state-space modeling with Monte-Carlo techniques : a survey and comparative study. *In* D. N. Shanbhag and C. R. Rao (Eds.), *Handbook of Statistics 21. Stochastic processes : Modelling and Simulation*, Elsevier. pp. 871–929. 41

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. MIT press, Cambridge, Massachusetts, USA. 15, 19

Van der Merwe, R., Doucet, A., De Freitas, N., and Wan, E. (2000). The unscented particle filter. *In* T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Adv. Neural Inf. Process. Syst.*, vol. 13, MIT Press. 16, 18, 20, 22, 54, 149

Van der Merwe, R. and Wan, E. (2003). Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models. *In Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Hong Kong, pp. 701–704. 149

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. 99

Verweij, B., Ahmed, S., Kleywegt, A., Nemhauser, G., and Shapiro, A. (2003). The sample average approximation method applied to stochastic routing problems : A computational study. *Computational Optimization and Applications*, **24**(2), 289–333. 158

Wang, Q., Niemi, J., Tan, C.-M., You, L., and West, M. (2009). Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy. *Synthetic Biology*, **to appear**. 15, 19

Wei, G. C. G. and Tanner, M. A. (1991). A Monte-Carlo implementation of the EM algorithm and the poor man's Data Augmentation algorithms. *J. Am. Statist. Assoc.*, **85**, 699–704. 99

Whitley, D. (1994). A genetic algorithm tutorial. *Stat. Comput.*, **4**, 65–85. 70, 73

Williams, D. (1991). *Probability with Martingales*. Cambridge University Press. 38

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103. 153, 154

Zaritskii, V., Svetnik, V., and Shimelevich, L. (1975). Monte-Carlo techniques in problems of optimal data processing. *Autom. Remote Control*, **12**, 2015–2022. 41, 48

Zhang, J. L. and Liu, J. S. (2002). A new sequential importance sampling method and its application to the two-dimensional hydrophobic-hydrophilic model. *J. Chem. Physics*, **117**(7). 18, 21, 120, 121, 125