Monte Carlo Methods for Machine Learning : Practical and Theoretical Contributions for Importance Sampling and Sequential Methods

Yazid Janati

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAS008

Thèse de doctorat

TELECOM SudParis

IP PARIS

Monte Carlo Methods for Machine Learning : Practical and Theoretical Contributions for Importance Sampling and Sequential Methods

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°574 École Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Sorbonne université, campus UPMC, le 28/09/2023, par

YAZID JANATI EL IDRISSI

Composition du Jury :

Jean Marc Bardet Professeur, Université Panthéon-Sorbonne	Président
Anthony Lee Professor, University of Bristol	Rapporteur
Jean-Michel Marin Professeur, Université de Montpellier	Rapporteur
Marylou Gabrié Maîtresse de conférences, Ecole polytechnique	Examinatrice
Sylvain Le Corff Professeur, Sorbonne Université	Directeur de thèse
Yohan Petetin Maître de conférences, Télécom SudParis	Co-directeur de thèse
Arnaud Guyader Professeur, Sorbonne Université	Invité

A mon très cher papa, Driss

Remerciements

Je tiens tout d'abord à adresser mes plus sincères remerciements à mes deux directeurs de thèse, Sylvain Le Corff et Yohan Petetin. Grâce à vous, ces années de thèse ont toujours été agréables et je n'aurais pas pu espérer mieux. Votre contribution intellectuelle et humaine a grandement enrichi ma thèse, et je suis reconnaissant d'avoir eu la chance de travailler avec vous. Vous avez su me proposer les bons problèmes dès la première année, ce qui m'a rapidement permis de comprendre ce qui me plaît réellement. Vous êtes tous les deux de très bonnes rencontres, tant sur le plan humain que scientifique. Sylvain, tu m'as constamment poussé à donner le meilleur de moi-même grâce à tes conseils précieux empreints de bienveillance et à ton souci du détail exemplaire. Tu n'as jamais hésité à me mettre en avant, et pour cca, je t'en suis profondément reconnaissant. Grâce à toi, j'ai également eu la chance de faire de belles rencontres scientifiques qui ont enrichi mon parcours académique. Yohan, sans toi, ce projet de thèse n'aurait probablement jamais vu le jour. C'est grâce à tes cours en première année d'école que ma passion pour le Monte Carlo est née. Nos discussions de longues heures sur des détails fins ont toujours été de vrais moments de plaisir et ont eu une grande résonance sur mes travaux. J'apprécie beaucoup l'originalité de tes idées, et je te remercie pour l'importance que tu accordes aux miennes.

Je souhaite remercier Anthony Lee et Jean-Michel Marin d'avoir accepté de rapporter ma thèse. Thank you, Anthony, for accepting to report my thesis. Je remercie également Arnaud, Jean-Marc et Marylou d'avoir accepté de faire partie du jury.

Un remerciement particulier à Eric Moulines et Alain Durmus, auprès desquels une partie non négligeable de cette thèse a été réalisée. Eric, ton immense générosité et ta curiosité sans limites sont une véritable source d'inspiration. Je suis toujours impressionné par l'ampleur de tes connaissances (notamment sur le Maroc, que tu connais bien mieux que moi, à tel point que j'en ai parfois honte) et par ta capacité à poser les questions les plus pertinentes. Alain, je suis content d'être venu te poser la question sur le Langevin à Ben-Guerir, cela a ouvert la porte à une collaboration fructueuse au cours de laquelle j'ai beaucoup appris à tes côtés. Tes tours de force mathématiques m'impressionneront toujours, et je n'oublierai pas les longues séances au tableau ensemble, c'était un vrai plaisir. Je te remercie chaleureusement pour toutes les opportunités que tu m'as offertes, notamment celle de voyager à Cambridge avec Gabriel, un périple dont je ne conserverai que des souvenirs précieux. J'écris ces mots sans trop de nostalgie car notre réelle collaboration commence la semaine prochaine.

Je tiens également à exprimer ma gratitude envers Randal Douc, qui m'a accueilli en stage il y a quelque temps déjà. Sa contribution à ma passion pour la simulation ne fait aucun doute. Je garde un excellent souvenir de l'après-midi que nous avons passée ensemble à Madrid.

J'en viens maintenant à Gabriel, ah sacré Gabriel, collègue devenu un très bon ami en si peu de temps. Copilote de ma troisième année de thèse, je te suis infiniment reconnaissant pour tout ce que tu m'as appris. J'espère que tu me pardonnes mon entêtement parfois excessif, cf

le fameux débat de 11h à 15h à Londres qui nous a fait rater le déjeuner. Merci pour tous les cafés que tu m'as préparés les dimanches avant les deadlines, dans la tasse NeurIPS. Chaque histoire a son emblématique duo : Batman et Robin, Tom & Jerry, Mario et Luigi. Nous aurons été Astérix et Obélix de Jussieu (je te laisse deviner qui est qui). Obrigado por tudo, e sim, concordo que devemos nos livrar do frown.

Je souhaite également remercier mes co-auteurs Achille, Arnaud, Charles, Christian, Franccois, Jimmy et Julien. Achille, j'ai eu beaucoup de chance de travailler à tes côtés pendant la première année de thèse. Cela a été important pour moi, et je t'en remercie. Merci de m'avoir sauvé le jour où j'ai supprimé la soumission ICML ! Ta gentillesse est rare. Jimmy, I enjoyed spending time with you in Cambridge. I hope that we will have the opportunity to collaborate again in the future. Julien, c'est un vrai plaisir de travailler avec toi. Je te remercie pour ta bienveillance.

Je remercie aussi Elouan, mon coauteur sur Messenger. C'est bien dommage que l'on n'ait jamais transformé toutes les idées qu'on s'envoie par message en papier. J'espère que cca finira par se concrétiser un jour. Merci d'être toujours à l'écoute.

Ces trois années n'auraient pas été ce qu'elles sont sans mes chers co-doctorants. Je remercie tout d'abord l'équipe Lagrange, Achille, Lisa, Louis, Maxence, Mehdi, Pablo, Pierre, Thomas, Tom, Valentin et Vincent pour tous les bons moments passés ensemble. Merci à Sasila qui a toujours été là, même au nord du Maroc! On a passé de très bons moments ensemble. Merci à Rémi pour les quelques bons moments passés ensemble, heureux d'avoir ta connaissance! Merci aussi à Fabiola et Mahdi, mes très bon amis d'école. Je remercie l'équipe Jussieu de m'avoir chaleureusement accueilli cette année. J'ai vraiment apprécié faire partie de votre quotidien. Ariane et Iqraa, je suis très content d'avoir fait votre connaissance. Je vous remercie pour les innombrables pauses déjeuners (très drôles à chaque fois) qui se sont (un peu trop souvent) étendues en de très longues pauses. Antonio, j'ai beaucoup apprécié partager le bureau avec toi. J'admire ton dynamisme, et je suis ravi de pouvoir continuer l'aventure avec toi. Mathis, mon binôme de lissage, j'apprécie toutes nos discussions. Je remercie aussi Pablo, Camilla, Miguel, Grâce, Lucas, Ludovic, Romain, Pierre et Alexis. Je souhaite aussi remercier Anna pour sa bienveillance constante ainsi qu'Antoine, Arnaud, Badr, Claire, Maxime et Stéphane sans qui l'ambiance au LPSM ne serait probablement pas aussi agréable. Je veux aussi remercier Pierre Gloaguen, avec qui j'ai passé de bons moments!

Je veux aussi remercier Julie Bonnet et Laura Landes qui ont toujours su m'aider à résoudre mes problèmes administratifs.

Je tiens à exprimer ma sincère gratitude envers mes amis, les *Lascars*, Younes, Mouhdi, Salim et Othmane. Vous avez été les meilleurs compagnons possibles pendant ces trois années de thèse. Vous êtes mes repères et votre amitié précieuse a facconné la personne que je suis devenu. Hakim, au-delà de notre amitié exceptionnelle que je n'ai pas besoin de commenter, c'est à tes côtés que j'ai appris à aimer les mathématiques. Les longues heures passées à explorer le fameux PDF pendant l'été restent des souvenirs chers à mon cœur. Je te suis extrêmement reconnaissant. Ismaïl, tu es un pillier dans ma vie. Les moments que je passe avec toi, à discuter de la vie, parler de musique ou travailler à Jussieu ou à la BSG, sont toujours d'excellents souvenirs. A toi aussi, je dois énormèment. Younes, je suis reconnaissant de t'avoir à mes côtés et de pouvoir compter sur toi. Tu es un formidable ami, merci pour tout et en particulier les années à Evry. Adam, Sara, Nouhaila, Mamoun, Sasila, Najib, Taha, Inchaouh, Jad ainsi qu'à tous les autres qui sont à mes côtés depuis le début, je vous adresse un immense merci. Votre amitié a été un précieux équilibre dans ma vie.

A ma famille, ma très cher Maman, Nor, Ghita, May et Rokia, je vous doit tout. Merci d'avoir

fait de moi la personne que je suis aujourd'hui et de continuer à me combler. Nor, tout a réellement commencé par toi, je t'en remercie infiniment.

Enfin, ma précieuse Soukaina, sans toi je ne peux rien faire, pas même cette thèse. Tout cela n'a de saveur que grâce à toi. Merci pour tout ce que tu fais pour moi depuis de si longues années.

Contents

Notation

The following notations are used throughout the introduction of this thesis.

- i.i.d.: independent and identically distributed.
- For $r, s \in \mathbb{N}$ such that $r < s$, we write $[r : s] = \{r, r+1, \dots, s\}$. This notation extends to collections of variables, e.g. we will sometimes write $X^{r:s}$ for $(X^r, X^{r+1}, \dots, X^s)$.
- We denote by $\mathsf{M}_1(\mathcal{X})$ the set of probability measures on $(\mathsf{X}, \mathcal{X})$. We denote by $\mathsf{F}(\mathcal{X})$ the set of bounded measurable functions. For the particular case of $\mathbb{R}^d$ we write $\mathsf{M}_1(\mathbb{R}^d)$ and $\mathsf{F}(\mathbb{R}^d)$.
- For $\mu \in \mathsf{M}_1(\mathcal{X})$ and $h \in \mathsf{F}(\mathcal{X})$ we write $\mu(h) = \int h(x)\mu(\mathrm{d}x)$.
- For $\mu, \nu \in \mathsf{M}_1(\mathcal{X})^2$, we write $(\mu \otimes \nu)$ for the tensor product defined for all $(A, B) \in \mathcal{X}^2$ by $(\mu \otimes \nu)(A \times B) = \int \mathbb{1}_A(x)\mathbb{1}_B(y)\mu(\mathrm{d}x)\nu(\mathrm{d}y)$.
- For a Markov kernel K from $(\mathsf{X}, \mathcal{X})$ to another measurable space $(\mathsf{Y}, \mathcal{Y})$, we write for all $h \in \mathsf{F}(\mathcal{Y})$, $K(h) : \mathsf{X} \ni x \mapsto \int h(y)K(x, \mathrm{d}y)$.
- If $\mu \in \mathsf{M}_1(\mathcal{X})$, then μK is a probability measure in $\mathsf{M}_1(\mathcal{Y})$ defined for all $A \in \mathcal{Y}$ by $\mu K(A) = \int \mathbb{1}_A(y)\mu(\mathrm{d}x)K(x, \mathrm{d}y)$.
- $x \mapsto \mathcal{N}(x; \mu, \sigma^2)$ is the Gaussian probability density with mean μ and variance σ^2.

Chapter 1

Introduction

"Tout le malheur des hommes vient de l'espérance." - Albert Camus.

This introduction surveys the main results of this thesis. We begin with a general discussion of the challenges addressed by Monte Carlo simulation. Subsequently, we give a brief introduction to the three subjects studied during the Ph.D.: importance sampling, sequential Monte Carlo, and generative modeling. The outline of the remaining chapters and our contributions are summarized at the end of the chapter.

1.1 Introduction générale

Dans de nombreuses applications, l'intégration par rapport à une mesure de probabilité et l'échantillonnage sont primordiaux; ils permettent le calcul de probabilités, l'estimation de paramètres et d'états inconnus, ainsi que la comparaison de différents modèles. Dans la plupart des scénarios réels, les mesures de probabilité d'intérêt n'ont pas de formes analytiques simples et vivent dans des espaces de grandes dimensions, ce qui rend les méthodes traditionnelles d'échantillonnage et d'intégration inapplicables. Les méthodes de Monte Carlo se sont imposées comme des outils puissants pour résoudre ces problèmes complexes.

Pour illustrer la pertinence du problème d'échantillonnage, considérons par exemple le cas de l'inférence bayésienne. Soient Y et X deux variables aléatoires dépendantes. Nous supposons que Y est une observation incomplète de X. L'objectif est d'échantillonner les reconstructions les plus plausibles de X en exploitant les connaissances préalables encodées dans sa loi, appelée loi a priori, et l'observation Y. Ce cadre, en apparence simple, est utile pour de nombreuses applications. En tomodensitométrie, des représentations précises de la structure interne du corps d'un patient sont reconstruites à partir de mesures limitées de rayons X. Pour le suivi d'objets, des observations incomplètes et potentiellement bruitées, ainsi que des connaissances sur la dynamique de l'objet, sont utilisées pour estimer sa position et sa trajectoire dans des séquences vidéo.

Si $X \in \mathbb{R}^{d_x}$ et $Y \in \mathbb{R}^{d_y}$, la mesure de probabilité jointe de (X, Y) s'écrit :

$$q_{X,Y}(\mathrm{d}x, \mathrm{d}y) = q_X(\mathrm{d}x) q_{Y|X}(\mathrm{d}y|x),$$

où q_X est la loi a priori de X sur $(\mathbb{R}^{d_x}, \mathcal{B}(\mathbb{R}^{d_x}))$ et $q_{Y|X}$ est un noyau de transition sur $\mathbb{R}^{d_x} \times \mathcal{B}(\mathbb{R}^{d_y})$. Cette loi jointe décrit le processus génératif de la paire (X, Y). Trouver les reconstructions les plus crédibles de X étant donnée l'observation Y revient donc à inverser ce

processus génératif tout en maintenant la bonne loi jointe. Cela implique d'échantillonner la distribution conditionnelle suivante, appelée *loi a posteriori*, définie pour $y \in \mathbb{R}^{d_y}$ par

$$q_{X|Y}(\mathrm{d}x|y) = \frac{q_{Y|X}(y|x)q_X(\mathrm{d}x)}{q_Y(y)}, \quad \text{où} \quad q_Y(y) = \int q_{Y|X}(y|x)q_X(\mathrm{d}x). \tag{1.1.1}$$

La loi a posteriori (1.1.1) pondère la distribution q_X de X avec la vraisemblance conditionnelle, ce qui rend l'échantillonnage potentiellement difficile, même lorsque q_X est relativement facile à échantillonner. Au fil des années, différentes techniques et méthodologies ont été développées pour relever ce défi, notamment les méthodes de Monte Carlo par chaînes de Markov (MCMC) telles que les algorithmes de Metropolis-Hastings Metropolis et al. (1953) et l'échantillonneur de Gibbs Gelfand and Smith (1990) ainsi que l'échantillonnage préferentiel Kahn (1949); Goertzel (1949) et les méthodes de Monte Carlo séquentielles Handschin and Mayne (1969); Gordon et al. (1993b). Ces approches ont permis aux praticiens de générer des échantillons approximatifs à partir de distributions a posteriori d'intérêt et sont au cœur de nombreuses avancées scientifiques.

Dans l'exemple de reconstruction que nous venons de décrire, X est la variable aléatoire d'intérêt. Pour les modèles génératifs, les rôles sont inversés et X est une variable latente dont le but est de complexifier le modèle proposé pour la loi de Y. Dans ce contexte, q_X et $q_{Y|X}$ sont souvent choisies faciles à échantillonner, de sorte à ce que l'échantillonnage de q_Y soit simple. Le véritable défi réside plutôt dans l'estimation de la densité marginale $q_Y(y)$ pour tout $y \in \mathbb{R}^{d_y}$, ce qui est un problème d'intégration. Nous mettons en évidence ce défi en considérant l'exemple suivant.

Example 1.1.1 (Modèles à variables latentes profonds (DLVM))**.** *Dans Kingma and Welling (2013), la paramétrisation suivante de la loi marginale de Y est considérée. Soit Θ un ensemble de paramètres et $\mu : \Theta \times \mathbb{R}^{d_x} \ni (\theta, x) \mapsto \mu_\theta(x)$, $\sigma^2 : \Theta \times \mathbb{R}^{d_x} \ni (\theta, x) \mapsto \sigma^2_\theta(x)$. Supposons que $X \sim \mathcal{N}(\mathbf{0}_{d_x}, \mathbf{I}_{d_x})$ et que conditionnellement à X, Y a pour loi $\mathcal{N}(\mu_\theta(X), \mathbf{I}_{d_y} \cdot \sigma^2_\theta(X))$. $\mu_\theta(x)$ et $\sigma^2_\theta(x)$ sont généralement les sorties d'un réseau de neurones et x sert d'entrée. La densité marginale résultante q^θ_Y est alors un mélange infini et est donc très expressive. Cela signifie que si les réseaux neuronaux μ_θ, σ^2_θ ont suffisamment de profondeur, q^θ_Y peut modéliser n'importe quelle densité de probabilité positive presque partout.*

Étant donné N observations i.i.d. $(Y^1, \dots, Y^N)$ échantillonnées à partir d'une loi inconnue π_Y, le DLVM de l'exemple précédent peut être appris pour approcher π_Y, c'est-à-dire trouver un paramètre θ_* de telle sorte que q^θ_Y soit une bonne approximation de π_Y. Avec cette loi qui se substitue à π_Y, nous pouvons ensuite générer de nouveaux échantillons qui sont approximativement distribués sous π_Y. Cependant, l'apprentissage de cette loi présente un défi important; nous ne pouvons pas apprendre le paramètre θ par maximum de vraisemblance car q^θ_Y est une intégrale (1.4.6) et n'a pas de forme analytique directe si μ_θ et σ^2_θ sont des réseaux de neurones. On pourrait alors penser à une estimation de type Monte Carlo de q^θ_Y car q_X est facile à échantillonner. Cependant, cela pose problème en grande dimension car $x \mapsto q^\theta_{Y|X}(y|x)$ est susceptible de prendre des valeurs élevées uniquement dans un petit sous-ensemble A de $\mathbb{R}^{d_x}$ qui a de plus une probabilité très faible selon q_X. Il est alors probable qu'aucun échantillon de X ne tombe dans cet ensemble. Ainsi, le défi pratique réside dans la taille d'échantillon substantielle nécessaire pour obtenir une approximation fiable. Pour les problèmes de grande dimension, cette exigence devient irréalisable, rendant l'approche Monte Carlo standard impraticable. Il est raisonnable de penser que si nous sommes capables d'échantillonner à partir d'une mesure de probabilité proche de la loi a posteriori, nous pouvons peut-être estimer la probabilité marginale $q_Y(y)$ avec une grande précision en n'utilisant une modeste taille d'échantillon, et c'est en effet le cas et le principal objectif de l'*échantillonnage préferentiel* (Kahn, 1949; Goertzel, 1949) qui

est présenté dans la Section 1.3.

En résumé, le problème de l'intégration repose sur celui de l'échantillonnage, qui à son tour dépend de la recherche d'une approximation appropriée de la mesure de probabilité cible. Il peut s'agir soit d'une chaîne de Markov qui converge vers la distribution cible souhaitée (MCMC), soit d'une loi de proposition soigneusement construite (échantillonnage par importance et méthodes de Monte Carlo séquentielles) dont un ensemble d'échantillons est pondéré pour former une approximation empirique de la cible.

Cette thèse vise à contribuer à la problématique de la construction de lois de proposition et d'estimateurs efficaces pour l'échantillonnage par importance et les méthodes de Monte Carlo séquentielles. Dans les chapitres 2 et 3, nous étudions ce problème dans le contexte de l'échantillonnage par importance en l'abordant sous deux angles différents. Dans le chapitre 2, nous cherchons une loi de proposition peu coûteuse et basée sur des étapes d'optimisation qui peut être utilisée pour l'apprentissage de modèles génératifs ou pour l'estimation en temps réel dans le cas des méthodes de Monte Carlo séquentielles. Dans le chapitre 3, nous proposons un nouveau schéma pour l'apprentissage de lois de proposition. Les chapitres 4, 5 et 6 sont ensuite dédiés aux méthodes de Monte Carlo séquentielles. Dans le chapitre 4, notre objectif est de fournir de meilleurs estimateurs de la variance asymptotique qui apparaît dans le théorème central limite pour le filtre à particules. En tant que deuxième contribution, nous obtenons le premier estimateur pour la variance asymptotique du lisseur à particules. Dans le chapitre 5, nous étudions le problème de l'estimation du gradient de la logvraisemblance de modèles de Markov cachés et obtenons une procédure d'estimation avec un biais réduit dont nous prouvons la convergence. Enfin, dans le chapitre 6, nous proposons une approche basée sur les méthodes de Monte Carlo séquentielles et les modèles de diffusion pour résoudre des problèmes inverses linéaires bayésiens.

1.2 General introduction

In many applications, integration with respect to a probability measure and sampling are paramount; they enable the computation of probabilities, estimation of unknown parameters and states as well as the comparison of different models. In most real-world scenarios, the underlying probability distributions of interest lack simple analytical forms and are high dimensional, making traditional sampling and integration methods infeasible. Monte Carlo methods have emerged as powerful and versatile tools for tackling these complex problems.

To illustrate the relevance of the sampling problem, consider for instance the case of Bayesian inference. Let Y and X be two dependent random variables. We assume that Y is an incomplete observation of X, e.g. a low dimensional representation corrupted with noise. The goal is to sample the most plausible reconstructions of X by leveraging prior knowledge encoded in its law, known as the *prior*, and the observation Y. This seemingly simple program encompasses important applications. In Computed Tomography, accurate representations of the internal structure of a patient's body are reconstructed from limited X-ray measurements. For object tracking, incomplete and potentially noisy observations, along with knowledge about the object's dynamics, are employed to estimate its position and trajectory in video sequences.

If $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$, the joint probability measure of (X, Y) is written

$$q_{X,Y}(\mathrm{d}x, \mathrm{d}y) = q_X(\mathrm{d}x) q_{Y|X}(\mathrm{d}y|x),$$

where q_X is the prior distribution of X on $(\mathbb{R}^{d_x}, \mathcal{B}(\mathbb{R}^{d_x}))$ and $q_{Y|X}$ is a transition kernel on $\mathbb{R}^{d_x} \times \mathcal{B}(\mathbb{R}^{d_y})$. This joint distribution describes the generative process of the pair (X, Y). The

observed Y is a sample from $q_{Y|X}(\cdot|X)$ for some given X sampled from q_X. Finding the most credible reconstructions of X given Y then boils down to reversing this generative process while maintaining the correct joint distribution. This involves sampling the following conditional distribution, known as the *posterior*, defined for $y \in \mathbb{R}^{d_y}$ by

$$q_{X|Y}(\mathrm{d}x|y) = q_{Y|X}(y|x)q_X(\mathrm{d}x)/q_Y(y)\,, \quad \text{where} \quad q_Y(y) = \int q_{Y|X}(y|x)q_X(\mathrm{d}x)\,. \tag{1.2.1}$$

The posterior (1.2.1) weights the distribution q_X of X with the conditional likelihood making it likely to be challenging to sample, even when q_X is relatively easy to sample from. Over the years, various techniques and methodologies have been developed to address this challenge, including Markov chain Monte Carlo (MCMC) methods such as the Metropolis-Hastings algorithms Metropolis et al. (1953) and Gibbs sampling Gelfand and Smith (1990) as well as importance sampling Kahn (1949); Goertzel (1949) ans Sequential Monte Carlo Handschin and Mayne (1969); Gordon et al. (1993b). These approaches have enabled practitioners to generate approximate samples from posterior distributions of interest and are at the heart of many scientific advances.

In the reconstruction example we have just described X is the random variable of interest. In generative modeling, the roles are inverted and X is a latent variable whose purpose is to design a more complex law for Y. In this context, q_X and $q_{Y|X}$ are often chosen to be simple to sample from and so overall, sampling from q_Y is straightforward. The real challenge instead lies in the estimation of the marginal density $q_Y(y)$ for all $y \in \mathbb{R}^{d_y}$, which is an integration problem. We highlight this challenge by considering the following example.

Example 1.2.1 (Deep latent variable models (DLVM))**.** *In Kingma and Welling (2013) the following parameterization for the marginal law of Y is considered. Let Θ be a set of parameters and $\mu : \Theta \times \mathbb{R}^{d_x} \ni (\theta, x) \mapsto \mu_\theta(x)$, $\sigma^2 : \Theta \times \mathbb{R}^{d_x} \ni (\theta, x) \mapsto \sigma^2_\theta(x)$. Let $X \sim \mathcal{N}(\mathbf{0}_{d_x}, \mathbf{I}_{d_x})$ and the conditional distribution of Y given X be $\mathcal{N}(\mu_\theta(X), \mathbf{I}_{d_y} \cdot \sigma^2_\theta(X))$. $\mu_\theta(x)$ and $\sigma^2_\theta(x)$ are typically the outputs of a neural network and x serves as the input. The resulting marginal density q^θ_Y is an infinite mixture and is thus highly expressive meaning that if the neural networks μ_θ, σ^2_θ are given enough depth, q^θ_Y can model any probability density that is positive everywhere.*

Given N i.i.d. observations $(Y^1, \ldots, Y^N)$ sampled from an unknown data distribution π_Y, the DLVM of the previous example can be "learned" to approximate π_Y, i.e. find a parameter θ_* so that $q^{\theta_*}_Y$ is a good approximation of π_Y. With this surrogate we can then generate new approximate samples from π_Y. However, learning this surrogate presents a significant challenge; we cannot learn the parameter θ through maximum likelihood since q^θ_Y is an integral (1.4.6) and does not have a straightforward analytical form if μ_θ and σ^2_θ are neural networks. We could think of a Monte Carlo estimate of q^θ_Y as q_X is easy to sample from. This is flawed however in high dimensions as $x \mapsto q^\theta_{Y|X}(y|x)$ is likely to take large values only in a small subset A of $\mathbb{R}^{d_x}$ that furthermore has a very small probability under q_X. It is likely then that no samples from X will fall in this subset. As such, the practical challenge arises from the substantial sample size needed to achieve a reliable approximation. Particularly for high-dimensional problems, this requirement becomes computationally infeasible, rendering the standard Monte Carlo approach impractical. As we have mentionned before, the posterior reverses the generative process and localizes the sampling in the regions of $\mathbb{R}^{d_x}$ that have likely generated Y. It is then sensible to assume that if we are able to sample from a probability measure that is close to the posterior then perhaps we can estimate the marginal probability $q_Y(y)$ with high accuracy using only a small sample size, and this is indeed the case and the main point of *importance sampling* (Kahn, 1949; Goertzel, 1949) which is presented in the next section.

In summary, the problem of integration relies on the problem of sampling, which, in turn,

hinges on finding an appropriate surrogate for the target probability measure. It can either be a Markov chain that converges to the desired target distribution (MCMC) or, a carefully crafted proposal (importance sampling and sequential Monte Carlo) of which a pool of samples is weighted to form a consistent empirical approximation.

This thesis aims at contributing to the problem of constructing efficient proposals and estimators for importance sampling and sequential Monte Carlo methods. In Chapters 2 and 3 we study this problem in the context of importance sampling and approach it from two different angles. In Chapter 2, we seek a lightweight optimization based proposal that can be used for learning generative models or for real-time estimation in Sequential Monte Carlo and in Chapter 3 we derive a new scheme for learning sharp importance proposals. Chapters 4, 5 and 6 are dedicated to Sequential Monte Carlo. In Chapter 4, we aim at providing better estimators of the asymptotic variance that appears in the Central Limit Theorem for the particle filter. As a second contribution we derive the first estimator for the asymptotic variance of the particle smoother. In Chapter 5, we study the problem of particle smoothing estimation in the context of parameter learning and derive a procedure with provably reduced bias. Finally, in Chapter 6 we devise Sequential Monte Carlo-based approach for solving linear Bayesian inverse problems with generative model priors.

1.3 Technical background

1.3.1 Importance sampling

Principle

Importance sampling (IS) is a technique with roots dating back to Kahn (1949); Goertzel (1949). We denote by π a target probability measure defined on a measurable space $(\mathsf{X}, \mathcal{X})$. Our goal is to compute integrals of the form

$$\pi(f) = \int f(x)\pi(\mathrm{d}x)$$

where f is a real valued measurable function. IS comes into play when either: (1) it is not possible to sample from π, or (2) it is possible to sample from π but the vanilla Monte Carlo estimator performs poorly for a fixed computational budget. The latter happens for example when f takes non-zero values only in the tails of π. IS consists in introducing an **easy to sample** importance distribution μ with support including the support of π and applying a change of measure to estimate the expectation of interest.

Let μ be a measure defined on $(\mathsf{X}, \mathcal{X})$ and dominating π, i.e. for any $A \in \mathcal{X}$, $\pi(A) > 0$ implies that $\mu(A) > 0$. Denote by $\mathrm{d}\pi/\mathrm{d}\mu$ the Radon-Nikodym derivative which satisfies the change of measure identity $\pi(\mathrm{d}x) = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\mu(\mathrm{d}x)$. For any π-integrable function f we have that $\pi(f) = \int f(x)\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\mu(\mathrm{d}x)$, and hence, by drawing M i.i.d. samples from μ we obtain the following estimator $\pi_\mu^M(f)$ of $\pi(f)$,

$$\pi_\mu^M(f) = M^{-1}\sum_{i=1}^{M} f(X^i)\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X^i)\,, \quad \text{where} \quad (X^1, \ldots, X^M) \overset{\text{iid}}{\sim} \mu\,. \tag{1.3.1}$$

The importance distribution μ is a free parameter that may be optimized to improve the efficiency of the estimator (1.3.1), e.g. accurate estimation with M as small as possible. By inspecting the variance of $\pi_\mu^M(f)$ we see that it is equal to $M^{-1}(\mu(f^2\frac{\mathrm{d}\pi}{\mathrm{d}\mu}^2) - \pi(f)^2)$ and hence if $f \geq 0$ for example, setting $\mu = f \cdot \pi/\pi(f)$ achieves a zero variance estimator. In the more

general case, setting $\mu \propto |f| \cdot \pi$ minimizes the variance. This means in practice that μ should assign high probability in the regions of the space where the measure $|f| \cdot \pi$ has large mass. This ideal importance distribution is however intractable and one should aim at setting μ as close as possible to this optimal choice. In fact, this should be done by paying a particular attention to the tails of the importance distribution, as tails lighter than those of the target distribution may result in an estimator with infinite variance. If one wants to estimate integrals with different test functions f using the same samples, then it is best to choose μ as close as possible to π.

When π is known only up to a normalizing constant, e.g. π is a posterior distribution, $\mathrm{d}\pi/\mathrm{d}\mu$ is also known up to a scaling factor. We can still estimate $\pi(f)$ by noticing that $\pi(f)/\pi(\mathbf{1})$ is free of any scaling factor and equal to $\pi(f)$. Estimating both the numerator and denominator with IS using the *same* samples $X^{1:M} \overset{\text{iid}}{\sim} \mu$ then yields the estimator

$$\overline{\pi}_\mu^M(f) = \sum_{i=1}^M \frac{\widetilde{\omega}^i}{\sum_{j=1}^M \widetilde{\omega}^j} f(X^i)\,, \quad \text{where} \quad \widetilde{\omega}^i = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X^i)\,, \tag{1.3.2}$$

known as the *self-normalized importance sampling* (SNIS) estimator. In contrast with (1.3.1), it is *biased*, i.e. $\mathbb{E}[\overline{\pi}_\mu^M(f)] \neq \pi(f)$. The non-asymptotic properties of the SNIS estimator are also well understood, see Cappe et al. (2005); Agapiou et al. (2017); Chatterjee and Diaconis (2018). Let us first formally define the α-Rényi and KL divergence which will allow us to quantify the discrepancy between two probability measures. The most fundamental properties of such divergences are covered in Van Erven and Harremos (2014).

Definition 1.3.1. *The α-Rényi divergence between $\pi \in \mathsf{M}_1$ and $\mu \in \mathsf{M}_1$ with $\mu \gg \pi$ is defined for any $\alpha \in (0,1) \cup (1,\infty)$ by*

$$\mathcal{R}_\alpha(\pi \parallel \mu) := \frac{1}{\alpha - 1} \log \int \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)^\alpha \mu(\mathrm{d}x)\,. \tag{1.3.3}$$

Its extension to the order $\alpha = 1$ is the backward *Kullback-Leibler divergence*

$$\mathsf{KL}(\pi \parallel \mu) := \int \log \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x) \pi(\mathrm{d}x)\,. \tag{1.3.4}$$

For any $\alpha \in (0,\infty)$, $\mathcal{R}_\alpha(\pi \parallel \mu) = 0$ if and only if $\pi = \mu$.

Theorem 1.3.2 (Agapiou et al. (2017))**.** *Assume that $\mu \gg \pi$. The bias and MSE of the SNIS estimator over bounded test functions is given by*

$$\sup_{|f| \leq 1} \left| \mathbb{E}\left[\overline{\pi}_\mu^M(f) - \pi(f)\right] \right| \leq \frac{12 \exp \mathcal{R}_2(\pi \parallel \mu)}{M}\,,$$
$$\sup_{|f| \leq 1} \mathbb{E}\left[\left(\overline{\pi}_\mu^M(f) - \pi(f)\right)^2\right] \leq \frac{4 \exp \mathcal{R}_2(\pi \parallel \mu)}{M}\,.$$

Impact of the dimension

As for the standard IS estimator, the sufficient sample size M that guarantees a small variance and bias for a fixed importance distribution μ is proportional to the exponential of the 2-Rényi divergence, or the variance of the importance weight. The 2-Rényi divergence captures the intuitive idea that the importance proposal should have the same regions of high probability as π. This is observed by recognizing that $\mathcal{R}_2(\pi \parallel \mu) = \log \int \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \mathrm{d}\pi$ and hence it takes large values if $\frac{\mathrm{d}\pi}{\mathrm{d}\mu}$ is large in regions of high probability under π. The same intuition holds for the KL divergence.

Theorem 1.3.2 also highlights how the dimension of the ambient space X impacts the IS and SNIS estimators. Indeed, assume that the target distribution is instead $\pi^{\otimes d}(\mathrm{d}x_{1:d}) = \prod_{i=1}^{d} \pi(\mathrm{d}x_i)$ and the importance distribution is $\mu^{\otimes d}$; we replicate d times the original target π and importance distribution μ. Then according to the 2-Rényi divergence criterion in Theorem 1.3.2, the sample size M is now $\exp(d \cdot \mathcal{R}_2(\pi \parallel \mu))$ since $\mathcal{R}_2(\pi^{\otimes d} \parallel \mu^{\otimes d}) = d \cdot \mathcal{R}_2(\pi \parallel \mu)$; it scales exponentially with the dimension. This a necessary and sufficient sample size for the IS estimator (1.3.1). If one is however interested in being close to $\pi(f)$ in high probability then the sample size prescribed by the variance is often much larger than what is required. In Chatterjee and Diaconis (2018) it is shown, under the assumption that $\log \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(Y)$, where $Y \sim \pi$, is concentrated around its mean $\mathsf{KL}(\pi \parallel \mu)$, that the necessary and sufficient sample size is in fact of the order $\exp(\mathsf{KL}(\pi \parallel \mu))$. This confirms that importance sampling intriniscally suffers from exponential dependence on the dimension since $\mathsf{KL}(\pi^{\otimes d} \parallel \mu^{\otimes d}) = d \cdot \mathsf{KL}(\pi \parallel \mu)$. While this discussion only holds for probability measures that factorize over the dimensions, the curse of dimensionality is also observed empirically for probability measures that have *intrinsic* dimension smaller than that of the ambient space.

IS-based samplers

Interestingly, the SNIS estimator defines a consistent empirical approximation of π given by $\overline{\pi}_\mu^M(\mathrm{d}x) = \sum_{i=1}^M \omega^i \delta_{X^i}(\mathrm{d}x)$ where $\omega^i = \widetilde{\omega}^i / \sum_{j=1}^M \widetilde{\omega}^j$. This suggests that we can draw approximate samples from π as follows; sample $(\mathrm{J}_I^1, \ldots, \mathrm{J}_I^{\tilde{M}}) \overset{\text{iid}}{\sim} \text{Categorical}(\{\omega^i\}_{i=1}^M)$ and set $\tilde{X}^j = X^{I^j}$ for all $j \in [1 : \tilde{M}]$. This results in a *unweighted* particle approximation $\widetilde{\pi}_\mu^M$ of π given by $\widetilde{\pi}_\mu^M = \tilde{M}^{-1} \sum_{i=1}^{\tilde{M}} \delta_{\tilde{X}^i}(\mathrm{d}x)$ where the samples $\tilde{X}^i$ are dependent however. Nonetheless, they become i.i.d. when the sample size M of the underlying SNIS estimator goes to infinity, see (Cappe et al., 2005, Chapter 9).

We now describe the *iterated SIR* (iSIR) algorithm Andrieu et al. (2010) which defines, with finite M, a geometrically ergodic Markov chain, i.e. one that gets arbitrarily close to π in finite time. iSIR proceeds iteratively as follows; given a state $\tilde{X}_k$ at iteration k, (1) set $X_{k+1}^1 = \tilde{X}_k$ and sample $(X_{k+1}^2, \ldots, X_{k+1}^M) \overset{\text{iid}}{\sim} \mu$; (2) compute the unnormalized weights $\widetilde{\omega}_k^i = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X_{k+1}^i)$ and set $\omega_{k+1}^i = \widetilde{\omega}_{k+1}^i / \sum_{j=1}^M \widetilde{\omega}_{k+1}^j$; (3) sample $I_{k+1} \sim \text{Categorical}(\{\omega_{k+1}^i\}_{i=1}^M)$ and set $\tilde{X}_{k+1} = X_{k+1}^{I_{k+1}}$. Basically at each step a promising particle is selected according to its importance weight and propagated to the next iteration. This procedure defines a π-invariant Markov chain $(\tilde{X}_k)_{k \in \mathbb{N}}$ with transition kernel defined by

$$P_M(x, \mathrm{d}y) = \int \left\{ \sum_{i=1}^M \frac{\mathrm{d}\pi/\mathrm{d}\mu(x_{k+1}^i)}{\sum_{j=1}^M \mathrm{d}\pi/\mathrm{d}\mu(x_{k+1}^j)} \delta_{x_{k+1}^i}(\mathrm{d}y) \right\} \delta_x(\mathrm{d}x_{k+1}^1) \mu^{\otimes M-1}(\mathrm{d}x_{k+1}^{2:M}) \,.$$

Theorem 1.3.3 (Andrieu et al. (2018a))**.** *Assume that* $w_\infty = \sup_{x \in \mathsf{X}} \mathrm{d}\pi/\mathrm{d}\mu(x) < \infty$*. Then for any initial distribution* $\xi \in \mathsf{M}_1(\mathcal{X})$ *and all* $M > 1$*,*

$$\|\pi - \xi P_M^k\|_{\mathrm{TV}} \leq \rho_M^k \,,$$

where $\rho_M = 1 - (M-1)/(2w_\infty + M - 2)$*.*

The Markov chain resulting from the iSIR procedure thus converges to π geometrically fast. The mixing rate ρ decreases with M but depends on w_∞ which is exponential in the dimension, making the use of iSIR prohibitive in large dimensions, similarly to SNIS. Indeed, in high dimensions the variance of the importance ratio increases exponentially and iSIR gets stuck in points with large importance ratio, which are difficult to escape. It has recently been shown in Samsonov et al. (2022) that its performance can be improved by combining it with a *local*

MCMC kernel, i.e. by considering instead the kernel $K_M(x, \mathrm{d}y) = \int P_M(x, \mathrm{d}z)R(z, \mathrm{d}y)$. R can be for example the Metropolis Adjusted Langevin algorithm (MALA) or Hamiltonian Monte Carlo (HMC). By incorporating a local MCMC step, the sampler can eventually move to areas where points have small importance ratio and escape at the next iSIR step. Finally, Compared to traditional MCMC algorithms, iSIR chooses one sample $\tilde{X}_{k+1}$ from N proposals $(X_k^1, \dots, X_k^N)$ and thus incurs a non-negligible computational waste since $N-1$ samples are discarded at each step. Naesseth et al. (2020); Cardoso et al. (2022) propose to recycle the candidate pool $(X_k^1, \dots, X_k^N)$ and their normalized weights to form a SNIS estimator. Specifically, the following *roll-out* estimator is considered in Cardoso et al. (2022)

$$\overline{\pi}^M_{\mu,k_0:k}(f) = \frac{1}{k-k_0} \sum_{\ell=k_0+1}^{k} \left\{ \sum_{i=1}^{M} \frac{\widetilde{\omega}_\ell^i}{\sum_{j=1}^M \widetilde{\omega}_\ell^j} f(X_\ell^i) \right\}, \quad \text{where} \quad \widetilde{\omega}_\ell^i = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X_\ell^i),$$

and is shown to have a geometrically decreasing bias, thus inheriting the properties of iSIR.

Optimization of the importance distribution

In light of these results, a crucial and natural question arises: *how can we minimize the KL or 2-Rényi divergence between π and μ within a specific family of probability measures?* We now review the existing methods that explicitly focus on doing so.

In Cappé et al. (2008) the authors propose a method for minimizing $\mu \mapsto \mathsf{KL}(\pi \parallel \mu)$ within the family of probability measures written as $\mu^\theta(\mathrm{d}x) = \int \mu^\theta_{X|Z}(\mathrm{d}x|z)\mu^\theta_Z(\mathrm{d}z)$, where $\mu^\theta_{X|Z}$ and μ^θ_Z are chosen within some parametric classes of probability measures on $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Z}, \mathcal{Z})$ respectively. For such probability measures we have that

$$\mathsf{KL}(\pi \parallel \mu^\theta) \leq \int\int \log \frac{\pi(x)\pi_{Z|X}(z|x)}{\mu^\theta_{X|Z}(x|z)\mu^\theta_Z(z)} \pi(\mathrm{d}x)\pi_{Z|X}(\mathrm{d}z|x) := \mathsf{KL}(\pi \otimes \pi_{Z|X} \parallel \mu^\theta_{X|Z} \otimes \mu^\theta_Z),$$

for all kernels $\pi_{Z|X}$, and we can thus minimize the upper bound with respect to θ and the kernel $x \mapsto \pi_{Z|X}(\mathrm{d}z|x)$ in a coordinate ascent fashion. For a fixed θ_0, the minimum w.r.t. to $\pi_{Z|X}(\cdot|x)$ for all $x \in \mathsf{X}$ is attained at $\mu^{\theta_0}_{Z|X}(\mathrm{d}z|x) \propto \mu^{\theta_0}_{X|Z}(x|z)\mu^{\theta_0}_Z(\mathrm{d}z)$, in which case the upper bound is equal to $\mathsf{KL}(\pi \parallel \mu^{\theta_0})$. As a consequence, optimizing the upper bound is equivalent to solving iteratively,

$$\theta_{t+1} = \operatorname*{argmin}_\theta - \int\int \log \mu^\theta_{X|Z}(x|z)\mu^\theta_Z(z)\mu^{\theta_t}_{Z|X}(\mathrm{d}z|x)\pi(\mathrm{d}x), \tag{1.3.5}$$

and induces the decrease of $\mathsf{KL}(\pi \parallel \mu^\theta)$ at each step. This procedure is the exact equivalent of the *Expectation-Maximization* algorithm Dempster et al. (1977). In particular, it offers **closed form** integrated updates for the means, covariances and weights when μ^θ is a mixture of Gaussian distributions, thus avoiding any parameterization of the weights and covariances. The updates are however given in terms of expectations over π. They are in practice estimated with SNIS at step $t+1$ using μ^{θ_t}, which in some sense is the best available importance distribution *up to step t* that one can choose. This procedure lies more broadly within the family of *Adaptive importance sampling* (AIS) Oh and Berger (1993); Cappé et al. (2004); Douc et al. (2007a); Cornuet et al. (2012); Daudel et al. (2021b) methods, where the importance distributions are updated adaptively using samples and weights from previous iterations. See Bugallo et al. (2017); Elvira and Martino (2021) for a detailed overview. Their convergence properties are also well understood, see Douc et al. (2007a); Marin et al. (2019); Portier and Delyon (2018).

More recently, AIS has shifted towards the use of MCMC, SMC algorithms and deep learning for optimizing the importance distribution. While we have omitted its discussion in the previous

paragraph, the choice of the family on which one optimizes is of the utmost importance. The integrated EM method we have just presented relies on a specific factorization of the importance distribution. Furthermore, in order to obtain closed form updates, $\mu^{\theta}_{X|Z}$ needs to lie within the exponential family and μ^{θ}_{Z} has to be a discrete measure with a number of components chosen beforehand. One also needs to account for the tails of the target distribution. Optimizing within families of probability measures that avoid these constraints altogether while having at the same time a strong *approximation power* is thus crucial. *Normalizing flows* (NF) Rippel and Adams (2013); Dinh et al. (2014); Papamakarios et al. (2021) are able to represent arbitrarily complex probability distributions using deep learning architectures, thus allowing for **automatic** adaptation to the tails and multimodality of the target distribution.

A NF uses invertible neural networks $\mathrm{T}_\theta : \mathsf{X} \to \mathsf{X}$ with **easy to compute** Jacobian to represent probability densities positive everywhere on X. The idea is then to *push* a base distribution p (often a Normal distribution, hence the *normalizing* terminology) to the target distribution π, i.e. fit the parameters of the invertible neural network T_θ so that the law of $\mathrm{T}_\theta(X)$, $X \sim p$, which we denote by $\mathrm{T}_{\theta\#}p$, is approximately π. $\mathrm{T}_{\theta\#}p$ is easy to sample and furthermore its density is tractable. Indeed, by the change of variable formula, see (Bogachev and Ruas, 2007, Chapter 3), $\mathrm{T}_{\theta\#}p(x) = p(\mathrm{T}_\theta^{-1}(x))\left|\mathbf{J}_{\mathrm{T}_\theta^{-1}}(x)\right|$, where $x \mapsto \left|\mathbf{J}_{\mathrm{T}_\theta^{-1}}(x)\right|$ is the determinant of the Jacobian.

Similarly to standard AIS, when fitting $\mathrm{T}_{\theta\#}p$ to π with gradient descent one is faced with an intractable expectation w.r.t. π;

$$\nabla_\theta \mathsf{KL}(\pi \parallel \mathrm{T}_{\theta\#}p) = -\int \nabla_\theta \log p(\mathrm{T}_\theta^{-1}(x)) + \log\left|\mathbf{J}_{\mathrm{T}_\theta^{-1}}(x)\right| \pi(\mathrm{d}x)\,. \tag{1.3.6}$$

At this point, one may estimate the gradient at θ_t using either importance sampling, MCMC or SMC with $\mathrm{T}_{\theta_{t-1}\#}p$ as initial proposal. We focus first on MCMC and SMC based estimates which have been exploited recently in Naesseth et al. (2020); Gabrié et al. (2022); Arbel et al. (2021); Samsonov et al. (2022). See Grenioux et al. (2023) for a more detailed survey and comparison. In Naesseth et al. (2020) the authors use iSIR with proposal $\mathrm{T}_{\theta_{t-1}\#}p$. Gabrié et al. (2022) alternate between an Independent Metropolis-Hastings with proposal $\mathrm{T}_{\theta_{t-1}\#}p$ and a MALA or Unadjusted Langevin algorithm (ULA) initialized at samples from $\mathrm{T}_{\theta_{t-1}\#}p$. Samsonov et al. (2022) use iSIR combined with MALA and Arbel et al. (2021) devise an SMC sampler with a normalizing flow approximating the intermediate bridge distributions thus facilitating their sampling when combined with an MCMC kernel. The convergence of the MCMC only adaptive samplers Naesseth et al. (2020); Gabrié et al. (2022); Samsonov et al. (2022) is undertaken in Kim et al. (2022).

These methods have shown superior performance compared to those which estimate the gradient using importance sampling Müller et al. (2019a); Prangle and Viscardi (2023). The latter also suffer from exponential dependence on the dimension due to the importance weight and hence the optimization procedure fails when the initial distribution $\mathrm{T}_{\theta_0\#}p$ is not already close to π. See Geffner and Domke (2021); Dhaka et al. (2021) for a more thorough disccusion of this matter. On the other hand, MCMC based adaptive IS may not encounter the same issue since MCMC samplers, at least when π is *log concave*, exhibit more favourable dependence on the dimension. Finally, let us emphasize that using MCMC algorithms to learn importance proposals, when we could just use MCMC algorithms directly, is not contradictory in itself. Indeed, after the map T_θ is learned, we can use it to obtain i.i.d. (although approximate) samples *on the fly* and estimate normalizing constants unbiasedly.

We end this section with two research directions that we believe are relevant in the context of the previous discussion.

(**Q1**) Gradient-based samplers such as the unadjusted Langevin algorithm Besag (1994) and Hamiltonian Monte Carlo Neal et al. (2011) have gained widespread popularity. However, the density of the t-th iterate for these samplers is generally intractable, making them unsuitable for use as importance proposals. One possible approach to overcome this limitation is by expanding the dimension of the state space and introducing reverse kernels Neal (2001a). Unfortunately, devising such reverse kernels is challenging in most cases.

Are there any gradient-based samplers that have tractable density? Furthermore, can we leverage all the orbits (i.e. the whole chain) on the trajectory of these samplers in a principled manner to estimate normalizing constants unbiasedly?

(**Q2**) The use of MCMC, weighting mechanisms and resampling has been useful for adressing the various challenges of adaptive importance sampling. Is it possible to identify an appropriate weighting mechanism that, when combined with MCMC and resampling, improves the exploration of the target probability measures?

1.3.2 Sequential Monte Carlo

Importance sampling can be implemented *sequentially* to solve a specific type of problems, called non-linear filtering, involving a time component. As the time horizon expands, so does the dimension of the state space, leading to a degradation of the resulting importance sampling estimates. The solution of Gordon et al. (1993b) that we shall present now allows the rejuvenation of the samples by duplicating those with high importance weights and discarding those with low weights.

Setting

Let M_0 be an initial probability measure. For all $t \in \mathbb{N}_{>0}$ define the Markov kernel $M_t : \mathsf{X} \times \mathcal{X} \ni (x, A) \mapsto \int \mathbb{1}_A(y) M_t(x, \mathrm{d}y)$ and for all $t \in \mathbb{N}$ the potential function $g_t : \mathsf{X} \ni x \mapsto g_t(x)$. Define also for all $t \in \mathbb{N}_{>0}$ the measure, or Feynman-Kac model

$$\gamma_{0:t}(\mathrm{d}x_{0:t}) = M_0(\mathrm{d}x_0) \prod_{s=1}^{t} g_{t-1}(x_{t-1}) M_t(x_{t-1}, \mathrm{d}x_t)\,, \tag{1.3.7}$$

and $\gamma_0(\mathrm{d}x_0) = M_0(\mathrm{d}x_0)$. We subsequently write $\gamma_t(\mathrm{d}x_t) = \int \mathbb{1}_{\mathsf{X}^t}(x_{0:t-1}) \gamma_{0:t}(\mathrm{d}x_{0:t})$. Define respectively the *predictive* and *filtering* probability measures

$$\eta_t(\mathrm{d}x_t) = \gamma_t(\mathrm{d}x_t)/\gamma_t(\mathbb{1}_{\mathsf{X}}), \quad \phi_t(\mathrm{d}x_t) = g_t(x_t)\eta_t(\mathrm{d}x_t)/\eta_t(g_t)\,. \tag{1.3.8}$$

The following recursion linking the predictive and filtering distributions is satisfied

$$\eta_t(\mathrm{d}x_t) = \int \phi_{t-1}(\mathrm{d}x_{t-1}) M_t(x_{t-1}, \mathrm{d}x_t). \tag{1.3.9}$$

From the definitions in (1.3.8) and (1.3.7), we observe that if we have an empirical approximation of ϕ_{t-1}, we can use it to approximate η_t and consequently ϕ_t using (1.3.8). Before delving into the details of how these approximations are obtained, let us provide two examples that demonstrate the utility of this framework for Bayesian inference and the approximation of target distributions known up to a normalizing constant.

Example 1.3.4 (Hidden Markov models)**.** *HMMs consist of an unobserved state process* $\{X_t\}_{t\in\mathbb{N}}$ *and observations* $\{Y_t\}_{t\in\mathbb{N}}$*. They respectively evolve in two general measurable spaces* $(\mathsf{X}, \mathcal{X})$ *and* $(\mathsf{Y}, \mathcal{Y})$*. It is assumed that* $\{X_t\}_{t\in\mathbb{N}}$ *is a Markov chain with transition kernels* $(M_{t+1})_{t\in\mathbb{N}}$ *and initial distribution* M_0*. Given the states* $\{X_t\}_{t\in\mathbb{N}}$*, the observations* $\{Y_t\}_{t\in\mathbb{N}}$ *are independent and*

*for all $t \in \mathbb{N}$, the conditional distribution of the observation Y_t only depends on the current state X_t. This distribution is written $G_t(X_t, .)$ and admits the potential $g_t(x_t, .)$ as density (the dependency in Y_t is made implicit and we drop the second argument). Given a realized observation record $Y_{0:t}$, the predictive and filtering distributions (1.3.8) are then the distributions of X_t given $Y_{0:t-1}$ and X_t given $Y_{0:t}$ respectively. Interestingly, the recursions (1.3.8) and (1.3.7) show that it is even possible to build **online** approximations of the posterior; when a new observation Y_{t+1} is available, we can approximate the distribution of X_{t+1} given $Y_{0:t+1}$ using the approximation of that of X_t given $Y_{0:t}$.*

Example 1.3.5 (SMC sampler Del Moral et al. (2006a))**.** *Let $T \in \mathbb{N}$ and π be some target probability measure defined on $(\mathsf{X}, \mathcal{X})$. Consider a sequence of probability measures $\{\pi_t\}_{t=0}^T$ also defined on $(\mathsf{X}, \mathcal{X})$ with $\pi_{t+1} \ll \pi_t$, π_0 easy to sample and $\pi_T = \pi$, and let $\{K_t\}_{t=0}^T$ be a sequence of Markov kernels such that K_t is π_t-invariant. Using that $\pi_t = \pi_t K_t$ we find*

$$\pi_t(\mathrm{d}x_t) = \int \frac{\frac{\mathrm{d}\pi_t}{\mathrm{d}\pi_{t-1}}(x_{t-1})}{\int \frac{\mathrm{d}\pi_t}{\mathrm{d}\pi_{t-1}}(z_{t-1})\pi_{t-1}(\mathrm{d}z_{t-1})} \pi_{t-1}(\mathrm{d}x_{t-1}) K_t(x_{t-1}, \mathrm{d}x_t)\,,$$

and letting $M_t(x_{t-1}, \mathrm{d}x_t) = K_t(x_{t-1}, \mathrm{d}x_t)$, $g_t(x_t) = (\mathrm{d}\pi_{t+1}/\mathrm{d}\pi_t)(x_t)$ we see that $\eta_t = \phi_{t-1} = \pi_t$. A popular example of sequence $\{\pi_t\}_{t=0}^T$ is the annealing sequence defined by $\pi_t(x) \propto \pi_0(x)^{1-\gamma_t}\pi(x)^{\gamma_t}$ where $\gamma_0 = 0 < \ldots < \gamma_T = 1$ control the interpolation between π_0 and π. The invariant kernel K_t can be a Metropolis-Hastings kernel, which does not require the knowledge of the normalizing constant of each π_t. If the discrepancy between π_{t+1} and π_t is small enough, then it is possible to build efficient particle approximations of each π_t and hence π. This provides an interesting alternative to the SNIS particle approximation where an artificial time component is introduced.

Particle filtering

Let us now detail how the particle approximations of η_t and ϕ_t are obtained sequentially. We recommend Del Moral (2004); Douc et al. (2014) for an in-depth description of the theoretical properties of SMC algorithms and Chopin and Papaspiliopoulos (2020) for a more recent account balancing theoretical aspects, algorithmic details and implementation.

Assume that we have at hand a consistent particle approximation $\phi_t^N = \sum_{i=1}^N \omega_t^i \delta_{\xi_t^i}$ of ϕ_t. By plugging this approximation in (1.3.9) we obtain the mixture

$$\phi_t^N M_{t+1}(\mathrm{d}x_{t+1}) = \sum_{i=1}^N \omega_t^i M_{t+1}(\xi_t^i, \mathrm{d}x_{t+1})\,, \tag{1.3.10}$$

approximating η_{t+1} and which serves as a basis for obtaining its particle approximation. Indeed, this is done by sampling from the same mixture, which boils down to first sampling conditionally i.i.d. ancestor indexes $(A_t^1, \ldots, A_t^N) \overset{\text{iid}}{\sim} \text{Categorical}(\{\omega_t^j\}_{j=1}^N)$ and then sampling for all $i \in [1:N]$, $X_{t+1}^i \sim M_{t+1}(\xi_t^{A_t^i}, \cdot)$. This results in an *unweighted* particle approximation $\eta_{t+1}^N = N^{-1}\sum_{i=1}^N \delta_{\xi_{t+1}^i}$. The particle approximation of ϕ_{t+1} is then obtain by plugging η_{t+1}^N in (1.3.8), which yields

$$\phi_{t+1}^N(\mathrm{d}x_{t+1}) = \sum_{i=1}^N \omega_{t+1}^i \delta_{\xi_{t+1}^i}(\mathrm{d}x_{t+1}), \quad \text{where} \quad \omega_{t+1}^i \propto g_{t+1}(\xi_{t+1}^i)\,.$$

The initial particle approximation is obtained by simply drawing N particle $\xi_0^{1:N} \sim M_0^{\otimes N}$ and then weighting according to g_0. This procedure coincides with the *bootstrap particle filter* with

multinomial resampling Gordon et al. (1993b). Other possible resampling schemes are also possible Douc and Cappé (2005). Another approach to obtaining the particle approximation of η_{t+1} differently consists in simply sampling $\xi_{t+1}^i \sim M_{t+1}(\xi_t^i, \cdot)$ for all $i \in [1:N]$ and setting $\eta_t^N = \sum_{i=1}^N \omega_t^i \delta_{\xi_{t+1}^i}$. The particle approximation of ϕ_{t+1} is then $\phi_{t+1}^N = \sum_{i=1}^N \omega_{t+1}^i \delta_{\xi_{t+1}^i}$ where $\omega_{t+1}^i \propto \omega_t^i g_{t+1}(\xi_{t+1}^i)$. As a result, the weight at step $t+1$ is $\omega_{t+1}^i \propto \prod_{s=0}^{t+1} g_s(\xi_s^i)$ and quickly degenerates as t grows. This phenomenon is in fact directly related to the degeneracy of the importance weights discussed in the previous section, where, after a few time steps, all normalized weights except one are equal to 0, see (Douc et al., 2014, Chapter 10) for numerical evidence. By resampling according to the weights $\{\omega_t^i\}_{i=1}^N$ Gordon et al. (1993b), the particles with small importance weights are eliminated whereas those with large importance weights are replicated with weights reset to $1/N$. Interestingly, the resampling step increases the variance *locally* while at the same time guaranteeing that the particle approximations do not degenerate *globally*.

The particle approximation of the measure γ_t is obtained as a byproduct of η_t^N and is given by

$$\gamma_t^N(\mathrm{d}x_t) = \left\{ \prod_{s=0}^{t-1} \frac{1}{N} \sum_{i=1}^N \omega_s^i \right\} \eta_t^N(\mathrm{d}x_t)\,. \tag{1.3.11}$$

Surprisingly, it provides *unbiased* estimates of $\gamma_t(f)$ Crisan et al. (1998) for any bounded test function f. The particular case of $f = 1$ allows for unbiased estimation of the likelihood of the observations in HMMs in Example 1.3.4 and the unbiased estimation of the normalizing constant in Example 1.3.5.

The theoretical properties of the particle filter (PF) are well understood. More specifically, $\eta_t^N(f)$, $\phi_t^N(f)$ and $\gamma_t^N(f)$ all converge almost surely to $\eta_t(f)$, $\phi_t(f)$ and $\gamma_t(f)$ respectively for bounded test functions f. Deviation inequalities are given in Del Moral and Guionnet (1998) and the particle approximations satisfy central limit theorems with explicit expressions of the asymptotic variances Del Moral and Guionnet (1999); Chopin (2004). Stability results, which are fundamental for assessing the reliability of the particle approximations, are also relatively well understood by now Del Moral and Guionnet (2001); Oudjane and Rubenthaler (2005); Van Handel (2009); Whiteley (2013); Douc et al. (2014). Indeed, at the most basic level, a particle filter consists in the accumulation of approximations; we use the approximation of the filter at step t to obtain the one at step $t+1$. One may then expect the overall approximation error to blow up very fast, in which case the PF would be a truly useless algorithm for Bayesian inference tasks that involve streaming data. Fortunately, the *ergodicity* of the signal $\{X_t\}_{t\in\mathbb{N}}$ and observation process $\{Y_t\}_{t\in\mathbb{N}}$, the most basic assumptions which have been relaxed by now, allow showing that the particle filter converges to the true filter *uniformly* in time. Furthermore, the stability of the PF is also observed for more general models that are not even close to satisfying the strong mixing assumptions.

Particle smoothing

The poor man smoother. We now turn to *smoothing* which is highly relevant in Bayesian inference. Smoothing corresponds to approximating the conditional distribution of the state X_s, or in fact a subsequence of states $X_{s:t}$ for $s \leq t$, given the observation record $Y_{0:t}$. As we will see, smoothing distributions are particularly useful for parameter learning via the EM algorithm. To simplify the discussion, we will focus only on the joint smoothing distribution: i.e. the conditional distribution of $X_{0:t}$ given $Y_{0:t}$.

In the frameowrk of Example 1.3.4 the distribution of $X_{0:t}$ given the observations $Y_{0:t}$ coincides with

$$\phi_{0:t|t}(\mathrm{d}x_{0:t}) = g_t(x_t)\gamma_{0:t}(\mathrm{d}x_{0:t})/\gamma_t(g_t)\,.$$

The filtering distribution ϕ_t at step t is its t-th marginal. Assume that the particle approximation at step t is $\phi^N_{0:t|t}(\mathrm{d}x_{0:t}) = \sum_{i=1}^N \omega_t^i \delta_{\xi^i_{0:t}}(\mathrm{d}x_{0:t})$ where $\xi^i_{0:t} \in \mathsf{X}^{t+1}$ is a particle path. Since we have

$$\phi_{0:t+1|t+1}(\mathrm{d}x_{0:t+1}) \propto g_{t+1}(x_{t+1}) M_{t+1}(x_t, \mathrm{d}x_{t+1}) \phi_{0:t|t}(\mathrm{d}x_{0:t}) \,,$$

we may obtain an estimate of $\phi_{0:t+1|t+1}$ by resampling and propagating according to M_{t+1}, similarly to the particle filter. The key difference is that instead of resampling particles, we select ancestor paths at each step. The particle smoothing approximation is hence given by

$$\phi_{0:t+1|t+1}(\mathrm{d}x_{0:t+1}) = \sum_{i=1}^N \omega_{t+1}^i \delta_{(\xi^{A^i_t}_{0:t}, \xi^i_{t+1})}(\mathrm{d}x_{0:t+1}) \,, \tag{1.3.12}$$

and recursively we see that a particle path $\xi^i_{0:t}$ is made of the ancestors from $t-1$ to 0 of the particle ξ^i_t, i.e.

$$\xi^i_{0:t} = (\xi^i_t, \xi^{\mathrm{E}^i_{t,t-1}}_{t-1}, \ldots, \xi^{\mathrm{E}^i_{t,0}}_0) \,, \quad \text{where} \quad \mathrm{E}^i_{t,s} = A^{\mathrm{E}^i_{t,s+1}}_s \mathbb{1}_{[0:t)}(s) + i \cdot \mathbb{1}_{\{t\}}(s) \,. \tag{1.3.13}$$

Unfortunately, the simplicity of this procedure has a cost; as the time horizon t grows larger than the number of particles N and we keep selecting from the previous paths, the pool of particles with time index far from t in the support of $\phi^N_{0:t|t}$ keeps shrinking. As a result, there likely exists a timestep $t_N \in [1:t]$ (typically $t_N = \mathcal{O}(N)$, see Koskela et al. (2020)) such that $\mathrm{E}^1_{t,t_N} = \ldots = \mathrm{E}^N_{t,t_N}$. Consequently, for s much smaller than t we get a particle approximation $\phi_{0:s|t}$ that is supported only on one particle path. This makes the particle smoothing estimate unpractical for any task involving smoothing over long time horizons.

Forward filtering backward smoothing (FFBS). The FFBS algorithm has been proposed to overcome the degeneracy of the vanilla particle smoother Godsill et al. (2004). The FFBS relies on the following *backward* decomposition of the smoothing distribution, which assumes that the transition kernel M_t admits a density m_t with respect to some dominating measure:

$$\phi_{0:t|t}(\mathrm{d}x_{0:t}) = \phi_t(\mathrm{d}x_t) \prod_{s=0}^{t-1} \mathbf{B}_{\phi_s}(x_{s+1}, \mathrm{d}x_s) \,, \tag{1.3.14}$$

where $\mathbf{B}_{\phi_s}(x_{s+1}, \mathrm{d}x_s) \propto m_{s+1}(x_s, x_{s+1}) \phi_s(\mathrm{d}x_s)$ is the *backward* kernel. As it is expressed in terms of the filter ϕ_s we can estimate it by plugging in the particle approximation ϕ^N_s,

$$\mathbf{B}^N_{\phi_s}(x_{s+1}, \mathrm{d}x_s) = \sum_{i=1}^N \frac{\omega_s^i m_{s+1}(\xi^i_s, x_{s+1})}{\sum_{j=1}^N \omega_s^j m_{s+1}(\xi^j_s, x_{s+1})} \delta_{\xi^i_s}(\mathrm{d}x_s) \,.$$

The FFBS particle approximation $\phi^{N,\mathrm{FFBS}}_{0:t|t}$ of $\phi_{0:t|t}$ is then

$$\begin{aligned}
\phi^{N,\mathrm{FFBS}}_{0:t|t}(\mathrm{d}x_{0:t}) &= \phi^N_t(\mathrm{d}x_t) \prod_{s=0}^{t-1} \mathbf{B}^N_{\phi_s}(x_{s+1}, \mathrm{d}x_s) \\
&= \sum_{i_{0:t} \in [1:N]^{t+1}} \left\{ \omega_t^{i_t} \prod_{s=1}^t \beta_s^{\mathsf{BS}}(i_s, i_{s-1}) \right\} \delta_{(\xi^{i_0}_0, \ldots, \xi^{i_t}_t)}(\mathrm{d}x_{0:t}) \,,
\end{aligned}$$

where for all $(i,j) \in [1:N]^2$, $\beta_s^{\mathsf{BS}}(i,j) \propto \omega^j_{s-1} m_s(\xi^j_{s-1}, \xi^i_s)$. The resulting smoothing estimator is no longer supported on the ancestral paths; instead, its support is made of N^{t+1} weighted particle paths which enables it to avoid the particle degeneracy problem. Interestingly, the FFBS

estimator can also be obtained by marginalizing the vanilla smoothing estimator (1.3.12) w.r.t. to the ancestor random variables $(A_0^{1:N}, \dots, A_{t-1}^{1:N})$ and so its mean squared error is smaller. It is however impractical for any problem of interest that involves a time horizon t large enough as the cost of evaluating all the weight products is $\mathcal{O}(N^{t+1})$. Godsill et al. (2004) propose the *forward filtering backward simulation* (FFBSi) estimator to approximate the FFBS by drawing N particle index trajectories $\{(B_0^i, \dots, B_t^i)\}_{i=1}^N$ according to the weight product. The i-th path for each $i \in [1:M]$ is obtained as follows: draw $B_t^i \sim \text{Categorical}(\{\omega_t^j\}_{j=1}^N)$ and then for all $s \in [0:t-1]$ draw $B_s^i \sim \text{Categorical}(\{\beta_{s+1}^{\mathsf{BS}}(B_{s+1}^i, j)\}_{j=1}^N)$. The resulting particle approximation is then

$$\phi_{0:t|t}^{N,\text{FFBSi}}(\mathrm{d}x_{0:t}) = \frac{1}{N}\sum_{i=1}^N \delta_{(\xi_0^{B_0^i}, \dots, \xi_t^{B_t^i})}(\mathrm{d}x_{0:t})\,.$$

The FFBSi estimator has of course a larger variance than the FFBS but the overall cost for computing an estimate with N paths is now $\mathcal{O}(N^2 t)$). Douc et al. (2011a) propose to further bring down the complexity by using the *accept-reject* algorithm. Indeed, assuming that the transition density m_s is bounded above, i.e. $m_s(x_{s-1}, x_s) \leq c_+$ for all $(x_{s-1}, x_s) \in \mathsf{X}^2$, then we have that $\omega_{s-1}^i m_s(\xi_{s-1}^i, \xi_s^j) \leq c_+ \omega_{s-1}^i$ for all $(i,j) \in [1:N]^2$ and we can thus draw the indexes according to $\text{Categorical}(\{\omega_{s-1}^i\}_{i=1}^N)$ and then accept or reject the proposals. If the kernels are bounded from below, i.e. $m_s(x_{s-1}, x_s) \geq c_-$ for all $(x_{s-1}, x_s) \in \mathsf{X}^2$, the complexity is then provably linear in the number of particles N. However, when this assumption does not hold, the running time may be heavy-tailed, in which case on may resort to a *hybrid rejection sampler*, see Taghavi et al. (2013); Olsson and Westerborn (2017); Dau and Chopin (2022). It was then later suggested to instead sample the s-th index of the i-th backward index trajectory by running a few steps of *Independent Metropolis-Hastings* targeting $\beta_{s+1}^{\mathsf{BS}}(B_{s+1}^i, \cdot)$ with $\text{Categorical}(\{\omega_s^i\}_{i=1}^N)$ as proposal Bunch and Godsill (2012), and later mentioned in Gloaguen et al. (2022) in the context of online smoothing. In this setting, the backward index B_s^i are sampled from the following MH kernel (or its iterations) initialized at $i_s \in [1:N]$,

$$\begin{aligned} K_s(i_s, \mathrm{d}b_s) = &\sum_{k \in [1:N] \setminus i_s} \omega_s^k \min\left(1, m_{s+1}(\xi_s^k, \xi_{s+1}^{B_{s+1}^i}) / m_{s+1}(\xi_s^{i_s}, \xi_{s+1}^{B_{s+1}^i})\right) \delta_k(\mathrm{d}b_s) \\ &+ \left\{1 - \sum_{\ell \in [1:N] \setminus i_s} \omega_s^\ell \min\left(1, m_{s+1}(\xi_s^\ell, \xi_{s+1}^{B_{s+1}^i}) / m_{s+1}(\xi_s^{i_s}, \xi_{s+1}^{B_{s+1}^i})\right)\right\} \delta_{i_s}(\mathrm{d}b_s)\,. \end{aligned}$$

Dau and Chopin (2022) show that using only one step of IMH initialized at $i_s = A_s^{B_{s+1}^i}$, the ancestor of B_{s+1}^i, is enough to produce a fast, consistent and stable estimator.

In contrast with the particle filter, the FFBS as presented above is essentially an offline algorithm. Indeed, there is no obvious way to use it in a context where observations are processed in real time without incurring a prohibitively large computational cost. Furthermore, even when used in an offline context both its memory and computational costs grow linearly with the time horizon t. It can be made online if we restrict ourselves to computing smoothing estimates of *additive functionals* Del Moral et al. (2010b). We say that $h_{0:t}$ is an additive functional if

$$h_{0:t}(x_{0:t}) = \sum_{s=0}^{t-1} \tilde{h}_s(x_s, x_{s+1})\,, \quad \forall x_{0:s} \in \mathsf{X}^{t+1}\,. \tag{1.3.15}$$

As we will now see, it is possible to obtain $\phi_{0:t+1|t+1}^{N,\text{FFBS}}(h_{0:t+1})$ from $\phi_{0:t|t}^{N,\text{FFBS}}(h_{0:t})$ for such functionals with $\mathcal{O}(N^2)$ memory cost and operations that can be further reduced with additional approximations. Note that smoothed expectations of such functionals appear naturally in the context of parameter learning for HMMs with the EM algorithm; the *E-step* is in fact a smoothed

expectation of the joint log likelihood, which is an additive functional. In pairwise marginal smoothing, any expectation w.r.t. $\phi_{s-1:s|t}$ for $s \in [1:t]$ is the expectation of an additive functional.

Define for all $t \in \mathbb{N}_{>0}$, $\mathbf{T}_t(x_t, \mathrm{d}x_{0:t-1}) = \prod_{\ell=1}^{t} \mathbf{B}_{\phi_\ell}(x_\ell, \mathrm{d}x_{\ell-1})$. The decomposition (1.3.14) is alternatively written as $\phi_{0:t|t}(\mathrm{d}x_{0:t}) = \phi_t(\mathrm{d}x_t)\mathbf{T}_t(x_t, \mathrm{d}x_{0:t-1})$. We write

$$\mathbf{T}_t[h_{0:t}](x_t) := \int h_{0:t}(x_{0:t})\mathbf{T}_t(x_t, \mathrm{d}x_{0:t-1})\,.$$

For all $t \in \mathbb{N}_{>1}$, $\mathbf{T}_t[h_{0:t}]$ satisfies the following recursion,

$$\begin{aligned}\mathbf{T}_t[h_{0:t}](x_t) &= \int \left\{h_{0:t-1}(x_{0:t-1}) + \tilde{h}_{t-1}(x_{t-1}, x_t)\right\} \mathbf{B}_{\phi_{t-1}}(x_t, \mathrm{d}x_{t-1})\mathbf{T}_{t-1}(x_{t-1}, \mathrm{d}x_{0:t-2}) \\ &= \int \left\{\mathbf{T}_{t-1}[h_{0:t-1}](x_{t-1}) + \tilde{h}_{t-1}(x_{t-1}, x_t)\right\} \mathbf{B}_{\phi_{t-1}}(x_t, \mathrm{d}x_{t-1})\,.\end{aligned}$$

Consequently, starting from the initial particle approximation,

$$\begin{aligned}\mathbf{T}_1^N[h_{0:1}](x_1) &:= \int \tilde{h}_0(x_0, x_1)\mathbf{B}_{\phi_0}^N(x_1, \mathrm{d}x_0) \\ &= \sum_{i=1}^{N} \frac{\omega_0^i m_1(\xi_0^i, x_1)}{\sum_{j=1}^N \omega_0^j m_1(\xi_0^j, x_1)} \tilde{h}_0(\xi_{t-1}^i, x)\,,\end{aligned}$$

we can build recursively the particle approximations of each $\mathbf{T}_t[h_{0:t}]$

$$\mathbf{T}_t^N[h_{0:t}](x_t) := \sum_{i=1}^{N} \frac{\omega_{t-1}^i m_t(\xi_{t-1}^i, x_t)}{\sum_{j=1}^N \omega_{t-1}^j m_t(\xi_{t-1}^j, x_t)} \left\{\mathbf{T}_{t-1}^N[h_{0:t-1}](\xi_{t-1}^i) + \tilde{h}_{t-1}(\xi_{t-1}^i, x_t)\right\}\,. \tag{1.3.16}$$

The particle approximation $\phi_{0:t|t}^{N,\mathrm{FFBS}}(h_{0:t})$ is recovered by integrating $\mathbf{T}_t^N[h_{0:t}]$ w.r.t. ϕ_t^N. From (1.3.16) we see that in order to approximate the smoothed expectations online we only need to evaluate the functionals $\mathbf{T}_t^N[h_{0:t}]$ in the particles $\xi_t^{1:N}$. Thus, the forward smoother can run in parallel of the particle filter. Letting $\tau_t^i := \mathbf{T}_t^N[h_{0:t}](\xi_t^i)$, this forward version of the FFBS then boils down to the following recursion

$$\tau_t^i = \sum_{j=1}^{N} \beta_t^{\mathsf{BS}}(i,j) \left\{\tau_{t-1}^j + \tilde{h}_{t-1}(\xi_{t-1}^j, \xi_t^i)\right\}\,. \tag{1.3.17}$$

The per time step cost of this forward version of the FFBS is $\mathcal{O}(N^2)$ but can be reduced by approximating the update (1.3.14) by means of additional Monte Carlo simulation Olsson and Westerborn (2017), yielding an "FFBSi"-like forward-only smoother. It is obtained as follows; at each step t and for all $i \in [1:N]$, sample $(J_{i,t-1}^1, \ldots, J_{i,t-1}^M) \overset{\mathrm{iid}}{\sim} \beta_t^{\mathsf{BS}}(i, \cdot)$ (or approximately) and set

$$\widetilde{\tau}_t^i = \frac{1}{M} \sum_{k=1}^{M} \widetilde{\tau}_{t-1}^{J_{i,t-1}^k} + \tilde{h}_{t-1}(\xi_{t-1}^{J_{i,t-1}^k}, \xi_t^i)\,. \tag{1.3.18}$$

The smoothing approximation is then $\phi_{0:t|t}^{M,\mathtt{PARIS}}(h_{0:t}) := \sum_{i=1}^N \omega_t^i \widetilde{\tau}_t^i$. Similarly to the FFBSi, the indexes can be sampled either by accept-reject, hybrid accept-reject or IMH, thus reducing the complexity if M is small enough. The key feature of this smoothing estimator is that M does not necessarily need to be large in order to ensure performances comparable to those of the forward smoother with (1.3.16). Theoretical results and numerical experiments in Olsson and Westerborn (2017) illustrate that setting $M \in \{2, 3\}$ is enough to produce a stable estimator

with good performance. Interestingly, the case $M = 1$ yields an estimator that degenerates but not in the same way as the vanilla particle smoother (1.3.12). In this case, the support of the particle approximation $\phi_{0:t|t}^{M,\texttt{PARIS}}$ is made of particle trajectories that correspond to what one would obtain by running the vanilla particle smoother (1.3.12) and selecting ancestor particle trajectories according to the backward weights. The resulting estimator degenerates eventually and the PaRIS circumvents this issue by selecting more than one ancestor.

Monte Carlo error

As for static Monte Carlo, assessing the error of SMC estimators is crucial for obtaining confidence intervals and comparing the performance of two different estimators, e.g. in the context of auxiliary PF Pitt and Shephard (1999). The particles involved in SMC estimates are far from being i.i.d. which renders the theoretical expression of their asymptotic variances (given by the CLT) quite convoluted Chopin (2004); Douc et al. (2011a). For the asymptotic variance of the particle filter, see Section B.4 where we derive its theoretical expression from first principles. The main challenge is then to estimate the asymptotic variances using only a **single** run, as running multiple instances of the PF or particle smoother in parallel may not be feasible due to computational constraints.

The first breakthrough in estimating the asymptotic variance of the PF was achieved in Chan and Lai (2013) where a strikingly simple consistent estimator is provided. For the asymptotic variance of the predictive distribution particle approximation, their estimator reads

$$\mathcal{V}_{\eta,t}^{N}(h) := -N^{-1} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i \neq \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}\,, \tag{1.3.19}$$

where for $i \in [1:N]$, $\mathrm{E}_{t,0}^i$ is time 0 ancestor of ξ_t^i defined formally in (1.3.13). (1.3.19) allows tracking the asymptotic variance **online** since $\mathrm{E}_{t,0}^i$ can be computed recursively by adding one line of code to the original particle filter. This estimator was later refined and extended in Lee and Whiteley (2018); Du and Guyader (2021). Lee and Whiteley (2018) provide a different derivation and proof of consistency for (1.3.14) using techniques developed in Cérou et al. (2011); Andrieu et al. (2018a) as well as a novel estimator. Du and Guyader (2021) extend all the previous work to the adaptive SMC framework Beskos et al. (2016).

Unfortunately, the simplicity of (1.3.19) comes at a cost; as we have mentioned earlier, after $\mathcal{O}(N)$ time steps all the particles end up with the same ancestor at time 0, i.e. $\mathrm{E}_{t,0}^i = \mathrm{E}_{t,0}^j$ for all $(i,j) \in [1:N]^2$, t larger than N, and (1.3.19) collapses to 0. Inspired by the *fixed-lag* smoother Kitagawa (1993), Olsson and Douc (2019) workaround the degeneracy by only considering the ancestors up to some time $t-\lambda$ where $\lambda \in [0:t]$ is known as the *lag*. Their estimator reads

$$\mathcal{V}_{\eta,t}^{N,\lambda}(h) := -N^{-1} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,t-\lambda}^i \neq \mathrm{E}_{t,t-\lambda}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}\,. \tag{1.3.20}$$

In the regime where (1.3.19) degenerates, (1.3.20) can be made stable provided that the lag λ is chosen carefully such that there is little asymptotic bias.

(**Q3**) As we have seen, the FFBS solves to a certain extent the degeneracy problem of the vanilla particle smoother. Is it then possible to derive a FFBS version of (1.3.19)? Can it be made online?

(**Q4**) While there exists an explicit expression of the asymptotic variance for the FFBS Douc et al. (2014), there is currently no estimator in the literature. Can the techniques developed in the aforementionned papers be extended to derive consistent asymptotic variance estimators for the FFBS?

Particle Gibbs

Similar to importance sampling, the particle filter enables the definition of an ergodic sampler that targets $\phi_{0:t|t}$. This sampler is known as the *Particle Gibbs* (PG) or *conditional particle filter* (CPF) and is the sequential counterpart of iSIR Andrieu et al. (2010). At each step, a promising trajectory $(\zeta_0, \dots, \zeta_t)$ is selected by sampling from the particle smoothing approximation and then inserted in the pool of particle trajectories at the next iteration. More formally, let $k \in \mathbb{N}$ be the iteration index and $\zeta_{0:t}[k] := (\zeta_0[k], \dots, \zeta_t[k])$ the trajectory selected at the same iteration. PG consists in running a modified particle filter where at step $s \in [1:t]$ we set $\xi_s^1 = \zeta_s[k]$, $A_s^1 = 1$, sample the remaining $N-1$ particles from $\phi_{s-1}^N M_s$ and set $\phi_s^N = \sum_{i=1}^N \omega_s^i \delta_{\xi_s^i}$. $\zeta_{0:t}[k+1]$ is then obtained by sampling from the vanilla particle smoothing approximation $\phi_{0:t|t}^N$ (1.3.12) where by construction $\xi_{0:t}^1 = \zeta_{0:t}[k]$. The procedure we have just described defines a Markov chain that converges geometrically fast to $\phi_{0:t|t}$ under standard strong mixing assumptions.

As is the case of particle smoothing, sampling from $\phi_{0:t|t}^N$ results in bad mixing since $\phi_{0:s|t}^N$ has large variance if $s \ll t$. This may be overcome by instead sampling the particle trajectory with backward sampling Whiteley (2010). The resulting sampler is still geometrically ergodic, is provably better than the vanilla PG and only requires the number of particles to scale linearly with the trajectories length t Lee et al. (2020).

(**Q5**) Similar to iSIR, the PG suffers from significant computational waste as only one trajectory is kept at each iteration. Is it possible to recycle the successive particle clouds generated at each step of the PG to generate smoothing estimates with reduced bias, thereby extending the approach presented in Cardoso et al. (2022) to state space models?

1.3.3 Generative modeling

Generative modeling consists in finding a suitable model of observed data which can then be used to generate new samples that resemble the original data distribution. The basic underlying idea is that a dataset $Y^{1:N} := (Y^1, \dots, Y^N)$ is assumed to be N i.i.d. realizations of some unknown probability distribution π_Y defined on a measurable space $(\mathsf{Y}, \mathcal{T})$. As an illustration, a datum $Y^i = [Y_1^i, \dots, Y_d^i]$ could be images in which case Y may be $[0,1]^d$ for simplicity and π_Y exhibits spatial correlation, or $Y^i \in \mathsf{Y} = \mathbb{R}^T$ is a time series or natural language text and π_Y may be in this case Markovian. More generally, the data generating distribution is highly complex and the current effort in generative modeling lies in designing powerful parametric approximations q_Y^θ of π_Y which are easy to sample with possibly tractable density. Here approximation means that $D(q_Y^\theta, \pi_Y)$ is relatively small, where D is some (pseudo-)distance or discrepancy measure between two probability distributions.

Of course, the design choices for q_Y^θ are restricted by the fact that it needs to be a valid probability measure and hence be non-negative and integrate to one; i.e. $\int_{\mathsf{Y}} q_Y^\theta(\mathrm{d}y) = 1$. By now there are many ways for enforcing these two constraints and each design has its own tradeoffs. We herebelow detail three different design choices. To keep the presentation simple, we use the standard notations for conditional densities.

Variational auto-encoder

Variational auto-encoders (VAE) Kingma and Welling (2013) model q_Y^θ through a parameterized *latent variable model* (LVM),

$$q_Y^\theta(\mathrm{d}y) = \int q_{Y|X}^\theta(\mathrm{d}y|x) q_X(\mathrm{d}x)\,, \tag{1.3.21}$$

where q_X is a probability measure on a measurable space $(\mathsf{X}, \mathcal{X})$ and the dimension of the latent space X need not be the same as that of Y. A possible choice for q_X and $q^\theta_{Y|X}$ when $\mathsf{Y} = \mathbb{R}^d$ and $\mathsf{X} = \mathbb{R}^\ell$ is $q_X(x) = \mathcal{N}(x; \mathbf{0}_\ell, \mathbf{I}_\ell)$ and $q^\theta_{Y|X}(y|x) = \mathcal{N}(y; \mu_\theta(x), \mathbf{I}_d \cdot \sigma^2_\theta(x))$ where for all $\theta \in \Theta$, $\mu_\theta : \mathbb{R}^\ell \to \mathbb{R}^d$, $\sigma^2_\theta : \mathbb{R}^\ell \to \mathbb{R}^d$ are neural networks. In this case, (1.3.21) may thus be interpreted as an infinite mixture of Gaussian distributions. VAEs are trained by maximizing the following lower bound on the log-likelihood coined *Evidence LOwer Bound* (ELBO) obtained by introducing a parametric *importance distribution* $\nu^\varphi_{X|Y}$ and using Jensen's inequality

$$\begin{aligned}\sum_{i=1}^N \log q^\theta_Y(Y^i) &= \sum_{i=1}^N \log \int \frac{q^\theta_{Y|X}(Y^i|x) q^\theta_X(x)}{\nu^\varphi_{X|Y}(x|Y^i)} \nu^\varphi_{X|Y}(\mathrm{d}x|Y^i) \\ &\geq \sum_{i=1}^N \int \log \frac{q^\theta_{Y|X}(Y^i|x) q^\theta_X(x)}{\nu^\varphi_{X|Y}(x|Y^i)} \nu^\varphi_{X|Y}(\mathrm{d}x|Y^i) =: \mathcal{L}^N(\theta, \varphi)\,.\end{aligned} \tag{1.3.22}$$

In Kingma and Welling (2013), the authors choose $\nu^\varphi_{X|Y}(\cdot|y)$ to be the density of a multivariate Gaussian with mean $\mu_\varphi(y)$ and covariance $\mathbf{I}_\ell \cdot \sigma^2_\varphi(y)$ where $\mu_\varphi : \mathbb{R}^d \to \mathbb{R}^\ell$ and $\sigma^2_\varphi : \mathbb{R}^d \to \mathbb{R}^\ell$ are neural networks. The gradient of the lower bound in (1.3.22) can then be estimated by drawing samples from $\nu^\varphi_{X|Y}(\cdot|y)$ and noting that a sample $X \sim \nu^\varphi_{X|Y}(\cdot|y)$ can be written as $\mu_\varphi(y) + \sigma_\varphi(y)Z \sim \nu^\varphi_{X|Y}(\cdot|y)$ where $Z \sim \mathcal{N}(\mathbf{0}_\ell, \mathbf{I}_\ell)$, thus allowing the decoupling of the randomness (Z) from the parameter (φ).

Choosing a coordinate ascent point of view we see that the optimization of the lower bound in (1.3.22) involves iteratively minimizing the Kullback-Leibler (KL) divergence between the importance distribution and the true posterior, and maximizing the expected log-likelihood of the data given the latent variables,

$$\varphi_{k+1} = \operatorname*{argmin}_\varphi \sum_{i=1}^N \mathsf{KL}(\nu^\varphi(\cdot|Y^i) \parallel q^{\theta_k}_{X|Y}(\cdot|Y^i))\,, \tag{1.3.23}$$

$$\theta_{k+1} = \arg\max_\theta \sum_{i=1}^N \int \log q^\theta_{Y|X}(Y^i|x) \nu^{\varphi_{k+1}}(\mathrm{d}x|Y^i)\,. \tag{1.3.24}$$

This procedure recovers the well known *Expectation Maximization* algorithm Dempster et al. (1977) if the variational family in which the variational posteriors $\nu^\varphi_{X|Y}(\cdot|y)$ lie contains $\{q^\theta_{Y|X}(\cdot|y) : y \in \mathsf{Y}, \theta \in \Theta\}$. For this reason, a subset of recent works have focused on departing from the vanilla *mean-field* parameterization of Kingma and Welling (2013) and building more expressive variational posteriors Rezende and Mohamed (2015); Salimans et al. (2015); Wolf et al. (2016); Hoffman (2017); Thin et al. (2020); Papamakarios et al. (2021).

Energy based models

Energy based models (EBM) Ackley et al. (1985) draw inspiration from Boltzmann distributions in statistical physics to introduce

$$q^\theta_Y(y) = \frac{\exp(-E_\theta(y))}{\mathcal{Z}_\theta}\,, \quad \text{where} \quad \mathcal{Z}_\theta = \int \exp(-E_\theta(y))\mathrm{d}y\,. \tag{1.3.25}$$

The normalizing constant $\mathcal{Z}_\theta$ is also known as the *partition function*. In terms of flexibility, the EBM model (1.3.25) is generally more versatile than (1.3.21), as the latter is bottlenecked by the less flexible parametrization of the variational posterior. However, this increased flexibility comes with a tradeoff since the normalizing constant $\mathcal{Z}_\theta$ is intractable and (1.3.25) is unavailable

for sampling. This has implications for maximum likelihood training through SGD of the EBM as we now see.

Consider the gradient of the log-likelihood objective function:

$$\nabla_\theta \frac{1}{N}\sum_{i=1}^{N} \log q_Y^\theta(Y^i) \mid_{\theta_0} = -\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta E_\theta(Y^i)_{|\theta_0} + \nabla_\theta \log \mathcal{Z}_\theta \mid_{\theta_0} .$$

By utilizing *Fisher's identity*, we can express the gradient of the partition function as follows

$$\begin{aligned}\nabla_\theta \log \mathcal{Z}_\theta \mid_{\theta_0} &= \frac{\int -\nabla_\theta \exp\big(-E_\theta(y)\big)_{|\theta_0} \mathrm{d}y}{\int \exp(-E_{\theta_0}(y))\mathrm{d}y} = \frac{-\int \nabla_\theta E_\theta(y)_{|\theta_0} \exp\big(-E_{\theta_0}(y)\big)\mathrm{d}y}{\int \exp(-E_{\theta_0}(y))} \\ &= -\int \nabla_\theta E_\theta(y)_{|\theta_0} q_Y^{\theta_0}(\mathrm{d}y)\,,\end{aligned}$$

where we have used common differentiability assumptions. Hence, the gradient of the partition function may be estimated by drawing independent approximate samples from $q_Y^{\theta_0}(\mathrm{d}y)$ using MCMC Tieleman (2008); Qiu et al. (2020), IS or SMC Carbone et al. (2023) which do not require the knowledge of $\mathcal{Z}_{\theta_0}$.

Denoising diffusion probabilistic models

Denoising diffusion probabilistic models (DDPM) or diffusion models Ho et al. (2020); Song et al. (2021c) rely on a forward noising process which slowly turns the data distribution π_Y into pure noise; i.e. a standard Gaussian. The goal is then to *reverse* this noising process so that we can turn noise into samples from the data distribution. The forward noising process is a Markov chain with transition

$$q_{t+1}(x_{t+1}|x_t) := \mathcal{N}(x_{t+1}; (1-\beta_{t+1})^{1/2} x_t, \beta_t \mathbf{I}_d)\,,$$

where $\{\beta_t\}_{t\in\mathbb{N}} \subset (0,1)$. The joint distribution of the noising process is then given by

$$\mathsf{q}_{0:n}(\mathrm{d}x_{0:n}) := \pi_Y(\mathrm{d}x_0) \prod_{s=0}^{n-1} q_{s+1}(\mathrm{d}x_{s+1}|x_s)\,, \tag{1.3.26}$$

where $n \in \mathbb{N}$ is some final time step which we assume to be large enough. The marginal distributions of this joint process $\mathsf{q}_s(\mathrm{d}x_s) := \int \mathsf{q}_{0:n}(\mathrm{d}x_{0:n})$ have an intuitive interpretation; they slowly bridge between the initial distribution π_Y and the terminal one q_n, which is approximately a standard Gaussian if n is large enough. The forward process (1.3.26) can be reversed; we can write that

$$\mathsf{q}_{0:n}(\mathrm{d}x_{0:n}) = \mathsf{q}_n(\mathrm{d}x_n) \prod_{s=0}^{n-1} q_s(\mathrm{d}x_s|x_{s+1})\,, \quad \text{where} \quad q_s(\mathrm{d}x_s|x_{s+1}) := \frac{q_{s+1}(x_{s+1}|x_s)\mathsf{q}_s(\mathrm{d}x_s)}{\mathsf{q}_{s+1}(x_{s+1})}\,.$$

The backward decomposition above suggests that we can turn pure noise into samples from π_Y if we can approximate the backward kernels. For this purpose the following parameterized backward process is introduced

$$\mathsf{p}_{0:n}^\theta(\mathrm{d}x_{0:n}) := \mathsf{p}_n(\mathrm{d}x_n) \prod_{s=0}^{n-1} p_s^\theta(\mathrm{d}x_s|x_{s+1})\,, \tag{1.3.27}$$

where for all $s \in [1:n]$,

$$p_s^\theta(x_{s-1}|x_s) := \mathcal{N}\left(x_{s-1}; \frac{1}{(1-\beta_s)^{1/2}}\left\{x_s - \frac{1-\beta_s}{(1-\bar{\alpha}_s)^{1/2}}\epsilon_s^\theta(x_s)\right\}, \beta_s \mathbf{I}_d\right), \tag{1.3.28}$$

where $\bar{\alpha_s} = \prod_{\ell=1}^s (1-\beta_\ell)$, $\{\epsilon_s^\theta\}_{s=1}^n$ are neural networks and $\mathsf{p}_n := \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. See Ho et al. (2020) for more details on this particular parameterization. Denote by p_0^θ the time 0 marginal of (1.3.27). The backward process is learned by minimizing an the KL divergence between $\mathsf{q}_{0:n}$ and $\mathsf{p}_{0:n}^\theta$ which upperbounds the KL between π_Y and p_0^θ,

$$\mathsf{KL}(\pi_Y \parallel \mathsf{p}_0^\theta) \leq \mathsf{KL}(\mathsf{q}_{0:n} \parallel \mathsf{p}_{0:n}^\theta)\,.$$

Then, using that under the forward process (1.3.26) X_s given X_{s+1} and X_0 is distributed according to a Gaussian distribution with mean linear in X_{s+1} and X_0, see Ho et al. (2020), we find that

$$\begin{aligned}
&\operatorname*{argmin}_\theta \mathsf{KL}(\mathsf{q}_{0:n} \parallel \mathsf{p}_{0:n}^\theta) \\
&= \operatorname*{argmin}_\theta \int \frac{\beta_s}{2(1-\beta_s)(1-\bar{\alpha}_s)} \sum_{s=1}^n \|\epsilon - \epsilon_s^\theta(\bar{\alpha}_s^{1/2} x_0 + (1-\bar{\alpha}_s)^{1/2}\epsilon)\|_2^2 \pi_Y(\mathrm{d}x_0)\mathcal{N}(\epsilon; \mathbf{0}_d, \mathbf{I}_d)\mathrm{d}\epsilon\,.
\end{aligned} \tag{1.3.29}$$

In short, the *score* networks $\{\epsilon_s^\theta\}_{s=1}^n$ aim at predicting ϵ from the input $\bar{\alpha}_s^{1/2} x_0 + (1-\bar{\alpha}_s)^{1/2}\epsilon$. They can also be interpreted as approximations to $\nabla \log \mathsf{q}_s$ (up to a constant). To see why this is the case, note that by Fisher's identity

$$\nabla \log \mathsf{q}_s(x_s) = \int \nabla \log q_{s|0}(x_s|x_0) q_{0|s}(\mathrm{d}x_0|x_s)\,,$$

where $q_{s|0}(\mathrm{d}x_s|x_0) := \int \prod_{\ell=0}^{s-1} q_{\ell+1}(\mathrm{d}x_{\ell+1}|x_\ell) = \mathcal{N}(\mathrm{d}x_s; \bar{\alpha}_s^{1/2} x_0, (1-\bar{\alpha}_s)\mathbf{I}_d)$ and $q_{0|s}(\mathrm{d}x_0|x_s) \propto \pi_Y(\mathrm{d}x_0) q_{s|0}(x_s|x_0)$. Hence, using the previous definitions and the fact that

$$\pi_Y(\mathrm{d}x_0) q_{s|0}(\mathrm{d}x_s|x_0) = \mathsf{q}_s(\mathrm{d}x_s) q_{0|s}(\mathrm{d}x_0|x_s)\,,$$

we find

$$\begin{aligned}
&\mathbb{E}\left[\|\epsilon - \epsilon_s^\theta(\bar{\alpha}_s^{1/2} X_0 + (1-\bar{\alpha}_s)^{1/2}\epsilon)\|_2^2\right] \\
&= \mathbb{E}\left[\|(1-\bar{\alpha}_s)^{1/2}\nabla \log q_{s|0}(X_s|X_0) + \epsilon_s^\theta(X_s)\|_2^2\right] \\
&= \mathbb{E}\left[\|\epsilon_s^\theta(X_s)\|_2^2\right] + 2(1-\bar{\alpha}_s)^{1/2} \int \nabla \log q_{s|0}(x_s|x_0)^\intercal \epsilon_s^\theta(x_s) \mathsf{q}_s(\mathrm{d}x_s) q_{0|s}(\mathrm{d}x_0|x_s) + C_1 \\
&= \mathbb{E}\left[\left\|(1-\bar{\alpha}_s)^{1/2}\nabla \log \mathsf{q}_s(X_s) + \epsilon_s^\theta(X_s)\right\|_2^2\right] + C_2\,,
\end{aligned}$$

where C_1 and C_2 are constant independent of θ. This insight is also in agreement with the SDE interpretation of diffusion models Song et al. (2021c). The forward process we have introduced is simply the discretized version of an Ornstein-Uhlenbeck process initialized at π_Y and the associated reverse SDE has a drift involving $\nabla \log \mathsf{q}_s$. The backward process (1.3.27) can thus be seen as a discretization of the reverse process started at p_n.

By now DDPMs have outperformed GANs and VAEs, which were state of the art just two years ago, and are the backbones of the very recent *stable diffusion* Rombach et al. (2022) which produces stunning high resolution images. It is thus interesting to understand the key

distinctions between a DDPM (Denoising Diffusion Probabilistic Model) and a VAE (Variational Autoencoder) since both are latent variable models. A VAE learns to transform random noise into the desired data distribution (1.3.21) in a single step using $q^{\theta}_{Y|X}$ and it attempts to reverse this process in a single step as well through $\nu^{\varphi}_{X|Y}$ which is assumed to be Gaussian and is known to be suboptimal. On the other hand, DDPMs rely on multiple simple transformations during the forward process that are comparatively easier to reverse and, for which the Gaussian approximation (1.3.28) is no longer suboptimal according to the SDE interpretation. As a result however, it is more expensive to train and sample from a DDPM since n needs to be large enough for the approximation (1.3.27) to be valid. This divide and conquer approach, which involves sampling from intermediate distributions which are simple to bridge, has also been used successfully in the Monte Carlo literature Neal (2001a); Del Moral et al. (2006b).

DDPMs have allowed interesting developments in *controlled generation*. Assume that we are interested in sampling from $\phi^y_0(\mathrm{d}x_0) \propto g^y_0(x_0)\pi_Y(\mathrm{d}x_0)$. The potential g^y_0 is interpreted as the likelihood of some observation y given x_0 and can be given for example by a Gaussian linear inverse problem. Denote by ϕ^y_s the marginals of the forward noising process applied to ϕ^y_0, i.e.

$$\phi^y_s(\mathrm{d}x_s) := \int \phi^y_0(\mathrm{d}x_0) q_{s|0}(\mathrm{d}x_s|x_0)\,.$$

Following the previous developments, if we are able to approximate $\nabla \log \phi^y_s$ then we can obtain approximate samples from ϕ^y_0 using a DDPM. By reversing the forward process and using the definition of ϕ^y_0, we have that

$$\phi^y_s(\mathrm{d}x_s) \propto \mathsf{q}_s(\mathrm{d}x_s) \int g^y_0(x_0) q_{0|s}(\mathrm{d}x_0|x_s)\,,$$

and thus $\nabla \log \phi^y_s(x_s) = \nabla \log \mathsf{q}_s(x_s) + \nabla \log \int g^y_0(x_0) q_{0|s}(\mathrm{d}x_0|x_s)$. In Ho et al. (2022); Chung et al. (2023) it is proposed to approximate the intractable integral $\int g^y_0(x_0) q_{0|s}(\mathrm{d}x_0|x_s)$ by $g^y_0(\mathbb{E}[X_0|X_s])$ where by Tweedie's formula, we have under the forward process that

$$\mathbb{E}[X_0|X_s] = \{X_s + (1-\bar{\alpha}_s)\nabla \log \mathsf{q}_s(X_s)\}/\bar{\alpha}_s^{1/2}\,. \tag{1.3.30}$$

As a result, we can approximate the conditional score $\nabla \log \phi^y_s$ if we have score networks $\{\epsilon^{\theta}_s\}_{s=1}^n$ approximating $\nabla \log \mathsf{q}_s$, i.e.

$$\nabla \log \phi^y_s(x_s) \approx -(1-\bar{\alpha}_s)^{-1/2}\epsilon^{\theta}_s(x_s) + \nabla \log g^y_0(\boldsymbol{\chi}^{\theta}_s(x_s))\,, \tag{1.3.31}$$

where

$$\boldsymbol{\chi}^{\theta}_s(x_s) := \left\{x_s - (1-\bar{\alpha}_s)^{1/2}\epsilon^{\theta}_s(x_s)\right\}/\bar{\alpha}_s^{1/2}$$

approximates (1.3.30). In practical terms, the significance of this approach lies in its ability to tackle various controlled generation tasks using the same data distribution π_Y. This is made possible by relying solely on the score networks $\{\epsilon^{\theta}_s\}_{s=1}^n$ that approximate the data distribution. To illustrate its importance, consider its application in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), where the reconstruction of medical images from partial measurements is crucial. Traditionally, supervised learning methods for image reconstruction in CT and MRI heavily depend on paired datasets consisting of both partial measurements and corresponding complete medical images. However, this approach presents significant drawbacks as it imposes stringent restrictions and incurs high costs. The requirement of pairing partial measurements with their corresponding complete images poses a considerable challenge in data acquisition, preparation, and annotation. The approach we have just described does not bear such constraints.

(**Q6**) Undoubtedly, the approximation (1.3.31) introduces a non-negligible bias and is not guaranteed to sample from the targeted posterior. This, in turn, gives rise to safety concerns, particularly in sensitive applications like medical imaging. There is thus an urgent need to develop controlled generation methods that have at least asymptotic guarantees. On the other hand, sampling approximately from ϕ_0^y using bridge distributions is the kind of tasks that are handled best with SMC samplers Del Moral et al. (2006b). How can we leverage the bridge distributions of DDPM and SMC to devise an asymptotically exact method for sampling from the posterior of the diffusion model $\phi_0^y(\mathrm{d}x_0) \propto g_0^y(x_0)\mathsf{p}_0^{\theta_*}(\mathrm{d}x_0)$ where θ_* is the approximate solution of (1.3.29)?

1.4 Outline and contributions of this thesis.

The content of the present thesis is motivated by the research questions (**Q**1-2-3-4-5-6) studied in the following papers which constitute the five remaining chapters of this thesis.

1. NEO: Non-equilibrium sampling on the orbit of a deterministic transform (Thin et al., 2021).
   Achille Thin, **Yazid Janati El Idrissi**, Sylvain Le Corff, Charles Ollion, Eric Moulines, Arnaud Doucet, Alain Durmus, Christian X. Robert.
   Advances in Neural Information Processing Systems 35 (NeurIPS) (2021).
2. Entropic Mirror Monte Carlo.
   Yazid Janati El Idrissi, Alain Durmus, Sylvain Le Corff, Yohan Petetin, Julien Stoehr.
   Preliminary work.
3. Variance estimation for SMC algorithms: a backward sampling approach (Janati et al., 2023).
   Yazid Janati El Idrissi, Sylvain Le Corff, Yohan Petetin.
   To appear in Bernoulli.
4. State and parameter learning with the PaRIS particle Gibbs (Cardoso et al., 2023a).
   Gabriel Cardoso, **Yazid Janati El Idrissi**, Sylvain Le Corff, Eric Moulines, Jimmy Olsson.
   International Conference in Machine Learning 40 (ICML) (2023).
5. Monte Carlo guided Diffusion for Bayesian linear inverse problems (Cardoso et al., 2023b).
   Gabriel Cardoso, **Yazid Janati El Idrissi**, Sylvain Le Corff, Eric Moulines.
   Under review.

While not present in this thesis, I have also co-authored the following paper on variational inference for *jump state space models*:

- Structured variational Bayesian inference for Gaussian state-space models with regime switching (Petetin et al., 2021).
  Yohan Petetin, **Yazid Janati El Idrissi**, Franccois Desbouvries.
  IEEE Signal Processing Letters.

Below, we provide a summary of the contributions made in each chapter. Please note that we introduce notations in each chapter, although there may be some overlap. These notations are always defined at the beginning of each chapter.

Chapter 2 / (Q1) - Non-equilibrium sampling on the orbit of a deterministic transform

Langevin and Hamiltonian-like dynamics have now become widely popular sampling algorithms due to their ability to efficiently explore high-dimensional spaces and sample from complex distributions. However, they cannot be used as an importance sampling proposal as their density cannot be evaluated. They are instead used in adaptive importance sampling schemes to build proposals well tailored to the target π Hoffman et al. (2019); Noé et al. (2019); Gabrié et al. (2022). The computational cost of learning the proposal can be prohibitive and too slow for applications where accurate (and potentially unbiased) estimates need to be computed on the fly.

In this chapter we derive a novel *unbiased* importance sampling estimator of normalizing constants based on a well chosen deterministic and invertible transform $\mathrm{T} : \mathbb{R}^d \mapsto \mathbb{R}^d$ with iterates (or orbits) $\{\mathrm{T}^\ell(x)\}_{\ell=1}^K$, with $x \in \mathbb{R}^d$, that define a trajectory that explores the target distribution π. Our estimator combines these orbits initialized at some (not necessarily well adapted to π) initial proposal ρ and weights them in a principled way so that the overall estimator is unbiased.

Define $\rho_{\mathrm{T}} = K^{-1} \sum_{\ell=1}^K \mathrm{T}^\ell_\# \rho$ where $\mathrm{T}^\ell_\# \rho(x) = \rho(\mathrm{T}^{-\ell}(x)) |\mathbf{J}_{\mathrm{T}^{-\ell}}(x)|$. Our IS estimator is based on the following identity valid for any $A \in \mathcal{B}(\mathbb{R}^d)$

$$\int \mathbb{1}_A(x)\pi(\mathrm{d}x) = \frac{1}{K}\sum_{\ell=1}^K \int \mathbb{1}_A(\mathrm{T}^\ell(x)) \frac{\pi(\mathrm{T}^\ell(x))}{\rho_{\mathrm{T}}(\mathrm{T}^\ell(x))}\rho(\mathrm{d}x)\,,$$

and which straightforwardly defines an unbiased estimator by drawing M i.i.d. samples from ρ. If the trajectories of the transform are informed with the target π, e.g. through $\nabla \log \pi$, then this estimator may improve over the vanilla IS estimator with ρ as proposal since π is evaluated at the orbits. In fact, it is possible to define such a transform by taking inspiration from Hamiltonian Monte Carlo Neal et al. (2011). We examine the non-asymptotic properties of NEO-IS and its SNIS counterpart. Additionally, we leverage this estimator to develop a new geometrically ergodic MCMC sampler akin to iSIR, which can be directly compared to HMC. Through numerical experiments, we demonstrate that both NEO-IS and NEO-MCMC exhibit competitive performance when compared to several other well-known methods.

Chapter 3 / (Q2) - Entropic Mirror Monte Carlo

We continue our study of importance sampling by focusing in this chapter on building adaptive importance sampling methdods with improved exploration of the state space. Our AIS scheme is based on the iterates of the following mapping

$$\mathcal{F}_{\mathsf{em}}(\mu; \lambda, K_\pi, \varepsilon)(\mathrm{d}x) = \frac{\lambda}{\int \frac{\pi(y)^\varepsilon}{\mu(y)^\varepsilon}\mu(\mathrm{d}y)} \frac{\pi(x)^\varepsilon}{\mu(x)^\varepsilon}\mu(\mathrm{d}x) + \frac{1-\lambda}{\int \frac{\pi(y)^\varepsilon}{\mu(y)^\varepsilon}\mu K_\pi(\mathrm{d}y)} \frac{\pi(x)^\varepsilon}{\mu(x)^\varepsilon}\mu K_\pi(\mathrm{d}x)\,,$$

where K_π is a Markov transition kernel, $\mu_t K_\pi(\mathrm{d}x) = \int \mu_t(\mathrm{d}z) K_\pi(z, \mathrm{d}x)$ and $\lambda_t \in [0,1]$. Its iterates are defined recursively by

$$\mu_{t+1}(\mathrm{d}x) = \mathcal{F}_{\mathsf{em}}(\mu_t; \lambda_t, K_\pi, \varepsilon), \quad \text{where} \quad \lambda_t \in [0,1]\,. \tag{1.4.1}$$

If $\lambda_t = 1$ for all $t \in \mathbb{N}$ we recover the updates obtained by minimizing $\mu \mapsto \mathsf{KL}(\pi \parallel \mu)$ with entropic mirror descent Beck and Teboulle (2003); Dai et al. (2016); Korba and Portier (2022).

To see why (1.4.1) can be of interest, consider its particle approximation obtained by drawing N i.i.d. samples from μ_t (which we assume to be feasible for now),

$$\mu_{t+1}^N(\mathrm{d}x) = \lambda_t \sum_{i=1}^N \omega_t^i \delta_{X^i}(\mathrm{d}x) + (1-\lambda_t) \sum_{i=1}^N \varpi_t^i \delta_{Y^i}(\mathrm{d}x)\,, \tag{1.4.2}$$

where

$$\omega_t^i = \frac{\pi(X^i)^\varepsilon}{\mu_t(X^i)^\varepsilon} \Big/ \sum_{j=1}^N \frac{\pi(X^j)^\varepsilon}{\mu_t(X^j)^\varepsilon}\,, \quad \text{where} \quad X^{1:N} \overset{\text{iid}}{\sim} \mu_t\,, \tag{1.4.3}$$

$$\varpi_t^i = \frac{\pi(Y^i)^\varepsilon}{\mu_t(Y^i)^\varepsilon} \Big/ \sum_{j=1}^N \frac{\pi(Y^j)^\varepsilon}{\mu_t(Y^j)^\varepsilon}\,, \quad \text{where} \quad Y^{1:N} \sim K_\pi(X^1,\cdot) \otimes \ldots \otimes K_\pi(X^N,\cdot)\,. \tag{1.4.4}$$

The weight $\pi(X^i)^\varepsilon/\mu_t(X^i)^\varepsilon$ (1.4.3) is regularized and has less variance than the classical importance weight (Korba and Portier, 2022). The second normalized weight (1.4.4) is non-standard in the sense that μ_t in the denominator is evaluated in samples Y^i from $\mu_t K_\pi$ and not μ_t. If K_π is informed with the target π then the samples from $\mu_t K_\pi$ that are in the regions of the space where π is likely and μ_t is inadequate will have a larger weight. Thus, the second component of the mixture allows a greedy approach for exploring the target π if the involved Markov kernel K_π can make global moves.

The contributions of this chapter are the following. We start by showing that the introduced sequence is principled. In particular, we show that if the sequence $(\lambda_t)_{t\in\mathbb{N}}$ is chosen appropriately and the transition kernel K_π is π-invariant, i.e. $\pi K_\pi = \pi$, then the sequence $(\mu_t)_{t\in\mathbb{N}}$ enjoys a geometric convergence to π in backward KL divergence. If K_π is the unadjusted Langevin kernel, which is not π-invariant, we provide quantitative bounds in total variation distance under appropriate assumptions on π.

These results suggest that we may be able to build a suitable importance proposal for π by considering the scheme

$$\mu_{\theta_s} = \operatorname*{argmin}_{\mu\in\mathsf{F}_\Theta} \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu)\,, \tag{1.4.5}$$

where $\widetilde{\mathcal{R}}_\pi$ is some criterion and F_Θ is a parametric family of densities. We then introduce a novel criterion coined *skew Rényi projection* which guarantees that when $\widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_{\theta_s})$ is sufficiently small for all $s \in [1:T]$ (where T is the number of iterations), $\mathsf{KL}(\pi \parallel \mu_{\theta_T})$ is also small. To the best of our knowledge, this is not guaranteed if $\widetilde{\mathcal{R}}_\pi$ is the KL divergence. Finally, we devise a practical algorithm for solving (1.4.5) approximately by leveraging the particle approximations (1.4.2).

We provide preliminary numerical results. We demonstrate on highly challenging experiments that the final parametric approximation is close to what one would obtain by learning directly π via maximum likelihood. We also provide encouraging normalizing constants estimates obtained with our proposal using a moderate sample size relative to the dimension of the ambient space.

Chapter 4 / (Q3-4) - Variance estimation for SMC algorithms: a backward sampling approach.

The first contribution of this chapter is to propose a parameter free estimator of the asymptotic variance of the particle filter with multinomial resampling that trades computational cost for stability and reduced variance. The construction of our estimator starts from the observation

that the degeneracy of the current estimators Chan and Lai (2013); Lee and Whiteley (2018); Du and Guyader (2021) is similar to that of the vanilla particle smoother (1.3.12) and Particle Gibbs. In both cases, a backward sampling step which aims at diversifying the particle trajectories has shown to be a reliable workaround that decreases the (theoretical) variance of the estimators at the expense of higher computational cost. We thus aim at introducing such a mechanism in the estimation of the asymptotic variance. The construction of our estimator relies on the analysis conducted in Lee and Whiteley (2018) in which it is shown that the estimator of Chan and Lai (2013) can be interpreted as a conditional expectation with respect to the indices that retrace the genealogy of the particles, given all the particles and ancestors. We show that this construction still holds when the distribution of the indices relies on the backward importance weights. The resulting estimator is computed by averaging auxiliary statistics that are very similar to those of the *forward implementation* of the FFBS for additive functionals Del Moral et al. (2010c); Dubarry and Le Corff (2013) and can be thus updated online. The time complexity per update of our estimator is of order $\mathcal{O}(N^3)$. Driven by the efficient implementation of the FFBS for additive functionals developed in Olsson and Westerborn (2017), we show that the computational cost of our estimator can be reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ by means of additional Monte Carlo simulation while remaining as competitive in terms of bias and variance.

We next focus on the FFBS algorithm for the estimation of smoothed functionals (1.3.15). Despite the fact that a CLT has been obtained for estimators based on the FFBS Douc et al. (2011a), no variance estimator has been proposed in the literature. We show that our previous construction enables us to fill this gap and we thus provide a consistent estimator in the case of additive functionals. Again, this estimator can be computed online and in the particular case of marginal smoothing its computational cost can be drastically reduced. We validate our results with numerical experiments. Notably, we show empirically that our novel estimator for the filter has a favourable dependence on the time horizon t in comparison with the existing estimators.

Chapter 5 / (Q5) - State and parameter learning with the PaRIS particle Gibbs.

In this chapter, we consider the problem of parameter learning with stochastic gradient algorithms. We set the focus on learning the parameter of a function whose gradient is the smoothed expectation of an additive functional (1.3.15).

In this specific context, where a smoothing estimator is employed repeatedly to produce gradient estimates, controlling the bias and the MSE of the estimator becomes critical, see Karimi et al. (2019). This learning problem is usually tackled using either the Particle Gibbs Lindholm and Lindsten (2018), or classical smoothing algorithms such as the FFBSi or the PaRIS Olsson and Westerborn (2017). While the former has exponentially decreasing bias (w.r.t the number of iterates) under standard assumptions, it usually results in high variance and a non-negligible waste of the particle cloud generated. The latter is biased, since it is self-normalised but results in smaller variance than the particle Gibbs. Recently, zero bias estimators Jacob et al. (2020); Lee et al. (2020) have been proposed based on the coupling of the particle Gibbs that could be used in this framework, but they suffer from having a random computational complexity and high variance.

We propose a new algorithm combining the PaRIS and the PG algorithms. The conditional particle cloud resulting from the PG is now used not only to generate the next conditioning trajectory as in the usual PG but it is also used to generate a smoothing estimate, reducing waste of computational work.

This leads to a batch mode *PaRIS particle Gibbs* (`PPG`) *sampler*, which we furnish with an

upper bound on the bias that decreases inversely proportional to the number N of particles and exponentially fast with the particle Gibbs iteration index (under the assumption that the particle Gibbs sampler is uniformly ergodic), while keeping the MSE comparable to that of the underlying backward smoother. Furthermore, in the context of score ascent with the PPG we provide a non-asymptotic bound for the expectation of the squared gradient in terms of bias and MSE of the PPG. This bound establishes an $\mathcal{O}(\log(n)/\sqrt{n})$ convergence of the learning procedure.

Chapter 6 / (Q6) - Monte Carlo guided Diffusion for Bayesian linear inverse problems.

In this chapter we consider the problem of sampling from the posterior of a diffusion model given by

$$\phi_0^y(\mathrm{d}x_0) \propto g_0^y(x_0)\mathsf{p}_0^{\theta_*}(\mathrm{d}x_0)\,. \tag{1.4.6}$$

where $\mathsf{p}_0^{\theta_*}$ is the time 0 marginal of (1.3.27) and θ_* is the approximate solution of (1.3.29). We focus on potentials g_0^y that are given by a linear Gaussian inverse problem

$$Y = AX + \sigma Z, \quad \text{where} \quad A \in \mathbb{R}^{d_y \times d_x}, \quad Z \sim \mathcal{N}(\mathbf{0}_{d_y}, \mathbf{I}_{d_y}), \quad \sigma \geq 0\,.$$

Current methods Song et al. (2021a); Kawar et al. (2022); Lugmayr et al. (2022); Chung et al. (2023) aiming to sample from (1.4.6), including those that rely on approximations of the conditional score (1.3.31), introduce an irreducible bias rendering them unreliable for critical applications such as medical imaging. Our goal in this chapter is to devise a sequential Monte Carlo sampler that returns a consistent particle approximation of (1.4.6), ensuring that asymptotically we sample from the target posterior. For this purpose we introduce a sequence of distributions $\{g_s^y\}_{s=1}^n$ given by

$$\phi_s^y(\mathrm{d}x_s) \propto g_s^y(x_s)\mathsf{p}_s^{\theta_*}(\mathrm{d}x_s)\,,$$

and with potentials $\{g_s^y\}_{s=1}^n$ chosen so as to ensure that the discrepancy between ϕ_s^y and ϕ_{s+1}^y is small for all $s \in [0:n-1]$. Our first contribution is a specific choice of potential in the case where the inverse problem is "noiseless", i.e. $\sigma = 0$, that is simple to compute. We then show that more generally, a "noisy" Gaussian linear inverse problem with $\sigma > 0$ can be seen as a noiseless inverse problem on the extended states with prior $\mathsf{p}_{0:n}^{\theta_*}$. In both cases we devise a SMC sampler that targets the posterior and we furnish it with a non-asymptotic bound on the KL divergence between the target posterior and the *expected* particle approximation.

To evaluate the performance of our algorithms, we perform numerical simulations on several examples (in high-dimension) for which the target posterior distribution is known. Simulation results support our theoretical results, i.e. the empirical distribution of samples from our algorithms converge to the target posterior distributions. This is **not** the case for the competing methods (using the same denoising diffusion generative priors) which are shown, when run with random initialization of the denoising diffusion, to generate a significant number of samples outside the support of the target posterior.

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Agakov, F. V. and Barber, D. (2004). An auxiliary variational method. In *International Conference on Neural Information Processing*, pages 561–566. Springer.

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Akyildiz, Ö. D. and Míguez, J. (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(2):1–17.

Anderson, G. D. and Qiu, S.-L. (1997). A monotonicity property of the gamma function. *Proc. Amer. Math. Soc.*, 125(11):3355–3362.

Andrieu, C. and Doucet, A. (2003). Online Expectation–Maximization type algorithms for parameter estimation in general state space models. volume 6, pages 69–72.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(3):269–342.

Andrieu, C., Lee, A., and Vihola, M. (2018a). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842 – 872.

Andrieu, C., Lee, A., and Vihola, M. (2018b). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842 – 872.

Arbel, M., Matthews, A., and Doucet, A. (2021). Annealed flow transport monte carlo. In *International Conference on Machine Learning*, pages 318–330. PMLR.

Arjomand Bigdeli, S., Zwicker, M., Favaro, P., and Jin, M. (2017). Deep mean-shift priors for image restoration. *Advances in Neural Information Processing Systems*, 30.

Bakry, D., Gentil, I., Ledoux, M., et al. (2014). *Analysis and geometry of Markov diffusion operators*, volume 103. Springer.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Benton, J., Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2022). From denoising diffusions to denoising markov models. *arXiv preprint arXiv:2211.03595*.

Besag, J. (1994). Comments on "Representations of knowledge in complex systems" by U. Grenander and M. Miller. *J. Roy. Statist. Soc. Ser. B*, 56:591–592.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Beskos, A., Jasra, A., Kantas, N., and Thiery, A. (2016). On the convergence of adaptive sequential monte carlo methods.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(1):973–978.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bogachev, V. I. and Ruas, M. A. S. (2007). *Measure theory*, volume 1. Springer.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79.

Bunch, P. and Godsill, S. (2012). Improved particle approximations to the joint smoothing distribution using markov chain monte carlo. *IEEE Transactions on Signal Processing*, 61(4):956–963.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *The 4th International Conference on Learning Representations (ICLR)*.

Cappé, O. (2001). Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods Appl.*, 7(1–2):81–92.

Cappé, O. (2011). Online EM algorithm for hidden Markov models. *J. Comput. Graph. Statist.*, 20(3):728–749.

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.

Cappé, O., Godsill, S. J., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *IEEE Proceedings*, 95(5):899–924.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Cappé, O. and Moulines, E. (2005). On the use of particle filtering for maximum likelihood parameter estimation. In *European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey.

Cappe, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Carbone, D., Hua, M., Coste, S., and Vanden-Eijnden, E. (2023). Efficient training of energy-based models using jarzynski equality. *arXiv preprint arXiv:2305.19414*.

Cardoso, G., El Idrissi, Y. J., Le Corff, S., Moulines, É., and Olsson, J. (2023a). State and parameter learning with paris particle gibbs. In *International Conference on Machine Learning*, pages 3625–3675. PMLR.

Cardoso, G., Idrissi, Y. J. E., Corff, S. L., and Moulines, E. (2023b). Monte carlo guided diffusion for bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*.

Cardoso, G., Samsonov, S., Thin, A., Moulines, E., and Olsson, J. (2022). BR-SNIS: Bias reduced self-normalized importance sampling. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Chan, H. P. and Lai, T. L. (2013). A general theory of particle filters in hidden Markov models and some applications. *Ann. Statist.*, 41(6):2877 – 2904.

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135.

Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y., and Bengio, Y. (2020). Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint arXiv:2003.06060*.

Cheney, M. and Borden, B. (2009). *Fundamentals of radar imaging*. SIAM.

Chewi, S., Erdogdu, M. A., Li, M., Shen, R., and Zhang, S. (2022). Analysis of langevin monte carlo from poincare to log-sobolev. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1–2. PMLR.

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Ann. Statist.*, 32(6):2385–2411.

Chopin, N. and Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer International Publishing.

Chopin, N., Papaspiliopoulos, O., et al. (2020). *An introduction to sequential Monte Carlo*, volume 4. Springer.

Chopin, N. and Robert, C. P. (2010). Properties of nested sampling. *Biometrika*, 97(3):741–755.

Chopin, N. and Singh, S. S. (2015a). On particle Gibbs sampling. *Bernoulli*, 21(3):1855 – 1883.

Chopin, N. and Singh, S. S. (2015b). On particle gibbs sampling. *Bernoulli*, 21(3):1855–1883.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. (2023). Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*.

Cornish, R., Caterini, A., Deligiannidis, G., and Doucet, A. (2020). Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pages 2133–2143. PMLR.

Cornuet, J.-M., Marin, J.-M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

Craiu, R. V. and Lemieux, C. (2007). Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Statistics and Computing*, 17(2):109.

Cremer, C., Morris, Q., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*.

Crisan, D., Del Moral, P., and Lyons, T. (1998). *Discrete filtering using branching and interacting particle systems*. Université de Toulouse. Laboratoire de Statistique et Probabilités [LSP].

Cérou, F., Del Moral, P., and Guyader, A. (2011). A nonasymptotic theorem for unnormalized Feynman–Kac particle models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(3):629 – 649.

Dai, B., He, N., Dai, H., and Song, L. (2016). Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pages 985–994. PMLR.

Dashti, M. and Stuart, A. M. (2017). The bayesian approach to inverse problems. In *Handbook of uncertainty quantification*, pages 311–428. Springer.

Dau, H.-D. and Chopin, N. (2022). On the complexity of backward smoothing algorithms. *arXiv preprint arXiv:2207.00976*.

Daudel, K., Douc, R., and Portier, F. (2021a). Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250–2270.

Daudel, K., Douc, R., and Roueff, F. (2021b). Monotonic alpha-divergence minimisation. *arXiv preprint arXiv:2103.05684*.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67.

Del Moral, P. (2004). Feynman-kac formulae. In *Feynman-Kac Formulae*, pages 47–93. Springer.

Del Moral, P. (2013). *Mean Field Simulation for Monte Carlo Integration*. CRC Press.

Del Moral, P., Doucet, A., and Jasra, A. (2006a). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Del Moral, P., Doucet, A., and Jasra, A. (2006b). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B*, 68(3):411–436.

Del Moral, P., Doucet, A., and Singh, S. (2010a). Forward smoothing using sequential monte carlo. *arXiv preprint arXiv:1012.5390*.

Del Moral, P., Doucet, A., and Singh, S. S. (2010b). A backward particle interpretation of feynman-kac formulae. *ESAIM: Math. Model. Num. Analysis*, 44(5):947–975.

Del Moral, P., Doucet, A., and Sumeetpal, S. (2010c). Forward smoothing using sequential monte carlo. *ArXiv:1012.5390*.

Del Moral, P. and Guionnet, A. (1998). Large deviations for interacting particle systems: applications to non-linear filtering. *Stochastic processes and their applications*, 78(1):69–95.

Del Moral, P. and Guionnet, A. (1999). Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab.*, pages 275–297.

Del Moral, P. and Guionnet, A. (2001). On the stability of interacting processes with applications to filtering and genetic algorithms. In *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, volume 37, pages 155–194. Elsevier.

Del Moral, P. and Jasra, A. (2018). A sharp first order analysis of Feynman–Kac particle models, part II: Particle Gibbs samplers. 128(1):354–371.

Del Moral, P., Kohn, R., and Patras, F. (2016). On particle Gibbs samplers. 52(4):1687–1733.

Delyon, B. and Portier, F. (2021). Safe adaptive importance sampling: A mixture approach. *The Annals of Statistics*, 49(2):885 – 917.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J. H., and Vehtari, A. (2021). Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems*.

Ding, X. and Freedman, D. J. (2019). Learning deep generative models with annealed importance sampling. *arXiv preprint arXiv:1906.04904*.

Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE.

Douc, R., Garivier, A., Moulines, E., and Olsson, J. (2011a). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.*, 21(6):2109 – 2145.

Douc, R., Garivier, A., Moulines, E., Olsson, J., et al. (2011b). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Annals of Applied Probability*, 21(6):2109–2145.

Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2007a). Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35(1):420–448.

Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2007b). Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics*, 11:427–447.

Douc, R. and Moulines, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, 36(5):2344–2376.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018a). *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018b). *Markov chains*. Springer.

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear time series: Theory, methods and applications with R examples*. CRC press.

Doucet, A., De Freitas, N., Gordon, N. J., et al. (2001). *Sequential Monte Carlo methods in practice*, volume 1. Springer.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Stat. Comput.*, 10(3):197–208.

Du, Q. and Guyader, A. (2021). Variance estimation in adaptive sequential Monte Carlo. *Ann. Appl. Probab.*, 31(3):1021 – 1060.

Dubarry, C. and Le Corff, S. (2013). Non-asymptotic deviation inequalities for smoothed additive functionals in nonlinear state-space models. *Bernoulli*, 19(5B):2222 – 2249.

Durmus, A. and Moulines, É. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854 – 2882.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850.

Elbakri, I. A. and Fessler, J. A. (2002). Statistical image reconstruction for polyenergetic x-ray computed tomography. *IEEE transactions on medical imaging*, 21(2):89–99.

Elvira, V. and Martino, L. (2021). Advances in importance sampling. *arXiv preprint arXiv:2102.05407*.

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2017). Improving population monte carlo: Alternative weighting and resampling schemes. *Signal Processing*, 131:77–91.

Erdogdu, M. A., Hosseinzadeh, R., and Zhang, S. (2022). Convergence of langevin monte carlo in chi-squared and rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 8151–8175. PMLR.

Fearnhead, P., Wyncoll, D., and Tawn, J. (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika*, 97(2):447–464.

Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., and Freeman, W. T. (2006). Removing camera shake from a single photograph. In *Acm Siggraph 2006 Papers*, pages 787–794.

Figueiredo, M. A., Bioucas-Dias, J. M., and Nowak, R. D. (2007). Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image processing*, 16(12):2980–2991.

Fort, G., Moulines, E., and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, 39(6):3262 – 3289.

Franca, G., Sulam, J., Robinson, D. P., and Vidal, R. (2020). Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124008.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2022). Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119.

Geffner, T. and Domke, J. (2021). On the difficulty of unbiased alpha divergence minimization. In *International Conference on Machine Learning*, pages 3650–3659. PMLR.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gloaguen, P., Le Corff, S., and Olsson, J. (2022). A pseudo-marginal sequential monte carlo online smoothing algorithm. *Bernoulli*, 28(4):2606–2633.

Godsill, S. J., Doucet, A., and West, M. (2004). Monte carlo smoothing for nonlinear time series. *J. Amer. Statist. Assoc.*, 99(465):156–168.

Goertzel, G. (1949). Quota sampling and importance functions in stochastic solution of particle problems. Technical report.

González, R. C., Woods, R. E., and Masters, B. R. (2009). Digital image processing, third edition. *Journal of Biomedical Optics*, 14:029901.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993a). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F*, volume 140, pages 107–113. IET.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993b). Novel approach to nonlinear/ non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140(2):107–113.

Grenioux, L., Durmus, A., Moulines, É., and Gabrié, M. (2023). On sampling with approximate transport maps. *arXiv preprint arXiv:2302.04763*.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Handschin, J. E. and Mayne, D. Q. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International journal of control*, 9(5):547–559.

Hartman, P. (1982). *Ordinary Differential Equations: Second Edition*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Ho, J., Salimans, T., Gritsenko, A. A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*.

Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. (2019). Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv preprint arXiv:1903.03704*.

Hoffman, M. D. (2017). Learning deep latent gaussian models with markov chain monte carlo. In *International conference on machine learning*, pages 1510–1519. PMLR.

Holley, R. and Stroock, D. W. (1986). Logarithmic sobolev inequalities and stochastic ising models.

Huggins, J. H. and Roy, D. M. (2019). Sequential Monte Carlo as approximate sampling: bounds, adaptive resampling via ∞-ESS, and an application to particle Gibbs. *Bernoulli*, 25(1):584 – 622.

Idier, J. (2013). *Bayesian approach to inverse problems*. John Wiley & Sons.

Ivanov, O., Figurnov, M., and Vetrov, D. (2018). Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Janati, Y., Le Corff, S., and Petetin, Y. (2023). Variance estimation for sequential monte carlo algorithms: a backward sampling approach. *To appear, Bernoulli*.

Jia, H. and Seljak, U. (2020). Normalizing constant estimation with Gaussianized bridge sampling. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–14. PMLR.

Kahn, H. (1949). Stochastic (monte carlo) attenuation analysis. Technical report, RAND CORP SANTA MONICA CALIF.

Kaipio, J. P., Kolehmainen, V., Somersalo, E., and Vauhkonen, M. (2000). Statistical inversion and monte carlo sampling methods in electrical impedance tomography. *Inverse problems*, 16(5):1487.

Kaltenbach, S., Perdikaris, P., and Koutsourelakis, P.-S. (2023). Semi-supervised invertible neural operators for bayesian inverse problems. *Computational Mechanics*, pages 1–20.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statist. Sci.*, 30(3):328–351.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019). Non-asymptotic analysis of biased stochastic approximation scheme. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1944–1974. PMLR.

Kawar, B., Elad, M., Ermon, S., and Song, J. (2022). Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606.

Kim, K., Oh, J., Gardner, J., Dieng, A. B., and Kim, H. (2022). Markov chain score ascent: A unifying framework of variational inference with markovian gradients. *Advances in Neural Information Processing Systems*, 35:34802–34816.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Kitagawa, G. (1993). A monte carlo filtering and smoothing method for non-gaussian nonlinear state space models. In *Proceedings of the 2nd US-Japan joint seminar on statistical time series analysis*, volume 2, pages 110–131.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979.

Korba, A. and Portier, F. (2022). Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR.

Koskela, J., Jenkins, P. A., Johansen, A. M., and Spano, D. (2020). Asymptotic genealogies of interacting particle systems with an application to sequential monte carlo. *Ann. Statist.*, 48(1):560–583.

Künsch, H. R. (2005). Recursive monte carlo filters: algorithms and theoretical analysis. *Ann. Statist.*, 33(5):1983–2021.

Lawson, D., Tucker, G., Dai, B., and Ranganath, R. (2019). Energy-inspired models: Learning with sampler-induced distributions. *arXiv preprint arXiv:1910.14265*.

Le Corff, S. and Fort, G. (2013). Online expectation maximization based algorithms for inference in hidden markov models. *Electronic Journal of Statistics*, 7:763–792.

Lee, A., Singh, S. S., and Vihola, M. (2020). Coupled conditional backward sampling particle filter. *Ann. Statist.*, 48(5):3066–3089.

Lee, A. and Whiteley, N. (2018). Variance estimation in the particle filter. *Biometrika*, 105(3):609–625.

Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. (2018). Generalizing Hamiltonian Monte Carlo with neural networks. In *International Conference on Learning Representations*.

Lindholm, A. and Lindsten, F. (2018). Learning dynamical systems with particle stochastic approximation em.

Lindsten, F., Douc, R., and Moulines, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797.

Lindsten, F., Jordan, M. I., and Schön, T. B. (2014). Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.*, 15(1):2145–2184.

Lindsten, F. and Schön, T. B. (2012). On the use of backward simulation in the particle gibbs sampler. In *2012 IEEE ICASSP*, pages 3845–3848.

Liu, J. and West, M. (2001). Sequential monte carlo methods in practice. *Statistics for Engineering and Information Science, edited by A. Doucet, N. Freitas, and N. Gordon (Springer, New York, 2001)*, pages 225–246.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. (2018). Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*.

Marin, J.-M., Pudlo, P., and Sedki, M. (2019). Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, 25(3):1977 – 1998.

Marnissi, Y., Zheng, Y., Chouzenoux, E., and Pesquet, J.-C. (2017). A variational bayesian approach for image restoration—application to image deblurring with poisson–gaussian noise. *IEEE Transactions on Computational Imaging*, 3(4):722–737.

Mastrototaro, A. and Olsson, J. (2023). Adaptive online variance estimation in particle filters: the alvar estimator. *Statistics and Computing*, 33(4):1–26.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. (2019a). Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5):1–19.

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. (2019b). Neural importance sampling. *ACM Transactions on Graphics*, 38(145).

Naesseth, C., Lindsten, F., and Blei, D. (2020). Markovian score climbing: Variational inference with kl (p|| q). *Advances in Neural Information Processing Systems*, 33:15499–15510.

Neal, R. M. (2001a). Annealed importance sampling. *Statistics and computing*, 11:125–139.

Neal, R. M. (2001b). Annealed importance sampling. *Statistics and Computing*, 11:125–139.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Neddermeyer, J. C. (2009). Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association*, 104(486):788–802.

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147.

Oh, M.-S. and Berger, J. O. (1993). Integration of multimodal functions by monte carlo importance sampling. *Journal of the American Statistical Association*, 88(422):450–456.

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential monte carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179.

Olsson, J. and Douc, R. (2019). Numerically stable online estimation of variance in particle filters. *Bernoulli*, 25(2):1504 – 1535.

Olsson, J. and Westerborn, J. (2017). Efficient particle-based online smoothing in general hidden Markov models: The PaRIS algorithm. *Bernoulli*, 23(3):1951 – 1996.

Oudjane, N. and Rubenthaler, S. (2005). Stability and uniform particle approximation of nonlinear filters in case of non ergodic signals. *Stochastic Analysis and applications*, 23(3):421–448.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680.

Peng, J., Liu, D., Xu, S., and Li, H. (2021). Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784.

Petetin, Y., Janati, Y., and Desbouvries, F. (2021). Structured variational bayesian inference for gaussian state-space models with regime switching. *IEEE Signal Processing Letters*, 28:1953–1957.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.*, 94(446):590–599.

Portier, F. and Delyon, B. (2018). Asymptotic optimality of adaptive importance sampling. *Advances in Neural Information Processing Systems*, 31.

Poyiadjis, G., Doucet, A., and Singh, S. S. (2005). Particle methods for optimal filter derivative: application to parameter estimation. pages v/925–v/928.

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

Prangle, D. (2019). Distilling importance sampling. *arXiv preprint arXiv:1910.03632*.

Prangle, D. and Viscardi, C. (2023). Distilling importance sampling for likelihood free inference. *Journal of Computational and Graphical Statistics*, pages 1–11.

Qiu, Y., Zhang, L., and Wang, X. (2020). Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rippel, O. and Adams, R. P. (2013). High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Rotskoff, G. and Vanden-Eijnden, E. (2019). Dynamical computation of the density of states and Bayes factors using nonequilibrium importance sampling. *Physical Review Letters*, 122(15):150602.

Royden, H. L. and Fitzpatrick, P. (1988). *Real analysis*, volume 32. Macmillan New York.

Rubin, D. B. (1987). Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):542–543.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and computing in applied probability*, 1:127–190.

Ruiz, F. J., Titsias, M. K., Cemgil, T., and Doucet, A. (2021). Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. In *Uncertainty in Artificial Intelligence*.

Sahlström, T. and Tarvainen, T. (2023). Utilizing variational autoencoders in the bayesian inverse problem of photoacoustic tomography. *SIAM Journal on Imaging Sciences*, 16(1):89–110.

Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *International conference on machine learning*, pages 1218–1226. PMLR.

Samsonov, S., Lagutin, E., Gabrié, M., Durmus, A., Naumov, A., and Moulines, E. (2022). Local-global mcmc kernels: the best of both worlds. *Advances in Neural Information Processing Systems*, 35:5178–5193.

Särkkä, S. (2013). *Bayesian filtering and smoothing*. Number 3. Cambridge university press.

Shestopaloff, A. Y., Neal, R. M., et al. (2018). Sampling latent states for high-dimensional non-linear state space models with the embedded hmm method. *Bayesian Analysis*, 13(3):797–822.

Shin, H. and Choi, M. (2023). Physics-informed variational inference for uncertainty quantification of stochastic differential equations. *Journal of Computational Physics*, page 112183.

Singh, S. S., Lindsten, F., and Moulines, E. (2017). Blocking strategies and stability of particle gibbs samplers. *Biometrika*, 104(4):953–969.

Skare, Ø., Bølviken, E., and Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859.

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88.

So, M. K. (2006). Bayesian analysis of nonlinear and non-Gaussian state space models via multiple-try sampling methods. *Statistics and Computing*, 16(2):125–141.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

Song, J., Meng, C., and Ermon, S. (2021a). Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021b). Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428.

Song, Y., Shen, L., Xing, L., and Ermon, S. (2022). Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021c). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Stuart, A. M. (2010). Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559.

Su, J., Xu, B., and Yin, H. (2022). A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65.

Tabak, E. G., Vanden-Eijnden, E., et al. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233.

Taghavi, E., Lindsten, F., Svensson, L., and Schön, T. B. (2013). Adaptive stopping for fast particle smoothing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6293–6297. IEEE.

Tanizaki, H. and Mariano, R. (1994). Prediction, filtering and smoothing in non-linear and non-normal cases using monte carlo integration. *J. Appl. Econometrics*, 9(2):163–79.

Thin, A., Janati El Idrissi, Y., Le Corff, S., Ollion, C., Moulines, E., Doucet, A., Durmus, A., and Robert, C. X. (2021). Neo: non equilibrium sampling on the orbits of a deterministic transform. *Advances in Neural Information Processing Systems*, 34:17060–17071.

Thin, A., Kotelevskii, N., Denain, J.-S., Grinsztajn, L., Durmus, A., Panov, M., and Moulines, E. (2020). Metflow: A new efficient method for bridging the gap between markov chain monte carlo and variational inference. *arXiv preprint arXiv:2002.12253*.

Tieleman, T. (2008). Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. S. (2023). Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*.

Turner, R., Hung, J., Frank, E., Saatchi, Y., and Yosinski, J. (2019). Metropolis–Hastings generative adversarial networks. In *International Conference on Machine Learning*, pages 6345–6353. PMLR.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Van Handel, R. (2009). Uniform time average consistency of monte carlo particle filters. *Stochastic Processes and their Applications*, 119(11):3835–3861.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32.

Vlaardingerbroek, M. T. and Boer, J. A. (2013). *Magnetic resonance imaging: theory and practice*. Springer Science & Business Media.

Wan, Z., Zhang, J., Chen, D., and Liao, J. (2021). High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701.

Wei, X., van Gorp, H., Gonzalez-Carabarin, L., Freedman, D., Eldar, Y. C., and van Sloun, R. J. (2022). Deep unfolding with normalizing flow priors for inverse problems. *IEEE Transactions on Signal Processing*, 70:2962–2971.

Whiteley, N. (2010). Discussion on particle markov chain monte carlo methods. pages 306–307.

Whiteley, N. (2013). Stability properties of some particle filters. *The Annals of Applied Probability*, 23(6):2500 – 2537.

Wirnsberger, P., Ballard, A. J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., Jimenez Rezende, D., and Blundell, C. (2020). Targeted free energy estimation via learned mappings. *The Journal of Chemical Physics*, 153(14):144112.

Wolf, C., Karl, M., and van der Smagt, P. (2016). Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*.

Wu, H., Köhler, J., and Noe, F. (2020). Stochastic normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. (2016). On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*.

Xiang, H., Zou, Q., Nawaz, M. A., Huang, X., Zhang, F., and Yu, H. (2023). Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046.

Yeh, R. A., Lim, T. Y., Chen, C., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. (2018). Image restoration with deep generative models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6772–6776. IEEE.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514.

Zeng, Y., Fu, J., Chao, H., and Guo, B. (2022). Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, G., Ji, J., Zhang, Y., Yu, M., Jaakkola, T., and Chang, S. (2023). Towards coherent image inpainting using denoising diffusion implicit models. *arXiv preprint arXiv:2304.03322*.

Zhao, Y., Nassar, J., Jordan, I., Bugallo, M., and Park, I. M. (2022). Streaming variational monte carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1150–1161.

Zheng, C., Cham, T.-J., and Cai, J. (2019). Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447.

Zhihang, X., Yingzhi, X., and Qifeng, L. (2023). A domain-decomposed vae method for bayesian inverse problems. *arXiv preprint arXiv:2301.05708*.

Chapter 2

NEO: Non-equilibrium sampling on the orbit of a deterministic transform

2.1 Introduction

Consider a target distribution of the form $\pi(x) \propto \rho(x)\mathrm{L}(x)$ where ρ is a probability density function (pdf) on $\mathbb{R}^d$ and L is a nonnegative function. Typically, in a Bayesian setting, π is a posterior distribution associated with a prior distribution ρ and a likelihood function L. Another situation of interest is generative modeling where π is the distribution implicitly defined by a Generative Adversarial Networks (GAN) discriminator-generator pair where ρ is the distribution of the generator and L is derived from the discriminator Turner et al. (2019); Che et al. (2020). In a Variational Auto Encoder (VAE) context Kingma and Welling (2014); Burda et al. (2016), π could be the true posterior distribution, ρ the approximate posterior distribution output by the encoder and L an importance weight between the true posterior and approximate posterior distributions. We are interested in this chapter in sampling from π and approximating its intractable normalizing constant $\mathcal{Z} = \int \rho(x)\mathrm{L}(x)\mathrm{d}x$. These problems arise in many applications in statistics, molecular dynamics or machine learning, and remain challenging.

Many approaches to compute normalizing constants are based on Importance Sampling (IS) - see Agapiou et al. (2017); Akyildiz and Míguez (2021) and the references therein - and its variations, among others, Annealed Importance Sampling (AIS) Neal (2001b); Wu et al. (2016); Ding and Freedman (2019) and Sequential Monte Carlo (SMC) Del Moral et al. (2006b). More recently, Neural IS has also become very popular in machine learning; see e.g. El Moselhy and Marzouk (2012); Müller et al. (2019b); Papamakarios et al. (2019); Prangle (2019); Wirnsberger et al. (2020); Wu et al. (2020). Neural IS is an adaptive IS which relies on an importance function obtained by applying a normalizing flow to a reference distribution. The parameters of this normalizing flow are chosen by minimizing a divergence between the proposal and the target (such as the Kullback–Leibler Müller et al. (2019b) or the χ^2-divergence Agapiou et al. (2017)). Recent work on the subject proposes to add stochastic moves in order to enhance the performance of the normalizing flows Wu et al. (2020).

More recently, the *Non-Equilibrium IS* (NEIS) method has been introduced by Rotskoff and Vanden-Eijnden (2019) as an alternative to these approaches. Similar to Neural IS, NEIS consists in transporting samples $\{X^i\}_{i=1}^N$ from a reference distribution using a family of deterministic mappings. For NEIS, this family is chosen to be an homogeneous differential flow $(\phi_t)_{t\in\mathbb{R}}$. In

contrast to Neural IS, for any $i \in [N]$, the sample X^i is propagated both forward and backward in time along the orbits associated with $(\phi_t)_{t\in\mathbb{R}}$ until stopping conditions are met. Moreover, the resulting estimator of the normalizing constant is obtained by computing weighted averages of the whole orbit $(\phi_t(X^i))_{t\in[\tau_{+,i},\tau_{-,i}]}$, where $\tau_{+,i}, \tau_{-,i}$ are the resulting stopping times, and not only the endpoints $\phi_{\tau_{+,i}}(X^i), \phi_{\tau_{-,i}}(X^i)$. In Rotskoff and Vanden-Eijnden (2019), the authors provide an application of NEIS with $(\phi_t)_{t\in\mathbb{R}}$ associated with a conformal Hamiltonian dynamics, and reports impressive numerical results on difficult normalizing constants estimation problems, in particular for high-dimensional multimodal distributions.

We propose in this work NEO-IS which alleviates the shortcomings of NEIS. Similar to NEIS, samples are drawn from a reference distribution, typically set to ρ, and are propagated under the forward and backward orbits of a *discrete-time* dynamical system associated with an invertible transform T. An estimator of the normalizing constant is obtained by reweighting all the points on the whole orbits using the IS rule. Contrary to NEIS, the NEO-IS estimator of Z is unbiased under assumptions that are mild and easy to verify. It is more flexible than NEIS because it does not rely on the accuracy of the discretization of a continuous-time dynamical system.

We then show how it is possible to leverage the unbiased estimator of Z defined by NEO-IS to obtain NEO-MCMC, a novel massively parallel MCMC algorithm to sample from π. In a nutshell, NEO-MCMC relies on parallel walkers which each estimates the normalizing constant but are allowed to interact through a resampling mechanism.

Our contributions can be summarized as follows.

- **(i)** We present a novel class of IS estimators of the normalizing constant $\mathcal{Z}$ referred to as NEO-IS. More broadly, a small modification of this algorithm also allows us to estimate integrals with respect to π. Both finite sample and asymptotic guarantees are provided for these two methodologies.
- **(ii)** We develop a new massively parallel MCMC method, NEO-MCMC. NEO-MCMC combines NEO-IS unbiased estimator of the normalizing constant with iterated sampling-importance resampling methods. We prove that it is π-reversible and ergodic under very general conditions. We derive also conditions which imply that NEO-MCMC is uniformly geometrically ergodic (with an explicit expression of the mixing time).
- **(iii)** We illustrate our findings using numerical benchmarks which show that both NEO-IS and NEO-MCMC outperform state-of-the-art (SOTA) methods in difficult settings.

2.2 NEO-IS algorithm

In this section, we derive the NEO-IS algorithm. The two key ingredients for this algorithm are (1) the reference distribution ρ and (2) a transformation T assumed to be a C^1-diffeomorphism with inverse T^{-1}. Write, for $k \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$, $\mathrm{T}^k = \mathrm{T} \circ \mathrm{T}^{k-1}$, $\mathrm{T}^0 = \mathrm{Id}_d$ and similarly $\mathrm{T}^{-k} = \mathrm{T}^{-1} \circ \mathrm{T}^{-(k-1)}$. For any $k \in \mathbb{Z}$, denote by $\rho_k : \mathbb{R}^d \to \mathbb{R}_+$ the pushforward of ρ by T^k, defined for $x \in \mathbb{R}^d$ by $\rho_k(x) = \rho(\mathrm{T}^{-k}(x))\mathbf{J}_{\mathrm{T}^{-k}}(x)$, where $\mathbf{J}_\Phi(x) \in \mathbb{R}^+$ is the absolute value of the Jacobian determinant of $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ evaluated at x. In line with multiple importance sampling *à la* Owen and Zhou Owen and Zhou (2000), we introduce the proposal density

$$\rho_{\mathrm{T}}(x) = \Omega^{-1} \sum\nolimits_{k\in\mathbb{Z}} \varpi_k \rho_k(x)\,, \tag{2.2.1}$$

where $\{\varpi_k\}_{k\in\mathbb{Z}}$ is a nonnegative sequence and $\Omega = \sum_{k\in\mathbb{Z}} \varpi_k$. Note that we assume in the sequel that the support of the weight sequence defined as $\{k \in \mathbb{Z} : \varpi_k \neq 0\}$ is finite. Thus, the mixture distribution in (2.2.1) is a **finite mixture**. Given $x \in \mathbb{R}^d$, $\rho_{\mathrm{T}}(x)$ is a function of the forward and backward orbit of T through x.

For any nonnegative function f, the definition of ρ_{T} implies that

$$\int f(y)\rho_{\mathrm{T}}(y)\mathrm{d}y = \Omega^{-1}\int \sum_{k\in\mathbb{Z}} \varpi_k f(\mathrm{T}^k(x))\rho(x)\mathrm{d}x\,.$$

Assuming that $\varpi_0 > 0$, the ratio $\rho(x)/\rho_{\mathrm{T}}(x) \leq \varpi_0^{-1}\Omega < \infty$ is bounded. We can therefore apply the IS principle which allows to write the identity

$$\boxed{\int f(x)\rho(x)\mathrm{d}x = \int \left(f(y)\frac{\rho(y)}{\rho_{\mathrm{T}}(y)}\right)\rho_{\mathrm{T}}(y)\mathrm{d}y = \int \sum_{k\in\mathbb{Z}} f(\mathrm{T}^k(x))w_k(x)\rho(x)\mathrm{d}x\,,} \tag{2.2.2}$$

where the weights are given by (see Section A.1.2 for a detailed derivation),

$$w_k(x) = \varpi_k\rho(\mathrm{T}^k(x))/\{\Omega\rho_{\mathrm{T}}(\mathrm{T}^k(x))\} = \varpi_k\rho_{-k}(x)\Big/\sum\nolimits_{i\in\mathbb{Z}}\varpi_{k+i}\rho_i(x)\ . \tag{2.2.3}$$

We assume in the sequel that $\varpi_0 > 0$. In particular, note that under this condition, the weights w_k are also upper bounded uniformly in x: for any $x \in \mathbb{R}^d$, $w_k(x) \leq \varpi_k/\varpi_0$. Equations (2.2.2) and (2.2.3) suggest to estimate the integral $\int f(x)\rho(x)\mathrm{d}x$ by

$$I^{\mathrm{NEO}}_{\varpi,N}(f) = N^{-1}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)f(\mathrm{T}^k(X^i))\,,$$

where $\{X^i\}_{i=1}^N$ are i.i.d. samples from the proposal ρ, which is denoted by $X^{1:N} \overset{\mathrm{iid}}{\sim} \rho$.

Algorithm 1 NEO-IS Sampler

1. Sample $X^{1:N} \overset{\mathrm{iid}}{\sim} \rho$ for $i \in [N]$.
2. For $i \in [N]$, compute the path $(\mathrm{T}^j(X^i))_{j\in\mathbb{Z}}$ and weights $(w_j(X^i))_{j\in\mathbb{Z}}$.
3. $I^{\mathrm{NEO}}_{\varpi,N}(f) = N^{-1}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)f(\mathrm{T}^k(X^i))$.

This estimator is obtained by a weighted combination of the elements of the independent forward and backward orbits $\{\mathrm{T}^k(X^i)\}_{k\in\mathbb{Z}}$ with $X^{1:N} \overset{\mathrm{iid}}{\sim} \rho$. This estimator is referred to as NEO-IS. Choosing $f \equiv \mathrm{L}$ provides the NEO-IS estimator of the normalizing constant of π:

$$\widehat{\mathcal{Z}}^{\varpi}_{X^i} = \textstyle\sum_{k\in\mathbb{Z}} \mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)\,, \quad \widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}} = N^{-1}\sum_{i=1}^{N}\widehat{\mathcal{Z}}^{\varpi}_{X^i}\,. \tag{2.2.4}$$

We now study the performance of the NEO-IS estimator. The following two quantities play a fundamental role in the analysis:

$$E^{\varpi}_{\mathrm{T}} = \mathbb{E}_{X\sim\rho}\big[\big(\textstyle\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))/\mathcal{Z}\big)^2\big], M^{\varpi}_{\mathrm{T}} = \sup_{x\in\mathbb{R}^d}\sum_{k\in\mathbb{Z}} w_k(x)\mathrm{L}(\mathrm{T}^k(x))/\mathcal{Z}\,. \tag{2.2.5}$$

Theorem 2.2.1. *$\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ is an unbiased estimator of $\mathcal{Z}$. If $E^{\varpi}_{\mathrm{T}} < \infty$, then, $\mathbb{E}[|\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}/\mathcal{Z} - 1|^2] = N^{-1}(E^{\varpi}_{\mathrm{T}} - 1)$. If $M^{\varpi}_{\mathrm{T}} < \infty$, then, for any $\delta \in (0,1)$, with probability $1-\delta$, $\sqrt{N}\left|\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}/\mathcal{Z} - 1\right| \leq M^{\varpi}_{\mathrm{T}}\sqrt{\log(2/\delta)/2}$.*

The proof is postponed to Section A.1.3. E^{ϖ}_{T} plays the role of the second-order moment of the importance weights $\mathbb{E}_{X\sim\rho}[\mathrm{L}^2(X)]$ which is key to the performance of IS algorithms Agapiou et al. (2017); Akyildiz and Míguez (2021). In addition, since the NEO-IS estimator $\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ is unbiased, the Cauchy–Schwarz inequality implies that $\mathbb{E}_{X\sim\rho}\big[\big(\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))\big)^2\big] \geq \mathcal{Z}^2$ and hence that $E^{\varpi}_{\mathrm{T}} \geq 1$. Note that if $\|\mathrm{L}\|_\infty = \sup_{x\in\mathbb{R}^d}\mathrm{L}(x) < \infty$, then since the weights are uniformly bounded by $\Omega\varpi_0^{-1}$, we have $M^{\varpi}_{\mathrm{T}} \leq \|\mathrm{L}\|_\infty\Omega\varpi_0^{-1}/\mathcal{Z}$.

Using the NEO-IS estimate $\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ of the normalizing constant, we can construct a self-normalized IS estimate of $\int f(x)\pi(x)\mathrm{d}x$:

$$J^{\mathrm{NEO}}_{\varpi,N}(f) = N^{-1}\sum_{i=1}^{N}\frac{\widehat{\mathcal{Z}}^{\varpi}_{X^i}}{\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}}\sum_{k\in\mathbb{Z}}\frac{\mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)}{\widehat{\mathcal{Z}}^{\varpi}_{X^i}}f(\mathrm{T}^k(X^i))\,, \tag{2.2.6}$$

referred to as NEO-SNIS estimator. This expression may seem unnecessarily complicated but highlights the hierarchical structure of the estimator. We combine estimators

$$(\widehat{\mathcal{Z}}^{\varpi}_{X^i})^{-1}\sum_{k\in\mathbb{Z}}\mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)f(\mathrm{T}^k(X^i))$$

evaluated on the forward and backward orbits through the points $\{X^i\}_{i=1}^N$ using the normalized weights $\{\widehat{\mathcal{Z}}^{\varpi}_{X^i}/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}\}_{i=1}^N$. Although the NEO-IS estimator is unbiased, the NEO-SNIS is in general biased. However, for bounded functions, both the bias and the variance of the NEO-SNIS estimator are $O(N^{-1})$, with constants proportional to E^{ϖ}_{T}. For g a π-integrable function, we set $\pi(g) = \int g(x)\pi(x)\mathrm{d}x$.

Theorem 2.2.2. *Assume that $E^{\varpi}_{\mathrm{T}} < \infty$. Then, for any function g satisfying $\sup_{x\in\mathbb{R}^d}|g(x)| \leq 1$ on $\mathbb{R}^d$, and $N \in \mathbb{N}$*

$$\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)|^2\right] \leq 4\cdot N^{-1}E^{\varpi}_{\mathrm{T}}\,, \tag{2.2.7}$$

$$\left|\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)\right]\right| \leq 2\cdot N^{-1}E^{\varpi}_{\mathrm{T}}\,. \tag{2.2.8}$$

If $M^{\varpi}_{\mathrm{T}} < \infty$, then for $\delta \in (0,1]$, with probability at least $1-\delta$,

$$\sqrt{N}|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)| \leq \|g\|_\infty M^{\varpi}_{\mathrm{T}}\sqrt{32\log(4/\delta)}\,. \tag{2.2.9}$$

The proof is postponed to Section A.1.4. These results extend to NEO-SNIS estimators the results known for self-normalized IS estimators; see e.g., Agapiou et al. (2017); Akyildiz and Míguez (2021) and the references therein. The upper bounds stated in this result suggest it is good practice to keep $E^{\varpi}_{\mathrm{T}}/N$ small in order to obtain sensible approximations.

Lemma 2.2.3. *For any nonnegative sequence $(\varpi_k)_{k\in\mathbb{Z}}$, we have $E^{\varpi}_{\mathrm{T}} \leq \exp\left(\mathcal{R}_2(\pi \parallel \rho_{\mathrm{T}})\right)$.*

The proof is postponed to Section A.1.5. Lemma 2.2.3 suggests that accurate sampling requires N to scale linearly with the exponential of the 2-Rényi divergence between the target π and the extended proposal ρ_{T}.

Remark 2.2.4. *We can extend* NEO *to non homogeneous flows, replacing the family $\{\mathrm{T}^k : k \in \mathbb{Z}\}$ with a collection of mappings $\{\mathsf{T}_k : k \in \mathbb{Z}\}$. This would allow us to consider further flexible classes of transformations such as normalizing flows; see e.g. Papamakarios et al. (2019). The 2-Rényi divergence provides a natural criterion for learning the transformation. We leave this extension to future work.*

Conformal Hamiltonian transform The efficiency of NEO relies heavily on the choice of T. Intuitively, a sensible choice of T requires that (i) E^{ϖ}_{T} is small, i.e. ρ_{T} should be close to π by Lemma 2.2.3 (see (2.2.5)), (ii) the inverse T^{-1} and the Jacobian of T are easy to compute. Following Rotskoff and Vanden-Eijnden (2019), we use for T a discretization of a conformal Hamiltonian dynamics. Assume that $U(\cdot) = -\log\pi(\cdot)$ is continuously differentiable. We consider the augmented distribution $\tilde{\pi}(q,p) \propto \exp\{-U(q) - K(p)\}$ on $\mathbb{R}^{2d}$, where q is the position, p is the momentum, and $K(p) = p^T\mathrm{M}^{-1}p/2$ is the kinetic energy, with M a positive

definite mass matrix. By construction, the marginal distribution of the momentum under $\tilde{\pi}$ is the target pdf $\pi(q) = \int \tilde{\pi}(q,p)\mathrm{d}p$. The conformal Hamiltonian ODE associated with $\tilde{\pi}$ is defined by

$$\begin{aligned} \mathrm{d}q_t/\mathrm{d}t &= \nabla_p H(q_t,p_t) = \mathrm{M}^{-1}p_t\,, \\ \mathrm{d}p_t/\mathrm{d}t &= -\nabla_q H(q_t,p_t) - \gamma p_t = -\nabla U(q_t) - \gamma p_t\,, \end{aligned} \tag{2.2.10}$$

where $H(q,p) = U(q) + K(p)$, and $\gamma > 0$ is a damping constant. Any solution $(q_t,p_t)_{t\geq 0}$ of (2.2.10) satisfies setting $H_t = H(q_t,p_t)$, $\mathrm{d}H_t/\mathrm{d}t = -\gamma p_t^T \mathrm{M}^{-1} p_t \leq 0$. Hence, all orbits converge to fixed points that satisfy $\nabla U(q) = 0$ and $p = 0$; see e.g. Franca et al. (2020); Maddison et al. (2018).

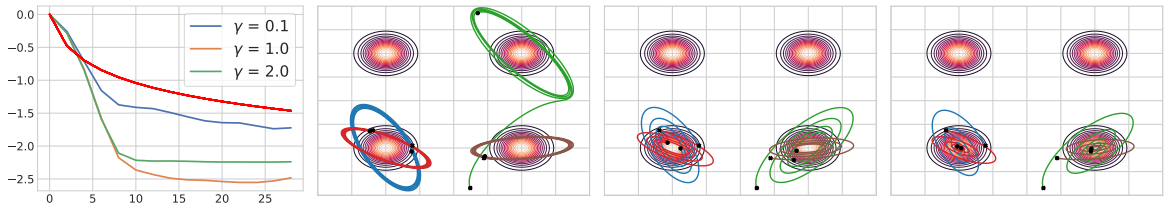

Figure 2.1: Left: $E_{\mathrm{T}_h}^{\mathbb{1}_{[K]}}(K) - 1$ *vs* $E^{\mathrm{IS}}(K) - 1$ *(red) in* $\log_{10}$*-scale as a function of the length of trajectories* K *(the lower the better). Second left to right: Four examples of orbits with the same random seed for different values of* γ *(from left to right,* $\gamma = 0.1, 1, 2$*).*

In the applications below, we consider the conformal version of the symplectic Euler (SE) method of (2.2.10), see Franca et al. (2020). This integrator can be constructed as a splitting of the two conformal and conservative parts of the system (2.2.10). When composing a dissipative with a symplectic operator, we set for all $(q,p) \in \mathbb{R}^{2d}$, $\mathrm{T}_h(q,p) = (q + h\mathrm{M}^{-1}\{\mathrm{e}^{-h\gamma}p - h\nabla U(q)\}, \mathrm{e}^{-h\gamma}p - h\nabla U(q))$, where $h > 0$ is a discretization stepsize. This transformation can be connected with classical momentum optimization schemes, see (Franca et al., 2020, Section 4). By (Franca et al., 2020, Section 3), for any $h > 0$ T_h is a C^1-diffeomorphism on $\mathbb{R}^{2d}$ with Jacobian given by $\mathbf{J}_{\mathrm{T}_h}(q,p) = \mathrm{e}^{-\gamma h d}$. In addition, its inverse is $\mathrm{T}_h^{-1}(q,p) = (q - h\mathrm{M}^{-1}p, \mathrm{e}^{\gamma h}\{p + h\nabla U(q - h\mathrm{M}^{-1}p)\})$. Therefore, the weight (2.2.3) of the NEO estimator is given by

$$w_k(q,p) = \frac{\varpi_k \tilde{\rho}(\mathrm{T}_h^k(q,p))\mathrm{e}^{-\gamma k h d}}{\sum_{j\in\mathbb{Z}} \varpi_{k+j}\tilde{\rho}(\mathrm{T}_h^{-j}(q,p))\mathrm{e}^{\gamma j h d}}\,,$$

where $\tilde{\rho}(q,p) \propto \rho(q)\mathrm{e}^{-K(p)}$. Figure 2.1 displays for different values of γ on a log-scale the bound $E_{\mathrm{T}_h}^{\mathbb{1}_{[0:K]}} - 1$ appearing in Theorem 2.2.1 as a function of K, here we use the sequence of weights $(\varpi_k)_{k\in\mathbb{Z}} = (\mathbb{1}_{[0:K]}(k))_{k\in\mathbb{Z}}$ (i.e. only the $K+1$ first elements of the forward orbits are used and are equally weighted). For comparison, we also present on the same plot the bounds achieved by averaging $K+1$ independent IS estimates, $E^{\mathrm{IS}}(K) - 1 = (K+1)^{-1}\mathbb{E}_{X\sim\rho}[\mathrm{L}(X)^2]$. Interestingly, Figure 2.1 shows that there is a trade-off in the choice of γ which controls the exploration of the state space by the Hamiltonian dynamics since the higher γ, the faster the orbits converge towards the modes. This fast convergence prevents a "good" exploration of the space; e.g. $E_{\mathrm{T}_h}^{\mathbb{1}_{[0:K]}}$ is smaller for $\gamma = 1.0$ than for $\gamma = 2.0$ when $K > 7$. the evolution of $E_{\mathrm{T}}^{\mathbb{1}_{[K]}}$ shows that the use of the forward orbit of a conformal Hamiltonian with an appropriately chosen damping factor outperforms the IS estimator. The associated trajectories are plotted for different values of γ (the higher γ, the faster the orbits converge towards the modes).

2.3 NEO-MCMC algorithm

We now derive an MCMC method to sample from π based on the NEO-IS estimator. A natural idea consists in adapting the Sampling Importance Resampling procedure (SIR) (see for example Rubin (1987); Skare et al. (2003)) to the NEO framework.

Algorithm 2 NEO-MCMC Sampler

At step $n \in \mathbb{N}^*$, given the conditioning orbit point Y_{n-1}.

Step 1: Update the conditioning point

1. Set $X_n^1 = Y_{n-1}$ and for any $i \in \{2, \dots, N\}$, sample $X_n^i \overset{\text{iid}}{\sim} \rho$.
2. Sample the orbit index I_n with probability proportional to $(\widehat{\mathcal{Z}}^{\varpi}_{X_n^i})_{i\in[N]}$, (2.2.4).
3. Set $Y_n = X_n^{I_n}$.

Step 2: Output a sample

4. Sample index K_n with probability proportional to $\{w_k(Y_n)\mathrm{L}(\mathrm{T}^k(Y_n))/\widehat{\mathcal{Z}}^{\varpi}_{Y_n}\}_{k\in\mathbb{Z}}$
5. Output $U_n = \mathrm{T}^{K_n}(Y_n)$.

The SIR method to sample $J^{\mathrm{NEO}}_{\varpi,N}$ (see (2.2.6)) consists of 4 steps.

(SIR-1) Draw independently $X^{1:N} \overset{\text{iid}}{\sim} \rho$ and compute the associated forward and backward orbits $\{\mathrm{T}^k(X^i)\}_{k\in\mathbb{Z}}$ of the point.
(SIR-2) Compute the normalizing constants associated with each orbit $\{\widehat{\mathcal{Z}}^{\varpi}_{X^i}\}_{i=1}^N$.
(SIR-3) Sample an orbit index $I^N \in [N]$ with probability $\{\widehat{\mathcal{Z}}^{\varpi}_{X^i}/\sum_{j=1}^N \widehat{\mathcal{Z}}^{\varpi}_{X^j}\}_{i=1}^N$.
(SIR-4) Draw the iteration index K^N on the I^N-th orbit with probability

$$\{\mathrm{L}(\mathrm{T}^k(X^{I^N}))w_k(X^{I^N})/\widehat{\mathcal{Z}}^{\varpi}_{X^{I^N}}\}_{k\in\mathbb{Z}}\,.$$

The resulting draw is denoted by $U^N = \mathrm{T}^{K^N}(X^{I^N})$. By construction, for any bounded function f, we get that $\mathbb{E}\left[f(U^N)\middle|X^{1:N}, I^N\right] = \{\widehat{\mathcal{Z}}^{\varpi}_{X^{I^N}}\}^{-1}\sum_{k\in\mathbb{Z}} w_k(X_{I^N})\mathrm{L}(\mathrm{T}^k(X_{I^N}))$ which implies $\mathbb{E}\left[f(U^N)\middle|X^{1:N}\right] = J^{\mathrm{NEO}}_{\varpi,N}(f)$ (see (2.2.6)). Using Theorem 2.2.2, we therefore obtain $|\mathbb{E}[f(U^N)] - \int f(z)\pi(z)\mathrm{d}z| \le 10^{1/2}\|f\|_\infty E^{\varpi}_{\mathrm{T}} N^{-1}$, showing that the law of the random variable $\mu_N = \mathrm{Law}(U^N)$ converges in total variation to π as $N \to \infty$,

$$\|\mu_N - \pi\|_{\mathsf{TV}} = \sup_{\|f\|_\infty \le 1} |\mu_N(f) - \pi(f)| \le 10^{1/2} E^{\varpi}_{\mathrm{T}} N^{-1}\,. \tag{2.3.1}$$

Based on Andrieu et al. (2010), we now derive the NEO-MCMC procedure, which in a nutshell consists in iterating the SIR procedure while keeping a conditioning point (or equivalently, orbit); see Section A.3. The convergence of NEO-MCMC does not rely on letting $N \to \infty$: the NEO-MCMC works as soon as $N \ge 2$, although as we will see below the mixing time decreases as N increases.

This procedure is summarized in Algorithm 2. The NEO-MCMC procedure is an iterated algorithm which produces a sequence $\{(Y_n, U_n)\}_{n\in\mathbb{N}}$ of points in $\mathbb{R}^d$. The n-th iteration of the NEO-MCMC algorithm consists in two main steps: 1) updating the conditioning point $Y_{n-1} \to Y_n$ 2) sampling U_n by selecting a point in the orbit $\{\mathrm{T}^k(Y_n)\}_{k\in\mathbb{Z}}$ of the conditioning point. Compared to SIR, only the generation of the points (step (SIR-1)) is modified: we set $X_n^1 = Y_{n-1}$ (the **conditioning point**), and then draw $X_n^{2:N} \overset{\text{iid}}{\sim} \rho$.

The sequence $\{Y_n\}_{n\in\mathbb{N}}$ defined by Algorithm 2 is a Markov chain:

$$\mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_{0:n-1}\right) = \mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_{n-1}\right) = P(Y_n, \mathsf{A})\,,$$

where

$$P(y,\mathsf{A}) = \int \delta_y(\mathrm{d}x^1) \prod_{j=2}^{N} \rho(x^j)\mathrm{d}x^j \sum_{i=1}^{N} \frac{\widehat{\mathcal{Z}}^{\varpi}_{x^i}}{\sum_{j=1}^{N} \widehat{\mathcal{Z}}^{\varpi}_{x^j}} \mathbb{1}_{\mathsf{A}}(x^i)\,, \quad y \in \mathbb{R}^d\,, \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\,. \tag{2.3.2}$$

Note that this Markov kernel describes the way, at stage $n+1$, the conditioning point Y_{n+1} is selected given Y_n, which **depends only on** the estimator of the normalizing constants associated with each orbit, **but not** on the sample U_n selected on the conditioning orbit. In addition, given the conditioning point Y_n at the n-th iteration, the conditional distribution of the output sample U_n is $\mathbb{P}\left(U_n \in \mathsf{B} \,|\, I_n, X_n^{1:N}\right) = \mathbb{P}\left(U_n \in \mathsf{B} \,|\, Y_n\right) = Q(Y_n, \mathsf{B})$ where

$$Q(y,\mathsf{B}) = \sum_{k\in\mathbb{Z}} \frac{w_k(y)\mathrm{L}(\mathrm{T}^k(y))}{\widehat{\mathcal{Z}}^{\varpi}_{y}} \mathbb{1}_{\mathsf{B}}(\mathrm{T}^k(y))\,, \quad y \in \mathbb{R}^d\,, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)\,. \tag{2.3.3}$$

With these notations, if the Markov chain is started at $Y_0 = y$, then for any $n \in \mathbb{N}$, the law of the n-th conditioning point is $\mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_0 = y\right) = P^n(y,\mathsf{A})$ and the law of the n-th sample is $\mathbb{P}\left(U_n \in \mathsf{B} \,|\, Y_0\right) = P^nQ(y,\mathsf{B})$. Define $\tilde{\pi}$ the pdf given for $y \in \mathbb{R}^d$ by

$$\tilde{\pi}(y) = \frac{\rho(y)}{\mathcal{Z}} \sum_{k\in\mathbb{Z}} w_k(y)\mathrm{L}(\mathrm{T}^k(y)) = \frac{\rho(y)\widehat{\mathcal{Z}}^{\varpi}_{y}}{\mathcal{Z}}\,. \tag{2.3.4}$$

The following theorem shows that, for any initial condition $y \in \mathbb{R}^d$, the distribution of the variable Y_n converges in total variation to $\tilde{\pi}$ and that the distribution of U_n converges to π.

Theorem 2.3.1. *The Markov kernel P is reversible w.r.t. the distribution $\tilde{\pi}$, ergodic and Harris positive,* i.e., *for all $y \in \mathbb{R}^d$, $\lim_{n\to\infty} \|P^n(y,\cdot) - \tilde{\pi}\|_{\mathsf{TV}} = 0$. In addition, $\pi = \tilde{\pi}Q$ and $\lim_{n\to\infty} \|P^nQ(y,\cdot) - \pi\|_{\mathsf{TV}} = 0$. Moreover, for any bounded function g and any $y \in \mathbb{R}^d$, $\lim_{n\to\infty} n^{-1}\sum_{i=0}^{n-1} g(U_i) = \pi(g)$, $\mathbb{P}$-almost surely, where $\{U_i\}_{i\in\mathbb{N}}$ is defined in Algorithm 2 with $Y_0 = y$.*

The proof is postponed to Section A.1.6.

Remark 2.3.2. We may provide another sampling procedure of $\{Y_n\}_{n\in\mathbb{N}}$. Define the pdf on the extended space $[N] \times \mathbb{R}^{dN}$ by $\breve{\pi}(i, x^{1:N}) = N^{-1}\tilde{\pi}(x^i)\prod_{j=1,j\neq i}^{N}\rho(x^j)$. Consider a Gibbs sampler targeting $\breve{\pi}$ consisting in (a) sampling $X_n^{1:N\setminus\{I_{n-1}\}}|(I_{n-1}, X_{n-1}) \sim \prod_{j\neq I_{n-1}}\rho(x^j)$, (b) sampling $I_n|X_n^{1:N} \sim \mathrm{Cat}(\{\widehat{\mathcal{Z}}^{\varpi}_{X_n^i}/\sum_{j=1}^{N}\widehat{\mathcal{Z}}^{\varpi}_{X_n^j}\}_{i=1}^{N}$ and (c) set $Y_n = X_n^{I_n}$. This algorithm is a Gibbs sampler on $\breve{\pi}$ and we easily verify that the distribution of $\{Y_n\}_{n\in\mathbb{N}}$ is the same as Algorithm 2.

The next theorem provides non asymptotic quantitative bounds on the convergence in total variation. The main interest of NEO-MCMC algorithm is motivated empirically from observed behaviour: the mixing time of the corresponding Markov chain improves as N increases. This behaviour is quantified theoretically in the next theorem. Moreover, this improvement is obtained with little extra computational overhead, since sampling N points from the proposal distribution ρ, computing the forward and backward orbits of the points and evaluating the normalizing constants $\{\widehat{\mathcal{Z}}^{\varpi}_{X_n^i}\}_{i=1}^{N}$ can be performed in parallel.

Theorem 2.3.3. *Assume that $M_{\mathrm{T}}^{\varpi} < \infty$, see (2.2.5). Set $\epsilon_N = (N-1)/(2M_{\mathrm{T}}^{\varpi} + N - 2)$ and $\kappa_N = 1 - \epsilon_N$. Then, for any $y \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\|P^k(y,\cdot) - \tilde{\pi}\|_{\mathsf{TV}} \leq \kappa_N^k$ and $\|P^kQ(y,\cdot) - \pi\|_{\mathsf{TV}} \leq \kappa_N^k$.*

Instead of sampling the new points $X_n^{2:N}$ independently from ρ (Step 1 in Algorithm 2), it is possible to draw the proposals $X_n^{1:N}$ conditional to the current point Y_{n-1}; see So (2006); Craiu

and Lemieux (2007); Shestopaloff et al. (2018); Ruiz et al. (2021) for related works. Following Ruiz et al. (2021), we use a reversible Markov kernel w.r.t. the proposal ρ, i.e., such that $\rho(x)m(x,x') = \rho(x')m(x',x)$, assuming for simplicity that this kernel has density $m(x,x')$. If $\rho = \mathrm{N}(0,\sigma^2\mathrm{I}_d)$, an appropriate choice is an autoregressive kernel $m(x,x') = \mathrm{N}(x';\alpha x,\sigma^2(1-\alpha^2)\mathrm{I}_d)$. More generally, we can use a Metropolis–Hastings kernel with invariant distribution ρ. In this case, $r_1(x^1,x^{1:N\setminus\{1\}}) = \prod_{j=2}^N m(x^{j-1},x^j)$ and for each $i \in [2:N]$,

$$r_i(x^i,x^{1:N\setminus\{i\}}) = \prod_{j=1}^{i-1} m(x^{j+1},x^j) \prod_{j=i+1}^{N} m(x^{j-1},x^j)\,. \tag{2.3.5}$$

Since m is reversible w.r.t. ρ, for all $i,j \in [N]$, $\rho(x^i)r_i(x^i,x^{1:N\setminus\{i\}}) = \rho(x^j)r_j(x^j,x^{1:N\setminus\{j\}})$ where $r_i(x^i;x^{1:N\setminus\{i\}})$ defines the the conditional distribution of $X^{1:N\setminus\{i\}}$ given $X^i = x^i$. The only modification in Algorithm 2 is Step 1, which is replaced by: *Draw $U_n \in [N]$ uniformly, set $X_n^{U_n} = Y_{n-1}$ and sample $X_n^{1:N\setminus\{U_n\}} \sim r_{U_n}(X_n^{U_n},\cdot)$*. The validity of this procedure is established in Section A.1.6.

2.4 Continuous-time version of NEO and NEIS

The NEO framework can be thought of as an extension of NEIS introduced in Rotskoff and Vanden-Eijnden (2019). NEIS focuses on normalizing constant estimation and should be therefore compared with NEO-IS. In Rotskoff and Vanden-Eijnden (2019), the authors do not consider possible extensions of these ideas to sampling problems. We consider here how NEO could be adapted to continuous-time dynamical system. Proofs of the statements and detailed technical conditions are postponed to Section A.2.

Consider the Ordinary Differential Equation (ODE) $\dot{x}_t = b(x_t)$, where $b\colon \mathbb{R}^d \to \mathbb{R}^d$ is a smooth vector field. Denote by $(\phi_t)_{t\in\mathbb{R}}$ the flow of this ODE (assumed to be well-behaved). Under appropriate regularity condition $\mathbf{J}_{\phi_t}(x) = \exp(\int_0^t \nabla\cdot b(\phi_s(x))\mathrm{d}s)$; see Lemma A.2.2. Let $\varpi : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative smooth function with finite support, with $\Omega^c = \int_{-\infty}^{\infty} \varpi(t)\mathrm{d}t$. The continuous-time counterpart of the proposal distribution (2.2.1) is $\rho_{\mathrm{T}}^{\mathrm{c}}(x) = (\Omega^c)^{-1}\int_{-\infty}^{\infty} \varpi(t)\rho(\phi_{-t}(x))\mathbf{J}_{\phi_{-t}}(x)\mathrm{d}t$, which is a continuous mixture of the pushforward of the proposal ρ by the flow of $(\phi_s)_{s\in\mathbb{R}}$. Assuming for simplicity that $\rho(x) > 0$ for all $x \in \mathbb{R}^d$, then $\rho_{\mathrm{T}}^{\mathrm{c}}(x) > 0$ for all $x \in \mathbb{R}^d$, and using again the IS formula, for any nonnegative function f,

$$\int f(x)\rho(x)\mathrm{d}x = \int f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}^{\mathrm{c}}(x)}\rho_{\mathrm{T}}^{\mathrm{c}}(x)\mathrm{d}x = \int \left[\int_{-\infty}^{\infty} w_t^{\mathrm{c}}(x) f(\phi_t(x))\mathrm{d}t\right]\rho(x)\mathrm{d}x\,, \tag{2.4.1}$$

$$w_t^{\mathrm{c}}(x) = \varpi(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\Big/\int_{-\infty}^{\infty}\varpi(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s\,. \tag{2.4.2}$$

These relations are the continuous-time counterparts of (2.2.2). Eqs. (2.4.1)-(2.4.2) define a version of NEIS Rotskoff and Vanden-Eijnden (2019), with a finite support weight function ϖ; see Sections A.2.2 and A.2.3 for weight functions with infinite support. This identity is of theoretical interest but must be discretized to obtain a computationally tractable estimator. For $h > 0$, denote by T_h an integrator with stepsize $h > 0$ of the ODE $\dot{x} = b(x)$. We may construct NEO-IS and NEO-SNIS estimators based on the transform $\mathrm{T} \leftarrow \mathrm{T}_h$ and weights $\varpi_k \leftarrow \varpi(kh)$. We might show that for any bounded function f and for any $x \in \mathbb{R}^d$, $\lim_{h\downarrow 0}\sum_{k\in\mathbb{Z}} w_k(x)f(\mathrm{T}_h^k(x)) = \int_{-\infty}^{\infty} w_t^{\mathrm{c}}(x)f(\phi_t(x))\mathrm{d}t$, where we omitted here the dependency in h of w_k. Therefore, taking $h \downarrow 0^+$, the NEO-IS converges to the continuous-version (2.4.1)-(2.4.2). There is however an important difference between NEO and the NEIS method in Rotskoff and Vanden-Eijnden (2019) which stems from the way (2.4.1)-(2.4.2) are discretized. Compared to

NEIS, NEO-IS using $\mathrm{T} \leftarrow \mathrm{T}_h$ and weights $\varpi_k \leftarrow \varpi(kh)$ is unbiased for any stepsize $h > 0$. NEIS uses an approach inspired by the nested-sampling approach, which amounts to discretizing the integral in (2.4.1) also in the state-variable x; see Skilling (2006); Chopin and Robert (2010). This discretization is biased which prevents the use of this approach to develop MCMC sampling algorithm; see Section A.2.

2.5 Experiments and Applications

Normalizing constant estimation The performance of NEO-IS is assessed on different normalizing constant estimation benchmarks; see Jia and Seljak (2020). We focus on two challenging examples. Additional experiments and discussion on hyperparameter choice are given in the supplementary material, see Section A.4.1.

(1) Mixture of Gaussian `(MG25)`: $\pi(x) = P^{-1} \sum_{i=1}^{P} \mathrm{N}(x; \mu_{i,j}, D_d)$, where $d \in \{10, 20, 45\}$, $D_d = \mathrm{diag}(0.01, 0.01, 0.1, \ldots, 0.1)$ and $\mu_{i,j} = [i, j, 0, \ldots, 0]^T$ with $i, j \in \{-2, \ldots, 2\}$.

(2) Funnel distribution `(Fun)` $\pi(x) = \mathrm{N}(x_1; 0, a^2) \prod_{i=1}^{d} \mathrm{N}(x_i; 0, \mathrm{e}^{2bx_1})$ with $d \in \{10, 20, 45\}$, $a = 1$, and $b = 0.5$. In both case, the proposal is $\rho = \mathrm{N}(0, \sigma_\rho^2 \mathrm{I}_d)$ with $\sigma_\rho^2 = 5$.

The NEO-IS estimator is compared with (i) the IS estimator using the proposal ρ, (ii) the Adaptive Importance Sampling (AIS) estimator of Tokdar and Kass (2010), (iii) Stochastic Normalizing Flows (SNF)[1] and (iv) the Neural Importance Sampling (NIS)[2]. For NEO-IS, we use $\varpi_k = \mathbb{1}_{[K]}(k)$ with $K = 10$ (ten steps on the forward orbit), and conformal Hamiltonian dynamics $\gamma = 1$, $M = 5 \cdot \mathrm{I}_d$ for dimensions $d = \{10, 20\}$, and $\gamma = 2.5$ for $d = 45$ (where γ is the damping factor, M the mass matrix, h is the stepsize of the integrator). The parameters of AIS are set to obtain a complexity comparable to NEO-IS; see Section A.4.1. For NIS, we use the default parameters and for SNF we used the same architectures as in Wu et al. (2020). In `Fun`, we set $\gamma = 0.2$, $K = 10$, $M = 5 \cdot \mathrm{I}_d$, and $h = 0.3$. The IS estimator was based on $5 \cdot 10^5$ samples, and NIS, NEO-IS and AIS were computed with $5 \cdot 10^4$ samples. Figure 2.2 shows that NEO-IS consistently outperforms the competing methods. NEIS is run with the default parameters of the implementation with $2 \cdot 10^4$ samples (to get the wall clock run time)

Sampling NEO-MCMC is assessed for the distributions `(MG25)` ($d = 40$) and `Fun` ($d = 20$). NEO-MCMC sampler is compared with (i) the No-U-Turn Sampler - Pyro library Bingham et al. (2019) - and (ii) i-SIR algorithm Ruiz et al. (2021). The proposal distribution is $\rho = \mathrm{N}(0, \sigma_\rho^2 \mathrm{I}_d)$ with $\sigma_\rho^2 = 5$. Dependent proposals are used (see (2.3.5)) with $m(x, x') = \mathrm{N}(x'; \alpha x, \sigma_\rho^2(1 - \alpha^2)\mathrm{I}_d)$ with $\alpha = 0.99$. For NUTS, the default parameters are used. For i-SIR, we use the same number of proposals $N = 10$, proposal distribution and dependent proposal as for NEO-MCMC. To perform a fair comparison, we use the same clock time for all three algorithms. The number of iterations for correlated i-SIR, NEO-MCMC, and NUTS are $n = 4 \cdot 10^6$, $n = 4 \cdot 10^5$, and $n = 5 \cdot 10^5$, respectively. Figure 2.3 displays the empirical two-dimensional histograms of the two first coordinates of samples from the ground truth, i-SIR, NUTS and NEO-MCMC sampler. It is worthwhile to note that NEO-MCMC algorithm performs much better for `MG25` which is a very challenging distribution, even for SOTA algorithm such as NUTS, which struggles to cross energy barriers between modes. For `Fun`, NEO-MCMC performs favourably w.r.t. NUTS, which is well adapted for this type of distributions.

[1] Implementation available at https://github.com/noegroup/stochastic_normalizing_flows.
[2] Implementation available at https://github.com/ndeutschmann/zunis.

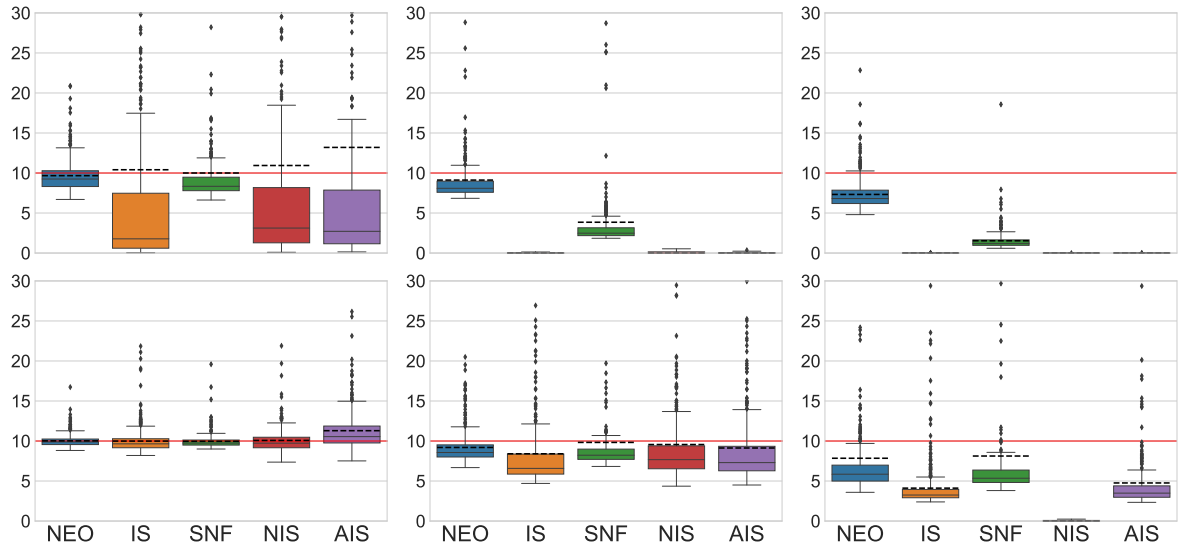

Figure 2.2: Boxplots of 500 independent estimations of the normalizing constant in dimension $d = \{10, 20, 45\}$ *(from left to right) for* `MG25` *(top) and* `Fun` *(bottom). The true value is given by the red line. The figure displays the median (solid lines), the interquartile range, and the mean (dashed lines) over the 500 runs.*

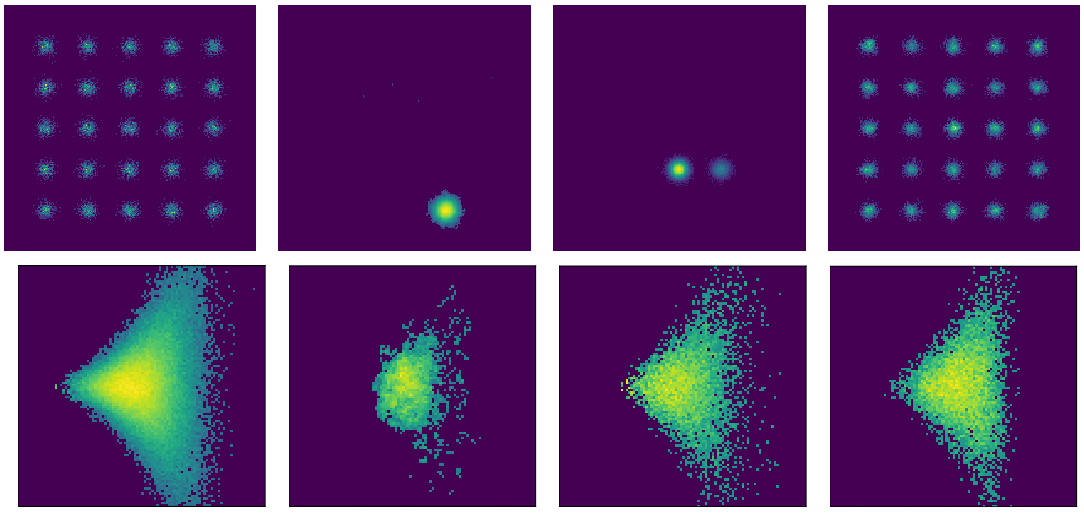

Figure 2.3: Empirical 2-D histogram of the samples of different algorithms targeting `MG25` *(top) and* `Fun` *(bottom). Left to right: samples from the target distribution, correlated i-SIR, NUTS,* NEO-*MCMC.*

Block Gibbs Inpainting with Deep Generative models and NEO**-MCMC** We apply NEO-MCMC to the task of sampling the posterior of a deep latent variable model. To be consistent with the rest of the chapter, we use non-standard notation here with x being the latent variable and z the observation. More precisely, we assume that $x \sim \mathrm{N}(0, \mathrm{I}_d)$ and a conditional distribution $p(z \mid x)$ which generates an image $z = (z^1, \ldots, z^D) \in \mathbb{R}^D$. Given a family of parametric *decoders* $\{x \mapsto p_\theta(z \mid x),\ \theta \in \Theta\}$, and a training set $\mathcal{D} = \{z_i\}_{i=1}^M$, training involves finding the MLE $\theta^* = \arg\max_{\theta \in \Theta} p_\theta(\mathcal{D})$. As $p_\theta(z) = \int p_\theta(z \mid x) p(x) \mathrm{d}x$, the likelihood is intractable and to alleviate this problem, Kingma and Welling (2014) proposed to train jointly an approximate posterior $q_\phi(x|z)$ that maximizes a tractable lower-bound on

the log-likelihood: $\mathrm{ELBO}(z, \theta, \phi) = \mathbb{E}_{X \sim q_\phi(\cdot|z)}[\log p_\theta(z, X)/q_\phi(X|z)] \leq p_\theta(z)$, where $q_\phi(x \mid z)$ is a tractable conditional distribution with parameters $\phi \in \Phi$. It is assumed in the sequel that conditional to the latent variable x, the coordinates are independent, *i.e.* $p_\theta(z \mid x) = \prod_{i=1}^{D} p_\theta(z^i|x)$.

Note that it is possible to train VAE with the NEO algorithm, using the unbiased estimate of the normalizing constant to construct an ELBO. This approach is described in the supplement Section A.5. We do not focus on this approach here and assume that the VAE has been trained and we are only interested in the sampling problem. In our experiment, we use a VAE trained on CelebA dataset [3] Liu et al. (2018). We consider the Block Gibbs inpainting task introduced in (Levy et al., 2018, Section 5.2.2). Given an image z, denote by $[z^t, z^b]$ the top and the bottom half pixels. Assume only $z^t_\star$ is observed, then we are interested in in-painting the bottom of an image by the posterior distribution of z^b given $z^t_\star$. This is achieved using Block Gibbs sampling. A two-stage Gibbs sampler amounts to (a) sampling $p_{\theta^*}(x|z^t, z^b)$ and (b) sampling $p_{\theta^*}(z^b|x, z^t) = p_{\theta^*}(z^b|x)$ (since z^b and z^t are independent conditional on x). Given $z_k = (z^t_\star, z^b_k)$, we sample at each step $x_k \sim p_{\theta^*}(x \mid z_k)$ and then $z^b_{k+1} \sim p_{\theta^*}(z^b \mid x_k)$. We then set $z_{k+1} = (z^t_*, z^b_{k+1})$. Stage (b) is elementary but stage (a) is challenging. We use an MCMC-within-Gibbs scheme using different samplers. We use the following decomposition of $p_{\theta^*}(x \mid z) \propto \rho(x)\mathrm{L}(x)$ for $\rho(x) \propto q^\beta_{\phi^*}(x \mid z)$ and $\mathrm{L}(x) = p_{\theta^*}(x, z)/q^\beta_{\phi^*}(x \mid z)$ with $\beta \in (0, 1)$. It is possible to sample from $\rho(x)$ as $q_{\phi^*}(x \mid z)$ is Gaussian. In our experiments with CelebA and the chosen trained VAE, we have $x \in \mathbb{R}^{10}$ (recall that x is our latent variable here), $z \in \mathbb{R}^{12288}$, and use $\beta = 0.1$. We then compare i-SIR, HMC and NEO-MCMC sampler in stage (a), with the same computational complexity ($N = 10$, $K = 12$, $\gamma = 0.2$ for NEO-MCMC, $N = 120$ for i-SIR, and HMC is run with $K = 20$ leap-frog steps). Again, NEO-MCMC and i-SIR use dependent proposals, with m a Random Walk Metropolis kernel with stepsize 0.1. For each algorithm, 10 steps are performed. Figure A.4 displays the evolution of the resulting Markov chains. The samples clearly illustrate that NEO-MCMC mixes better than i-SIR and HMC. More details and examples are presented in the supplementary.

2.6 Conclusion and perspectives

In this chapter, we have proposed a new family of algorithms, NEO, for computing normalizing constants and sampling from complex distributions. This methodology comes with asymptotic and non-asymptotic convergence guarantees. For normalizing constant estimation, NEO-IS compares favorably to state-of-the-art algorithms on difficult benchmarks. NEO-MCMC is also able to sample some complex distributions: it is particularly well-adapted to sampling multimodal distributions, thanks to its proposal mechanism which avoids being trapped in local modes. There are numerous potential extensions to this work. For example, it would be interesting to consider deterministic transformations other than conformal Hamiltonian dynamics integrators. These transformations could be trained, as for Neural IS, using a variation lower bound. It would also be interesting to further investigate the influence of the mixture weights $\{\varpi_k\}_{k \in \mathbb{Z}}$ on the efficiency of NEO. Also, while it is not investigated in the present chapter, we believe that the NEO estimator with the Hamiltonian transform can be useful in particle filtering Pitt and Shephard (1999). Finally, our bounds are valid for any transform T and it would be interesting to derive specific bounds for the Hamiltonian transform in order to quantify the improvements over the vanilla SNIS estimator with proposal ρ observed in Figure 2.1.

[3] Publicly available online, see https://github.com/YannDubs/disentangling-vae/tree/master/results/betaH_celeba

Figure 2.4: Two examples for the Gibbs inpainting task for CelebA dataset. From top to bottom (twice) : i-SIR, HMC and NEO-*MCMC: From left to right, original image, blurred image to reconstruct, and output every 5 iterations of the Markov chain. Last line: a forward orbit used in* NEO-*MCMC for the second example.*

Chapter 3

Entropic Mirror Monte Carlo

3.1 Introduction

In Bayesian statistics, inference on unknown quantities often requires sampling from complex posterior densities, which are frequently intractable. As an alternative to the dominant paradigm of Markov chain Monte Carlo (MCMC), Variational Inference (VI) methods aim to approximate the posterior density. They achieve this by seeking the best variational density, with respect to some user-specified criterion, within a set of probability densities F_Θ parameterized by variational parameters θ that lie in a parameter space Θ. This approach effectively formulates an optimization problem to find the most suitable approximation. The criterion optimized is defined by the specific task being solved. When learning the parameters of a hierarchical model, the forward Kullback-Leibler (KL) divergence appears naturally as the appropriate criterion through the *evidence lower bound* (Kingma and Welling, 2013; Blei et al., 2017). When the goal is to estimate expectations with respect to the posterior by means of importance sampling (IS), which will be the focus of this paper, the backward KL and the 2-Rényi divergence are instead the natural criteria (Cappe et al., 2005; Agapiou et al., 2017; Chatterjee and Diaconis, 2018).

More formally, let π be a target distribution known up to a constant and μ be a proposal distribution dominating π and from which sampling is tractable. Using the change of measure $\pi(\mathrm{d}x) = \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)\mu(\mathrm{d}x)$, any expectation $\pi(f) := \int f(x)\pi(\mathrm{d}x) = \mu(f\frac{\mathrm{d}\pi}{\mathrm{d}\mu})$ can be estimated using N independent samples $(X^1, \dots, X^N)$ drawn from μ,

$$\pi_\mu^N(f) := \sum_{i=1}^N \omega^i f(X^i)\,, \quad \text{where} \quad \omega^i \propto \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(X^i)\,,$$

and $\sum_{i=1}^n \omega^i = 1$. The so-called *self normalized importance sampling* estimator is useful whenever sampling directly from π is impossible but can also be seen as a variance reduction method of the crude Monte Carlo (MC) estimator, in particular for rare event simulation. Over broad classes of functions, the performance of $\pi_{n,\mu}(f)$ relies on the quality of the importance distribution μ and is captured by the following bound (Agapiou et al., 2017),

$$\sup_{|f|\leq 1} \left\|\pi_\mu^N(f) - \pi(f)\right\|_{2,\mu^{\otimes N}}^2 \leq \frac{4}{N}\exp\left(\mathcal{R}_2(\pi \parallel \mu)\right),$$

where $\|.\|_{2,\mu^{\otimes N}}$ is the 2-norm with respect to the product measure $\mu^{\otimes N}$ and $\mathcal{R}_2(\pi \parallel \mu)$ the 2-Rényi divergence. Hence, the sample size needs to grow exponentially with the 2-Rényi divergence to get a sharp upper-bound on the Mean Squared Error (MSE) between the importance

and the target distributions. Moreover, according to Chatterjee and Diaconis (2018), a log sample size set to the Kullback-Leibler (KL) divergence between π and ν is also necessary for guarantees in high probability. These results highlight the relevance of optimizing $\mathcal{R}_2(\pi \parallel \mu)$ and $\mathsf{KL}(\pi \parallel \mu)$.

One of the earliest attempts to optimize the backward KL divergence can be traced back to the cross-entropy (CE) method (Rubinstein, 1999; De Boer et al., 2005). CE is a stochastic optimization procedure for the minimization of $\mu \mapsto \mathsf{KL}(\pi \parallel \mu)$ within a family of parametric probability measures $\mathsf{F}_\Theta = \{\mu_\theta \in \mathsf{M}_1 \,:\, \theta \in \Theta\}$ for which sampling and density evaluation are straightforward. As $\mathsf{KL}(\pi \parallel \mu_\theta)$ is an expectation with respect to π from which sampling is intractable, the optimization is carried out using a reference probability measure $\mu_{\mathrm{ref}} \in \mathsf{F}_\Theta$ as importance proposal from which samples are drawn. The generalization of this methodology is known as *adaptive importance sampling* (AIS) (Oh and Berger, 1993; Cappé et al., 2004; Cornuet et al., 2012). AIS is an iterative procedure in which the importance proposal is dynamically adjusted using weighted samples from the proposal of the previous iteration. In contrast, the proposal in CE is updated at each step using samples from the fixed reference measure. Traditional AIS methods differ by the weighting scheme (Cappé et al., 2004, 2008; Cornuet et al., 2012; Elvira et al., 2017; Korba and Portier, 2022) and on how the proposal is updated (Douc et al., 2007b; Cappé et al., 2008; Cornuet et al., 2012). The parametric family F_Θ is often chosen to be that of the mixture of an exponential family distribution and specific procedures have been developed to handle these cases. In Cappé et al. (2008), an *integrated* Expectation-Maximization (EM) algorithm for the optimization of $\mathsf{F}_\Theta \ni \mu_\theta \mapsto \mathsf{KL}(\pi \parallel \mu_\theta)$ is developed, yielding explicit updates for the means, weights and covariance matrices for Gaussian and Student-t mixtures. This procedure effectively avoids the need to parameterize the weights and covariance matrices during the optimization.

Arguably, one of the main bottlenecks when performing AIS is the family of proposals on which the optimization is performed. In designing a parametric family F_Θ, two key considerations are vital; the densities within the chosen family must have more modes and heavier tails than the target distribution. Unfortunately, obtaining such information beforehand is challenging, as we only have access to the density and estimating any additional statistics necessitates is not at all trivial. To circumvent these hurdles, non-parametric AIS methods based on kernel density estimates, which can atleast adapt to the multimodality of the target, have been considered (Neddermeyer, 2009; Dai et al., 2016; Delyon and Portier, 2021; Korba and Portier, 2022). In more recent years, the particular aspect of designing an appropriate parametric family for AIS has benefitted from the advances in probabilistic modelling driven by variational inference and generative modelling. Normalizing flows (Tabak et al., 2010; Papamakarios et al., 2021), which can represent arbitrarily complex probability measure with positive density using deep learning architectures, have emerged as efficient tools for *automatic* adaptation to the tails and multimodality of the target distribution, thus paving the way for *black-box* AIS. This recent research direction have resulted in a few computationally efficient methods that scale well with the dimension and have helped bridge the gap between AIS and adaptive Markov chain Monte Carlo (MCMC) methods. In Naesseth et al. (2020); Gabrié et al. (2022); Samsonov et al. (2022), a normalizing flow is adapted by minimizing the backward KL divergence estimated using MCMC kernels with proposal the normalizing flow at the previous parameter, thus resulting in an AIS proposal as well as an adaptive MCMC algorithm. In a more traditional way, Prangle and Viscardi (2023) estimate the backward KL using importance sampling with the normalizing flow at the previous parameter as proposal and by truncating the importance weights. Arbel et al. (2021) optimize the backward KL using the main ingredients of a sequential Monte Carlo (SMC) sampler, i.e. resampling and MCMC kernels. Although these methods may not explicitly

identify themselves as AIS techniques, they still belong to its broader framework; obtaining the subsequent proposal in these approaches involves learning from samples that have been transformed to resemble samples from π as closely as possible. They are transformed either through weighting, MCMC kernels, non-linear transformations or reweighting.

Contributions

In this paper, we develop a novel AIS scheme by recursively defining a sequence of probability measures $\{\mu_t\}_{t\in\mathbb{N}}$ by $\mu_{t+1} = \mathcal{F}(\mu_t)$ where $\mathcal{F}$ is a well-designed functional on the set of probability measures $\mathcal{P}(\mathsf{X})$ ensuring contraction of the iterates with respect to the backward KL. The functional considered builds upon *Entropic Mirror Descent* (EMD) updates considered in Beck and Teboulle (2003); Dai et al. (2016); Daudel et al. (2021a); Korba and Portier (2022). While this sequence results in the geometric decrease of the backward KL along the iterates, its practical implementation can lead to poor parametric approximations of the target distribution (see Figure 3.2 below). To address this problem, we augment the original EMD updates with Markovian dynamics that leave invariant the target density π in order to improve the exploration of the state space. The resulting AIS scheme informs the samples of the proposal by combining regularized weights, that achieve a bias variance trade-off, and a non-standard weight which allows the greedy exploration of the target if the involved Markov kernel allows global moves. In this context, our contributions are the following:

- In Section 3.2 we define a principled objective, coined *skewed Rényi divergence*, to minimize in order obtain an appropriate practical implementation of a sequence of iterates that contracts with respect to the backward KL.
- Motivated by the failure of the optimization procedure in the case of EMD, which is due to the nature of the sequence and not the objective, we derive in Section 3.2.2 a novel sequence akin to EMD that incorporates Markovian moves. We show its backward KL contraction under mild assumptions, mainly the π-invariance of the underlying Markov kernel. When the Markov kernel is that of the Unadjusted Langevin Algorithm (ULA), we show contraction in total variation distance up to a discretization error explicit in the step-size and the dimension of the ambient space.
- In Section 3.2.3 we define the vanilla stochastic version of our algorithm and improve it by means of additional resampling and local MCMC steps. The numerical experiments, which are presented in Section 6.3, show that our algorithm outperforms existing AIS methods on highly challenging examples.

3.2 Entropic Mirror Monte Carlo

3.2.1 General framework

In this chapter, we are interested in mappings $\mathcal{F} : \mathsf{M}_\pi \to \mathsf{M}_\pi$ that induce a systematic decrease of the backward KL divergence, i.e. such that for all $\mu \in \mathsf{M}_\pi$

$$\mathsf{KL}(\pi \parallel \mathcal{F}(\mu)) \leq \rho \mathsf{KL}(\pi \parallel \mu), \tag{3.2.1}$$

where $\rho \in [0, 1)$. Let us start by providing one notable example of such a map.

Entropic Mirror Descent. For all $\mu \in \mathsf{M}_\pi$ write:

$$\mathcal{F}_{\mathsf{e}}(\mu) = \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)^{\varepsilon} \mu \Big/ \int_{\mathbb{R}^d} \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)^{\varepsilon} \mathrm{d}\mu\,, \tag{3.2.2}$$

where $\varepsilon \in (0,1]$. The functional $\mathcal{F}_{\mathsf{e}}$ corresponds to one iteration of the *Mirror Descent* algorithm applied to the convex map $\mu \mapsto \mathsf{KL}(\mu \parallel \pi)$ Beck and Teboulle (2003). In Korba and Portier (2022), it is shown that (3.2.1) is satisfied with $\rho = 1-\varepsilon$. As a result, the closer ε to 1 is, the faster the convergence. On the other hand, since $\varepsilon \in [0,1]$, by Jensen's inequality $\int (\mathrm{d}\pi/\mathrm{d}\mu)^{\varepsilon}\mathrm{d}\mu \leq 1$ and

$$\mathsf{KL}(\mu \parallel \mathcal{F}_{\mathsf{e}}(\mu)) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^{\varepsilon} \mathrm{d}\mu + \log \int \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)^{\varepsilon} \mathrm{d}\mu \leq \varepsilon \mathsf{KL}(\mu \parallel \pi)\,, \tag{3.2.3}$$

the value of ε controls the discrepancy between a probability measure $\mu \in \mathsf{M}_\pi$ and its update $\mathcal{F}_{\mathsf{e}}(\mu)$ in forward KL. These two observations naturally lead in practice to a tradeoff between speed of convergence and the quality of the approximation of $\mathcal{F}_{\mathsf{e}}(\mu)$.

While appealing theoretically, the iterates of a mapping satisfying (3.2.1) are in general intractable for both sampling and density evaluation. Therefore, in order to approximate these updates we plan on projecting each application of the mapping $\mathcal{F}$ on a family of probability measures $\mathsf{F}_\Theta = \{\mu_\theta \in \mathsf{M}_1, \theta \in \Theta\}$. Note that Korba and Portier (2022) focus instead on building non-parametric approximations with kernel density estimates.

We give below one notable example of family F_Θ.

Example 3.2.1 (Normalizing flows)**.** *Let $\mathrm{T}_\theta : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a C^1-diffeomorphism for any $\theta \in \Theta$. For any $\mu \in \mathsf{M}_1$ admitting a density w.r.t. the Lebesgue measure, the pushforward $\mathrm{T}_\theta \# \mu$, which corresponds to the law of $\mathrm{T}_\theta(X)$ where $X \sim \mu$, has density*

$$\mathrm{T}_{\theta\#}\mu(y) = \mu(\mathrm{T}_\theta^{-1}(y))|\mathbf{J}_{\mathrm{T}_\theta^{-1}}(y)|\,,$$

where $|\mathbf{J}_{\mathrm{T}_\theta^{-1}}(y)|$ is the determinant of the Jacobian at $y \in \mathbb{R}^d$. By taking μ to be a Gaussian (or in fact any simple distribution) and T_θ an invertible neural network with cheap to compute Jacobian, we can model complex distributions with $\mathrm{T}_{\theta\#}\mu$ using SGD Papamakarios et al. (2021).

We may consider two different choices for the projection step as we will now see. In either case, the projection step will be solved approximately once we are able to obtain an empirical approximation ζ_t^N of $\mathcal{F}(\mu_{\theta_t})$,

$$\zeta_t^N = \sum_{i=1}^N \omega_t^i \delta_{X_t^i}\,, \tag{3.2.4}$$

where $(X_t^1, \ldots, X_t^N)$ are random variables and the weights $\{\omega_t^i\}_{i=1}^N$ sum to one.

(i) *Backward KL projection.* As we will essentially work with the particle approximations of the updated measures $\mathcal{F}(\mu)$, weighted maximum likelihood seems to be the most straightforward way for obtaining a parametric approximation. The population criterion is then

$$\mu_{\theta_{t+1}} = \operatorname*{argmin}_{\nu \in \mathsf{F}_\Theta} \mathsf{KL}(\mathcal{F}(\mu_{\theta_t}) \parallel \nu)\,, \tag{3.2.5}$$

which is solved approximately by replacing $\mathcal{F}(\mu_{\theta_t})$ with its particle approximation (3.2.4);

$$\mu_{\theta_t} = \operatorname*{argmin}_{\nu \in \mathsf{F}_\Theta} \int \log \frac{\mathrm{d}\mathcal{F}(\mu_{\theta_t})}{\mathrm{d}\nu} \mathrm{d}\zeta_t^N = \operatorname*{argmin}_{\nu \in \mathsf{F}_\Theta} -\sum_{i=1}^n \omega_t^i \log \nu(X_t^i)\,. \tag{3.2.6}$$

Of course, the cost of replacing (3.2.5) with (3.2.6) heavily depends on the quality of the empirical approximation (3.2.4) which is discussed later on.

(ii) *Skewed Rényi projection.* Let $\nu \in \mathsf{M}_\pi$. For any $\mu \in \mathsf{M}_\pi$ we define the *skewed Rényi divergence*

$$\widetilde{\mathcal{R}}_\pi(\nu \parallel \mu) = \log \int \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \mathrm{d}\nu\,. \tag{3.2.7}$$

If $\nu = \pi$ then $\widetilde{\mathcal{R}}_\pi(\pi \parallel \mu) = \mathcal{R}_2(\pi \parallel \mu)$. For $\mu = \pi$ or $\mu = \nu$ we get that $\widetilde{\mathcal{R}}_\pi(\nu \parallel \mu) = 0$, and by strict convexity of $x \mapsto 1/x$ on $\mathbb{R}_{>0}$ we also get that $\widetilde{\mathcal{R}}_\pi(\nu \parallel \alpha\pi + (1-\alpha)\nu) < 0$ for $\alpha \in (0,1)$ which contrasts with the non-negativity of the usual Rényi divergence. We motivate the introduction of this functional with the following inequality. For all $\mu \in \mathsf{M}_\pi$,

$$\begin{aligned}\mathsf{KL}(\pi \parallel \mu) &= \mathsf{KL}(\pi \parallel \mathcal{F}(\mu_{\theta_{t-1}})) + \int \log\left(\mathrm{d}\mathcal{F}(\mu_{\theta_{t-1}})/\mathrm{d}\mu\right)\mathrm{d}\pi \\ &\leq \rho\mathsf{KL}(\pi \parallel \mu_{\theta_{t-1}}) + \log \int \left(\mathrm{d}\mathcal{F}(\mu_{\theta_{t-1}})/\mathrm{d}\mu\right)\mathrm{d}\pi \\ &= \rho\mathsf{KL}(\pi \parallel \mu_{\theta_{t-1}}) + \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{t-1}}) \parallel \mu)\,.\end{aligned}$$

where we have applied Jensen's inequality and the contraction of the mapping $\mathcal{F}$ (3.2.1). Iterating the bound, we get that for any sequence $\{\mu_{\theta_s}\}_{s=0}^t$,

$$\mathsf{KL}(\pi \parallel \mu_{\theta_t}) \leq \rho^t \mathsf{KL}(\pi \parallel \mu_{\theta_0}) + \sum_{s=1}^{t} \rho^{t-s} \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_{\theta_s})\,, \tag{3.2.8}$$

which hints that one way to achieve a small backward KL between π and μ_{θ_t} is to take for all $s \in [1:t]$

$$\mu_{\theta_s} = \operatorname*{argmin}_{\mu \in \mathsf{F}_\Theta} \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu)\,. \tag{3.2.9}$$

Furthermore, any mismatches resulting from $\widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_{\theta_s})$ being too far from the minimum are forgotten geometrically fast as t grows. If we are able to achieve a uniform error Δ when estimating each $\mathcal{F}(\mu_{\theta_s})$, i.e. $\widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_{\theta_s}) \leq \Delta$ for all $s \in [1:T]$ where T is the total number of steps, then the final approximation μ_{θ_T} satisfies

$$\mathsf{KL}(\pi \parallel \mu_{\theta_T}) \leq \rho^T \mathsf{KL}(\pi \parallel \mu_{\theta_0}) + \frac{\Delta}{1-\rho}\,. \tag{3.2.10}$$

Let us now detail how to approximate the solution to (3.2.9) when Gradient Descent is performed on the parameters of the parametric family. Under common differentiability assumptions we have that

$$\nabla_\theta \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_\theta) = -\int \left\{ \frac{\mathrm{d}\pi/\mathrm{d}\mu_\theta(x)}{\int (\mathrm{d}\pi/\mathrm{d}\mu_\theta)\mathrm{d}\mathcal{F}(\mu_{\theta_{s-1}})} \nabla_\theta \log \mu_\theta(x) \right\} \mathcal{F}(\mu_{\theta_{s-1}})(\mathrm{d}x)\,, \tag{3.2.11}$$

and a biased Monte Carlo approximation of the gradient is thus obtained by plugging in an empirical approximation (3.2.24) of $\mathcal{F}(\mu_{\theta_{s-1}})$,

$$\widehat{\nabla}_\theta \widetilde{\mathcal{R}}_\pi(\mathcal{F}(\mu_{\theta_{s-1}}) \parallel \mu_\theta) = -\sum_{i=1}^{N} \frac{\omega_{s-1}^i \mathrm{d}\pi/\mathrm{d}\mu_\theta(X_{s-1}^i)}{\sum_{j=1}^n \omega_{s-1}^j \mathrm{d}\pi/\mathrm{d}\mu_\theta(X_{s-1}^j)} \nabla \log \mu_\theta(X_{s-1}^i)\,. \tag{3.2.12}$$

In the case of (3.2.2), we may consider the following weighted empirical approximation obtained by sampling from $\mu_{\theta_{t-1}}$

$$\zeta_{t-1}^N = \sum_{i=1}^{N} \frac{\mathrm{d}\pi/\mathrm{d}\mu_{\theta_{t-1}}(X_{t-1}^i)^\varepsilon}{\sum_{j=1}^n \mathrm{d}\pi/\mathrm{d}\mu_{\theta_{t-1}}(X_{t-1}^j)^\varepsilon} \delta_{X_{t-1}^i}, \quad \text{where} \quad X_{t-1}^1, \ldots, X_{t-1}^N \overset{\text{iid}}{\sim} \mu_{\theta_{t-1}}\,. \tag{3.2.13}$$

In the next example we illustrate that when N is not large enough (3.2.13) can be a poor approximation even if ε is chosen so that the forward KL between $\mu_{\theta_{t-1}}$ and $\mathcal{F}(\mu_{\theta_{t-1}})$ is small following (3.2.3), resulting in a non-convergent scheme.

Example 3.2.2. *We consider two target distributions,* $\pi_1 = \mathcal{N}(10, 1)$ *and* $\pi_2 = 0.5 \cdot \mathcal{N}(0, 1) + 0.5 \cdot \mathcal{N}(10, 1)$*. In Figure 3.1 and 3.2 we display the iterates of the EMD mapping (3.2.2)* $\mathcal{F}_{\mathsf{e}}$*,* $\mu_t = \mathcal{F}_{\mathsf{e}}(\mu_{t-1})$ *targeting each distribution and their approximations* μ_{θ_t}*. The parametric family is respectively that of the unimodal and bimodal Gaussians. We use the Expectation-Maximization algorithm to solve (3.2.6) using* $N = 5000$ *samples. In the unimodal case, there is overlap between two consecutive iterates so that the updates are well approximated. In the bimodal example, this is not the case anymore and the approximate iterates fail to converge as sampling from the second mode remains unlikely.*

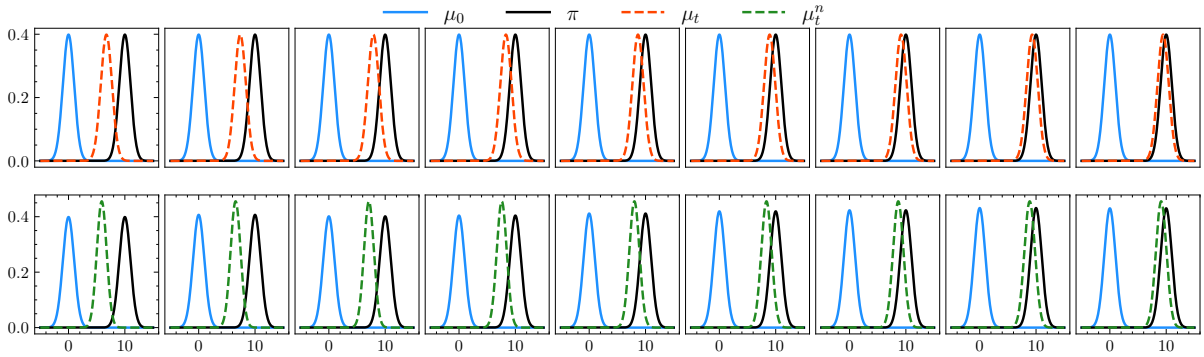

Figure 3.1: Unimodal target. Top plot: the exact iterates of μ_t *(3.2.2), bottom plot: approximate iterates* μ_{θ_t} *(3.2.6). We only display the iterates from* $t = 6$ *to* $t = 15$*.* ε *is set to* 0.2*.*

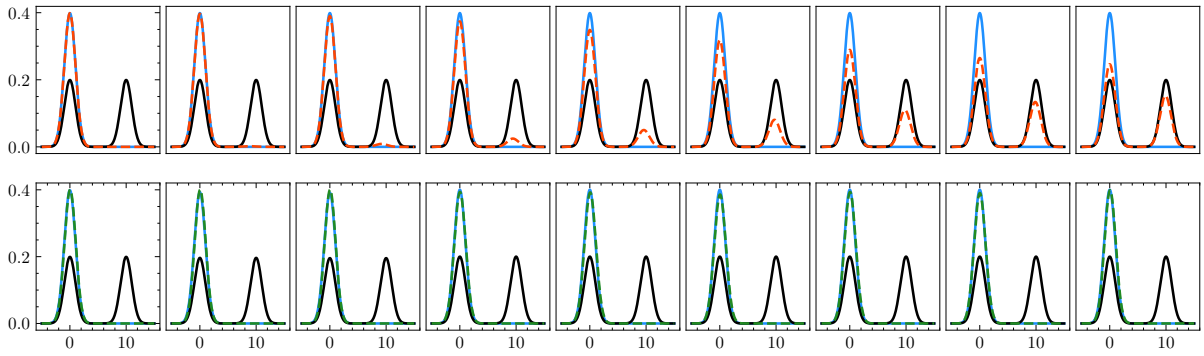

Figure 3.2: Bimodal target. The setup is the same as in Figure 3.1.

The issue underlying the previous example is the same as in importance sampling. Indeed, if the (conditional) variance of the unnormalized weights $\mathrm{d}\pi/\mathrm{d}\mu_{\theta_{t-1}}(X_{t-1}^i)^\varepsilon$ that appears in (3.2.13) is too large then (3.2.13) will be a poor approximation which in turn hinders the convergence to π.

3.2.2 Entropic Mirror Descent with Markov kernels

In this section we introduce our modification of the mapping (3.2.2) using a π-invariant Markov kernel K_π. Let $\varepsilon \in [0, 1]$ and $\lambda \in [0, 1]$. For all $\mu \in \mathsf{M}_\pi$ write

$$\mathcal{F}_{\mathsf{em}}(\mu; \lambda, K_\pi, \varepsilon) = \lambda \cdot \mathcal{F}_{\mathsf{e}}(\mu) + (1 - \lambda) \cdot \mathcal{F}_{K_\pi}(\mu)\,. \tag{3.2.14}$$

where $\mathcal{F}_{\mathsf{e}}$ is (3.2.2) and

$$\mathcal{F}_{K_\pi}(\mu) = \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)^{\varepsilon} \mu K_\pi \Big/ \int_{\mathbb{R}^d} \left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)^{\varepsilon} \mathrm{d}\mu K_\pi \,.$$

Note that K_π can be an ℓ-th iterate of a π-invariant Markov kernel. As a result, if such iterates are able to effectively explore π, then we can expect that $\mu_t K_\pi$ will have high density regions similar to those of π. To a certain extent, this can potentially alleviate the issues raised in the previous section. However, using $\mu_t K_\pi$ alone in (3.2.14) does not necessarily yield the contraction (3.2.1) which comes from the fact that $\mu_t K_\pi$ is different from π. As we will see in the following, the weight inspired by (3.2.2) $(\mathrm{d}\pi/\mathrm{d}\mu)^{\varepsilon}$ acts as a correction that will ensure the contraction (3.2.1).

For notational convenience we may write $\mathcal{F}_{\mathsf{em}}(\mu)$ instead of $\mathcal{F}_{\mathsf{em}}(\mu; \lambda, K_\pi, \varepsilon)$ whenever there is no ambiguity. Let us provide a sufficient and weak condition under which the iterates

$$\mu_t := \mathcal{F}_{\mathsf{em}}(\mu_{t-1}; \lambda_t, K_\pi, \varepsilon)\,, \tag{3.2.15}$$

with $\lambda_t \in (0, 1]$ are well defined.

Lemma 3.2.3. *If $\mu \in \mathsf{M}_\pi$ is such that $\|\mathrm{d}\pi/\mathrm{d}\mu\|_\infty < \infty$ and $\lambda \in (0, 1]$ then $\mathcal{F}_{\mathsf{em}}(\mu; \lambda)$ is well defined and is a probability measure in M_π satisfying $\|\mathrm{d}\pi/\mathrm{d}\mathcal{F}_{\mathsf{em}}(\mu; \lambda)\|_\infty < \infty$.*

Proof. If $\|\mathrm{d}\pi/\mathrm{d}\mu\|_\infty < \infty$ then $\int (\mathrm{d}\pi/\mathrm{d}\mu)^{\varepsilon} \mathrm{d}\mu K_\pi < \infty$ and thus $\mathcal{F}_{\mathsf{em}}(\mu; \lambda)$ is indeed a probability measure. Next, since $\varepsilon \in [0, 1]$ and $\lambda \in (0, 1]$, we have that

$$\sup_{x \in \mathbb{R}^d} \frac{\mathrm{d}\pi}{\mathrm{d}\mathcal{F}_{\mathsf{em}}(\mu)}(x) \leq \sup_{x \in \mathbb{R}^d} \left\{ \frac{1}{\lambda} \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x)^{1-\varepsilon} \int \frac{\mathrm{d}\pi}{\mathrm{d}\mu}(y)^{\varepsilon} \mu(\mathrm{d}y) \right\} < \infty \,.$$

□

Based on the previous Lemma, starting from an initial distribution μ_0 such that $\|\mathrm{d}\pi/\mathrm{d}\mu_0\|_\infty < \infty$, we get that $\|\mathrm{d}\pi/\mathrm{d}\mu_t\|_\infty < \infty$ and in the sequel we will write $f_t(x) = \mathrm{d}\pi/\mathrm{d}\mu_t(x)$ for the Radon-Nikodym derivatives in order to simplify the notations.

We are now ready to state the first result of this chapter; we identify sufficient conditions on the kernel K_π and the sequence $(\lambda_t)_{t \in \mathbb{N}}$ under which the iterates (3.2.15) converge geometrically fast to π in backward KL.

Proposition 3.2.4. *Let $\mu_0 \in \mathsf{M}_\pi$ such that $\|\mathrm{d}\pi/\mathrm{d}\mu_0\|_\infty < \infty$ and $\beta_t \in (0, 1)$. If the kernel K_π is a π-invariant Markov kernel then the iterates (3.2.15) with*

$$\lambda_t = \begin{cases} \dfrac{\log \int f_t^{\varepsilon} \mathrm{d}\mu_t K_\pi}{\log \int f_t^{\varepsilon} \mathrm{d}\mu_t K_\pi - \log \int f_t^{\varepsilon} \mathrm{d}\mu_t}, & \text{if } \log \int f_t^{\varepsilon} \mathrm{d}\mu_t K_\pi > 0 \\ \beta_t & \text{otherwise,} \end{cases}$$

satisfy

$$\mathsf{KL}(\pi \parallel \mu_t) \leq (1 - \varepsilon)^t \mathsf{KL}(\pi \parallel \mu_0)\,.$$

Proof. If μ_t is well defined then by Jensen's inequality we have that $\int (\mathrm{d}\pi/\mathrm{d}\mu_t)^{\varepsilon} \mathrm{d}\mu_t \leq 1$ since $\varepsilon \in [0, 1]$. A straightforward induction shows that $\lambda_t \in (0, 1]$ for all $t \in \mathbb{N}$ and that μ_t is indeed a mixture of two probability measures. Next, by convexity of the KL divergence, we have that

$$\begin{aligned} \mathsf{KL}(\pi \parallel \mu_{t+1}) &\leq \lambda_t \mathsf{KL}(\pi \parallel \mathcal{F}_{\mathsf{e}}(\mu_t)) + (1 - \lambda_t) \mathsf{KL}(\pi \parallel \mathcal{F}_{K_\pi}(\mu_t)) \\ &\leq \lambda_t (1 - \varepsilon) \mathsf{KL}(\pi \parallel \mu_t) + (1 - \lambda_t)(\mathsf{KL}(\pi \parallel \mu_t K_\pi) - \varepsilon \mathsf{KL}(\pi \parallel \mu_t)) \\ &\quad + \lambda_t \log \int f_t^{\varepsilon} \mathrm{d}\mu_t + (1 - \lambda_t) \log \int f_t^{\varepsilon} \mathrm{d}\mu_t K_\pi \,. \end{aligned}$$

By the Data Processing inequality Van Erven and Harremos (2014) and the fact that π is invariant for K_π we have that

$$\mathsf{KL}(\pi \parallel \mu_t K_\pi) = \mathsf{KL}(\pi K_\pi \parallel \mu_t K_\pi) \leq \mathsf{KL}(\pi \parallel \mu_t)\,,$$

and thus

$$\mathsf{KL}(\pi \parallel \mu_{t+1}) \leq (1-\varepsilon)\mathsf{KL}(\pi \parallel \mu_t) + \lambda_t \log \int f_t^\varepsilon \mathrm{d}\mu_t + (1-\lambda_t) \log \int f_t^\varepsilon \mathrm{d}\mu_t K_\pi\,.$$

Finally, by definition of λ_t and using that $\int f_t^\varepsilon \mathrm{d}\mu_t \leq 1$ we get

$$\lambda_t \log \int f_t^\varepsilon \mathrm{d}\mu_t + (1-\lambda_t) \log \int f_t^\varepsilon \mathrm{d}\mu_t K_\pi \leq 0\,,$$

from which the contraction follows. □

In what follows we write $\mathcal{Z}_{\mu_t} = \int f_t^\varepsilon \mathrm{d}\mu_t$ and $\mathcal{Z}_{\mu_t K_\pi} = \int f_t^\varepsilon \mathrm{d}\mu_t K_\pi$. As it can be noted from the proof, the particular choice of λ_t ensures that the backward KL decreases along the iterates, taming the behavior of the second term in (3.2.14). Indeed, $\log \mathcal{Z}_{\mu_t K_\pi}$ may not always be negative; taking $K_\pi(x,\cdot) = \pi$, we get $\log \mathcal{Z}_{\mu_t K_\pi} = \varepsilon \mathcal{R}_{1+\varepsilon}(\pi \parallel \mu_t) \geq 0$ which may prevent the desired contraction (3.2.1). In fact, it is not difficult to come up with examples in the discrete case where setting $\lambda_t = 0$ for all $t \in \mathbb{N}$ results in a non-convergent scheme. Our choice of λ_t provided by Proposition 3.2.4 is the smallest parameter, when $\log \mathcal{Z}_{\mu_t K_\pi} > 0$, that ensures the contraction holds and thus maximizes the weight of the second term in (3.2.14) under this same constraint.

We may investigate the range of λ_t in the case $K_\pi(x,\cdot) = \pi$. Indeed, using that the Rényi divergence is increasing with respect to its order we get that

$$\frac{\log \mathcal{Z}_{\mu_t K_\pi}}{\log \mathcal{Z}_{\mu_t K_\pi} - \log \mathcal{Z}_{\mu_t}} = \frac{\varepsilon \mathcal{R}_{1+\varepsilon}(\pi \parallel \mu_t)}{\varepsilon \mathcal{R}_{1+\varepsilon}(\pi \parallel \mu_t) + (1-\varepsilon)\mathcal{R}_\varepsilon(\pi \parallel \mu_t)} \geq \varepsilon\,,$$

which shows that $\lambda_t \in [\varepsilon, 1]$.

Proposition 3.2.4 requires minimal assumptions regarding the choice of the Markov kernel K_π, except that it is π-invariant. Therefore there is a diverse range of options to choose from, e.g., MALA Besag (1994), HMC Neal et al. (2011) or iSIR Andrieu et al. (2010). However, we may also be interested in using kernels that are not π-invariant such as the unadjusted Langevin Algorithm (ULA). ULA corresponds to the Euler-Maruyama discretization of the Langevin diffusion

$$\mathrm{d}X_s = \nabla \log \pi(X_s)\mathrm{d}s + \sqrt{2}\mathrm{d}B_s\,, \tag{3.2.16}$$

with $(B_s)_{s\geq 0}$ a d-dimensional standard Brownian motion, and defines the discrete time Markov chain

$$X_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma} Z_{k+1}\,, \tag{3.2.17}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian vectors and γ is a fixed positive stepsize. We write R_γ for the Markov kernel corresponding to the ULA recursion: $\mathsf{R}_\gamma(x,\cdot)$ is the Gaussian distribution with mean $x + \gamma \nabla \log \pi(x)$ and covariance matrix $2\gamma \mathbf{I}_d$.

We now give guarantees for the iterates of (3.2.14) using ULA, defined by

$$\mu_t^{\mathsf{R}} = \mathcal{F}_{\mathsf{em}}(\mu_{t-1}^{\mathsf{R}}; \lambda_t^\star, \mathsf{R}_\gamma)\,. \tag{3.2.18}$$

where

$$\lambda_t^\star = \begin{cases} \dfrac{\log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}} \mathsf{R}_\gamma}{\log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}} \mathsf{R}_\gamma - \log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}}}, & \text{if } \log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}} \mathsf{R}_\gamma > 0\,, \\ \beta, & \text{otherwise,} \end{cases}$$

and $f_t = \mathrm{d}\pi/\mathrm{d}\mu_t^{\mathsf{R}}$. For this purpose, we strengthen the assumptions on the target distribution π.

(**A1**) The target π satisfies a log-Sobolev inequality; there exists $C_{\mathrm{LS}} > 0$ such that for all smooth functions $g : \mathbb{R}^d \mapsto \mathbb{R}$

$$\mathsf{Ent}_\pi(g^2) \leq C_{\mathrm{LS}} \int \|\nabla g\|^2 \mathrm{d}\pi\,. \tag{3.2.19}$$

(**A2**) The log target is continuously differentiable and L-smooth; there exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla \log \pi(x) - \nabla \log \pi(y)\| \leq L\|x - y\|\,.$$

Assumptions (**A1**) and (**A2**) are standard in the sampling literature Durmus and Moulines (2019); Vempala and Wibisono (2019); Chewi et al. (2022); Erdogdu et al. (2022). In particular, Assumption (**A1**) provides a powerful tool for analyzing the exponential convergence of Markov semi-groups; see Bakry et al. (2014). This condition holds for a large class of probability measures, including log concave distributions and is stable under bounded perturbations Holley and Stroock (1986). Combining (**A1**) and (**A2**) enables the derivation of non-asymptotic convergence bounds for ULA in the forward Kullback-Leibler divergence and more broadly, the forward α-Rényi divergence for $\alpha > 1$ Vempala and Wibisono (2019); Chewi et al. (2022); Erdogdu et al. (2022).

To ensure the existence of an invariant distribution π_γ for R_γ we consider the following drift condition Meyn and Tweedie (2012).

(**A3**) There exist $V : \mathbb{R}^d \mapsto \mathbb{R}$ such that $V \geq 1$, $b \geq 0$ and a compact set $C \subset \mathbb{R}^d$ such that

$$\mathsf{R}_\gamma V \leq V - 1 + b\mathbb{1}_C\,. \tag{3.2.20}$$

Condition (**A3**) is ''almost" a necessary condition for R_γ to admit an invariant distribution and to be ergodic. Indeed, Using (**A2**) it is easily shown that for any compact set $\mathsf{K} \subset \mathbb{R}^d$ there exists $C \geq 0$ such that for all $x \in \mathsf{K}$ and $A \in \mathcal{B}(\mathbb{R}^d)$

$$\mathsf{R}_\gamma(x, A) \geq C_{\mathsf{K}} \mathrm{Leb}(A \cap \mathsf{K})\,, \tag{3.2.21}$$

and thus R_γ is Leb-irreducible and strongly aperiodic. Therefore, by (Meyn and Tweedie, 2012, Theorem 14.0.1) R_γ is positive recurrent and admits a unique invariant distribution if and only if (3.2.20) is satisfied for some petite set C. In (**A3**), we only strengthen this condition on C and suppose that it is compact which is satisfied in most applications.

We rely on the following result which is a consequence of (Vempala and Wibisono, 2019, Theorem 1).

Proposition 3.2.5. *Assume (**A1**-**2**-**3**) and that $\gamma \in (0, 1/(2C_{\mathrm{LS}}L^2)]$. Then R_γ has a unique stationary distribution π_γ that satisfies*

$$\mathsf{KL}(\pi_\gamma \parallel \pi) \leq 4\gamma d L^2 C_{\mathrm{LS}}\,.$$

Proof. As mentionned before, (3.2.21) implies that any compact set is small for R_γ and the Lebesgue measure is an irreducibility measure. Together with (**A3**) and (Meyn and Tweedie, 2012, Theorem 14.0.1) this implies that for π_γ-a.e. $x \in \mathbb{R}^d$,

$$\lim_{\ell \to \infty} \|\mathsf{R}_\gamma^\ell(x, \cdot) - \pi_\gamma\|_{\mathsf{TV}} = 0 \,.$$

Furthermore, for all $x \in \mathbb{R}^d$ and $A \in \mathcal{B}(\mathbb{R}^d)$ such that $\mathrm{Leb}(A) > 0$, $\mathsf{R}_\gamma(x, A) > 0$ which implies that $\pi_\gamma(A) = \pi_\gamma \mathsf{R}_\gamma(A) > 0$ and thus $\pi_\gamma \gg \mathrm{Leb}$. Finally, for any $\nu \in \mathsf{M}_1$

$$\|\nu \mathsf{R}_\gamma^\ell - \pi_\gamma\|_{\mathsf{TV}} \le \int \nu(\mathrm{d}x) \|\mathsf{R}_\gamma^\ell(x, \cdot) - \pi_\gamma\|_{\mathsf{TV}} \,,$$

and by the Lebesgue dominated convergence theorem it follows that

$$\lim_{\ell \to \infty} \|\nu \mathsf{R}_\gamma^\ell - \pi_\gamma\|_{\mathsf{TV}} = 0. \tag{3.2.22}$$

In (Vempala and Wibisono, 2019, Theorem 1) it is shown that for all $\ell > 0$

$$\mathsf{KL}(\nu \mathsf{R}_\gamma^\ell \parallel \pi) \le \exp(-C_{\mathrm{LS}} \gamma \ell / 2) \mathsf{KL}(\nu \parallel \pi) + 4\gamma d L^2 C_{\mathrm{LS}} \,.$$

Thus, the lower semi-continuity of the KL for the weak topology (Van Erven and Harremos, 2014, Theorem 19) and the convergence in total variation distance (3.2.22) imply that

$$\mathsf{KL}(\pi_\gamma \parallel \pi) \le \liminf_{\ell \to \infty} \mathsf{KL}(\nu \mathsf{R}_\gamma^\ell \parallel \pi) \le 4\gamma d L^2 C_{\mathrm{LS}} \,.$$

where ν is such that $\mathsf{KL}(\nu \parallel \pi) < \infty$. □

Theorem 3.2.6. *Assume (**A1-2-3**). Let $\mu_0^{\mathsf{R}} \in \mathsf{M}_\pi$ such that $\|\mathrm{d}\pi / \mathrm{d}\mu_0^{\mathsf{R}}\|_\infty < \infty$ and let $\gamma \in (0, 1/(2C_{\mathrm{LS}} L^2)]$. The iterates (3.2.18) satisfy*

$$\|\pi - \mu_t^{\mathsf{R}}\|_{\mathsf{TV}} \le (1 - \varepsilon)^{t/2} \sqrt{2\mathsf{KL}(\pi_\gamma \parallel \mu_0^{\mathsf{R}})} + 4\sqrt{2\gamma d L^2 C_{\mathrm{LS}}} \,.$$

Theorem 3.2.6 gives quantitative convergence bounds for the iterates (3.2.18) in the total variation distance. In contrast to (3.2.4), the sequence $(\mu_t^{\mathsf{R}})_t$ does not converge to π. This comes from the fact that the stationary distribution of ULA π_γ is different from π which introduces a bias at each iteration. However, Theorem 3.2.6 establishes that the accumulated error is bounded by $\sqrt{\gamma d L^2 C_{\mathrm{LS}}}$. This type of result is typical in studies of ULA, and as a result, for a precision δ it is enough to take $\gamma \le \delta^2 / 8 d L^2 C_{\mathrm{LS}}$ and $t \ge \log\left(8\mathsf{KL}(\pi_\gamma \parallel \mu_0^{\mathsf{R}})/\delta^2\right) / |\log(1 - \varepsilon)|$ to obtain $\|\pi - \mu_t^{\mathsf{R}}\|_{\mathsf{TV}} \le \delta$. It is worth noting that we do not establish a bound in KL divergence. Since μ_t^{R} is obtained by initializing ULA at μ_{t-1}^{R} doing so would require the control of the discretization error of ULA with initial distribution μ_t^{R} in Rényi divergence through Girsanov's Theorem and comes at the cost of much stronger assumptions on the tails of each iterate μ_t^{R}. Here we show convergence under a weaker measure of discrepancy by only requiring mild conditions on the initial measure μ_0^{R}.

Proof. By convexity of the KL divergence and the definition of $\lambda_t^\star$ (3.2.4),

$$\begin{aligned}
\mathsf{KL}(\pi_\gamma \parallel \mu_{t+1}^{\mathsf{R}}) &\le \lambda_t^\star \int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d} f_t^\varepsilon \mu_t^{\mathsf{R}}} \mathrm{d}\pi_\gamma + (1 - \lambda_t^\star) \int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d} f_t^\varepsilon \mu_t^{\mathsf{R}} \mathsf{R}_\gamma} \mathrm{d}\pi_\gamma \\
&\quad + \lambda_t^\star \log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}} + (1 - \lambda_t^\star) \log \int f_t^\varepsilon \mathrm{d}\mu_t^{\mathsf{R}} \mathsf{R}_\gamma \\
&\le \lambda_t^\star \int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d} f_t^\varepsilon \mu_t^{\mathsf{R}}} \mathrm{d}\pi_\gamma + (1 - \lambda_t^\star) \int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d} f_t^\varepsilon \mu_t^{\mathsf{R}} \mathsf{R}_\gamma} \mathrm{d}\pi_\gamma \,,
\end{aligned}$$

and we have that

$$\int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d}f_t^\varepsilon \mu_t^{\mathsf{R}}} \mathrm{d}\pi_\gamma = (1-\varepsilon)\mathsf{KL}(\pi_\gamma \parallel \mu_t^{\mathsf{R}}) + \varepsilon \mathsf{KL}(\pi_\gamma \parallel \pi)\,,$$
$$\int \log \frac{\mathrm{d}\pi_\gamma}{\mathrm{d}f_t^\varepsilon \mu_t^{\mathsf{R}} \mathsf{R}_\gamma} \mathrm{d}\pi_\gamma \leq (1-\varepsilon)\mathsf{KL}(\pi_\gamma \parallel \mu_t^{\mathsf{R}}) + \varepsilon \mathsf{KL}(\pi_\gamma \parallel \pi)\,,$$

where the second inequality follows using the Data Processing inequality and the fact that R_γ is π_γ-invariant. Thus,

$$\mathsf{KL}(\pi_\gamma \parallel \mu_{t+1}^{\mathsf{R}}) \leq (1-\varepsilon)\mathsf{KL}(\pi_\gamma \parallel \mu_t^{\mathsf{R}}) + \varepsilon \mathsf{KL}(\pi_\gamma \parallel \pi) \leq (1-\varepsilon)^t \mathsf{KL}(\pi_\gamma \parallel \mu_0^{\mathsf{R}}) + \mathsf{KL}(\pi_\gamma \parallel \pi)\,.$$

The desired bound follows by Pinsker's inequality and Proposition 3.2.5,

$$\begin{aligned}\|\pi - \mu_t^{\mathsf{R}}\|_{\mathsf{TV}} &\leq \|\pi - \pi_\gamma\|_{\mathsf{TV}} + \|\pi_\gamma - \mu_t^{\mathsf{R}}\|_{\mathsf{TV}} \\ &\leq (1-\varepsilon)^{t/2}\sqrt{2\mathsf{KL}(\pi_\gamma \parallel \mu_0^{\mathsf{R}})} + 2\sqrt{2\mathsf{KL}(\pi_\gamma \parallel \pi)}\,.\end{aligned}$$

□

Algorithm 1: Entropic Mirror Monte Carlo (`EM2C`)

Input: μ_0, ε, N, global kernel K_π, local kernel L_π, family F_Θ, number of iterations T.
Output: μ_{θ_T}

for $t \in [0 : T-1]$ **do**
- Sample $X_t^{1:N} \overset{\text{iid}}{\sim} \mu_{\theta_t}$
- Set $\widetilde{\omega}_t^i = (\mathrm{d}\pi/\mathrm{d}\mu_{\theta_t})(X_t^i)^\varepsilon$ and $\omega_t^i = \widetilde{\omega}_t^i / \sum_{j=1}^N \widetilde{\omega}_t^j$.
- Sample $Y_t^{1:N} \sim K_\pi(X_t^1, \cdot) \otimes \cdots \otimes K_\pi(X_t^N, \cdot)$
- Set $\widetilde{\varpi}_t^i = (\mathrm{d}\pi/\mathrm{d}\mu_{\theta_t})(X_t^i)^\varepsilon$ and $\varpi_t^i = \widetilde{\varpi}_t^i / \sum_{j=1}^N \widetilde{\varpi}_t^j$
- **if** $N^{-1}\sum_{i=1}^N \widetilde{\varpi}_t^i > 1$ **then**
  - Set $\lambda_{t+1,N}^\star = \log\left(N^{-1}\sum_{i=1}^N \widetilde{\varpi}_t^i\right) / \left\{\log\left(\sum_{i=1}^n \widetilde{\varpi}_t^i\right) - \log\left(\sum_{i=1}^n \widetilde{\omega}_t^i\right)\right\}$
- **else**
  - Set $\lambda_{t+1,N}^\star = \beta$
- Sample $Z_t^{1:N} \overset{\text{iid}}{\sim} \left\{\lambda_{t,N}^\star \sum_{i=1}^N \omega_t^i \delta_{X_t^i} + (1-\lambda_{t+1,N}^\star)\sum_{i=1}^N \varpi_t^i \delta_{Y_t^i}\right\} L_\pi$
- Set $\mu_{\theta_{t+1}} = \operatorname{argmin}_{\mu \in \mathsf{F}_\Theta} -N^{-1}\sum_{i=1}^N \log \mu(Z_t^i)$

3.2.3 Stochastic updates

We now detail the practical implementation of the sequence (3.2.14) with the criterion (3.2.5). The complete procedure is summarized in Algorithm 1. Given $\mu_{\theta_t} \in \mathsf{M}_\pi$ the update $\mathcal{F}_{\mathsf{em}}(\mu_{\theta_t}; \lambda_{t+1}^\star)$ is approximated as follows. The first step consists in obtaining an empirical approximation of the update by drawing N samples $X_t^{1:N}$ from μ_{θ_t} and then sampling $Y_t^i \sim K_\pi(X_t^i, \cdot)$ for $i \in [1:n]$. These samples are then used to estimate the intractable mixture weight $\lambda_{t+1}^\star$ (3.2.4) with $\lambda_{t+1,N}^\star$ given by

$$\lambda_{t+1,N}^\star := \begin{cases} \dfrac{\log N^{-1}\sum_{i=1}^N f_t(Y_t^i)^\varepsilon}{\log \sum_{i=1}^N f_t(Y_t^i)^\varepsilon - \log \sum_{i=1}^N f_t(X_t^i)^\varepsilon}\,, & \text{if } \log N^{-1}\sum_{i=1}^N f_t(Y_t^i)^\varepsilon > 0\,, \\ \beta_t & \text{otherwise.} \end{cases} \tag{3.2.23}$$

where $f_t : x \mapsto \mathrm{d}\pi/\mathrm{d}\mu_{\theta_t}(x)$. An empirical version of $\mathcal{F}_{\mathsf{em}}(\mu_{\theta_t}; \lambda^\star_{t+1,N})$ is then

$$\zeta_t^N = \lambda^\star_{t+1,N} \sum_{i=1}^N \omega_t^i \delta_{X_t^i} + (1 - \lambda^\star_{t+1,N}) \sum_{i=1}^N \varpi_t^i \delta_{Y_t^i}\,, \tag{3.2.24}$$

where $\omega_t^i = f_t(X_t^i)^\varepsilon / \sum_{j=1}^n f_t(X_t^j)^\varepsilon$ and $\varpi_t^i = f_t(Y_t^i)^\varepsilon / \sum_{j=1}^n f_t(Y_t^j)^\varepsilon$. In comparison with $f_t(X_t^i)$, the unnormalized weights $f_t(Y_t^i)$ are not classical importance weights since Y_t^i is sampled from $\mu_{\theta_t} K_\pi$. Consequently, if $\mu_{\theta_t} K_\pi$ has non-negligible mass in the effective support of π where μ_{θ_t} does not, it is expected that the distribution of $f_t(Y)^\varepsilon$ with $Y \sim \mu_{\theta_t} K_\pi$ exhibits heavier tails than $f_t(X)^\varepsilon$ when $X \sim \mu_{\theta_t}$. This provides an intuitive interpretation of (3.2.24). The first component of the mixture $\sum_{i=1}^N \omega_t^i \delta_{X_t^i}$ adapts the samples from μ_{θ_t} to the target π through the importance weights raised to ε, thus providing a *more biased* particle approximation of π with *less variance*, see Korba and Portier (2022). This also helps avoiding the situation where one sample has a normalized weight close to 1, which often happens in large dimensions. The second component $\sum_{i=1}^N \varpi_t^i \delta_{Y_t^i}$ on the otherhand corresponds to a greedy exploration step; the samples Y_t^i that land in regions of the effective support of π where μ_{θ_t} is unlikely are assigned a large unnormalized weight. The counterpart of this greedy approach is that, more often than not, one sample will have a normalized weight close to 1. This an undesirable feat as it may result in poor parametric approximations. To circumvent this issue and benefit from such samples, we add a *resample-move* step; we sample from the mixture (3.2.24) and diversify the samples by running a local MCMC algorithm with kernel L_π leaving π invariant, such as the Random walk Metropolis or MALA. Formally, this additional step boils down to sampling from $\zeta_t^N L_\pi$; first sample $A_t^{1:N} \overset{\text{iid}}{\sim} \zeta_t^N$ and then sample $Z_t^1, \ldots, Z_t^N \sim L_\pi(X_t^1, \cdot) \otimes \ldots L_\pi(X_t^N, \cdot)$. In Example (3.2.8) we illustrate the benefits of this additional step.

The scheme we have just described can be thought of as an approximation of the updates

$$\mu_{t+1}(\mathrm{d}x) = \int \mathcal{F}_{\mathsf{em}}(\mu_t; \lambda^\star_{t+1}, K_\pi)(\mathrm{d}y) L_\pi(y, \mathrm{d}x)\,. \tag{3.2.25}$$

Since L_π is π-invariant, by the Data Processing inequality, the KL contraction in Proposition 3.2.4 still holds. The bound in Theorem 3.2.6 also holds by applying the Data Processing inequality for the total variation distance.

Remark 3.2.7. *We can further improve our algorithm by allowing sampling from some heavy tailed distribution h_t (which may eventually depend on the iteration number) and considering instead the iterates*

$$\mu_{t+1}(\mathrm{d}x) = \int \{\lambda^\star_{t+1} \cdot \mathcal{F}_{\mathsf{e}}(\mu_t) + (1 - \lambda^\star_{t+1}) \cdot \mathcal{F}_{K_\pi}(\tilde{\mu}_t)\}(\mathrm{d}y) L_\pi(y, \mathrm{d}x)\,,$$

where $\tilde{\mu}_t = \alpha_t \mu_t + (1-\alpha_t) h_t$ and $\alpha_t \in [0,1]$. Mixing the iterate μ_t with a heavy tailed distribution allows for a better exploration of π observed numerically at the cost of a bias in the bounds of Proposition 3.2.4, Theorem 3.2.6 (this may aswell be an artifact of the proof techniques).

Example 3.2.8. *Assume that*

$$\pi = 0.8 \cdot \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2) + 0.2 \cdot \mathcal{N}\left(12 \cdot \mathbf{1}_2, \begin{bmatrix} 6 & -5 \\ -5 & 5 \end{bmatrix}\right)$$

and $K_\pi = \mathsf{R}_\gamma^\ell$ and L_π is the composition of ℓ random walk metropolis steps. Both are run with step size $\gamma = 0.05$ and $\ell = 100$. The initial distribution μ_0 is a standard Gaussian. In Figure 3.2.8 we display the effect of the resample-move step described above.

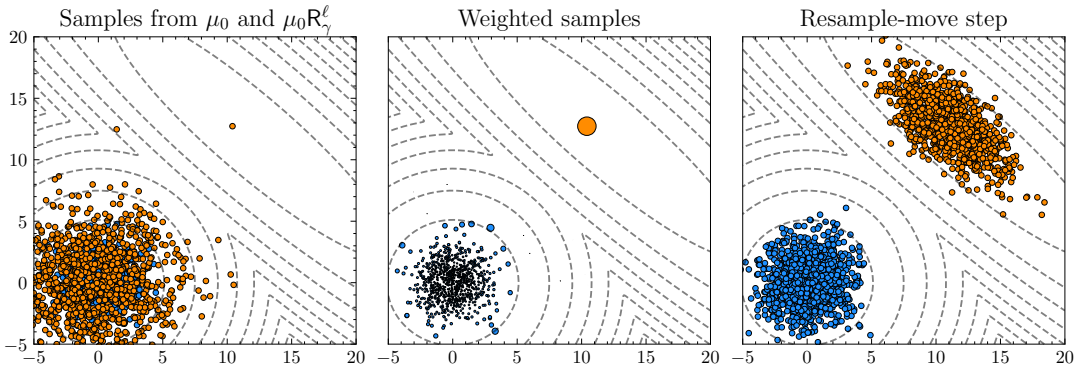

Figure 3.3: Left: samples from μ_0 (in blue) and samples from $\mu_0 K_\pi$ (in orange). Middle: same samples weighted according to $\{\omega_1^i\}_{i=1}^n$ (in blue) and $\{\varpi_1^i\}_{i=1}^n$ (in orange). The size of each dot is proportional to the weight of the associated sample. Right: samples from $\zeta_t^n L_\pi$ (3.2.24). The blue and orange samples come from the first and second component of (3.2.24) respectively. The contour plot of $\log \pi$ is in dashed lines.

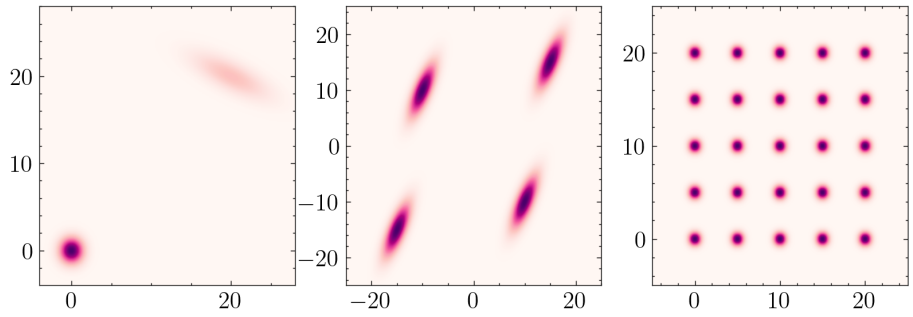

Figure 3.4: The $\widetilde{\pi}_i$ associated with the target distributions.

3.3 Numerical experiments

In this section we assess the performance of our algorithm on three different synthetic examples. We would like to stress that these are only preliminary experiments. For each example we compute the normalizing constant $\mathcal{Z} = \int \pi(\mathrm{d}x) = 1$ using importance sampling and estimate the inclusive KL using exact samples from π.

We test the robustness of our method with respect to: (i) initial distribution μ_0, (ii) dimension d of the ambient space and (iii) model specification. In all experiments we use $\varepsilon = 0.5$ and the kernel K_π is the composition of 40 of `ULA` with step-size $\gamma_k = 1/k^{0.2}$. The kernel L_π is the composition of 100 steps of random walk Metropolis. We implement the scheme described in Remark 3.2.7 with $h_t = \text{Student}(\mathbf{0}_d, \mathbf{I}_d)$ using the KL projection (3.2.5). N is set to 1000 in each experiment and we compute the normalizing constants using $M = 5000$ samples.

The target distributions considered factorize as $\pi_i = \widetilde{\pi}_i^{\otimes d/2}$ and thus we only specify $\widetilde{\pi}_i$ in the next examples. The targets $\widetilde{\pi}_{1:3}$ are shown in Figure 3.4.

(1) Two component GM with unbalanced modes (`GM2`).

$$\widetilde{\pi}_1 = 0.20 \cdot \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2) + 0.80 \cdot \mathcal{N}(20 \cdot \mathbf{1}_2, \begin{bmatrix} 10 & -4 \\ -4 & 3 \end{bmatrix}).$$

(2) Four component anisotropic GM (`GM4`).

$$\widetilde{\pi}_2 = 0.25 \sum_{i=1}^{4} \mathcal{N}(m_i, \begin{bmatrix} 3 & 4 \\ 4 & 10 \end{bmatrix}),$$

where $m_1 = (-10, 10), m_2 = (10, -10), m_3 = (15, 15), m_4 = (-15, -15)$.

(3) Twenty five component GM (`GM25`).

$$\widetilde{\pi}_3 = 0.2 \sum_{\ell=0}^{4} \sum_{k=0}^{4} \mathcal{N}(\begin{bmatrix} 5\ell \\ 5k \end{bmatrix}, 0.25 \cdot \mathbf{I}_2).$$

We denote by GM_ℓ the family of Gaussian mixtures with ℓ components and full covariance matrices, $\mathsf{GM}_\ell = \{\sum_{i=1}^{\ell} w_i \mathcal{N}(m_i, \Sigma_i) : w_{1:\ell}, m_{1:\ell}, \Sigma_{1:\ell} \in \mathcal{S}_\ell \times \mathbb{R}^{d\times\ell} \times (\mathsf{S}_d^{++})^\ell\}$, where $\mathcal{S}_\ell$ is the ℓ-simplex and S_d^{++} is the set of $d \times d$ positive definite matrices. We use the *Expectation-Maximization* (`EM`) algorithm for the maximum likelihood step with Gaussian mixtures in step 11 of Algorithm 1.

We first investigate if our parametric approximations are able to learn unbalanced targets **(1)**. We let $\mathsf{F}_\Theta = \mathsf{GM}_2$. Let $(w_0^*, w_{20}^*, m_0^*, m_{20}^*, \Sigma_0^*, \Sigma_{20}^*)$ denote the parameters output by `EM2C` where $(w_0^*, m_0^*, \Sigma_0^*)$ (resp. $(w_{20}^*, m_{20}^*, \Sigma_{20}^*)$) are the weight, mean and covariance matrix associated with the component that is the closest to that with mean $\mathbf{0}_d$ (resp $20 \cdot \mathbf{1}_d$). The results are given in Table 3.1 for `GM2` where we write D_m for $\|m_0^* + m_{20}^* - 20 \cdot \mathbf{1}_d\|_2$ and D_Σ for $\|(\Sigma_0^* + \Sigma_{20}^*) - (\mathbf{I}_2 + \Sigma)\|_2$. Remarkably, even in $d = 100$ `EM2C` is able to recover the correct weight of the mixture while also being very close in mean and covariance. In Table 3.2 we provide the results for the estimation of $\mathcal{Z}$.

Next, we consider the example `GM4` and test the robustness of `EM2C` to mispecification of the parametric family. We learn parametric approximations within the families GM_2 (underspecified), GM_4 (perfectly specified), GM_8 (overspecified). Here the point of comparison is the `EM` algorithm; we compare the quality of the approximations obtained by `EM2C` to those obtained with `EM` using $N = 10000$ *exact* samples from π. The results are given in Figure 3.5. The dashed

Table 3.1: Bias correction step. Results are given in mean ± standard deviation format

	$d = 10$	$d = 50$	$d = 100$
w_0^*	0.19 ± 0.01	0.19 ± 0.01	0.19 ± 0.01
D_m	0.11 ± 0.07	0.59 ± 0.28	0.60 ± 0.30
D_Σ	1.10 ± 0.25	10.15 ± 1.50	7.89 ± 4.71

Table 3.2: `GM2` example. Estimates of $\mathcal{Z}$ computed with $N = 5000$ samples. Results are given in mean ± standard deviation format

	$d = 20$	$d = 50$	$d = 100$
GM_2	1.00 ± 0.00	1.00 ± 0.02	1.00 ± 0.10

lines represent the median value obtained by `EM` for each example and dimension d, over a total of 30 runs. In this example, the initial distribution for `EM2C` is set to be a standard Gaussian with mean $-10 \cdot \mathbf{I}_d$. Notably, in higher dimensions `EM2C` is competitive with `EM` which uses real samples from the target. The results for the estimation of $\mathcal{Z}$ are given in Table 3.3.

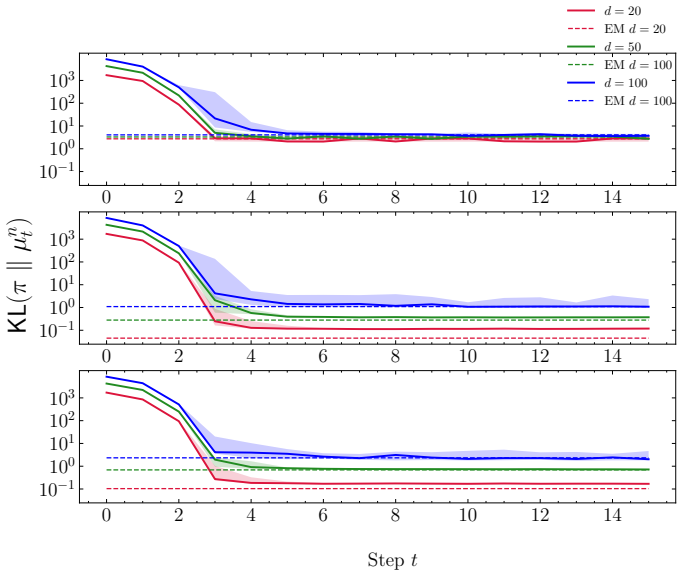

Figure 3.5: `GM4` example. Top to bottom: `EM2C` and `EM` with variational families GM_2, GM_4 and GM_8.

Finally, we consider the most challenging example, `GM25`. We run `EM2C` on the variational family GM_{10} and then use the resulting variational approximation as a warm start in a MALA. We run 20000 parallel chains with 5000 steps each and then use the obtained samples to perform maximum likelihood within the family GM_{25} and also with an `RNVP` Dinh et al. (2016). Strikingly, under perfect specification, we are able to achieve a KL of 1.13 in $d = 50$ and accurate normalizing constant estimates. For the `RNVP`, we do not match the performance of the Gaussian mixture but this is to be expected, since normalizing flows inherently struggle with multimodal distributions Cornish et al. (2020).

Table 3.3: `GM4` example. Estimates of $\mathcal{Z}$ computed with $M = 5000$ samples. Results are given in mean $\pm$ standard deviation format

	$d = 20$	$d = 50$	$d = 100$
GM_2	1.00 ± 0.09	1.00 ± 0.07	1.00 ± 0.25
GM_4	1.00 ± 0.00	1.00 ± 0.02	1.00 ± 0.24
GM_8	1.00 ± 0.00	1.00 ± 0.02	0.90 ± 0.23

	$d = 20$	$d = 50$
GM_{10}	0.99 ± 0.05	1.03 ± 0.13
GM_{25}	1.00 ± 0.01	1.01 ± 0.12
`RNVP`	0.99 ± 0.23	0.97 ± 0.63

	$d = 20$	$d = 50$
GM_{10}	2.42 ± 0.20	3.44 ± 0.23
GM_{25}	0.25 ± 0.05	1.13 ± 0.06
`RNVP`	3.24 ± 0.25	5.78 ± 0.94

Table 3.4: `GM25` example. Left table: normalizing constant estimates. Right table: inclusive KL value. The results are given in mean $\pm$ standard deviation format, obtained over 100 seeds. The estimates are computed with $M = 5000$.

3.4 Conclusion and perspectives

We have presented a novel sequence that extends the entropic mirror descent sequence by incorporating Markovian moves. We have shown that it converges geometrically fast to the target distribution in KL divergence under mild assumptions, mainly the invariance with respect to π of the Markov kernel. For ULA, which converges to a biased limit, we showed that the iterates converge geometrically fast to π up to a bias explicit in the smoothness and log-Sobolev constants of π, the step size and the dimension. We have derived a principled objective to minimize aswell as a pratical stochastic scheme. Current numerical experiments show that the proposed algorithm performs well on difficult normalizing constant estimation tasks.

We believe that a few fundamental questions remain to be answered. First, do the projected iterates (3.2.5) or (3.2.9) converge towards some biased limit? If yes, can we characterize this limit, i.e. is it for example $\operatorname{argmin}_{\mu \in \mathsf{F}_\Theta} \mathsf{KL}(\pi \parallel \mu)$, which is what we are interested in? Finally, the ultimate goal is to be able to obtain bounds in backward KL divergence between π and the iterates of the stochastic scheme. For now we have no clear path on how this can be achieved.

Chapter 4

Variance estimation for SMC algorithms: a backward sampling approach

4.1 Introduction

Sequential Monte Carlo (SMC) methods offer a flexible framework for the approximation of posterior distributions in the context of Bayesian inference, for instance in *Hidden Markov Models* (HMM). These models presuppose that the observations are defined using an unobserved process assumed to be a Markov chain. In such a setting, we are particularly interested in estimating the law of a hidden state given all past observations referred to as the filtering distribution and the laws of sequences of states given past and future observations, referred to as smoothing distributions. These distributions can be approximated by weighted empirical measures associated with random samples, usually known as *particles*. All SMC methods are based on successive importance sampling and resampling steps. When a new observation is available, new particles are sampled according to an importance distribution and then weighted to match the target distribution. Finally, through a resampling scheme, particles with large weights are duplicated while low weighted particles are discarded. This general procedure has been used in a wide range of applications such as signal processing, target tracking, econometrics, biology, see Cappe et al. (2005); Douc et al. (2014); Chopin and Papaspiliopoulos (2020) and the references therein.

The quantification of the Monte Carlo error of SMC estimators is a major challenge. For a variety of SMC methods such as the *bootstrap filter* Gordon et al. (1993a) or the *Forward Filtering Backward Smoothing* (FFBS) Tanizaki and Mariano (1994) algorithms, *Central Limit Theorems* (CLT) with theoretical expressions of the asymptotic variance (in the number of particles) have been derived Del Moral and Guionnet (1999); Chopin (2004); Künsch (2005); Douc et al. (2011a). However, these expressions are not computable in practice and give rise to the natural question of their estimation. Since this problem appears in an online context, a critical constraint is that the samples produced by the original SMC algorithm should be recycled to compute such estimates.

This problem has received some satisfactory solutions recently: a simple estimator of the asymptotic variance when multinomial resampling is used has been proposed in Chan and Lai (2013) for the bootstrap filter algorithm and has since been refined in Lee and Whiteley (2018); Olsson and Douc (2019); Du and Guyader (2021). The computation of the associated asymptotic

variance estimator at time t is based on tracing the genealogy of each particle down to time 0. Although it has been shown to be consistent as the number of particles N grows to infinity, it is particularly prone to instability when t is large, as the successive resampling steps lead to the well known path degeneracy issue. After a few time steps, particles are likely to share the same ancestor at time 0 which in turn makes the asymptotic variance estimates collapse. To overcome this degeneracy issue, Olsson and Douc (2019) proposed to only trace a part of the genealogy of each particle according to a fixed-lag parameter following the fixed-lag smoothing approach introduced in Olsson et al. (2008). As long as the number of particles is balanced with the chosen lag, the bias introduced by considering only the most recent ancestors of each particle can be controlled as shown in Olsson and Douc (2019). However, while this alternative estimator remains easily computable, choosing an optimal lag is a non trivial task which makes this approach hard to tune in practice. We thus address the limitations of current asymptotic variance estimators in this chapter.

The first contribution of this chapter is to propose a parameter free estimator of the asymptotic variance associated with the bootstrap particle filter with multinomial resampling that trades computational cost for stability and reduced variance. The construction of our estimator starts from the observation that the aforementioned degeneracy is similar to that of classical SMC-based smoothing algorithms Doucet et al. (2000); Fearnhead et al. (2010); Poyiadjis et al. (2011) or Particle Markov Chain Monte Carlo algorithms such as the Particle Gibbs sampler Andrieu et al. (2010). In both cases, a backward sampling step which aims at diversifying the particle trajectories has shown to be a reliable workaround that decreases the (theoretical) variance of the estimators at the expense of higher computational cost Godsill et al. (2004); Douc et al. (2011a); Olsson and Westerborn (2017); Del Moral et al. (2010b); Lindsten and Schön (2012); Chopin and Singh (2015a). We thus aim at introducing such a mechanism in the estimation of the asymptotic variance. The construction of our estimator relies on the analysis conducted in Lee and Whiteley (2018) in which it is shown that the estimator of Chan and Lai (2013) can be interpreted as a conditional expectation with respect to the indices that retrace the genealogy of the particles, given all the particles and ancestors. We show that this construction still holds when the distribution of the indices relies on the backward importance weights. The resulting estimator is computed by averaging auxiliary statistics that are very similar to those of the *forward implementation* of the FFBS for additive functionals Del Moral et al. (2010c) and can be thus updated online. The time complexity per update of our estimator is of order N^3. Driven by the efficient implementation of the FFBS for additive functionals developed in Olsson and Westerborn (2017), we show that the computational cost of our estimator can be reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ by means of additional Monte Carlo simulation while remaining as competitive in terms of bias and variance.

We next focus on the FFBS algorithm for the estimation of smoothing estimators. Despite the fact that a CLT has been obtained for estimators based on the FFBS, no variance estimator has been proposed in the literature. We show that our previous construction enables us to fill this gap and we thus provide a consistent estimator in the case of additive functionals which are particularly critical, for instance in the Expectation Maximization framework. Again, this estimator can be computed online and in the particular case of marginal smoothing its computational cost can be drastically reduced.

The chapter is organized as follows. In Section 4.3 we briefly review the SMC framework and discuss the current estimators of the asymptotic variance proposed in Chan and Lai (2013); Lee and Whiteley (2018); Olsson and Douc (2019). In Section 4.4, we introduce our estimator based on the backward weights, propose an online implementation and establish its asymptotic properties. In Section 4.5, we extend our derivations to the FFBS algorithm and provide

a consistent asymptotic variance estimator. We finally validate our results with numerical experiments in Section 4.6. Notably, we show empirically that our novel estimator for the filter has a favourable dependence on the time horizon t in comparison with the existing estimators. All additional proofs and discussions can be found in the appendix.

4.2 Notation

For any measurable space $(\mathsf{E}, \mathcal{E})$, we denote by $\mathsf{F}(\mathcal{E})$ the set of $\mathbb{R}$ valued, $\mathcal{E}$-measurable functions, by $\mathsf{F}_b(\mathcal{E})$ the subset of $\mathsf{F}(\mathcal{E})$ of bounded functions on E, and by $\mathsf{M}_1(\mathsf{E})$ the set of measures on E. For any $\mu \in \mathsf{M}_1(\mathrm{E})$ and $h \in \mathsf{F}(\mathcal{E})$, we write

$$\mu(h) := \int_{\mathsf{E}} h(x)\mu(\mathrm{d}x)\,.$$

For any transition kernel M from $(\mathsf{E}, \mathcal{E})$ to another measurable space $(\mathsf{G}, \mathcal{G})$, define

$$M[h](x) := \int_{\mathsf{G}} M(x, \mathrm{d}y)h(y)\,, \quad \forall h \in \mathsf{F}(\mathcal{G})\,, \quad \forall x \in \mathsf{E}\,,$$

and write μM the measure defined on G by

$$\mu M(A) := \int_{\mathsf{E}} \mu(\mathrm{d}x)M(x, A)\,, \quad \forall A \in \mathcal{G}\,.$$

If M_1 is a transition kernel from $(\mathsf{E}, \mathcal{E})$ to $(\mathsf{G}, \mathcal{G})$ and M_2 a transition kernel from $(\mathsf{G}, \mathcal{G})$ to a measurable space $(\mathsf{H}, \mathcal{H})$, then $M_1 M_2$ is the transition kernel from $(\mathsf{E}, \mathcal{E})$ to $(\mathsf{H}, \mathcal{H})$ defined by

$$M_1 M_2(x, A) = \int_{\mathsf{G}} M_1(x, \mathrm{d}y)M_2(y, A)\,, \quad \forall x \in \mathsf{E}\,, \quad \forall A \in \mathcal{H}\,.$$

In addition, $M_1 \otimes M_2$ is the transition kernel from $(\mathsf{E}, \mathcal{E})$ to $(\mathsf{G} \times \mathsf{H}, \mathcal{G} \otimes \mathcal{H})$ defined by

$$M_1 \otimes M_2(x, A) := \int \mathbb{1}_A(y, z)M_1(x, \mathrm{d}y)M_2(y, \mathrm{d}z)\,, \quad \forall x \in \mathsf{E}, \quad A \in \mathcal{G} \otimes \mathcal{H}\,.$$

In particular, for all $N \geq 1$, we will write $M^{\otimes N}$ for $\bigotimes_{i=1}^{N} M$. For two $\mathcal{E}$-measurable functions f, g the tensor product is defined as

$$f \otimes g : \mathsf{E}^2 \ni (x, y) \mapsto f(x)g(y)\,.$$

The sets $\mathbb{N}$ and $\mathbb{N}_{>0}$ are respectively the sets of natural numbers and positive natural numbers. The $\mathbf{L}_q$ norm of a random variable X is $\|X\|_q := \mathbb{E}\big[|X|^q\big]^{1/q}$. The supremum norm of $f \in \mathsf{F}_b(\mathsf{E})$ is denoted by $|f|_\infty$. The unit function $\mathbf{1}$ is such that $\mathbf{1}(x) = 1$ for all $x \in \mathsf{E}$. The transpose of a matrix A is denoted $A^\top$. If $A, B \in \mathbb{R}^{M \times N}$ are two matrices, then the Hadamard product $A \odot B$ is the element-wise product, i.e for all $1 \leq i \leq M$ and $1 \leq j \leq N$, $(A \odot B)_{i,j} = a_{i,j}b_{i,j}$. If $A \in \mathbb{R}^{N \times N}$, then $\mathrm{Diag}(A)$ is the $N \times N$ diagonal matrix such that for all $1 \leq i \leq N$, $\mathrm{Diag}(A)_{i,i} = A_{i,i}$ and if $\boldsymbol{x} \in \mathbb{R}^{N \times 1}$ then $\mathrm{Diag}(\boldsymbol{x})$ is the $N \times N$ diagonal matrix such that $\mathrm{Diag}(\boldsymbol{x})_{i,i} = \boldsymbol{x}_i$. For $(a, b) \in \mathbb{N}^2$, $[a : b] := \mathbb{N} \cap [a, b]$ and $[b] := [1 : b]$. If f is a mapping from $[N]^2$ to $\mathbb{R}$, we denote by $\boldsymbol{f}$ the associated $N \times N$ matrix such that $\boldsymbol{f}_{i,j} = f(i, j)$. Finally, we adopt the following conventions. Given some set $\{\xi_s^i\}_{s \in [0:t], i \in [N]}$, we write $\xi_t^{1:N} := (\xi_t^1, \cdots, \xi_t^N)$, $\xi_{0:t}^{k_{0:t}} := (\xi_0^{k_0}, \cdots, \xi_t^{k_t})$, $\boldsymbol{\xi}_{0:t}^{1:N} := \{\xi_{0:t}^{k_{0:t}}\}_{k_{0:t} \in [N]^{t+1}}$.

4.3 Sequential Monte Carlo

In this section, we first review the bootstrap particle filter methodology and recall the main asymptotic results associated with the estimates produced by this algorithm. A state of the art and a discussion of the estimation of the asymptotic variances related to this algorithm are also presented.

4.3.1 Definitions

Let $(\mathsf{X}, \mathcal{X})$ be a general measurable space. Let M_0 and $(M_t)_{t\in\mathbb{N}}$ be a probability measure on $(\mathsf{X}, \mathcal{X})$ and a sequence of Markov transition kernels on $\mathsf{X} \times \mathcal{X}$, respectively. Consider also a family $(g_t)_{t\in\mathbb{N}}$ of non-negative $\mathcal{X}$-measurable functions, referred to as potentials. Throughout this chapter, we make the following assumptions on $\{M_t\}_{t\in\mathbb{N}}$ and $\{g_t\}_{t\in\mathbb{N}}$.

(**A4**) The probability measure M_0 admits m_0 as probability density with respect to some reference measure $\nu \in \mathsf{M}_1(\mathcal{X})$. For all $t \in \mathbb{N}$ and $x_t \in \mathsf{X}$, $M_{t+1}(x_t, .)$ admits $m_{t+1}(x_t, .)$ as probability density with respect to ν.

(**A5**) There exists a constant $G_\infty > 0$ such that for all $t \in \mathbb{N}$ and $x \in \mathsf{X}$, $0 < g_t(x) \leq G_\infty$.

Define the sequence of unnormalized transition kernels $(\mathbf{Q}_{t+1})_{t\in\mathbb{N}}$ where, for all $t \in \mathbb{N}$, $x_t \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\mathbf{Q}_{t+1}(x_t, A) := g_t(x_t) M_{t+1}(x_t, A) \,,$$

and, for any $s, t \in \mathbb{N}^2$,

$$\mathbf{Q}_{s:t} := \begin{cases} \mathbf{Q}_s \otimes \cdots \otimes \mathbf{Q}_t & \text{if } s \leq t \,, \\ \mathrm{Id} & \text{otherwise} \,. \end{cases}$$

Let $\overline{\mathbf{Q}}_{s:t}$ denote its marginal with respect to the variable x_t, i.e. for all $x_{s-1} \in \mathsf{X}$ and all measurable set A,

$$\overline{\mathbf{Q}}_{s:t}(x_{s-1}, A) := \int_{\mathsf{X}^{t-s+1}} \mathbf{Q}_{s:t}(x_{s-1}, \mathrm{d}x_s, \ldots, \mathrm{d}x_t) \mathbb{1}_A(x_t).$$

Define recursively the sequence of measures $(\gamma_{0:t})_{t\in\mathbb{N}}$ by

$$\gamma_0(\mathrm{d}x_0) := M_0(\mathrm{d}x_0) \,, \quad \gamma_{0:t}(\mathrm{d}x_{0:t}) := \gamma_{0:t-1}(\mathrm{d}x_{0:t-1}) \mathbf{Q}_t(x_{t-1}, \mathrm{d}x_t) \,, \tag{4.3.1}$$

and let $\gamma_t(A) = \int_{\mathsf{X}^{t+1}} \gamma_{0:t}(\mathrm{d}x_{0:t}) \mathbb{1}_A(x_t)$ for all measurable sets A. Sequential Monte Carlo algorithms aim at solving recursively the filtering problem, i.e. at estimating the sequence of probability measures defined as

$$\eta_t(\mathrm{d}x_t) := \gamma_t^{-1}(\mathbf{1}) \gamma_t(\mathrm{d}x_t), \quad \phi_t(\mathrm{d}x_t) := g_t(x_t) \eta_t(\mathrm{d}x_t) / \eta_t(g_t) \,, \tag{4.3.2}$$

respectively called the *predictive* and *filtering* measures. Note that η_t can be computed recursively using

$$\eta_t(\mathrm{d}x_t) = \int \phi_{t-1}(\mathrm{d}x_{t-1}) M_t(x_{t-1}, \mathrm{d}x_t) \,. \tag{4.3.3}$$

We motivate these definitions with the following example.

Example 4.3.1. *Hidden Markov models consist of an unobserved state process $\{X_t\}_{t\in\mathbb{N}}$ and observations $\{Y_t\}_{t\in\mathbb{N}}$. They respectively evolve in two general measurable spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$. It is assumed that $\{X_t\}_{t\in\mathbb{N}}$ is a Markov chain with transition kernels $(M_{t+1})_{t\in\mathbb{N}}$ and initial distribution M_0. Given the states $\{X_t\}_{t\in\mathbb{N}}$, the observations $\{Y_t\}_{t\in\mathbb{N}}$ are independent and for all $t \in \mathbb{N}$, the conditional distribution of the observation Y_t only depends on the current*

state X_t. This distribution is written $G_t(X_t, .)$ and admits the potential $g_t(x_t, .)$ as density (the dependency in Y_t is made implicit and we drop the second argument). Given an observation record $Y_{0:t}$, the predictive and filtering distributions (4.3.2) are then the distributions of X_t given $Y_{0:t-1}$ and X_t given $Y_{0:t}$ respectively.

These two distributions are of considerable interest in Bayesian filtering as they enable the estimation of the hidden states through the observed data record. Unfortunately only in a few cases, such as discrete state spaces or linear and Gaussian HMM, can they be obtained in closed form, see Cappe et al. (2005); Chopin and Papaspiliopoulos (2020) for a complete overview.

4.3.2 Particle filter

We now illustrate how to obtain empirical estimates of η_t and ϕ_t in an online manner through Monte Carlo simulation. Assume that at time t the empirical measure

$$\eta_t^N(\mathrm{d}x_t) := N^{-1} \sum_{i=1}^N \delta_{\xi_t^i}(\mathrm{d}x_t)$$

based on random samples $\{\xi_t^i\}_{1 \leq i \leq N}$ approximates $\eta_t(\mathrm{d}x_t)$. Plugging η_t^N in (4.3.2) provides an approximation of $\phi_t(\mathrm{d}x_t)$,

$$\phi_t^N(\mathrm{d}x_t) := \sum_{i=1}^N \omega_t^i \delta_{\xi_t^i}(\mathrm{d}x_t)\,,$$

where $\omega_t^i := \Omega_t^{-1} \widetilde{\omega}_t^i$, $\widetilde{\omega}_t^i := g_t(\xi_t^i)$ and $\Omega_t := \sum_{i=1}^N \widetilde{\omega}_t^i$. Replacing ϕ_t by ϕ_t^N in (4.3.3), we obtain the mixture $\phi_t^N M_{t+1}$ which allows to construct η_{t+1}^N by drawing N samples from it by performing for all $i \in [N]$

$$\xi_{t+1}^i \sim M_{t+1}(\xi_t^{A_t^i}, \cdot), \quad \text{where } A_t^i \sim \text{Categorical}(\omega_t^{1:N})\,.$$

The algorithm is initialized with the approximation $\eta_0^N := N^{-1} \sum_{i=1}^N \delta_{\xi_0^i}(\mathrm{d}x_t)$ where $\xi_0^{1:N} \sim M_0^{\otimes N}$ and coincides with the *bootstrap* algorithm with multinomial resampling Gordon et al. (1993a). Note that this mechanism has been extended in many directions in the past decades Pitt and Shephard (1999); Douc and Cappé (2005); Chopin and Papaspiliopoulos (2020).
Alongside η_t^N and ϕ_t^N, the particle approximation of the unnormalized marginal $\gamma_t(\mathrm{d}x_t)$ is given by

$$\gamma_t^N(\mathrm{d}x_t) := \left\{ \prod_{s=0}^{t-1} N^{-1} \Omega_s \right\} \eta_t^N(\mathrm{d}x_t)\,, \quad \forall t > 0\,, \tag{4.3.4}$$

and $\gamma_0^N(\mathrm{d}x_0) = \eta_0^N(\mathrm{d}x_0)$. In particular, $\gamma_t^N(h)$ is unbiased for any function h Del Moral (2004).

In the remainder of this chapter, we denote by $\mathcal{F}_t^N$ the σ-field containing all the particles and ancestors up to time t, i.e.

$$\mathcal{F}_t^N := \sigma\left(\boldsymbol{\xi}_{0:t}^{1:N}, \boldsymbol{A}_{0:t-1}^{1:N}\right)\,.$$

4.3.3 Asymptotic variance estimation in particle filters

The particle filter described above yields consistent estimators, see for instance Cappe et al. (2005); Douc et al. (2014); Chopin and Papaspiliopoulos (2020); Liu and West (2001); Del Moral (2004) and references therein for a complete overview. Indeed, for a test function $h \in \mathsf{F}(\mathcal{X})$ and under assumption (**A5**), the SMC estimators satisfy a Strong Law of Large Numbers when the number of particles N goes to infinity, i.e.

$$\gamma_t^N(h) \xrightarrow[N\to\infty]{a.s.} \gamma_t(h), \quad \eta_t^N(h) \xrightarrow[N\to\infty]{a.s.} \eta_t(h)\,, \quad \phi_t^N(h) \xrightarrow[N\to\infty]{a.s.} \phi_t(h)\,. \tag{4.3.5}$$

Under the same assumptions, CLTs for $\gamma_t^N(h)$, $\eta_t^N(h)$ and $\phi_t^N(h)$ are also available Del Moral and Guionnet (1999); Chopin (2004):

$$\begin{cases} \sqrt{N}(\gamma_t^N(h) - \gamma_t(h)) & \underset{N\to\infty}{\Longrightarrow} \mathcal{N}(0, \mathcal{V}_{\gamma,t}^{\infty}(h)) \,, \\ \sqrt{N}(\eta_t^N(h) - \eta_t(h)) & \underset{N\to\infty}{\Longrightarrow} \mathcal{N}(0, \mathcal{V}_{\eta,t}^{\infty}(h)) \,, \\ \sqrt{N}(\phi_t^N(h) - \phi_t(h)) & \underset{N\to\infty}{\Longrightarrow} \mathcal{N}(0, \mathcal{V}_{\phi,t}^{\infty}(h)) \,, \end{cases} \tag{4.3.6}$$

where $\Longrightarrow$ denotes weak convergence and

$$\mathcal{V}_{\gamma,t}^{\infty}(h) = \sum_{s=0}^{t} \{\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h]^2) - \gamma_t(h)^2\} \,, \tag{4.3.7}$$

$$\mathcal{V}_{\eta,t}^{\infty}(h) = \sum_{s=0}^{t} \frac{\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h - \eta_t(h)]^2)}{\gamma_t(\mathbf{1})^2} \,, \tag{4.3.8}$$

$$\mathcal{V}_{\phi,t}^{\infty}(h) = \sum_{s=0}^{t} \frac{\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[g_t\{h - \phi_t(h)\}]^2)}{\gamma_{t+1}(\mathbf{1})^2} \,. \tag{4.3.9}$$

An intuitive derivation of (4.3.7) is proposed in Section B.4 of the Appendix. The authors of Chan and Lai (2013) propose to estimate (4.3.7) online using the samples produced by the particle filter described above. Their estimator is based on the genealogy of the particle system induced by the successive resampling steps of the particle filter. From the indices A_t^i, it is possible to trace back the ancestors of each particle and deduce the corresponding ancestor at time $t = 0$. More interestingly, these ancestors can be computed in a forward way by introducing the Eve indices $\mathrm{E}_{t,0}^i$. For all $i \in [1 : N]$, $\mathrm{E}_{t,0}^i$ describes the index of the ancestor at time 0 of particle ξ_t^i and can be computed from

$$\mathrm{E}_{t,0}^i = \mathrm{E}_{t-1,0}^{A_{t-1}^i} \mathbb{1}_{t>0} + i\mathbb{1}_{t=0} \,. \tag{4.3.10}$$

The asymptotic variance estimator of $\eta_t^N(h)$ obtained in Chan and Lai (2013) reads (see Section B.2.1 of the supplementary for a proof):

$$\mathcal{V}_{\eta,t}^{N}(h) := -N^{-1} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i \neq \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\} \,. \tag{4.3.11}$$

We sometimes refer to $\mathcal{V}_{\eta,t}^{N}(h)$ as the CLE (Chan & Lai Estimator). Note that it can be computed online in a remarkably simple way since the Eve indices (4.3.10) are computed recursively. However, the counterpart of its computational simplicity is that it degenerates as soon as the ancestral paths coalesce. Indeed, it is widely known in the SMC literature that all lineages eventually end up with the same ancestor when t is large enough with respect to the number of samples N (see e.g. (Fearnhead et al., 2010, Section 2.2) for a more detailed explanation). This means that for a fixed N, as t grows and $s \ll t$, $E_{s,0}^i = E_{s,0}^j$ for all $(i, j) \in \mathbb{N}^2$ and $\mathcal{V}_{\eta,t}^{N}(h) = 0$ for any test function h.

The degeneracy problem concerning (4.3.11) is partially addressed in Olsson and Douc (2019) by truncating the genealogy of the particle system. Denoting $\lambda \in [t]$ the lag and $\mathrm{E}_{t,t-\lambda}^i$ the ancestor of ξ_t^i at time $t - \lambda$, their estimator reads

$$\mathcal{V}_{\eta,t}^{N,\lambda}(h) := -N^{-1} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,t-\lambda}^i \neq \mathrm{E}_{t,t-\lambda}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\} \,. \tag{4.3.12}$$

In the regime where (4.3.11) degenerates, (4.3.12) can be made stable provided that the lag λ is chosen such that there is little asymptotic bias. However, besides the strong mixing case for

which the authors propose a heuristic, the practical choice of such a λ, although crucial, is a non trivial task.

In Lee and Whiteley (2018), the CLE is revisited using different techniques based on Cérou et al. (2011); Andrieu et al. (2010, 2018a). These tools enable them to derive a weakly consistent term by term estimator of (4.3.7) based on the unbiased estimation of $\gamma_t(h)^2$ and of each $\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h]^2)$, for all $s \in [0:t]$. The construction of this second estimator is appealing and insightful in that it helps identifying the deep root of the degeneracy in the CLE. They indeed show that (4.3.11) and each $\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h]^2)$ can be interpreted as a conditional expectation with respect to particle indices that retrace the ancestral paths. Indeed, by introducing discrete random variables $K^1_{0:t}$ and $K^2_{0:t}$ such that conditionally on $\mathcal{F}^N_t$, K^1_t and K^2_t are distributed uniformly on $[N]$ and such that for any $s \in [0:t-1]$,

$$K^1_s = A^{K^1_{s+1}}_s, \quad K^2_s = \mathbb{1}_{K^1_{s+1} \neq K^2_{s+1}} A^{K^2_{s+1}}_s + \mathbb{1}_{K^1_{s+1}=K^2_{s+1}} C_s \,,$$

where $C_s \sim \text{Categorical}(\omega^{1:N}_s)$, then for example

$$\mathcal{V}^N_{\eta,t}(h) = -N\mathbb{E}\left[\prod_{s=0}^{t} \mathbb{1}_{K^1_s \neq K^2_s}\{h(\xi^{K^1_t}_t) - \eta^N_t(h)\}\{h(\xi^{K^2_t}_t) - \eta^N_t(h)\} \middle| \mathcal{F}^N_t\right] .$$

An intuitive extension is to replace the deterministic assignments $K^1_s = A^{K^1_{s+1}}_s$ and $K^2_s = A^{K^2_{s+1}}_s$ by random ones based on backward sampling. Essentially, backward sampling consists in sampling at time s a particle index i starting from the j-th particle at time $s+1$ with probability proportional to $\widetilde{\omega}^i_s m_{s+1}(\xi^i_s, \xi^j_{s+1})$ and thus allows considering relevant trajectories which are not necessarily ancestral trajectories.

4.4 Variance estimation with backward sampling

In this section we present three variance estimators for the bootstrap particle filter. In Section 4.4.1, we lay out our methodology and derive a term by term variance estimator; its computation is detailed in Section 4.4.2. In Sections 4.4.3 and 4.4.4, we provide two additional estimators that have a lower computational cost. All estimators and justifications are provided for the distribution (4.3.4). We give the expressions for the variance estimators of the predictor and filter and provide their justification in Section B.2.2 of the Appendix.

4.4.1 Term by term variance estimator

For any $t \in \mathbb{N}$, let $\mathcal{B}_t := \{0,1\}^{t+1}$. Denote by $\mathbf{0}$ the null vector in $\mathcal{B}_t$ and e_s the vector with 1 at position s and 0 elsewhere. Let $(X_s, X'_s)_{s\in[0:t]}$ be a bivariate Markov chain in $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ and depending on $b \in \mathcal{B}_t$ with initial distribution $\mathcal{M}^{b_0}_0$ and transition kernels $\mathcal{M}^{b_t}_t$, $t \geq 1$, where

$$\begin{aligned} \mathcal{M}^{b_0}_0(\mathrm{d}x_0, \mathrm{d}x'_0) &:= M_0(\mathrm{d}x_0)\{\mathbb{1}_{b_0=0} M_0(\mathrm{d}x'_0) + \mathbb{1}_{b_0=1}\delta_{x_0}(\mathrm{d}x'_0)\}\,, \\ \mathcal{M}^{b_t}_t(x, x'; \mathrm{d}z, \mathrm{d}z') &:= M_t(x, \mathrm{d}z)\{\mathbb{1}_{b_t=0} M_t(x', \mathrm{d}z') + \mathbb{1}_{b_t=1}\delta_z(\mathrm{d}z')\}\,, \quad \forall t \geq 1\,. \end{aligned} \tag{4.4.1}$$

Define also for any $b \in \mathcal{B}_t$ the measure $\mathcal{Q}_{b,t}$ by $\mathcal{Q}_{b,0}(\mathrm{d}x_0, \mathrm{d}x'_0) = \mathcal{M}^{b_0}_0(\mathrm{d}x_0, \mathrm{d}x'_0)$ and for $t \geq 1$:

$$\mathcal{Q}_{b,t}(\mathrm{d}x_{0:t}, \mathrm{d}x'_{0:t}) := \mathcal{M}^{b_0}_0(\mathrm{d}x_0, \mathrm{d}x'_0) \prod_{s=0}^{t-1} g^{\otimes 2}_s(x_s, x'_s) \prod_{s=1}^{t} \mathcal{M}^{b_s}_s(x_{s-1}, x'_{s-1}; \mathrm{d}x_s, \mathrm{d}x'_s)\,. \tag{4.4.2}$$

The measure $\mathcal{Q}_{b,t}$ is the joint distribution (4.3.1) of the Feynman-Kac model defined by the initial distribution $\mathcal{M}_0^{b_0}$, the transition kernels $\{\mathcal{M}_s^{b_s}\}_{s\in[1:t]}$ and by the potential functions $\{g_s^{\otimes 2}\}_{s\in[0:t-1]}$. Remark that for any $h \in \mathsf{F}(\mathcal{X})$, writing $h_t : x_{0:t} \mapsto h(x_t)$, we have that $\mathcal{Q}_{\mathbf{0},t}(h_t^{\otimes 2}) = \gamma_t(h)^2$ and

$$\mathcal{Q}_{e_s,t}(h_t^{\otimes 2}) = \gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h]^2)\,. \tag{4.4.3}$$

A generalization of (4.4.3) is proved in Proposition 4.5.1. Consequently, for $h \in \mathsf{F}(\mathcal{X})$ $\mathcal{V}_{\gamma,t}^{\infty}(h)$ in (4.3.7) can be rewritten as

$$\mathcal{V}_{\gamma,t}^{\infty}(h) = \sum_{s=0}^{t} \{\mathcal{Q}_{e_s,t}(h_t^{\otimes 2}) - \mathcal{Q}_{\mathbf{0},t}(h_t^{\otimes 2})\}\,, \tag{4.4.4}$$

where $h_t : x_{0:t} \mapsto h(x_t)$. Following this observation, for a given b, an estimator of $\mathcal{Q}_{b,t}(h_t)$ could be obtained with a single run of a bootstrap particle filter in augmented dimension (i.e. relying on the the bivariate transition $\mathcal{M}_t^{b_t}$ at time t and on the weighing of the associated particles with $g_t^{\otimes 2}$). As a direct extension of (4.3.4), this estimator would be unbiased and as a byproduct, we would get an unbiased estimator of (4.4.4). However, this procedure is not in line with our initial objective in the sense that we aim at estimating (4.4.4) with the particles and indices already available.

In order to motivate and introduce our approach, let us consider the static situation where $t = 0$, $b_0 = 0$ and let $(h, f) \in \mathsf{F}(\mathcal{X})^2$. Then,

$$\frac{1}{N(N-1)} \sum_{i,j\in[N]^2} \mathbb{1}_{i\neq j} h(\xi_0^i) f(\xi_0^j), \quad \xi_0^{1:N} \overset{\text{iid}}{\sim} M_0\,, \tag{4.4.5}$$

is an unbiased and almost sure convergent estimator of $\mathcal{Q}_{0,0}(h \otimes f) = M_0^{\otimes 2}(h \otimes f)$ and only relies on i.i.d. samples from M_0 rather than $M_0^{\otimes 2}$. Note that for $h = f$, we thus get an unbiased estimator of $\gamma_0(h)^2$. If $b_0 = 1$, then $N^{-1}\sum_{i=1}^N h(\xi_0^i) f(\xi_0^i)$ is an unbiased and consistent estimator of $\mathcal{Q}_{1,0}(h \otimes f) = M_0(hf)$.

Taking advantage of the fact that the particles at time t are i.i.d. conditionally on $\mathcal{F}_{t-1}^N$ as we use multinomial resampling for the particle filter, we can carry these simple observations to the sequential case in two directions as we now detail. Define for any $t \in \mathbb{N}_{>0}$ the functional version of the backward weights:

$$\beta_t^N(x, y) := \frac{g_{t-1}(y) m_t(y, x)}{\sum_{\ell=1}^N \widetilde{\omega}_{t-1}^\ell m_t(\xi_{t-1}^\ell, x)}, \quad \forall (x, y) \in \mathsf{X}^2\,. \tag{4.4.6}$$

Assume that $t = 1$ and define for any $(k_0^1, k_0^2) \in [N]^2$ the following random variables that involve the backward weights

$$\mathcal{E}_0^{\mathsf{BS}}(k_0^1, k_0^2) := \frac{\Omega_0^2}{N(N-1)} \sum_{k_1^{1:2}\in[N]^2} \mathbb{1}_{k_1^1\neq k_1^2} \beta_1^N(\xi_1^{k_1^1}, \xi_0^{k_0^1}) \beta_1^N(\xi_1^{k_1^2}, \xi_0^{k_0^2}) h(\xi_1^{k_1^1}) f(\xi_1^{k_1^2})\,, \tag{4.4.7}$$

$$\mathcal{E}_1^{\mathsf{BS}}(k_0^1, k_0^2) := \frac{\Omega_0^2}{N} \sum_{k_1^{1:2}\in[N]^2} \mathbb{1}_{k_1^1=k_1^2} \beta_1^N(\xi_1^{k_1^1}, \xi_0^{k_0^1}) \omega_0^{k_0^2} h(\xi_1^{k_1^1}) f(\xi_1^{k_1^2})\,, \tag{4.4.8}$$

and also the following which involve the ancestors

$$\mathcal{E}_0^{\mathsf{GT}}(k_0^1, k_0^2) := \frac{\Omega_0^2}{N(N-1)} \sum_{k_1^{1:2}\in[N]^2} \mathbb{1}_{A_0^{k_1^1}=k_0^1, A_0^{k_1^2}=k_0^2, k_1^1\neq k_1^2} h(\xi_1^{k_1^1}) f(\xi_1^{k_1^2})\,, \tag{4.4.9}$$

$$\mathcal{E}_1^{\mathsf{GT}}(k_0^1, k_0^2) := \frac{\Omega_0^2}{N} \sum_{k_1^{1:2}\in[N]^2} \mathbb{1}_{k_0^1=A_0^{k_1^1}, k_1^1=k_1^2} \omega_0^{k_0^2} h(\xi_1^{k_1^1}) f(\xi_1^{k_1^2})\,. \tag{4.4.10}$$

Here, BS and GT correspond to backward sampling and genealogy tracing, respectively. Consider also Lemma 4.4.1 which states a crucial identity involving the backward weights.

Lemma 4.4.1. *For all* $t \in \mathbb{N}_{>0}$, $(x, y) \in \mathsf{X}^2$,

$$\beta_t^N(x, y)\phi_{t-1}^N M_t(\mathrm{d}x) = \frac{g_{t-1}(y)}{\Omega_{t-1}} M_t(y, \mathrm{d}x)\,, \tag{4.4.11}$$

and for any $(k_{t-1}, k_t) \in [N]^2$ *and* $h \in \mathsf{F}(\mathcal{X})$,

$$\mathbb{E}[\beta_t^N(\xi_t^{k_t}, \xi_{t-1}^{k_{t-1}})h(\xi_t^{k_t})|\mathcal{F}_{t-1}^N] = \mathbb{E}[\mathbb{1}_{k_{t-1}=A_{t-1}^{k_t}} h(\xi_t^{k_t})|\mathcal{F}_{t-1}^N] = \omega_{t-1}^{k_{t-1}} M_t[h](\xi_{t-1}^{k_{t-1}})\,. \tag{4.4.12}$$

The proof is postponed to Section B.1.2 in the Appendix. Applying Lemma 4.4.1 to (4.4.7) and (4.4.9) and using that given $\mathcal{F}_0^N$ the particles at $t = 1$ are i.i.d., we get

$$\mathbb{E}[\mathcal{E}_0^{\mathsf{BS}}(k_0^1, k_0^2)|\mathcal{F}_0^N] = \mathbb{E}[\mathcal{E}_0^{\mathsf{GT}}(k_0^1, k_0^2)|\mathcal{F}_0^N] = g_0^{\otimes 2}(\xi_0^{k_0^1}, \xi_0^{k_0^2})\mathcal{M}_1^0[h \otimes f](\xi_0^{k_0^1}, \xi_0^{k_0^2})\,,$$

and

$$\begin{aligned}
\mathbb{E}\Big[\sum_{k_0^{1:2}\in[N]^2} \mathbb{1}_{k_0^1\neq k_0^2}\mathcal{E}_0^{\mathsf{BS}}(k_0^1, k_0^2)\Big] &= \mathbb{E}\Big[\sum_{k_0^{1:2}\in[N]^2} \mathbb{1}_{k_0^1\neq k_0^2}\mathbb{E}[\mathcal{E}_0^{\mathsf{BS}}(k_0^1, k_0^2)|\mathcal{F}_0^N]\Big] \\
&= \mathbb{E}\Big[\sum_{k_0^{1:2}\in[N]^2} \mathbb{1}_{k_0^1\neq k_0^2}\mathbb{E}[\mathcal{E}_0^{\mathsf{GT}}(k_0^1, k_0^2)|\mathcal{F}_0^N]\Big] \\
&= N(N-1)\mathcal{Q}_{\mathbf{0},1}(h \otimes f)\,.
\end{aligned}$$

Similarly,

$$\mathbb{E}[\mathcal{E}_1^{\mathsf{BS}}(k_0^1, k_0^2)|\mathcal{F}_0^N] = \mathbb{E}[\mathcal{E}_1^{\mathsf{GT}}(k_0^1, k_0^2)|\mathcal{F}_0^N] = g_0^{\otimes 2}(\xi_0^{k_0^1}, \xi_0^{k_0^2})\mathcal{M}_1^1[h \otimes f](\xi_0^{k_0^1}, \xi_0^{k_0^2})\,,$$

and

$$\mathbb{E}\Big[\sum_{k_0^{1:2}\in[N]^2} \mathbb{1}_{k_0^1\neq k_0^2}\mathcal{E}_1^{\mathsf{BS}}(k_0^1, k_0^2)\Big] = N(N-1)\mathcal{Q}_{e_1,1}(h \otimes f)\,.$$

Therefore, it is possible to derive unbiased estimators of $\mathcal{Q}_{\mathbf{0},1}(h \otimes f)$ and $\mathcal{Q}_{e_1,1}(h \otimes f)$ (but also $\mathcal{Q}_{e_0,1}(h \otimes f)$ and $\mathcal{Q}_{(1,1),1}(h \otimes f)$) with a single run of the particle filter in two different ways: either by using the backward weights and (4.4.7)-(4.4.8), or by using directly the ancestry of the particles and (4.4.9)-(4.4.10). The asymptotic variance estimators proposed in Chan and Lai (2013); Lee and Whiteley (2018); Olsson and Douc (2019); Du and Guyader (2021) are all based on the latter solution, while in this chapter we instead focus on estimators based on the backward weights.

We now generalize the derivations performed in the case $t = 1$. Denote by $\Lambda_{1,t}$ and $\Lambda_{2,t}$ the discrete measures conditioned on $\mathcal{F}_t^N$ and defined by

$$\Lambda_{1,t}(k_{0:t}) := N^{-1}\prod_{s=1}^{t}\beta_s(k_s, k_{s-1})\,, \tag{4.4.13}$$

$$\Lambda_{2,t}(k_{0:t}^1; k_{0:t}^2) := N^{-1}\prod_{s=1}^{t}\left\{\mathbb{1}_{k_s^2=k_s^1}\omega_{s-1}^{k_{s-1}^2} + \mathbb{1}_{k_s^2\neq k_s^1}\beta_s(k_s^2, k_{s-1}^2)\right\}. \tag{4.4.14}$$

Specific choices of kernels $\{\beta_s\}_{s=1}^t$ are, for all $(k, \ell) \in [N]^2$,

$$\beta_s^{\mathsf{GT}}(k, \ell) := \mathbb{1}_{\ell=A_{s-1}^k}, \quad \beta_s^{\mathsf{BS}}(k, \ell) := \frac{\widetilde{\omega}_{s-1}^\ell m_s(\xi_{s-1}^\ell, \xi_s^k)}{\sum_{j=1}^N \widetilde{\omega}_{s-1}^j m_s(\xi_{s-1}^j, \xi_s^k)} = \beta_s^N(\xi_s^k, \xi_{s-1}^\ell)\,,$$

where here again GT stands for *genealogy tracing* and BS for *backward sampling*. When the conditional distribution given $\mathcal{F}_t^N$ is $\Lambda_{1,t}^{\mathsf{BS}} \otimes \Lambda_{2,t}^{\mathsf{BS}}$ (resp. $\Lambda_{1,t}^{\mathsf{GT}} \otimes \Lambda_{2,t}^{\mathsf{GT}}$) we write $\mathbb{E}_{\mathsf{BS}}[\cdot|\mathcal{F}_t^N]$ (resp. $\mathbb{E}_{\mathsf{GT}}[\cdot|\mathcal{F}_t^N]$). Define also for any $b \in \mathcal{B}_t$ the coalescence function:

$$\mathrm{I}_{b,s} : ([N]^{s+1})^2 \ni (k_{0:s}^1, k_{0:s}^2) \mapsto \prod_{\ell=0}^{s} \{\mathbb{1}_{k_\ell^1 = k_\ell^2} \mathbb{1}_{b_\ell = 1} + \mathbb{1}_{k_\ell^1 \neq k_\ell^2} \mathbb{1}_{b_\ell = 0}\}, \quad \forall s \in [0:t], \tag{4.4.15}$$

for any $h \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$ the random variable

$$\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) := \prod_{s=0}^{t} N^{b_s} \left(\frac{N}{N-1}\right)^{1-b_s} \gamma_t^N(\mathbf{1})^2 \mathbb{E}_{\mathsf{BS}}[\mathrm{I}_{b,t}(K_{0:t}^1, K_{0:t}^2) h(\xi_{0:t}^{K_{0:t}^1}, \xi_{0:t}^{K_{0:t}^2}) | \mathcal{F}_t^N], \tag{4.4.16}$$

and denote by $\mathcal{Q}_{b,t}^{N,\mathsf{GT}}$ the counterpart where the expectation on the r.h.s. is $\mathbb{E}_{\mathsf{GT}}$.

Remark 4.4.2. *By (4.4.15), the random variable $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ remains defined for any $b \in \mathcal{B}_r$ with $r > t$ and $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) = \mathcal{Q}_{b_{0:t},t}^{N,\mathsf{BS}}(h)$ where $b_{0:t}$ is the truncation of b to the $t+1$ first terms.*

Finally, define for any $h \in \mathsf{F}(\mathcal{X})$, using $h_t : x_{0:t} \mapsto h(x_t)$,

$$\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h) := \sum_{s=0}^{t} \{\mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(h_t^{\otimes 2}) - \mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h_t^{\otimes 2})\}. \tag{4.4.17}$$

Proposition 4.4.3. *Let $t \in \mathbb{N}$. For any $b \in B_t$ and any $h \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$,*

(i) $\mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)|\mathcal{F}_{t-1}^N] = \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h])$ *for all* $t \in \mathbb{N}_{>0}$.

(ii) $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ *is an unbiased estimator of* $\mathcal{Q}_{b,t}(h)$.

(iii) If $h \in \mathsf{F}(\mathcal{X})$, $\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h)$ *is an unbiased estimator of* $\mathcal{V}_{\gamma,t}^{\infty}(h)$.

By convention, we used in (i) the following notation:

$$g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h] : (x_{0:t-1}, x'_{0:t-1}) \mapsto g_{t-1}^{\otimes 2}(x_{t-1}, x'_{t-1}) \int h(x_{0:t}, x'_{0:t}) \mathcal{M}_t^{b_t}(x_{t-1}, x'_{t-1}; \mathrm{d}x_t, \mathrm{d}x'_t).$$

The proof is provided in Section B.1.2 of the Appendix. First, (ii) is a generalization of (Lee and Whiteley, 2018, Lemma 2) which states that $\mathcal{Q}_{b,t}^{N,\mathsf{GT}}(h)$ is also an unbiased estimator $\mathcal{Q}_{b,t}(h)$. Its proof, see (Lee and Whiteley, 2018, Supplementary), is based on a doubly conditional SMC argument Andrieu et al. (2018a) and while this scheme can be replicated to our estimator based on backward weights, we rather propose an alternative and elementary proof that also extends straightforwardly to GT and which is based on our previous discussion. From (4.4.4), (iii) is a direct consequence of (ii) and provides an estimator of (4.4.4) based on a single particle run.

Theorem 4.4.4 deals with the convergence of $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ for bounded h. The convergence in $\mathbf{L}_2$ is stated under assumptions (**A5-6-7**) which are standard and the convergence rate is obtained under the additional assumption (**A8**). The equivalent result for $\mathcal{Q}_{b,t}^{N,\mathsf{GT}}$ is stated in Lee and Whiteley (2018) and is proved under (**A5**) alone. From a technical point of view, this is possible because the use of indicators instead of backward weights allows for cancellations that simplify the analysis significantly. Assumption (**A7**) enables us to show that the additional terms that come with the use of backward weights go to zero.

(**A6**) For all $t > 0$ and $(x, x') \in \mathsf{X}^2$, $m_t(x', x) > 0$.

(**A7**) There exists $\sigma_+ > 0$ such that for all $t \geq 1$, $\sup_{x,x' \in \mathsf{X}} m_t(x', x) \leq \sigma_+$.

(**A8**) There exists $0 < \sigma_- < \sigma_+$ such that for all $t \geq 1$, $\inf_{x,x' \in \mathsf{X}} m_t(x', x) \geq \sigma_-$.

Assumption (**A8**) is a strong assumption that is typically verified in models where the state space X is compact. This assumption, together with (**A7**), are now classic and have been widely used to obtain quantitative bounds in the SMC literature Dubarry and Le Corff (2013); Douc et al. (2011a); Lee et al. (2020).

Theorem 4.4.4. *Assume that (**A5-6-7**) hold. For any $t \in \mathbb{N}$, $b \in \mathcal{B}_t$ and $h \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$,*

$$\lim_{N \to \infty} \|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t}(h)\|_2 = 0\,. \tag{4.4.18}$$

*In addition, if (**A8**) holds the convergence rate is $\mathcal{O}(1/\sqrt{N})$.*

Remark 4.4.5. *The dependence on the time horizon t of the $\mathbf{L}_2$ bound is difficult to analyze and we did not undertake it in the proof. Adapting the proofs of the existing analysis Del Moral et al. (2010c); Dubarry and Le Corff (2013) is not trivial as our smoothing estimators are non standard. Furthermore, the time dependence of the GT counterpart has not been analyzed neither, which renders the comparison with our approach even more difficult.*

The proof can be found in Section B.1.4 of the Appendix. As a straightforward consequence, the term by term estimator (4.4.17) of the asymptotic variance is weakly consistent. It remains to detail how it can be computed. The next section is devoted to the exact computation of the estimators $\mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h)$ and $\mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(h)$ that appear in its expression.

4.4.2 Computation for $b = 0$ and $b = e_s$

We now derive practical expressions of $\mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h)$ and $\mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(h)$ in the practical case where $h : x_{0:t}, x'_{0:t} \mapsto h(x_t, x'_t) \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$. Define, for any $b \in \mathcal{B}_t$ and any $t \geq 0$,

$$\mathcal{T}_t^b(K_t^1, K_t^2) := \mathbb{E}_{\mathsf{BS}}\left[\mathrm{I}_{b,t}(K_{0:t}^1, K_{0:t}^2) | \mathcal{F}_t^N, K_t^1, K_t^2\right]. \tag{4.4.19}$$

Then, by the tower property, $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ in (4.4.16) can be rewritten as

$$\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) = \prod_{s=0}^{t} N^{b_s} \left(\frac{N}{N-1}\right)^{1-b_s} \frac{\gamma_t^N(\mathbf{1})^2}{N^2} \sum_{k,\ell \in [N]^2} \mathcal{T}_t^b(k,\ell) h(\xi_t^k, \xi_t^\ell)\,. \tag{4.4.20}$$

Next, define for any $t \in \mathbb{N}$ and $(k, \ell) \in [N]^2$,

$$S_t(k,\ell) := \sum_{s=0}^{t} \mathcal{T}_t^{e_s}(k,\ell)\,. \tag{4.4.21}$$

Plugging (4.4.20) in (4.4.17), $\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h)$ can be rewritten as

$$\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h) = \frac{N^{t-1}\gamma_t^N(\mathbf{1})^2}{(N-1)^t} \sum_{k,\ell \in [N]^2} \left\{ S_t(k,\ell) - \frac{t+1}{N-1} \mathcal{T}_t^{\mathbf{0}}(k,\ell) \right\} h(\xi_t^k) h(\xi_t^\ell)\,. \tag{4.4.22}$$

The sequential computation of $\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h)$ relies on that of $S_t(k,\ell)$ in (4.4.21), and so on that of $\mathcal{T}_t^{e_s}(k,\ell)$ and $\mathcal{T}_t^{\mathbf{0}}(k,\ell)$. By the tower property, we obtain the following recursions for $\mathcal{T}_t^b$:

$$\begin{cases} \mathcal{T}_0^b(k,\ell) = \mathbb{1}_{k \neq \ell, b_0 = 0} + \mathbb{1}_{k=\ell, b_0 = 1}\,, \\ \mathcal{T}_t^b(k,\ell) = \mathbb{1}_{k \neq \ell} \sum_{i,j \in [N]^2} \beta_t^{\mathsf{BS}}(k,i) \beta_t^{\mathsf{BS}}(\ell,j) \mathcal{T}_{t-1}^b(i,j) & \text{if} \quad b_t = 0\,, \\ \mathcal{T}_t^b(k,\ell) = \mathbb{1}_{k=\ell} \sum_{i,j \in [N]^2} \beta_t^{\mathsf{BS}}(k,i) \omega_{t-1}^j \mathcal{T}_{t-1}^b(i,j) & \text{if} \quad b_t = 1\,, \end{cases} \tag{4.4.23}$$

for all $(k,\ell) \in [N]^2$ and $t \in \mathbb{N}_{>0}$. In particular, if $b = \mathbf{0}$,

$$\mathcal{T}_t^{\mathbf{0}}(k,\ell) = \mathbb{1}_{k\neq\ell} \sum_{i,j\in[N]^2} \beta_t^{\mathsf{BS}}(k,i)\beta_t^{\mathsf{BS}}(\ell,j)\mathcal{T}_{t-1}^{\mathbf{0}}(i,j)\,, \tag{4.4.24}$$

and if $b = e_s$,

$$\mathcal{T}_t^{e_s}(k,\ell) = \begin{cases} \mathbb{1}_{k\neq\ell} \sum_{i,j\in[N]^2} \beta_t^{\mathsf{BS}}(k,i)\beta_t^{\mathsf{BS}}(\ell,j)\mathcal{T}_{t-1}^{e_s}(i,j) & t > s\,, \\ \mathbb{1}_{k=\ell} \sum_{i,j\in[N]^2} \beta_t^{\mathsf{BS}}(k,i)\omega_{t-1}^j\mathcal{T}_{t-1}^{\mathbf{0}}(i,j) & t = s\,, \\ \mathcal{T}_t^{\mathbf{0}}(k,\ell) & t < s\,. \end{cases} \tag{4.4.25}$$

Next, combining (4.4.21)-(4.4.25) we obtain the online update of S_t:

$$S_t(k,\ell) = \mathcal{T}_t^{e_t}(k,\ell) + \mathbb{1}_{k\neq\ell} \sum_{i,j\in[N]^2} \beta_t^{\mathsf{BS}}(k,i)\beta_t^{\mathsf{BS}}(\ell,j)S_{t-1}(i,j)\,, \tag{4.4.26}$$

for any $(k,\ell) \in [N]^2$ and $t \in \mathbb{N}$. We have shown that despite the sum over s that appears in (4.4.17) we are still able to update (4.4.22) at a computational cost independent of the time horizon t by propagating S_t and $\mathcal{T}_t^{\mathbf{0}}$. Note that the algorithm provided in (Lee and Whiteley, 2018, Algorithm 3, Supplementary) does not compute $\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{GT}}$ sequentially since it relies on the computation of each $\mathcal{Q}_{e_s,t}^{N,\mathsf{GT}}(h_t^{\otimes 2})$ and $\mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{GT}}(h_t^{\otimes 2})$ from scratch whenever a new observation is available. In Section B.2.3 of the supplementary material we show how it can be computed online using the same ideas behind the previous derivations.

The computation of the estimates $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ and $\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{BS}}(h)$ can benefit from parallelization by implementing the updates (4.4.24)-(4.4.25) with matrix operations:

$$\begin{cases} \boldsymbol{\mathcal{T}}_t^b = \boldsymbol{\beta}_t^{\mathsf{BS}}\boldsymbol{\mathcal{T}}_{t-1}^b\boldsymbol{\beta}_t^{\mathsf{BS}\top} - \mathrm{Diag}(\boldsymbol{\beta}_t^{\mathsf{BS}}\boldsymbol{\mathcal{T}}_{t-1}^b\boldsymbol{\beta}_t^{\mathsf{BS}\top}) & \text{if} \quad b_t = 0\,, \\ \boldsymbol{\mathcal{T}}_t^b = \mathrm{Diag}(\boldsymbol{\beta}_t^{\mathsf{BS}}\boldsymbol{\mathcal{T}}_{t-1}^b\boldsymbol{\omega}_{t-1}^{1:N}) & \text{if} \quad b_t = 1\,. \end{cases}$$

4.4.3 Variance estimators with reduced computational cost

In this section we derive a second estimator that relies only on the update of $\mathcal{T}_t^{\mathbf{0}}$. Let $h \in \mathsf{F}(\mathcal{X})$. By (4.3.6), $\sqrt{N}(\gamma_t^N(h) - \gamma_t(h))$ converges in distribution; moreover, $N(\gamma_t^N(h) - \gamma_t(h))^2$ is uniformly integrable, using for instance a Hoeffding type inequality (see Douc et al. (2014)). Hence $N\mathbb{E}[(\gamma_t^N(h) - \gamma_t(h))^2]$ converges to the asymptotic variance $\mathcal{V}_{\gamma,t}^{\infty}(h)$. On the other hand, using the lack of bias of $\gamma_t^N(h)$,

$$N\mathbb{E}\left[\left(\gamma_t^N(h) - \gamma_t(h)\right)^2\right] = N\left(\mathbb{E}\left[\gamma_t^N(h)^2\right] - \gamma_t(h)^2\right) = N\left(\mathbb{E}\left[\gamma_t^N(h)^2\right] - \mathcal{Q}_{\mathbf{0},t}(h_t^{\otimes 2})\right)\,.$$

A natural estimator of this quantity is obtained by replacing $\mathbb{E}\left[\gamma_t^N(h)^2\right]$ and $\mathcal{Q}_{\mathbf{0},t}(h_t^{\otimes 2})$ by their unbiased estimators $\gamma_t^N(h)^2$ and $\mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h_t^{\otimes 2})$, respectively,

$$\begin{aligned} \mathcal{V}_{\gamma,t}^{N,\mathsf{BS}}(h) &:= N(\gamma_t^N(h)^2 - \mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h_t^{\otimes 2})) \\ &= N\gamma_t^N(\mathbf{1})^2\left(\eta_t^N(h)^2 - \frac{N^{t-1}}{(N-1)^{t+1}} \sum_{i,j\in[N]^2} \mathcal{T}_t^{\mathbf{0}}(i,j)h(\xi_t^i)h(\xi_t^j)\right)\,. \end{aligned} \tag{4.4.27}$$

For the sake of completeness we also provide the estimator for the predictor and filter and defer their justification to the Section B.2.2 of the supplementary material,

$$\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}(h) := \frac{-N^t}{(N-1)^{t+1}} \sum_{i,j\in[N]^2} \mathcal{T}_t^{\mathbf{0}}(i,j)\{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}\,, \tag{4.4.28}$$

$$\mathcal{V}_{\phi,t}^{N,\mathsf{BS}}(h) := \frac{-N^{t+2}}{(N-1)^{t+1}} \sum_{i,j\in[N]^2} \omega_t^i\omega_t^j\mathcal{T}_t^{\mathbf{0}}(i,j)\{h(\xi_t^i) - \phi_t^N(h)\}\{h(\xi_t^j) - \phi_t^N(h)\}\,. \tag{4.4.29}$$

Remark 4.4.6. *It is worthwhile to note the parallel between* (4.4.28) *and* (4.3.11) *(up to a negligible term depending on N); the indicator is replaced by the backward statistic $\mathcal{T}_t^{\mathbf{0}}(i,j)$ which is the conditional probability of having two disjoint backward trajectories starting from ξ_t^i and ξ_t^j.*

The convergence of (4.4.27) stated in Theorem 4.4.7 stems from the following identity which also appears in Lee and Whiteley (2018); Du and Guyader (2021) and dates back to Cérou et al. (2011):

$$\begin{aligned}&\sum_{b\in\mathcal{B}_t}\left\{\prod_{s=0}^{t}\frac{1}{N^{b_s}}\left(\frac{N-1}{N}\right)^{1-b_s}\right\}\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h_t^{\otimes 2})\\ &\quad=\gamma_t^N(\mathbf{1})^2\mathbb{E}_{\mathsf{BS}}\left[\sum_{b\in\mathcal{B}_t}\mathrm{I}_{b,t}(K_{0:t}^1,K_{0:t}^2)h(\xi_t^{K_t^1})h(\xi_t^{K_t^2})\middle|\mathcal{F}_t^N\right]=\gamma_t^N(\mathbf{1})^2\eta_t^N(h)^2=\gamma_t^N(h)^2\,.\end{aligned}\tag{4.4.30}$$

Theorem 4.4.7. *Let (**A5-6-7**) hold. For any $h\in\mathsf{F}(\mathcal{X})$, $\mathcal{V}_{\gamma,t}^{N,\mathsf{BS}}(h)$ converges in probability to $\mathcal{V}_{\gamma,t}^{\infty}(h)$.*

The proof is in Section B.1.5 of the Appendix. The main advantage of (4.4.27) w.r.t. (4.4.22) is the computational cost. Indeed, remark that (4.4.27) only relies on the sequential update of $\mathcal{T}_t^{\mathbf{0}}$, contrary to (4.4.22) which also relies on that of $\mathcal{T}_t^{e_s}$. Consequently, the computational time of (4.4.27) is roughly twice lower. In addition, experiments show that the difference in performance is negligible so (4.4.27) is to be preferred in practice.

Remark 4.4.8. *This alternative estimator does not invalidate the relevance of* (4.4.17)*. Indeed, remember that* (4.4.17) *is an unbiased estimator. Moreover, the asymptotic variance estimator of the FFBS algorithm that we provide in Section 4.5.2 is a term by term estimator that can be updated online in a way similar to* (4.4.17)*.*

4.4.4 A PaRIS variance estimator

Let us discuss how the computational cost of (4.4.27) and (4.4.22) can be further reduced *à la PaRIS* Olsson and Westerborn (2017); Gloaguen et al. (2022). In Olsson and Westerborn (2017), the forward only implementation of the FFBS algorithm is sped up by replacing the backward statistics by a conditionally unbiased estimator obtained by sampling particle indices according to the backward probabilities β_t^{BS} through rejection sampling. We therefore apply the same idea here by letting $\widetilde{\mathcal{T}}_0^{\mathbf{0}}:=\mathcal{T}_0^{\mathbf{0}}$ and replacing $\mathcal{T}_t^b$ with

$$\begin{aligned}\widetilde{\mathcal{T}}_t^b(k,\ell)&:=\frac{\mathbb{1}_{k\neq\ell}}{M}\sum_{i=1}^{M}\widetilde{\mathcal{T}}_{t-1}^b(J_{k,t-1}^i,J_{\ell,t-1}^i)\qquad\text{if}\quad b_t=0\,,\\ \widetilde{\mathcal{T}}_t^b(k,\ell)&:=\frac{\mathbb{1}_{k=\ell}}{M}\sum_{i=1}^{M}\sum_{j=1}^{N}\omega_{t-1}^j\widetilde{\mathcal{T}}_{t-1}^b(J_{k,t-1}^i,j)\quad\text{if}\quad b_t=1\,,\end{aligned}$$

where for any $k\in[N]$, $J_{k,t-1}^{1:M}$ are i.i.d. samples according to $\beta_t^{\mathsf{BS}}(k,.)$. For $h\in\mathsf{F}(\mathcal{X}^{\otimes 2})$, the *PaRIS* estimator of $\mathcal{Q}_{b,t}(h)$ is, for any $b\in\mathcal{B}_t$

$$\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)=\left\{\prod_{s=0}^{t}N^{b_s}\left(\frac{N}{N-1}\right)^{1-b_s}\right\}\frac{\gamma_t^N(1)^2}{N^2}\sum_{i,j\in[N]^2}\widetilde{\mathcal{T}}_t^b(i,j)h(\xi_t^i,\xi_t^j)\,,\tag{4.4.31}$$

and the *PaRIS* variance estimators are

$$\overline{\mathsf{V}}_{\gamma,t}^{N,M}(h) = \sum_{s=0}^{t} \{\widetilde{\mathcal{Q}}_{e_s,t}^{N,M}(h^{\otimes 2}) - \widetilde{\mathcal{Q}}_{\mathbf{0},t}^{N,M}(h^{\otimes 2})\}\,, \tag{4.4.32}$$

$$\mathsf{V}_{\gamma,t}^{N,M}(h) = N(\gamma_t^N(h)^2 - \widetilde{\mathcal{Q}}_{\mathbf{0},t}^{N,M}(h^{\otimes 2}))\,, \tag{4.4.33}$$

where $M > 1$ refers to the number of sampled indices. The computation of (4.4.27) and (4.4.33) is summarized in Algorithm 2.

Algorithm 2: Update at step $t+1$ of the variance estimators (4.4.27) and (4.4.33) associated to $\gamma_{t+1}^N(h)$

Input: $M, \widetilde{\omega}_t^{1:N}, \xi_t^{1:N}, \xi_{t+1}^{1:N}$, $\boldsymbol{\mathcal{T}}_t^{\mathbf{0}}$ and $\gamma_t^N(\mathbf{1})$

Output: $N\gamma_{t+1}^N(\mathbf{1})^2\{\eta_{t+1}^N(h)^2 - N^t \sum_{i,j\in[N]^2} \boldsymbol{\mathcal{Q}}_{i,j}/(N-1)^{t+2}\}$, $\boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}}$.

Compute $\boldsymbol{\beta}_{t+1}^{\mathsf{BS}}$

if *PaRIS* **then**

  for $k \in [1:N]$ **do**

    Sample $J_{k,t}^{1:M} \overset{\text{iid}}{\sim} \beta_{t+1}^{\mathsf{BS}}(k,.)$

  for $(k,\ell) \in [1:N]^2$ **do**

    Set $\mathcal{T}_{t+1}^{\mathbf{0}}(k,\ell) = \mathbb{1}_{k\neq\ell} \sum_{i=1}^{M} \mathcal{T}_t^{\mathbf{0}}(J_{k,t}^i, J_{\ell,t}^i)/M$

else

  Compute $\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}} \boldsymbol{\mathcal{T}}_t^{\mathbf{0}} \boldsymbol{\beta}_{t+1}^{\mathsf{BS}\prime}$.

  Set $\boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}} = \overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} - \mathrm{Diag}(\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}})$.

Compute $\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}} \odot [h(\xi_{t+1}^{1:N}) h(\xi_{t+1}^{1:N})^\top]$.

We are able to reduce the time complexity of computing $\mathcal{T}_t^b$ to $\mathcal{O}(MN^2)$. The key feature of the *PaRIS* approach is that M does not necessarily need to be large (see (Olsson and Westerborn, 2017, Section 3.1) for a discussion on this matter). We impose $M > 1$ because then in the case $b = \mathbf{0}$, which is the case we are the most interested in, the support of $\mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h)$ is made of $N^2 M^{t+1}$ terms whereas when $M = 1$ it is only N^2. We show empirically in our experiments that setting $M = 3$ is sufficient to provide good results for the asymptotic variance estimation.

While it is not needed to obtain a $\mathcal{O}(MN^2)$ time complexity, the indices $J_{k,t-1}^{1:M}$ can be sampled using an accept-reject procedure with the filtering weights as proposals, if the transition densities m_t are upper bounded. This approach does not require the computation of the normalizing constant of the backward weights (4.4.6). The computational time is then random but if the transition kernels are strongly mixing it can be provably further reduced Douc et al. (2011a). Theorem 4.4.9 is concerned with the convergence of $\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)$ for any bounded h and for any *fixed* $M > 1$. Its proof bears some similarity with that of Theorem 4.4.4 with the exception that the additional sampling introduces non trivial terms that need to be handled carefully. As a straightforward consequence, we obtain the convergence in probability of $\overline{\mathsf{V}}_{\gamma,t}^{N,M}(h)$ for any $h \in \mathsf{F}(\mathcal{X})$. The weak consistency of (4.4.33) in Theorem 4.4.10 is however less straightforward than that of Theorem 4.4.7 and relies on the insight that the identity (4.4.30) still holds when $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}$ are replaced with their *PaRIS* versions. The proofs are provided respectively in Section B.1.6 and B.1.7 of the Appendix.

Theorem 4.4.9. *Assume that (**A5-6-7**) hold. For any $t \in \mathbb{N}$, $b \in \mathcal{B}_t$, $M > 1$ and $h \in \mathsf{F}(\mathcal{X}^{\otimes 2})$,*

$$\lim_{N\to\infty} \|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \mathcal{Q}_{b,t}(h)\|_2 = 0\,. \tag{4.4.34}$$

*In addition, if (**A8**) holds the convergence rate is* $\mathcal{O}(1/\sqrt{N})$.

Theorem 4.4.10. *Let (**A5-6-7**) hold. For all* $t \in \mathbb{N}$, $M > 1$ *and* $h \in \mathsf{F}(\mathcal{X})$, $\mathsf{V}_{\gamma,t}^{N,M}(h)$ *converges in probability to* $\mathcal{V}_{\gamma,t}^{\infty}(h)$ *when* N *goes to infinity.*

4.5 Application to the FFBS

In this section, we derive an estimator for the asymptotic variance of the *Forward Filtering Backward Smoothing* algorithm. We start by giving a short presentation of the FFBS algorithm and we next derive an estimator of the asymptotic variance for additive functionals.

4.5.1 FFBS algorithm

The FFBS algorithm aims at solving the well known degeneracy problem associated with the particle filter of Section 4.3.2 when it is used for approximating smoothing distributions. It relies on the following backward decomposition of the joint smoothing distribution:

$$\phi_{0:t|t}(h) = \int h(x_{0:t})\phi_t(\mathrm{d}x_t)\mathbf{T}_t(x_t, \mathrm{d}x_{0:t-1})\,, \tag{4.5.1}$$

where $\mathbf{T}_t$ is the backward transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{X}^t, \mathcal{X}^{\otimes t})$: $\mathbf{T}_0 := \mathrm{Id}$ and for $t > 0$,

$$\mathbf{T}_t := \mathbf{B}_{\phi_{t-1}} \otimes \cdots \otimes \mathbf{B}_{\phi_0}\,,$$

and $\mathbf{B}_{\phi_s}$ is the backward kernel defined by

$$\mathbf{B}_{\phi_s}(x_{s+1}, A) := \frac{\int m_{s+1}(x_s, x_{s+1})\mathbb{1}_A(x_s)\phi_s(\mathrm{d}x_s)}{\phi_s(m_{s+1}(., x_{s+1}))}, \quad \forall A \in \mathcal{X}\,, \forall x_{s+1} \in \mathsf{X}\,.$$

Denote by $\mathbf{T}_t^N$ the particle approximation of $\mathbf{T}_t$ where each backward kernel $\mathbf{B}_{\phi_s}$ is replaced by plugging in the particle approximation of the filter. This yields for any $A \in \mathcal{X}$ and $x_{s+1} \in \mathsf{X}$,

$$\mathbf{B}_{\phi_s}^N(x_{s+1}, A) := \frac{\int m_{s+1}(x_s, x_{s+1})\mathbb{1}_A(x_s)\phi_s^N(\mathrm{d}x_s)}{\phi_s^N(m_{s+1}(., x_{s+1}))} = \sum_{i=1}^N \frac{\widetilde{\omega}_s^i m_{s+1}(\xi_s^i, x_{s+1})}{\sum_{j=1}^N \widetilde{\omega}_s^j m_{s+1}(\xi_s^j, x_{s+1})}\mathbb{1}_A(\xi_s^i)\,.$$

Plugging this approximation and that of the filtering distribution in (4.5.1) yields

$$\phi_{0:t|t}^{N,\mathrm{FFBS}}(h) := \sum_{i_0=1}^N \cdots \sum_{i_t=1}^N \widetilde{\Lambda}_t(i_{0:t})h(\xi_0^{i_0}, \cdots, \xi_t^{i_t})\,, \tag{4.5.2}$$

where $\widetilde{\Lambda}_t(i_{0:t}) := \omega_t^{i_t}\prod_{s=1}^t \beta_t^{\mathsf{BS}}(i_s, i_{s-1})$. In the following, we write $\phi_{0:t|t}^N$ for $\phi_{0:t|t}^{N,\mathrm{FFBS}}$ and if h is such that $h : x_{0:t} \mapsto h(x_{s:\ell})$ with $0 \le s \le \ell \le t$, we will instead write $\phi_{s:\ell|t}^N(h)$.

The theoretical properties of the FFBS are well understood in both the asymptotic regimes of N and t Douc et al. (2011a); Del Moral et al. (2010a,b); Dubarry and Le Corff (2013); Olsson and Westerborn (2017); Douc et al. (2014). In particular, a Central Limit Theorem with an explicit expression of the asymptotic variance is established for any $h \in \mathsf{F}(\mathcal{X}^{\otimes t+1})$ under (**A5**) in (Douc et al., 2011a, Theorem 8),

$$\sqrt{N}(\phi_{0:t|t}^N(h) - \phi_{0:t|t}(h)) \Longrightarrow \mathcal{N}(0, \mathcal{V}_{0:t|t}^{\mathrm{FFBS}}(h))\,, \tag{4.5.3}$$

where

$$\mathcal{V}^{\mathrm{FFBS}}_{0:t|t}(h) := \sum_{s=0}^{t} \frac{\eta_s(\mathbf{G}_{s,t}[g_t\{h - \phi_{0:t|t}(h)\}]^2)}{\eta_s(\overline{\mathbf{Q}}_{s+1:t}[g_t])^2}\,, \tag{4.5.4}$$

and $\mathbf{G}_{s,t}$ is the kernel that integrates h forward and backward starting from x_s, i.e.

$$\mathbf{G}_{s,t}[h](x_s) := \mathbf{T}_s[\mathbf{Q}_{s+1:t}[h]](x_s) = \int h(x_{0:t})\mathbf{T}_s(x_s, \mathrm{d}x_{0:s-1})\mathbf{Q}_{s+1:t}(x_s, \mathrm{d}x_{s+1:t})\,,$$

for any $s \in [0:t]$ and $x_s \in \mathsf{X}$.

Unlike the asymptotic variance of filtering algorithms, no estimator of (4.5.4) exists in the literature, even though the FFBS and its variants are of significant importance in marginal smoothing and parameter estimation in HMMs Kantas et al. (2015). In this section, we bridge this gap by providing an online estimator for additive functionals h of the form

$$h_{0:t}(x_{0:t}) = \sum_{s=0}^{t-1} \tilde{h}_s(x_s, x_{s+1})\,, \tag{4.5.5}$$

where for $s \in [0:t-1]$, we assume that $\tilde{h}_s$ is bounded. For such functionals, the FFBS can be computed online with a $\mathcal{O}(N^2)$ time complexity per time step, i.e. whenever a new observation is processed. For $0 \leq s < r \leq t$, we write $\tilde{h}_{s:r}(x_{s:r}) = \sum_{\ell=s}^{r-1} \tilde{h}_\ell(x_\ell, x_{\ell+1})$. Expectations of functionals of the form (4.5.5) include marginal smoothing, pairwise marginal smoothing and the E-step of the Expectation Maximization algorithm.

Before we derive our estimator, let us first recall why the FFBS can be indeed computed online in this case. For more details on the forward only implementation of the FFBS and its variants we refer the reader to Douc et al. (2014); Olsson and Westerborn (2017). For any $t > 0$ and any additive functional $h_{0:t}$,

$$\begin{aligned}\mathbf{T}_t[h_{0:t}](x_t) &= \int \left\{\tilde{h}_{0:t-1}(x_{0:t-1}) + \tilde{h}_{t-1}(x_{t-1}, x_t)\right\} \mathbf{B}_{\phi_{t-1}}(x_t, \mathrm{d}x_{t-1})\mathbf{T}_{t-1}(x_{t-1}, \mathrm{d}x_{0:t-2}) \\ &= \mathbf{B}_{\phi_{t-1}}[\mathbf{T}_{t-1}[\tilde{h}_{0:t-1}] + \tilde{h}_{t-1}(., x_t)](x_t)\,.\end{aligned}$$

Then, plugging in the particle approximations, we obtain the following recursion

$$\mathbf{T}^N_t[h_{0:t}](x_t) = \sum_{i=1}^{N} \frac{\widetilde{\omega}^i_{t-1} m_t(\xi^i_{t-1}, x_t)}{\sum_{j=1}^N \widetilde{\omega}^j_{t-1} m_t(\xi^j_{t-1}, x_t)} \{\mathbf{T}^N_{t-1}[\tilde{h}_{0:t-1}](\xi^i_{t-1}) + \tilde{h}_{t-1}(\xi^i_{t-1}, x_t)\}\,, \tag{4.5.6}$$

and then $\phi^N_{0:t|t}(h_{0:t}) = \sum_{i=1}^N \omega^i_t \mathbf{T}^N_t[h_{0:t}](\xi^i_t)$. Therefore, $\mathbf{T}^N_t[h_{0:t}]$ needs only to be estimated at the particle locations and smoothing estimates for additive functionals can be computed with the forward pass and has $\mathcal{O}(N^2)$ complexity per time step.

4.5.2 Asymptotic variance estimator

From now on we will assume that $h_{0:t}$ satisfies (4.5.5). Our estimator is based on the following alternative expression of the asymptotic variance (4.5.4)

$$\mathcal{V}^{\mathrm{FFBS}}_{0:t|t}(h) = \sum_{s=0}^{t} \frac{\gamma_s(\mathbf{1})\gamma_s(\mathbf{G}_{s,t}[g_t\{h_{0:t} - \phi_{0:t|t}(h_{0:t})\}]^2)}{\gamma_{t+1}(\mathbf{1})^2}\,, \tag{4.5.7}$$

which is deduced using the definitions given in Section 4.3. This expression is motivated by Proposition 4.5.1 in which we express the numerators that appear in (4.5.7) in terms of expectations with respect to $\mathcal{Q}_{e_s,t}$. The proof is given in Section B.1.3 of the Appendix.

Proposition 4.5.1. *For any $s \in [0:t]$ and any additive functional $h_{0:t} \in \mathsf{F}(\mathcal{X}^{\otimes t+1})$,*

$$\gamma_s(\mathbf{1})\gamma_s(\mathbf{G}_{s,t}[h_{0:t}]^2) = \mathcal{Q}_{e_s,t}([\mathbf{T}_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t}]^{\otimes 2}) \,. \tag{4.5.8}$$

By Theorem 4.4.4, for any additive functional $h_{0:t}$ as in (4.5.5), we have that $\mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([\mathbf{T}_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t}]^{\otimes 2})$ is a consistent estimator of $\mathcal{Q}_{e_s,t}([\mathbf{T}_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t}]^{\otimes 2})$, but $\mathbf{T}_s[h_{0:s}]$ is intractable and we only have access to its particle approximation $\mathbf{T}^N_s[h_{0:s}]$. Our proposed estimator of the asymptotic variance (4.5.3) is then

$$\mathcal{V}^{N,\mathsf{BS}}_{0:t|t}(h_t) := \sum_{s=0}^{T} \frac{\mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t\{\mathbf{T}^N_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t} - \phi^N_{0:t|t}(h_{0:t})\}]^{\otimes 2})}{\gamma^N_{t+1}(\mathbf{1})^2} \,, \tag{4.5.9}$$

where we have replaced $\phi_{0:t|t}(h_t)$ by its FFBS estimator. Remark that Theorem 4.4.4 cannot be applied to $\mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t\{\mathbf{T}^N_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t}]^{\otimes 2})$ because its proof relies on the fact that the function h integrated by $\mathcal{Q}^{N,\mathsf{BS}}_{b,t}$ does not depend on the particles.

Theorem 4.5.2 proved in Section B.1.8 of the supplementary material shows that weak consistency still holds under the assumptions of Theorem 4.4.4. The proof proceeds in three steps. We first establish that for all $s > 0$ and additive functional $h_{0:s}$, $\mathbf{T}^N_s[h_{0:s}](x_s)$ converges $\mathbb{P}$-a.s. to $\mathbf{T}_s[h_{0:s}](x_s)$ for any $x_s \in \mathsf{X}$. Then, we use it to show that at $t = s$, the distance in $\mathbf{L}_2$ between $\mathcal{Q}^{N,\mathsf{BS}}_{e_s,s}([\mathbf{T}^N_s[h_{0:s}]c_s + \tilde{h}_s] \otimes [\mathbf{T}^N_s[f_{0:s}]d_s + \tilde{f}_s])$ and the "idealized" consistent estimator $\mathcal{Q}^{N,\mathsf{BS}}_{e_s,s}([\mathbf{T}_s[h_{0:s}]c_s + \tilde{h}_s] \otimes [\mathbf{T}_s[f_{0:s}]d_s + \tilde{f}_s])$, goes to 0. Finally, we extend the result to $t > s$ by induction, similarly to Theorem 4.4.4.

Theorem 4.5.2. *Assume that (**A5-6-7**) hold. For any $t \in \mathbb{N}$, $s \in [0:t]$, $(\tilde{h}_{s:t}, \tilde{f}_{s:t}) \in \mathsf{F}(\mathcal{X}^{\otimes t-s+1})^2$, $(c_t, d_t) \in \mathsf{F}(\mathcal{X})^2$ and additive functionnals $(h_{0:s}, f_{0:s})$ (4.5.5),*

$$\lim_{N\to\infty} \big\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([\mathbf{T}^N_s[h_{0:s}]c_t + \tilde{h}_{s:t}] \otimes [\mathbf{T}^N_s[f_{0:s}]d_t + \tilde{f}_{s:t}]) - \mathcal{Q}_{e_s,t}([\mathbf{T}_s[h_{0:s}]c_t + \tilde{h}_{s:t}] \otimes [\mathbf{T}_s[f_{0:s}]d_t + \tilde{f}_{s:t}]) \big\|_2 = 0 \,, \tag{4.5.10}$$

and for any additive functional (4.5.5), $\mathcal{V}^{\mathsf{BS}}_{0:t|t}(h_{0:t})$ converges in probability to $\mathcal{V}^{\mathrm{FFBS}}_{0:t|t}(h_{0:t})$.

4.5.3 Algorithm for marginal smoothing

We now provide an algorithm for the case $h_\ell : x_{0:t} \mapsto h_\ell(x_\ell)$ known as the marginal smoothing problem. For such functions (4.5.9) is defined for $t \geq \ell$ and simplifies to

$$\begin{aligned} \mathcal{V}^{N,\mathsf{BS}}_{\ell|t}(h_\ell) := \frac{1}{\gamma^N_{t+1}(\mathbf{1})^2} \Big\{ & \sum_{s=0}^{\ell} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t\{h_\ell - \phi^N_{\ell|t}(h_\ell)\}]^{\otimes 2}) \\ & + \sum_{s=\ell+1}^{t} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t\{\mathbf{T}^N_s[h_\ell] - \phi^N_{\ell|t}(h_\ell)\}]^{\otimes 2}) \Big\} \,. \end{aligned} \tag{4.5.11}$$

When $\ell = t$ we recover the term by term estimator of the filter which is consistent with the fact that $\phi^N_{t|t}(h) = \phi^N_t(h)$. Using the bilinearity of $\mathcal{Q}^{N,\mathsf{BS}}_{b,t}$ yields

$$\mathcal{V}^{\mathsf{BS}}_{\ell|t}(h_\ell) = R^\ell_{1,t} - \phi^N_{\ell|t}(h_\ell)R^\ell_{2,t} + \phi^N_{\ell|t}(h_\ell)^2 R_t \,, \tag{4.5.12}$$

where $R_t := \sum_{s=0}^t \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(g_t^{\otimes 2})$ and

$$\begin{cases} R^1_{\ell,t} & := \sum_{s=0}^{\ell} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t h_\ell]^{\otimes 2}) + \sum_{s=\ell+1}^{t} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}([g_t\mathbf{T}^N_s[h_\ell]]^{\otimes 2}) \,, \\ R^2_{\ell,t} & := \{ \sum_{s=0}^{\ell} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(g_t h_\ell \otimes g_t) + \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(g_t \otimes g_t h_\ell) \} \\ & \qquad + \{ \sum_{s=\ell+1}^{t} \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(g_t\mathbf{T}^N_s[h_\ell] \otimes g_t) + \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(g_t \otimes g_t\mathbf{T}^N_s[h_\ell]) \} \,. \end{cases}$$

Mirroring (4.4.25), define for any $t \in \mathbb{N}$, $n \in [0:t]$ and $f_n : x_{0:t}, x'_{0:t} \mapsto f(x_n, x'_n)$ the random variable

$$\mathcal{T}_t^{e_s}[f_n](K_t^1, K_t^2) := \mathbb{E}[\mathrm{I}_{e_s,t}(K_{0:t}^1, K_{0:t}^2) f_n(\xi_n^{K_n^1}, \xi_n^{K_n^2}) | \mathcal{F}_{t-1}^N, K_t^1, K_t^2] \,, \tag{4.5.13}$$

and also write $S_{t,t}^1(K_t^1, K_t^2) = S_t(K_t^1, K_t^2) h_t^{\otimes 2}(\xi_t^{K_t^1}, \xi_t^{K_t^2})$, $S_{t,t}^2(K_t^1, K_t^2) = S_t(K_t^1, K_t^2) h_t^{\oplus 2}(\xi_t^{K_t^1}, \xi_t^{K_t^2})$ and for any $t > \ell$,

$$\begin{aligned}
S_{\ell,t}^1(K_t^1, K_t^2) &= \sum_{s=0}^{\ell} \mathcal{T}_t^{e_s}[h_\ell^{\otimes 2}](K_t^1, K_t^2) + \sum_{s=\ell+1}^{t} \mathcal{T}_t^{e_s}[\mathbf{T}_s^N[h_\ell]^{\otimes 2}](K_t^1, K_t^2), \\
S_{\ell,t}^2(K_t^1, K_t^2) &= \sum_{s=0}^{\ell} \mathcal{T}_t^{e_s}[h_\ell^{\oplus 2}](K_t^1, K_t^2) + \sum_{s=\ell+1}^{t} \mathcal{T}_t^{e_s}[\mathbf{T}_s^N[h_\ell]^{\oplus 2}](K_t^1, K_t^2),
\end{aligned}$$

where for any f_ℓ, $f_\ell^{\oplus 2} : x_\ell, x'_\ell \mapsto f_\ell(x_\ell) + f_\ell(x'_\ell)$ and S_ℓ is defined in (4.4.21). Applying the tower property,

$$\mathcal{V}_{\ell|t}^{\mathsf{BS}}(h_t) = N \left(\frac{N}{N-1} \right)^t \sum_{i,j \in [N]^2} \omega_t^i \omega_t^j \{ S_{\ell,t}^1(i,j) - \phi_{\ell|t}^N(h_\ell) S_{\ell,t}^2(i,j) + \phi_{\ell|t}^N(h_\ell)^2 S_t(i,j) \} \,.$$

The quantities $S_{\ell,t+1}^1$, $S_{\ell,t+1}^2$ may be updated online using the following recursions which are again obtained by applying the tower property

$$\begin{aligned}
S_{\ell,t+1}^1(i,j) := \mathcal{T}_{t+1}^{e_{t+1}}(i,j) \mathbf{T}_{t+1}^N[h_\ell](\xi_{t+1}^i) \mathbf{T}_{t+1}^N[h_\ell](\xi_{t+1}^j) \\
+ \mathbb{1}_{i \neq j} \sum_{m,n \in [N]^2} \beta_{t+1}^{\mathsf{BS}}(i,m) \beta_{t+1}^{\mathsf{BS}}(j,n) S_{\ell,t}^1(m,n) \,,
\end{aligned} \tag{4.5.14}$$

and

$$\begin{aligned}
S_{\ell,t+1}^2(i,j) := \mathcal{T}_{t+1}^{e_{t+1}}(i,j) \{ \mathbf{T}_{t+1}^N[h_\ell](\xi_{t+1}^i) + \mathbf{T}_{t+1}^N[h_\ell](\xi_{t+1}^j) \} \\
+ \mathbb{1}_{i \neq j} \sum_{m,n \in [N]^2} \beta_{t+1}^{\mathsf{BS}}(i,m) \beta_{t+1}^{\mathsf{BS}}(j,n) S_{\ell,t}^2(m,n) \,.
\end{aligned} \tag{4.5.15}$$

The updates of $S_{\ell,t+1}^1(i,j)$ and of $S_{\ell,t+1}^2(i,j)$ are thus similiar to that of S_t in (4.4.26). The computation of the variance estimator is described in Algorithm 3.

4.6 Numerical simulations

We now demonstrate our estimators on particle filtering and smoothing examples in HMMs (see Ex. 4.3.1). We assume in this section that $\mathsf{X} = \mathsf{Y} = \mathbb{R}$ and that the dominating measure is the Lebesgue measure. The model considered is the *stochastic volatility model* with, for all $n \geq 1$,

$$X_{n+1} = \varphi X_n + \sigma U_{n+1} \quad \text{and} \quad Y_n = \beta \exp(X_n/2) V_n \,, \tag{4.6.1}$$

with $(\varphi, \beta, \sigma) = (.975, .641, .165)$, $\{U_n\}_{n \in \mathbb{N}}$ and $\{V_n\}_{n \in \mathbb{N}}$ are two sequences of independent standard Gaussian noises and U_n is independent of V_m for all $(n,m) \in \mathbb{N}^2$. The state process $\{X_n\}_{n \in \mathbb{N}}$ is initialized with a Gaussian distribution with zero mean and variance $\sigma^2/(1-\varphi^2)$. These are the exact values and initialization used in Olsson and Douc (2019). The assumptions on the model under which Olsson and Douc (2019) conduct their theoretical analysis and (**A4-5-6-7**) are satisfied for this model. All the simulations are run on GPU and the implementations using matrix operations are available at https://github.com/yazidjanati/asymptoticvariance.

Algorithm 3: Update at step $t+1$ of the variance estimator for marginal smoothing

Input: $\boldsymbol{\omega}_{t+1}^{1:N}, \boldsymbol{\omega}_t^{1:N}, \beta_{t+1}^{\mathsf{BS}}, \mathbf{T}_{t+1}^N[h_\ell], \boldsymbol{\mathcal{T}}_t^{\mathbf{0}}, \boldsymbol{S}_{\ell,t}^1, \boldsymbol{S}_{\ell,t}^2, \phi_{\ell|t+1}^N(h_\ell)$

Output: $-N^{t+2}/(N-1)^{t+1} \sum_{i,j\in[N]^2} \omega_{t+1}^i \omega_{t+1}^j \overline{S}_{\ell,t+1}(i,j), \mathcal{T}_{t+1}^{\mathbf{0}}, \boldsymbol{S}_{\ell,t+1}^1, \boldsymbol{S}_{\ell,t+1}^2, \boldsymbol{S}_{t+1}.$

Compute $\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{e_{t+1}} = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}} \boldsymbol{\mathcal{T}}_t^{\mathbf{0}} \boldsymbol{\omega}_t^{1:N}, \quad \overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}} \boldsymbol{\mathcal{T}}_t^{\mathbf{0}} \boldsymbol{\beta}_{t+1}^{\mathsf{BS}\top}, \quad \widetilde{\boldsymbol{S}}_{t+1} = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}} \boldsymbol{S}_t \boldsymbol{\beta}_{t+1}^{\mathsf{BS}\top}$

Set $\boldsymbol{\mathcal{T}}_{t+1}^{e_{t+1}} = \mathrm{Diag}(\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{e_{t+1}}), \quad \boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}} = \overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} - \mathrm{Diag}(\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}}),$
$\boldsymbol{S}_{t+1} = \widetilde{\boldsymbol{S}}_{t+1} - \mathrm{Diag}(\widetilde{\boldsymbol{S}}_{t+1}) + \mathcal{T}_{t+1}^{e_{t+1}}$

for $i \in \{1,2\}$ **do**

  Compute $\widetilde{\boldsymbol{S}}_{\ell,t+1}^i = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}} \boldsymbol{S}_{\ell,t}^i \boldsymbol{\beta}_{t+1}^{\mathsf{BS}\top}$.

  Set $\widetilde{\boldsymbol{S}}_{\ell,t+1}^i = \widetilde{\boldsymbol{S}}_{t+1}^i - \mathrm{Diag}(\widetilde{\boldsymbol{S}}_{t+1}^i)$

Set $\boldsymbol{S}_{\ell,t+1}^1 = \widetilde{\boldsymbol{S}}_{\ell,t+1}^1 + \boldsymbol{\mathcal{T}}_{t+1}^{e_{t+1}} \odot [\mathbf{T}_{t+1}[h_\ell] \mathbf{T}_{t+1}[h_\ell]^\top]$

Set $\boldsymbol{S}_{\ell,t+1}^2 = \widetilde{\boldsymbol{S}}_{\ell,t+1}^2 + \boldsymbol{\mathcal{T}}_{t+1}^{e_{t+1}} \odot [\mathbf{T}_{t+1}[h_\ell] + \mathbf{T}_{t+1}[h_\ell]^\top]$

Set $\overline{\boldsymbol{S}}_{\ell,t+1} = \boldsymbol{S}_{\ell,t+1}^1 - \phi_{\ell|t+1}^N(h_\ell) \boldsymbol{S}_{\ell,t+1}^2 + \phi_{\ell|t+1}^N(h_\ell)^2 \boldsymbol{S}_{t+1}$

4.6.1 Asymptotic variance of the predictor

We are interested in the estimation of the asymptotic variance of the predictor $\eta_t^N(\mathrm{Id})$ *at each time step* t. The estimator is given in Section B.2.2 of the Appendix. We use synthetic datasets sampled from (4.6.1). The real asymptotic variances are intractable and we estimate them by repeating independently and a thousand times the computation of each predictor mean $\eta_t^N(\mathrm{Id})$ with $N = 10000$ and then multiplying the sample variance by N.

We first investigate how the backward sampling variance estimators $\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}$, $\mathsf{V}_{\eta,t}^{N,M}$ and $\overline{\mathcal{V}}_{\eta,t}^{N,\mathsf{BS}}$ behave in terms of computational time, bias and variance. The *PaRIS* estimator is used with $M = 3$ without rejection sampling. For this first experiment, we sampled 750 observations from (4.6.1) and ran 50 particle filters with $N = 3000$ from which we obtained 50 replicates of each asymptotic variance estimate. The results are reported in Figure 4.1. The three estimators exhibit approximately the same variance but the term by term version becomes slightly more biased at t increases. Strikingly, the *PaRIS* estimator $\mathsf{V}_{\eta,t}^{N,M}$ behaves similarly to $\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}$ with a much lower computational time and complexity as can be seen on the left plot of Figure B.1 in the supplementary material.

We now compare $\mathsf{V}_{\eta,t}^{N,M}$ with $M = 3$ to Chan and Lai (2013); Lee and Whiteley (2018); Olsson and Douc (2019) on two different observation records of different length. For the lag size parameter λ, we found that $\lambda = 20$ has the best bias-variance trade-off by comparing the obtained fixed lag estimates with the crude asymptotic variance estimator. Note that in realistic situations choosing the right λ is non trivial (besides the strong mixing case, as argued in Olsson and Douc (2019)) and for this reason we conducted the experiments with two additional lag values, $\lambda \in \{100, 200\}$. For moderately long observation records ($t \in [750]$) and with $N = 3000$, $\mathsf{V}_{\eta,t}^{N,M}$ compares favorably in terms of bias-variance trade-off with the best lagged estimator and even has similar computational time on GPU. The results are reported in Figure 4.2. The five different estimators are computed with the same particle cloud and replicated 50 times. As expected, when the lag is increased the fixed lag estimates exhibit more variance because of the particle degeneracy. In the extreme case where the lag is set to 750 (CLE) bias and variance both increase significantly, as showcased in the fifth plot.

For the longer time horizon $t \in [3000]$ we set $N = 5000$ and picked three time steps in order to monitor the bias and variance closely, see Figure 4.3. The variance of $\mathsf{V}_{\eta,t}^{N,M}$ remains steady while the bias increases gradually but slowly. This is attributed to the fact that our estimator is

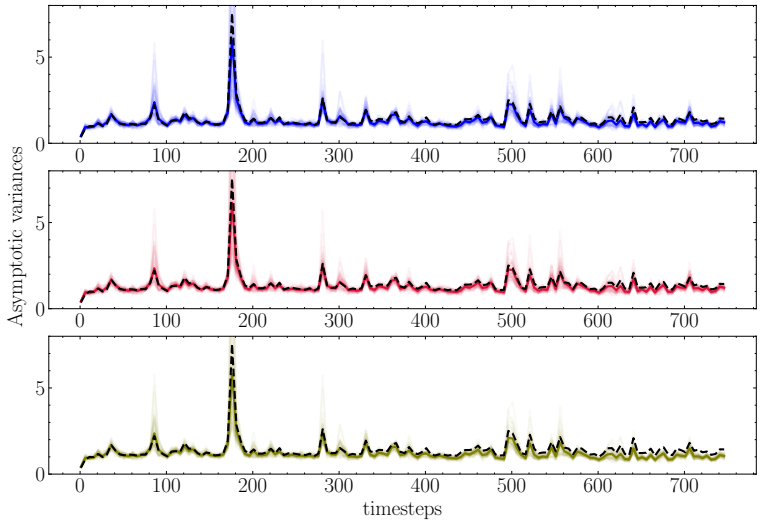

Figure 4.1: Long-term behavior of $\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}$ *(top),* $\mathsf{V}_{\eta,t}^{N,M}$ *with* $M = 3$ *(middle) and* $\overline{\mathcal{V}}_{\eta,t}^{N,\mathsf{BS}}$ *(bottom). The black dashed line is the asymptotic variance estimated using brute force. The number of particles is set to* $N = 2000$*.*

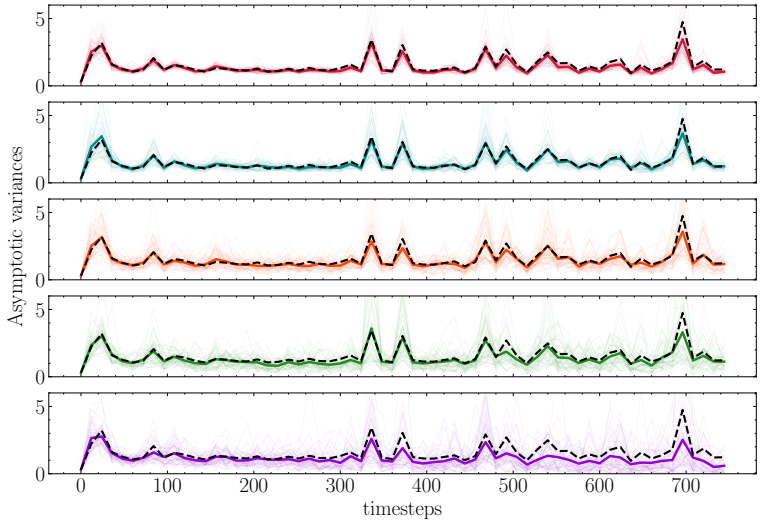

Figure 4.2: Long-term behavior of the asymptotic variance estimators up to $t = 750$*. From top to bottom: PaRIS version of* $\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}$ *with* $M = 3$*, lagged estimators with (in order)* $\lambda \in \{20, 100, 200, 750\}$*. The case* $\lambda = 750$ *corresponds to the CLE estimator. For each estimator, the blurred colored lines represent each run out of fifty runs and solid colored lines correspond to their average. The black dashed line is the asymptotic variance obtained by brute force. The number of particles* N *is set to* 2000*.*

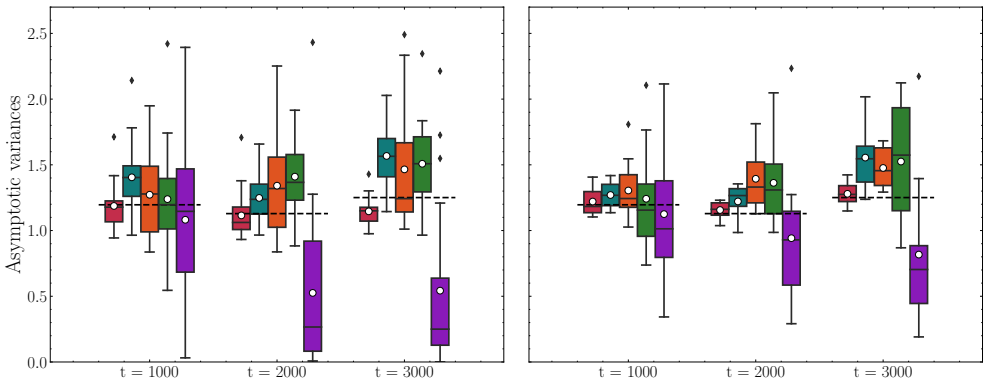

Figure 4.3: Long-term behavior of the asymptotic variance estimates up to $t = 3000$. White dots represent the average of the asymptotic variance estimates of each algorithm. The dashed black lines correspond to the asymptotic variances estimated by brute force. N is set to 5000 on the left boxplot and 10000 on the right one. The boxplots at each time step from left to right are: $\mathsf{V}_{\eta,t}^{N,M}$ with $M = 3$ and then the lagged CLEs with $\lambda \in \{20, 100, 200, 3000\}$.

a ratio of two estimators with increasing bias and variance. Nonetheless, our estimator remains competitive with the best fixed lag estimator. Doubling the number of particles decreases both bias and variance, as highlighted by the right plot. The computational cost of $\mathsf{V}_{\eta,t}^{N,M}$ is approximately twice larger than that of the genealogy tracing estimators when $N = 5000$ as shown in Figure B.1 in the supplementary material. However, we are able to maintain a small variance and bias without having to tune any other parameter besides the sample size. In more complicated models or realistic scenarios, we do not know in advance which lag size is suitable and using an inappropriate lag might yield poor estimates. We further investigate the stability of our estimator with respect to t by comparing

$$D_N^{\mathsf{BS}}(t) := \sum_{i,j\in[N]^2} \mathcal{T}_t^{\mathbf{0}}(i,j)/N(N-1), \quad D_N^{\mathsf{GT}}(t) := \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i \neq \mathrm{E}_{t,0}^j}/N(N-1)\,,$$

which are central in the expression of the variance estimators (see Remark 4.4.6). We also compare $E_N^{\mathsf{BS}}(t) := \left|(\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}(\mathrm{Id})/\mathcal{V}_{\mathrm{t}}^{\infty}(\mathrm{Id})) - 1\right|$ and $E_N^{\mathsf{GT}}(t)$ which is defined in an analogous way. For the CLE (4.3.11), although it is expected to collapse to 0 after $\mathcal{O}(N)$ timesteps following Koskela et al. (2020), the estimator starts to exhibit high bias and variance much before as the set of time 0 ancestors depletes at a fast rate. This is illustrated on the left plot of Figure 4.4 where we fix N to 1000 and vary t between 0 and 3000. We see that $D_N^{\mathsf{GT}}(t)$ decreases much faster than $D_N^{\mathsf{BS}}(t)$ and this in turn translates into longer stability for our estimator as can be seen on the right plot of the same figure.

4.6.2 Asymptotic variance of the smoother

Here we are interested in the estimation of the asymptotic variance associated to the FFBS estimates of the marginal means $\phi_{\ell|t}(\mathrm{Id})$ with ℓ fixed and $t \geq \ell$ varying using Algorithm 3. For this example we sampled four different observation records of length 160 each and ℓ is set to 100. The real asymptotic variances of each $\phi_{\ell|t}(\mathrm{Id})$ are intractable and they are estimated using 1000 independent replicates of the marginal means $\phi_{\ell|t}^N(\mathrm{Id})$ with $N = 10000$. We then multiply the obtained sample variance by N. The results are reported in Figure 4.5. As expected, the crude estimates of the asymptotic variances all stagnate after some time t due to the incoming observations becoming less and less informative as t grows and thus no longer influencing the value of $\phi_{\ell|t}(\mathrm{Id})$.

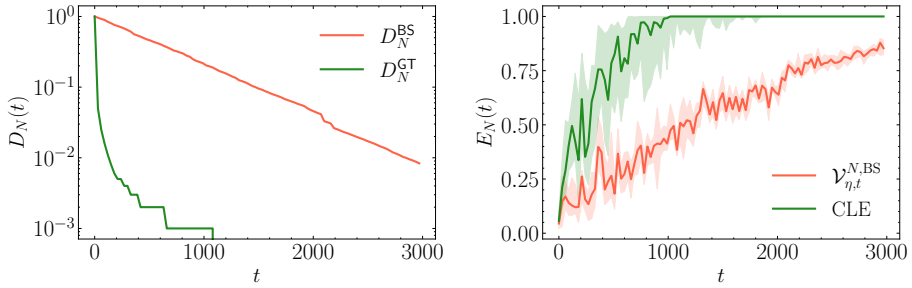

Figure 4.4: Dependency on the time t of the variance estimators. The right plot displays the empirical error $E_N(t)$ for both BS *and* GT *with N fixed to* 1000*. We display the median and the interquartile range over* 30 *runs. The left plot displays the median of $D_N(t)$ associated with the* BS *and* GT *variance estimators used on the right plot.*

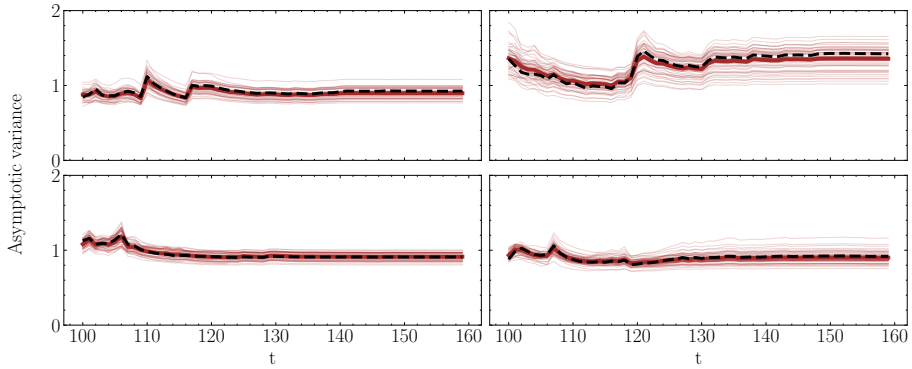

Figure 4.5: Asymptotic variance estimates for four different observation records of the marginal mean $\phi_{100|t}^N(\mathrm{Id})$ where $t \in [100, 160]$. The blurred brown lines on the left plot represent 50 runs and the solid brown line their average. The black dashed line is the crude variance estimator. The number of particles N is set to 5000*.*

The estimator proposed in 4.5.3 captures well the behavior of the asymptotic variance with little variance and also stagnates at the same time. We observed in our experiments that, in comparison with the variance estimators of filtering algorithms, for this estimator to provide good performance more samples are required for shorter time horizons. Nonetheless, the increased computational time incurred by the increase of the number of particles is to be compared with the time it takes to compute the crude variance estimates up to $t = 160$, which is about 1 hour when running on GPU. In comparison, one run of our estimator takes 3 minutes.

4.7 Conclusion and perspectives

We have derived a novel estimator for the asymptotic variance of the particle filter relying on backward sampling instead of genealogy tracking, thus extending the works of Chan and Lai (2013); Lee and Whiteley (2018); Du and Guyader (2021); Olsson and Douc (2019). Our estimator has many similarities with the forward only FFBS for additive functionals and is treated as such; we also derived a PaRIS version Olsson and Westerborn (2017) of the estimator. As a second contribution, we have derived the first consistent asymptotic variance estimator for the FFBS algorithm for which we gave a practical online implementation in the case of additive smoothing.

After this paper was released, a novel estimator for the asymptotic variance of the particle filter

was released Mastrototaro and Olsson (2023). They extend the fixed lag estimator of Olsson and Douc (2019) into an estimator that adaptively calibrates the lag and is essentially free of any tuning parameter. We believe that their methodology can be readily combined with our estimator for the FFBS variance and that our proof technique can be adapted.

Chapter 5

State and parameter learning with PaRIS particle Gibbs

5.1 Introduction

Sequential Monte Carlo (SMC) *methods*, or *particle filters*, are simulation-based approaches used for the online approximation of posterior distributions in the context of Bayesian inference in state space models. In nonlinear *hidden Markov models* (HMM), they have been successfully applied for approximating online the typically intractable posterior distributions of sequences of unobserved states $(X_{s_1}, \dots, X_{s_2})$ given observations $(Y_{t_1}, \dots, Y_{t_2})$ for $0 \leq s_1 \leq s_2$ and $0 \leq t_1 \leq t_2$. Standard SMC methods use Monte Carlo samples generated recursively by means of sequential importance sampling and resampling steps. A particle filter approximates the flow of marginal posteriors by a sequence of occupation measures associated with a sequence $\{\xi_t^i\}_{i=1}^N$, $t \in \mathbb{N}$, of Monte Carlo samples, each *particle* ξ_t^i being a random draw in the state space of the hidden process. Particle filters revolve around two operations: a *selection step* duplicating/discarding particles with large/small importance weights, respectively, and a *mutation step* evolving randomly the selected particles in the state space. Applying alternatingly and iteratively selection and mutation results in swarms of particles being both temporally and spatially dependent. The joint state posteriors of an HMM can also be interpreted as laws associated with a Markovian backward dynamics; this interpretation is useful, for instance, when designing backward-sampling-based particle algorithms for nonlinear smoothing Douc et al. (2011a); Del Moral et al. (2010c).

Throughout the years, several convergence results as the number N of particles tends to infinity have been established; see, *e.g.*, Del Moral (2004); Douc and Moulines (2008); Cappe et al. (2005) and the references therein. In addition, a number of non-asymptotic results have been established, including time-uniform bounds on the SMC L_p error and bias as well as bounds describing the propagation of chaos among the particles. Extensions to the backward-sampling-based particle algorithms can also be found for instance in Douc et al. (2011a); Del Moral et al. (2010c); Dubarry and Le Corff (2013).

In this chapter, we consider the problem of parameter learning with stochastic gradient algorithms. We set the focus on learning the parameter of a function whose gradient is the smoothed expectation of an additive functional, i.e. can be written $\eta_{0:t} h_t = \mathbb{E}[h_t(X_{0:t}) \mid Y_{0:t}]$ for additive functionals h_t in the form

$$h_t(x_{0:t}) := \sum_{s=0}^{t-1} \tilde{h}_s(x_s, x_{s+1}), \tag{5.1.1}$$

where $X_{0:n}$ and $Y_{0:n}$ denote vectors of states and observations (see below for precise definitions). Such expectations appear frequently in the context of maximum-likelihood parameter estimation in nonlinear HMMs, for instance, when computing the score function (the gradient of the log-likelihood function) or the Expectation Maximization intermediate quantity; see Cappé (2001); Andrieu and Doucet (2003); Poyiadjis et al. (2005); Cappé (2011); Poyiadjis et al. (2011); Le Corff and Fort (2013). In this specific context, where a smoothing estimator is employed repeatedly to produce mean-field estimates, controlling the bias and the MSE of the estimator becomes critical (see Karimi et al. (2019)). This learning problem is usually tackled using either the Particle Gibbs Lindholm and Lindsten (2018), or classical smoothing algorithms such as the FFBSi or the `PARIS` Olsson and Westerborn (2017). While the former has exponentially decreasing bias (w.r.t the number of iterates) under standard assumptions, it usually results in high variance and a huge waste of the particle cloud generated. The latter is biased, since it is self normalised but results in smaller variance that the particle Gibbs. Recently, zero bias estimators (see Jacob et al. (2020); Lee et al. (2020)) have been proposed based on the coupling of the particle Gibbs that could be used in this framework, but they suffer from having a random computational complexity and high variance.

We propose a new algorithm combining the `PARIS` and the particle Gibbs algorithms. The conditional particle cloud resulting from the particle Gibbs is now used not only to generate the next conditioning trajectory as in the usual particle Gibbs but it is also used to generate a smoothing estimate, reducing waste of computational work.

This leads to a batch mode *`PARIS` particle Gibbs* (`PPG`) *sampler*, which we furnish with an upper bound on the bias that decreases inversely proportionally to the number N of particles and exponentially fast with the particle Gibbs iteration index (under the assumption that the particle Gibbs sampler is uniformly ergodic), while keeping the MSE comparable to that of the underlying backward smoother. Furthermore, in the context of score ascent with the `PPG` we provide a non-asymptotic bound for the expectation of the squared gradient in terms of bias and MSE of the `PPG`. This bound establishes an $\mathcal{O}(\log(n)/\sqrt{n})$ convergence of the learning procedure. This chapter and its contributions are structured as follows.

- In Section 5.3, we lay out the methodology of our Particle Gibbs within smoothing algorithm, coined the `PPG` algorithm. We then provide an upper bound on its bias and MSE as a function of the number of particles and the iteration index of the Gibbs algorithm, see Theorem 5.3.1.
- In Section 5.4, we undertake the learning problem and present the second result of this chapter, a $\mathcal{O}(\log(n)/\sqrt{n})$ non-asymptotic bound on the expectation of the squared gradient norm taken at a random index K, see Theorem 5.4.1.
- In Section 6.3, we illustrate our results through numerical experiments, showing that our algorithm is empirically grounded.

Notation. For a given measurable space $(\mathsf{X}, \mathcal{X})$, where $\mathcal{X}$ is a countably generated σ-algebra, we denote by $\mathsf{F}(\mathcal{X})$ the set of bounded $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable functions on X. For any $h \in \mathsf{F}(\mathcal{X})$, we let $\|h\|_\infty := \sup_{x\in\mathsf{X}} |h(x)|$ and $\mathrm{osc}(h) := \sup_{(x,x')\in\mathsf{X}^2} |h(x) - h(x')|$ denote the supremum and oscillator norms of h, respectively. Let $\mathsf{M}(\mathcal{X})$ be the set of σ-finite measures on $(\mathsf{X}, \mathcal{X})$ and $\mathsf{M}_1(\mathcal{X}) \subset \mathsf{M}(\mathcal{X})$ the probability measures. For any $h \in \mathsf{F}(\mathcal{X})$ and $\mu \in \mathsf{M}(\mathcal{X})$ we write $\mu(h) = \int h(x)\mu(\mathrm{d}x)$. For a Markov kernel K from $(\mathsf{X}, \mathcal{X})$ to another measurable space $(\mathsf{Y}, \mathcal{Y})$, we define the measurable function $Kh : \mathsf{X} \ni x \mapsto \int h(y)K(x, \mathrm{d}y)$. The composition μK is a probability measure on $(\mathsf{Y}, \mathcal{Y})$ such that $\mu K : \mathcal{X} \ni A \mapsto \int \mu(\mathrm{d}x)K(x, \mathrm{d}y)\mathbb{1}_A(y)$. For all sequences $\{a_u\}_{u\in\mathbb{Z}}$ and $\{b^u\}_{u\in\mathbb{Z}}$, and all $s \leq t$ we write $a_{s:t} = \{a_s, \ldots, a_t\}$ and $b^{s:t} = \{b^s, \ldots, b^t\}$.

5.2 Background

5.2.1 Hidden Markov models

Hidden Markov models consist of an unobserved state process $\{X_t\}_{t\in\mathbb{N}}$ and observations $\{Y_t\}_{t\in\mathbb{N}}$, where, at each time $t \in \mathbb{N}$, the unobserved state X_t and the observation Y_t are assumed to take values in some general measurable spaces $(\mathsf{X}_t, \mathcal{X}_t)$ and $(\mathsf{Y}_t, \mathcal{Y}_t)$, respectively. It is assumed that $\{X_t\}_{t\in\mathbb{N}}$ is a Markov chain with transition kernels $\{M_t\}_{t\in\mathbb{N}}$ and initial distribution η_0. Given the states $\{X_t\}_{t\in\mathbb{N}}$, the observations $\{Y_t\}_{t\in\mathbb{N}}$ are assumed to be independent and such that for all $t \in \mathbb{N}$, the conditional distribution of the observation Y_t depends only on the current state X_t. This distribution is assumed to admit a density $g_t(X_t, \cdot)$ with respect to some reference measure. In the following we assume that we are given a fixed sequence $\{y_t\}_{t\in\mathbb{N}}$ of observations and define, abusing notations, $g_t(\cdot) = g_t(\cdot, y_t)$ for each $t \in \mathbb{N}$. We denote, for $0 \leq s \leq t$, $\mathsf{X}_{s:t} := \prod_{u=s}^{t} \mathsf{X}_u$ and $\mathcal{X}_{s:t} := \bigotimes_{u=s}^{t} \mathcal{X}_u$. Consider the unnormalized transition kernel

$$Q_s : \mathsf{X}_s \times \mathcal{X}_{s+1} \ni (x, A) \mapsto g_s(x) M_s(x, A) \tag{5.2.1}$$

and let

$$\gamma_{0:t} : \mathcal{X}_{0:t} \ni A \mapsto \int \mathbb{1}_A(x_{0:t})\, \eta_0(\mathrm{d}x_0) \prod_{s=0}^{t-1} Q_s(x_s, \mathrm{d}x_{s+1}). \tag{5.2.2}$$

Using these quantities, we may define the *joint-smoothing* and *predictor distributions* at time $t \in \mathbb{N}$ as

$$\eta_{0:t} : \mathcal{X}_{0:t} \ni A \mapsto \frac{\gamma_{0:t}(A)}{\gamma_{0:t}(\mathsf{X}_{0:t})}, \tag{5.2.3}$$

$$\eta_t : \mathcal{X}_t \ni A \mapsto \eta_{0:t}(\mathsf{X}_{0:t-1} \times A), \tag{5.2.4}$$

respectively. It can be shown (see (Cappe et al., 2005, Section 3)) that $\eta_{0:t}$ and η_t are the conditional distributions of $X_{0:t}$ and X_t given $Y_{0:t-1}$ respectively, evaluated at $y_{0:t-1}$. Unfortunately, these distributions, which are vital in Bayesian smoothing and filtering as they enable the estimation of hidden states through the observed data stream, are available in a closed form only in the cases of linear Gaussian models or models with finite state spaces; see Cappe et al. (2005) for a comprehensive coverage.

5.2.2 Particle filters

For most models of interest in practice, the joint smoothing and predictor distributions are intractable, and so are also any expectation associated with these distributions. Still, such expectations can typically be efficiently estimated using *particle methods*, which are based on the predictor recursion $\eta_{t+1} = \eta_t Q_t / \eta_t g_t$. At time t, if we assume that we have at hand a consistent particle approximation of η_t, formed by N random draws $\{\xi_t^i\}_{i=1}^N$, so-called *particles*, in X_t and given by $\eta_t^N = N^{-1} \sum_{i=1}^N \delta_{\xi_t^i}$, plugging η_t^N into the recursion tying η_{t+1} and η_t yields the mixture $\eta_t^N Q_t$, from which a sample of N new particles can be drawn in order to construct η_{t+1}^N. To do so, we sample, for all $1 \leq i \leq N$, ancestor indices $\alpha_t^i \sim \mathrm{Categorical}(\{g_t(\xi_t^\ell)\}_{\ell=1}^N)$ and then propagate $\xi_{t+1}^i \sim M_t(\xi_t^{\alpha_t^i}, \cdot)$. This procedure, which is initialized by sampling the initial particles $\{\xi_0^i\}_{i=1}^N$ independently from η_0, describes the particle filter with multinomial resampling and produces consistent estimators such that for every $h \in \mathsf{F}(\mathsf{X}_t)$, $\eta_t^N(h)$ converges almost surely to $\eta_t(h)$ as the number N of particles tends to infinity.

This procedure can also be extended to produce particle approximations of the joint-smoothing distributions $\{\eta_{0:t}\}_{t\in\mathbb{N}}$. Note that the successive ancestor selection steps described previously

generates an ancestor line for each terminal particle ξ_t^i, which we denote by $\xi_{0:t}^i$. It can then be easily shown that $\eta_{0:t}^N = N^{-1} \sum_{i=1}^N \delta_{\xi_{0:t}^i}$ forms a particle approximation of the joint-smoothing distribution $\eta_{0:t}$. However, it is well known that the same selection operation also depletes the ancestor lines, since, at each step, two different particles are likely to originate from the same ancestor from the previous generation. Thus, eventually, all the particles end up having a large portion of their initial ancestry in common. This means that in practice, this naive approach, which we refer to as the *poor man's smoother*, suffers generally from high variance when used for estimating joint-smoothing expectations of objective functionals depending on the whole state trajectory.

5.2.3 Backward smoothing and the `PARIS` algorithm

We now discuss how to avoid the problem of particle degeneracy relative to the smoothing problem by means of so-called *backward sampling*. While this line of research has broader applicability, we restrict ourselves for the sake of simplicity to the case of *additive state functionals* in the form

$$h_t(x_{0:t}) := \sum_{s=0}^{t-1} \tilde{h}_s(x_{s:s+1}), \quad x_{0:t} \in \mathsf{X}_{0:t}. \tag{5.2.5}$$

Appealingly, using the poor man's smoother described in the previous section, smoothing of additive functionals can be performed online alongside the particle filter by letting, for each s,

$$\eta_{0:s}^N h_s := N^{-1} \sum_{i=1}^N \beta_s^i, \tag{5.2.6}$$

where the statistics $\{\beta_s^i\}_{i=1}^N$ satisfy the recursion

$$\beta_{s+1}^i = \beta_s^{\alpha_s^i} + \tilde{h}_s(\xi_s^{\alpha_s^i}, \xi_{s+1}^i), \tag{5.2.7}$$

where α_s^i is, as described, the ancestor at time s of particle ξ_{s+1}^i.

As mentioned above, the previous estimator suffers from high variance when s is relatively large with respect to N. However, assume now that the model is *fully dominated* in the sense that each state process kernel M_s has a transition density m_s with respect to some reference measure; then, interestingly, it is easily seen that the conditional probability that $\alpha_s^i = j$ given the offspring ξ_{s+1}^i and the ancestors $\{\xi_s^\ell\}_{\ell=1}^N$ is given by

$$\mathbf{\Lambda}_s(i,j) := \frac{g_s(\xi_s^j) m_s(\xi_s^j, \xi_{s+1}^i)}{\sum_{\ell=1}^N g_s(\xi_s^\ell) m_s(\xi_s^\ell, \xi_{s+1}^i)}. \tag{5.2.8}$$

Here $\mathbf{\Lambda}_s$ forms a backward Markov transition kernel on $[1:N] \times [1:N]$. Using this observation, we may avoid completely the particle-path degeneracy of the poor man's smoother by simply replacing the naive update (5.2.7) by the Rao–Blackwellized counterpart

$$\beta_{s+1}^i = \sum_{j=1}^N \mathbf{\Lambda}_s(i,j) \{\beta_s^j + \tilde{h}_s(\xi_s^j, \xi_{s+1}^i)\}. \tag{5.2.9}$$

This approach, proposed in Del Moral et al. (2010c), avoids elegantly the path degeneracy as it eliminates the ancestral connection between the particles by means of averaging. Furthermore, it is entirely online since at step s only the particle populations $\xi_s^{1:N}$ and $\xi_{s+1}^{1:N}$ are needed to perform the update. Still, a significant drawback is the overall $\mathcal{O}(N^2)$ complexity for the computation of $\beta_t^{1:N}$, since the calculation of each β_{s+1}^i in (5.2.9) involves the computation of N^2

terms, which can be prohibitive when the number N of particles is large. Thus, in Olsson and Westerborn (2017), the authors propose to sample $M \ll N$ conditionally independent indices $\{J_s^{i,j}\}_{j=1}^M$ from the distribution $\mathbf{\Lambda}_s(i, \cdot)$ and to update the statistics according to

$$\beta_{s+1}^i = M^{-1} \sum_{j=1}^{M} \left(\beta_s^{J_s^{i,j}} + \tilde{h}_s(\xi_s^{J_s^{i,j}}, \xi_{s+1}^i) \right). \tag{5.2.10}$$

The key aspect of this approach is that the number M of sampled indices at each step can be very small; indeed, for any fixed $M \geq 2$, the algorithm, which is referred to as the `PARIS`, can be shown to be stochastically stable with an $\mathcal{O}(t)$ variance (see (Olsson and Westerborn, 2017, Section 1) for details), and setting M to 2 or 3 yields typically fully satisfying results.

Let us end this section by mentioning that the `PARIS` estimator can be viewed as an alternative to the FFBSm Doucet et al. (2000), rather than the FFBSi Godsill et al. (2004). Even if the `PARIS` and FFBSi are both randomised versions of the FFBSm estimator, the `PARIS` is of a different nature than the FFBSi. The `PARIS` approximates the forward-only FFBSm online in the context of additive functionals by approximating each updating step by additional Monte Carlo sampling. The sample size M is an accuracy parameter that determines the precision of this approximation, and by increasing M the statistical properties of the `PARIS` approaches those of the forward-only FFBSm (see (Olsson and Westerborn, 2017, Theorem 8)). On the other hand, as shown in (Douc et al., 2011a, Corollary 9), the asymptotic variance of FFBSi is always larger than that of the FFBSm, with a gap given by the variance of the state functional under the joint-smoothing distribution. Thus, we expect, especially in the case of a low signal-to-noise ratio, the `PARIS` estimator to be more accurate than the FFBSi for a given computational budget. The methodology we develop next can be seamlessly extended to the FFBSm and FFBSi algorithms but since the `PARIS` has a practical edge w.r.t. the FFBSi, we chose to center our contribution around it although the main idea behind our chapter is more general.

5.3 `PARIS` particle Gibbs

5.3.1 Particle Gibbs methods

The *conditional particle filter* (CPF) introduced in Andrieu et al. (2010) serves the basis of a particle-based MCMC algorithm targeting the joint-smoothing distribution $\eta_{0:t}$. Let $\ell \in \mathbb{N}^*$ be an iteration index and $\zeta_{0:t}[\ell]$ a conditional path used at iteration ℓ of the CPF to construct a particle approximation of $\eta_{0:t}$ as follows. At step $s \in [1:t]$ of the CPF, a randomly selected particle, with uniform probability $1/N$, is set to $\zeta_s[\ell]$, whereas the remaining $N-1$ particles are all drawn from the mixture $\eta_{s-1}^N Q_{s-1}$. At the final step, a new particle path $\zeta_{0:t}[\ell+1]$ is drawn either:

- by selecting randomly, again with uniform probability $1/N$, a genealogical trace from the ancestral tree of the particles $\{\xi_s^{1:N}\}_{s=0}^t$ produced by the CPF, as in the vanilla particle Gibbs sampler;
- or by generating the path by means of backward sampling, *i.e.*, by drawing indices $J_{0:t}$ backwards in time according to $J_t \sim \text{Categorical}(\{1/N\}_{i=1}^N)$ and, conditionally to J_{s+1}, $J_s \sim \mathbf{\Lambda}_s(J_{s+1}, \cdot)$, $s \in [0:t-1]$, and letting $\zeta_{0:t}[\ell+1] = (\xi_0^{J_0}, \dots \xi_t^{J_t})$, where the transition kernels $\{\mathbf{\Lambda}_s\}_{s=0}^t$, defined by (5.2.8), are induced by the particles produced by the CPF, as proposed in Whiteley (2010).

The theoretical properties of the different versions of the particle Gibbs sampler are well studied Singh et al. (2017); Chopin and Singh (2015b); Andrieu et al. (2018a). In short, the produced

conditional paths $(\zeta_{0:t}[\ell])_{\ell\in\mathbb{N}}$ form a Markov chain whose marginal law converges geometrically fast in total variation to the target distribution $\eta_{0:t}$. As it is the case for smoothing algorithms, the vanilla particle Gibbs sampler suffers from bad mixing due to particle path degeneracy while its backward-sampling counterpart exhibits superior performance as t increases Lee et al. (2020).

5.3.2 The PPG algorithm

Remarkably, for the standard particle Gibbs samplers to output a single conditional path, a whole particle cloud $\{\xi_s^{1:N}\}_{s=0}^t$ is generated and then discarded, resulting in significant waste of computational work. Thus, we now introduce a variant of the PARIS algorithm, coined the PARIS particle Gibbs (PPG), in which the conditional path of particle Gibbs with backward sampling is merged with the intermediate particles, ensuring less computational waste and reduced bias with respect to the vanilla PARIS.

In the following we let $t \in \mathbb{N}$ be a fixed time horizon, and describe in detail how the PPG approximates iteratively $\eta_{0:t}h_t$, where h_t is an additive functional in the form (5.2.5). Using a given conditional path $\zeta_{0:t}[\ell-1]$ as input, the ℓ-th iteration of the PPG outputs a many-body system $((\xi_{0:t}^1,\beta_t^1),\dots,(\xi_{0:t}^N,\beta_t^N))$ comprising N backward particle paths $\{\xi_{0:t}^i\}_{i=1}^N$ with associated PARIS statistics $\{\beta_t^i\}_{i=1}^N$ (5.2.10). This is the so-called *conditional PARIS update* detailed in Algorithm 3. After this, an updated conditional path is selected with probability $1/N$ among the N particle paths $\{\xi_{0:t}^i\}_{i=1}^N$ and used as input in the next conditional PARIS operation. At each iteration, the produced statistics $\{\beta_t^i\}_{i=1}^N$ provide an approximation of $\eta_{0:t}h_t$ according to (5.2.6). The overall algorithm is summarized in Algorithm 4. The function CPF_s describes one internal step of the conditional particle filter and is given in Algorithm 8 of the supplementary material. In addition, the PPG algorithm defines a Markov chain with Markov transition kernel denoted by $\mathbb{K}_t$ and detailed in (C.1.17).

Algorithm 3 One conditional PARIS update (cPaRIS)

Input: $\{(\xi_{0:s}^i,\beta_s^i)\}_{i=1}^{\mathbb{N}}$, ζ_{s+1}, $\tilde{h}_s$
Result: $\{(\xi_{0:s+1}^i,\beta_{s+1}^i)\}_{i=1}^{\mathbb{N}}$
draw $\xi_{s+1}^{1:N} \sim \mathsf{CPF}_{s+1}(\zeta_{s+1})$
for $i \leftarrow 1$ **to** $\mathbb{N}$ **do**
  draw $\{J_s^{i,\ell}\}_{\ell=1}^M \sim \mathbf{\Lambda}(i,\cdot)^{\otimes M}$
  set $\beta_{s+1}^i \leftarrow M^{-1}\sum_{\ell=1}^M \left(\beta_s^{J_s^{i,\ell}} + \tilde{h}_s(\xi_s^{J_s^{i,\ell}},\xi_{s+1}^i)\right)$
  set $\xi_{0:s+1}^i \leftarrow (\xi_{0:s}^{J_s^{i,1}},\xi_{s+1}^i)$

Algorithm 4 One iteration of PPG

Input: Initial path $\zeta_{0:t}$, $\{\tilde{h}_s\}_{s=0}^{t-1}$
Result: $\{\beta_t^i\}_{i=1}^N$, $\zeta'_{0:t}$
draw $\xi_0^{1:N} \sim \mathsf{CPF}_0(\zeta_0)$
set $\beta_0^i \leftarrow 0$ for $i \in [1:N]$
for $s \leftarrow 0$ **to** $t-1$ **do**
  set $\{(\xi_{0:s+1}^i,\beta_{s+1}^i)\}_{i=1}^N \leftarrow \mathsf{cPaRIS}(\{(\xi_{0:s}^i,\beta_s^i)\}_{i=1}^N,\zeta_{s+1},\tilde{h}_s)$
draw $\zeta'_{0:t} \sim N^{-1}\sum_{i=1}^N \delta_{\xi_{0:t}^i}$

As performing k steps of the PPG results in k many-body systems, it is natural to consider the

following *roll-out estimator* which combines the backward statistics from step $k_0 < k$ to k:

$$\Pi_{(k_0,k),N}(h_t) = [N(k-k_0)]^{-1} \sum_{\ell=k_0+1}^{k} \sum_{j=1}^{\mathbb{N}} \beta_t^j[\ell]. \tag{5.3.1}$$

The total number of particles used in this estimator is $C = (\mathbb{N}-1)k$ per time step. We denote by $\upsilon = (k-k_0)/k$ the ratio of the number of particles used in the estimator to the total number of sampled particles.

We now state the first main results of the present chapter, in the form of theoretical bounds on the bias and mean-squared error (MSE) of the roll-out estimator (5.3.1). These results are obtained under the following *strong mixing* assumptions, which are now standard in the literature (see Del Moral (2004); Douc and Moulines (2008); Del Moral (2013); Del Moral et al. (2016); Lee et al. (2020)). It is crucial for obtaining quantitative bounds for particle smoothing algorithms, see Olsson and Westerborn (2017) or Gloaguen et al. (2022) but also for the coupled conditional backward sampling particle filter Lee et al. (2020).

(**A9**) For every $s \in \mathbb{N}$ there exist $\underline{\tau}_s$, $\bar{\tau}_s$, $\underline{\sigma}_s$, and $\bar{\sigma}_s$ in $\mathbb{R}_{>0}$ such that

(i) $\underline{\tau}_s \leq g_s(x_s) \leq \bar{\tau}_s$ for every $x_s \in \mathsf{X}_s$,
(ii) $\underline{\sigma}_s \leq m_s(x_s, x_{s+1}) \leq \bar{\sigma}_s$ for every $(x_s, x_{s+1}) \in \mathsf{X}_{s:s+1}$.

Under (**A9**), define, for every $s \in \mathbb{N}$, $\rho_s := \max_{m\in[0:s]} \bar{\tau}_m \bar{\sigma}_m (\underline{\tau}_m \underline{\sigma}_m)^{-1}$ and, for every $\mathbb{N} \in \mathbb{N}_*$ and $t \in \mathbb{N}$ such that $\mathbb{N} > \mathbb{N}_t := (1 + 5\rho_t^2/2) \vee 2t(1+\rho_t^2)$,

$$\kappa_{N,t} := 1 - \frac{1 - (1 + 5t\rho_t^2/2)/N}{1 + 4t(1+2\rho_t^2)/N}. \tag{5.3.2}$$

Note that $\kappa_{N,t} \in (0,1)$ for all $\mathbb{N}$ and t as above.

Theorem 5.3.1. *Assume (**A9**). Then for every $t \in \mathbb{N}$, $M \in \mathbb{N}_*$, $\xi \in \mathsf{M}_1(\mathcal{X}_{0:t})$, $k_0 \in \mathbb{N}_*$, $k > k_0$ and $\mathbb{N} \in \mathbb{N}_*$ such that $N > N_t$,*

$$\begin{aligned} \left| \mathbb{E}_\xi[\Pi_{(k_0,k),N}(h_t)] - \eta_{0:t} h_t \right| &\leq \sigma_{bias} \\ \mathbb{E}_\xi\left[\left(\Pi_{(k_0,k),N}(h_t) - \eta_{0:t} h_t \right)^2 \right] &\leq \sigma_{mse}^2, \end{aligned} \tag{5.3.3}$$

where

$$\sigma_{bias} := \frac{\mathsf{c}_t^{bias} \kappa_{t,N}^{k_0} \sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty}{(k-k_0)(1-\kappa_{t,N})N}, \quad \sigma_{mse}^2 := \frac{(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty)^2}{N(k-k_0)} \left(\mathsf{c}_t^{mse} + \frac{2\mathsf{c}_t^{cov}}{N^{1/2}(1-\kappa_{t,N})} \right)$$

and $\mathsf{c}_t^{bias}, \mathsf{c}_t^{mse}$ and c_t^{cov} are constants that do not depend on N and $\mathbb{E}_\xi$ denotes the expectation under the law of the Markov chain formed by the PPG *when initialized according to ξ.*

The proof is provided in the supplementary material. One of the important ingredients for the proof is that under the smoothing distribution $\eta_{0:t}$, the PPG estimates are unbiased (see theorem C.1.6). Importantly, (5.3.3) provides a bound on the bias of the roll-out estimator that decreases exponentially with the burn-in period k_0 and is inversely proportional to the number N of particles. This means that we can improve the bias of the PARIS estimator with a better allocation of the computational resources.

5.4 Parameter learning with PPG

We now turn to parameter learning using PPG and gradient-based methods. We set the focus on learning the parameter θ of a function $V(\theta)$ whose gradient is the smoothed expectation of

an additive functional $s_{0:t,\theta}$ in the form (5.2.5). Algorithm 6 defines a stochastic approximation (SA) scheme where the noise forms a parameter dependent Markov chain with associated invariant measure π_θ. We follow the approach of Karimi et al. (2019) to establish a non-asymptotic bound over the mean field $h(\theta) := \pi_\theta s_{0:t,\theta}$. Such a setting encompasses for instance the following estimation procedures.

(1) *Score ascent.* In the case of fully dominated HMMs, we are often interested in optimizing the log-likelihood of the observations given by $V(\theta) = \log \int \gamma_{0:t,\theta}(\mathrm{d}x_{0:t})$. By applying *Fisher's identity*, we may express its gradient as a smoothed expectation of an additive functional according to

$$\begin{aligned}\nabla_\theta V(\theta) &= \int \nabla_\theta \log \gamma_{0:t}(x_{0:t})\, \eta_{0:t,\theta}(\mathrm{d}x_{0:t}),\\ &= \int \sum_{\ell=0}^{t-1} s_{\ell,\theta}(x_\ell, x_{\ell+1})\, \eta_{0:t,\theta}(\mathrm{d}x_{0:t}),\end{aligned}$$

where $s_{\ell,\theta} : \mathsf{X}_{\ell:\ell+1} \ni (x, x') \mapsto \nabla_\theta \log\{g_{\ell,\theta}(x) m_{\ell,\theta}(x, x')\}$ and $s_{0:t,\theta} := \sum_{\ell=0}^{t-1} s_{\ell,\theta}$.

(2) *Backward KL surrogates.* Inspired by Naesseth et al. (2020), we may consider the problem of learning a surrogate model for $\eta_{0:t,\theta}$ in the form $q_\phi(x_{0:t}) = q_\phi(x_0) \prod_{\ell=0}^{t-1} q_\phi(x_{\ell+1}, x_\ell)$ by minimizing $V(\phi) = \mathrm{KL}(\eta_{0:t,\theta}, q_\phi)$.

Algorithm 5 Gradient estimation with roll-out PPG ($\widehat{\mathsf{Gd}}$)

Input: $\theta, \zeta_{0:t}[0]$, $s_{0:t,\theta}$, number k of PPG iterations, burn-in k_0.
Result: $\beta_t^{1:N}[k_0 : k], \zeta_{0:t}[k]$

for $\ell \leftarrow 0$ **to** $k-1$ **do**
  $(\tilde{\beta}_t^{1:N}[\ell+1], \zeta_{0:t}[\ell+1]) \leftarrow \mathsf{PPG}(\theta; \zeta_{0:t}[\ell], s_{0:t,\theta})$
  if $\ell \geq k_0 - 1$ **then**
    set $\beta_t^{1:N}[\ell+1] = \tilde{\beta}_t^{1:N}[\ell+1]$

Algorithm 6 Score ascent with PPG.

Input: θ_0, $\zeta_{0:t}[0]$, number k of PPG iterations, burn-in k_0, number of SA iterations n, learning-rate sequence $\{\gamma_\ell\}_{\ell \in \mathbb{N}}$.
Result: θ_n

for $i \leftarrow 0$ **to** $n-1$ **do**
  $\beta_t^{1:N}[k_0 : k], \zeta_{0:t}[i+1] \leftarrow \widehat{\mathsf{Gd}}(\theta_i, \zeta_{0:t}[i], s_{0:t,\theta_i}, k, k_0)$
  set $\Pi_{(k_0,k),N}(s_{0:t,\theta_i}) = \frac{1}{N(k-k_0)} \sum_{\ell=k_0}^{k-1} \sum_{j=1}^{N} \beta_t^j[\ell]$
  set $\theta_{i+1} \leftarrow \theta_i + \gamma_{i+1} \Pi_{(k_0,k),N}(s_{0:t,\theta_i})$

Note that Algorithm 5 defines a (collapsed) Markov kernel $\mathbb{P}_{\theta,t}$ defining for each path $\zeta_{0:t}$ a measure $\mathbb{P}_{\theta,t}(\zeta_{0:t}, \mathrm{d}(\tilde{\zeta}_{0:t}, \tilde{\beta}_t^{1:N}[k_0 : k]))$ over the extended space of paths and sufficient statistics. Note that by evaluating the function $b_t^{1:N}[k_0 : k] \mapsto [N(k-k_0)]^{-1} \sum_{\ell=k_0+1}^{k} \sum_{j=1}^{N} b_t^j[\ell]$ at a realisation of this kernel gives the roll-out estimator whose properties are analysed in Theorem 5.3.1. The Markov kernel $\mathbb{P}_{\theta,t}$ is detailed in (C.2.4).

The following assumptions, are vital when analysing the convergence of Algorithm 6.

(**A10**) (i) The function $\theta \mapsto V(\theta)$ is L^V-smooth.
(ii) The function $\theta \mapsto \eta_{0:t,\theta}$ is L^η-Lipschitz in total variation distance.

(iii) For each path $\zeta_{0:t} \in \mathsf{X}_{0:t}$, the function $\theta \mapsto K_{\theta,t}(\zeta_{0:t}, \mathrm{d}\tilde{\zeta}_{0:t})$ is L_1^P-Lipschitz in total variation distance, where $K_{\theta,t}$ is the path-marginalized Markov transition kernel associated with the PPG algorithm when the model is parameterized by θ, see (C.1.17).

(iv) For each path $\zeta_{0:t} \in \mathsf{X}_{0:t}$, the function

$$\theta \mapsto \mathbb{P}_{\theta,t}\Pi_{k_0-1,k,N}(s_{0:t,\theta})(\zeta_{0:t}) \tag{5.4.1}$$

is L_2^P-Lipschitz in total variation distance.

In the case of score ascent we check, in Section C.2, that these assumptions hold if the strong mixing assumption (**A9**) is satisfied uniformly in θ, and with additional assumptions on the model. We are now ready to state a bound on the mean field $h(\theta)$ for Algorithm 6.

Theorem 5.4.1. *Assume (**A9**) uniformly in θ and (**A10**) and suppose that the stepsizes $\{\gamma_{\ell+1}\}_{\ell\in[0:n]}$ satisfy $\gamma_{\ell+1} \leq \gamma_\ell$, $\gamma_\ell < a\gamma_{\ell+1}$, $\gamma_\ell - \gamma_{\ell+1} < a'\gamma_\ell^2$ and $\gamma_1 \leq 0.5(L^V + C_h)$ for some $a > 0$, $a' > 0$ and all $n \in \mathbb{N}$. Then,*

$$\mathbb{E}\left[\|h(\theta_\varpi)\|^2\right] \leq 2\frac{V_{0,n} + C_{0,n} + C_{0,\gamma}\sum_{k=0}^n \gamma_{k+1}^2}{\sum_{k=0}^n \gamma_{k+1}}\,, \tag{5.4.2}$$

where $V_{0,n} = \mathbb{E}\left[V(\theta) - V(\theta_n)\right]$ and

$$C_{0,n} := \gamma_1 h(\theta_0)C_0 + \sigma_{bias}(\gamma_1 - \gamma_{n+1} + 1)\delta_{k,N,t}^{-1}\,, \tag{5.4.3}$$

$$C_{0,\gamma} := \sigma_{mse}^2 L^V + \sigma_{mse}C_1 + \sigma_{bias}L^V\delta_{k,N,t}^{-1} + \sigma_{mse}\sigma_{bias}\left(L^V + \frac{C_2}{1-\kappa_{N,t}}\right)\delta_{k,N,t}^{-1}\,, \tag{5.4.4}$$

$$C_h := (L^V + a' + 1)\sigma_{bias}\delta_{k,N,t}^{-1} + \left(C_1 + \frac{\sigma_{bias}C_2}{(1-\kappa_{N,t})\delta_{k,N,t}}\right)\left[\frac{a+1}{2} + a\sigma_{mse}\right]\,, \tag{5.4.5}$$

$$C_1 = L_2^P\left[1 + \kappa_{N,t}^k\delta_{k,N,t}^{-1}\right] + L^V \tag{5.4.6}$$

$$C_2 = L_1^P\delta_{k,N,t}^{-1} + L^\eta\kappa_{N,t}^k\,. \tag{5.4.7}$$

where C_0 is independent of $\sigma_{bias}, \sigma_{mse}, N$ and where $\delta_{k,N,t} = 1 - \kappa_{N,t}^k$.

Theorem 5.4.1 establishes not only the convergence of Algorithm 6, but also illustrates the impact of the bias and the variance of the PPG on the convergence rate.

Remark 5.4.2. *Under additional assumptions on the model (cf Section C.2), if we consider $\gamma_1 \leq 0.5(L^V + C_h)$, $\gamma_\ell = \gamma_1\ell^{-1/2}$ for all $\ell \in [1:n]$, then $\sum_{k=0}^n \gamma_{k+1}^2 / \sum_{k=0}^n \gamma_{k+1} \sim \log n/\sqrt{n}$, showing that $\mathbb{E}\left[\|h(\theta_\varpi)\|^2\right]$ is $\mathcal{O}(\log n/\sqrt{n})$, where the leading constant depends on σ_{bias} and σ_{mse}.*

Remark 5.4.2 establishes the rate of convergence of Algorithm 6. In principle we could try to optimize the parameters k, k_0 and N of the algorithm using these bounds, but one of the main challenges with this approach is the determination of the mixing rate, which is crudely upper bounded by $\kappa_{N,t}$. Still, our bound provides interesting information of the role of both bias and MSE.

5.5 Numerical experiments

In this section, we focus on the numerical analysis of the two main results of the chapter, namely the bias and MSE bounds of the roll-out estimator established in Theorem 5.3.1 and the efficiency of using PPG for learning in the framework developed in Section 5.4. For the latter, we will restrict ourselves to the case of parameter learning via score ascent. The code used in

this section is available [1]. Throughout this section, we set $M = 2$ for the PPG algorithm. In this setting, the competing method that corresponds most closely to the one presented here consists of using, as presented in Algorithm 7, a standard particle Gibbs sampler Π_θ instead of the PPG. One of the most common such samplers is the *particle Gibbs with ancestor sampling* (PGAS) presented in Lindsten et al. (2014). In Lindholm and Lindsten (2018), the PGAS is used for parameter learning in HMMs via the Expectation Maximization (EM) algorithm.

Algorithm 7 Score ascent with particle Gibbs kernel.

Data: $\zeta_{0:t}[0]$, θ_0, number k of paths per trajectory, burn-in k_0, number n of SA iterations, learning-rate sequence $\{\gamma_\ell\}_{\ell\in\mathbb{N}}$, $\Pi_\theta(\zeta_{0:t}, \mathrm{d}\tilde{\zeta}_{0:t})$ a Markov kernel targeting $\eta_{0:t}$.

Result: θ_n

for $i \leftarrow 0$ **to** $n-1$ **do**
  for $j \leftarrow 0$ **to** $k-1$ **do**
    sample $\tilde{\zeta}_{0:t}[j+1] \sim \Pi_\theta(\tilde{\zeta}_{0:t}[j], \cdot)$
  set $\theta_{i+1} \leftarrow \theta_i + \frac{\gamma_{i+1}}{k-k_0} \sum_{\ell=k_0+1}^{k} s_{0:t,\theta_i}(\tilde{\zeta}_{0:t}[\ell])$
  set $\zeta_{0:t}[i+1] = \tilde{\zeta}_{0:t}[k]$

5.5.1 PPG

Linear Gaussian state-space model (LGSSM). We first consider a linear Gaussian HMM

$$X_{m+1} = AX_m + Q\epsilon_{m+1}, \quad Y_m = BX_m + R\zeta_m, \quad m \in \mathbb{N}, \tag{5.5.1}$$

where $\{\epsilon_m\}_{m\in\mathbb{N}_*}$ and $\{\zeta_m\}_{m\in\mathbb{N}}$ are sequences of independent standard normally distributed random variables, independent of X_0. The coefficients A, Q, B, and R are assumed to be known and equal to 0.97, 0.60, 0.54, and 0.33, respectively. Using this parameterisation, we generate, by simulation, a record of $t = 999$ observations.

In this setting, we aim at computing smoothed expectations of the state one-lag covariance $h_t(x_{0:t}) := \sum_{m=0}^{t-1} x_m x_{m+1}$. In the linear Gaussian case, the *disturbance smoother* (see (Cappe et al., 2005, Algorithm 5.2.15)) provides the exact values of the smoothed sufficient statistics, which allows us to study the bias of the estimator for a given computational budget C. Figure 5.1 displays, for three different total budgets C, the distribution of estimates of $\eta_{0:n}h_n$ using the PARIS as well as three different configurations of the PPG corresponding to $k \in \{2, 4, 10\}$ (and $N = C/k$) with $k_0 = k/2$ and $k_0 = k/4$. The reference value is shown as a red-dashed line and the mean value of each distribution is shown as a black-dashed line. Each boxplot is based on 1000 independent replicates of the corresponding estimator. We observe that in this example, all configurations of the PPG are less biased than the equivalent PARIS estimator while maintaining comparable variance. The illustration of the bounds from Theorem 5.3.1 is postponed to Section C.3.3.

5.5.2 Score ascent

LGSSM. We consider the LGSSM with state and observation spaces being $\mathbb{R}^5$. We assume that the parameters R and Q are known and consider the inference of $\theta = (A, B)$ on the basis of a simulated sequence of $n = 999$ observations. In this setting, the M-step of the EM algorithm can be solved exactly with the disturbance smoother (Cappe et al., 2005, Chapter 11). The parameter obtained by this procedure (denoted θ_{mle}) is the reference value for any likelihood

[1] https://anonymous.4open.science/r/ppg/

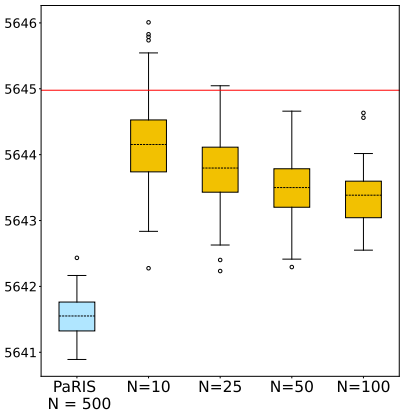

Figure 5.1: `PARIS` and `PPG` outputs for the LGSSM for $C = 500$, yellow boxes correspond to `PPG` outputs produced using $N \in \{10, 25, 50, 100\}$ iterations and $k = C/N$ particles with $k_0 = k/2$.

maximization algorithm. Table 5.1 shows the L_2 distance between the singular values of θ_{mle} and those of the parameters obtained by Algorithm 6 and Algorithm 7. The CLT confidence intervals were obtained on the basis of 25 replicates. The configurations of the `PPG` estimators respect a given particle budget $kN = C = 1024$. For a fair comparison, for each configuration of the `PPG` estimator, we run an equivalent w.r.t. clock time `PGAS` estimator. The time needed for one gradient step for each estimator averaged over 100 replicates is reported in Table 5.1. The choice of keeping $k_0 = k/2$ is a heuristic rule to achieve a good bias–variance trade-off, but other combinations of k_0 and k may lead to better performance for different problems. We analyse the impact of the different settings for the LGSMM in Section C.3.4.All settings are the same for both algorithms and are described in Section C.3.4. The `PPG` achieves consistently a smaller distance to θ_{mle}. Figure 5.2 displays, for each estimator and configuration, the evolution of the distance to the MLE estimator as a function of the iteration index.

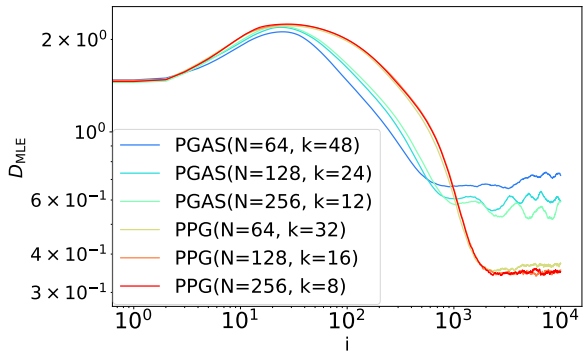

Figure 5.2: Distance to the MLE estimator as a function of the iteration step for the PGAS and `PPG` configurations from table 5.1. The solid lines and the shaded region represent the mean and CLT confidence intervals obtained with 25 replications.

CRNN. We consider now the problem of inference in a non-linear HMM and in particular the chaotic recurrent neural network introduced by Zhao et al. (2022). We use the same setting

Algorithm	N	k_0	k	D_{mle}	δt(s)
PGAS	64	24	48	0.72 ± 0.04	5.66
PGAS	128	12	24	0.59 ± 0.04	2.84
PGAS	256	6	12	0.59 ± 0.05	1.42
PPG	64	16	32	0.37 ± 0.03	4.56
PPG	128	8	16	0.36 ± 0.04	2.37
PPG	256	4	8	0.35 ± 0.04	1.57

Table 5.1: Distance to θ_{MLE} (D_{mle}) for each configuration in the LGSSM case. $\delta t(s)$ represents the average running time for each configuration.

Algorithm	N	k_0	k	NLL	δt(s)
PGAS	32	32	64	31887 ± 128	3.90
PGAS	64	16	32	31269 ± 254	1.99
PGAS	128	8	16	30994 ± 288	1.16
PPG	32	16	32	22292 ± 48	2.79
PPG	64	8	16	22315 ± 25	1.39
PPG	128	4	8	22353 ± 39	0.92

Table 5.2: Per configuration negative loglikelihood for the CRNN model.

as in the original chapter. The state and observation equations are

$$\begin{aligned} X_{m+1} &= X_m + \tau^{-1}\Delta\left(-X_m + \gamma W \tanh(X_m)\right) + \epsilon_{m+1}, \\ Y_m &= BX_m + \zeta_m, \quad m \in \mathbb{N}, \end{aligned}$$

where $\{\epsilon_m\}_{m\in\mathbb{N}_*}$ is a sequence of 20-dimensional independent multivariate Gaussian random variables with zero mean and covariance $0.01\mathbf{I}$ and $\{\zeta_m\}_{m\in\mathbb{N}}$ is a sequence of independent random variables where each component is distributed independently according to a Student's t-distribution with scale 0.1 and 2 degrees of freedom. We consider $\theta = (W, B)$.

In this case, the natural metric used to evaluate the different estimators is the negative log likelihood (NLL). We use the unbiased estimator of the likelihood given by the mean of the log weights produced by a particle filter (Douc et al., 2014, Section 12.1) using $N = 10^4$ particles. Table 5.2 shows the results obtained for 25 different replications for several different configurations of `PPG` while keeping total budget of particles fixed. As for the LGSSM, for each configuration of the `PPG` we run the time-equivalent `PGAS` estimator. Further numerical details and the system configuration used in the experiments are given in Section C.3.4. We observe that `PPG` achieves a considerably lower NLL than `PGAS` in all configurations.

5.6 Conclusion and perspectives

We have presented a new algorithm, referred to as `PPG` as well as bounds on its bias and MSE in Theorem 5.3.1. We then propose a way of using `PPG` in a learning framework and derive a non-asymptotic bound over the gradient of the updates when doing score ascent with the `PPG` with explicit dependence on the bias and MSE of the estimator. We provide numerical simulations to support our claims, and we show that our algorithm outperforms the current competitors in the two different examples analysed.

In Cardoso et al. (2022) they use a bootstrap, or reshuffling, approach to reduce the variance of their bias reduced SNIS estimator. It is not clear if such approach can be extended to

the sequential case. Note also that the Lipschitz constant of $\mathbb{K}_{\theta,t}$ possesses an unexpected dependence on $N-1$ (see Corollary C.2.8). One would expect it not to be true in that we know that $\mathbb{K}_{\theta,t}$ converges geometrically fast and uniformly to $\eta_{0:t}$ and this is faster as N gets bigger. Therefore, for large N the Lipschitz constant is expected to converge to that of $\eta_{0:t}$ whose Lipschitz constant is independent of N.

Chapter 6

Monte Carlo guided Diffusion for Bayesian linear inverse problems

6.1 Introduction

This paper is concerned with linear inverse problems $y = \mathrm{A}x + \sigma_y \varepsilon$, where $y \in \mathbb{R}^{\mathsf{d}}_y$ is a vector of indirect observations, $x \in \mathbb{R}^{\mathsf{d}_x}$ is the vector of unknowns, $\mathrm{A} \in \mathbb{R}^{\mathsf{d}_y \times \mathsf{d}_x}$ is the linear forward operator and $\varepsilon \in \mathbb{R}^{\mathsf{d}_y}$ is an unknown noise vector. This general model is used throughout computational imaging, including various tomographic imaging applications such as common types of magnetic resonance imaging Vlaardingerbroek and Boer (2013), X-ray computed tomography Elbakri and Fessler (2002), radar imaging Cheney and Borden (2009), and basic image restoration tasks such as deblurring, superresolution, and image inpainting González et al. (2009). The classical approach to solving linear inverse problems relies on prior knowledge about x, such as its smoothness, sparseness in a dictionary, or its geometric properties. These approaches attempt to estimate a $\widehat{x}$ by minimizing a regularized inverse problem, $\hat{x} = \mathrm{argmin}_x\{\|y - \mathrm{A}x\|^2 + \mathrm{Reg}(x)\}$, where Reg is a regularization term that balances data fidelity and noise while enabling efficient computations. However, a common difficulty in the regularized inverse problem is the selection of an appropriate regularizer, which has a decisive influence on the quality of the reconstruction.

Whereas regularized inverse problems continue to dominate the field, many alternative **statistical formulations** have been proposed; see Besag et al. (1991); Idier (2013); Marnissi et al. (2017) and the references therein - see also Stuart (2010) for a mathematical perspective. A main advantage of **statistical approaches** is that they allow for **uncertainty quantification** in the reconstructed solution; see Dashti and Stuart (2017). The **Bayes' formulation** of the regularized inverse problem is based on considering the indirect measurement Y, the state X and the noise ε as random variables, and to specify $p(y|x)$ the *likelihood* (the conditional distribution of Y at X) and the prior $p(x)$ (the distribution of the state). One can use Bayes' theorem to obtain the **posterior distribution** $p(x|y) \propto p(y|x)p(x)$, where " $\propto$ " means that the two sides are equal to each other up to a multiplicative constant that does not depend on x. Moreover, the use of an appropriate method for Bayesian inference allows the quantification of the uncertainty in the reconstructed solution x. A variety of priors are available, including but not limited to Laplace priors Figueiredo et al. (2007), total variation (TV) priors Kaipio et al. (2000) and mixture-of-Gaussians priors Fergus et al. (2006). In the last decade, a variety of techniques have been proposed to design and train generative models (GM) capable of producing perceptually realistic images Kingma and Welling (2019); Kobyzev et al. (2020); Gui

et al. (2021). Denoising diffusion models have been shown to be particularly effective generative models in this context Sohl-Dickstein et al. (2015); Song et al. (2021c,a,b); Benton et al. (2022). These models convert noise into natural images through a series of denoising steps. A popular approach is to use a fixed, generic diffusion model that has been pre-trained for image generation, eliminating the need for re-training and making the process more efficient and versatile Trippe et al. (2023); Zhang et al. (2023).

Although this was not the main motivation for developing GM, these models can of course be used as prior distributions in Bayesian inverse problems. This simple observation has led to a new, fast-growing line of research on how linear inverse problems can benefit from the flexibility and expressive power of the recently introduced deep generative models; see Arjomand Bigdeli et al. (2017); Wei et al. (2022); Su et al. (2022); Kaltenbach et al. (2023); Shin and Choi (2023); Zhihang et al. (2023); Sahlström and Tarvainen (2023) and the references therein.

Contributions

- We propose `MCGdiff`, a novel algorithm for sampling from the Bayesian posterior of Gaussian linear inverse problems with denoising diffusion model priors. `MCGdiff` specifically exploits the structure of both the linear inverse problem and the denoising diffusion generative model to design an efficient Sequential Monte Carlo (SMC) sampler.

- We establish under sensible assumptions that the empirical distribution of the samples produced by `MCGdiff` converges to the target posterior when the number of particles goes to infinity. To the best of our knowledge, `MCGdiff` is the first provably consistent algorithm for conditional sampling from the denoising diffusion posteriors.

- To evaluate the performance of `MCGdiff`, we perform numerical simulations on several examples (in high-dimension) for which the target posterior distribution is known. Simulation results support our theoretical results, i.e. the empirical distribution of samples from `MCGdiff` converges to the target posterior distribution. This is **not** the case for the competing methods (using the same denoising diffusion generative priors) which are shown, when run with random initialization of the denoising diffusion, to generate a significant number of samples outside the support of the target posterior.

- We perform experimental evaluations on inpainting problems on CelebA-HQ, showing that `MCGdiff` generates both diverse and realistic reconstructions, which are coherent with the observations.

Background and notations.

This section provides a concise overview of the diffusion model framework and notations used in this paper. We cover only the elements that are important for understanding our approach, and we recommend that readers refer to the original papers for complete details and derivations Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021c,a). A denoising diffusion model is a generative model that consists of a forward and a backward process. The forward noising process involves sampling a data point $X_0 \sim \mathsf{q}_{\text{data}}$ from the data distribution, which is then converted to a sequence $X_{1:n}$ of recursively corrupted versions of X_0. On the other hand, the backward denoising process involves sampling X_n according to an easy-to-sample reference distribution on $\mathbb{R}^{\mathsf{d}_x}$ and generating $X_0 \in \mathbb{R}^{\mathsf{d}_x}$ by a sequence of denoising steps. Following Sohl-Dickstein et al. (2015); Song et al. (2021a), the forward noising process can be chosen as a

Markov chain with joint distribution

$$\mathsf{q}_{0:n}(x_{0:n}) = \mathsf{q}_{\text{data}}(x_0) \prod_{t=1}^{n} q_t(x_t|x_{t-1}), \quad q_t(x_t|x_{t-1}) = \mathcal{N}(x_t; (1-\beta_t)^{1/2} x_{t-1}, \beta_t \, \mathrm{I}_{\mathsf{d}_x}) \,, \tag{6.1.1}$$

where $\mathrm{I}_{\mathsf{d}_x}$ is the identity matrix of size d_x, $\{\beta_t, t \in \mathbb{N}\} \subset (0,1)$ is a non-increasing sequence and $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ is the p.d.f. of the Gaussian distribution with mean μ and covariance matrix Σ (assumed to be non-singular) evaluated at $\mathbf{x}$. For all $t > 0$, set $\bar{\alpha}_t = \prod_{\ell=1}^{t}(1-\beta_\ell)$ with the convention $\alpha_0 = 1$. We have for all $0 \le s < t \le n$,

$$q_{t|s}(x_t|x_s) := \int \prod_{\ell=s+1}^{t} q_\ell(x_\ell|x_{\ell-1}) \mathrm{d}x_{s+1:t-1} = \mathcal{N}(x_t; (\bar{\alpha}_t/\bar{\alpha}_s)^{1/2} x_s, (1-\bar{\alpha}_t/\bar{\alpha}_s) \, \mathrm{I}_{\mathsf{d}_x}) \,. \tag{6.1.2}$$

For the standard choices of $\bar{\alpha}_t$, the sequence of distributions $(\mathsf{q}_t)_t$ converges weakly to the standard normal distribution as $t \to \infty$, which we chose as the reference distribution. For the reverse process, Song et al. (2021a,b) introduce an *inference distribution* $q^{\sigma}_{1:n|0}(x_{1:n}|x_0)$, depending on a sequence $\{\sigma_t, t \in \mathbb{N}\}$ of hyperparameters satisfying $\sigma_t^2 \in [0, 1-\bar{\alpha}_{t-1}]$ for all $t \in \mathbb{N}^*$, and defined as

$$q^{\sigma}_{1:n|0}(x_{1:n}|x_0) = q^{\sigma}_{n|0}(x_n|x_0) \prod_{t=n}^{2} q^{\sigma}_{t-1|t,0}(x_{t-1}|x_t, x_0) \,,$$

where $q^{\sigma}_{n|0}(x_n|x_0) = \mathcal{N}\left(x_n; \bar{\alpha}_n^{1/2} x_0, (1-\bar{\alpha}_n) \, \mathrm{I}\right)$ and

$$q^{\sigma}_{t-1|t,0}(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \boldsymbol{\mu}_t(x_0, x_t), \sigma_t^2 \, \mathrm{I}_d\right) \,, \tag{6.1.3}$$

where

$$\boldsymbol{\mu}_t(x_0, x_t) = \bar{\alpha}_{t-1}^{1/2} x_0 + (1 - \bar{\alpha}_{t-1} - \sigma_t^2)^{1/2} (x_t - \bar{\alpha}_t^{1/2} x_0)/(1-\bar{\alpha}_t)^{1/2} \,. \tag{6.1.4}$$

For $t = n-1, \ldots, 1$, we define by backward induction the sequence:

$$q^{\sigma}_{t|0}(x_t|x_0) = \int q^{\sigma}_{t|t+1,0}(x_t|x_{t+1}, x_0) q^{\sigma}_{t+1|0}(x_{t+1}|x_0) \mathrm{d}x_{t+1}. \tag{6.1.5}$$

It is shown in (Song et al., 2021a, Lemma 1) that for all $t \in [1:n]$, the distributions of the forward and inference process conditioned on the initial state coincide, i.e. that

$$q^{\sigma}_{t|0}(x_t|x_0) = q_{t|0}(x_t|x_0). \tag{6.1.6}$$

The backward denoising process is derived from the inference distribution by replacing, for each $t \in [2:n]$, x_0 in the definition $q^{\sigma}_{t-1|t,0}(x_{t-1}|x_t, x_0)$ with a prediction

$$\boldsymbol{\chi}^{\theta}_{0|t}(x_t) := \bar{\alpha}_t^{-1/2} \left(x_t - (1-\bar{\alpha}_t)^{1/2} \mathbf{e}^{\theta}(x_t, t)\right) \,, \tag{6.1.7}$$

where $\mathbf{e}^{\theta}(x,t)$ is typically a neural network parameterized by θ. More formally, the backward distribution is defined as

$$\mathsf{p}^{\theta}_{0:n}(x_{0:n}) = \mathsf{p}_n(x_n) \prod_{t=0}^{n+1} p^{\theta}_t(x_t|x_{t+1}) \,,$$

where $\mathsf{p}_n(x_n) = \mathcal{N}(x_n; 0_{\mathsf{d}_x}, \mathrm{I}_{\mathsf{d}_x} \ \mathrm{I}_{\mathsf{d}_x})$ and for all $t \in [1:n-1]$,

$$\begin{aligned} p^{\theta}_t(x_t|x_{t+1}) &:= q^{\sigma}_{t|t+1,0}(x_t|x_{t+1}, \boldsymbol{\chi}^{\theta}_{0|t+1}(x_{t+1})) \\ &= \mathcal{N}(x_t, \mathsf{m}^{\theta}_{t+1}(x_{t+1}), \sigma^2_{t+1} \mathrm{I}_{\mathsf{d}_x}) \,, \end{aligned} \tag{6.1.8}$$

where $\mathsf{m}_{t+1}(x_{t+1}) := \boldsymbol{\mu}(\boldsymbol{\chi}^{\theta}_{0|t+1}(x_{t+1}), x_{t+1})$ and 0_{d_x} is the null vector of size d_x. At step 0, we set $p_0(x_0|x_1) := \mathcal{N}(x_0; \boldsymbol{\chi}^{\theta}_{0|1}(x_1), \sigma_1^2 \mathrm{I}_{\mathsf{d}_x})$. The parameter θ is obtained (see (Song et al., 2021a, Theorem 1)) by solving the following optimization problem:

$$\theta_* \in \operatorname*{argmin}_{\theta} \sum_{t=1}^{n} \frac{1}{2\mathsf{d}_x \sigma_t^2 \alpha_t} \int \|\epsilon - \mathbf{e}^{\theta}(\sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, t)\|_2^2 \mathcal{N}(\epsilon; 0_{\mathsf{d}_x}, \mathrm{I}_{\mathsf{d}_x}) \mathsf{q}_{\mathrm{data}}(\mathrm{d}x_0)\mathrm{d}\epsilon\,. \quad (6.1.9)$$

Thus, $\mathbf{e}^{\theta_*}(X_t, t)$ might be seen as the predictor of the noise added to X_0 to obtain X_t (in the forward pass) and justifies the "prediction" terminology for (6.1.7). The time 0 marginal $\mathsf{p}_0^{\theta_*}(x_0) = \int \mathsf{p}_{0:n}^{\theta_*}(x_{0:n})\mathrm{d}x_{1:n}$ which we will refer to as the *prior* is used as an approximation of $\mathsf{q}_{\mathrm{data}}$ and the time s marginal is $\mathsf{p}_s^{\theta_*}(x_s) = \int \mathsf{p}_{0:n}^{\theta_*}(x_{0:n})\mathrm{d}x_{1:s-1}\mathrm{d}x_{s+1:n}$. In the rest of the paper we drop the dependence on the parameter θ_*. We define for all $v \in \mathbb{R}^{\ell}, w \in \mathbb{R}^{k}$, the concatenation operator $v^{\frown}w = [v^T, w^T]^T \in \mathbb{R}^{\ell+k}$. For $i \in [1:\ell]$, we let $v[i]$ the i-th coordinate of the vector v.

Related works.

The subject of Bayesian problems is very vast, and it is impossible to discuss here all the results obtained in this very rich literature. We will focus on image restoration problems, (deblurring, denoising inpainting) a challenging problem in computer vision that involves restoring a partially observed degraded image. Deep learning techniques are widely used for this task Arjomand Bigdeli et al. (2017); Yeh et al. (2018); Xiang et al. (2023); Wei et al. (2022) with many of them relying on auto-encoders, VAEs Ivanov et al. (2018); Peng et al. (2021); Zheng et al. (2019), GANs Yeh et al. (2018); Zeng et al. (2022), or autoregressive transformers Yu et al. (2018); Wan et al. (2021).

In what follows, we focus on methods based on denoising diffusion that has recently emerged as a way to produce high-quality realistic images on par with the best GANs in terms of image generation, without the intricacies of adversarial training; see Sohl-Dickstein et al. (2015); Song et al. (2021c, 2022). Diffusion-based approaches do not require specific training for degradation types, making them much more versatile and computationally efficient. In Song et al. (2022), noisy linear inverse problems are proposed to be solved by diffusing the degraded observation forward, leading to intermediate observations $\{y_s\}_{s=0}^n$, and then running a modified backward process that promotes consistency with y_s at each step s. The Denoising-Diffusion-Restoration model (`DDRM`) Kawar et al. (2022) also modifies the backward process so that the unobserved part of the state follows the backward process while the observed part is obtained as a noisy weighted sum between the noisy observation and the prediction of the state. As observed by Lugmayr et al. (2022), `DDRM` is very efficient, but the simple blending used occasionally causes inconsistency in the restoration process. The recently introduced `DPS` Chung et al. (2023) considers a backward process targeting the posterior. `DPS` approximates the score of the posterior using the Tweedie formula, which incorporates the learned score of the prior. The approximation error is quantified and shown to decrease when the noise level is large, i.e., when the posterior is close to the prior distribution. As shown in Section 6.3 with a very simple example, neither `DDRM` nor `DPS` can be used to sample the target posterior and therefore do not solve the Bayesian recovery problem (even if we run `DDRM` and `DPS` several time with independent initializations). Indeed, we show that `DDRM` and `DPS` produce samples under the "prior" distribution (which is generally captured very well by the denoising diffusion model), but which are not consistent with the observations (many samples land in areas with very low likelihood).

In Trippe et al. (2023) the authors introduce `SMCdiff`, a Sequential Monte Carlo-based denoising diffusion model that aims at solving specifically the *inpainting problem*. `SMCdiff` produces

a particle approximation of the conditional distribution of the non observed part of the state conditionally on a forward-diffused trajectory of the observation. The resulting particle approximation is shown to converge to the true posterior of the GM under the assumption that the joint laws of the forward and backward processes coincide, which fails to be true in realistic setting. Other than being restricted to the noiseless inpainting problem, their assumption cannot be guaranteed to hold in realistic scenarios. In comparison with `SMCdiff`, `MCGdiff` is a versatile approach that solves any Bayesian linear inverse problem while being consistent under practically no assumption.

6.2 The `MCGdiff` algorithm

In this section we present our methodology for the inpainting problem (6.2.1), both with noise and without noise. The more general case is treated in Section 6.2.3. Let $\mathsf{d}_y \in [1 : \mathsf{d}_x - 1]$. In what follows we may denote the d_y top coordinates of a vector $x \in \mathbb{R}^{\mathsf{d}_x}$ by $\overline{x}$ and the remaining coordinates by $\underline{x}$, so that $x = \overline{x}^\frown \underline{x}$. The inpainting problem is defined as

$$Y = \overline{X} + \sigma_y \varepsilon \,, \quad \varepsilon \sim \mathcal{N}(0, \mathrm{I}_{\mathsf{d}_y}) \,, \quad \sigma \geq 0 \,, \tag{6.2.1}$$

where $\overline{X}$ are the first d_y coordinates of a random variable $X \sim \mathsf{p}_0$. The goal is then to recover the law of the complete state X given a realization $\mathbf{y}$ of the incomplete observation $\mathbf{Y}$ and the model (6.2.1).

6.2.1 Noiseless case

We begin by considering the case in which $\sigma_y = 0$. Since the first d_y coordinates are observed exactly, we aim at infering the remaining coordinates of the random variable, which correspond to $\underline{X}$. As such, given an observation y, we aim at sampling from the posterior given by $\phi_0^y(\underline{x}_0) \propto \mathsf{p}_0(y^\frown \underline{x}_0)$ and which has integral form

$$\phi_0^y(\underline{x}_0) \propto \int \mathsf{p}_n(x_n) \left\{ \prod_{s=1}^{n-1} p_s(x_s | x_{s+1}) \right\} p_0(y^\frown \underline{x}_0 | x_1) \mathrm{d}x_{1:n} \,. \tag{6.2.2}$$

To solve this problem, we propose to use Sequential Monte Carlo (SMC) algorithms Doucet et al. (2001); Cappe et al. (2005); Chopin et al. (2020), where a set of N random samples, referred to as particles, is iteratively updated to approximate the posterior distribution. The updates involve, at iteration s, selecting promising particles from $\xi_{s+1}^{1:N} = (\xi_{s+1}^1, \dots, \xi_{s+1}^N)$ based on a weight function $\widetilde{\omega}_s$, to which we then apply a Markov transition p_s^y to obtain the samples $\xi_s^{1:N}$. The transition $p_s^y(x_s | x_{s+1})$ is designed to follow the backward process while guiding the d_y top coordinates of the pool of particles $\xi_s^{1:N}$ towards the measurement y. Before we proceed to define the transition kernels, note that under the backward dynamics (6.1.8), $\overline{X}_t$ and $\underline{X}_t$ are independent conditionally on X_{t+1} with transition kernels respectively

$$\overline{p}_t(\overline{x}_t | x_{t+1}) := \mathcal{N}(\overline{x}_t; \overline{\mathsf{m}}_{t+1}(x_{t+1}), \sigma_{t+1}^2 \mathrm{I}_{\mathsf{d}_y}) \,, \quad \underline{p}_t(\underline{x}_t | x_{t+1}) := \mathcal{N}(\underline{x}_t; \underline{\mathsf{m}}_{t+1}(x_{t+1}), \sigma_{t+1}^2 \mathrm{I}_{\mathsf{d}_x - \mathsf{d}_y})$$

where $\overline{\mathsf{m}}_{t+1}(x_{t+1}) \in \mathbb{R}^{\mathsf{d}_y}$ and $\underline{\mathsf{m}}_{t+1}(x_{t+1}) \in \mathbb{R}^{\mathsf{d}_x - \mathsf{d}_y}$ are such that $\mathsf{m}_{t+1}(x_{t+1}) = \overline{\mathsf{m}}_{t+1}(x_{t+1})^\frown \underline{\mathsf{m}}_{t+1}(x_{t+1})$ and the above kernels satisfy $p_t(x_s | x_{s+1}) = \overline{p}_t(\overline{x}_t | x_{t+1}) \underline{p}_t(\underline{x}_t | x_{t+1})$.

We consider the following proposal kernels for the steps $t \in [1 : n]$,

$$p_s^y(x_t | x_{t+1}) \propto p_t(x_t | x_{t+1}) \overline{q}_{t|0}(\overline{x}_t | y) \,, \quad \text{where} \quad \overline{q}_{t|0}(\overline{x}_t | y) := \mathcal{N}(\overline{x}_t; \bar{\alpha}_t^{1/2} y, (1 - \bar{\alpha}_t) \mathrm{I}_{\mathsf{d}_y}) \,, \tag{6.2.3}$$

and $p_n^y(x_n) \propto \mathsf{p}_n(x_n)\overline{q}_{n|0}(\overline{x}_n|y)$. For the final step, we use the kernel $\underline{p}_0^y(\underline{x}_0|x_1) = \underline{p}_0(\underline{x}_0|x_1)$. Using standard Gaussian conjugation formulas, we find that

$$p_t^y(x_t|x_{t+1}) = \underline{p}_t(\underline{x}_t|x_{t+1}) \cdot \mathcal{N}\left(\overline{x}_t; \mathsf{K}_t\alpha_t^{1/2}y + (1-\mathsf{K}_t)\overline{\mathsf{m}}_{t+1}(x_{t+1}), (1-\bar{\alpha}_t)\mathsf{K}_t \cdot \mathrm{I}_{\mathsf{d}_y}\right),$$
$$\mathsf{p}_n(x_n) = \mathcal{N}(\underline{x}_n; 0_{\mathsf{d}_x-\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_x-\mathsf{d}_y}) \cdot \mathcal{N}(\overline{x}_n; \mathsf{K}_n\bar{\alpha}_n^{1/2}y, (1-\bar{\alpha}_n)\mathsf{K}_n \cdot \mathrm{I}_{\mathsf{d}_y})$$

where $\mathsf{K}_t := \sigma_{t+1}^2/(\sigma_{t+1}^2 + 1 - \alpha_t)$. For this procedure to target the posterior ϕ_0^y, the weight function $\widetilde{\omega}_s$ is chosen as follows; we set

$$\begin{aligned}\widetilde{\omega}_{n-1}(x_n) &:= \int p_{n-1}(x_{n-1}|x_n)\overline{q}_{n-1|0}(\overline{x}_{n-1}|y)\mathrm{d}x_{n-1} \\ &= \mathcal{N}\left(\alpha_{n-1}^{1/2}y; \overline{\mathsf{m}}_n(x_n), \sigma_n^2 + 1 - \alpha_s\right)\end{aligned}$$

and for $t \in [1:n-2]$,

$$\begin{aligned}\widetilde{\omega}_t(x_{t+1}) &:= \int \overline{p}_t(\overline{x}_t|x_{t+1})\overline{q}_{t|0}(\overline{x}_t|y)\mathrm{d}x_t/\overline{q}_{t+1|0}(\overline{x}_{t+1}|y) \\ &= \frac{\mathcal{N}\left(\alpha_t^{1/2}y; \overline{\mathsf{m}}_{t+1}(x_{s+1}), (\sigma_{t+1}^2 + 1 - \alpha_t)\mathrm{I}_{\mathsf{d}_y}\right)}{\mathcal{N}\left(\alpha_{t+1}^{1/2}y; \overline{x}_{t+1}, (1-\alpha_{t+1})\mathrm{I}_{\mathsf{d}_y}\right)}.\end{aligned} \tag{6.2.4}$$

For the final step, we set $\widetilde{\omega}_0(x_1) := \overline{p}_0(y|\overline{x}_1)/\overline{q}_{1|0}(\overline{x}_1|y)$. The overall SMC algorithm targeting ϕ_0^y using the instrumental kernel (6.2.3) and weight function (6.2.4) is summarized in Algorithm 4.

Algorithm 4: `MCGdiff` ($\sigma = 0$)

Input: Number of particles N
Output: $\xi_0^{1:N}$
`// Operations involving index` i `are repeated for each` $i \in [1:N]$

$\overline{z}_n^i \sim \mathcal{N}(\mathbf{0}_{\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_y})$, $\underline{z}_n^i \sim \mathcal{N}(\mathbf{0}_{\mathsf{d}_x-\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_x-\mathsf{d}_y})$;
$\overline{\xi}_n^i = \mathsf{K}_n\bar{\alpha}_n^{1/2}y + (1-\bar{\alpha}_n)\mathsf{K}_n\overline{z}_n^i$;
Set $\xi_n^i = \overline{\xi}_n^i \frown \underline{z}_n^i$;
for $s \leftarrow n-1 : 0$ **do**
  if $s = n-1$ **then**
    $\widetilde{\omega}_{n-1}(\xi_n^i) = \mathcal{N}(\bar{\alpha}_n^{1/2}y; \overline{\mathsf{m}}_n(\xi_n^i), 2 - \bar{\alpha}_n)$;
  else
    $\widetilde{\omega}_s(\xi_{s+1}^i) = \mathcal{N}\left(\bar{\alpha}_s^{1/2}y; \overline{\mathsf{m}}_{s+1}(\xi_{s+1}^i), \sigma_{s+1}^2 + 1 - \bar{\alpha}_s\right) / \mathcal{N}\left(\bar{\alpha}_{s+1}^{1/2}y; \overline{\xi}_{s+1}^i, 1 - \bar{\alpha}_{s+1}\right)$;
  $A_{s+1}^i \sim \mathrm{Categorical}(\{\widetilde{\omega}_s(\xi_{s+1}^j)/\sum_{k=1}^N \widetilde{\omega}_s(\xi_{s+1}^k)\}_{j=1}^N)$;
  $\overline{z}_s^i \sim \mathcal{N}(\mathbf{0}_{\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_y})$, $\underline{z}_s^i \sim \mathcal{N}(\mathbf{0}_{\mathsf{d}_x-\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_x-\mathsf{d}_y})$;
  $\overline{\xi}_s^i = \mathsf{K}_s\bar{\alpha}_s^{1/2}y + (1-\mathsf{K}_s)\overline{\mathsf{m}}_{s+1}(\xi_{s+1}^i) + (1-\alpha_s)^{1/2}\mathsf{K}_s^{1/2}\overline{z}_s^i$;
  $\underline{\xi}_s^i = \underline{\mathsf{m}}_{s+1}(\xi_{s+1}^i) + \sigma_{s+1}\underline{z}_s^i$;
  Set $\xi_s^i = \overline{\xi}_s^i \frown \underline{\xi}_s^i$;

We now provide a justification to Algorithm 4. Let $\{g_s^y\}_{s=1}^n$ be a sequence of positive functions. Consider the sequence of distributions $\{\phi_s^y\}_{s=1}^n$ defined as follows; $\phi_n^y(x_n) \propto \mathsf{p}_n(x_n)g_n^y(x_n)$ and for $t \in [1:n-1]$

$$\phi_t^y(x_t) \propto \int \frac{g_t^y(x_t)}{g_{t+1}^y(x_{t+1})} p_t(x_t|x_{t+1})\phi_{t+1}^y(\mathrm{d}x_{t+1}), \tag{6.2.5}$$

By construction, the time t marginal (6.2.5) is $\phi_t^y(x_t) \propto \mathsf{p}_t(x_t) g_t^y(x_t)$ for all $t \in [1:n]$. Then, using ϕ_1^y and (6.2.2), we have that

$$\phi_0^y(\underline{x}_0) \propto \int \frac{\overline{p}_0(y|\overline{x}_1)}{g_1^y(x_1)} \underline{p}_0(\underline{x}_0|x_1) \phi_1^y(\mathrm{d}x_1)\,. \tag{6.2.6}$$

The recursion (6.2.5) suggests a way of obtaining a particle approximation of ϕ_0^y; by sequentially approximating each ϕ_t^y we can effectively derive a particle approximation of the posterior using (6.2.6). To construct the intermediate particle approximations we use the framework of *auxiliary particle filters* (APF) Pitt and Shephard (1999). We focus on the particular case where $g_t^y(x_t) = \overline{q}_{t|0}(\overline{x}_t|y)$ which corresponds to Algorithm 4. The initial particle approximation ϕ_n^y is obtained by drawing N i.i.d. samples $(\xi_n^1, \ldots, \xi_n^N)$ from p_n^y and setting $\phi_n^N = N^{-1} \sum_{i=1}^N \delta_{\xi_n^i}$ where δ_ξ is the Dirac mass at ξ . Assume that the empirical approximation of ϕ_{t+1}^y is

$$\phi_{t+1}^N = N^{-1} \sum_{i=1}^N \delta_{\xi_{t+1}^i}\,,$$

where $(\xi_{t+1}^1, \ldots, \xi_{t+1}^N)$ are N random variables. Substituting ϕ_{t+1}^N into the recursion (6.2.5) and introducing the instrumental kernel (6.2.3), we obtain the following mixture

$$\widehat{\phi}_t^N(x_t) = \sum_{i=1}^N \frac{\widetilde{\omega}_t(\xi_{t+1}^i)}{\sum_{j=1}^N \widetilde{\omega}_t(\xi_{t+1}^j)} p_t^y(x_t|\xi_{t+1}^i)\,. \tag{6.2.7}$$

Then, a particle approximation of (6.2.7) is obtained by sampling N conditionally i.i.d. ancestor indices

$$A_{t+1}^{1:N} i.i.d. \text{Categorical}(\{\widetilde{\omega}_t(\xi_{t+1}^i)/\sum_{j=1}^N \widetilde{\omega}_t(\xi_{t+1}^j)\}_{i=1}^N)$$

and then propagating each ancestor particle $\xi_{s+1}^{I_{t+1}^i}$ according to the instrumental kernel (6.2.3). The final particle approximation is given by

$$\phi_0^N = N^{-1} \sum_{i=1}^N \delta_{\underline{\xi}_0^i}\,, \quad \text{where} \quad \underline{\xi}_0^i \sim \underline{p}_0(\cdot|\xi_1^{A_0^i})\,, \quad A_0^i \sim \text{Categorical}(\{\widetilde{\omega}_0(\xi_1^k)/\sum_{j=1}^N \widetilde{\omega}_0(\xi_1^j)\}_{k=1}^N)\,.$$

The potential $g_t^y(x_t) = \overline{q}_{t|0}(\overline{x}_t|y)$, and hence Equations (6.2.5) and (6.2.6), is motivated by considering the posterior of the state under the forward process (6.1.1) $\rho_t^y(x_t) := \int \phi_0^y(\underline{x}_0) q_{t|0}(x_t|y^\frown \underline{x}_0)\, \mathrm{d}\underline{x}_0$. Indeed, first note that the bridge kernel decomposes across the dimensions as follows,

$$q_{t-1|t,0}^\sigma(x_{t-1}|x_t, x_0) = \overline{q}_{t-1|t,0}^\sigma(\overline{x}_{t-1}|\overline{x}_t, \overline{x}_0) \underline{q}_{t-1|t,0}^\sigma(\underline{x}_{t-1}|\underline{x}_t, \underline{x}_0)$$

where

$$\begin{aligned}
\overline{q}_{t|t+1,0}^\sigma(\overline{x}_t|\overline{x}_{t+1}, \overline{x}_0) &= \mathcal{N}(\overline{x}_t; \boldsymbol{\mu}_t(\overline{x}_0, \overline{x}_{t+1}), \sigma_{t+1}^2 \mathrm{I}_{\mathsf{d}_y})\,, \\
\underline{q}_{t-1|t,0}^\sigma(\underline{x}_{t-1}|\underline{x}_t, \underline{x}_0) &= \mathcal{N}(\underline{x}_{t-1}; \boldsymbol{\mu}_t(\underline{x}_0, \underline{x}_t), \sigma_t^2 \mathrm{I}_{\mathsf{d}_x - \mathsf{d}_y})\,,
\end{aligned}$$

and $\boldsymbol{\mu}_{t+1}$ is defined in (6.1.4). It is then easily seen that

$$\begin{aligned}
\overline{q}_{t|0}(\overline{x}_t|\overline{x}_0) &= \int \overline{q}_{t|t+1,0}^\sigma(\overline{x}_t|\overline{x}_{t+1}, \overline{x}_0) \overline{q}_{t+1|0}(\mathrm{d}\overline{x}_{t+1}|\overline{x}_0) \\
\underline{q}_{t|0}(\underline{x}_t|\underline{x}_0) &= \int \underline{q}_{t|t+1,0}^\sigma(\underline{x}_t|\underline{x}_{t+1}, \underline{x}_0) \underline{q}_{t+1|0}(\mathrm{d}\underline{x}_{t+1}|\underline{x}_0)\,,
\end{aligned}$$

where $\underline{q}_{t|0}$ is defined analoguously to $\overline{q}_{t|0}$ in (6.2.3). Hence, using that $q_{t|0}(x_t|x_0) = \overline{q}_{t|0}(\overline{x}_t|\overline{x}_0)\underline{q}_{t|0}(\underline{x}_t|\underline{x}_0)$, we see that

$$\begin{aligned}\rho_t^y(x_t) &= \int \phi_0^y(\mathrm{d}\underline{x}_0)\overline{q}_{t|0}(\overline{x}_t|y)\underline{q}^{\sigma}_{t|t+1,0}(\underline{x}_t|\underline{x}_{t+1},\underline{x}_0)\overline{q}_{t+1|0}(\mathrm{d}\underline{x}_{t+1}|\underline{x}_0)\\ &= \int \phi_0^y(\mathrm{d}\underline{x}_0)\overline{q}_{t|0}(\overline{x}_t|y)\underline{q}^{\sigma}_{t|t+1,0}(\underline{x}_t|\underline{x}_{t+1},\underline{x}_0)q_{t+1|0}(\mathrm{d}x_{t+1}|y^\frown x_0)\,.\end{aligned}$$

Finally, replacing $\underline{x}_0$ by the vector made of the last $\mathsf{d}_x - \mathsf{d}_y$ coordinates of its prediction $\boldsymbol{\chi}_{0|s+1}(x_{s+1})$, which we denote by $\underline{\boldsymbol{\chi}}_{0|s+1}(x_{s+1})$, we see that ρ_s^y satisfies

$$\rho_s^y(x_s) \approx \int \overline{q}_{s|0}(\overline{x}_s|y)\underline{p}_s(\underline{x}_s|x_{s+1})\rho_{s+1}^y(\mathrm{d}x_{s+1})\,. \tag{6.2.8}$$

According to this recursion, the transition from $\underline{x}_s$ should follow the backward process conditioned by x_{s+1}, while that from $\overline{x}_s$ should follow the forward one conditioned on the measurement y at time 0. Equation (6.2.5) reflects this behaviour, while (6.2.6) ensures that the procedure ultimately produces the posterior as the final marginal. The idea of using the forward diffused observation to guide the observed part of the state, as we do here through $\overline{q}_t(\overline{x}_t|y)$, has been exploited in prior works but in a different way. For instance, in Song et al. (2021c, 2022) the observed part of the state is directly replaced by the forward noisy observation and, as it has been noted Trippe et al. (2023), this introduces an irreducible error resulting in a procedure that fails to sample from the posterior. Instead, `MCGdiff` weights the backward process by the density of the forward one conditioned on y, resulting in a natural and consistent algorithm.

We now establish the convergence of `MCGdiff` with a general sequence of *potentials* $\{g_s^y\}_{s=1}^n$. We consider the following assumption on the sequence of potentials $\{g_t^y\}_{t=1}^n$.

(**A11**) (i) $\sup_{x_1\in\mathbb{R}^{\mathsf{d}_x}} \overline{p}_0(y|x_1)/g_1^y(x_1) < \infty$,

(ii) $\sup_{x_{t+1}\in\mathbb{R}^{\mathsf{d}_x}} \int g_t^y(x_t)p_t(x_t|x_{t+1})\mathrm{d}x_t/g_{t+1}^y(x_{t+1}) < \infty$ for all $t \in [1:n-1]$.

The following exponential deviation inequality is standard and is a direct application of (Douc et al., 2014, Theorem 10.17). In particular, it implies a $\mathcal{O}(1/\sqrt{N})$ bound on the mean squared error $\|\phi_0^N(h) - \phi_0^y(h)\|_2$.

Proposition 6.2.1. *Assume (**A11**). There exist constants $c_{1,n}, c_{2,n} \in (0,\infty)$ such that, for all $N \in \mathbb{N}$, $\varepsilon > 0$ and bounded function $h : \mathbb{R}^{\mathsf{d}_x} \mapsto \mathbb{R}$,*

$$\mathbb{P}\left[|\phi_0^N(h) - \phi_0^{\mathbf{y}}(h)| \geq \varepsilon\right] \leq c_{1,n}\exp(-c_{2,n}N\varepsilon^2/|h|_\infty^2)$$

where $|h|_\infty := \sup_{x\in\mathbb{R}^{\mathsf{d}_x}} |h(x)|$.

We also furnish our estimator with an explicit non-asymptotic bound on its bias. Define $\Phi_0^N = \mathbb{E}[\phi_0^N]$ where $\phi_0^N = N^{-1}\sum_{i=1}^N \delta_{\underline{\xi}_0^i}$ is the particle approximation produced by Algorithm 4 and the expectation is with respect to the law of $(\xi_0^{1:N}, \ldots, \xi_n^{1:N}, A_0^{1:N}, \ldots, A_{n-1}^{1:N})$. Define for all $t \in [1:n]$,

$$\phi_t^\star(x_t) \propto \mathsf{p}_t(x_t)\int \delta_y(\mathrm{d}\overline{x}_0)p_{0|t}(x_0|x_t)\mathrm{d}\underline{x}_0\,,$$

where $p_{0|t}(x_0|x_t) := \int \left\{\prod_{s=0}^{t-1} p_s(x_s|x_{s+1})\right\}\mathrm{d}x_{1:t-1}$.

Proposition 6.2.2. *It holds that*

$$\mathsf{KL}(\phi_0^y \parallel \Phi_0^N) \leq \frac{\mathsf{C}_{0:n}^y}{N-1} + \frac{\mathsf{D}_{0:n}^y}{N^2}\,, \tag{6.2.9}$$

where $\mathsf{D}^y_{0:n} > 0$,

$$\mathsf{C}^y_{0:n} := \sum_{t=1}^{n} \int \frac{\mathcal{Z}_t/\mathcal{Z}_0}{g^y_t(z_t)} \left\{ \int \delta_y(\mathrm{d}\overline{x}_0) p_{0|t}(x_0|z_t) \mathrm{d}\underline{x}_0 \right\} \phi^\star_t(\mathrm{d}z_t)\,, \tag{6.2.10}$$

and $\mathcal{Z}_t := \int g^y_t(x_t)\mathsf{p}_t(\mathrm{d}x_t)$ *for all* $t \in [1:n]$ *and* $\mathcal{Z}_0 := \int \delta_y(\mathrm{d}\overline{x}_0)\mathsf{p}_0(x_0)\mathrm{d}\underline{x}_0$. *If furthermore (**A11**) holds then both* $\mathsf{C}^y_{0:n}$ *and* $\mathsf{D}^y_{0:n}$ *are finite.*

The proof of Proposition 6.2.2 is postponed to Section D.2.1. (**A11**) is an assumption on the equivalent of the weights $\{\widetilde{\omega}_t\}_{t=0}^n$ with a general sequence of potentials $\{g^y_t\}_{t=1}^n$ and is not restrictive as it can be satisfied by setting for example $g^y_s(x_s) = \overline{q}_{s|0}(\overline{x}_s|y)+\delta$ where $\delta > 0$. The resulting algorithm is then only a slight modification of the one described above, see Section D.2.1 for more details. It is also worth noting that Proposition 6.2.2 combined with Pinsker's inequality implies that the bias of `MCGdiff` goes to 0 with the number of particle samples N for fixed n. We have chosen to present a bound in Kullback–Leibler (KL) divergence, inspired by Andrieu et al. (2018b); Huggins and Roy (2019), as it allows an explicit dependence on the modeling choice $\{g^y_s\}_{s=1}^n$, see Lemma D.2.2. Finally, unlike the theoretical guarantees established for `SMCdiff` in Trippe et al. (2023), proving the asymptotic exactness of our methodology w.r.t. to the generative model posterior does not require having $\mathsf{p}_{s+1}(x_{s+1})p_s(x_s|x_{s+1}) = \mathsf{p}_s(x_s)q_{s+1}(x_{s+1}|x_s)$ for all $s \in [0:n-1]$, i.e. that the one step forward kernel is the time reversal of the backward one, which does not in hold in practice. As such, `SMCdiff` exhibits a non-vanishing asymptotic bias.

6.2.2 Noisy case

We now turn to the case $\sigma_y > 0$. The posterior density we consider in this section is given by

$$\phi^y_0(x_0) \propto g^y_0(\overline{x}_0)\mathsf{p}_0(x_0), \quad \text{where} \quad g^y_0 : x_0 \mapsto \mathcal{N}(y; \overline{x}_0, \sigma_y^2 \mathrm{I}_{\mathsf{d}_y})\,. \tag{6.2.11}$$

In what follows we assume that there exists a timestep $\tau \in [1:n]$ such that $\sigma^2 = (1-\bar{\alpha}_\tau)/\bar{\alpha}_\tau$. We discuss this assumption in the numerical section. In what follows we denote $\tilde{y}_\tau = \bar{\alpha}_\tau^{1/2} y$. We can then write that

$$\begin{aligned} g^y_0(\overline{x}_0) &= \bar{\alpha}_\tau^{1/2} \cdot \mathcal{N}(\tilde{y}_\tau; \bar{\alpha}_\tau^{1/2} x_0, (1-\bar{\alpha}_\tau)\cdot \mathrm{I}_{\mathsf{d}_y}) \\ &= \bar{\alpha}_\tau^{1/2} \cdot \overline{q}_{\tau|0}(\tilde{y}_\tau|\overline{x}_0)\,, \end{aligned} \tag{6.2.12}$$

which hints that the likelihood function g^y_0 is closely related to the forward process (6.1.1). We may then write the posterior (6.2.11) as follows

$$\begin{aligned} \phi^y_0(x_0) &\propto \int \delta_{\tilde{y}_\tau}(\mathrm{d}\overline{x}_s)\overline{q}_{\tau|0}(\overline{x}_\tau|\overline{x}_0)\mathsf{p}_0(x_0) \\ &\propto \int \delta_{\tilde{y}_\tau}(\mathrm{d}\overline{x}_\tau) q_{\tau|0}(x_\tau|x_0)\mathsf{p}_0(x_0)\mathrm{d}\underline{x}_\tau\,, \end{aligned}$$

Next, let us assume that the forward process (6.1.1) is the reverse of the backward one (6.1.8), i.e. that

$$\mathsf{p}_t(x_t)q_{t+1}(x_{t+1}|x_t) = \mathsf{p}_{t+1}(x_{t+1})p_t(x_t|x_{t+1})\,, \quad \forall t \in [0:n-1]\,. \tag{6.2.13}$$

This is similar to the assumption made in `SMCdiff` Trippe et al. (2023). Then, it is easily seen that it implies $\mathsf{p}_0(x_0)q_{\tau|0}(x_\tau|x_0) = \mathsf{p}_\tau(x_\tau)p_{0|\tau}(x_0|x_\tau)$ and thus

$$\begin{aligned} \phi^y_0(x_0) &= \frac{\int p_{0|\tau}(x_0|x_\tau)\delta_{\tilde{y}_\tau}(\mathrm{d}\overline{x}_\tau)\mathsf{p}_\tau(x_\tau)\mathrm{d}\underline{x}_\tau}{\int \delta_{\tilde{y}_\tau}(\mathrm{d}\overline{z}_\tau)\mathsf{p}_\tau(z_\tau)\mathrm{d}\underline{z}_\tau} \\ &= \int p_{0|\tau}(x_0|\tilde{y}_\tau \frown \underline{x}_\tau)\phi^{\tilde{y}_\tau}_\tau(\mathrm{d}\underline{x}_\tau)\,, \end{aligned} \tag{6.2.14}$$

where $\phi_\tau^{\tilde{y}_\tau}(\underline{x}_\tau) \propto \mathsf{p}_\tau(\tilde{y}_\tau \frown \underline{x}_\tau)$. (6.2.14) highlights that solving the inverse problem (6.2.1) with $\sigma_y > 0$ is equivalent to solving an inverse problem on the intermediate state $X_\tau \sim \mathsf{p}_\tau$ with *noiseless* observation $\tilde{y}_\tau$ of the d_y top coordinates and then propagating the resulting posterior back to time 0 with the backward kernel $p_{0|\tau}$. To obtain a particle approximation of $\phi_\tau^{\tilde{y}_\tau}$ we may then use the methodology of the previous section with the potentials $g_t^y : x_t \mapsto \overline{q}_{t|\tau}(x_t|\tilde{y}_\tau)$ for $t \in [\tau{+}1 : n]$. Indeed, consider the sequence $\{\rho_t^{\tilde{y}_\tau}\}_{t=\tau}^n$ defined by $\rho_t^{\tilde{y}_\tau}(x_t) = \int \phi_\tau^{\tilde{y}_\tau}(\mathrm{d}\underline{x}_\tau) q_{t|\tau}(x_t|\tilde{y}_\tau \frown \underline{x}_\tau)$. Using that $\mathsf{p}_\tau(x_\tau) = \int \mathsf{p}_0(\mathrm{d}x_0)\ q_{\tau|0}(x_\tau|x_0)$ by assumption and the identity (6.1.5), we find that

$$\begin{aligned}\rho_t^{\tilde{y}_\tau}(x_t) &\propto \int \mathsf{p}_0(\mathrm{d}x_0)\overline{q}_{\tau|0}(\tilde{y}_\tau|\overline{x}_0)\overline{q}_{t|\tau}(\overline{x}_t|\tilde{y}_\tau)\underline{q}_{t|0}(\underline{x}_t|\underline{x}_0) \\ &\propto \int \mathsf{p}_0(\mathrm{d}x_0)\overline{q}_{\tau|0}(\tilde{y}_\tau|\overline{x}_0)\overline{q}_{t|\tau}(\overline{x}_t|\tilde{y}_\tau)\overline{q}_{t+1|\tau}(\mathrm{d}\overline{x}_{t+1}|\tilde{y}_\tau)\underline{q}^\sigma_{t|t+1,0}(\underline{x}_t|\underline{x}_{t+1},\underline{x}_0)\underline{q}_{t+1|0}(\mathrm{d}\underline{x}_{t+1}|\underline{x}_0)\,,\end{aligned}$$

and replacing $\underline{x}_0$ in the bridge kernel $\underline{q}^\sigma_{t|t+1,0}$ by $\underline{\boldsymbol{\chi}}_{0|s+1}(x_{s+1})$, we obtain the recursion

$$\rho_t^{\tilde{y}_\tau}(x_t) \approx \int \overline{q}_{t|\tau}(\overline{x}_t|\tilde{y}_\tau)\underline{p}_t(\underline{x}_t|x_{t+1})\rho_{t+1}^{\tilde{y}_\tau}(\mathrm{d}x_{t+1})\,, \tag{6.2.15}$$

which provides the intuition for our choice of potential.

The assumption (6.2.13) regarding the reversal of the backward process holds only approximately in realistic settings. Therefore, while (6.2.14) also holds only approximately in practice, we can still use it as inspiration for designing potentials when the assumption is not valid. Consider then $\{g_t^y\}_{t=\tau}^n$ and sequence of probability measures $\{\phi_t^y\}_{t=\tau}^n$ defined for all $t \in [\tau : n]$ as

$$\phi_t^y(x_t) \propto g_t^y(x_t)\mathsf{p}_t(x_t)\,, \quad \text{where} \quad g_t^y : x_t \mapsto \mathcal{N}\left(x_t; \tilde{y}_\tau, (1-(1-\kappa)\bar{\alpha}_t/\bar{\alpha}_\tau)\mathrm{I}_{\mathsf{d}_y}\right)\,, \quad \kappa \geq 0\,. \tag{6.2.16}$$

In the particular case of $\kappa = 0$, we have that $g_t^y(x_t) = \overline{q}_{t|\tau}(\overline{x}_t|\tilde{y}_\tau)$ for $t \in [\tau{+}1 : n]$ and $\phi_\tau^y = \phi_\tau^{\tilde{y}_\tau}$. The recursion (6.2.5) holds for $t \in [\tau : n]$ and assuming that $\kappa > 0$, we find that

$$\phi_0^y(x_0) \propto \int \frac{g_0^y(x_0)}{g_\tau^y(x_\tau)} p_{0|\tau}(x_0|x_\tau)\phi_\tau^y(\mathrm{d}x_\tau)\,,$$

which resembles the recursion (6.2.14). In practice we take κ to be small in order to mimick the Dirac delta mass at $\overline{x}_\tau$ in (6.2.14). Having obtained a particle approximation $\phi_\tau^N = N^{-1}\sum_{i=1}^N \delta_{\xi_\tau^i}$ of ϕ_τ^y by adapting Algorithm 4, we estimate ϕ_0^y with

$$\phi_0^N = \sum_{i=1}^N \omega_0^i \delta_{\xi_0^i}\,, \quad \text{where} \quad \xi_0^i \sim p_{0|\tau}(\cdot|\xi_\tau^i)\,, \quad \omega_0^i = \frac{g_0^y(\xi_0^i)/g_\tau^y(\xi_\tau^i)}{\sum_{j=1}^N g_0^y(\xi_0^j)/g_\tau^y(\xi_\tau^j)}\,.$$

In the next section we extend this methodology to general linear Gaussian observation models. Finally, note that (6.2.14) allows us to extend `SMCdiff` to handle noisy inverse problems in a principled manner. This is detailed in Section D.1.

6.2.3 Extension to general linear inverse problems

We now extend `MCGdiff` to a general linear Gaussian observation model. Consider $Y = \mathrm{A}X + \sigma_y\varepsilon$ where $\mathrm{A} \in \mathbb{R}^{\mathsf{d}_y \times \mathsf{d}_x}$, $\varepsilon \sim \mathcal{N}(0_{\mathsf{d}_y}, \mathrm{I}_{\mathsf{d}_y})$ and $\sigma_y \geq 0$ and the singular value decomposition (SVD) $\mathrm{A} = \mathrm{U}\mathrm{S}\overline{\mathrm{V}}^T$, where $\overline{\mathrm{V}} \in \mathbb{R}^{\mathsf{d}_x \times \mathsf{d}_\mathbf{y}}$, $\mathrm{U} \in \mathbb{R}^{\mathsf{d}_y \times \mathsf{d}_\mathbf{y}}$ are two orthonormal matrices, and $\mathrm{S} \in \mathbb{R}^{\mathsf{d}_\mathbf{y} \times \mathsf{d}_\mathbf{y}}$ is diagonal. For simplicity, it is assumed that the singular values are all distinct $s_1 > s_2 > \cdots > s_{\mathsf{d}_\mathbf{y}} > 0$. Set $\mathsf{b} = \mathsf{d}_x - \mathsf{d}_\mathbf{y}$. Let $\underline{\mathrm{V}} \in \mathbb{R}^{\mathsf{d}_x \times \mathsf{b}}$ be an orthonormal matrix of which the columns

complete those of $\overline{\mathrm{V}}$ into an orthonormal basis of $\mathbb{R}^{\mathsf{d}_x}$, i.e. $\underline{\mathrm{V}}^T\underline{\mathrm{V}} = \mathrm{I}_{\mathsf{b}}$ and $\underline{\mathrm{V}}^T\overline{\mathrm{V}} = \mathbf{0}_{\mathsf{b},\mathsf{d}_{\mathbf{y}}}$. We define $\mathrm{V} = [\overline{\mathrm{V}}, \underline{\mathrm{V}}] \in \mathbb{R}^{\mathsf{d}_x \times \mathsf{d}_x}$. In what follows, for a given $\mathbf{x} \in \mathbb{R}^{\mathsf{d}_x}$ we write $\overline{\mathbf{x}} \in \mathbb{R}^{\mathsf{d}_{\mathbf{y}}}$ for its top $\mathsf{d}_{\mathbf{y}}$ coordinates and $\underline{\mathbf{x}} \in \mathbb{R}^{\mathsf{b}}$ for the remaining coordinates. Multiplying the measurement equation by $\mathrm{S}^{-1}\mathrm{U}^T$ yields

$$\mathbf{Y} = \overline{\mathbf{X}} + \sigma_y \mathrm{S}^{-1}\widetilde{\varepsilon}\,, \quad \widetilde{\varepsilon} \sim \mathcal{N}(0, \mathrm{I}_{\mathsf{d}_{\mathbf{y}}})\,,$$

where $\mathbf{X} := \mathrm{V}^T X$ and $\mathbf{Y} := \mathrm{S}^{-1}\mathrm{U}^T Y$. In this section, we focus on solving this linear inverse problem in the orthonormal basis defined by V using the methodology developed in the previous sections. This prompts us to define the diffusion based generative model in this basis. As V is an orthonormal matrix, the law of $\mathbf{X}_0 = \mathrm{V}^T X_0$ is $\mathfrak{p}_0(\mathbf{x}_0) := \mathfrak{p}_0(\mathrm{V}\mathbf{x}_0)$. By definition of $\mathfrak{p}_0$ and the fact that $\|\mathrm{V}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{R}^{\mathsf{d}_x}$ we have that

$$\begin{aligned}\mathfrak{p}_0(\mathbf{x}_0) &= \int p_0(\mathrm{V}\mathbf{x}_0|x_1) \left\{\prod_{s=1}^{n-1} p_s(\mathrm{d}x_s|x_{s+1})\right\} \mathfrak{p}_n(\mathrm{d}x_n) \\ &= \int \lambda_0(\mathbf{x}_0|\mathbf{x}_1) \left\{\prod_{s=1}^{n-1} \lambda_s(\mathrm{d}\mathbf{x}_s|\mathbf{x}_{s+1})\right\} \mathfrak{p}_n(\mathrm{d}\mathbf{x}_n)\end{aligned}$$

where for all $s \in [0 : n-1]$,

$$\lambda_s(\mathbf{x}_s|\mathbf{x}_{s+1}) := \mathcal{N}(\mathbf{x}_s; \mathfrak{m}_{s+1}(\mathbf{x}_{s+1}), \sigma^2_{t+1}\mathrm{I}_{\mathsf{d}_x})\,, \quad \text{where} \quad \mathfrak{m}_{s+1}(\mathbf{x}_{s+1}) := \mathrm{V}^T\mathfrak{m}_{s+1}(\mathrm{V}\mathbf{x}_{s+1})\,.$$

The transition kernels $\{\lambda_s\}_{s=0}^n$ thus define a diffusion based model in the basis V. In the following we shall write $\overline{\mathfrak{m}}_{s+1}(\mathbf{x}_{s+1})$ for the first $\mathsf{d}_{\mathbf{y}}$ coordinates of $\mathfrak{m}_{s+1}(\mathbf{x}_{s+1})$ and $\underline{\mathfrak{m}}_{s+1}(\mathbf{x}_{s+1})$ the last b coordinates. We denote by $\mathfrak{p}_s$ the time s marginal of the backward process,

Noiseless. In this case the target posterior is $\phi_0^{\mathbf{y}}(\mathbf{x}_0) \propto \mathfrak{p}_0(\mathbf{y}^\frown\underline{\mathbf{x}}_0)$. The extension of the algorithm described in section 6.2.1 is thus straighforward; it is enough to replace y with $\mathbf{y}$ ($=\mathrm{S}^{-1}\mathrm{U}^T y$) and the backward kernels $\{p_t\}_{t=0}^{n-1}$ with $\{\lambda_t\}_{t=0}^{n-1}$.

Noisy. The posterior density is then

$$\phi_0^{\mathbf{y}}(\mathbf{x}_0) \propto g_0^{\mathbf{y}}(\overline{\mathbf{x}}_0)\mathfrak{p}_0(\mathbf{x}_0), \quad \text{where} \quad g_0^{\mathbf{y}}(\overline{\mathbf{x}}_0) = \prod_{i=1}^{\mathsf{d}_{\mathbf{y}}} \mathcal{N}(\mathbf{y}[i]; \overline{\mathbf{x}}_0[i], (\sigma_y/s_i)^2)\,.$$

We now generalize Section 6.2.2. Assume that there exists $\{\tau_i\}_{i=1}^{\mathsf{d}_{\mathbf{y}}} \subset [1 : n]$ such that $\bar{\alpha}_{\tau_i}\sigma_y^2 = (1-\bar{\alpha}_{\tau_i})s_i^2$. Define for all $i \in [1 : \mathsf{d}_{\mathbf{y}}]$, $\tilde{\mathbf{y}}_i := \bar{\alpha}_{\tau_i}^{1/2}\mathbf{y}[i]$. Then we can write the potential $g_0^{\mathbf{y}}$ as the product of forward processes from time 0 to each time step τ_i, i.e.

$$g_0^{\mathbf{y}}(\mathbf{x}_0) = \prod_{i=1}^{\mathsf{d}_{\mathbf{y}}} \bar{\alpha}_{\tau_i}^{1/2}\mathcal{N}(\tilde{\mathbf{y}}_i; \bar{\alpha}_{\tau_i}^{1/2}\mathbf{x}_0[i], (1-\bar{\alpha}_{\tau_i}))\,.$$

Writing the potential this way allows us to generalize (6.2.14) as follows. Denote for $\ell \in [1 : \mathsf{d}_x]$, $\mathbf{x}^{\setminus\ell} \in \mathbb{R}^{\mathsf{d}_x - 1}$ the vector $\mathbf{x}$ with its ℓ-th coordinate removed. Define

$$\phi^{\tilde{\mathbf{y}}}_{\tau_1:n}(\mathrm{d}\mathbf{x}_{\tau_1:n}) \propto \left\{\prod_{i=1}^{\mathsf{d}_{\mathbf{y}}-1} \lambda_{\tau_i|\tau_{i+1}}(\mathbf{x}_{\tau_i}|\mathbf{x}_{\tau_{i+1}})\delta_{\tilde{\mathbf{y}}_i}(\mathrm{d}\mathbf{x}_{\tau_i}[i])\mathrm{d}\mathbf{x}_{\tau_i}^{\setminus i}\right\} \mathfrak{p}_{\tau_{\mathsf{d}_{\mathbf{y}}}}(\mathbf{x}_{\tau_{\mathsf{d}_{\mathbf{y}}}})\delta_{\tilde{\mathbf{y}}_{\mathsf{d}_{\mathbf{y}}}}(\mathrm{d}\mathbf{x}_{\tau_{\mathsf{d}_{\mathbf{y}}}}[\mathsf{d}_{\mathbf{y}}])\mathrm{d}\mathbf{x}_{\tau_{\mathsf{d}_{\mathbf{y}}}}^{\setminus \mathsf{d}_{\mathbf{y}}}\,,$$

which corresponds to the posterior of a noiseless inverse problem on the joint states $\mathbf{X}_{\tau_1:n} \sim \mathfrak{p}_{\tau_1:n}$ with noiseless observations $\tilde{\mathbf{y}}_{\tau_i}$ of $\mathbf{X}_{\tau_i}[i]$ for all $i \in [1 : \mathsf{d}_{\mathbf{y}}]$.

Proposition 6.2.3. *Assume that* $\mathfrak{p}_{s+1}(\mathbf{x}_{s+1})\lambda_s(\mathbf{x}_s|\mathbf{x}_{s+1}) = \mathfrak{p}_s(\mathbf{x}_s)q_{s+1}(\mathbf{x}_{s+1}|\mathbf{x}_s)$ *for all* $s \in [0 : n-1]$. *Then it holds that*

$$\phi_0^{\mathbf{y}}(\mathbf{x}_0) \propto \int \lambda_{0|\tau_1}(\mathbf{x}_0|\mathbf{x}_{\tau_1})\phi^{\tilde{\mathbf{y}}}_{\tau_1:n}(\mathrm{d}\mathbf{x}_{\tau_1:n})\,.$$

The proof of Proposition 6.2.3 can be found in Section D.2.2. We have thus shown that sampling from the posterior $\phi_0^{\mathbf{y}}$ is equivalent to sampling from the posterior $\phi^{\tilde{\mathbf{y}}}_{\tau_1:n}$ then propagating the final state $\mathbf{X}_{\tau_1}$ up to time 0 according to the backward kernel $\lambda_{0|\tau_1}$. Furthermore, we can extend the approximate recursion (6.2.15). For all $t \in [\tau_1 : n]$ define $\rho_t^{\tilde{\mathbf{y}}}(\mathrm{d}\mathbf{x}_t) := \int \phi^{\tilde{\mathbf{y}}}_{\tau_1:n}(\mathrm{d}\mathbf{x}_{\tau_1:n})$, and for all $s \in [0 : n-1]$, $z_s \in \mathbb{R}$ and $\ell \in [1 : \mathsf{d}_{\mathbf{y}}]$, let $\lambda_s^\ell(z_s|\mathbf{x}_{s+1}) := \mathcal{N}(z_s; \mathfrak{m}_{s+1}(\mathbf{x}_{s+1})[\ell], \sigma^2_{t+1})$. By adapting the derivations of the previous section and using Lemma D.2.4, we find that the following approximate recursion is satisfied; for all $k \in [1 : \mathsf{d}_{\mathbf{y}}]$ and $t \in [\tau_k + 1, \tau_{k+1} - 1]$,

$$\rho_t^{\tilde{\mathbf{y}}}(\mathbf{x}_t) \approx \int \underline{\lambda}_t(\underline{\mathbf{x}}_t|\mathbf{x}_{t+1}) \prod_{\ell=\tau(t)+1}^{\mathsf{d}_{\mathbf{y}}} \lambda_t^\ell(\mathbf{x}_t[\ell]|\mathbf{x}_{t+1}) \prod_{j=1}^{\tau(t)} q^j_{t|\tau_j}(\mathbf{x}_t[j]|\tilde{\mathbf{y}}_j)\rho_{t+1}^{\tilde{\mathbf{y}}}(\mathrm{d}\mathbf{x}_{t+1})\,, \tag{6.2.17}$$

where for all $t \in [\tau_1 : n]$, $\tau(t) := \max\{k \in [1 : \mathsf{d}_{\mathbf{y}}] : \tau_k \le t\}$ and for all $j \in [1 : \mathsf{d}_{\mathbf{y}}]$, and $t, s \in [1 : n]$ such that $t > s$, $q^j_{t|s}$ denotes the density of the j-th coordinate of the forward process from s to t. If $t = \tau_k$, then

$$\rho_t^{\tilde{\mathbf{y}}}(\mathrm{d}\mathbf{x}_t) \approx \int \underline{\lambda}_t(\mathrm{d}\underline{\mathbf{x}}_t|\mathbf{x}_{t+1})\delta_{\tilde{\mathbf{y}}_k}(\mathrm{d}\mathbf{x}_t[k]) \prod_{\ell=\tau(t)+1}^{\mathsf{d}_{\mathbf{y}}} \lambda_t^\ell(\mathrm{d}\mathbf{x}_t[\ell]|\mathbf{x}_{t+1}) \prod_{j=1}^{\tau(t)-1} q^j_{t|\tau_j}(\mathrm{d}\mathbf{x}_t[j]|\tilde{\mathbf{y}}_j)\rho_{t+1}^{\tilde{\mathbf{y}}}(\mathrm{d}\mathbf{x}_{t+1})\,.$$

We target the posterior $\phi_0^{\mathbf{y}}$ by mimicking this recursion. Consider then $\{g_t^y\}_{t=\tau}^n$ and sequence of probability measures $\{\phi_t^{\mathbf{y}}\}_{t=\tau}^n$ defined for all $t \in [\tau_1 : n]$ by $\phi_t^{\mathbf{y}}(\mathbf{x}_t) \propto g_t^{\mathbf{y}}(\mathbf{x}_t)\mathsf{p}_t(\mathbf{x}_t)$ and

$$g_t^{\mathbf{y}} : \mathbf{x}_t \mapsto \prod_{i=1}^{\tau(t)} \mathcal{N}\left(\mathbf{x}_t; \tilde{\mathbf{y}}_i, 1 - (1-\kappa)\bar{\alpha}_t/\bar{\alpha}_{\tau_i}\right)\,, \quad \kappa > 0\,. \tag{6.2.18}$$

We obtain a particle approximation of $\phi_{\tau_1}^{\mathbf{y}}$ using a particle filter with proposal kernel and weight function

$$\lambda_t^{\mathbf{y}}(\mathbf{x}_t|\mathbf{x}_{t+1}) \propto g_t^{\mathbf{y}}(\mathbf{x}_t)p_t(\mathbf{x}_t|\mathbf{x}_{t+1})\,, \quad \widetilde{\omega}_t(\mathbf{x}_{t+1}) = \frac{\int g_t^{\mathbf{y}}(\mathbf{x}_t)p_t(\mathrm{d}\mathbf{x}_t|\mathbf{x}_{t+1})}{g_{t+1}^{\mathbf{y}}(\mathbf{x}_{t+1})}\,,$$

which are both available in closed form. Indeed, using standard Gaussian conjugation formulas, we find that

$$\lambda_t^{\mathbf{y}}(\mathbf{x}_t|\mathbf{x}_{t+1}) = \underline{\lambda}_t(\underline{\mathbf{x}}_t|\mathbf{x}_{t+1}) \prod_{k=\tau(t)+1}^{\mathsf{d}_{\mathbf{y}}} \lambda_t^k(\mathbf{x}_t[k]|\mathbf{x}_{t+1}) \prod_{\ell=1}^{\tau(t)} \lambda_t^{\mathbf{y},\ell}(\mathbf{x}_t[\ell]|\mathbf{x}_{t+1})\,, \tag{6.2.19}$$

where, by letting $\sigma^2_{t|\tau_\ell} := 1 - (1-\kappa)\bar{\alpha}_t/\bar{\alpha}_{\tau_\ell}$ and $\mathsf{K}_{t|\tau_\ell} = \sigma^2_{t+1}/(\sigma^2_{t+1} + \sigma^2_{t|\tau_\ell})$,

$$\lambda_t^{\mathbf{y},\ell}(\mathbf{x}_t[\ell]|\mathbf{x}_{t+1}) = \mathcal{N}(\mathbf{x}_t[\ell]; \mathsf{K}_{t|\tau_\ell}\tilde{\mathbf{y}}_\ell + (1 - \mathsf{K}_{t|\tau_\ell})\mathfrak{m}_{t+1}(\mathbf{x}_{t+1})[\ell], \mathsf{K}_{t|\tau_\ell}\sigma^2_{t|\tau_\ell})\,, \tag{6.2.20}$$

and

$$\widetilde{\omega}_t(\mathbf{x}_{t+1}) = \prod_{\ell=1}^{\tau(t)} \frac{\mathcal{N}(\tilde{\mathbf{y}}_\ell; \mathfrak{m}_{t+1}(\mathbf{x}_{t+1})[\ell], \sigma^2_{t+1} + \sigma^2_{t|\tau_\ell})}{\mathcal{N}(\mathbf{x}_{t+1}[\ell]; \tilde{\mathbf{y}}_\ell, \sigma^2_{t|\tau_\ell})}\,. \tag{6.2.21}$$

Thus, applying Algorithm 4 with the transition kernels $\{\lambda_t\}_{t=\tau_1}^{n-1}$ and weight function $\{\widetilde{\omega}_t\}_{t=\tau_1}^{n-1}$ yields the particle approximation $\phi^N_{\tau_1} = N^{-1}\sum_{i=1}^N \delta_{\xi^i_{\tau_1}}$ and that of $\phi_0^{\mathbf{y}}$ is given by

$$\phi_0^N = \sum_{i=1}^N \omega_0^i \delta_{\xi_0^i}\,, \quad \text{where} \quad \xi_0^i \sim \lambda_{0|\tau}(\cdot|\xi^i_{\tau_1})\,, \quad \omega_0^i = \frac{g_0^y(\xi_0^i)/g^y_{\tau_1}(\xi^i_{\tau_1})}{\sum_{j=1}^N g_0^y(\xi_0^j)/g^y_{\tau_1}(\xi^j_{\tau_1})}\,.$$

6.3 Numerical experiments

Gaussian Mixture Model.

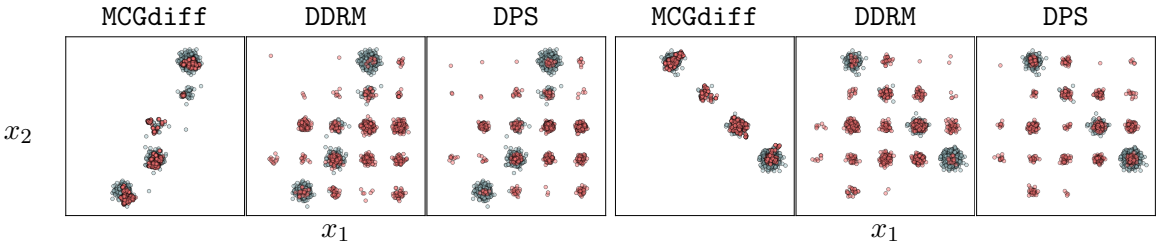

Figure 6.1: We display the first two dimensions of the GMM inverse problem for one of the measurement models tested. The blue dots represent samples from the exact posterior, while the red dots correspond to samples generated by each of the algorithms used (the names of the algorithms are given at the top of each column). The first three columns correspond to $(\mathsf{d}_x, \mathsf{d}_\mathbf{y}) = (80, 1)$ *and the last three to* $(\mathsf{d}_x, \mathsf{d}_\mathbf{y}) = (800, 1)$*. 20 steps of* `DDIM` *were used.*

d	d_y	MCGdiff	DDRM	DPS
8	1	**2.6 ± 0.8**	6.9 ± 3.4	6.7 ± 1.1
8	2	**1.0 ± 0.4**	4.6 ± 0.9	5.4 ± 1.3
8	4	**0.5 ± 0.2**	1.8 ± 0.8	4.3 ± 1.1
80	1	**2.2 ± 0.6**	7.7 ± 1.1	6.1 ± 1.2
80	2	**0.9 ± 0.4**	9.9 ± 1.3	7.4 ± 1.4
80	4	**0.4 ± 0.1**	7.8 ± 1.1	4.4 ± 1.1
800	1	**2.3 ± 0.8**	7.7 ± 0.8	6.5 ± 0.8
800	2	**1.8 ± 0.8**	8.6 ± 0.8	6.5 ± 1.1
800	4	**0.7 ± 0.5**	9.5 ± 0.9	5.5 ± 0.9

d	d_y	MCGdiff	DDRM	DPS
8	1	**1.8 ± 0.6**	4.5 ± 1.5	3.3 ± 0.8
8	2	**1.1 ± 0.5**	4.5 ± 1.3	3.2 ± 1.1
8	4	**0.2 ± 0.0**	1.2 ± 0.6	0.9 ± 0.5
80	1	**1.3 ± 0.3**	6.4 ± 1.2	3.1 ± 1.4
80	2	**1.1 ± 0.7**	9.2 ± 1.1	2.8 ± 1.2
80	4	**0.3 ± 0.0**	7.1 ± 1.0	1.4 ± 0.6
800	1	**2.6 ± 0.9**	7.2 ± 1.4	2.7 ± 1.0
800	2	**1.3 ± 0.8**	8.6 ± 1.0	1.8 ± 0.9
800	4	**0.3 ± 0.0**	8.3 ± 1.0	0.4 ± 0.2

Table 6.1: Sliced Wasserstein for the GMM case. Table on the left correspond to 20 *steps of* `DDIM` *with* $\eta = 0.6$ *and on the right to* 100 *steps and* $\eta = 0.85$*.*

We present an example where the data distribution q_{data} is a mixture of 25 Gaussian distributions. The means and variances of the components of the mixture are given in Section D.2.3. In this case, for each choice of forward operator A and measurement noise standard deviation $\sigma_y > 0$, the target posterior distribution $\phi_0^{\mathbf{y}}$ can be computed explicitly, see Section D.2.3. Moreover, it is possible to explicitly minimize each term occurring in the denoising problem (6.1.9), so that the choice of the weighting scheme $\{\sigma_t,\, t \in \mathbb{N}\}$ is irrelevant. To investigate the performance of posterior sampling methods, for each pair of dimensions $(\mathsf{d}_x, \mathsf{d}_\mathbf{y}) \in \{8, 80, 800\} \times \{1, 2, 4\}$ we randomly generate multiple measurement models $(y, \mathrm{A}, \sigma_y) \in \mathbb{R}^{\mathsf{d}_\mathbf{y}} \times \mathbb{R}^{\mathsf{d}_\mathbf{y} \times \mathsf{d}_x} \times [0, 1]$, and we also randomly choose the weight associated with each component of the Gaussian mixture. Details (distribution of A, σ_y, etc...) are given in Section D.2.3. This example is particularly interesting because it allows us to study the behaviour of our method on difficult and ill-posed problems in high dimensions while having access to the exact posterior with which to compare – obtaining a "ground truth" for the posterior is not feasible for most problems. By varying the dimension, the forward operator, the noise level, and the mixture weight, we gain insight into the performance of posterior sampling methods under varying conditions and with different levels of posterior multimodality.

To compare the posterior distribution estimated by each algorithm with the exact target posterior distribution, we use the sliced Wasserstein (SW) distance defined in Section D.2.3. We

Figure 6.2: Inpainting with different masks on the CelebA test set.

use 10^4 slices for the SW distance and compare 1000 samples of `MCGdiff`, `DPS`, and `DDRM` with 1000 samples of the true posterior distribution obtained using 20 `DDIM` steps and 100 `DDIM` steps for generation. The variance reduction `DDIM` parameter η is set to 0.6 and 0.85 for the 20 and 100 `DDIM` steps, respectively. When choosing the timesteps of `DDIM` we want to emulate the constraint $\sigma_y^2 \alpha_{\tau_i} = (1 - \alpha_{\tau_i}) s_i^2$. Therefore, we include the timesteps t that minimizes $|\sigma_y^2 \alpha_t - (1 - \alpha_t) s_i^2|$ for each $i \in [1 : \mathtt{d_y}]$. More details are given in Section D.2.3.

Table 6.1 indicates the CLT 95% confidence intervals obtained by considering 20 randomly selected measurement models $(y, \mathrm{A}, \sigma_y)$ for each setting $(\mathtt{d}_x, \mathtt{d_y})$. Figure 6.1 shows the first two dimensions of the estimated posterior distributions corresponding to the configurations $(80, 1)$ and $(800, 1)$ from Table 6.1 for one of the randomly generated measurement model $(y, \mathrm{A}, \sigma_y)$ in the case of 20 `DDIM` steps. Illustration of other settings are given in Section D.2.3. These illustrations give us insight into the behaviour of the algorithms and their ability to accurately estimate the posterior distribution. We see that `MCGdiff` is more precise at estimating the posterior distribution and does not sample outside of the support of the posterior distribution in all scenarios tested. Table 6.1 also shows that the difference in performance is greater in the more extreme settings where the problem is ill-posed ($\mathtt{d_y} = 1, 2$) and the number of `DDIM` steps is limited (20).

Inpainting.

We consider the inpainting problem on the CelebA dataset. The images dimension are $3 \times 256 \times 256$ and we use pretrained denoising network available at https://github.com/bahjat-kawar/ddrm for all the methods. All methods share the same `DDIM` parameters, with 250 sample steps between $[0 : 1000]$ and $\eta = 1$. For `MCGdiff` and `SMCdiff` , a total of $N = 384$ particles is used. For `DPS` we set the learning rate parameter to $\zeta_i = 0.5$ and for `DDRM` we consider the configuration proposed on Kawar et al. (2022). We also use several different masks on images from the CelebA test set in fig. 6.2. The first row corresponds to samples from the most ill-posed problem.

Next, we compare `MCGdiff` to `DPS` and `DDRM` on several different noisy inverse problems over several different image datasets (section 6.3). For each algorithm, we generate 1000 samples and we show the pair of samples that are the furthest appart in L^2 norm from each other in the pool of samples. For `MCGdiff` we ran several parallel particle filters with $N = 64$ to generate

1000 samples.

	CIFAR-10	Flowers	Cats	Bedroom	Church	CelebaHQ
(W, H, C)	$(32, 32, 3)$	$(64, 64, 3)$	$(128, 128, 3)$	$(256, 256, 3)$	$(256, 256, 3)$	$(256, 256, 3)$

Table 6.2: The datasets used for the inverse problems over image datasets.

Super Resolution.

We compare for the super resolution problem. We set $\sigma_y = 0.05$ for all the datasets and $\zeta_{\text{coeff}} = 0.1$ for `DPS` . We use 100 steps of `DDIM` with $\eta = 1$. The results are shown in Figure 6.3. We use a downsampling ratio of 4 for the CIFAR-10 dataset, 8 for both Flowers and Cats datasets and 16 for the others.

Gaussian 2D debluring.

We consider a Gaussian 2D square kernel with sizes $(w/6, h/6)$ and standard deviation $w/30$ where (w, h) are the width and height of the image. We set $\sigma_y = 0.1$ for all the datasets and $\zeta_{\text{coeff}} = 0.1$ for `DPS` . We use 100 steps of `DDIM` with $\eta = 1$. The results are shown in Figure 6.4.

6.4 Conclusion

In this chapter, we have introduced `MCGdiff` a novel method for solving Bayesian linear Gaussian inverse problems with a denoising diffusion based generative model prior. We have shown that `MCGdiff` is theoretically grounded and provided numerical experiments that reflect the adequacy of `MCGdiff` in a Bayesian framework, as opposed to recent works. This difference is of the uttermost importance when the relevance of the generated samples is hard to verify, as in safety critical applications. `MCGdiff` is a first step towards robust approaches for addressing the challenges of Bayesian linear inverse problems with denoising diffusion based generative model priors.

Finally, our work can be of course improved by considering better choices of potentials $\{g_s^y\}_{s=1}^n$ and proposals. While the backward process involving the approximate conditional score (1.3.31) could be a good proposal, it nonetheless involves taking the gradient of the score network with respect to its input which is prohibitive computationally and limits the number of particles N that can be used. We thus believe that deriving more computationally efficient approximations of the conditional score is the main area of improvement.

Figure 6.3: Ratio 4 for CIFAR, 8 for flowers and Cats and 16 for CELEB


CIFAR 10
Flowers
Cats
Bedroom
Church
CelebaHQ
sample
observation
DPS
DPS
DDRM
DDRM
MCGdiff
MCGdiff

Figure 6.4

Appendices

Appendix A

Appendix of Chapter 2

A.1 Proofs

A.1.1 Additional notation

By abuse of notation, we denote by ρ and $\tilde{\pi}$ the probability measures with density w.r.t. the Lebesgue measure ρ and $\tilde{\pi}$ respectively.

A.1.2 Proof of (2.2.3)

The second expression of w_k follows from $\mathbf{J}_{\mathrm{T}^{-j}}(\mathrm{T}^k(x)) = \mathbf{J}_{\mathrm{T}^{k-j}}(x)/\mathbf{J}_{\mathrm{T}^k}(x)$ which implies

$$\begin{aligned} w_k(x) &= \varpi_k \rho(\mathrm{T}^k(x)) / \sum_{j\in\mathbb{Z}} \varpi_j \rho(\mathrm{T}^{k-j}(x)) \mathbf{J}_{\mathrm{T}^{-j}}(\mathrm{T}^k(x)) \,, \\ &= \varpi_k \rho(\mathrm{T}^k(x)) \mathbf{J}_{\mathrm{T}^k}(x) / \sum_{j\in\mathbb{Z}} \varpi_j \rho(\mathrm{T}^{k-j}(x)) \mathbf{J}_{\mathrm{T}^{k-j}}(x) = \varpi_k \rho_{-k}(x) \Big/ \sum_{i\in\mathbb{Z}} \varpi_{k+i} \rho_i(x) \quad . \end{aligned}$$

A.1.3 Proof of Theorem 2.2.1

The unbiasedness of $\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ follows directly from (2.2.2). Moreover, as $\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ is unbiased and $E^{\varpi}_{\mathrm{T}} < \infty$, we can write

$$\mathrm{Var}_{\rho}[\widehat{\mathcal{Z}}^{\varpi}_{X}/\mathcal{Z}] = \mathbb{E}_{\rho}[(\widehat{\mathcal{Z}}^{\varpi}_{X}/\mathcal{Z})^2] - 1 = E^{\varpi}_{\mathrm{T}} - 1 \,. \tag{A.1.1}$$

As $X^{1:N} \overset{\text{iid}}{\sim} \rho$, $\mathrm{Var}_{\rho}[\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}/\mathcal{Z}] = N^{-1}\,\mathrm{Var}_{\rho}[\widehat{\mathcal{Z}}^{\varpi}_{X}/\mathcal{Z}]$. Finally, if $M^{\varpi}_{\mathrm{T}} < \infty$, then Hoeffding's inequality applies and we can write for any $\epsilon > 0$,

$$\mathbb{P}(|\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}/\mathcal{Z} - 1| > \epsilon) \leq 2\exp(-2N\epsilon^2/(M^{\varpi}_{\mathrm{T}})^2) \,. \tag{A.1.2}$$

Writing $\delta = 2\exp(-2N\epsilon^2/(M^{\varpi}_{\mathrm{T}})^2)$, we identify $\log(2/\delta) = 2N\epsilon^2/(M^{\varpi}_{\mathrm{T}})^2$ and $\epsilon = M^{\varpi}_{\mathrm{T}}\sqrt{\log(2/\delta)/(2N)}$. Plugging this expression of ϵ in (A.1.2) concludes the proof.

A.1.4 Proof of Theorem 2.2.2

We first present two auxiliary lemmas necessary to establish Theorem 2.2.2.

Lemma A.1.1. *Let A, B be two integrable random variables satisfying $|A/B| \leq M$ almost*

surely and denote $a = \mathbb{E}[A]$, $b = \mathbb{E}[B]$. *Then,*

$$|\mathbb{E}[A/B] - a/b| \leq \frac{\sqrt{\mathrm{Var}(A/B)\,\mathrm{Var}(B)}}{b}\,, \tag{A.1.3}$$

$$\mathrm{Var}(A/B) \leq \mathbb{E}\left[|A/B - a/b|^2\right] \leq \frac{2}{B^2}\left(\mathbb{E}\left[|A_N - A|^2\right] + M^2\mathbb{E}\left[|B_N - B|^2\right]\right). \tag{A.1.4}$$

Proof. Write first, using the Cauchy-Schwarz inequality,

$$\begin{aligned}\left|\mathbb{E}\left[\frac{A}{B}\right] - \frac{a}{b}\right| &= \left|\mathbb{E}\left[\frac{A}{B}\right] - \frac{\mathbb{E}\left[A\right]}{b}\right| = \left|\mathbb{E}\left[A\left(\frac{1}{B} - \frac{1}{b}\right)\right]\right|,\\ &= \left|\mathbb{E}\left[\frac{A}{B}\left(\frac{b-B}{b}\right)\right]\right| = \left|\mathbb{E}\left[\left(\frac{A}{B} - \mathbb{E}\left[\frac{A}{B}\right]\right)\left(\frac{B-b}{b}\right)\right]\right|,\\ &\leq \frac{\sqrt{\mathrm{Var}(A/B)}\sqrt{\mathrm{Var}(B)}}{b}\,.\end{aligned}$$

Moreover, using $|A/B| \leq M$ yields

$$\begin{aligned}\left|\frac{A}{B} - \frac{a}{b}\right| = \left|\frac{1}{b}(A-a) + A\left(\frac{1}{B} - \frac{1}{b}\right)\right| &\leq \frac{1}{b}|A-a| + \frac{|A|}{Bb}|B-b|\,,\\ &\leq \frac{1}{b}|A-a| + \frac{M}{b}|B-b|\,.\end{aligned}$$

Therefore,

$$|A/B - a/b|^2 \leq \frac{2}{b^2}\left(|A-a|^2 + M^2|B-b|^2\right),$$

Using that $\mathbb{E}\left[|A/B - a/b|^2\right] = \mathrm{Var}(A/B) + |\mathbb{E}[A/B] - a/b|^2$ concludes the proof. □

We get the following lemma from (Douc et al., 2011b, Lemma 4).

Lemma A.1.2. *Assume that* A *and* B *are random variables and that there exist positive constants* b, M, C, K *such that*

(i) $|A/B| \leq M$, $\mathbb{P}$*-a.s.*,

(ii) for all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left(|B-b| > \epsilon\right) \leq K\exp(-R\epsilon^2)$,

(iii) for all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left(|A| > \epsilon\right) \leq K\exp\left(-R\epsilon^2/M^2\right)$,

then,

$$\mathbb{P}(|A/B| \geq \epsilon) \leq 2K\exp(-Rb^2\epsilon^2/4M^2)\,.$$

Proof. By the triangle inequality,

$$\begin{aligned}|A/B| &= \left|\frac{A}{B}(b-B)b^{-1} + b^{-1}A\right|,\\ &\leq b^{-1}\left|A/B\right|\left|b-B\right| + b^{-1}\left|A\right| \leq Mb^{-1}\left|b-B\right| + b^{-1}\left|A\right|\,.\end{aligned}$$

Therefore,

$$\{|A/B| \geq \epsilon\} \subseteq \left\{|B-b| \geq \frac{\epsilon b}{2M}\right\} \cup \left\{|A| \geq \frac{\epsilon b}{2}\right\}.$$

Then, conditions (ii) and (iii) imply that

$$\begin{aligned}\mathbb{P}\left(|A/B| \geq \epsilon\right) &\leq \mathbb{P}\left(|B-b| \geq \frac{\epsilon b}{2M}\right) + \mathbb{P}\left(|A| \geq \frac{\epsilon b}{2}\right),\\ &\leq 2K\exp(-Rb^2\epsilon^2/(4M^2))\,.\end{aligned}$$

□

Proof of Theorem 2.2.2. Let $g : \mathbb{R}^d \to \mathbb{R}$ such that $\sup_{x\in\mathbb{R}^d} |g|(x) \leq 1$ and denote $\pi(g) = \int g \mathrm{d}\pi$. We use Lemma A.1.1 with $A = A_N$ and $B = \widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$ where

$$A_N = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))g(\mathrm{T}^k(X^i))\,,\ \widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\,. \quad \text{(A.1.5)}$$

By construction, since $\sup_{x\in\mathbb{R}^d} |g|(x) \leq 1$, almost surely $A_N/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}} \leq 1$ and $\mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}) = N^{-1}\mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1})$. Then, using (2.2.2) with $a = \mathbb{E}[A_N] = \mathcal{Z}\pi(g)$ and $b = \mathbb{E}[\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}] = \mathcal{Z}$, Lemma A.1.1 implies

$$\left|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)\right| = \left|\mathbb{E}[A_N/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}] - a/b\right| \leq N^{-1/2}\sqrt{\mathrm{Var}(A_N/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}})\mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1})}\,. \quad \text{(A.1.6)}$$

On the other hand,

$$\mathbb{E}\left[|A_N - a|^2\right] = N^{-1}\mathbb{E}_{X\sim\rho}\left[\left\{\textstyle\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))g(\mathrm{T}^k(X)) - \mathcal{Z}\pi(g)\right\}^2\right] \leq N^{-1}\mathcal{Z}^2 E^{\varpi}_{\mathrm{T}}\,.$$

These inequalities yield using $\mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1}) \leq E^{\varpi}_{\mathrm{T}}$ and Lemma A.1.1 again:

$$\mathbb{E}\left[|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)|^2\right] \leq \frac{2}{N}(E^{\varpi}_{\mathrm{T}} + \mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1})) \leq \frac{4}{N}E^{\varpi}_{\mathrm{T}}\,,$$
$$|\mathbb{E}\left[J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)\right]| \leq \frac{\sqrt{2(E^{\varpi}_{\mathrm{T}} + \mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1}))\mathrm{Var}(\widehat{\mathcal{Z}}^{\varpi}_{X^1})}}{N} \leq \frac{2E^{\varpi}_{\mathrm{T}}}{N}\,,$$

which concludes the proof.

Define

$$\tilde{A}_N = N^{-1}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\left(g(\mathrm{T}^k(X^i)) - \pi(g)\right)\,.$$

With this notation, the proof of (2.2.9) relies on the application of Lemma A.1.2 to $A = \tilde{A}_N$ and $B = \widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}$, since

$$J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g) = A_N/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}\,.$$

As $\sup_{x\in\mathbb{R}^d} |g|(x) \leq 1$, we get that $\tilde{A}_N/\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}} \leq 2$. By (2.2.2), $\mathbb{E}[\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}}] = \mathcal{Z}$ and $\widehat{\mathcal{Z}}^{\varpi}_{X^{1:N}} = N^{-1}\sum_{i=1}^{N} W_i$ with $W_i = \sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i)) \leq M^{\varpi}_{\mathrm{T}}$. Then, by Hoeffding's inequality, for all $\varepsilon > 0$,

$$\mathbb{P}(|B_N - \mathcal{Z}| > \varepsilon) \leq 2\exp(-2N(\varepsilon/M^{\varpi}_{\mathrm{T}})^2)\,.$$

Similarly, A_N is centered and $A_N = N^{-1}\sum_{i=1}^{N} U_i$ with

$$U_i = \sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\{g(\mathrm{T}^k(X^i)) - \pi(g)\}$$

and $|U_i| \leq 2M^{\varpi}_{\mathrm{T}}$ almost surely. By Hoeffding's inequality, for all $\varepsilon > 0$,

$$\mathbb{P}(|A_N| > \epsilon) \leq 2\exp(-N\varepsilon^2/(8(M^{\varpi}_{\mathrm{T}})^2))\,.$$

The assumptions of Lemma A.1.2 are met so that

$$\mathbb{P}(|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)| > \varepsilon) \leq 4\exp(-\varepsilon^2 N\mathcal{Z}^2/[32(M^{\varpi}_{\mathrm{T}})^2])\,,$$

which concludes the proof. □

A.1.5 Proof of Lemma 2.2.3

As $w_k(x) = \varpi_k \rho(\mathrm{T}^k(x))/\{\Omega \rho_{\mathrm{T}}(\mathrm{T}^k(x))\}$, by Jensen's inequality,

$$E_{\mathrm{T}}^{\varpi} = \int \left(\sum_{k\in\mathbb{Z}} w_k(x) \mathrm{L}(\mathrm{T}^k(x))/\mathcal{Z} \right)^2 \rho(x)\mathrm{d}x = \int \left(\sum_{k\in\mathbb{Z}} \frac{\varpi_k}{\Omega} \frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x)\mathrm{d}x \,,$$
$$\leq \int \sum_{k\in\mathbb{Z}} \frac{\varpi_k}{\Omega} \left(\frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x)\mathrm{d}x \,,$$
$$\leq \Omega^{-1} \sum_{k\in\mathbb{Z}} \varpi_k \int \left(\frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x)\mathrm{d}x \,.$$

Using the change of variables $y = \mathrm{T}^k(x)$ yields, by (2.2.1),

$$E_{\mathrm{T}}^{\varpi} \leq \Omega^{-1} \sum_{k\in\mathbb{Z}} \varpi_k \int \left(\frac{\pi(y)}{\rho_{\mathrm{T}}(y)} \right)^2 \rho(\mathrm{T}^{-k}(y)) \mathbf{J}_{\mathrm{T}^{-k}}(y)\mathrm{d}y \leq \int \left(\frac{\pi(y)}{\rho_{\mathrm{T}}(y)} \right)^2 \rho_{\mathrm{T}}(y)\mathrm{d}y \,.$$

A.1.6 Proofs of NEO MCMC sampler

Proof of Theorem 2.3.1. Note first that by symmetry, we have

$$P(y, \mathsf{A}) = N^{-1} \int \sum_{i=1}^{N} \delta_y(\mathrm{d}x^i) \prod_{j=1, j\neq i}^{N} \rho(x^j)\mathrm{d}x^j \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}_{x^k}^{\varpi}}{\sum_{j=1}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^k) \,. \tag{A.1.7}$$

We begin with the proof of reversibility of P w.r.t. $\tilde{\pi}$. Let f, g be nonnegative measurable functions. By definition of P,

$$\int \tilde{\pi}(\mathrm{d}y) P(y, \mathrm{d}y') f(y) g(y') = \frac{1}{N\mathcal{Z}} \int \sum_{i=1}^{N} \rho(\mathrm{d}y) \widehat{\mathcal{Z}}_y^{\varpi} f(y) \delta_y(\mathrm{d}x^i) \prod_{l=1, l\neq i}^{N} \rho(\mathrm{d}x^l) \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}_{x^k}^{\varpi}}{\sum_{j=1}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} g(x^k) \,,$$
$$= \frac{1}{N\mathcal{Z}} \int \sum_{i=1}^{N} \widehat{\mathcal{Z}}_{x^i}^{\varpi} f(x^i) \prod_{l=1}^{N} \rho(\mathrm{d}x^l) \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}_{x^k}^{\varpi}}{\sum_{j=1}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} g(x^k) \,,$$
$$= \frac{1}{N\mathcal{Z}} \int \prod_{l=1}^{N} \rho(\mathrm{d}x^l) \frac{\sum_{i=1}^{N} \widehat{\mathcal{Z}}_{x^i}^{\varpi} f(x^i) \sum_{k=1}^{N} \widehat{\mathcal{Z}}_{x^k}^{\varpi} g(x^k)}{\sum_{j=1}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \,,$$
$$= \int \tilde{\pi}(\mathrm{d}y) P(y, \mathrm{d}y') f(y') g(y) \,,$$

which shows that P is $\tilde{\pi}$-reversible. We now establish that P is $\tilde{\pi}$-irreducible. We have for $y \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$P(y, \mathsf{A}) = \int \delta_y(\mathrm{d}x^1) \sum_{i=1}^{N} \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{N \widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^{N} \rho(\mathrm{d}x^j)$$
$$= \int \frac{\widehat{\mathcal{Z}}_y^{\varpi}}{\widehat{\mathcal{Z}}_y^{\varpi} + \sum_{j=2}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x) \prod_{j=2}^{N} \rho(\mathrm{d}x^j) + \int \sum_{i=2}^{N} \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{\widehat{\mathcal{Z}}_y^{\varpi} + \sum_{j=2}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^{N} \rho(\mathrm{d}x^j)$$
$$\geq \sum_{i=2}^{N} \int \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \sum_{j=2, j\neq i}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^{N} \rho(\mathrm{d}x^j)$$
$$\geq \sum_{i=2}^{N} \int \tilde{\pi}(\mathrm{d}x^i) \mathbb{1}_{\mathsf{A}}(x^i) \int \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \sum_{j=2, j\neq i}^{N} \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \prod_{j=2, j\neq i}^{N} \rho(\mathrm{d}x^j) \,.$$

Since the function $f\colon z \mapsto (z+a)^{-1}$ is convex on $\mathbb{R}_+$ for $a>0$, we get for $i \in \{2,\dots,N\}$,

$$\int \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \sum_{j=2,j\neq i}^N \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \prod_{j=2,j\neq i}^N \rho(\mathrm{d}x^j) \geq \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \int \sum_{j=2,j\neq i}^N \widehat{\mathcal{Z}}_{x^j}^{\varpi} \prod_{j=2,j\neq i}^N \rho(\mathrm{d}x^j)}$$
$$\geq \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \mathcal{Z}(N-2)}\,. \quad \text{(A.1.8)}$$

Therefore, for $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ satisfying $\tilde{\pi}(\mathsf{A})>0$, we get $P(y,\mathsf{A})>0$ for any $y\in\mathbb{R}^d$ since $\widehat{\mathcal{Z}}_x^{\varpi} < \infty$ for any $x\in\mathbb{R}^d$. By definition, P is $\tilde{\pi}$-irreducible.

We show that P is Harris recurrent using (Tierney, 1994, Corollary 2). To this end, since P is $\tilde{\pi}$-irreducible, it is sufficient to show that P is a Metropolis type kernel. Define $\alpha(x^1,x^2) = (N-1)\int \prod_{j=3}^N \rho(\mathrm{d}x^j)\widehat{\mathcal{Z}}_{x^2}^{\varpi}/\sum_{j=1}^N \widehat{\mathcal{Z}}_{x^j}^{\varpi}$ for $x^1,x^2\in\mathbb{R}^d$ and $\rho_{2:N}(\mathrm{d}x^{2:N}) = \{\prod_{j=2}^N \rho_{2:N}(x^j)\}\mathrm{d}x^{2:N}$. Then, by (2.3.2), we get with this notation, for $y\in\mathbb{R}^d$, $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}
&P(y,\mathsf{A})\\
&= \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \sum_{i=2}^N \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^i) + \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{\mathcal{Z}}_{x^1}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^1)\\
&= \sum_{i=2}^N \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^i) + \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{\mathcal{Z}}_{x^1}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^1)\\
&= \sum_{i=2}^N \int \delta_y(\mathrm{d}x^1)\rho(\mathrm{d}x^i) \int \prod_{j=2,j\neq i}^N \rho(x^j)\mathrm{d}x^j \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}\mathbb{1}_{\mathsf{A}}(x^i)}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} + \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{\mathcal{Z}}_{x^1}^{\varpi}\mathbb{1}_{\mathsf{A}}(x^1)}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}}\\
&= \sum_{i=2}^N \int \frac{\alpha(y,x^i)}{(N-1)}\mathbb{1}_{\mathsf{A}}(x^i)\rho(\mathrm{d}x^i) + \int \delta_y(\mathrm{d}x^1)\rho_{2:N}(\mathrm{d}x^{2:N}) \left\{1 - \sum_{i=2}^N \frac{\widehat{\mathcal{Z}}_{x^i}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}}\right\} \mathbb{1}_{\mathsf{A}}(x^1)\\
&= \int_{\mathsf{A}} \alpha(y,y')\rho(y')\mathrm{d}y' + \left(1 - \int \alpha(y,y')\rho(y')\mathrm{d}y'\right)\delta_y(\mathsf{A})\,. \quad \text{(A.1.9)}
\end{aligned}$$

With the terminology of (Tierney, 1994, Corollary 2), P is Metropolis type kernel and therefore is Harris recurrent.

Note that Algorithm 2 defines a Markov chain $\{Y_i,U_i\}_{i\in\mathbb{N}}$ taking for U_0 an arbitrary initial point with Markov kernel denoted by $\tilde{P}$. By abuse of notation, we denote by $\{Y_i,U_i\}_{i\in\mathbb{N}}$ the canonical process on the canonical space $(\mathbb{R}^d\times\mathbb{R}^d)^{\mathbb{N}}$ endowed with the corresponding σ-field and denote by $\mathbb{P}_{y,u}$ the distribution associated with the Markov chain with kernel $\tilde{P}$ and initial distribution $\delta_y\otimes\delta_u$. Denote for any $y\in\mathbb{R}^d$ by $\mathbb{P}_y$ the marginal distribution of $\mathbb{P}_{y,u}$ with respect to $\{Y_i\}_{i\in\mathbb{N}}$, *i.e.* $\mathbb{P}_y(\mathsf{A}) = \mathbb{P}_{(y,u)}(\{Y_i\}_{i\in\mathbb{N}}\in\mathsf{A})$ for $u\in\mathbb{R}^d$, noting that by definition, $\mathbb{P}_{(y,u)}(\mathsf{A}\times(\mathbb{R}^d)^{\mathbb{N}})$ does not depend on u. In addition, under $\mathbb{P}_y$, $\{Y_i\}_{i\in\mathbb{N}}$ is a Markov chain associated with P. Therefore, since P is $\tilde{\pi}$-irreducible and Harris recurrent, we get by (Douc et al., 2018a, Theorem 11.3.1) and (Tierney, 1994, Theorem 2, 3) for any $y\in\mathbb{R}^d$, $\lim_{k\to\infty}\|\delta_y P^k - \tilde{\pi}\|_{\mathsf{TV}} = 0$ and for any bounded and measurable function g,

$$n^{-1}\sum_{k=1}^n g(Y_k) = \tilde{\pi}(g)\,, \qquad \mathbb{P}_y\text{-almost surely}\,. \quad \text{(A.1.10)}$$

We now turn to proving the properties regarding Q. For any $\mathsf{B}\in\mathcal{B}(\mathbb{R}^d)$, using (2.2.2), we obtain

$$\int \tilde{\pi}(y)Q(y,\mathsf{B})\mathrm{d}y = \mathcal{Z}^{-1}\int \rho(y)\sum_{k\in\mathbb{Z}} w_k(y)\mathrm{L}(\mathrm{T}^k(y))\mathbb{1}_{\mathsf{B}}(\mathrm{T}^k(y))\mathrm{d}y = \pi(\mathsf{B})\,.$$

Using for all $y \in \mathbb{R}^d$, $\lim_{n\to\infty} \|P^n(y,\cdot) - \tilde{\pi}\|_{\mathsf{TV}} = 0$, we get $\lim_{n\to\infty} \|P^n Q(y,\cdot) - \pi\|_{\mathsf{TV}} = 0$. It remains to show the stated Law of Large Numbers. Let $y, u \in \mathbb{R}^d$ and g be a bounded measurable function. Define for any $i \in \mathbb{N}^*$, $\tilde{U}_i = g(U_i) - Qg(Y_i)$. By definition, for any $i \in \mathbb{N}^*$, $\left|\tilde{U}_i\right| \leq 2\sup_{x\in\mathbb{R}^d} |g(x)|$ and $\mathbb{E}_{(y,u)}[\tilde{U}_i|\mathcal{F}_{i-1}] = 0$, where $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ is the canonical filtration. Therefore, $\{\tilde{U}_i\}_{i\in\mathbb{N}^*}$ are $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$-martingale increments and $\{S_k = \sum_{i=1}^k \tilde{U}_i\}_{k\in\mathbb{N}}$ is a $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$-martingale. Using (Hall and Heyde, 1980, Theorem 2.18), we get

$$\lim_{n\to\infty} \{S_n/n\} = 0\,, \quad \mathbb{P}_{(y,u)}\text{-almost surely}\,. \tag{A.1.11}$$

The proof is completed using that $\lim_{n\to\infty}\{n^{-1}\sum_{i=1}^n Qg(Y_i)\} = \tilde{\pi}(Qg) = \pi(g)$, $\mathbb{P}_y$-almost surely by (A.1.10) and therefore by definition, $\mathbb{P}_{(y,u)}$-almost surely. □

Proof of Theorem 2.3.3. We have for $(x, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$,

$$P(y,\mathsf{A}) \geq \sum_{i=2}^N \int \tilde{\pi}(\mathrm{d}x^i)\mathbb{1}_{\mathsf{A}}(x^i) \int \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \sum_{j=2,j\neq i}^N \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \prod_{j=2,j\neq i}^N \rho(\mathrm{d}x^j)\,.$$

Moreover, as for any $x \in \mathbb{R}^d$, $\widehat{\mathcal{Z}}_x^{\varpi}/\mathcal{Z} \leq M_{\mathrm{T}}^{\varpi}$,

$$\int \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \sum_{j=2,j\neq i}^N \widehat{\mathcal{Z}}_{x^j}^{\varpi}} \prod_{j=2,j\neq i}^N \rho(\mathrm{d}x^j) \geq \frac{\mathcal{Z}}{\widehat{\mathcal{Z}}_y^{\varpi} + \widehat{\mathcal{Z}}_{x^i}^{\varpi} + \mathcal{Z}(N-2)} \geq \frac{1}{2M_{\mathrm{T}}^{\varpi} + N - 2}\,.$$

We finally obtain the inequality

$$P(x,\mathsf{A}) \geq \tilde{\pi}(\mathsf{A}) \times \frac{N-1}{2M_{\mathrm{T}}^{\varpi} + N - 2} = \epsilon_N \tilde{\pi}(\mathsf{A})\,. \tag{A.1.12}$$

The proof for P is concluded from (Douc et al., 2018a, Theorem 18.2.4).

As $\|P^k(y,\cdot) - \tilde{\pi}\|_{\mathsf{TV}} \leq \kappa_N^k$, for any bounded function f, $\|f\|_\infty \leq 1$, we have $|P^k f(y) - \tilde{\pi}(f)| \leq \kappa_N^k$, by definition of the Total Variation Distance. Then, writing $f = Qg$ for any bounded function g, $\|g\|_\infty \leq 1$, we have $\|f\|_\infty \leq 1$ and

$$|P^k f(y) - \tilde{\pi}(f)| = |P^k Qg(y) - \tilde{\pi}Q(g)| = |P^k Qg(y) - \pi(g)| \leq \kappa_N^k\,. \tag{A.1.13}$$

□

Write now P the Markov kernel extending to correlated proposals: for $y \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$P(y,\mathsf{A}) = N^{-1} \int \sum_{i=1}^N \delta_y(\mathrm{d}x^i) r_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^N \frac{\widehat{\mathcal{Z}}_{x^k}^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \mathbb{1}_{\mathsf{A}}(x^k)\,, \tag{A.1.14}$$

where the Markov kernels R_i are defined by $R_i(x^i, \mathrm{d}x^{1:N\setminus\{i\}}) = r_i(x^i, x^{1:N\setminus\{i\}})\mathrm{d}x^{1:N\setminus\{i\}}$ and r_i by (2.3.5).

Theorem A.1.3. *P is $\tilde{\pi}$-invariant.*

Proof. Define the Nd-dimensional probability measure $\bar{\rho}_N(\mathrm{d}x^{1:N}) = \rho(\mathrm{d}x^1)R_1(x^1, \mathrm{d}x^{2:n})$. Let

$\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$. Then, we have

$$\begin{aligned}
\tilde{\pi}P(\mathsf{A}) &= N^{-1} \int \tilde{\pi}(\mathrm{d}y) \int \sum_{i=1}^{N} \delta_y(\mathrm{d}x^i) R_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}^{\varpi}_{x^k}}{N\widehat{\mathcal{Z}}^{\varpi}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (N\mathcal{Z})^{-1} \int \sum_{i=1}^{N} \rho(\mathrm{d}x^i) \widehat{\mathcal{Z}}^{\varpi}_{x^i} R_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}^{\varpi}_{x^k}}{N\widehat{\mathcal{Z}}^{\varpi}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (N\mathcal{Z})^{-1} \int \bar{\rho}_N(\mathrm{d}x^{1:N}) \sum_{i=1}^{N} \widehat{\mathcal{Z}}^{\varpi}_{x^i} \sum_{k=1}^{N} \frac{\widehat{\mathcal{Z}}^{\varpi}_{x^k}}{N\widehat{\mathcal{Z}}^{\varpi}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (N\mathcal{Z})^{-1} \int \sum_{k=1}^{N} \widehat{\mathcal{Z}}^{\varpi}_{x^k} \bar{\rho}_N(\mathrm{d}x^{1:N}) \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (N\mathcal{Z})^{-1} \int \sum_{k=1}^{N} \widehat{\mathcal{Z}}^{\varpi}_{x^k} \rho(\mathrm{d}x^k) \mathbb{1}_{\mathsf{A}}(x^k) = \tilde{\pi}(\mathsf{A}) \,.
\end{aligned}$$

□

A.2 Continuous-time limit of NEO and NEIS

A.2.1 Proof for the continuous-time limit

Consider $\bar{h} > 0$ and a family $\{\mathrm{T}_h : h \in \left(0, \bar{h}\right]\}$ of C^1-diffeomorphisms. For $N \in \mathbb{N}^*$ and a bounded and continuous $f : \mathbb{R}^d \to \mathbb{R}$, write

$$I^{\mathrm{NEO}}_{\varpi,N,h}(f) = N^{-1} \sum_{i=1}^{N} \sum_{k\in\mathbb{Z}} w_{k,h}(X^i) f(\mathrm{T}^k_h(X^i)) \,, \tag{A.2.1}$$

where $\{X_i\}_{i=1}^N \overset{\text{iid}}{\sim} \rho$ and for some weight function $\varpi^{\mathrm{c}} : \mathbb{R} \to \mathbb{R}_+$ with bounded support (see (**A14**)), $k \in \mathbb{Z}$ and $h > 0$, setting $\varpi_{k,h} = \varpi^{\mathrm{c}}(kh)$,

$$w_{k,h}(x) = \varpi_{k,h}\rho_{-k}(x) \Big/ \sum\nolimits_{i\in\mathbb{Z}} \varpi_{k+i,h}\rho_i(x) \quad . \tag{A.2.2}$$

We show in this section the convergence of the sequence of NEO-IS estimators $\{I^{\mathrm{NEO}}_{\varpi,N,h}(f) : h \in \left(0, \bar{h}\right]\}$ as $h \downarrow 0$ to its continuous counterpart, the version (2.4.1) of NEIS Rotskoff and Vanden-Eijnden (2019), with weight function ϖ, in the case where for any $h \in \left(0, \bar{h}\right]$, T_h corresponds to one step of a discretization scheme with stepsize h of the ODE

$$\dot{x}_t = b(x_t) \,, \tag{A.2.3}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is a drift function. We are particularly interested in the case where (A.2.3) corresponds to the conformal Hamilonian dynamics (2.2.10) and $\{\mathrm{T}_h : h \in \left(0, \bar{h}\right]\}$ to its conformal symplectic Euler discretization: for all $(q, p) \in \mathbb{R}^{2d}$,

$$\mathrm{T}_h(q, p) = (q + h\mathrm{M}^{-1}\{\mathrm{e}^{-h\gamma}p - h\nabla U(q)\}, \mathrm{e}^{-h\gamma}p - h\nabla U(q)) \,. \tag{A.2.4}$$

We make the following conditions on b, ρ, ϖ^{c} and $\{\mathrm{T}_h : h \in \left(0, \bar{h}\right]\}$.

(**A12**) The function b is continuously differentiable and L_b-Lipschitz.

Under (**A12**), consider $(\phi_t)_{t\geq 0}$ the differential flow associated with (A.2.3), *i.e.* $\phi_t(x) = x_t$ where $(x_t)_{t\in\mathbb{R}}$ is the solution of (A.2.3) starting from x. Note that (**A12**) implies that $(t,x) \mapsto \phi_t(x)$ is continuously differentiable on $\mathbb{R}\times\mathbb{R}^d$, see (Hartman, 1982, Theorem 4.1 Chapter V).

(**A12**) is satisfied in the case of the conformal Hamiltonian dynamics if the potential U is continuously differentiable and with Lipschitz gradient, that is there exists $L_U \in \mathbb{R}_+^*$ such that for any $x_1, x_2 \in \mathbb{R}^d$, $\|\nabla U(x_1) - \nabla U(x_2)\| \leq L_U \|x_1 - x_2\|$.

(**A13**) For any $h \in \left(0, \bar{h}\right]$, $\mathrm{T}_h : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1-diffeomorphism. In addition, it holds:

(i) there exist $C \geq 0$ and $\delta \in (0,1]$ such that for any $x \in \mathbb{R}^d$,

$$\|\mathrm{T}_h(x) - (x + hb(x))\| \leq Ch^{1+\delta}(1 + \|x\|)\,;$$

(ii) for any $x \in \mathbb{R}^d$ and $T \in \mathbb{R}_+^*$,

$$\lim_{h\downarrow 0} \max_{k\in[-\lfloor T/h\rfloor:\lfloor T/h\rfloor]} \|\mathbf{J}_{\phi_{kh}}(x) - \mathbf{J}_{\mathrm{T}_h^k}(x)\| = 0\,.$$

Note that (**A13**) is automatically satisfied for the conformal symplectic Euler discretization (A.2.4) of the conformal Hamiltonian dynamics. Indeed, in that case $\mathcal{R}_b((\| \phi)_t(x)) = \gamma d$, and therefore $\mathbf{J}_{\phi_t}(x) = e^{\gamma dt}$ for $t \in \mathbb{R}$, and for any $h > 0, k \in \mathbb{Z}$, $\mathbf{J}_{\mathrm{T}_h^k}(x) = e^{\gamma dhk}$; see Franca et al. (2020).

Define

$$\mathrm{support}(\varpi^{\mathrm{c}}) = \{t \in \mathbb{R} \,:\, \varpi^{\mathrm{c}}(t) \neq 0\}\,. \tag{A.2.5}$$

(**A14**) (i) ρ is continuous and positive on $\mathbb{R}^d$

(ii) ϖ^{c} is piecewise continuous on $\mathbb{R}$, its support $\mathrm{support}(\varpi^{\mathrm{c}})$ is bounded and $\sup_{(s,t)\in\mathsf{A}_\varpi} \varpi^{\mathrm{c}}(t)/\varpi^{\mathrm{c}}(t+s) = m < \infty$ where

$$\mathsf{A}_\varpi = \{(s,t) \in \mathbb{R}^2;\, t \in \mathrm{support}(\varpi^{\mathrm{c}}),\, (s+t) \in \mathrm{support}(\varpi^{\mathrm{c}})\}\,.$$

(iii) Moreover, for any $x \in \mathbb{R}^d$, we have $\rho_{\mathrm{T}}^c(x) = \int \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t > 0$.

Note that (**A14**) implies that $\sup_{t\in\mathbb{R}} |\varpi^{\mathrm{c}}(t)| < +\infty$. (**A14**) is automatically satisfied for example in the case $\varpi^{\mathrm{c}} = \mathbb{1}_{[-T_1,T_2]}$ for $T_1, T_2 \geq 0$.

Theorem A.2.1. *Assume (**A12**), (**A13**), (**A14**). For any $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded,*

$$\lim_{h\downarrow 0} \left| \sum_{k\in\mathbb{Z}} w_{k,h}(x) f(\mathrm{T}_h^k(x)) - \int_{-\infty}^{\infty} w_t^{\mathrm{c}}(x) f(\phi_t(x))\mathrm{d}t \right| = 0\,,$$

where $\{w_{k,h}\}_{k\in\mathbb{Z}}$ and w_t^{c} are defined in (A.2.2) and (2.4.2) respectively, i.e. *for $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$,*

$$w_t^{\mathrm{c}}(x) = \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) \Big/ \int_{-\infty}^{\infty} \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s\ . \tag{A.2.6}$$

Proof. Let f be a bounded continuous function, $x \in \mathbb{R}^d$. Setting

$$g_{k,h}(x) = \rho(\mathrm{T}_h^k(x))\varpi^{\mathrm{c}}(kh)\mathbf{J}_{\mathrm{T}_h^k}(x) f(\mathrm{T}_h^k(x))$$
$$h\Delta_{k,h}(x) = h\sum_{i\in\mathbb{Z}} \rho(\mathrm{T}_h^i(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}_h^i(x)}\,,$$

we have that

$$\sum_{k\geq 0}\frac{hg_{k,h}(x)}{h\Delta_{k,h}(x)}=\int_0^{T_\varpi}\frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)}g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t+\int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h}\frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)}g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t\,,$$

as $g_{k,h}(x)=0$ when $k>\lfloor T_\varpi/h\rfloor$. Therefore, we can consider the following decomposition,

$$\left|\sum_{k\geq 0}\frac{\rho(\mathrm{T}_h^k(x))\varpi^{\mathrm{c}}(kh)\mathbf{J}_{\mathrm{T}_h^k}(x)f(\mathrm{T}_h^k(x))}{\sum_{i\in\mathbb{Z}}\rho(\mathrm{T}_h^i(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}_h^i(x)}}-\int_0^{T_\varpi}\frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))\mathrm{d}t}{\int\varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s}\right|\leq A+B$$

with

$$A=\left|\int_0^{T_\varpi}\frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)}\left\{g_{\lfloor t/h\rfloor,h}(x)-\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))\right\}\mathrm{d}t\right|$$
$$+\left|\int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h}\frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)}g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t\right|,$$

and

$$B=\int_0^{T_\varpi}\left|\frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))\mathrm{d}t}{h\Delta_{\lfloor t/h\rfloor,h}(x)}-\frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))}{\int\varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s}\right|\mathrm{d}t\,,$$

We bound those terms separately. First of all, under (**A14**)-(ii), for any k such that $kh\in[0,T_\varpi]$, we have $h\Delta_{k,h}(x)\geq hm^{-1}\Delta_{0,h}(x)$. Second, as $\lim_{h\downarrow 0}h\Delta_{0,h}(x)=\int_0^{T_\varpi}\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\varpi^{\mathrm{c}}(s)\mathrm{d}s>0$, there exists some $\tilde{h}>0$ and $c>0$ such that for all $k\in\mathbb{Z}$, $h<\tilde{h}$ implies

$$\int_0^{T_\varpi}\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t>c\,,\quad h\Delta_{k,h}(x)\geq hm^{-1}\Delta_{0,h}(x)>c\,.\tag{A.2.7}$$

Then, for $h<\tilde{h}$,

$$A\leq c^{-1}\int_0^{T_\varpi}|g_{\lfloor t/h\rfloor,h}(x)-\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))|\mathrm{d}t$$
$$+c^{-1}\int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h}\left|g_{\lfloor t/h\rfloor,h}(x)\right|\mathrm{d}t\,.$$

By (**A12**) and (**A14**), the function $t\to\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))$ is continuous on the compact $[0,2T_\varpi]$ and thus is bounded. Therefore, for any $h\in\left(0,\bar{h}\right)$,

$$\sup_{t\in[0,2T_\varpi]}|\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))|\leq\sup_{t\in\mathbb{R}}|\varpi^{\mathrm{c}}|\sup_{x\in\mathbb{R}^d}|f(x)|\sup_{t\in[0,2T_\varpi]}|\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)|<\infty\,.\tag{A.2.8}$$

Under (**A13**), (A.2.8) and Lemma A.2.5 imply that

$$\sup_{t\in[0,h\lfloor T_\varpi/h\rfloor+h)}g_{\lfloor t/h\rfloor,h}(x)$$
$$\leq\sup_{t\in\mathbb{R}}|\varpi^{\mathrm{c}}(t)|\sup_{x\in\mathbb{R}^d}|f(x)|\sup_{t\in[0,h\lfloor T_\varpi/h\rfloor+h)}\rho(\mathrm{T}_h^{\lfloor t/h\rfloor}(x))\mathbf{J}_{\mathrm{T}_h^{\lfloor t/h\rfloor}}(x)<\infty\,,$$

Then, $\lim_{h\downarrow 0}\int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h}\left|g_{\lfloor t/h\rfloor,h}(x)\right|\mathrm{d}t=0$. Finally, Lemma A.2.6 implies that $\lim_{h\downarrow 0}A=0$.

Moreover, setting for $t \in [0, T_\varpi]$,

$$\begin{aligned}
&\Delta^B_{t,h}(x) \qquad\qquad (A.2.9)\\
&= \int |\rho(\phi_{h\lfloor s/h\rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h\rfloor + \lfloor t/h\rfloor))\mathbf{J}_{\phi_{h\lfloor s/h\rfloor}(x)} - \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)|\mathbb{1}_{\mathsf{A}_\varpi}(s,t)\mathrm{d}s\\
&\quad + \int_{T_\varpi - h\lfloor t/h\rfloor}^{h(\lfloor T_\varpi/h\rfloor - \lfloor t/h\rfloor + 1)} |\rho(\phi_{h\lfloor s/h\rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h\rfloor + \lfloor t/h\rfloor))\mathbf{J}_{\phi_{h\lfloor s/h\rfloor}(x)}|\mathbb{1}_{\mathsf{A}_\varpi}(s,t)\mathrm{d}s\,,
\end{aligned}$$

we have for $h < \tilde{h}$, by (A.2.7) and (**A14**)-(ii),

$$\begin{aligned}
B &= \int_0^{T_\varpi} \left| \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{h\Delta_{\lfloor t/h\rfloor,h}(x)} - \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{\int \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s} \right| \mathrm{d}t\\
&\le \int_0^{T_\varpi} \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{h\Delta_{\lfloor t/h\rfloor,h}(x)\int \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s} \Delta^B_{t,h}(x)\mathrm{d}t\\
&\le mc^{-2}\int_0^{T_\varpi} \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))\Delta^B_{t,h}(x)\mathrm{d}t\\
&\le mc^{-2}\sup_{t\in\mathbb{R}}|\varpi^{\mathrm{c}}(t)| \sup_{x\in\mathbb{R}^d}|f(x)| \sup_{t\in[0,T_\varpi]}|\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| \int_0^{T_\varpi} \Delta^B_{t,h}(x)\mathrm{d}t\,. \qquad (A.2.10)
\end{aligned}$$

By (**A12**) and (**A14**), the function $s \to \rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)$ is continuous on the interval $[-T_\varpi, T_\varpi]$ and thus is bounded. Therefore, for any $h \in \left(0, \bar{h}\right)$,

$$\begin{aligned}
&\sup_{(s,t)\in\mathsf{A}_\varpi} |\varpi^{\mathrm{c}}(h(\lfloor t/h\rfloor + \lfloor s/h\rfloor))\rho(\phi_{h\lfloor s/h\rfloor}(x))\mathbf{J}_{\phi_{h\lfloor s/h\rfloor}}(x)|\\
&\quad \le \sup_{(s,t)\in\mathsf{A}_\varpi} |\varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| < T_\varpi \sup_{s\in\mathbb{R}}|\varpi^{\mathrm{c}}(s)| \sup_{s\in[-T_\varpi,T_\varpi]}|\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| < \infty\,. \qquad (A.2.11)
\end{aligned}$$

This implies that

$$\lim_{h\downarrow 0} \int_{T_\varpi - h\lfloor t/h\rfloor}^{h(\lfloor T_\varpi/h\rfloor - \lfloor t/h\rfloor + 1)} |\rho(\phi_{h\lfloor s/h\rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h\rfloor + \lfloor t/h\rfloor))\mathbf{J}_{\phi_{h\lfloor s/h\rfloor}(x)}|\mathrm{d}s = 0\,.$$

Moreover, for any $t \in [0, T_\varpi]$, the function

$$s \mapsto |\varpi^{\mathrm{c}}(h(\lfloor t/h\rfloor + \lfloor s/h\rfloor))\rho(\phi_{h\lfloor s/h\rfloor}(x))\mathbf{J}_{\phi_{h\lfloor s/h\rfloor}}(x) - \varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)|\mathbb{1}_{\mathsf{A}_\varpi}(s,t)$$

converges pointwise to 0 for almost all $s \in \mathbb{R}$ when $h \downarrow 0$ using (**A12**), (**A14**) and the continuity of $s \mapsto \phi_s(x)$. The Lebesgue dominated convergence theorem applies and by (A.2.9), for all $t \in [0, T_\varpi]$,

$$\lim_{h\downarrow 0} \Delta^B_{t,h}(x) = 0\,.$$

Moreover, using $h\Delta_{k,h}(x) = h\sum_{i\in\mathbb{Z}} \rho(\mathrm{T}^i_h(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}^i_h(x)}$ and (A.2.11),

$$\sup_{t\in[0,T_\varpi]} \sup_{h\in\left(0,\bar{h}\right)} \Delta^B_{t,h}(x) < \infty\,.$$

The Lebesgue dominated convergence theorem and (A.2.10) show that $\lim_{h\downarrow 0} B = 0$ which concludes the proof. □

A.2.1.1 Supporting Lemmas

For $f \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, define $\mathfrak{J}_f(x)$ the Jacobian matrix of f evaluated at x and the divergence operator by $\mathcal{R}_f((\| x)) = \mathrm{tr}[\mathfrak{J}_f(x)]$.

Lemma A.2.2. *Let b be a C^1 vector field in $\mathbb{R}^d$ and $(\phi_t)_{t\in\mathbb{R}}$ be the flow of the ODE (A.2.3). For any $t \in \mathbb{R}$, the Jacobian of ϕ_t is given by*

$$\mathbf{J}_{\phi_t}(x) = \exp(\textstyle\int_0^t \mathcal{R}_b((\| \phi)_s(x))\mathrm{d}s)\,.$$

Proof. First, for $t \in \mathbb{R}$ and $x \in \mathbb{R}$, write $A(t,x) = \mathfrak{J}_{\phi_t}(x)$ the Jacobian matrix of ϕ_t evaluated at x. By Jacobi's formula, $\det \dot{A}(t,x) = \mathrm{tr}[\mathrm{adj}(A(t,x)) \cdot \dot{A}(t,x)]$, where $\mathrm{tr}[M]$ denotes the trace of a matrix M and $\mathrm{adj}(M)$ its adjugate, i.e. the transpose of the cofactor matrix of M such that $\mathrm{adj}(M)M = \det(M)\mathrm{I}$. Since for all t and x, $\dot{A}(t,x) = \mathfrak{J}_{b\circ\phi_t}(x) = \mathfrak{J}_b(\phi_t(x)) \cdot A(t,x)$, then

$$\dot{\mathbf{J}}_{\phi_t}(x) = \mathrm{tr}[\mathrm{adj}(A(t,x)) \cdot \mathfrak{J}_b(\phi_t(x)) \cdot A(t,x)] = \mathrm{tr}[\mathfrak{J}_b(\phi_t(x))]\mathbf{J}_{\phi_t}(x)\,. \tag{A.2.12}$$

Integrating this ODE yields $\mathbf{J}_{\phi_t}(x) = \exp(\int_0^t \mathcal{R}_b((\| \phi)_s(x))\mathrm{d}s)$. □

Lemma A.2.3. *Assume (**A12**). Then, there exists $C > 0$ such that for any $x \in \mathbb{R}^d, t \in \mathbb{R}$, $k \in \mathbb{Z}$, $h > 0$,*

$$\begin{aligned}\|\phi_t(x)\| &\leq Ce^{C|t|}(\|x\| + 1)\,,\\ \|\,\mathrm{T}_h^k(x)\| &\leq Ce^{C|kh|}(\|x\| + 1)\,.\end{aligned}$$

This lemma follows from Gronwall's inequality and (**A12**).

Lemma A.2.4. *Assume (**A12**) and (**A13**)-(i). There exists $C > 0$ such that for any $x \in \mathbb{R}^d, h \in \left(0, \bar{h}\right)$,*

$$\|\,\mathrm{T}_h(x) - \phi_h(x)\| \leq C\{1 + \|x\|\}\|h^{1+\delta}\,. \tag{A.2.13}$$

Proof. Under (**A12**) and (**A13**)-(i), we have

$$\|\,\mathrm{T}_h(x) - \phi_h(x)\| \leq \|x + hb(x) - \phi_h(x)\| + C_F h^{1+\delta}(1 + \|x\|)\,,$$

and as $\phi_h(x) = x + \int_0^h b(\phi_s(x))\mathrm{d}s$,

$$\begin{aligned}\|x + hb(x) - \phi_h(x)\| = \|hb(x) - \textstyle\int_0^h b(\phi_s(x))\| &\leq hL_b \sup_{s\in[0,h]} \|\phi_s(x) - x\|\\ &\leq L_b h^2\{L_b \sup_{s\in[0,h]} \phi_s(x) + \|b(0)\|\}\,.\end{aligned} \tag{A.2.14}$$

The proof is completed using Lemma A.2.3. □

Lemma A.2.5. *Assume (**A12**) and (**A13**)-(i). There exists $C > 0$ such that for any $x \in \mathbb{R}^d, k \in \mathbb{N}, h \in \left(0, \bar{h}\right)$, $kh \leq T_\varpi$,*

$$\|\,\mathrm{T}_h^k(x) - \phi_{kh}(x)\| \leq Ce^{khC}(1 + \|x\|)h^\delta\,. \tag{A.2.15}$$

Proof. Using Lemma A.2.4, (**A12**) and (**A13**)-(i), there exist $C_1, C_2, C_3 > 0$ such that for any $x \in \mathbb{R}^d, k \in \mathbb{N}, h \in \left(0, \bar{h}\right)$, $kh \leq T_\varpi$,

$$
\begin{aligned}
&\| \mathrm{T}_h^{k+1}(x) - \phi_{(k+1)h}(x)\| \leq \| \mathrm{T}_h^{k+1}(x) - \mathrm{T}_h \circ \phi_{kh}(x)\| + \| \mathrm{T}_h \circ \phi_{kh}(x) - \phi_{(k+1)h}(x)\| \\
&\leq (1 + hL_b) \| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \\
&\qquad\qquad + h^{1+\delta} C_1 \{2 + \| \mathrm{T}_h^k(x)\| + \|\phi_{kh}(x)\|\} + \| \mathrm{T}_h \circ \phi_{kh}(x) - \phi_{(k+1)h}(x)\| \\
&\leq (1 + hL_b) \| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| + h^{1+\delta} 2C_1 C_2 e^{C_2 T_\varpi} \{1 + \|x\|\} + C_3 \{1 + \|\phi_{kh}(x)\|\} h^{1+\delta} \\
&\leq (1 + hL_b) \| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \\
&\qquad\qquad + h^{1+\delta} 2C_1 C_2 e^{C_2 T_\varpi} \{1 + \|x\|\} + C_3 \{1 + C_2(1 + \|x\|)\} h^{1+\delta} e^{C_2 T_\varpi} \\
&\leq (1 + hL_b) \| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| + A_T \{1 + \|x\|\} h^{1+\delta} \,,
\end{aligned}
$$

with $A_T = (2C_1 C_2 + C_3(1 + C_2)) e^{C_2 T_\varpi}$. A straightforward induction yields

$$
\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \leq \frac{(1 + hL_b)^k}{L_b} A_T (1 + \|x\|) h^\delta \,.
$$

□

Lemma A.2.6. *Assume (**A12**), (**A13**), (**A14**) . For any $x \in \mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R}^d$ bounded and continuous,*

$$
\lim_{h \downarrow 0} \int_0^{T_\varpi} \left| \varpi^{\mathrm{c}}(h \lfloor t/h \rfloor) \rho(\mathrm{T}_h^{\lfloor t/h \rfloor}(x)) \mathbf{J}_{\mathrm{T}_h^{\lfloor t/h \rfloor}}(x) f(\mathrm{T}_h^{\lfloor t/h \rfloor}(x)) - \varpi^{\mathrm{c}}(t) \rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x) f(\phi_t(x)) \right| \mathrm{d}t = 0 \,.
$$

Proof. Let $x \in \mathbb{R}^d$. Consider the following decomposition, for any $h < \bar{h}$,

$$
\begin{aligned}
&\int_0^{T_\varpi} \left| \varpi^{\mathrm{c}}(h \lfloor t/h \rfloor) \rho(\mathrm{T}_h^{\lfloor t/h \rfloor}(x)) \mathbf{J}_{\mathrm{T}_h^{\lfloor t/h \rfloor}}(x) f(\mathrm{T}_h^{\lfloor t/h \rfloor}(x)) - \varpi^{\mathrm{c}}(t) \rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x) f(\phi_t(x)) \right| \mathrm{d}t \\
&\leq \tfrac{h}{T_\varpi} \textstyle\sum_{k \in \mathbb{Z}} \varpi^{\mathrm{c}}(kh) |\rho(\mathrm{T}_h^k(x)) \mathbf{J}_{\mathrm{T}_h^k}(x) f(\mathrm{T}_h^k(x)) - \rho(\phi_{kh}(x)) \mathbf{J}_{\phi_{kh}}(x) f(\phi_{kh}(x))| \\
&+ \textstyle\int_0^{T_\varpi} |\varpi^{\mathrm{c}}(t) \rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x) f(\phi_t(x)) - \varpi^{\mathrm{c}}(h \lfloor t/h \rfloor) \rho(\phi_{h \lfloor t/h \rfloor}(x)) \mathbf{J}_{\phi_{h \lfloor t/h \rfloor}}(x) f(\phi_{h \lfloor t/h \rfloor}(x))| \mathrm{d}t \,.
\end{aligned}
$$

The first term converges to 0 by Lemma A.2.5 and (**A13**)-(ii) as $\varpi^{\mathrm{c}}(kh) = 0$ for $kh > T_\varpi$. By (**A12**) and (**A14**), the function $t \to \varpi^{\mathrm{c}}(t) \rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x) f(\phi_t(x))$ is continuous on the compact $[0, T_\varpi]$ and thus is bounded. Therefore, for any $h \in \left(0, \bar{h}\right)$,

$$
\begin{aligned}
&\sup_{t \in [0, T_\varpi]} |\varpi^{\mathrm{c}}(h \lfloor t/h \rfloor) \rho(\phi_{h \lfloor t/h \rfloor}(x)) \mathbf{J}_{\phi_{h \lfloor t/h \rfloor}}(x) f(\phi_{h \lfloor t/h \rfloor}(x))| \\
&\qquad\qquad \leq \sup_{t \in \mathbb{R}} |\varpi^{\mathrm{c}}| \sup_{x \in \mathbb{R}^d} |f(x)| \sup_{t \in [0, T_\varpi]} |\rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x)| < \infty \,. \qquad \text{(A.2.16)}
\end{aligned}
$$

Moreover, $t \mapsto \varpi^{\mathrm{c}}(h \lfloor t/h \rfloor) \rho(\phi_{h \lfloor t/h \rfloor}(x)) \mathbf{J}_{\phi_{h \lfloor t/h \rfloor}}(x) f(\phi_{h \lfloor t/h \rfloor}(x))$ converges pointwise when $h \downarrow 0$ to $t \to \varpi^{\mathrm{c}}(t) \rho(\phi_t(x)) \mathbf{J}_{\phi_t}(x) f(\phi_t(x))$ by continuity, using (**A12**) and (**A14**). The Lebesgue dominated convergence theorem applies and the second term goes to 0 as $h \downarrow 0$. □

A.2.2 NEIS algorithm after Rotskoff and Vanden-Eijnden (2019)

Non Equilibrium Importance Sampling (NEIS) has been introduced in the pioneering work of Rotskoff and Vanden-Eijnden (2019). NEIS relies on the flow of the ODE $\dot{x}_t = b(x_t)$ and the introduction of a set $\mathsf{O} \subset \mathbb{R}^d$. As in Section A.2, we assume (**A12**) holds and denote by $(\phi_t)_{t \in \mathbb{R}}$ the flow of this ODE.

Define for $x \in \mathsf{O}$, the exit times $\tau^+(x) \geq 0$ (resp. $\tau^-(x) \leq 0$) satisfying

$$\tau^+(x) = \inf\{t \geq 0 \,:\, \phi_t(x) \notin \mathsf{O}\}\,,\ \tau^-(x) = \inf\{t \leq 0 \,:\, \phi_t(x) \notin \mathsf{O}\}\,. \tag{A.2.17}$$

The validity of NEIS relies on the following assumption.

(**A15**) The average time of an orbit in O is finite, *i.e.*

$$\mathcal{Z}_\tau = \int_{\mathsf{O}} (\tau^+(x) - \tau^-(x))\rho(x)\mathrm{d}x < \infty\,. \tag{A.2.18}$$

Under (**A15**), we can define the proposal distribution

$$\rho_{\mathrm{T}}(x) = \mathcal{Z}_\tau^{-1} \int_{\mathsf{O}} \mathbb{1}_{[\tau^-(x),\tau^+(x)]}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t\,. \tag{A.2.19}$$

Under (**A15**), (Rotskoff and Vanden-Eijnden, 2019, Equation (8)) derive the following estimator of $\rho(f)$, closely related to (2.4.1), in the case $\varpi \equiv 1$, on the restricted set $\mathsf{O} \subset \mathbb{R}^d$:

$$I_N^{\mathrm{NEIS}}(f) = \frac{1}{N}\sum_{i=1}^N \int_{\tau^-(X^i)}^{\tau^+(X^i)} w_t(X^i) f(\phi_t(X^i))\mathrm{d}t \tag{A.2.20}$$

$$w_t(x) = \frac{\rho(\phi_t(x))\mathbf{J}_{\phi_t(x)}}{\int_{\tau^-(x)}^{\tau^+(x)} \rho(\phi_t(x))\mathbf{J}_{\phi_t(x)}\mathrm{d}t}\,. \tag{A.2.21}$$

Note that in practice, in order for (**A15**) to be verified, one typically requires that O be bounded, as discussed in Rotskoff and Vanden-Eijnden (2019).

Following Rotskoff and Vanden-Eijnden (2019), consider a d-dimensional system with position $q \in \mathbb{R}^d$, momentum $p \in \mathbb{R}^d$ and Hamiltonian $H(p,q) = (1/2)\|p\|^2 + U(q)$ where $U(q)$ is a potential assumed to be bounded from below. Denote by $V(E)$ the volume of the phase-space below some threshold energy E,

$$V(E) = \int \mathbb{1}_{\{H(p,q)\leq E\}}\mathrm{d}p\mathrm{d}q\,. \tag{A.2.22}$$

To calculate (A.2.22), we set $x = (p,q)$, define $\mathsf{O} = \{x; H(x) \leq E_{\max}\}$ for some $E_{\max} < \infty$, and use the dissipative Langevin dynamics with $b(x) = (p, -\nabla U(q) - \gamma p)$, *i.e.*

$$\dot{q} = p\,, \quad \dot{p} = -\nabla U(q) - \gamma p\,,$$

for some friction coefficient $\gamma > 0$. With this choice, $\mathbf{J}_{\phi_t}(x) = \mathrm{e}^{-d\gamma t}$. Taking ρ to be the uniform distribution on the (bounded) set O, write the estimator for $E \leq E_{\max}$, $V(E)/V(E_{\max}) = \int \mathbb{1}_{\{H(p,q)\leq E\}}\rho(p,q)\mathrm{d}p\mathrm{d}q$, where $\rho(p,q) = \mathbb{1}_{\mathsf{O}}(p,q)/V(E_{\max})$, we get

$$\begin{aligned} V(E)/V(E_{\max}) &= \frac{1}{N}\sum_{i=1}^N \frac{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)}\mathbb{1}_{\{H(\phi_t(X^i))\leq E\}}\mathrm{d}t}{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)}\mathrm{d}t} \\ &= \frac{1}{N}\sum_{i=1}^N \frac{\int_{\tau^E(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)}\mathrm{d}t}{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)}\mathrm{d}t} = \frac{1}{N}\sum_{i=1}^N \mathrm{e}^{-d\gamma(\tau^E(X^i)-\tau^-(X^i))}\,, \end{aligned} \tag{A.2.23}$$

where $\tau^E(x)$ denotes the (possibly infinite) time for a trajectory initiated at $x = (p,q)$ to reach the energy $E \leq E_{\max}$.

Finally, to estimate the normalizing constant, Rotskoff and Vanden-Eijnden (2019) discretize the energy levels $\{E_0, \ldots, E_P\}$ and write their estimator as

$$\widehat{\mathcal{Z}}^{\mathrm{NEIS}}_{X^{1:N}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\ell=1}^{P} \mathrm{e}^{-d\gamma(\tau^E_\ell(X^i) - \tau^-(X^i))} (E_\ell - E_{\ell-1}) \,, \tag{A.2.24}$$

using an approximation of the identity

$$\mathcal{Z} = \int_{\mathsf{O}} \int_0^\infty \mathbb{1}_{\{\mathrm{L}(x) > L\}} \rho(x) \mathrm{d}L \mathrm{d}x = \int_0^\infty \mathbb{P}_{X \sim \rho}(\mathrm{L}(X) > L) \mathrm{d}L \,,$$

which is at the core of nested sampling Chopin and Robert (2010).

A.2.3 NEO with exit times

Consider $\mathsf{O} \subset \mathbb{R}^d$ and let T be a C^1-diffeomorphism on $\mathbb{R}^d$. We introduce here an estimator based on the forward and backward orbits in O associated with T. Define the exit times $\tau^+ : \mathbb{R}^d \to \mathbb{N}$ and $\tau^- : \mathbb{R}^d \to \mathbb{N}_-$, given, for all $x \in \mathbb{R}^d$, by

$$\tau^+(x) = \inf\{k \geq 1 \,:\, \mathrm{T}^k(x) \notin \mathsf{O}\} \,, \tag{A.2.25}$$

$$\tau^-(x) = \sup\{k \leq -1 \,:\, \mathrm{T}^k(x) \notin \mathsf{O}\} \,, \tag{A.2.26}$$

with the convention $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$, and set

$$\mathrm{I} = \{(x, k) \in \mathsf{O} \times \mathbb{Z} \,:\, k \in [\tau^-(x) + 1 : \tau^+(x) - 1]\} \,. \tag{A.2.27}$$

For any $k \in \mathbb{Z}$, define $\rho_k : \mathbb{R}^d \to \mathbb{R}_+$ by

$$\rho_k(x) = \rho(\mathrm{T}^{-k}(x)) \mathbf{J}_{\mathrm{T}^{-k}}(x) \mathbb{1}_{\mathrm{I}}(x, -k) \,. \tag{A.2.28}$$

The density ρ_k is the push-forward of $\mathbb{1}_{\mathrm{I}}(x, k)\rho(x)$ by T^k, *i.e.* for any $k \in \mathbb{Z}$ and any bounded function $g : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathsf{O}} g(y) \rho_k(y) \mathrm{d}y = \int_{\mathsf{O}} g(\mathrm{T}^k(x)) \mathbb{1}_{\mathrm{I}}(x, k) \rho(x) \mathrm{d}x \,. \tag{A.2.29}$$

Consider the following assumption:

(**A16**) The nonnegative sequence $(\varpi_k)_{k \in \mathbb{Z}}$ satisfies $\varpi_0 > 0$ and

$$\mathcal{Z}^{\varpi}_{\mathrm{T}} = \int_{\mathsf{O}} \sum_{k \in \mathbb{Z}} \varpi_k \rho_k(x) \mathrm{d}x = \int_{\mathsf{O}} \sum_{k \in \mathbb{Z}} \varpi_k \rho(\mathrm{T}^k(x)) \mathbf{J}_{\mathrm{T}^k}(x) \mathbb{1}_{\mathrm{I}}(x, k) \mathrm{d}x < \infty \,. \tag{A.2.30}$$

Consider the pdf

$$\rho_{\mathrm{T}}(x) = \frac{1}{\mathcal{Z}^{\varpi}_{\mathrm{T}}} \sum_{k \in \mathbb{Z}} \varpi_k \rho_k(x) \,, \tag{A.2.31}$$

where $\mathcal{Z}^{\varpi}_{\mathrm{T}}$ is the normalizing constant. This is a *non-equilibrium* distribution, since ρ_{T} is not invariant by T in general. Using ρ_{T} as an importance distribution to obtain an unbiased estimator of $\int f(x)\rho(x)\mathrm{d}x$ is feasible since as $\varpi_0 > 0$, $\sup_{x \in \mathsf{O}} \rho(x)/\rho_{\mathrm{T}}(x) \leq \mathcal{Z}_{\mathrm{T}}/\varpi_0 < \infty$, hence

$$\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x = \int_{\mathsf{O}} \left(f(x) \frac{\rho(x)}{\rho_{\mathrm{T}}(x)} \right) \rho_{\mathrm{T}}(x) \mathrm{d}x \,.$$

From (A.2.29), the right hand side can be computed using the following key result.

Theorem A.2.7. *For any $f : \mathbb{R}^d \to \mathbb{R}$ measurable bounded function, we have*

$$\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x = \int_{\mathsf{O}} \sum\nolimits_{k\in\mathbb{Z}} f(\mathrm{T}^k(x))w_k(x)\rho(x)\mathrm{d}x\,, \tag{A.2.32}$$

where, for any $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}$,

$$w_k(x) = \varpi_k\rho_{-k}(x) \Big/ \sum\nolimits_{j\in\mathbb{Z}} \varpi_{j+k}\rho_j(x) \quad . \tag{A.2.33}$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable bounded function. By (A.2.29), writing $g \leftarrow f\rho/\rho_{\mathrm{T}}$,

$$\begin{aligned}\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x &= \int_{\mathsf{O}} \left(f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}(x)}\right)\rho_{\mathrm{T}}(x)\mathrm{d}x \\ &= \int_{\mathsf{O}} \sum_{k\in\mathbb{Z}} \left(f(\mathrm{T}^k(x))\frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{\mathcal{Z}^{\varpi}_{\mathrm{T}}\rho_{\mathrm{T}}(\mathrm{T}^k(x))}\right)\rho(x)\mathrm{d}x\,.\end{aligned}$$

We now need to prove:

$$\frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{\mathcal{Z}^{\varpi}_{\mathrm{T}}\rho_{\mathrm{T}}(\mathrm{T}^k(x))} = \frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x))} = \frac{\varpi_k\rho_{-k}(x)}{\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho_j(x)} = w_k(x)\,,$$

with the convention $0/0 = 0$. We thus need to show that for any $x \in \mathsf{O}$, $k \in \mathbb{Z}$,

$$\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x)) = \frac{\mathbb{1}_{\mathrm{I}}(x,k)}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho_j(x)\,.$$

Using the identity $\mathbf{J}_{\mathrm{T}^{-i+k}}(x) = \mathbf{J}_{\mathrm{T}^{-i}}(\mathrm{T}^k(x))\mathbf{J}_{\mathrm{T}^k}(x)$, we obtain

$$\begin{aligned}\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x)) &= \sum_{i\in\mathbb{Z}}\mathbb{1}_{\mathrm{I}}(x,k)\varpi_i\rho(\mathrm{T}^{-i}(\mathrm{T}^k(x)))\mathbf{J}_{\mathrm{T}^{-i}}(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-i) \\ &= \frac{1}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{i\in\mathbb{Z}}\mathbb{1}_{\mathrm{I}}(x,k)\varpi_i\rho(\mathrm{T}^{-i+k}(x))\mathbf{J}_{\mathrm{T}^{-i+k}}(x)\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-i) \\ &= \frac{1}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho(\mathrm{T}^{-j}(x))\mathbf{J}_{\mathrm{T}^{-j}}(x)\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-j-k)\mathbb{1}_{\mathrm{I}}(x,k)\end{aligned}$$

Note that if $(x,k) \in \mathrm{I}$, we have $(x,-j) \in \mathrm{I}$ if and only if $(\mathrm{T}^k(x),-j-k) \in \mathrm{I}$ by definition of I (A.2.27). The proof is concluded by noting that:

$$\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x)),-j-k)\mathbb{1}_{\mathrm{I}}(x,k) = \mathbb{1}_{\mathrm{I}}(x,-j)\mathbb{1}_{\mathrm{I}}(x,k)\,.$$

□

A.3 Iterated SIR

Let us recall the principle of the Sampling Importance Resampling method (SIR; Rubin (1987); Smith and Gelfand (1992)) whose goal is to approximately sample from the target distribution π using samples drawn from a proposal distribution ρ.

In SIR, a N-i.i.d. sample $X^{1:N}$ is first generated from the proposal distribution ρ. A sample X^* is approximately drawn from the target π by choosing randomly a value in $X^{1:N}$ with probabilities

proportional to the importance weights $\{\mathrm{L}(X^i)\}_{i=1}^N$, where $\mathrm{L}(x) = \pi(x)/\rho(x)$. Note that the importance weights are required to be known only up to a constant factor.

For SIR, as $N \to \infty$, the sample X^* is *asymptotically* distributed according to π; see Smith and Gelfand (1992).

A subsequent algorithm is the *iterated SIR* (i-SIR) Andrieu et al. (2010). Here, N is not necessarily large ($N \geq 2$), the whole process of sampling a set of proposals, computing the importance weights, and picking a candidate, is iterated. At the n-th step of i-SIR, the active set of N proposals $X_n^{1:N}$ and the index $I_n \in [N]$ of the conditioning proposal are kept. First i-SIR updates the active set by setting $X_{n+1}^{I_n} = X_n^{I_n}$ (keep the conditioning proposal) and then draw independently $X_{n+1}^{1:N\setminus\{I_n\}}$ from ρ. Then it selects the next proposal index $I_{n+1} \in [N]$ by sampling with probability proportional to $\{\tilde{w}(X_{n+1}^i)\}_{i=1}^N$. As shown in Andrieu et al. (2010), this algorithm defines a partially collapsed Gibbs sampler (PCG) of the augmented distribution

$$\bar{\boldsymbol{\pi}}(x^{1:N}, i) = \frac{1}{N}\pi(x^i)\prod_{j\neq i}\rho(x^j) = \frac{1}{N}\tilde{w}(x^i)\prod_{j=1}^N \rho(x^j)\,.$$

The PCG sampler can be shown to be ergodic provided that ρ and π are continuous and ρ is positive on the support of π. If in addition the importance weights are bounded, the Gibbs sampler can be shown to be uniformly geometrically ergodic Lindsten et al. (2015); Andrieu et al. (2018a). It follows that the distribution of the conditioning proposal $X_n^* = X_n^{I_n}$ converges to π as the iteration index n goes to infinity. Indeed, for any integrable function f on $\mathbb{R}^d$, with $(X_{1:N}, I) \sim \bar{\boldsymbol{\pi}}$,

$$\mathbb{E}[f(X^I)] \;=\; \int \sum_{i=1}^N f(x^i)\bar{\boldsymbol{\pi}}(x^{1:N}, i)\mathrm{d}x^{1:N} \;=\; N^{-1}\sum_{i=1}^N \int f(x^i)\pi(x^i)\mathrm{d}x_i \;=\; \int f(x)\pi(x)\mathrm{d}x\,.$$

When the state space dimension d increases, designing a proposal distribution ρ guaranteeing proper mixing properties becomes more and more difficult. A way to circumvent this problem is to use dependent proposals, allowing in particular *local moves* around the conditioning orbit. To implement this idea, for each $i \in [N]$, we define a proposal transition, $r_i(x^i; x^{1:N\setminus\{i\}})$ which defines the the conditional distribution of $X^{1:N\setminus\{i\}}$ given $X^i = x^i$. The key property validating i-SIR with dependent proposals is that all one-dimensional marginal distributions are equal to ρ, which requires that for each $i, j \in [N]$,

$$\rho(x^i)r_i(x^i; x^{1:N\setminus\{i\}}) = \rho(x^j)r_j(x^j; x^{1:N\setminus\{j\}}) \tag{A.3.1}$$

The (unconditional) joint distribution of the particles is therefore defined as

$$\rho_N(x^{1:N}) = \rho(x^1)r_1(x^1; x^{1:N\setminus\{1\}})\,. \tag{A.3.2}$$

The resulting modification of the i-SIR algorithm is straightforward: $X^{1:N\setminus\{I_n\}}$ is sampled jointly from the conditional distribution $r_{I_n}(X_n^{I_n}, \cdot)$ rather than independently from ρ.

A.4 Additional Experiments

A.4.1 Normalizing constant estimation

We consider here the problem of the estimation of the normalizing constant of Cauchy mixtures. The Cauchy distribution with scale σ has a pdf defined by $\mathrm{Cauchy}(x; \mu, \sigma) = [\pi\sigma(1 + \{(x -$

$\mu)/\sigma\}^2]^{-1}$. The target distribution is a product of mixtures of two Cauchy distributions,

$$\pi(x) = \prod_{i=1}^{n} \frac{1}{2} \left[\text{Cauchy}\left(x_i; \mu, \sigma\right) + \text{Cauchy}\left(x_i; -\mu, \sigma\right)\right], \quad \mu = 5, \sigma = 1\,.$$

NEO-IS is compared with IS estimator using the same proposal ρ. We also compare NEO-IS to Neural IS Müller et al. (2019b) with a Cauchy as base distribution.

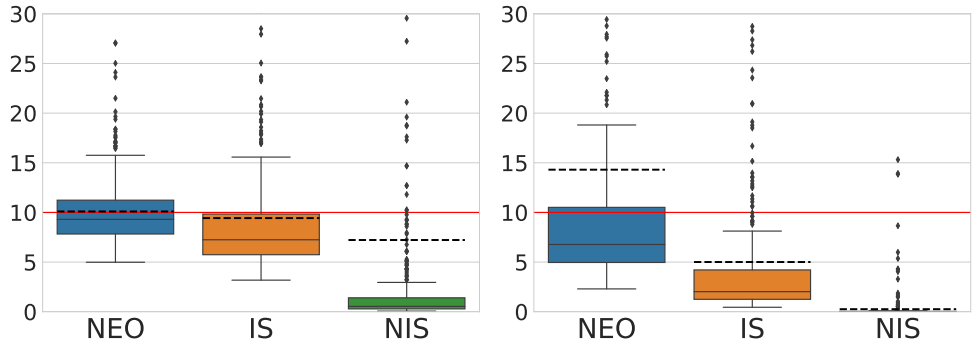

Figure A.1: Boxplots of 500 independent estimations of the normalizing constant of the Cauchy mixture in dimension $d = 10, 15$ (top, bottom). The true value is given by the red line. The figure displays the median (solid lines), the interquartile range, and the mean (dashed lines) over the 500 runs

Finally, we compare NEO-IS with NEIS[1]. We consider here `MG25` in dimension 5 and 10, where all the covariances of the Gaussian distributions are diagonal and equal to 0.005I. NEIS and NEO-IS are run for the same computational time. We add an IS scheme as a baseline for comparison. All algorithms (NEO-IS, NEIS, IS) are run for 7.20s and 11.30s wall clock time respectively for $d = 5$ and $d = 10$. For NEO-IS, we use a conformal transform with $h = 0.1$, $K = 10$ and $\gamma = 1$. For NEIS, we choose $\gamma = 1$ and consider a stepsize $h = 10^{-4}$ corresponding to an optimal trade-off between the discretization bias inherent to NEISand its computational budget. We can observe that NEO-IS always outperforms NEIS, which suffers from a non-negligeable bias if the stepsize h is not chosen small enough.

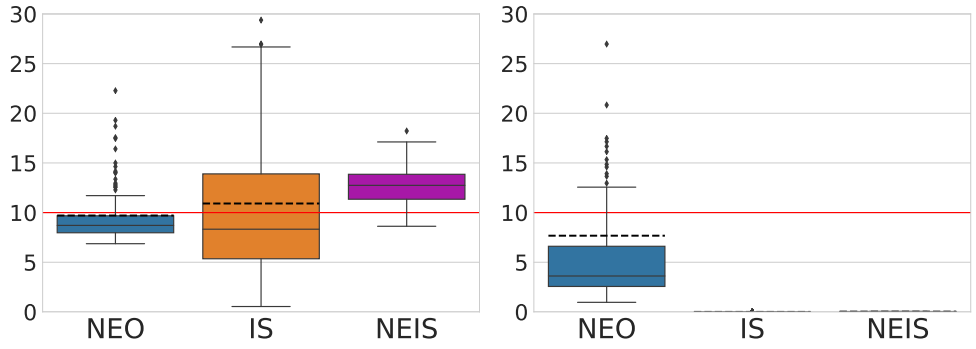

Figure A.2: NEO *v.* NEIS. *25 GM with $\sigma^2 = 0.005$, $d = 5$. 500 runs each.*

[1]The code from Rotskoff and Vanden-Eijnden (2019) we run is available at `https://gitlab.com/rotskoff/trajectory_estimators/-/tree/master`.

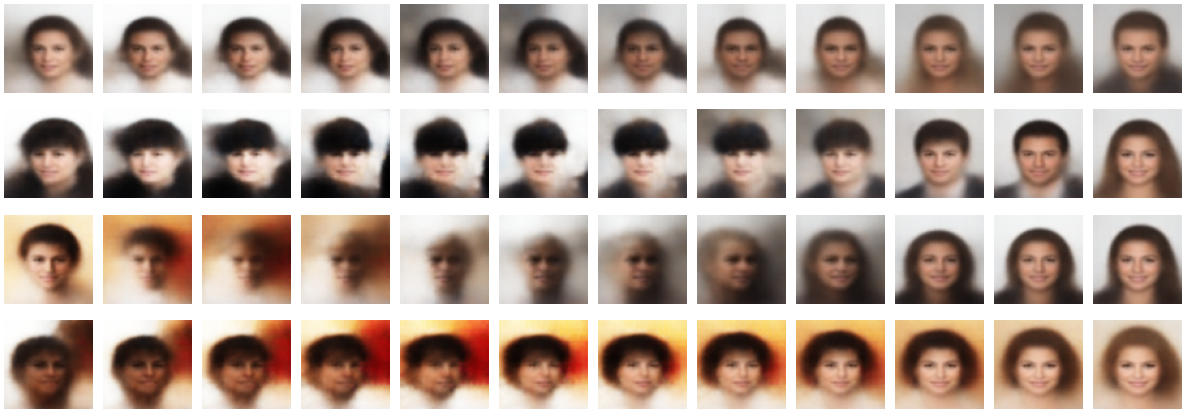

Figure A.3: Forward orbits of NEO-*MCMC.*

A.4.2 Gibbs inpainting

We display here additional results for the Gibbs inpainting experiment presented in Section 2.5. We emphasize that the starting images are chosen at random in the test set.

A.5 NEO and VAEs

Denote by $p_\theta(x, z)$ the joint distribution of the observation $z \in \mathbb{R}^p$ and the latent variable $x \in \mathbb{R}^d$. The marginal likelihood is given, for $z \in \mathbb{R}^p$ by $p_\theta(z) = \int p_\theta(x, z)\mathrm{d}x$. Given a training set $\mathcal{D} = \{z_i\}_{i=1}^M$, the objective is to estimate θ by maximizing the likelihood, *i.e.* maximizing $\log p_\theta(\mathcal{D}) = \sum_{i=1}^M \log p_\theta(z_i)$. We show two experiments in the following, first the evaluation of independently trained VAEs, and then the derivation and learning of a VAE based on NEO, and NEO-VAE.

A.5.1 Log-likelihood estimation

We present here first the evaluation of the log-likelihood of a trained VAE on the dynamically binarized MNIST dataset. The models we compare share the same architecture: the inference network q_ϕ is given by a convolutional network with 2 convolutional layers and one linear layer, which outputs the parameters $\mu_\phi(x), \sigma_\phi(x) \in \mathbb{R}^d$ of a factorized Gaussian distribution, while the generative model $p_\theta(\cdot|z)$ is given by another symmetrical convolutional network g_θ. This outputs the parameters for the factorized Bernoulli distribution (for MNIST dataset), that is

$$p_\theta(z|x) = \prod_{i=1}^N \mathrm{Ber}\Big(z^{(i)}|\big(g_\theta(x)\big)^{(i)}\Big)\,.$$

We here follow the experimental setting of Wu et al. (2016). Given a test set $\mathcal{T} = \{z_i\}_{i=1}^{M_\mathcal{T}}$, we estimate $\sum_{i=1}^{M_\mathcal{T}} \log p_{\theta^*}(z_i)$. We also estimate similarly the log-likelihood of an Importance Weighted Auto Encoder (IWAE) Burda et al. (2016). Following Wu et al. (2016), we compare IS, AIS, and NEO-IS. As previously, AIS, IS, and NEO-IS are given a similar computational budget, choosing here $K = 12$, $N = 5 \cdot 10^3$. For NEO, we choose $\gamma = 1.$ and $h = 0.2$. Similarly, the stepsize of HMC transitions in AIS is $h = 0.1$ in order to achieve an acceptance ratio of around 0.6 in the HMC transitions. We report in Table A.1 the log-likelihood computed on the test set for VAE, IWAE with latent dimension in $\{16, 32\}$. For the same computational budget, NEO-IS yields consistently better values for the estimation of the log-likelihood of the VAE.

Figure A.4: Additional examples for the Gibbs inpainting task for CelebA dataset. From top to bottom: i-SIR, HMC and NEO-*MCMC: From left to right, original image, blurred image to reconstruct, and output every 5 iterations of the Markov chain.*

Model	VAE, $d = 32$	VAE, $d = 16$	IWAE, $d = 32$	IWAE, $d = 16$
IS	-90.17	-90.44	-88.76	-90.13
AIS	-89.67	-89.97	-88.30	-89.61
NEO-IS	-88.81	-89.17	-87.46	-88.99

Table A.1: Evaluation of the log-likelihood (normalizing constant) of different Variational Auto Encoders.

A.5.2 Definition of a NEO-VAE

Variational inference (VI) provides us with a tool to simultaneously approximate the intractable posterior $p_\theta(x|z)$ and maximize the marginal likelihood $p_\theta(\mathcal{D})$ in the parameter θ. This is achieved by introducing a parametric family $\{q_\phi(x|z), \phi \in \Phi\}$ to approximate the posterior $p_\theta(x|z)$ and maximizing the Evidence Lower Bound (ELBO) (see Kingma and Welling (2019)) $\mathcal{L}_{\mathrm{ELBO}}(\mathcal{D}, \theta, \phi) = \sum_{i=1}^{M} \mathcal{L}_{\mathrm{ELBO}}(z_i, \theta, \phi)$ where

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}(z, \theta, \phi) &= \int \log\left(\frac{p_\theta(x, z)}{q_\phi(x \mid z)}\right) q_\phi(x \mid z)\mathrm{d}x \\ &= \log p_\theta(z) - \mathrm{KL}(q_\phi(\cdot \mid z) \| p_\theta(\cdot \mid z))\,,\end{aligned} \tag{A.5.1}$$

and KL is the Kullback–Leibler divergence. In the sequel, we set $\rho(x) = q_\phi(x \mid z)$ and $\mathrm{L}(x) = p_\theta(x, z)/q_\phi(x \mid z)$. In such a case, $\pi(x) = \rho(x)\mathrm{L}(x)/\mathcal{Z} = p_\theta(x \mid z)$ and $\mathcal{Z} = p_\theta(z)$ (in these notations, the dependence in the observation z is implicit).

We follow the the auxiliary variational inference framework (AVI) provided by Agakov and Barber (2004). We consider a joint distribution $\bar{p}_\theta(x, u, z)$ which is such that $p_\theta(z) = \int p_\theta(x, u, z)\mathrm{d}x\mathrm{d}u$ where $u \in \mathsf{U}$ is an auxiliary variable (the auxiliary variable can both have discrete and continuous components; when u has discrete components the integrals should be replaced by a sum). Then as the usual VI approach, we consider a parametric family $\{\bar{q}_\phi(x, u|z), \phi \in \Phi\}$. Introducing auxiliary variables loses the tractability of (A.5.1) but they allow for their own ELBO as suggested in Agakov and Barber (2004); Lawson et al. (2019) by minimizing

$$\mathrm{KL}(\bar{q}_\phi(\cdot \mid z) \| \bar{p}_\theta(\cdot \mid z)) = \int \bar{q}_\phi(x, u|z) \log\left(\frac{\bar{p}_\theta(x, u, z)}{\bar{q}_\phi(x, u|z)}\right) \mathrm{d}x\mathrm{d}u\,. \tag{A.5.2}$$

The auxiliary variable u is naturally associated with the extended target $\bar{p}$ defined similar to Remark 2.3.2,

$$\bar{p}_N([x, x^{1:N\setminus\{i\}}], i) = \check{\pi}(x^{1:N}, i) = \frac{\widehat{\mathcal{Z}}_x^{\varpi}}{N\mathcal{Z}} \rho_N(x^{1:N}) \tag{A.5.3}$$

with $(x, u) = ([x, x^{1:N\setminus\{i\}}], i)$, a shorthand notation for a N-tuple $x^{1:N}$ with $x^i = x$, and, with r_i defined in (2.3.5),

$$\rho_N(x^{1:N}) = \rho(x^1) r_1(x^1, x^{2:N}) = \rho(x^j) r_j(x^j, x^{1:N\setminus\{j\}})\,, \quad j \in \{1, \ldots, N\}\,, \tag{A.5.4}$$

generally for Markov transitions $\{r_j\}_{j\in[N]}$. We might write simply in the following

$$\rho_N(x^{1:N}) = \prod_{i=1}^{N} \rho(x^i)\,.$$

An extended proposal playing the role of $\bar{q}_\phi(x, u|z)$ is derived from the NEO-MCMC sampler, i.e.

$$\bar{q}_N([x, x^{1:N\setminus\{i\}}], i) = \frac{\widehat{\mathcal{Z}}_x^{\varpi}}{N\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}} \rho_N(x^{1:N})\,. \tag{A.5.5}$$

where $\widehat{\mathcal{Z}}_{x^{1:N}}^{\varpi}$ is the NEO estimator (2.2.4) of the normalizing constant. Note that, by construction,

$$\sum_{i=1}^{N} \bar{q}_N(x^{1:N}, i) = \rho_N(x^{1:N}) \tag{A.5.6}$$

Table A.2: Negative Log Likelihood estimates for VAE models for different latent space dimensions.

	$d = 4$		$d = 8$		$d = 16$		$d = 50$	
model	IS	NEO	IS	NEO	IS	NEO	IS	NEO
VAE	115.01	113.49	97.96	97.64	90.52	90.42	88.22	88.36
IWAE, $N = 5$	113.33	111.83	97.19	96.61	89.34	89.05	87.49	87.27
IWAE, $N = 30$	111.92	110.36	96.81	95.94	88.99	88.64	86.97	86.93
NEO VAE, $K = 3$	109.14	107.47	94.50	94.26	89.03	88.92	88.14	88.16
NEO VAE, $K = 10$	110.02	107.90	94.63	94.22	89.71	88.68	88.25	86.95

showing that this joint proposal can be sampled by drawing the proposals $x^{1:N} \sim \rho_N$, then sampling the path index $i \in [N]$ with probability proportional to $(\widehat{\mathcal{Z}}^{\varpi}_{x^i})_{i=1}^N$ (with $\widehat{\mathcal{Z}}^{\varpi}_x$ defined in (2.2.4)). The ratio of (A.5.3) over (A.5.5) is

$$\bar{p}_N(x^{1:N}, i)/\bar{q}_N(x^{1:N}, i) = \widehat{\mathcal{Z}}^{\varpi}_{x^{1:N}}/\mathcal{Z}\,. \tag{A.5.7}$$

Thus, we write the augmented ELBO (A.5.2)

$$\mathcal{L}_{\text{NEO}} = \int \rho_N(x^{1:N}) \log \widehat{\mathcal{Z}}^{\varpi}_{x^{1:N}} \mathrm{d}x^{1:N} = \log \mathcal{Z} - \mathrm{KL}(\bar{q}_N | \bar{p}_N)\,, \tag{A.5.8}$$

where we have used (A.5.6) and that the ratio $\bar{p}_N(x^{1:N}, i)/\bar{q}_N(x^{1:N}, i)$ does not depend on the path index i. When $\varpi_k = \delta_{k,0}$, where $\delta_{i,j}$ is the Kronecker symbol, and $\rho_N(x^{1:N}) = \prod_{j=1}^N \rho(x^j)$, we exactly retrieve the Importance Weighted AutoEncoder (IWAE); see e.g. Burda et al. (2016) and in particular the interpretation in Cremer et al. (2017).

Choosing the conformal Hamiltonian introduced in Section 2.2 allows for a family of invertible flows that depends on the parameter θ which itself is directly linked to the target distribution. Table A.2 displays the estimated NLL of all models provided by IS and the NEO method. It is interesting to note here again that NEO improves the training of the VAE when the dimension of the latent space is small to moderate.

Appendix B

Appendix of Chapter 4

B.1 Proofs

In this section we provide the proofs of Propositions 4.4.3, 4.5.1 and Theorems 4.4.4, 4.4.9, 4.4.10 and 4.5.2. The various Propositions and Lemmata used are stated and proved in Section B.1.9-B.1.10. Other intermediary technical results are provided in Section B.3. Equations, lemmata, propositions and theorems referred to without the prefix S are given in the main text.

B.1.1 Preliminaries

In order to simplify the notations, in what follows we write

$$\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) = \sum_{k_{0:t}^{1:2} \in [N]^{2(t+1)}} \overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^1, k_{0:t}^2) h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2}) \,,$$

where

$$\overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^1, k_{0:t}^2) := \prod_{s=0}^{t} N^{b_s} \left(\frac{N}{N-1}\right)^{1-b_s} \gamma_t^N(\mathbf{1})^2 \mathrm{I}_{b,t}(k_{0:t}^1, k_{0:t}^2) \Lambda_{1,t}^{\mathsf{BS}}(k_{0:t}^1) \Lambda_{2,t}^{\mathsf{BS}}(k_{0:t}^1; k_{0:t}^2) \,. \tag{B.1.1}$$

By (4.4.13),

$$\begin{aligned}
\Lambda_{1,t}^{\mathsf{BS}}(k_{0:t}^1) &= \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1) \Lambda_{1,t-1}^{\mathsf{BS}}(k_{0:t-1}^1) \,, \\
\Lambda_{2,t}^{\mathsf{BS}}(k_{0:t}^1; k_{0:t}^2) &= \{\beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2) \mathbb{1}_{k_t^2 \neq k_t^1} + \omega_{t-1}^{k_{t-1}^2} \mathbb{1}_{k_t^2 = k_t^1}\} \Lambda_{2,t-1}^{\mathsf{BS}}(k_{0:t-1}^1; k_{0:t-1}^2) \,.
\end{aligned}$$

and by (4.3.4) we have that $\gamma_t^N(\mathbf{1}) = \gamma_{t-1}^N(\mathbf{1}) N^{-1} \Omega_{t-1}$, hence, using (4.4.16) $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)$ becomes

$$\begin{aligned}
\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) &= \sum_{k_{0:t}^{1:2} \in [N]^{2(t+1)}} \overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^1, k_{0:t}^2) h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2}) \qquad \text{(B.1.2)} \\
&= \sum_{k_{0:t-1}^{1:2} \in [N]^{2t}} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1, k_{0:t-1}^2) \frac{\Omega_{t-1}^2}{N^2} \sum_{k_t^{1:2} \in [N]^2} \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1) \\
&\times \left[\frac{N}{N-1} \beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2) \mathbb{1}_{k_t^1 \neq k_t^2, b_t = 0} + N \omega_{t-1}^{k_{t-1}^2} \mathbb{1}_{k_t^1 = k_t^2, b_t = 1}\right] h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2}) \,.
\end{aligned}$$

B.1.2 Proof of Proposition 4.4.3

Proof of Lemma 4.4.1. By (**A4**), for any $(x, y) \in \mathsf{X}^2$,

$$\begin{aligned}\beta_t^N(x,y)\phi_{t-1}^N M_t(\mathrm{d}x) &= \beta_t^N(x,y)\sum_{i=1}^N \omega_{t-1}^i M_t(\xi_{t-1}^i, \mathrm{d}x) \\ &= \frac{g_{t-1}(y)m_t(y,x)}{\sum_{i=1}^N \widetilde{\omega}_{t-1}^i m_t(\xi_{t-1}^i, x)} \sum_{i=1}^N \omega_{t-1}^i m_t(\xi_{t-1}^i, x)\nu(\mathrm{d}x) \\ &= \frac{g_{t-1}(y)}{\Omega_{t-1}} M_t(y, \mathrm{d}x)\,.\end{aligned}$$

Consequently, for any $(k_{t-1}^1, k_t^1) \in \mathsf{X}^2$ and $h \in \mathsf{F}(\mathcal{X})$,

$$\begin{aligned}\mathbb{E}[\beta_t^N(\xi_t^{k_t}, \xi_{t-1}^{k_{t-1}})h(\xi_t^{k_t})|\mathcal{F}_{t-1}^N] &= \int \beta_t^N(x, \xi_{t-1}^{k_{t-1}})h(x)\phi_{t-1}^N M_t(\mathrm{d}x) \\ &= \int \frac{1}{\Omega_{t-1}} \widetilde{\omega}_{t-1}^{k_{t-1}^1} h(x) M_t(\xi_{t-1}^{k_{t-1}^1}, \mathrm{d}x) \\ &= \omega_{t-1}^{k_{t-1}} M_t[h](\xi_{t-1}^{k_{t-1}})\,.\end{aligned}$$

On the other hand,

$$\mathbb{E}[\mathbb{1}_{k_{t-1}=A_{t-1}^{k_t}} h(\xi_t^{k_t})|\mathcal{F}_{t-1}^N] = \int \sum_{i=1}^N \mathbb{1}_{k_{t-1}=i}\omega_{t-1}^i M_t(\xi_{t-1}^i, \mathrm{d}x)h(x) = \omega_{t-1}^{k_{t-1}} M_t[h](\xi_{t-1}^{k_{t-1}})\,.$$

□

Proof of Proposition 4.4.3. For the proof of i), note that conditionally on $\mathcal{F}_{t-1}^N$, $\xi_t^{1:N}$ are i.i.d. with distribution $\phi_{t-1}^N M_t$, hence,

$$\begin{aligned}&\mathbb{E}\left[\sum_{k_t^{1:2}\in[N]^2} \mathbb{1}_{k_t^1 \neq k_t^2} \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2)h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2}) \middle| \mathcal{F}_{t-1}^N\right] \\ &\quad = \sum_{k_t^{1:2}\in[N]^2} \int \mathbb{1}_{k_t^1 \neq k_t^2} \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2)h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2})\phi_{t-1}^N M_t(\mathrm{d}\xi_t^{k_t^1})\phi_{t-1}^N M_t(\mathrm{d}\xi_t^{k_t^2})\,,\end{aligned}$$

and by (4.4.11) in Lemma 4.4.1,

$$\begin{aligned}&\mathbb{E}\left[\sum_{k_t^{1:2}\in[N]^2} \mathbb{1}_{k_t^1 \neq k_t^2} \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2)h(\xi_{0:t}^{k_{0:t}^1}, \xi_{0:t}^{k_{0:t}^2}) \middle| \mathcal{F}_{t-1}^N\right] \\ &\qquad = \frac{N(N-1)}{\Omega_{t-1}^2}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])(\xi_{0:t-1}^{k_{0:t-1}^1}, \xi_{0:t-1}^{k_{0:t-1}^2})\,, \qquad \text{(B.1.3)}\end{aligned}$$

where

$$g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h] : (x_{0:t-1}, x'_{0:t-1}) \mapsto g_{t-1}^{\otimes 2}(x_{t-1}, x'_{t-1}) \int h(x_{0:t}, x'_{0:t}) M_t(x_{t-1}, \mathrm{d}x_t) M_t(x'_{t-1}, \mathrm{d}x'_t)\,.$$

On the other hand, by (4.4.11) in Lemma 4.4.1,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k_t^{1:2}\in[N]^2} \mathbb{1}_{k_t^1=k_t^2}\beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\omega_{t-1}^{k_{t-1}^2}h(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})\middle|\mathcal{F}_{t-1}^N\right] \\
&\quad= \sum_{k_t^{1:2}\in[N]^2}\int \mathbb{1}_{k_t^1=k_t^2}\beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\omega_{t-1}^{k_{t-1}^2}h(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})\phi_{t-1}^N M_t(\mathrm{d}\xi_t^{k_t^1})\delta_{\xi_t^{k_t^1}}(\mathrm{d}\xi_t^{k_t^2}) \\
&\quad= \frac{1}{\Omega_{t-1}^2}\sum_{k_t^{1:2}\in[N]^2}\mathbb{1}_{k_t^1=k_t^2}\widetilde{\omega}_{t-1}^{k_{t-1}^1}\widetilde{\omega}_{t-1}^{k_{t-1}^2}\int h(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})M_t(\xi_{t-1}^{k_{t-1}^1},\mathrm{d}\xi_t^{k_t^1})\delta_{\xi_t^{k_t^1}}(\mathrm{d}\xi_t^{k_t^2}) \\
&\quad= \frac{N}{\Omega_{t-1}^2}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})\,, \qquad \text{(B.1.4)}
\end{aligned}$$

where

$$g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h] : (x_{0:t-1},x'_{0:t-1}) \mapsto g_{t-1}^{\otimes 2}(x_{t-1},x'_{t-1})\int h(x_{0:t},x'_{0:t})M_t(x_{t-1},\mathrm{d}x_t)\delta_{x_t}(\mathrm{d}x'_t)\,.$$

If $b_t = 0$, since Ω_{t-1} and $\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)$ are $\mathcal{F}_{t-1}^N$ measurable for any $(k_{0:t-1}^1,k_{0:t-1}^2)\in[N]^{2t}$, by (B.1.2) and (B.1.3),

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)|\mathcal{F}_{t-1}^N\right] &= \sum_{k_{0:t-1}^{1:2}\in[N]^2}\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)\frac{\Omega_{t-1}^2}{N(N-1)} \\
&\qquad\times\mathbb{E}\left[\sum_{k_t^{1:2}\in[N]^2}\mathbb{1}_{k_t^1\neq k_t^2}\beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2,k_{t-1}^2)h(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})\middle|\mathcal{F}_{t-1}^N\right] \\
&= \sum_{k_{0:t-1}^{1:2}\in[N]^2}\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])(\xi_{0:t-1}^{k_{0:t-1}^1},\xi_{0:t-1}^{k_{0:t-1}^2}) \\
&= \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])\,.
\end{aligned}$$

If $b_t = 1$, again by (B.1.2) and (B.1.4),

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)|\mathcal{F}_{t-1}^N\right] &= \sum_{k_{0:t-1}^{1:2}\in[N]^2}\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)\frac{\Omega_{t-1}^2}{N} \\
&\qquad\times\mathbb{E}\left[\sum_{k_t^{1:2}\in[N]^2}\mathbb{1}_{k_t^1=k_t^2}\beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\omega_{t-1}^{k_{t-1}^2}h(\xi_{0:t}^{k_{0:t}^1},\xi_{0:t}^{k_{0:t}^2})\middle|\mathcal{F}_{t-1}^N\right] \\
&= \sum_{k_{0:t-1}^{1:2}\in[N]^2}\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])(\xi_{0:t-1}^{k_{0:t-1}^1},\xi_{0:t-1}^{k_{0:t-1}^2}) \\
&= \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])\,.
\end{aligned}$$

For the proof of ii), we proceed by induction. Let $t = 0$ and $h \in \mathsf{F}(\mathcal{X}^{\otimes 2})$. If $b_0 = 0$, since $\xi_0^{1:N} \overset{\text{iid}}{\sim} M_0$,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Q}_{0,0}^{N,\mathsf{BS}}(h)\right] &= \mathbb{E}\left[\frac{1}{N(N-1)}\sum_{i,j\in[N]^2}\mathbb{1}_{i\neq j}h(\xi_0^i,\xi_0^j)\right] \\
&= \frac{1}{N(N-1)}\sum_{i,j\in[N]^2}\mathbb{1}_{i\neq j}\mathcal{M}_0^0[h] = \mathcal{M}_0^0[h] = \mathcal{Q}_{b,0}(h)\,.
\end{aligned}$$

If $b_0 = 1$,

$$\mathbb{E}\left[\mathcal{Q}_{1,0}^{N,\mathsf{BS}}(h)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} h(\xi_0^i, \xi_0^i)\right] = \mathcal{M}_0^1[h] = \mathcal{Q}_{b,0}(h)\,.$$

Let $t \in \mathbb{N}_{>0}$ and $h \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$. Assume that $\mathbb{E}[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f)] = \mathcal{Q}_{b,t-1}(f)$ for any $b \in \mathcal{B}_t$ and $f \in \mathsf{F}(\mathcal{X}^{\otimes 2t})$. By (i) in Proposition 4.4.3 and the tower property

$$\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\right] = \mathbb{E}\left[\mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)|\mathcal{F}_{t-1}^N]\right] = \mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\right]$$

and $g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h] \in \mathsf{F}(\mathcal{X}^{\otimes 2t})$. Thus, by the induction hypothesis and the definition of $\mathcal{Q}_{b,t-1}$ we get

$$\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\right] = \mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\right] = \mathcal{Q}_{b,t-1}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h]) = \mathcal{Q}_{b,t}(h)\,,$$

which completes the proof. The proof of iii) is a direct consequence of (4.4.4), (4.4.17) and ii). □

B.1.3 Proof of Proposition 4.5.1

Let $h_{0:t}$ be an additive functional (4.5.5). By definition, for $s \in [t]$,

$$\begin{aligned}\mathbf{G}_{s,t}(x_s, h_{0:t}) &= \int \{\tilde{h}_{0:s}(x_{0:s}) + \tilde{h}_{s:t}(x_{s:t})\}\mathbf{T}_s(x_s, \mathrm{d}x_{0:s-1})\mathbf{Q}_{s+1:t}(x_s, \mathrm{d}x_{s+1:t}) \\ &= \int \{\mathbf{T}_s[\tilde{h}_{0:s}](x_s) + \tilde{h}_{s:t}(x_{s:t})\}\mathbf{Q}_{s+1:t}(x_s, \mathrm{d}x_{s+1:t})\,,\end{aligned}$$

and then, setting $\mathsf{H}_{s:t} : x_{s:t} \mapsto \mathbf{T}_s[\tilde{h}_{0:s}](x_s) + \tilde{h}_{s:t}(x_{s:t})$ we get

$$\begin{aligned}&\gamma_s(\mathbf{1})\gamma_s(\mathbf{G}_{s,t}[h_{0:t}]^2) \\ &= \gamma_s(\mathbf{1})\gamma_s(\mathbf{Q}_{s+1:t}[\mathbf{T}_s[\tilde{h}_{0:s}] + \tilde{h}_{s:t}]^2) \\ &= \int \gamma_{0:s-1}(\mathrm{d}x'_{0:s-1})g_{s-1}(x'_{s-1}) \int \mathbf{Q}_{s+1:t}[\mathsf{H}_{s:t}](x_s)\mathbf{Q}_{s+1:t}[\mathsf{H}_{s:t}](x_s)\gamma_{0:s}(\mathrm{d}x_{0:s}) \\ &= \int \gamma_{0:s-1}(\mathrm{d}x'_{0:s-1})g_{s-1}(x'_{s-1})\gamma_{0:s}(\mathrm{d}x_{0:s})\delta_{x_s}(\mathrm{d}x'_s)\mathbf{Q}_{s+1:t}[\mathsf{H}_{s:t}](x_s)\mathbf{Q}_{s+1:t}[\mathsf{H}_{s:t}](x'_s)\,,\end{aligned}$$

which establishes the result since by definition

$$\mathcal{Q}_{e_s,t}(\mathrm{d}x_{0:t}, \mathrm{d}x'_{0:t}) = \gamma_{0:s}(\mathrm{d}x_{0:s})\gamma_{0:s-1}(\mathrm{d}x'_{0:s-1})g_{s-1}(x'_{s-1})\delta_{x_s}(\mathrm{d}x'_s) \\ \mathbf{Q}_{s+1:t}(x_s, \mathrm{d}x_{s+1:t})\mathbf{Q}_{s+1:t}(x'_s, \mathrm{d}x'_{s+1:t})\,.$$

If $s = 0$, then $\mathbf{G}_{0,t}[h_{0:t}](x_0) = \int h_{0:t}(x_{0:t})\mathbf{Q}_{1:t}(x_0, \mathrm{d}x_{1:t})$ and

$$\begin{aligned}\gamma_0(\mathbf{1})\gamma_0(\mathbf{G}_{0:t}[h_{0:t}]^2) = \gamma_0(\mathbf{G}_{0:t}[h_{0:t}]^2) &= \int M_0(\mathrm{d}x_0)\mathbf{Q}_{1:t}[h_{0:t}](x_0)\mathbf{Q}_{1:t}[h_{0:t}](x_0) \\ &= \int M_0(\mathrm{d}x_0)\delta_{x_0}(\mathrm{d}x'_0)\mathbf{Q}_{1:t}[h_{0:t}](x_0)\mathbf{Q}_{1:t}[h_{0:t}](x'_0) \\ &= \mathcal{Q}_{e_0,t}(h_{0:t}^{\otimes 2})\,.\end{aligned}$$

B.1.4 Proof of Theorem 4.4.4

Let $m \in \mathbb{N}_{>0}$ and $N \geq 2$. Define

$$\mathcal{I}_0^m := \{k^{1:2m} \in [N]^{2m} : k^{2i-1} \neq k^{2i}, i \in [1:m]\}\,, \tag{B.1.5}$$
$$\mathcal{I}_1^m := \{k^{1:2m} \in [N]^{2m} : k^{2i-1} = k^{2i}, i \in [1:m]\}\,. \tag{B.1.6}$$

Define also for any $p \in [2m]$,

$$\mathcal{S}_m^p := \{k^{1:2m} \in [N]^{2m} : \mathrm{Card}(\{k^1, k^2, \cdots, k^{2m-1}, k^{2m}\}) = p\}\,.$$

Then $[N]^{2m} = \bigsqcup_{p=1}^{2m} \mathcal{S}_m^p$ and

$$\mathcal{I}_0^m = \bigsqcup_{p=1}^{2m} \mathcal{I}_0^m \cap \mathcal{S}_m^p = \bigsqcup_{p=2}^{2m} \mathcal{I}_0^m \cap \mathcal{S}_m^p\,, \quad \mathcal{I}_1^m = \bigsqcup_{p=1}^{2m} \mathcal{I}_1^m \cap \mathcal{S}_m^p = \bigsqcup_{p=1}^{m} \mathcal{I}_1^m \cap \mathcal{S}_m^p\,, \tag{B.1.7}$$

where $\bigsqcup$ means disjoint union. The first equality holds because the tuples in $\mathcal{I}_0^m$ must contain at least two different values and the second because tuples in $\mathcal{I}_1^m$ contain at most m different values. The proof of Theorem 4.4.4 is concerned with $m = 2$ and that of Proposition B.1.4 with $m \geq 2$. Example B.1.1 provides the intersections for the case $m = 2$.

Example B.1.1. *Choose $m = 2$ and $N \geq 4$. Then,*

$$\begin{aligned}
\mathcal{I}_0^2 \cap \mathcal{S}_2^2 &= \{k^{1:4} \in [N]^4 : k^1 \neq k^2, k^3 \neq k^4, \{k^3, k^4\} = \{k^1, k^2\}\}\,,\\
\mathcal{I}_0^2 \cap \mathcal{S}_2^3 &= \{k^{1:4} \in [N]^4 : k^1 \neq k^2, k^3 \neq k^4, k^3 \in \{k^1, k^2\}, k^4 \notin \{k^1, k^2\}\}\,,\\
&\qquad \sqcup \{k^{1:4} \in [N]^4 : k^1 \neq k^2, k^3 \neq k^4, k^3 \notin \{k^1, k^2\}, k^4 \in \{k^1, k^2\}\}\,,\\
\mathcal{I}_0^2 \cap \mathcal{S}_2^4 &= \{k^{1:4} \in [N]^4 : k^1 \neq k^2 \neq k^3 \neq k^4\}\,,
\end{aligned}$$

with $\mathrm{Card}(\mathcal{I}_0^2 \cap \mathcal{S}_2^2) = 2N(N-1)$, $\mathrm{Card}(\mathcal{I}_0^2 \cap \mathcal{S}_2^3) = 4N(N-1)(N-2)$ *and* $\mathrm{Card}(\mathcal{I}_0^2 \cap \mathcal{S}_2^4) = N(N-1)(N-2)(N-3)$. *As for* $\mathcal{I}_1^2$,

$$\begin{aligned}
\mathcal{I}_1^2 \cap \mathcal{S}_2^1 &= \{k^{1:4} \in [N]^4 : k^1 = k^2 = k^3 = k^4\}\,,\\
\mathcal{I}_1^2 \cap \mathcal{S}_2^2 &= \{k^{1:4} \in [N]^4 : k^1 = k^2, k^3 = k^4, k^1 \neq k^3\}\,.
\end{aligned}$$

with $\mathrm{Card}(\mathcal{I}_1^2 \cap \mathcal{S}_2^1) = N$ *and* $\mathrm{Card}(\mathcal{I}_1^2 \cap \mathcal{S}_2^2) = N(N-1)$.

Proof of Theorem 4.4.4. We proceed by induction. Throughout the proof we assume that $N \geq 4$ for the sake of simplicity. For $t = 0$ and $b = \mathbf{0}$, using (4.4.23) and (4.4.20),

$$\mathcal{Q}_{\mathbf{0},0}^{N,\mathsf{BS}}(h) = N^{-1}(N-1)^{-1} \sum_{i,j \in [N]^2} \mathbb{1}_{i \neq j} h(\xi_0^i, \xi_0^j)\,,$$

and $\mathcal{Q}_{\mathbf{0},0}(h) = M_0^{\otimes 2}(h)$.

$$\begin{aligned}
&\left\|\mathcal{Q}_{\mathbf{0},0}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{\mathbf{0},0}(h)\right\|_2^2\\
&\quad = \mathbb{E}\left[\frac{1}{N^2(N-1)^2} \sum_{i,j,i',j' \in [N]^4} \mathbb{1}_{i \neq j, i' \neq j'} \{h(\xi_0^i, \xi_0^j) - \mathcal{Q}_{\mathbf{0},0}(h)\}\{h(\xi_0^{i'}, \xi_0^{j'}) - \mathcal{Q}_{\mathbf{0},0}(h)\}\right]\\
&\quad = \frac{\tau_0 + \overline{\tau}_0}{N^2(N-1)^2}\,,
\end{aligned}$$

where

$$\tau_0 = \mathbb{E}\left[\sum_{i,j,i',j' \in \mathcal{I}_0^2 \cap \mathcal{S}_2^4} \{h(\xi_0^i, \xi_0^j) - \mathcal{Q}_{\mathbf{0},0}(h)\}\{h(\xi_0^{i'}, \xi_0^{j'}) - \mathcal{Q}_{\mathbf{0},0}(h)\}\right],$$

$$\overline{\tau}_0 = \mathbb{E}\left[\sum_{i,j,i',j' \in \mathcal{I}_0^2 \cap \overline{\mathcal{S}_2^4}} \{h(\xi_0^i, \xi_0^j) - \mathcal{Q}_{\mathbf{0},0}(h)\}\{h(\xi_0^{i'}, \xi_0^{j'}) - \mathcal{Q}_{\mathbf{0},0}(h)\}\right],$$

where $\mathcal{I}_0^2 \cap \mathcal{S}_2^4$ is defined in (B.1.5), (B.1.7) and explicited in Example B.1.1, and $\mathcal{I}_0^2 \cap \overline{\mathcal{S}_2^4} = (\mathcal{I}_0^2 \cap \mathcal{S}_2^2) \sqcup (\mathcal{I}_0^2 \cap \mathcal{S}_2^3)$. If $(i,j,i',j') \in \mathcal{I}_0^2 \cap \mathcal{S}_2^4$, then $\xi_0^i, \xi_0^j, \xi_0^{i'}$ and $\xi_0^{j'}$ are i.i.d. Therefore,

$$\mathbb{E}\left[\{h(\xi_0^i, \xi_0^j) - \mathcal{Q}_{\mathbf{0},0}(h)\}\{h(\xi_0^{i'}, \xi_0^{j'}) - \mathcal{Q}_{\mathbf{0},0}(h)\}\right] = \mathbb{E}\left[\{h(\xi_0^i, \xi_0^j) - \mathcal{Q}_{\mathbf{0},0}(h)\}\right]^2 = 0\,,$$

and $\tau_0 = 0$. Hence, using the fact that because h is bounded, $\|h - \mathcal{Q}_{\mathbf{0},0}(h)\|_\infty \leq 2\|h\|_\infty$ and that $\mathrm{Card}(\mathcal{I}_0^2 \cap \overline{\mathcal{S}_2^4}) = 4N(N-1)(N-2) + 2N(N-1)$ by Example B.1.1, we get

$$\begin{aligned}
\left\|\mathcal{Q}_{\mathbf{0},0}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{\mathbf{0},0}(h)\right\|_2^2 &\leq \frac{\overline{\tau}_0}{N^2(N-1)^2} \\
&\leq \frac{4\|h\|_\infty^2}{N^2(N-1)^2} \sum_{i,j,i',j' \in \mathcal{I}_0^2 \cap \overline{\mathcal{S}_2^4}} 1\,, \\
&= \frac{4\|h\|_\infty^2 [4N(N-1)(N-2) + 2N(N-1)]}{N^2(N-1)^2} = \mathcal{O}(N^{-1})\,,
\end{aligned}$$

which completes the proof for $b = 0$. For $b = 1$,

$$\mathcal{Q}_{1,0}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{1,0}(h) = N^{-1} \sum_{i=1}^N h(\xi_0^i, \xi_0^i) - \int h(x,x) M_0(\mathrm{d}x)\,,$$

and since h is bounded,

$$\left\|\mathcal{Q}_{1,0}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{1,0}(h)\right\|_2^2 = N^{-1} \mathbb{V}_{M_0}[h(\xi,\xi)] = \mathcal{O}(N^{-1})\,,$$

where $\mathbb{V}_{M_0}$ is the variance under M_0. This completes the proof for $t = 0$. Let $t > 0$, and assume now that (4.4.18) holds at time $t-1$. Consider the following decomposition

$$\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t}(h) = \mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) | \mathcal{F}_{t-1}^N] + \mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) | \mathcal{F}_{t-1}^N] - \mathcal{Q}_{b,t}(h)\,,$$

which, by Proposition 4.4.3, becomes

$$\begin{aligned}
\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t}(h) = \mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h]) \\
+ \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h]) - \mathcal{Q}_{b,t-1}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h])\,. \quad \text{(B.1.8)}
\end{aligned}$$

By the induction hypothesis, since h is bounded and also g_{t-1} by (**A5**), we have

$$\lim_{N\to\infty} \left\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h]) - \mathcal{Q}_{b,t-1}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h])\right\|_2 = 0\,,$$

hence, by Minkowski's inequality it remains to prove that

$$\lim_{N\to\infty} \left\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2} \mathcal{M}_t^{b_t}[h])\right\|_2 = 0\,. \quad \text{(B.1.9)}$$

By Proposition 4.4.3, $\mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)|\mathcal{F}_{t-1}^N] = \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])$ and

$$\mathbb{E}[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])] = \mathbb{E}[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])^2]\,,$$

hence,

$$\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 = \|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\|_2^2 - \|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2\,.$$

Consequently, by Proposition B.1.8, if $b_t = 0$,

$$\begin{aligned}
&\left\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{0}[h])\right\|_2^2 \\
&\leq \left(\frac{(N-2)(N-3)}{N(N-1)} - 1\right)\left\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{0}[h])\right\|_2^2 \\
&+ \frac{N-2}{N-1}G_\infty^3|h|_\infty^2\int \nu(\mathrm{d}x)\mathbb{E}\Bigg[\frac{\Omega_{t-1}}{N}\Bigg\{\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\beta_t^N(x,.)\otimes\mathbf{1}) \\
&+ \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes m_t(.,x))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\beta_t^N(x,.)\otimes\mathbf{1}) + \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1}) \\
&\qquad\times \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\beta_t^N(x,.)) + \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes m_t(.,x))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\beta_t^N(x,.))\Bigg\}\Bigg] \\
&+ G_\infty^2|h|_\infty^2\int \nu^{\otimes 2}(\mathrm{d}y,\mathrm{d}x)\mathbb{E}\Bigg[\frac{\Omega_{t-1}^2}{N(N-1)} \\
&\qquad\times\Bigg\{\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(x,.)\otimes\beta_t^N(y,.)\right) \\
&\qquad\quad+ \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(y,.)\otimes\beta_t^N(x,.)\right)\Bigg\}\Bigg]\,,
\end{aligned} \tag{B.1.10}$$

and if $b_t = 1$,

$$\begin{aligned}
\left\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{1}[h])\right\|_2^2 &\leq \left(\frac{N-1}{N} - 1\right)\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{1}[h])\|_2^2 \\
&+ G_\infty^3|h|_\infty^2\int \mathbb{E}\left[\frac{\Omega_{t-1}}{N}\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\beta_t^N(x,.)\otimes\mathbf{1})\right]\nu(\mathrm{d}x)\,.
\end{aligned} \tag{B.1.11}$$

By Proposition B.1.4, the first term in the r.h.s. of (B.1.10) and (B.1.11) is $\mathcal{O}(N^{-1})$ in both cases because $|g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h]|_\infty \leq G_\infty^2|h|_\infty < \infty$. We now show that the remaining terms go to zero when N goes to infinity. Define for any $x \in \mathsf{X}$ and $N \in \mathbb{N}_{>0}$,

$$\begin{aligned}
B_N(x) &:= \frac{\Omega_{t-1}}{N}\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\beta_t^N(x,.))\,, \\
\widetilde{B}_N(x) &:= \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\mathbf{1}\otimes\mathbf{1}\right)\,, \\
\widetilde{B}(x) &:= \mathcal{Q}_{b,t-1}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})\,.
\end{aligned} \tag{B.1.12}$$

We apply Theorem B.3.1 with $f_N = \mathbb{E}[B_N]$, $g_N = \mathbb{E}[\widetilde{B}_N]$, $g = \widetilde{B}$ and $f = 0$. To establish i), note that $\mathbb{E}[B_N(x)] \leq G_\infty\mathbb{E}[\widetilde{B}_N(x)]$ for all $N \in \mathbb{N}_{>0}$ and $x \in \mathsf{X}$, since for all $(x,i) \in \mathsf{X}\times[N]$, $\beta_t^N(x,\xi_{t-1}^i) \leq 1$ and $N^{-1}\Omega_{t-1} \leq G_\infty$. Then, to prove ii), for all $(h,f) \in \mathsf{F}(\mathcal{X}^{\otimes t})^2$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(h)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f) - \mathcal{Q}_{b,t-1}(h)\mathcal{Q}_{b,t-1}(f)\right]\right| \\
&\quad\leq \mathbb{E}\left[\left|\left(\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}(h)\right)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f)\right|\right] + \mathbb{E}\left[\left|\left(\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f) - \mathcal{Q}_{b,t-1}(f)\right)\mathcal{Q}_{b,t-1}(h)\right|\right] \\
&\quad\leq \|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}(h)\|_2\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f)\|_2 + \|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(f) - \mathcal{Q}_{b,t-1}(f)\|_2|\mathcal{Q}_{b,t-1}(h)|\,,
\end{aligned}$$

which goes to zero by the induction hypothesis, the fact that $\sup_{N\in\mathbb{N}} \left\|\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(f)\right\|_2 < \infty$ by Proposition B.1.4 and $|\mathcal{Q}_{b,t-1}(h)| < \infty$. Hence, for all $x \in \mathsf{X}$,

$$\lim_{N\to\infty} g_N(x) = g(x) \quad \text{and} \quad \lim_{N\to\infty} \mathbb{E}\left[\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2\right] = \mathcal{Q}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2 .$$

Added to the fact that $\int \widetilde{B}_N(x)\nu(\mathrm{d}x) = \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2$ and $\int \widetilde{B}(x)\nu(\mathrm{d}x) = \mathcal{Q}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2$, we get by applying Fubini's theorem

$$\lim_{N\to\infty} \int \mathbb{E}\left[\widetilde{B}_N(x)\right]\nu(\mathrm{d}x) = \lim_{N\to\infty} \mathbb{E}\left[\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2\right] = \mathcal{Q}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^2 = \int \widetilde{B}(x)\nu(\mathrm{d}x) .$$

Then, for iii), first we have that $\mathbb{E}[B_N(x)^{3/2}] \le G_\infty^{3/2}\mathbb{E}[\widetilde{B}_N(x)^{3/2}]$ and

$$\sup_{N\in\mathbb{N}} \mathbb{E}[\widetilde{B}_N(x)^{3/2}] \le \sigma_+^{3/2} \sup_{N\in\mathbb{N}} \mathbb{E}[\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^3] ,$$

where the r.h.s. is finite by choosing $m = 3$ in Proposition B.1.4. The family of non negative random variables $\{B_N(x)\}_{N\in\mathbb{N}}$ is then uniformly integrable for any $x \in \mathsf{X}$. Indeed, for any $x \in \mathsf{X}$, $\alpha \in \mathbb{R}_+^*$ and $N \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[B_N(x)\mathbb{1}_{B_N(x)\ge\alpha}\right] &\le \mathbb{E}[B_N(x)^{3/2}]/\sqrt{\alpha} \\ &\le \sigma_+^{3/2} G_\infty^{3/2} \sup_{N\in\mathbb{N}} \mathbb{E}[\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})^3]/\sqrt{\alpha} ,\end{aligned}$$

hence $\lim_{\alpha\to\infty} \sup_{N\in\mathbb{N}} \mathbb{E}\left[B_N(x)\mathbb{1}_{B_N(x)\ge\alpha}\right] = 0$. On the other hand,

$$B_N(x) = \frac{\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(\mathbf{1}\otimes g_{t-1}m_t(.,x)\right)}{N\phi^N_{t-1}\left(m_t(.,x)\right)} ,$$

and the induction hypothesis coupled with the fact that $\phi^N_{t-1}(m_t(.,x)) \xrightarrow{\mathbb{P}} \phi_{t-1}(m_t(.,x))$ with $\phi_{t-1}(m_t(.,x)) > 0$ by $(\mathbf{A5}:\mathbf{6})$ gives

$$\begin{aligned}&\frac{\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(\mathbf{1}\otimes g_{t-1}m_t(.,x)\right)}{\phi^N_{t-1}(m_t(.,x))} \\ &\qquad\qquad \xrightarrow{\mathbb{P}} \frac{\mathcal{Q}_{b,t-1}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}_{b,t-1}\left(\mathbf{1}\otimes g_{t-1}m_t(.,x)\right)}{\phi_{t-1}(m_t(.,x))} .\end{aligned}$$

Hence, $B_N(x) \xrightarrow{\mathbb{P}} 0$, and by uniform integrability, for any $x \in \mathsf{X}$

$$\lim_{N\to\infty} f_N(x) = \lim_{N\to\infty} \mathbb{E}[B_N(x)] = 0 .$$

Finally, by Theorem B.3.1 we deduce that

$$\lim_{N\to\infty} \int \mathbb{E}[B_N(x)]\nu(\mathrm{d}x) = \int \lim_{N\to\infty} \mathbb{E}[B_N(x)]\nu(\mathrm{d}x) = 0 . \tag{B.1.13}$$

The other similar terms are treated in the same way by adapting the definitions in (B.1.12). As for the second integral, define for any $(x,y) \in \mathsf{X}^2$,

$$R_N(x,y) := \frac{\Omega^2_{t-1}}{N(N-1)} \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes\beta^N_t(y,.)) .$$

Then, using that

$$\int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(x,.)\otimes\beta_t^N(y,.)\right)\nu(\mathrm{d}y) \\ \leq \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(x,.)\otimes\mathbf{1}\right),$$

together with Fubini's theorem we obtain, using $N \geq 4$,

$$0 \leq \int \mathbb{E}[R_N(x,y)]\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y) \leq \frac{4G_\infty}{3}\int \mathbb{E}[B_N(x)]\nu(\mathrm{d}x)\,,$$

and by (B.1.13) we get that

$$\lim_{N\to\infty}\int \mathbb{E}[R_N(x,y)]\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y) = 0\,.$$

The remaining term goes to zero by a similar reasoning. This completes the proof of (4.4.18). For the convergence rate, by the strong mixing assumption we have that

$$\beta_t^N(x,y) \leq \frac{G_\infty\sigma_+}{\sigma_-\Omega_{t-1}} \quad \forall (x,y)\in\mathsf{X}^2\,, \tag{B.1.14}$$

and in the case $b_t = 0$, we have for example that

$$\int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\beta_t^N(x,.)\otimes\mathbf{1})\nu(\mathrm{d}x) \\ \leq \frac{G_\infty\sigma_+}{\sigma_-\Omega_{t-1}}\int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\mathbf{1})\nu(\mathrm{d}x) \\ \leq \frac{G_\infty\sigma_+}{\sigma_-\Omega_{t-1}}\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\mathbf{1})^2\,.$$

and

$$\int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\beta_t^N(x,.)\otimes\beta_t^N(y,.))\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y) \\ \leq \frac{G_\infty^2\sigma_+^2}{\sigma_-^2\Omega_{t-1}^2}\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1}\otimes\mathbf{1})^2\,.$$

Thus, replacing in (i), we get

$$\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])\|_2^2 \\ \leq \left[\frac{(N-2)(N-3)}{N(N-1)} - 1\right]\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 \\ + \frac{2\sigma_+G_\infty^4|h|_\infty^2}{\sigma_-}\left[\frac{2(N-2)}{N(N-1)} + \frac{\sigma_+}{\sigma_-N(N-1)}\right]\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})\|_2^2\,.$$

The case $b_t = 1$ is handled similarly using (ii) which yields

$$\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) - \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])\|_2^2 \\ \leq \left[\frac{N-1}{N} - 1\right]\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])\|_2^2 + \frac{\sigma_+G_\infty^4|h|_\infty^2}{N}\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})\|_2^2\,.$$

Both upper bounds are $\mathcal{O}(N^{-1})$ by Proposition B.1.4. This concludes the proof. □

B.1.5 Proof of Theorem 4.4.7

The proof is a straightforward adaptation of the proof in (Lee and Whiteley, 2018, Theorem 1). Let $h \in \mathsf{F}(\mathcal{X})$. By (4.4.30),

$$\begin{aligned}\mathcal{V}_{\gamma,t}^{N,\mathsf{BS}}(h) &= \sum_{s=0}^{t} \left(\frac{N-1}{N}\right)^t \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(h_t^{\otimes 2}) + N\left[\left(\frac{N-1}{N}\right)^{t+1} - 1\right] \mathcal{Q}_{\mathbf{0},t}^{N,\mathsf{BS}}(h_t^{\otimes 2}) \\ &\quad + \sum_{b \in \mathcal{B}_t \setminus \{\mathbf{0}, e_{0:t}\}} N\left\{\prod_{s=0}^{t} \frac{1}{N^{b_s}} \left(\frac{N-1}{N}\right)^{1-b_s}\right\} \mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h_t^{\otimes 2}) \\ &\xrightarrow{\mathbb{P}} \sum_{s=0}^{t} \left\{\mathcal{Q}_{b,t}(h_t^{\otimes 2}) - \mathcal{Q}_{\mathbf{0},t}(h_t^{\otimes 2})\right\} = \mathcal{V}_{\gamma,t}^{\infty}(h)\,,\end{aligned}$$

where we have used that for any $b \in \mathcal{B}_t$, $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h_t^{\otimes 2}) \xrightarrow{\mathbb{P}} \mathcal{Q}_{b,t}(h_t^{\otimes 2})$ by Theorem 4.4.4.

B.1.6 Proof of Theorem 4.4.9

Define for any $t \in [N]$ and $b \in \mathcal{B}_t$,

$$\mathcal{G}_t^N := \sigma(\mathcal{G}_{t-1}^N \cup \sigma(\{J_{k,t-1}^i\}_{(i,k)\in[N]^2}) \cup \sigma(\{A_{t-1}^i, \xi_t^i\}_{i=1}^N))\,, \tag{B.1.15}$$

with $\mathcal{G}_0^N = \mathcal{F}_0^N$. In the following, we write

$$\mathcal{C}_{N,b,t} := \left\{\prod_{s=0}^{t} N^{b_s} \left(\frac{N}{N-1}\right)^{1-b_s}\right\} \gamma_t^N(\mathbf{1})^2/N^2\,. \tag{B.1.16}$$

The intermediary results used in the next proof are given in Section B.1.10.

Proof of Theorem 4.4.9. Let $h \in \mathsf{F}(\mathcal{X}^{\otimes 2})$. We proceed again by induction. The case $t = 0$ is a consequence of Theorem 4.4.4 since $\widetilde{\mathcal{Q}}_{b,0}^{N,M}(h) = \mathcal{Q}_{b,0}^{N,\mathsf{BS}}(h)$ for any $b \in \mathcal{B}_0$. Let $t > 0$. Similarly to Theorem 4.4.4 we make use of the following decomposition:

$$\begin{aligned}\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \mathcal{Q}_{b,t}(h) &= \widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h]) \\ &\quad + \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h]) - \mathcal{Q}_{b,t-1}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\,.\end{aligned}$$

By Minkowski's inequality and the induction hypothesis, it remains to prove that

$$\lim_{N\to\infty} \|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2 = 0\,. \tag{B.1.17}$$

By Lemma B.1.9, $\mathbb{E}[\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)|\mathcal{G}_{t-1}^N] = \widetilde{\mathcal{Q}}_{b,t}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])$ and

$$\mathbb{E}[\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])] = \mathbb{E}[\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])^2]\,,$$

hence,

$$\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \widetilde{\mathcal{Q}}_{b,t}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 = \|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)\|_2^2 - \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2\,.$$

By Proposition B.1.10, if $b_t = 0$,

$$\begin{aligned}\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)\|_2^2 &= \sum_{p=2}^{4} \mathbb{E}\left[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4)\right] \\ &\leq \frac{(N-2)(N-3)}{N(N-1)} \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])\|_2^2 \\ &\quad + |h|_\infty^2 \sum_{p=2}^{3} \mathbb{E}\left[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4)\right],\end{aligned}$$

and

$$\begin{aligned}
&\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \widetilde{\mathcal{Q}}_{b,t}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 \\
&\quad \le \left(\frac{(N-2)(N-3)}{N(N-1)} - 1\right) \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 \\
&\qquad\qquad + |h|_\infty^2 \sum_{p=2}^{3} \mathbb{E}\Bigg[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Bigg] .
\end{aligned}$$

By Proposition B.1.12, (**A5**) and the fact that h is bounded, $\sup_{N\in\mathbb{N}} \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 < \infty$ and

$$\lim_{N\to\infty} \left(\frac{(N-2)(N-3)}{N(N-1)} - 1\right) \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2 = 0 ,$$

and by (i) in Proposition B.1.11, the second term in the r.h.s. also goes to zero, which shows (B.1.17) when $b_t = 0$. If $b_t = 1$,

$$\begin{aligned}
\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\|_2^2 &= \sum_{p=1}^{2} \mathbb{E}\Bigg[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Bigg] \\
&\le \frac{N-1}{N}\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{1}[h])\|_2^2 + |h|_\infty^2 \mathbb{E}\Bigg[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^1} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Bigg] ,
\end{aligned}$$

and $\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h) - \widetilde{\mathcal{Q}}_{b,t}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])\|_2^2$ goes to zero similarly to the case $b_t = 0$ and by application of Proposition B.1.11.

The convergence rate follows straightforwardly by Proposition B.1.11 since for $p \in \{2,3\}$

$$\mathbb{E}\Bigg[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Bigg] = \mathcal{O}(N^{-1}) ,$$

and

$$\mathbb{E}\Bigg[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^1} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Bigg] = \mathcal{O}(N^{-1}) .$$

□

B.1.7 Proof of Theorem 4.4.10

The proof boils down to showing a *PaRIS* version of the identity (4.4.30). Let us first prove that for all $t \in \mathbb{N}$ and $(k_t^1, k_t^2) \in [N]^2$,

$$\sum_{b\in\mathcal{B}_t} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2) = 1 . \tag{B.1.18}$$

We proceed by induction. If $t = 0$,

$$\sum_{b\in\mathcal{B}_0} \widetilde{\mathcal{T}}_t^b(k_0^1,k_0^2) = \mathbb{1}_{k_0^1\neq k_0^2} + \mathbb{1}_{k_0^1=k_0^2} = 1 .$$

Let $t > 0$ and assume that (B.1.18) holds at $t-1$ for all $(k_{t-1}^1, k_{t-1}^2) \in [N]^2$. By the induction hypothesis,

$$
\begin{aligned}
&\sum_{b \in \mathcal{B}_t} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \\
&\quad = \sum_{b \in \mathcal{B}_{t-1}} M^{-1} \Big\{ \mathbb{1}_{k_t^1 \neq k_t^2} \sum_{i=1}^M \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^1,t-1}^i, J_{k_t^2,t-1}^i) + \mathbb{1}_{k_t^1 = k_t^2} \sum_{i=1}^M \sum_{n=1}^N \omega_{t-1}^n \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^1,t-1}^i, n) \Big\} \\
&\quad = M^{-1} \sum_{i=1}^M \{ \mathbb{1}_{k_t^1 \neq k_t^2} \sum_{b \in \mathcal{B}_{t-1}} \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^1,t-1}^i, J_{k_t^2,t-1}^i) + \mathbb{1}_{k_t^1 = k_t^2} \sum_{n=1}^N \omega_{t-1}^n \sum_{b \in \mathcal{B}_{t-1}} \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^1,t-1}^i, n) \Big\} \\
&\quad = \mathbb{1}_{k_t^1 \neq k_t^2} M^{-1} \sum_{i=1}^M 1 + \mathbb{1}_{k_t^1 = k_t^2} M^{-1} \sum_{n=1}^N \sum_{i=1}^M \omega_{t-1}^n \\
&\quad = \mathbb{1}_{k_t^1 \neq k_t^2} + \mathbb{1}_{k_t^1 = k_t^2} = 1 \,.
\end{aligned}
$$

which proves (B.1.18) at time t. Consequently, we have that for all $h \in \mathsf{F}(\mathcal{X})$

$$
\begin{aligned}
\sum_{b \in \mathcal{B}_t} \prod_{s=0}^t N^{-b_s} \left(\frac{N-1}{N} \right)^{1-b_s} \widetilde{\mathcal{Q}}_{b,t}^{N,M}(h^{\otimes 2}) &= \sum_{b \in \mathcal{B}_t} \frac{\gamma_t^N(\mathbf{1})^2}{N^2} \sum_{k_t^{1:2} \in [N]^2} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) h(\xi_t^{k_t^1}) h(\xi_t^{k_t^2}) \\
&= \frac{\gamma_t^N(\mathbf{1})^2}{N^2} \sum_{k_t^{1:2} \in [N]^2} h(\xi_t^{k_t^1}) h(\xi_t^{k_t^2}) \sum_{b \in \mathcal{B}_t} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \\
&= \gamma_t^N(\mathbf{1})^2 \eta_t^N(h)^2 = \gamma_t^N(h)^2 \,.
\end{aligned}
$$

The convergence in probability is then obtained by mimicking the proof of Theorem 4.4.7 and using Theorem 4.4.9.

B.1.8 Proof of Theorem 4.5.2

The proof of Theorem 4.5.2 requires the convergence of $\mathbf{T}_s^N[h_{0:s}](x)$ to $\mathbf{T}_s[h_{0:s}](x)$ $\mathbb{P}$-a.s. for any $x \in \mathsf{X}$.

Proposition B.1.2. *For any $s \in \mathbb{N}_{>0}$, any $x \in \mathsf{X}$ and additive functional $h_{0:s}$ (4.5.5),*

$$
\mathbf{T}_s^N[h_{0:s}](x) \xrightarrow[N \to \infty]{a.s.} \mathbf{T}_s[h_{0:s}](x) \,. \tag{B.1.19}
$$

Proof. Let $x \in \mathsf{X}$. Define

$$
\begin{cases}
a_N := \eta_{s-1}^N \left(g_{s-1} \{ \mathbf{T}_{s-1}^N[\tilde{h}_{0:s-1}] f_{s-1}^x + \tilde{f}_{s-1}^x \} \right) \,, \\
b_N := \eta_{s-1}^N \left(g_{s-1} m_s(., x) \right) \,, \\
b := \eta_{s-1} (g_{s-1} m_s(., x)) \,.
\end{cases}
$$

where $f_{s-1}^x : y \mapsto m_s(y, x)$ and

$$
\tilde{f}_{s-1}^x : y \mapsto m_s(y, x) \left\{ \tilde{h}_{s-1}(y, x) - \mathbf{T}_s[h_{0:s}](x) \right\} \,.
$$

Then, we have that $a_N / b_N = \mathbf{T}_s^N[h_{0:s}](x) - \mathbf{T}_s[h_{0:s}](x)$. By (**A7**), $(f_{s-1}^x, \tilde{f}_{s-1}^x) \in \mathsf{F}(\mathcal{X})^2$ for any $x \in \mathsf{X}$ and

$$
\eta_{s-1} \left(g_{s-1} \{ \mathbf{T}_{s-1}[\tilde{h}_{0:s-1}] f_{s-1}^x + \tilde{f}_{s-1}^x \} \right) = 0 \,.
$$

Hence, choosing $f_{s-1} = f^x_{s-1}$ and $\tilde{f}_{s-1} = \tilde{f}^x_{s-1}$ in Theorem B.3.2 (B.3.1), there exists $(d, \tilde{d}) \in (\mathbb{R}^*_+)^2$ such that

$$\mathbb{P}\left(|a_N| \geq \epsilon\right) \leq \tilde{d} \exp(-dN\epsilon^2) \ . \tag{B.1.20}$$

On the other hand, by choosing $f_{s-1} = f^x_{s-1}$ and $\tilde{f}_{s-1} = 0$, there exists $(d', \tilde{d}') \in (\mathbb{R}^*_+)^2$ such that

$$\mathbb{P}\left(|b_N - b| \geq \epsilon\right) \leq \tilde{d}' \exp(-d' N\epsilon^2) \, .$$

Finally, since $|a_N/b_N| \leq \left|\mathbf{T}^N_s[h_{0:s}](x)\right| + |\mathbf{T}_s[h_{0:s}](x)| \leq 2|h_{0:s}|_\infty$ $\mathbb{P}$-a.s. and $b > 0$ by $(\mathbf{A5}:\mathbf{6})$, there exist $(c_s, \tilde{c}_s) \in (\mathbb{R}^*_+)^2$ by Lemma B.3.3 such that

$$\mathbb{P}\left(|a_N/b_N| \geq \epsilon\right) = \mathbb{P}\left(\left|\mathbf{T}^N_s[h_{0:s}](x) - \mathbf{T}_s[h_{0:s}](x)\right| \geq \epsilon\right) \leq \tilde{c}_s \exp(-c_s N\epsilon^2) \, ,$$

from which (B.1.19) follows by applying the Borel-Cantelli Lemma. □

Proposition B.1.3. *For any $s \in \mathbb{N}_{>0}$ and additive functional $h_{0:s}$ (4.5.5)*

$$\lim_{N\to\infty} \mathbb{E}\left[\phi^N_{s-1} M_s\left[\left(\mathbf{T}^N_s[h_{0:s}] - \mathbf{T}_s[h_{0:s}]\right)^4\right]\right] = 0 \, . \tag{B.1.21}$$

Proof. The proof is a straightforward application of (Olsson and Westerborn, 2017, Lemma 17) (which dates back to Douc et al. (2011a)). We recall it with its proof for the sake of completeness. Define for any $x \in \mathsf{X}$

$$\begin{cases} A_N(x) & := \left|\mathbf{T}^N_s[h_{0:s}](x) - \mathbf{T}_s[h_{0:s}](x)\right|^4 \phi^N_{s-1}(m_s(., x)) \, , \\ \widetilde{A}_N(x) & := \phi^N_{s-1}(m_s(., x)) \, , \\ \widetilde{A}(x) & := \phi_{s-1}(m_s(., x)) \, . \end{cases}$$

We apply Theorem B.3.1 with $f_N = \mathbb{E}[A_N]$, $g_N = \mathbb{E}[\widetilde{A}_N]$, $f = 0$ and $g = \widetilde{A}$.

(i) For any $x_s \in \mathsf{X}$, $|\mathbf{T}^N_s[h_{0:s}]|(x_s) \leq \int |h_{0:s}|(x_{0:s}) \mathbf{T}^N_s(x_s, \mathrm{d}x_{0:s-1}) \leq |h_{0:s}|_\infty$, hence

$$\mathbb{E}[A_N(x)] \leq 16 |h_{0:s}|^4_\infty \mathbb{E}[\widetilde{A}_N(x)].$$

(ii) We have that $\widetilde{A}_N(x) \xrightarrow[N\to\infty]{a.s.} \widetilde{A}(x)$ for any $x \in \mathsf{X}$ by $(\mathbf{A7})$, (4.3.5) and by the dominated convergence theorem $\lim_{N\to\infty} \mathbb{E}[\widetilde{A}_N(x)] = \widetilde{A}(x)$. On the other hand, $\int \mathbb{E}[\widetilde{A}_N(x)]\nu(\mathrm{d}x) = \mathbb{E}[\phi^N_s(\mathbf{1})]$, $\int \widetilde{A}(x)\nu(\mathrm{d}x) = \phi_s(\mathbf{1})$ and $\lim_{N\to\infty} \mathbb{E}[\phi^N_s(\mathbf{1})] = \phi_s(\mathbf{1})$ again by the dominated convergence theorem. Hence

$$\lim_{N\to\infty} \int \mathbb{E}[\widetilde{A}_N(x)]\nu(\mathrm{d}x) = \lim_{N\to\infty} \mathbb{E}[\phi^N_s(\mathbf{1})] = \phi_s(\mathbf{1}) = \int \widetilde{A}(x)\nu(\mathrm{d}x) \, . \tag{B.1.22}$$

(iii) By Proposition B.1.2, $(\mathbf{A7})$ and (4.3.5)

$$A_N(x) \xrightarrow[N\to\infty]{a.s.} 0 \, ,$$

and since $A_N(x) \leq 16|h_{0:s}|^4_\infty \sigma_+$ $\mathbb{P}$-a.s., by the dominated convergence theorem we get $\lim_{N\to\infty} \mathbb{E}[A_N(x)] = 0$.

Finally, by Theorem B.3.1

$$\lim_{N\to\infty} \mathbb{E}\Big[\int A_N(x)\nu(\mathrm{d}x)\Big] = \lim_{N\to\infty} \mathbb{E}[\phi_{s-1}^N M_s[(\mathbf{T}_s^N[h_{0:s}] - \mathbf{T}_s[h_{0:s}])^4]] \tag{B.1.23}$$
$$= \int \lim_{N\to\infty} \mathbb{E}[A_N(x)]\nu(\mathrm{d}x) = 0\,.$$

□

Proof of Theorem 4.5.2. We write

$$\mathbf{H}_{s,t}^N := \mathbf{T}_s^N[h_{0:s}]c_t + \tilde{h}_{s:t}, \quad \mathbf{H}_{s,t} := \mathbf{T}_s[h_{0:s}]d_t + \tilde{h}_{s:t}$$
$$\mathbf{F}_{s,t}^N := \mathbf{T}_s^N[f_{0:s}]c_t + \tilde{f}_{s:t}, \quad \mathbf{F}_{s,t} := \mathbf{T}_s[f_{0:s}]d_t + \tilde{f}_{s:t}$$

We proceed by induction on $t \geq s$ with s fixed. By Theorem 4.4.4,

$$\lim_{N\to\infty} \left\| \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s} \otimes \mathbf{F}_{s,s}) - \mathcal{Q}_{e_s,s}(\mathbf{H}_{s,s} \otimes \mathbf{F}_{s,s}) \right\|_2^2 = 0\,.$$

Hence, by the triangle inequality it suffices to show that the difference with the "idealized" estimator goes to 0, i.e.

$$\lim_{N\to\infty} \left\| \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s}^N \otimes \mathbf{F}_{s,s}^N) - \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s} \otimes \mathbf{F}_{s,s}) \right\|_2^2 = 0\,. \tag{B.1.24}$$

For any $(h_{0:s}, f_{0:s})$ (4.5.5) and $(h_s, f_s) \in \mathsf{F}(\mathcal{X})^2$, by (4.4.20)

$$\begin{aligned}
&\left\| \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s}^N \otimes \mathbf{F}_{s,s}^N) - \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s} \otimes \mathbf{F}_{s,s}) \right\|_2 \\
&\leq N^{-2} \left\| \sum_{i,j\in[N]^2} \frac{N^{s+1}\gamma_s^N(\mathbf{1})^2}{(N-1)^s} \mathcal{T}_s^{e_s}(i,j) \left[\mathbf{H}_{s,s}^N(\xi_s^i)\mathbf{F}_{s,s}^N(\xi_s^j) - \mathbf{H}_{s,s}(\xi_s^i)\mathbf{F}_{s,s}(\xi_s^j) \right] \right\|_2 \\
&\leq N^{-1} \sum_{i,j\in[N]^2} \left\| \frac{N^s\gamma_s^N(\mathbf{1})^2}{(N-1)^s} \mathcal{T}_s^{e_s}(i,j) \right\|_4 \left\| \mathbf{H}_{s,s}^N(\xi_s^i)\mathbf{F}_{s,s}^N(\xi_s^j) - \mathbf{H}_{s,s}(\xi_s^i)\mathbf{F}_{s,s}(\xi_s^j) \right\|_4
\end{aligned}$$

by Cauchy-Schwarz inequality. We now show $\sup_{N\in\mathbb{N}} \left\| N^s(N-1)^{-s}\gamma_s^N(\mathbf{1})^2 \mathcal{T}_s^{e_s}(i,j) \right\|_4$ is bounded for all $(i,j) \in [N]^2$. We first show by induction that for any $n \in \mathbb{N}$, $\mathcal{T}_n^{\mathbf{0}}(i,j) \leq \mathbb{1}_{i\neq j}$. For all $(i,j) \in [N]^2$, $\mathcal{T}_0^{\mathbf{0}}(i,j) = \mathbb{1}_{i\neq j}$, and for any $n > 0$, by (4.4.23)

$$\begin{aligned}
\mathcal{T}_n^{\mathbf{0}}(i,j) &= \mathbb{1}_{i\neq j} \sum_{k,\ell\in[N]^2} \beta_n^{\mathsf{BS}}(i,k)\beta_n^{\mathsf{BS}}(j,\ell)\mathcal{T}_{n-1}^{\mathbf{0}}(k,\ell) \\
&\leq \mathbb{1}_{i\neq j} \sum_{k,\ell\in[N]^2} \beta_n^{\mathsf{BS}}(i,k)\beta_n^{\mathsf{BS}}(j,\ell)\mathbb{1}_{k\neq\ell} \\
&\leq \mathbb{1}_{i\neq j} \sum_{k,\ell\in[N]^2} \beta_n^{\mathsf{BS}}(i,k)\beta_n^{\mathsf{BS}}(j,\ell) \leq \mathbb{1}_{i\neq j}\,,
\end{aligned}$$

where we have used the induction hypothesis in the second line. This shows the result. Next, we have that

$$\begin{aligned}
\mathcal{T}_s^{e_s}(i,j) &= \mathbb{1}_{i=j} \sum_{k,\ell\in[N]^2} \beta_s^{\mathsf{BS}}(i,k)\omega_{s-1}^{\ell}\mathcal{T}_{s-1}^{\mathbf{0}}(k,\ell) \\
&\leq \mathbb{1}_{i=j} \sum_{k,\ell\in[N]^2} \beta_s^{\mathsf{BS}}(i,k)\omega_{s-1}^{\ell}\mathbb{1}_{k\neq\ell} \\
&\leq \mathbb{1}_{i=j} \sum_{k,\ell\in[N]^2} \beta_s^{\mathsf{BS}}(i,k)\omega_{s-1}^{\ell} = \mathbb{1}_{i=j}\,.
\end{aligned}$$

Hence, $\sup_{N\in\mathbb{N}} \left\| N^s \gamma_s^N(\mathbf{1})^2/(N-1)^s \mathcal{T}_s^{e_s}(i,j) \right\|_4 \leq (2G_\infty^2)^s \mathbb{1}_{i=j}$. Consequently,

$$\left\| \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,s}^N \otimes \mathbf{F}_{s,s}^N) - \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(\mathbf{H}_{s,s} \otimes \mathbf{F}_{s,s}) \right\|_2 \leq \frac{(2G_\infty^2)^s}{N} \sum_{i=1}^N \left\| \mathbf{H}_{s,s}^N(\xi_s^i)\mathbf{F}_{s,s}^N(\xi_s^i) - \mathbf{H}_{s,s}(\xi_s^i)\mathbf{F}_{s,s}(\xi_s^i) \right\|_4 , \quad \text{(B.1.25)}$$

and

$$\begin{aligned}
&\left\| \mathbf{H}_{s,s}^N(\xi_s^i)\mathbf{F}_{s,s}^N(\xi_s^j) - \mathbf{H}_{s,s}(\xi_s^i)\mathbf{F}_{s,s}(\xi_s^j) \right\|_4 \\
&\quad \leq \left\| (\mathbf{H}_{s,s}^N(\xi_s^i) - \mathbf{H}_{s,s}(\xi_s^i))\mathbf{F}_{s,s}^N(\xi_s^i) \right\|_4 + \left\| (\mathbf{F}_{s,s}^N(\xi_s^i) - \mathbf{F}_{s,s}(\xi_s^i))\mathbf{H}_{s,s}^N(\xi_s^i) \right\|_4 \\
&\quad \leq C_{f,c} \left\| \mathbf{T}_s^N[h_{0:s}](\xi_s^i) - \mathbf{T}_s[h_{0:s}](\xi_s^i) \right\|_4 + C_{h,d} \left\| \mathbf{T}_s^N[f_{0:s}](\xi_s^i) - \mathbf{T}_s[f_{0:s}](\xi_s^i) \right\|_4 .
\end{aligned}$$

where $C_{f,d} := |d_s|_\infty^4 (|f_{0:s}|_\infty + |f_s|_\infty)^4$ and $C_{h,c} := |c_s|_\infty^4 (|h_{0:s}|_\infty + |h_s|_\infty)^4$ which are finite because $\tilde{h}_s, \tilde{f}_s) \in \mathsf{F}(\mathcal{X}^{\otimes 2})^4$, $(c_s, d_s, h_s, f_s) \in \mathsf{F}(\mathcal{X})^4$. We have used that

$$|\mathbf{H}_{s,s}^N(\xi_s^i)| \leq \int |c_s|_\infty |h_{0:s}|_\infty \mathbf{T}_s^N(\xi_s^i, \mathrm{d}x_{0:s-1}) + |h_s|_\infty = |c_s|_\infty |h_{0:s}|_\infty + |h_s|_\infty = C_{h,c} .$$

For any $h_{0:s} \in \mathsf{F}(\mathcal{X}^{\otimes s+1})$,

$$\begin{aligned}
&\|\mathbf{T}_s^N[h_{0:s}](\xi_s^i) - \mathbf{T}_s[h_{0:s}](\xi_s^i)\|_4^4 \\
&\qquad = \mathbb{E}\left[\mathbb{E}[\mathbf{T}_s^N[h_{0:s}](\xi_s^i) - \mathbf{T}_s[h_{0:s}](\xi_s^i))^4 | \mathcal{F}_{t-1}^N] \right] \\
&\qquad = \mathbb{E}\left[\sum_{j=1}^N \omega_{s-1}^j \int (\mathbf{T}_s^N[h_{0:s}](x) - \mathbf{T}_s[h_{0:s}](x))^4 M_s(\xi_{s-1}^j, \mathrm{d}x) \right] \\
&\qquad = \mathbb{E}\left[\phi_{s-1}^N M_s \left[(\mathbf{T}_s^N[h_{0:s}] - \mathbf{T}_s[h_{0:s}])^4 \right] \right] ,
\end{aligned}$$

and replacing in (B.1.25) we get

$$\begin{aligned}
&\left\| \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(\mathbf{H}_{s,s}^N \otimes \mathbf{F}_{s,t}^N) - \mathcal{Q}_{e_s,s}^{N,\mathsf{BS}}(\mathbf{H}_{s,t} \otimes \mathbf{F}_{s,t}) \right\|_2^2 \\
&\qquad \leq (2G_\infty^2)^s \Big\{ C_{f,d} \mathbb{E}\left[\phi_{s-1}^N M_s \left[(\mathbf{T}_s^N[h_{0:s}] - \mathbf{T}_s[h_{0:s}])^4 \right] \right] \\
&\qquad\qquad + C_{h,c} \mathbb{E}\left[\phi_{s-1}^N M_s \left[(\mathbf{T}_s^N[f_{0:s}] - \mathbf{T}_s[f_{0:s}])^4 \right] \right] \Big\}
\end{aligned}$$

The upperbound goes to zero by Proposition B.1.3 and this finishes the proof of the initialization.

Let $t > s$ and $h_{0:s} \in \mathbb{A}_b(\mathcal{X}^{\otimes s+1})$ an additive functional. Assume that (B.1.24) holds at $t-1$. By the induction hypothesis

$$\lim_{N\to\infty} \left\| \mathcal{Q}_{e_s,t-1}^{N,\mathsf{BS}}(\mathbf{Q}_t\mathbf{H}_{s,t}^N \otimes \mathbf{Q}_t\mathbf{F}_{s,t}^N) - \mathcal{Q}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}_{s,t} \otimes \mathbf{Q}_t\mathbf{F}_{s,t}) \right\|_2 = 0 ,$$

where $\mathbf{Q}_t$ is defined in (4.3.1) and for example

$$\mathbf{Q}_t\mathbf{H}_{s,t}^N(x_{s:t-1}) = \mathbf{T}_s^N[h_{0:s}](x_s)\mathbf{Q}_t[c_t](x_{t-1}) + \mathbf{Q}_t[\tilde{h}_{s:t}](x_{s:t-1}) ,$$

where $\mathbf{Q}_t[c_t]$ and $\mathbf{Q}_t[\tilde{h}_{s:t}]$ are bounded by (**A5**), and by defintion of $\mathcal{Q}_{e_s,t}$ (4.3.1)

$$\mathcal{Q}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}_{s,t} \otimes \mathbf{Q}_t\mathbf{F}_{s,t}) = \mathcal{Q}_{e_s,t}(\mathbf{H}_{s,t} \otimes \mathbf{F}_{s,t}) .$$

Hence, to prove (4.5.10) it is enough to show

$$\lim_{N\to\infty} \left\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(\mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t}) - \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}) \right\|_2 = 0\,. \tag{B.1.26}$$

Because $\mathbf{T}^N_s[h_{0:s}]$ and $\mathbf{T}^N_s[f_{0:s}]$ are $\mathcal{F}^N_{t-1}$-measurable, by Proposition 4.4.3

$$\mathbb{E}[\mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(\mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t})|\mathcal{F}^N_{t-1}] = \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t})\,,$$

and thus

$$\begin{aligned}\left\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(\mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t}) - \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}) \right\|_2 \\ = \left\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(\mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t}) \right\|_2 - \left\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}) \right\|_2 .\end{aligned} \tag{B.1.27}$$

Now note that Proposition B.1.8 is still applicable with $h = \mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t}$ although there is a slight abuse because this specific h depends on the particles up to $s-1$ through $\mathbf{T}^N_s[h_{0:s}]$ and $\mathbf{T}^N_s[f_{0:s}]$. However, as they are $\mathcal{F}^N_{t-1}$-measurable, Proposition B.1.7 is still valid and hence Proposition B.1.8. Additionally, this specific h is bounded almost surely since for any $(x_{s:t}, x'_{s:t}) \in (\mathsf{X}^{t-s+1})^2$

$$|\mathbf{H}^N_{s,t}(x_{s:t})\mathbf{F}^N_{s,t}(x'_{s:t})| \le C := (|h_{0:s}|_\infty |c_t|_\infty + |\tilde{h}_{s:t}|_\infty)(|f_{0:s}|_\infty |d_t|_\infty + |\tilde{f}_{s:t}|_\infty)$$

and hence

$$\begin{aligned}&\left\| \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t}(\mathbf{H}^N_{s,t} \otimes \mathbf{F}^N_{s,t}) - \mathcal{Q}^{N,\mathsf{BS}}_{e_s,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}) \right\|_2 \\ &\quad\le \left[\frac{(N-2)(N-3)}{N(N-1)} - 1\right] \left\| \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}) \right\|_2^2 \\ &\qquad + \frac{N-2}{N-1} G_\infty^3 C^2 \int \nu(\mathrm{d}x)\mathbb{E}\Bigg[\frac{\Omega_{t-1}}{N}\Big\{ \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes\mathbf{1})\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes\mathbf{1}) \\ &\qquad + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes m_t(.,x))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes\mathbf{1}) + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes\mathbf{1}) \\ &\qquad \times \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\beta^N_t(x,.)) + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes m_t(.,x))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes\beta^N_t(x,.))\Big\}\Bigg] \\ &\quad + \int \nu^{\otimes 2}(\mathrm{d}y,\mathrm{d}x)\mathbb{E}\Bigg[\frac{G_\infty^2 C^2 \Omega_{t-1}^2}{N(N-1)} \\ &\qquad \times \Big\{ \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(\beta^N_t(x,.)\otimes\beta^N_t(y,.)\right) \\ &\qquad\quad + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}\left(\beta^N_t(y,.)\otimes\beta^N_t(x,.)\right)\Big\}\Bigg].\end{aligned}$$

The first term in the r.h.s. goes to zero by Proposition B.1.4 and the fact that $\mathbf{Q}_t\mathbf{H}^N_{s,t} \otimes \mathbf{Q}_t\mathbf{F}^N_{s,t}$ are bounded. The remaining terms are similar to those that appear in the proof of Theorem 4.4.4 up to some constants and thus go to zero.

For the second part, by Theorem B.3.2 and Borel-Cantelli Lemma, $\phi^N_{0:t|t}(h) \xrightarrow[N\to\infty]{a.s.} \phi_{0:t|t}(h)$. Then, by multiple applications of Theorem 4.5.2 and using the bilinearity of $\mathcal{Q}^{N,\mathsf{BS}}_{b,t}$ and $\mathcal{Q}_{b,t}$,

for any $s \in [0:t]$ and bounded additive functional h_t

$$\begin{aligned}
&\mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}([g_t\{\mathbf{T}_s^N[h_{0:s}] + \tilde{h}_{s:t} - \phi_{0:t|t}^N(h_t)\}]^{\otimes 2}) \\
&\quad = \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}([g_t\{\mathbf{T}_s^N[h_{0:s}] + \tilde{h}_{s:t}\}]^{\otimes 2}) - \phi_{0:t|t}^N(h_t)\Big(\mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}([g_t\{\mathbf{T}_s^N[h_{0:s}] + \tilde{h}_{s:t}\}] \otimes \mathbf{1}) + \\
&\qquad + \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(\mathbf{1} \otimes [g_t\{\mathbf{T}_s^N[h_{0:s}] + \tilde{h}_{s:t}\}])\Big) + \phi_{0:t|t}^N(h_t)^2 \mathcal{Q}_{e_s,t}^{N,\mathsf{BS}}(\mathbf{1} \otimes \mathbf{1}) \\
&\quad \xrightarrow{\mathbb{P}} \mathcal{Q}_{e_s,t}([g_t\{\mathbf{T}_s[h_{0:s}] + \tilde{h}_{s:t}\}]^{\otimes 2}) - \phi_{0:t|t}(h_t)\Big(\mathcal{Q}_{e_s,t}([g_t\{\mathbf{T}_s[h_{0:s}] + \tilde{h}_{s:t}\}] \otimes \mathbf{1}) + \\
&\qquad + \mathcal{Q}_{e_s,t}(\mathbf{1} \otimes [g_t\{\mathbf{T}_s[h_{0:s}] + \tilde{h}_{s:t}\}])\Big) + \phi_{0:t|t}(h_t)^2 \mathcal{Q}_{e_s,t}(\mathbf{1} \otimes \mathbf{1}) \\
&\quad = \mathcal{Q}_{e_s,t}([g_t\{\mathbf{T}_s[h_{0:s}] + \tilde{h}_{s:t} - \phi_{0:t|t}(h_t)\}]^{\otimes 2}),
\end{aligned}$$

from which the weak consistency of $\mathcal{V}_{0:t|t}^{N,\mathsf{BS}}(h)$ follows. □

B.1.9 Supporting results for Theorem 4.4.4

In this section we prove Proposition B.1.4 and the upperbound of $\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\|_2^2$ used in the proof of Theorem 4.4.4.

Proposition B.1.4. *Assume that* (**A5**) *holds. For any* $t \in \mathbb{N}$, $b \in \mathcal{B}_t$ *and* $m \in \mathbb{N}$,

$$\sup_{N \in \mathbb{N}} \mathbb{E}\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1})\|_m < \infty. \tag{B.1.28}$$

We preface the proof with supporting lemmata.

Lemma B.1.5. *For any* $p \geq 2$, *and* $N \geq 2m$

$$\mathrm{Card}(\mathcal{I}_0^m \cap \mathcal{S}_m^p) = \mathcal{O}(N^p) \quad \textit{and} \quad \mathrm{Card}(\mathcal{I}_1^m \cap \mathcal{S}_m^p) = \mathcal{O}(N^p).$$

Proof. The tuples in $\mathcal{S}_m^p$ contain p distinct elements. These p distinct elements can be selected in $\binom{N}{p}$ ways. For each of these tuples of size p, there are p^{2m} tuples of size $2m$ with each element taking one of the p values. These tuples of size $2m$ contain at most p distinct elements. Hence,

$$\mathrm{Card}(\mathcal{I}_0^m \cap \mathcal{S}_m^p) \leq \mathrm{Card}(\mathcal{S}_m^p) \leq \binom{N}{p} p^{2m} \leq \frac{N^p}{p!} p^{2m},$$

and similarly,

$$\mathrm{Card}(\mathcal{I}_1^m \cap \mathcal{S}_m^p) \leq \frac{N^p}{p!} p^{2m}.$$

□

Proposition B.1.6. *Let* $t \in \mathbb{N}_{>0}$, $m \in \mathbb{N}_{>0}$ *and* $(k_{t-1}^1, \cdots, k_{t-1}^{2m}) \in [N]^{2m}$.

(i) If $p \in [2:2m]$ *and* $(k_t^1, \cdots, k_t^{2m}) \in \mathcal{I}_0^m \cap \mathcal{S}_m^p$,

$$\mathbb{E}\left[\prod_{j=1}^{2m} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \middle| \mathcal{F}_{t-1}^N\right] \leq \frac{G_\infty^p}{\Omega_{t-1}^p}.$$

(ii) If $p \in [1:m]$ and $(k_t^1, \cdots, k_t^{2m}) \in \mathcal{I}_1^m \cap \mathcal{S}_m^p$,

$$\mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \leq \frac{G_\infty^{p+m}}{\Omega_{t-1}^{p+m}}.$$

Proof. Let $p \in [2:2m]$. By definition there are p distinct elements in each $\boldsymbol{k} := (k_t^1, \cdots, k_t^{2m}) \in \mathcal{I}_0^m \cap \mathcal{S}_m^p$. Let $\boldsymbol{k}_p := \{a_1, \cdots, a_p\} = \{k_t^1, \cdots, k_t^{2m}\}$ the set of cardinal p containing the p distinct elements in a tuple $\boldsymbol{k} \in \mathcal{S}_m^p$. Define for any $a_i \in \boldsymbol{k}_p$, $V_{a_i} := \{j \in [2m] : k_t^j = a_i\}$. Each V_{a_i} is non-empty so that it is possible to pick $j_i \in V_{a_i}$, and by (4.4.11) in Lemma 4.4.1 and the fact that different particles are i.i.d. conditionally to $\mathcal{F}_{t-1}^N$,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^{2m} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad= \prod_{i=1}^{p} \mathbb{E}\Big[\prod_{j \in V_{a_i}} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \Big| \mathcal{F}_{t-1}^N\Big] \\
&\quad= \prod_{i=1}^{p} \int \left\{\prod_{j \in V_{a_i}} \beta_t^N(\xi_t^{k_t^j}, \xi_{t-1}^{k_{t-1}^j}) \phi_{t-1}^N M_t(\mathrm{d}\xi_t^{a_i})\right\}, \\
&\quad= \prod_{i=1}^{p} \int \left\{\prod_{j \in V_{a_i} \setminus \{j_i\}} \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^j}) \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^{j_i}}) \phi_{t-1}^N M_t(\mathrm{d}\xi_t^{a_i})\right\}, \\
&\quad= \prod_{i=1}^{p} \int \left\{\prod_{j \in V_{a_i} \setminus \{j_i\}} \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^j}) \omega_{t-1}^{k_{t-1}^{j_i}} M_t(\xi_{t-1}^{k_{t-1}^{j_i}}, \mathrm{d}\xi_t^{a_i})\right\},
\end{aligned}$$

with the convention $\prod_\emptyset = 1$. Then, since for any $(k, \ell) \in [N]^2$, $\beta_t^{\mathsf{BS}}(\xi_t^k, \xi_{t-1}^\ell) \leq 1$, we get

$$\mathbb{E}\left[\prod_{j=1}^{2m} \beta_t^{\mathsf{BS}}(k_t^m, k_{t-1}^m) \middle| \mathcal{F}_{t-1}^N\right] \leq \prod_{i=1}^{p} \frac{\widetilde{\omega}_{t-1}^{k_{t-1}^{j_i}}}{\Omega_{t-1}} \int M_t(\xi_{t-1}^{k_{t-1}^{j_i}}, \mathrm{d}\xi_t^{a_i}) \leq \frac{G_\infty^p}{\Omega_{t-1}^p}.$$

Now let $p \in [1:m]$ and $\boldsymbol{k} = (k_t^1, \cdots, k_t^{2m}) \in \mathcal{I}_1^m \cap \mathcal{S}_m^p$. Define $\widetilde{V}_{a_i} := V_{a_i} \cap \{1, 3, \ldots, 2m-1\}$ for any $a_i \in \boldsymbol{k}_p$. These sets are non-empty since each V_{a_i} is non-empty and has as many even indices as odd indices, by definition of $\mathcal{I}_1^m \cap \mathcal{S}_m^p$. It is then possible to pick $j_i \in \widetilde{V}_{a_i}$ and, since

for any $i \in [N]$ ω_{t-1}^i is $\mathcal{F}_{t-1}^N$-measurable,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad = \prod_{j=1}^{m} \omega_{t-1}^{k_{t-1}^{2j}} \mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \middle| \mathcal{F}_{t-1}^N\right], \\
&\quad = \prod_{j=1}^{m} \omega_{t-1}^{k_{t-1}^{2j}} \prod_{i=1}^{p} \int \left\{ \prod_{j \in \widetilde{V}_{a_i}} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \phi_{t-1}^N M_t(\mathrm{d}\xi_t^{a_i}) \right\}, \\
&\quad = \prod_{j=1}^{m} \omega_{t-1}^{k_{t-1}^{2j}} \prod_{i=1}^{p} \int \left\{ \prod_{j \in \widetilde{V}_{a_i} \setminus \{j_i\}} \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^j}) \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^{j_i}}) \phi_{t-1}^N M_t(\mathrm{d}\xi_t^{a_i}) \right\}, \\
&\quad = \prod_{j=1}^{m} \omega_{t-1}^{k_{t-1}^{2j}} \prod_{i=1}^{p} \int \left\{ \prod_{j \in \widetilde{V}_{a^i} \setminus \{j_i\}} \beta_t^N(\xi_t^{a_i}, \xi_{t-1}^{k_{t-1}^j}) \omega_{t-1}^{k_{t-1}^{j_i}} M_t(\xi_{t-1}^{k_{t-1}^{j_i}}, \mathrm{d}\xi_t^{a_i}) \right\},
\end{aligned}$$

with the convention $\prod_{\emptyset} = 1$. Hence,

$$\mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \leq \frac{G_\infty^{p+m}}{\Omega_{t-1}^{p+m}}.$$

□

Proof of proposition B.1.4. Let $m \in \mathbb{N}$ and assume for now that $N \geq 2m$. We proceed by induction. For $t = 0$, and $b_0 = 0$ we have that $\mathcal{Q}_{\mathbf{0},0}^{N,\mathsf{BS}}(\mathbf{1}) = N^{-1}(N-1)^{-1} \sum_{i,j \in [N]^2} \mathbb{1}_{i \neq j} = 1$ which completes the proof. If $b_0 = 1$, $\mathcal{Q}_{1,0}^{N,\mathsf{BS}}(\mathbf{1}) = N^{-1} \sum_{i,j \in [N]^2} \mathbb{1}_{i=j} = 1$ and the result follows.

Let $t \in \mathbb{N}_{>0}$ and $b \in \mathcal{B}_t$. Assume (B.1.28) holds at time $t-1$. Again we treat the cases $b_t = 0$ and $b_t = 1$ separately. In the case $b_t = 0$, by (B.1.2),

$$\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1})^m\right] = \mathbb{E}\left[\sum_{k_{0:t}^{1:2m} \in [N]^{2m(t+1)}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^{2j-1}, k_{0:t}^{2j})\right]$$

and

$$\begin{aligned}
&\sum_{k_{0:t}^{1:2m} \in [N]^{2m(t+1)}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^{2j-1}, k_{0:t}^{2j}) \\
&\quad = \sum_{k_{0:t-1}^{1:2m} \in [N]^{2mt}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^{2j-1}, k_{0:t-1}^{2j}) \\
&\qquad \times \sum_{k_t^{1:2m} \in [N]^{2m}} \prod_{j=1}^{m} \frac{\Omega_{t-1}^2 \mathbb{1}_{k_t^{2j-1} \neq k_t^{2j}}}{N(N-1)} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \beta_t^{\mathsf{BS}}(k_t^{2j}, k_{t-1}^{2j}).
\end{aligned}$$

By Proposition B.1.6(i) and Lemma B.1.5, for any $k_{t-1}^{1:2m} \in [N]^{2m}$,

$$\begin{aligned}
&\sum_{k_t^{1:2m}\in[N]^{2m}} \mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1})\beta^{\mathsf{BS}}(k_t^{2j}, k_{t-1}^{2j}) \mathbb{1}_{k_t^{2j-1}\neq k_t^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad = \sum_{p=2}^{2m} \sum_{k_t^{1:2m}\in\mathcal{I}_0^m\cap\mathcal{S}_m^p} \mathbb{E}\left[\prod_{j=1}^{2m} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad \leq \sum_{p=2}^{2m} \sum_{k_t^{1:2m}\in\mathcal{I}_0^m\cap\mathcal{S}_m^p} \frac{G_\infty^p}{\Omega_{t-1}^p} \leq \sum_{p=2}^{2m} \frac{G_\infty^p \mathrm{Card}(\mathcal{I}_0^m\cap\mathcal{S}_m^p)}{\Omega_{t-1}^p} \leq C \sum_{p=2}^{2m} \frac{N^p}{\Omega_{t-1}^p},
\end{aligned}$$

where C is a constant independent of N. Consequently, using the fact that

$$\sum_{k_{0:t-1}^{1:2m}\in[N]^{2mt}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^{2j-1}, k_{0:t-1}^{2j}) = \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})^m$$

which is $\mathcal{F}_{t-1}^N$-measurable and that $\Omega_{t-1} \leq NG_\infty$ $\mathbb{P}$-a.s. by (**A5**), we get

$$\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1})^m\right] \leq C\mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})^m \frac{N^{2m-p}N^p}{N^m(N-1)^m}\right] \leq C\mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})^m\right],$$

which completes the proof. In the case $b_t = 1$, again by (B.1.2),

$$\begin{aligned}
&\sum_{k_{0:t}^{1:2m}\in[N]^{2m(t+1)}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t}^{1,2}(k_{0:t}^{2j-1}, k_{0:t}^{2j}) \\
&\quad = \sum_{k_{0:t-1}^{1:2m}\in[N]^{2mt}} \prod_{j=1}^{m} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^{2j-1}, k_{0:t-1}^{2j}) \\
&\qquad\qquad \times \sum_{k_t^{1:2m}\in[N]^{2m}} \prod_{j=1}^{m} \frac{\Omega_{t-1}^2 \mathbb{1}_{k_t^{2j-1}=k_t^{2j}}}{N} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}}.
\end{aligned}$$

Then, similarly to the case $b_t = 0$, by Proposition B.1.6-(ii) and Lemma B.1.5, for any $k_{t-1}^{1:2m} \in [N]^{2m}$,

$$\begin{aligned}
&\sum_{k_t^{1:2m}\in[N]^{2m}} \mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}} \mathbb{1}_{k_t^{2j-1}=k_t^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad = \sum_{p=1}^{m} \sum_{k_t^{1:2m}\in\mathcal{I}_m^1\cap\mathcal{S}_m^p} \mathbb{E}\left[\prod_{j=1}^{m} \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}} \middle| \mathcal{F}_{t-1}^N\right] \\
&\quad = \sum_{p=1}^{m} \sum_{k_t^{1:2m}\in\mathcal{I}_m^1\cap\mathcal{S}_m^p} \frac{G_\infty^{m+p}}{\Omega_{t-1}^{m+p}} \leq C \sum_{p=1}^{m} \frac{N^p}{\Omega_{t-1}^{m+p}},
\end{aligned}$$

where C is a constant independent of N, and

$$\mathbb{E}\left[\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1})^m\right] \leq C\mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})^m \sum_{p=1}^{m} \frac{\Omega_{t-1}^{2m}N^p}{N^m\Omega^{m+p}}\right] \leq C\mathbb{E}\left[\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(\mathbf{1})^m\right],$$

which completes the proof.

If $N < 2m$ (resp. $N < m$), then a tuple in $\mathcal{I}_0^m$ (resp. $\mathcal{I}_1^m$) contains at most N different elements and the proof proceeds similarly by truncating the sums over p. □

We now give more explicit computations for the case $m = 2$. The sets $\mathcal{I}_0^2 \cap \mathcal{S}_2^p$ and $\mathcal{I}_1^2 \cap \mathcal{S}_2^p$ are detailed in Example B.1.1.

Proposition B.1.7. *For any $h \in \mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$ and $(k_{0:t-1}^1, \cdots, k_{0:t-1}^4) \in [N]^{4t}$,*

$$\sum_{k_t^{1:4} \in [N]^4} \mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^4 \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \mathbb{1}_{k_t^1 \neq k_t^2, k_t^3 \neq k_t^4} \middle| \mathcal{F}_{t-1}^N \right]$$
$$\leq \frac{N(N-1)(N-2)(N-3)}{\Omega_{t-1}^4} (g_{t-1}^{\otimes 2} \mathcal{M}_t^0[h])^{\otimes 2}(\xi_{0:t-1}^{k_{0:t-1}^1}, \cdots, \xi_{0:t-1}^{k_{0:t-1}^4})$$
$$+ \frac{N(N-1)(N-2) G_\infty^3 |h|_\infty^2}{\Omega_{t-1}^3} \vartheta_t^N(\xi_{t-1}^{k_{t-1}^{1:4}}) + \frac{N(N-1) G_\infty^2 |h|_\infty^2}{\Omega_{t-1}^2} \upsilon_t^N(\xi_{t-1}^{k_{t-1}^{1:4}}),$$

and

$$\sum_{k_t^{1:4} \in [N]^4} \mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^2 \beta_t^{\mathsf{BS}}(k_t^{2j-1}, k_{t-1}^{2j-1}) \omega_{t-1}^{k_{t-1}^{2j}} \mathbb{1}_{k_t^1 = k_t^2, k_t^3 = k_t^4} \middle| \mathcal{F}_{t-1}^N \right]$$
$$\leq \frac{N(N-1)}{\Omega_{t-1}^4} (g_{t-1}^{\otimes 2} \mathcal{M}_t^1[h])^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) + \frac{N G_\infty^3}{\Omega_{t-1}^3} M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}).$$

where

$$\vartheta_t^N : x^{1:4} \mapsto M_t[\beta_t^N(., x^4)](x^1) + M_t[\beta_t^N(., x^4)](x^2) + M_t[\beta_t^N(., x^3)](x^1) + M_t[\beta_t^N(., x^3)](x^2),$$
$$\upsilon_t^N : x^{1:4} \mapsto M_t[\beta_t^N(., x^3)](x^1) M_t[\beta_t^N(., x^4)](x^2) + M_t[\beta_t^N(., x^4)](x^1) M_t[\beta_t^N(., x^3)](x^2).$$

Proof. Let $(k_{0:t-1}^1, \cdots, k_{0:t-1}^4) \in [N]^{4t}$. First note that

$$\sum_{k_t^{1:4} \in [N]^4} \mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^4 \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \mathbb{1}_{k_t^1 \neq k_t^2, k_t^3 \neq k_t^4} \middle| \mathcal{F}_{t-1}^N \right]$$
$$= \sum_{p=2}^4 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^p} \mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^4 \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \middle| \mathcal{F}_{t-1}^N \right].$$

We compute each term herebelow. For each $p \in [2:4]$ and $\boldsymbol{k} := (k_t^1, \cdots, k_t^4) \in \mathcal{I}_0^2 \cap \mathcal{S}_2^p$, let $\boldsymbol{k}_p := \{a_1, \cdots, a_p\} = \{k_t^1, \cdots, k_t^4\}$ the set of cardinal p containing the p distinct elements in a tuple $\boldsymbol{k} \in \mathcal{S}_2^p$. Define for any $a_i \in \boldsymbol{k}_p$, $V_{a_i} := \{k_t^j : j \in [1:4],\ k_t^j = a_i\}$.

– Let $(k_t^1, \cdots, k_t^4) \in \mathcal{I}_0^2 \cap \mathcal{S}_2^2$. Then, $\boldsymbol{k}_2 = \{k_t^1, k_t^2\}$ and we either have $V_{k_t^1} = \{k_t^1, k_t^3\}$ and $V_{k^2} = \{k_t^2, k_t^4\}$ or $V_{k_t^1} = \{k^1, k^4\}$ and $V_{k_t^2} = \{k_t^2, k_t^3\}$. Assume that $V_{k_t^1} = \{k_t^1, k_t^3\}$ and $V_{k_t^2} = \{k_t^2, k_t^4\}$. Then, by (4.4.11) in Lemma 4.4.1

$$\mathbb{E}\Big[\prod_{j=1}^4 \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \Big| \mathcal{F}_{t-1}^N \Big]$$
$$= \mathbb{E}\Big[\beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1) \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^3) | \mathcal{F}_{t-1}^N \Big] \mathbb{E}\Big[\beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^2) \beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^4) | \mathcal{F}_{t-1}^N \Big]$$
$$= \int \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^3) \omega_{t-1}^{k_{t-1}^1} M_t(\xi_{t-1}^{k_{t-1}^1}, \mathrm{d}\xi_t^{k_t^1}) \int \beta_t^{\mathsf{BS}}(k_t^2, k_{t-1}^4) \omega_{t-1}^{k_{t-1}^2} M_t(\xi_{t-1}^{k_{t-1}^2}, \mathrm{d}\xi_t^{k_t^2})$$
$$\leq \frac{G_\infty^2}{\Omega_{t-1}^2} M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}) M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^2}).$$

If $V_{k_t^1} = \{k_t^1, k_t^4\}$ and $V_{k_t^2} = \{k_t^2, k_t^3\}$,

$$\mathbb{E}\left[\prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j)\middle|\mathcal{F}_{t-1}^N\right] \leq \frac{G_\infty^2}{\Omega_{t-1}^2} M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^1}) M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^2}),$$

and

$$\begin{aligned}
&\sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^2} \mathbb{E}\left[\prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j)\middle|\mathcal{F}_{t-1}^N\right] \\
&\qquad \leq \frac{N(N-1)G_\infty^2}{\Omega_{t-1}^2}\Big\{M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}) M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^2}) \\
&\qquad\qquad + M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^1}) M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^2})\Big\}.
\end{aligned}$$

– Let $(k_t^1, \cdots, k_t^4) \in \mathcal{I}_0^2 \cap \mathcal{S}_2^3$. Then, either $\boldsymbol{k}_3 = \{k_t^1, k_t^2, k_t^3\}$ or $\boldsymbol{k}_3 = \{k_t^1, k_t^2, k_t^4\}$. Assume that $\boldsymbol{k}_3 = \{k_t^1, k_t^2, k_t^3\}$ and $V_{k_t^1} = \{k_t^1, k_t^4\}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j)\middle|\mathcal{F}_{t-1}^N\right] &= \omega_{t-1}^{k_{t-1}^1} \omega_{t-1}^{k_{t-1}^2} \omega_{t-1}^{k_{t-1}^3} M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^1}) \\
&\leq \frac{G_\infty^3}{\Omega_{t-1}^3} M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^1}).
\end{aligned}$$

Applying the same reasoning to all the combinations within $\mathcal{I}_0^2 \cap \mathcal{S}_2^3$ we get

$$\begin{aligned}
&\sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^3} \mathbb{E}\left[\prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j)\middle|\mathcal{F}_{t-1}^N\right] \\
&\qquad \leq \frac{N(N-1)(N-2)G_\infty^3}{\Omega_{t-1}^3}\Big\{M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^1}) + M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^4})\right](\xi_{t-1}^{k_{t-1}^2}) \\
&\qquad\qquad + M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}) + M_t\left[\beta_t^N(., \xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^2})\Big\}.
\end{aligned}$$

– Let $(k_t^1, \cdots, k_t^4) \in \mathcal{I}_0^4 \cap \mathcal{S}_2^4$. Then, $\boldsymbol{k}_4 = \{k_t^1, k_t^2, k_t^3, k_t^4\}$ and

$$\begin{aligned}
&\mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j)\middle|\mathcal{F}_{t-1}^N\right] \\
&\qquad = \int h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^{4} \beta_t^{\mathsf{BS}}(k_t^j, k_{t-1}^j) \phi_{t-1}^N M_t(\mathrm{d}\xi_t^{k_t^j}) \\
&\qquad = \int h^{\otimes 2}(\xi_{0:t}^{k_{0:t}^1}, \cdots, \xi_{0:t}^{k_{0:t}^4}) \prod_{j=1}^{4} \omega_{t-1}^{k_{t-1}^j} M_t(\xi_{t-1}^{k_{t-1}^j}, \mathrm{d}\xi_t^{k_t^j}). \\
&\qquad = (g_{t-1}^{\otimes 2} \mathcal{M}_t^0[h])^{\otimes 2}(\xi_{0:t-1}^{k_{0:t-1}^1}, \cdots, \xi_{0:t-1}^{k_{0:t-1}^4})/\Omega_{t-1}^4,
\end{aligned}$$

which completes the proof of the first inequality. For the second inequality, write

$$\sum_{k_t^{1:4}\in[N]^4} \mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_t^1},\cdots,\xi_{0:t}^{k_t^4})\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\mathbb{1}_{k_t^1=k_t^2,k_t^3=k_t^4}\middle|\mathcal{F}_{t-1}^N\right]$$
$$=\sum_{p=1}^{2}\sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^p}\mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_t^1},\cdots,\xi_{0:t}^{k_t^4})\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\middle|\mathcal{F}_{t-1}^N\right],$$

– Let $(k_t^1,\cdots,k_t^4)\in\mathcal{I}_1^2\cap\mathcal{S}_2^1$. Then $\boldsymbol{k}_1=\{k_t^1\}$ and

$$\mathbb{E}\left[\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\middle|\mathcal{F}_{t-1}^N\right]\le\omega_{t-1}^{k_{t-1}^1}\omega_{t-1}^{k_{t-1}^2}\omega_{t-1}^{k_{t-1}^4}M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})](\xi_{t-1}^{k_{t-1}^1})$$
$$\le\frac{G_\infty^3}{\Omega_{t-1}^3}M_t\left[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}),$$

and

$$\sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^1}\mathbb{E}\left[\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\middle|\mathcal{F}_{t-1}^N\right]\le\frac{NG_\infty^3}{\Omega_{t-1}^3}M_t\left[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})\right](\xi_{t-1}^{k_{t-1}^1}).$$

– Let $(k_t^1,\cdots,k_t^4)\in\mathcal{I}_1^2\cap\mathcal{S}_2^2$. Then, $\boldsymbol{k}_2=\{k_t^1,k_t^3\}$, $V_{k_t^1}=\{k_t^1,k_t^2\}$ and $V_{k_t^3}=\{k_t^3,k_t^4\}$. Hence,

$$\mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_t^1},\cdots,\xi_{0:t}^{k_t^4})\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\middle|\mathcal{F}_{t-1}^N\right]$$
$$=(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])^{\otimes 2}(\xi_{0:t-1}^{k_{0:t-1}^1},\cdots,\xi_{0:t-1}^{k_{0:t-1}^4})/\Omega_{t-1}^4,$$

and

$$\sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^2}\mathbb{E}\left[h^{\otimes 2}(\xi_{0:t}^{k_t^1},\cdots,\xi_{0:t}^{k_t^4})\prod_{j=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2j-1},k_{t-1}^{2j-1})\omega_{t-1}^{k_{t-1}^{2j}}\middle|\mathcal{F}_{t-1}^N\right]$$
$$=\frac{N(N-1)}{\Omega_{t-1}^4}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])^{\otimes 2}(\xi_{0:t-1}^{k_{0:t-1}^1},\cdots,\xi_{0:t-1}^{k_{0:t-1}^4}),$$

which completes the proof. □

Proposition B.1.8. *For any $t\in\mathbb{N}$, $h\in\mathsf{F}(\mathcal{X}^{\otimes 2(t+1)})$ and $b\in\mathcal{B}_t$,*

(i) If $b_t=0$,

$$\|\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h)\|_2^2\le\frac{(N-2)(N-3)}{N(N-1)}\|\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])\|_2^2+\frac{N-2}{N-1}G_\infty^3|h|_\infty^2\mathbb{E}\left[\frac{\Omega_{t-1}}{N}\nu(\Theta_{b,t}^N)\right]$$
$$+\mathbb{E}\left[\frac{G_\infty^2|h|_\infty^2\Omega_{t-1}^2}{N(N-1)}\nu^{\otimes 2}(\Upsilon_{b,t}^N)\right],$$

where

$$\Theta^N_{b,t} : x \mapsto [\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes \mathbf{1}) + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes m_t(.,x))] \\ \times [\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes \mathbf{1}) + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\mathbf{1}\otimes \beta^N_t(x,.))]\,,$$

$$\Upsilon^N_{b,t} : (x,y) \mapsto \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes \beta^N_t(y,.)) \\ + \mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(y,.)\otimes \beta^N_t(x,.))\,.$$

(ii) If $b_t = 1$,

$$\|\mathcal{Q}^{N,\mathsf{BS}}_{b,t}(h)\|_2^2 \le \frac{N-1}{N}\|\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(g^{\otimes 2}_{t-1}\mathcal{M}^1_t[h])\|_2^2 \\ + G^3_\infty |h|^2_\infty \int \mathbb{E}\left[\frac{\Omega_{t-1}}{N}\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(m_t(.,x)\otimes \mathbf{1})\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(\beta^N_t(x,.)\otimes \mathbf{1})\right]\nu(\mathrm{d}x)\,. \quad \text{(B.1.29)}$$

Proof. To prove (i), if $b_t = 0$, by (B.1.2),

$$\mathcal{Q}^{N,\mathsf{BS}}_{b,t}(h)^2 = \sum_{k^{1:4}_{0:t}\in[N]^{4(t+1)}} \overline{\Lambda}^{1,2}_{b,t}(k^1_{0:t},k^2_{0:t})\overline{\Lambda}^{1,2}_{b,t}(k^3_{0:t},k^4_{0:t})h^{\otimes 2}(\xi^{k^1_{0:t}}_{0:t},\cdots,\xi^{k^4_{0:t}}_{0:t}) \\ = \sum_{k^{1:4}_{0:t-1}\in[N]^{4t}} \overline{\Lambda}^{1,2}_{b,t}(k^1_{0:t-1},k^2_{0:t-1})\overline{\Lambda}^{1,2}_{b,t}(k^3_{0:t-1},k^4_{0:t-1}) \\ \times\left\{\sum_{p=2}^4 \sum_{k^{1:4}_t\in\mathcal{I}^2_0\cap\mathcal{S}^p_2} \frac{\Omega^4_{t-1}}{N^2(N-1)^2}\prod_{j=1}^4 \beta^{\mathsf{BS}}_t(k^j_t,k^j_{t-1})h^{\otimes 2}(\xi^{k^1_{0:t}}_{0:t},\cdots,\xi^{k^4_{0:t}}_{0:t})\right\},$$

and by Proposition B.1.7,

$$\|\mathcal{Q}^{N,\mathsf{BS}}_{b,t}(h)\|_2^2 \le \frac{(N-2)(N-3)}{N(N-1)}\left\|\mathcal{Q}^{N,\mathsf{BS}}_{b,t-1}(g^{\otimes 2}_{t-1}\mathcal{M}^0_t[h])\right\|_2^2 \\ + \frac{G^3_\infty|h|^2_\infty(N-2)}{N(N-1)}\mathbb{E}\left[\Omega_{t-1}\sum_{k^{1:4}_{0:t-1}\in[N]^{4t}}\overline{\Lambda}^{1,2}_{b,t-1}(k^1_{0:t-1},k^2_{0:t-1})\overline{\Lambda}^{1,2}_{b,t-1}(k^3_{0:t-1},k^4_{0:t-1})\right. \\ \times\left\{M_t\left[\beta^N_t(.,\xi^{k^4_{t-1}}_{t-1})\right](\xi^{k^1_{t-1}}_{t-1}) + M_t\left[\beta^N_t(.,\xi^{k^4_{t-1}}_{t-1})\right](\xi^{k^2_{t-1}}_{t-1})\right. \\ \left.\left.+M_t\left[\beta^N_t(.,\xi^{k^3_{t-1}}_{t-1})\right](\xi^{k^1_{t-1}}_{t-1}) + M_t\left[\beta^N_t(.,\xi^{k^3_{t-1}}_{t-1})\right](\xi^{k^2_{t-1}}_{t-1})\right\}\right] \\ + \frac{G^2_\infty|h|^2_\infty}{N(N-1)}\mathbb{E}\left[\Omega^2_{t-1}\sum_{k^{1:4}_{0:t-1}\in[N]^{4t}}\overline{\Lambda}^{1,2}_{b,t-1}(k^1_{0:t-1},k^2_{0:t-1})\overline{\Lambda}^{1,2}_{b,t-1}(k^3_{0:t-1},k^4_{0:t-1})\right. \\ \times\left\{M_t\left[\beta^N_t(.,\xi^{k^3_{t-1}}_{t-1})\right](\xi^{k^1_{t-1}}_{t-1})M_t\left[\beta^N_t(.,\xi^{k^4_{t-1}}_{t-1})\right](\xi^{k^2_{t-1}}_{t-1})\right. \\ \left.\left.+M_t\left[\beta^N_t(.,\xi^{k^4_{t-1}}_{t-1})\right](\xi^{k^1_{t-1}}_{t-1})M_t\left[\beta^N_t(.,\xi^{k^3_{t-1}}_{t-1})\right](\xi^{k^2_{t-1}}_{t-1})\right\}\right]\,.$$

Then using (**A4**),

$$\begin{aligned}
&\sum_{k_{0:t-1}^{1:4}\in[N]^{4t}} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^3,k_{0:t-1}^4) M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})](\xi_{t-1}^{k_{t-1}^1}) \\
&= \int \sum_{k_{0:t-1}^{1:4}\in[N]^{4t}} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^3,k_{0:t-1}^4)\beta_t^N(x,\xi_{t-1}^{k_{t-1}^3}) m_t(\xi_{t-1}^{k_{t-1}^1},x)\nu(\mathrm{d}x) \\
&= \int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(m_t(.,x)\otimes\mathbf{1}\right)\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(x,.)\otimes\mathbf{1}\right)\nu(\mathrm{d}x)\,,
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{k_{0:t-1}^{1:4}\in[N]^{4t}} \overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^1,k_{0:t-1}^2)\overline{\Lambda}_{b,t-1}^{1,2}(k_{0:t-1}^3,k_{0:t-1}^4) \\
&\qquad\qquad \times M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})](\xi_{t-1}^{k_{t-1}^1}) M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^4})](\xi_{t-1}^{k_{t-1}^2}) \\
&= \int \mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}(m_t(.,x)\otimes m_t(.,y))\mathcal{Q}_{b,t-1}^{N,\mathsf{BS}}\left(\beta_t^N(x,.)\otimes\beta_t^N(y,.)\right)\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y)\,.
\end{aligned}$$

which completes the proof of (i) by applying the same reasoning to the remaining terms. Item (ii) is obtained in the same way. □

B.1.10 Supporting results for Theorem 4.4.9

In this section, we prove the analogues of Propositions 4.4.3-B.1.4-B.1.6 and B.1.7 for $\widetilde{\mathcal{Q}}_{b,t}^{N,M}$. We remind the reader that the number of sampled indices M in the *PaRIS* estimator is *fixed* and that $\mathcal{G}_t^N$ is defined in (B.1.15). Let $\mathcal{G}_{t-1}^N \vee \xi_t^{1:N}$ be the following σ-algebra:

$$\mathcal{G}_{t-1}^N \vee \xi_t^{1:N} := \sigma(\mathcal{G}_{t-1}^N \cup \sigma(\xi_t^{1:N}))\,. \tag{B.1.30}$$

Lemma B.1.9. *For any $t\in\mathbb{N}_{>0}$, $h\in\mathsf{F}(\mathcal{X}^{\otimes 2})$ and $b\in\mathcal{B}_t$,*

(i) $\mathbb{E}[\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)|\mathcal{G}_{t-1}^N] = \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^{b_t}[h])$,

(ii) $\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)$ *is an unbiased estimator of* $\mathcal{Q}_{b,t}(h)$,

where $\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)$ *is defined in* (4.4.31).

Proof. We start with the case $b_t = 0$. For any $(k,\ell)\in[N]^2$, $J_{k,t-1}^i$ and $J_{\ell,t-1}^i$ are independent conditionally on $\mathcal{G}_{t-1}^N \vee \xi_t^{1:N}$ for any $i\in[M]$ if $k\neq\ell$ and $\widetilde{\mathcal{T}}_{t-1}^b$ is $\mathcal{G}_{t-1}^N$-measurable, hence, using (4.4.4),

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathcal{T}}_t^b(k,\ell)h(\xi_t^k,\xi_t^\ell)\middle|\mathcal{G}_{t-1}^N\right] \qquad\qquad (\text{B.1.31})\\
&\quad= \mathbb{E}[\mathbb{1}_{k\neq\ell}h(\xi_t^k,\xi_t^\ell)M^{-1}\sum_{i=1}^M \widetilde{\mathcal{T}}_{t-1}^b(J_{k,t-1}^i,J_{\ell,t-1}^i)|\mathcal{G}_{t-1}^N] \\
&\quad= \mathbb{E}\left[\mathbb{1}_{k\neq\ell}h(\xi_t^k,\xi_t^\ell)M^{-1}\sum_{i=1}^M\sum_{n,m\in[N]^2}\beta_t^{\mathsf{BS}}(k,n)\beta_t^{\mathsf{BS}}(\ell,m)\widetilde{\mathcal{T}}_{t-1}^b(n,m)|\mathcal{G}_{t-1}^N\right] \\
&\quad= \sum_{n,m\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(n,m)\mathbb{E}\left[\mathbb{1}_{k\neq\ell}\beta_t^{\mathsf{BS}}(k,n)\beta_t^{\mathsf{BS}}(\ell,m)h(\xi_t^k,\xi_t^\ell)|\mathcal{F}_{t-1}^N\right] \\
&\quad= \frac{\mathbb{1}_{k\neq\ell}}{\Omega_{t-1}^2}\sum_{n,m\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(n,m)(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])(\xi_{t-1}^n,\xi_{t-1}^m)\,.
\end{aligned}$$

If $b_t = 1$,

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\mathcal{T}}_t^b(k,\ell)h(\xi_t^k,\xi_t^\ell)\middle|\mathcal{G}_{t-1}^N\right] \qquad \text{(B.1.32)}\\
&\quad= \mathbb{E}[\mathbb{1}_{k=\ell}h(\xi_t^k,\xi_t^\ell)M^{-1}\sum_{i=1}^M\sum_{m=1}^N \omega_{t-1}^m\widetilde{\mathcal{T}}_{t-1}^b(J_{k,t-1}^i,m)|\mathcal{G}_{t-1}^N]\\
&\quad= \mathbb{E}\left[\mathbb{1}_{k=\ell}h(\xi_t^k,\xi_t^\ell)M^{-1}\sum_{i=1}^M\sum_{n,m\in[N]^2}\beta_t^{\mathsf{BS}}(k,n)\omega_{t-1}^m\widetilde{\mathcal{T}}_{t-1}^b(n,m)|\mathcal{G}_{t-1}^N\right]\\
&\quad= \sum_{n,m\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(n,m)\mathbb{E}\left[\mathbb{1}_{k=\ell}\beta_t^{\mathsf{BS}}(k,n)\omega_{t-1}^m h(\xi_t^k,\xi_t^\ell)|\mathcal{F}_{t-1}^N\right]\\
&\quad= \frac{\mathbb{1}_{k=\ell}}{\Omega_{t-1}^2}\sum_{n,m\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(n,m)(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])(\xi_{t-1}^n,\xi_{t-1}^m)\,.
\end{aligned}$$

Consequently, if $b_t = 0$, by (B.1.31),

$$\begin{aligned}
&\mathbb{E}[\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)|\mathcal{G}_{t-1}^N]\\
&\quad= \prod_{s=0}^{t-1}N^{b_s}\left(\frac{N}{N-1}\right)^{1-b_s}\frac{\gamma_{t-1}^N(\mathbf{1})^2}{N^2}\sum_{k,\ell\in[N]^2}\frac{\Omega_{t-1}^2}{N(N-1)}\mathbb{E}\left[\widetilde{\mathcal{T}}_t^b(k,\ell)h(\xi_t^k,\xi_t^\ell)|\mathcal{G}_{t-1}^N\right]\\
&\quad= \prod_{s=0}^{t-1}N^{b_s}\left(\frac{N}{N-1}\right)^{1-b_s}\frac{\gamma_{t-1}^N(\mathbf{1})^2}{N^2}\sum_{n,m\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(n,m)(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])(\xi_{t-1}^n,\xi_{t-1}^m)\\
&\quad= \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])\,,
\end{aligned}$$

and in a similar way, $\mathbb{E}[\widetilde{\mathcal{Q}}_{b,t}^{N,M}(h)|\mathcal{G}_{t-1}^N] = \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h])$ by (B.1.32) if $b_t = 1$. The second item follows straightforwardly by induction and the tower property. The induction is initialized by noting that $\widetilde{\mathcal{Q}}_{b,0}^{N,M}(h)$ is equal to $\mathcal{Q}_{b,0}^{N,\mathsf{BS}}(h)$ which is an unbiased estimator of $\mathcal{Q}_{b,0}(h)$ by Proposition 4.4.3. □

Proposition B.1.10. *Let $t > 0$ and $N \geq 4$.*

(i) If $b_t = 0$,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{C}_{N,b,t}^2\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^4}h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\middle|\mathcal{G}_{t-1}^N\right]\\
&\qquad= \frac{(N-2)(N-3)}{N(N-1)}\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h])^2\,, \qquad \text{(B.1.33)}
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{C}_{N,b,t}^2\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^2}\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\middle|\mathcal{G}_{t-1}^N\right] \leq \frac{G_\infty^2(M-1)\Omega_{t-1}^2}{MN(N-1)}\nu^{\otimes 2}(\widetilde{\Upsilon}_{b,t}^{N,M})\\
&\qquad+ \frac{G_\infty^2\Omega_{t-1}^2}{MN(N-1)}\mathcal{C}_{N,b,t-1}^2\sum_{k_{t-1}^{1:2}\in[N]^2}\mathsf{T}_{t-1,b}^{(1)}(k_{t-1}^1,k_{t-1}^2)\,, \qquad \text{(B.1.34)}
\end{aligned}$$

and

$$\mathbb{E}\Big[\mathcal{C}^2_{N,b,t}\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^3}\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]\le\frac{\Omega_{t-1}G_\infty^3(M-1)(N-2)}{MN(N-1)}\nu(\widetilde{\Theta}_{b,t}^{N,M})$$
$$+\frac{\Omega_{t-1}G_\infty^3(N-2)}{MN(N-1)}\mathcal{C}^2_{N,b,t-1}\sum_{k_{t-1}^{1:3}\in[N]^3}\mathsf{T}^{(2)}_{t-1,b}(k_{t-1}^{1:3})\,,\quad\text{(B.1.35)}$$

where

$$\widetilde{\Theta}_{b,t}^{N,M}:x\mapsto[\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes\mathbf{1})+\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1}\otimes m_t(.,x))]$$
$$\times[\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x,.)\otimes\mathbf{1})+\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1}\otimes\beta_t^N(x,.))]\,,$$
$$\widetilde{\Upsilon}_{b,t}^{N,M}:(x,y)\mapsto\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes m_t(.,y))\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x,.)\otimes\beta_t^N(y,.))$$
$$+\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes m_t(.,y))\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(y,.)\otimes\beta_t^N(x,.))\,,$$

and

$$\mathsf{T}^{(1)}_{b,t}:(k_t^1,k_t^2)\mapsto\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)^2+\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^2,k_t^1)\,,$$
$$\mathsf{T}^{(2)}_{b,t}:(k_t^1,k_t^2,k_t^3)\mapsto\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^1)+\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^3)$$
$$+\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^2,k_t^3)+\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^2)\,.$$

(ii) If $b_t=1$,

$$\mathbb{E}\Big[\mathcal{C}^2_{N,b,t}\sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^2}h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]$$
$$=\frac{N-1}{N}\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes2}\mathcal{M}_t^1[h])^2\,,\quad\text{(B.1.36)}$$

and

$$\mathbb{E}\Big[\mathcal{C}^2_{N,b,t}\sum_{k_t^{1:4}\in\mathcal{I}_1^2\cap\mathcal{S}_2^1}\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]\quad\text{(B.1.37)}$$
$$\le\frac{\Omega_{t-1}G_\infty^3}{N}\int\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes\mathbf{1})\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x,.)\otimes1)\nu(\mathrm{d}x)$$
$$+\frac{G_\infty^3\Omega_{t-1}}{MN}\mathcal{C}^2_{N,b,t-1}\sum_{k_{t-1}^{1,2,4}\in[N]^3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^4)\,.$$

Proof. We start with the case $b_t=0$.

– If $(k_t^1,\cdots,k_t^4)\in\mathcal{I}_0^2\cap\mathcal{S}_2^4$, then $k_t^1\neq k_t^2\neq k_t^3\neq k_t^4$ and conditionally on $\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}$, $J^i_{k_t^1,t-1}$, $J^i_{k_t^2,t-1}$, $J^j_{k_t^3,t-1}$ and $J^j_{k_t^4,t-1}$ are independent for any $(i,j)\in[M]^2$ and $J^{1:M}_{k_t^\ell,t-1}\overset{\text{iid}}{\sim}\beta_t^{\mathsf{BS}}(k_t^\ell,.)$ for any $\ell\in[1:4]$, thus, for any $(i,j)\in[M]^2$

$$\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\widetilde{\mathcal{T}}_t^b(J^j_{k_t^3,t-1},J^j_{k_t^4,t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]$$
$$=\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(J^i_{k_t^3,t-1},J^i_{k_t^4,t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]$$
$$=\sum_{k_{t-1}^{1:4}\in[N]^4}\prod_{n=1}^4\beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\,,$$

and

$$\begin{aligned}
&\mathbb{E}\Big[h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})M^{-2}\sum_{i,j\in[M]^2}\widetilde{\mathcal{T}}_t^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\widetilde{\mathcal{T}}_t^b(J^j_{k_t^3,t-1},J^j_{k_t^4,t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\
&=\mathbb{E}\Big[h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})M^{-2}\sum_{i,j\in[M]^2}\sum_{k_{t-1}^{1:4}\in[N]^4}\prod_{n=1}^4\beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\\
&\qquad\qquad\times\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\Big|\mathcal{G}_{t-1}^N\Big]\\
&=\sum_{k_{t-1}^{1:4}\in[N]^4}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\mathbb{E}\Big[\prod_{n=1}^4\beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\Big|\mathcal{G}_{t-1}^N\Big]\\
&=\frac{1}{\Omega_{t-1}^4}\sum_{k_{t-1}^{1:4}\in[N]^4}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\big(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h]\big)^{\otimes 2}(\xi_{t-1}^{k_{t-1}^1},\xi_{t-1}^{k_{t-1}^2},\xi_{t-1}^{k_{t-1}^3},\xi_{t-1}^{k_{t-1}^4}),
\end{aligned}$$

where we have used that $\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)$ and $\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)$ are $\mathcal{G}_{t-1}^N$-measurable in the third equality and then proceeded similarly to Proposition B.1.7. By Example B.1.1, $\mathrm{Card}(\mathcal{I}_0^2\cap\mathcal{S}_2^4)=N(N-1)(N-2)(N-3)$ and since $b_t=0$ implies that

$$\mathcal{C}_{N,b,t}^2=\mathcal{C}_{N,b,t-1}^2\frac{\Omega_{t-1}^4}{N^2(N-1)^2}, \tag{B.1.38}$$

we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^4}h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]\\
&=\mathcal{C}_{N,b,t}^2\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^4}\mathbb{E}\Big[h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\\
&\qquad\qquad\times M^{-2}\sum_{i,j\in[M]^2}\widetilde{\mathcal{T}}_t^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\widetilde{\mathcal{T}}_t^b(J^j_{k_t^3,t-1},J^j_{k_t^4,t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\
&=\frac{(N-2)(N-3)}{N(N-1)}\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}\big(g_{t-1}^{\otimes 2}\mathcal{M}_t^0[h]\big)^2.
\end{aligned}$$

– If $(k_t^1,\cdots,k_t^4)\in\mathcal{I}_0^2\cap\mathcal{S}_2^2$. Then $\boldsymbol{k}_2=\{k_t^1,k_t^2\}$ and we either have $V_{k_t^1}=\{k_t^1,k_t^3\}$ and $V_{k_t^2}=\{k_t^2,k_t^4\}$ or $V_{k_t^1}=\{k_t^1,k_t^4\}$ and $V_{k_t^2}=\{k_t^2,k_t^3\}$. Assume that $V_{k_t^1}=\{k_t^1,k_t^3\}$ and $V_{k_t^2}=\{k_t^2,k_t^4\}$. Taking into account that $J^i_{k_t^1,t-1}=J^i_{k_t^3,t-1}$ and $J^i_{k_t^2,t-1}=J^i_{k_t^4,t-1}$, we get

$$\begin{aligned}
&\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]\\
&\quad=\mathbb{E}\Big[M^{-2}\sum_{i,j\in[M]^2}\widetilde{\mathcal{T}}_{t-1}^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\widetilde{\mathcal{T}}_{t-1}^b(J^j_{k_t^3,t-1},J^j_{k_t^4,t-1})|\mathcal{G}_{t-1}^N]\\
&\quad=\mathbb{E}\Big[M^{-2}\sum_{i,j\in[M]^2}\mathbb{1}_{i=j}\sum_{k_{t-1}^{1:2}\in[N]^2}\beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)^2\\
&\qquad+M^{-2}\sum_{i,j\in[M]^2}\mathbb{1}_{i\neq j}\sum_{k_{t-1}^{1:4}\in[N]^4}\prod_{\ell=1}^4\beta_t^{\mathsf{BS}}(k_t^\ell,k_{t-1}^\ell)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)|\mathcal{G}_{t-1}^N\Big].
\end{aligned}$$

Then,

$$\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big] \le \frac{G_\infty^2}{M\Omega_{t-1}^2}\sum_{k_{t-1}^{1:2}\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)^2 + \frac{(M-1)G_\infty^2}{M\Omega_{t-1}^2}$$
$$\times \sum_{k_{t-1}^{1:4}\in[N]^4} M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})](\xi_{t-1}^{k_{t-1}^1})M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^4})](\xi_{t-1}^{k_{t-1}^2})\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\,.$$

Similarly, if $V_{k_t^1}=\{k_t^1,k_t^4\}$ and $V_{k_t^2}=\{k_t^2,k_t^3\}$, we obtain

$$\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]$$
$$\le \frac{G_\infty^2}{M\Omega_{t-1}^2}\sum_{k_{t-1}^{1:2}\in[N]^2}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^2,k_{t-1}^1) + \frac{(M-1)G_\infty^2}{M\Omega_{t-1}^2}$$
$$\times \sum_{k_{t-1}^{1:4}\in[N]^4} M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^4})](\xi_{t-1}^{k_{t-1}^1})M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^3})](\xi_{t-1}^{k_{t-1}^2})\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\,.$$

Consequently,

$$\mathbb{E}\Big[\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^2}\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]$$
$$\le \frac{G_\infty^2 N(N-1)}{M\Omega_{t-1}^2}\sum_{k_{t-1}^{1:2}\in[N]^2}\Big\{\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)^2+\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^2,k_{t-1}^1)\Big\}$$
$$+\frac{G_\infty^2 N(N-1)(M-1)}{M\Omega_{t-1}^2}\int\sum_{k_{t-1}^{1:4}\in[N]^4}\Big\{m_t(\xi_{t-1}^{k_{t-1}^1},x)\beta_t^N(x,\xi_{t-1}^{k_{t-1}^3})m_t(\xi_{t-1}^{k_{t-1}^2},y)\beta_t^N(y,\xi_{t-1}^{k_{t-1}^4})+$$
$$m_t(\xi_{t-1}^{k_{t-1}^1},x)\beta_t^N(x,\xi_{t-1}^{k_{t-1}^4})m_t(\xi_{t-1}^{k_{t-1}^2},y)\beta_t^N(y,\xi_{t-1}^{k_{t-1}^3})\Big\}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y)\,.$$

Then, using (B.1.38),

$$\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2\sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^2}\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big]$$
$$\le \frac{G_\infty^2\Omega_{t-1}^2}{MN(N-1)}\mathcal{C}_{N,b,t-1}^2\sum_{k_{t-1}^{1:2}\in[N]^2}\Big\{\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)^2+\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^2,k_{t-1}^1)\Big\}$$
$$+\frac{G_\infty^2(M-1)\Omega_{t-1}^2}{MN(N-1)}\int\Big\{\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes m_t(.,y))\,\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x,.)\otimes\beta_t^N(y,.))$$
$$+\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(.,x)\otimes m_t(.,y))\,\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(y,.)\otimes\beta_t^N(x,.))\Big\}\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y)\,,$$

which yields (B.1.34).

– If $(k_t^1,\cdots,k_t^4)\in\mathcal{I}_0^2\cap\mathcal{S}_2^3$. Then either $\boldsymbol{k}_3=\{k_t^1,k_t^2,k_t^3\}$ or $\boldsymbol{k}_3=\{k_t^1,k_t^2,k_t^4\}$. Assume that

$\boldsymbol{k}_3 = \{k_t^1, k_t^2, k_t^3\}$. Then $V_{k_t^1} = \{k_t^1, k_t^4\}$, $J^i_{k_t^1,t-1} = J^i_{k_t^4,t-1}$ for any $i \in [M]$ and

$$
\begin{aligned}
&\mathbb{E}\Big[\widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big|\mathcal{G}_{t-1}^N\Big] \\
&= \mathbb{E}\Big[M^{-2}\sum_{i,j\in[M]^2}\widetilde{\mathcal{T}}_{t-1}^b(J^i_{k_t^1,t-1},J^i_{k_t^2,t-1})\widetilde{\mathcal{T}}_{t-1}^b(J^j_{k_t^3,t-1},J^j_{k_t^4,t-1})|\mathcal{G}_{t-1}^N\Big] \\
&= \mathbb{E}\Big[M^{-2}\sum_{i,j\in[M]^2}\mathbb{1}_{i=j}\sum_{k_{t-1}^{1:3}\in[N]^3}\prod_{i=1}^{3}\beta_t^{\mathsf{BS}}(k_t^i,k_{t-1}^i)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^1) \\
&+ M^{-2}\sum_{i,j\in[M]^2}\mathbb{1}_{i\neq j}\sum_{k_{t-1}^{1:4}\in[N]^4}\prod_{\ell=1}^{4}\beta_t^{\mathsf{BS}}(k_t^\ell,k_{t-1}^\ell)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)|\mathcal{G}_{t-1}^N\Big] \\
&\leq \frac{G_\infty^3}{M\Omega_{t-1}^3}\sum_{k_{t-1}^{1:3}\in[N]^3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^1) \\
&\qquad + \frac{(M-1)G_\infty^3}{M\Omega_{t-1}^3}\sum_{k_{t-1}^{1:4}\in[N]^4}M_t[\beta_t^N(.,\xi_{t-1}^{k_{t-1}^4})](\xi_{t-1}^{k_{t-1}^1})\widetilde{\mathcal{T}}_t^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_t^b(k_{t-1}^3,k_{t-1}^4)\,.
\end{aligned}
$$

The remaining combinations are treated in the exact same way and (B.1.35) is obtained by using again (B.1.38).

Consider now the case $b_t = 1$ and let $(k_t^1,\cdots,k_t^4) \in \mathcal{I}_1^2 \cap \mathcal{S}_2^2$. Then $k_t^1 = k_t^2$, $k_t^3 = k_t^4$ and $k_t^1 \neq k_t^3$. Thus,

$$
\begin{aligned}
&\mathbb{E}\Big[h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})M^{-2}\sum_{i,j\in[M]^2}\sum_{k_{t-1}^{2,4}\in[N]^2}\omega_{t-1}^{k_{t-1}^2}\omega_{t-1}^{k_{t-1}^4} \\
&\qquad\qquad \times\widetilde{\mathcal{T}}_t^b(J^i_{k_t^1,t-1},k_{t-1}^2)\widetilde{\mathcal{T}}_t^b(J^j_{k_t^3,t-1},k_{t-1}^4)\Big|\mathcal{G}_{t-1}^N\Big] \\
&= \mathbb{E}\Big[h(\xi_t^{k_t^1},\xi_t^{k_t^2})h(\xi_t^{k_t^3},\xi_t^{k_t^4})\sum_{k_{t-1}^{1:4}\in[N]^4}\prod_{\ell=1}^{2}\beta_t^{\mathsf{BS}}(k_t^{2\ell-1},k_{t-1}^{2\ell-1})\omega_{t-1}^{k_{t-1}^{2\ell}} \\
&\qquad\qquad \times\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\Big|\mathcal{G}_{t-1}^N\Big] \\
&= \frac{1}{\Omega_{t-1}^4}\sum_{k_{t-1}^{1:4}\in[N]^4}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3,k_{t-1}^4)\big(g_{t-1}^{\otimes 2}\mathcal{M}_t^1[h]\big)^{\otimes 2}(\xi_{t-1}^{k_{t-1}^1},\xi_{t-1}^{k_{t-1}^2},\xi_{t-1}^{k_{t-1}^3},\xi_{t-1}^{k_{t-1}^4})\,.
\end{aligned}
$$

Since $b_t = 1$ implies that

$$
\mathcal{C}_{N,b,t}^2 = \mathcal{C}_{N,b,t-1}^2\frac{\Omega_{t-1}^4}{N^2}\,,
$$

and using $\mathrm{Card}(\mathcal{I}_1^1 \cap \mathcal{S}_2^2) = N(N-1)$, we get

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_1^2 \cap \mathcal{S}_2^2} h(\xi_t^{k_t^1}, \xi_t^{k_t^2}) h(\xi_t^{k_t^3}, \xi_t^{k_t^4}) \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4) \Big| \mathcal{G}_{t-1}^N \Big] \\
&= \mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_1^2 \cap \mathcal{S}_2^2} \mathbb{E}\Big[h(\xi_t^{k_t^1}, \xi_t^{k_t^2}) h(\xi_t^{k_t^3}, \xi_t^{k_t^4}) M^{-2} \sum_{i,j \in [M]^2} \sum_{k_{t-1}^{2,4} \in [N]^2} \omega_{t-1}^{k_{t-1}^2} \omega_{t-1}^{k_{t-1}^4} \\
&\qquad \times \widetilde{\mathcal{T}}_t^b(J^i_{k_t^1, t-1}, k_{t-1}^2) \widetilde{\mathcal{T}}_t^b(J^j_{k_t^3, t-1}, k_{t-1}^4) \Big| \mathcal{G}_{t-1}^N \Big] \\
&= \frac{N-1}{N} \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(g_{t-1}^{\otimes 2} \mathcal{M}_t^1[h])^2 \,,
\end{aligned}$$

which proves (B.1.36).

Let $(k_t^1, \cdots, k_t^4) \in \mathcal{I}_1^2 \cap \mathcal{S}_2^1$. Then $\boldsymbol{k}_1 = \{k_t^1\}$, $k_t^1 = k_t^2 = k_t^3 = k_t^4$ and $J^i_{k_t^1, t-1} = J^i_{k_t^3, t-1}$. Hence,

$$\begin{aligned}
&\mathbb{E}\Big[M^{-2} \sum_{i,j \in [M]^2} \sum_{k_{t-1}^{2,4} \in [N]^2} \omega_{t-1}^{k_{t-1}^2} \omega_{t-1}^{k_{t-1}^4} \widetilde{\mathcal{T}}_{t-1}^b(J^i_{k_t^1, t-1}, k_{t-1}^2) \widetilde{\mathcal{T}}_{t-1}^b(J^j_{k_t^3, t-1}, k_{t-1}^4) | \mathcal{G}_{t-1}^N \Big] \\
&= \mathbb{E}\Big[\frac{1}{M} \sum_{k_{t-1}^{1,2,4} \in [N]^3} \beta_t^{\mathsf{BS}}(k_t^1, k_{t-1}^1) \omega_{t-1}^{k_{t-1}^2} \omega_{t-1}^{k_{t-1}^4} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2) \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^4) \\
&+ \frac{M-1}{M} \sum_{k_{t-1}^{1:4} \in [N]^4} \prod_{\ell=1}^{2} \beta_t^{\mathsf{BS}}(k_t^{2\ell-1}, k_{t-1}^{2\ell-1}) \omega_{t-1}^{k_{t-1}^{2\ell}} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2) \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3, k_{t-1}^4) | \mathcal{G}_{t-1}^N \Big] \\
&= \frac{G_\infty^3}{M \Omega_{t-1}^3} \sum_{k_{t-1}^{1,2,4} \in [N]^3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2) \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^4) \\
&\qquad + \frac{G_\infty^3 (M-1)}{M \Omega_{t-1}^3} \int \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(., x) \otimes \mathbf{1}) \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x, .) \otimes \mathbf{1}) \nu(\mathrm{d}x) \,.
\end{aligned}$$

This yields

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_1^2 \cap \mathcal{S}_2^1} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4) | \mathcal{G}_{t-1}^N \Big] \\
&\leq \frac{G_\infty^3 \Omega_{t-1}}{MN} \mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1,2,4} \in [N]^3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2) \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^4) \\
&\qquad + \frac{G_\infty^3 (M-1) \Omega_{t-1}}{MN} \int \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(m_t(., x) \otimes \mathbf{1}) \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\beta_t^N(x, .) \otimes \mathbf{1}) \nu(\mathrm{d}x) \,,
\end{aligned}$$

which in turn proves (B.1.37). □

Proposition B.1.11. *Let* (**A5** : **7**) *hold. Let* $t \geq 0$. *For any* $b \in \mathcal{B}_t$

(i) If $b_t = 0$, *then for* $p \in \{2, 3\}$,

$$\lim_{N \to \infty} \mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^p} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4) \Big] = 0 \,. \tag{B.1.39}$$

(ii) If $b_t = 1$, *then*

$$\lim_{N \to \infty} \mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_1^2 \cap \mathcal{S}_2^1} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2) \widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4) \Big] = 0 \,. \tag{B.1.40}$$

Additionally, if (**A8**) *holds then the rate of convergence is* $\mathcal{O}(N^{-1})$.

Proof. We prove (i)-(ii) simultaneously by induction. Let $t = 0$ and $b \in \mathcal{B}_0$. By definition, $\widetilde{\mathcal{T}}_0^b = \mathcal{T}_0^b$ and, if $b_0 = 0$,

$$\mathcal{C}_{N,b,0}^2 \sum_{k_0^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^2} \widetilde{\mathcal{T}}_0^b(k_0^1, k_0^2)\widetilde{\mathcal{T}}_0^b(k_0^3, k_0^4) = \frac{2}{N^2(N-1)^2} \sum_{i,j \in [N]^2} \mathbb{1}_{i \neq j} = \mathcal{O}(N^{-2}), \tag{B.1.41}$$

$$\mathcal{C}_{N,b,0}^2 \sum_{k_0^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^3} \widetilde{\mathcal{T}}_0^b(k_0^1, k_0^2)\widetilde{\mathcal{T}}_0^b(k_0^3, k_0^4) = \frac{4}{N^2(N-1)^2} \sum_{i,j,k \in [N]^3} \mathbb{1}_{i \neq j}\mathbb{1}_{i \neq k} = \mathcal{O}(N^{-1}). \tag{B.1.42}$$

If $b_0 = 1$, then

$$\mathcal{C}_{N,b,0}^2 \sum_{k_0^{1:4} \in \mathcal{I}_1^2 \cap \mathcal{S}_2^1} \widetilde{\mathcal{T}}_0^b(k_0^1, k_0^2)\widetilde{\mathcal{T}}_t^b(k_0^3, k_0^4) = \frac{1}{N^2} \sum_{i,j \in [N]^2} \mathbb{1}_{i=j} = \mathcal{O}(N^{-1}). \tag{B.1.43}$$

Let $t > 0$ and assume that both (i)-(ii) hold at $t-1$. Define

$$\begin{aligned}
D_{1,b}^N &= \mathbb{E}\Big[\mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1,2,4} \in [N]^3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^4)\Big], \\
D_{2,b}^N &= \mathbb{E}\Big[\mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1:2} \in [N]^2} \mathsf{T}_{t-1,b}^{(1)}(k_{t-1}^1, k_{t-1}^2)\Big], \\
D_{3,b}^N &= \mathbb{E}\Big[\mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1:3} \in [N]^3} \mathsf{T}_{t-1,b}^{(2)}(k_{t-1}^1, k_{t-1}^2, k_{t-1}^3)\Big],
\end{aligned}$$

where $\mathsf{T}_{t-1,b}^{(1)}$ and $\mathsf{T}_{t-1,b}^{(2)}$ are defined in Proposition B.1.10. If $b_t = 0$, then by (i) in Proposition B.1.10, using that $\Omega_{t-1} \leq NG_\infty$,

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^2} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4)\Big] \\
&\qquad \leq \mathbb{E}\Big[\frac{\Omega_{t-1}^2 G_\infty^2 (M-1)}{MN(N-1)} \int \widetilde{\Upsilon}_{b,t}^{N,M}(x,y)\nu^{\otimes 2}(\mathrm{d}x, \mathrm{d}y)\Big] + 2\frac{G_\infty^4}{M} D_{2,b}^N,
\end{aligned} \tag{B.1.44}$$

and

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^3} \widetilde{\mathcal{T}}_t^b(k_t^1, k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3, k_t^4)\Big] \\
&\qquad \leq \mathbb{E}\Big[\frac{\Omega_{t-1} G_\infty^3 (M-1)(N-2)}{MN(N-1)} \int \widetilde{\Theta}_{b,t}^{N,M}(x)\nu(\mathrm{d}x)\Big] + \frac{G_\infty^4}{M} D_{3,b}^N,
\end{aligned} \tag{B.1.45}$$

where $\widetilde{\Theta}_{b,t}^{N,M}$ and $\widetilde{\Upsilon}_{b,t}^{N,M}$ are defined in Proposition B.1.10. We deal first with $D_{2,b}^N$ and $D_{3,b}^N$. If $b_{t-1} = 0$, then by definition of $\widetilde{\mathcal{T}}_{t-1}^b$ in (4.4.4),

$$\begin{aligned}
D_{2,b}^N &= \mathbb{E}\Big[\mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^2} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3, k_{t-1}^4)\Big], \\
D_{3,b}^N &= \mathbb{E}\Big[\mathcal{C}_{N,b,t-1}^2 \sum_{k_{t-1}^{1:4} \in \mathcal{I}_0^2 \cap \mathcal{S}_2^3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^1, k_{t-1}^2)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^3, k_{t-1}^4)\Big].
\end{aligned} \tag{B.1.46}$$

If $b_{t-1}=1$,

$$\begin{aligned} D^N_{2,b} &= 2\mathbb{E}\Big[\mathcal{C}^2_{N,b,t-1}\sum_{k^{1:4}_{t-1}\in\mathcal{I}^2_1\cap\mathcal{S}^1_2}\widetilde{\mathcal{T}}^b_{t-1}(k^1_{t-1},k^2_{t-1})\widetilde{\mathcal{T}}^b_{t-1}(k^3_{t-1},k^4_{t-1})\Big]\,,\\ D^N_{3,b} &= 4\mathbb{E}\Big[\mathcal{C}^2_{N,b,t-1}\sum_{k^{1:4}_{t-1}\in\mathcal{I}^2_1\cap\mathcal{S}^1_2}\widetilde{\mathcal{T}}^b_{t-1}(k^1_{t-1},k^2_{t-1})\widetilde{\mathcal{T}}^b_{t-1}(k^3_{t-1},k^4_{t-1})\Big]\,. \end{aligned} \tag{B.1.47}$$

In all cases, by the induction hypothesis we get for any $b\in\mathcal{B}_t$ with $b_t=0$,

$$\lim_{N\to\infty} D^N_{2,b}=0,\quad \lim_{N\to\infty} D^N_{3,b}=0\,.$$

Regarding the first terms in the r.h.s. of inequalities (B.1.44)-(B.1.45), they go to zero when N goes to infinity since they are, up to the constant $(M-1)/M\le 1$, the *PaRIS* counterpart of B_N (B.1.12) in the proof of Theorem 4.4.4 and are treated in the exact same way since $\sup_{N\in\mathbb{N}}\mathbb{E}[\widetilde{\mathcal{Q}}^{N,M}_{b,t}(\mathbf{1})^3]<\infty$ by Proposition B.1.12. If $b_t=1$, then, by Proposition B.1.10, using that $\Omega_{t-1}\le NG_\infty$,

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{C}^2_{N,b,t}\sum_{k^{1:4}_t\in\mathcal{I}^2_1\cap\mathcal{S}^1_2}\widetilde{\mathcal{T}}^b_t(k^1_t,k^2_t)\widetilde{\mathcal{T}}^b_t(k^3_t,k^4_t)\Big]\\ &\qquad\le \mathbb{E}\Big[\frac{\Omega_{t-1}G^3_\infty}{N}\int\widetilde{\mathcal{Q}}^{N,M}_{b,t-1}(m_t.,x)\otimes\mathbf{1})\widetilde{\mathcal{Q}}^{N,M}_{b,t-1}(\beta^N_t(x,.)\otimes 1)\nu(\mathrm{d}x)\Big]+\frac{G^4_\infty}{M}D^N_{1,b}\,. \end{aligned}$$

The first term goes to zero when N goes to infinity similarly to the case $b_t=0$. As for the second term, if $b_{t-1}=0$ then by definition of $D^N_{1,b}$

$$0\le D^N_{1,b}\le D^N_{3,b}\,,$$

and if $b_{t-1}=1$,

$$0\le D^N_{1,b}=\mathbb{E}\Big[\mathcal{C}^2_{N,b,t}\sum_{k^{1:2}_{t-1}\in[N]^2}\widetilde{\mathcal{T}}^b_{t-1}(k^1_{t-1},k^2_{t-1})^2\Big]\le D^N_{2,b}\,.$$

Hence, in both cases $D^N_{1,b}$ goes to zero by (B.1.10). This ends the proof of the first claim.

Now assume that (**A8**) holds also. We proceed by induction. At $t=0$ the rate of convergence is $\mathcal{O}(N^{-1})$ by (B.1.41)-(B.1.42) and (B.1.43). Let $t>0$ and assume that the rate of convergence in (i) and (ii) at $t-1$ is $\mathcal{O}(N^{-1})$. Assume that $b_t=0$. By the strong mixing assumption we have that

$$\beta^N_t(x,y)\le\frac{G_\infty\sigma_+}{\sigma_-\Omega_{t-1}}\,,\quad\forall(x,y)\in\mathsf{X}^2\,. \tag{B.1.48}$$

Using for example that

$$\int\widetilde{\mathcal{Q}}^{N,M}_{b,t}(m_t(.,x)\otimes\mathbf{1})\nu(\mathrm{d}x)=\int\widetilde{\mathcal{Q}}^{N,M}_{b,t}(m_t(.,x)\otimes m_t(.,y))\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y)=\widetilde{\mathcal{Q}}^{N,M}_{b,t}(\mathbf{1}\otimes\mathbf{1})\,,$$

and then bounding β^N_t using (B.1.48) we get that

$$\mathbb{E}\Big[\frac{\Omega^2_{t-1}G^2_\infty(M-1)}{MN(N-1)}\int\widetilde{\Upsilon}^{N,M}_{b,t}(x,y)\nu^{\otimes 2}(\mathrm{d}x,\mathrm{d}y)\Big]\le\frac{G^4_\infty\sigma^2_+(M-1)}{\sigma^2_-MN(N-1)}\|\widetilde{\mathcal{Q}}^{N,M}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})\|^2_2\,, \tag{B.1.49}$$

and

$$\mathbb{E}\Big[\frac{\Omega_{t-1}G^3_\infty(M-1)(N-2)}{MN(N-1)}\int\widetilde{\Theta}^{N,M}_{b,t}(x)\nu(\mathrm{d}x)\Big]\le\frac{4\sigma_+G^4_\infty(M-1)(N-2)}{\sigma_-MN(N-1)}\|\widetilde{\mathcal{Q}}^{N,M}_{b,t-1}(\mathbf{1}\otimes\mathbf{1})\|^2_2\,,$$

where both bounds are $\mathcal{O}(N^{-1})$ by Proposition B.1.12. Going back to (B.1.44)-(B.1.45), we obtain

$$\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^2} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big] \le \frac{2G_\infty^4\sigma_+^2(M-1)}{\sigma_-^2 MN(N-1)}\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1}\otimes\mathbf{1})\|_2^2 + 2\frac{G_\infty^4}{M}D_{2,b}^N\,,$$

and

$$\mathbb{E}\Big[\mathcal{C}_{N,b,t}^2 \sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^3} \widetilde{\mathcal{T}}_t^b(k_t^1,k_t^2)\widetilde{\mathcal{T}}_t^b(k_t^3,k_t^4)\Big] \le \frac{4\sigma_+G_\infty^4(M-1)(N-2)}{\sigma_- MN(N-1)}\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1}\otimes\mathbf{1})\|_2^2 + \frac{G_\infty^4}{M}D_{3,b}^N\,.$$

By (B.1.46)-(B.1.47) and the induction hypothesis, we get that $D_{2,b}^N$ and $D_{3,b}^N$ are both $\mathcal{O}(N^{-1})$. Finally, applying Proposition B.1.12 we get $\mathcal{O}(N^{-1})$ upper bounds. This ends the proof for the case $b_t = 0$. The case $b_t = 1$ follows the same steps. □

Proposition B.1.12. *Assume that* (**A5**) *holds. For all* $M > 1$, $t \in \mathbb{N}$ *and* $b \in \mathcal{B}_t$,

$$\sup_{N\in\mathbb{N}} \|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(\mathbf{1})\|_3 < \infty\,. \tag{B.1.50}$$

Proof. We proceed by induction on $t \in \mathbb{N}$. Assume for now that $N \ge 6$ and $M \ge 2$.

Since $\widetilde{\mathcal{Q}}_{b,0}^{N,M}(h) = \mathcal{Q}_{b,0}^{N,\mathsf{BS}}(h)$ for any h, the case $t = 0$ follows from Proposition B.1.4. Let $t > 0$ and assume that (B.1.50) holds at $t-1$. We only treat the case $b_t = 0$. The proof for the case $b_t = 1$ follows using the same steps. Since we have that

$$\begin{aligned}\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(\mathbf{1})\|_3^3 &= \mathbb{E}\Big[\mathcal{C}_{N,b,t}^3 \sum_{k_t^{1:6}\in[N]^6}\prod_{\ell=1}^3 \widetilde{\mathcal{T}}_t^b(k_t^{2\ell-1},k_t^{2\ell})\Big]\\ &= \mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3}\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\sum_{i^{1:3}\in[M]^3}\prod_{\ell=1}^3 \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big]\,,\end{aligned} \tag{B.1.51}$$

the proof proceeds by (i) splitting the sum in three parts with respect to the cardinal of the triplet $(i^1,i^2,i^3) \in [M]^3$, and (ii) bounding each term by $\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3$ up to a constant independent of N. Let $(i^1,i^2,i^3) \in [M]^3$. If $\mathrm{Card}(\{i^1,i^2,i^3\}) = 3$, then for all $k_t^{1:6} \in [N]^6$, $(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})_{\ell\in[1:3]}$ are mutually independent conditionally on $\mathcal{G}_{t-1}^N \vee \xi_t^{1:N}$, defined in (B.1.30), and

$$\begin{aligned}&\mathbb{E}\Big[\prod_{\ell=1}^3 \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\ &\quad= \mathbb{E}\Big[\prod_{\ell=1}^3 \mathbb{E}\Big[\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]\Big|\mathcal{G}_{t-1}^N\Big]\\ &\quad= \mathbb{E}\Big[\prod_{\ell=1}^3\sum_{k_{t-1}^{2\ell-1:2\ell}\in[N]^2}\beta_t^{\mathsf{BS}}(k_t^{2\ell-1},k_{t-1}^{2\ell-1})\beta_t^{\mathsf{BS}}(k_t^{2\ell},k_{t-1}^{2\ell})\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\Big|\mathcal{G}_{t-1}^N\Big]\\ &\quad= \sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^3\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\mathbb{E}\Big[\prod_{n=1}^6\beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\Big|\,\mathcal{F}_{t-1}^N\Big]\,.\end{aligned}$$

Hence, by Proposition B.1.6, Lemma B.1.5 and similarly to the proof of Proposition B.1.4,

$$\begin{aligned}
&\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\mathbb{E}\Big[\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\
&\quad=\sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\mathbb{E}\Big[\prod_{n=1}^{6}\beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\Big|\,\mathcal{F}_{t-1}^N\Big]\\
&\quad\le\sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\sum_{p=2}^{6}\sum_{k_t^{1:6}\in\mathcal{I}_0^3\cap\mathcal{S}_3^p}\frac{G_\infty^p}{\Omega_{t-1}^p}\\
&\quad\lesssim\Big(\sum_{p=2}^{6}\frac{N^p}{\Omega_{t-1}^p}\Big)\sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\,,
\end{aligned}$$

where $\lesssim$ means less than or equal up to a multiplicative constant independent of N. Consequently,

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3}\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\sum_{i^{1:3}\in[M]^3}\mathbb{1}_{i^1\neq i^2\neq i^3}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\
&\qquad=\frac{\mathcal{C}_{N,b,t}^3}{M^3}\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\sum_{i^{1:3}\in[M]^3}\mathbb{1}_{i^1\neq i^2\neq i^3}\mathbb{E}\Big[\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\Big]\\
&\qquad\lesssim\frac{M(M-1)(M-2)}{M^3}\Big(\sum_{p=2}^{6}\frac{N^p}{\Omega_{t-1}^p}\Big)\mathcal{C}_{N,b,t}^3\sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\,,
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{C}_{N,b,t}^3\Big(\sum_{p=2}^{6}\frac{N^p}{\Omega_{t-1}^p}\Big)&=\mathcal{C}_{N,b,t-1}^3\Big(\sum_{p=2}^{6}\frac{N^p\Omega_{t-1}^6}{\Omega_{t-1}^pN^3(N-1)^3}\Big)\\
&\le\mathcal{C}_{N,b,t-1}^3\Big(\sum_{p=2}^{6}\frac{G_\infty^{6-p}N^6}{N^3(N-1)^3}\Big)\lesssim\mathcal{C}_{N,b,t-1}^3\,.
\end{aligned}$$

Therefore,

$$\mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3}\sum_{k_t^{1:6}\in[N]^6}\sum_{i^{1:3}\in[M]^3}\mathbb{1}_{i^1\neq i^2\neq i^3}\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big]\lesssim\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3\,.\tag{B.1.52}$$

Assume now that $\mathrm{Card}(\{i^1,i^2,i^3\})=2$ and that $i^1=i^2$. For all $(k_t^1,\cdots,k_t^6)\in[N]^6$, conditionally on $\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}$, $\widetilde{\mathcal{T}}_{t-1}^b(J^{i^3}_{k_t^5,t-1},J^{i^3}_{k_t^6,t-1})$ is independent from $\widetilde{\mathcal{T}}_{t-1}^b(J^{i^1}_{k_t^1,t-1},J^{i^1}_{k_t^2,t-1}),\widetilde{\mathcal{T}}_{t-1}^b(J^{i^2}_{k_t^3,t-1},J^{i^2}_{k_t^4,t-1})$, hence, using (B.1.5),

$$\begin{aligned}
&\sum_{k_t^{1:6}\in\mathcal{I}_0^3}\mathbb{E}\Big[\prod_{\ell=1}^{3}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]\qquad\text{(B.1.53)}\\
&\quad=\Big\{\sum_{k_t^{1:4}\in\mathcal{I}_0^2}\mathbb{E}\Big[\prod_{\ell=1}^{2}\widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1},J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]\Big\}\\
&\qquad\qquad\times\Big\{\sum_{k_t^{5:6}\in\mathcal{I}_0^1}\mathbb{E}\Big[\widetilde{\mathcal{T}}_{t-1}^b(J^{i^3}_{k_t^5,t-1},J^{i^3}_{k_t^6,t-1})\Big|\mathcal{G}_{t-1}^N\vee\xi_t^{1:N}\Big]\Big\}\\
&\quad=(F_{2,b}^N+F_{3,b}^N+F_{4,b}^N)F_{1,b}^N\,,
\end{aligned}$$

with

$$F^N_{1,b} := \sum_{k_t^{5:6}\in\mathcal{I}_0^1} \mathbb{E}\Big[\widetilde{\mathcal{T}}^b_{t-1}(J^{i^3}_{k_t^5,t-1}, J^{i^3}_{k_t^6,t-1})\Big|\mathcal{G}^N_{t-1}\vee\xi_t^{1:N}\Big]$$
$$= \sum_{k_t^{5:6}\in[N]^2} \mathbb{1}_{k_t^5\neq k_t^6} \sum_{k_{t-1}^{5:6}\in[N]^2} \beta_t^{\mathsf{BS}}(k_t^5,k_{t-1}^5)\beta_t^{\mathsf{BS}}(k_t^6,k_{t-1}^6)\widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^5,k_{t-1}^6)\,,$$

and

$$\sum_{k_t^{1:4}\in\mathcal{I}_0^2} \mathbb{E}\Big[\prod_{\ell=1}^2 \widetilde{\mathcal{T}}^b_{t-1}(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}^N_{t-1}\vee\xi_t^{1:N}\Big]$$
$$= \sum_{p=2}^4 \sum_{k_t^{1:4}\in\mathcal{I}_0^2\cap\mathcal{S}_2^p} \mathbb{E}\Big[\prod_{\ell=1}^2 \widetilde{\mathcal{T}}^b_{t-1}(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big|\mathcal{G}^N_{t-1}\vee\xi_t^{1:N}\Big]$$
$$= F^N_{2,b} + F^N_{3,b} + F^N_{4,b}\,,$$

where

$$F^N_{2,b} := \sum_{k_t^{1:2}\in[N]^2} \mathbb{1}_{k_t^2\neq k_t^2} \sum_{k_{t-1}^{1:2}\in[N]^2} \beta_t^{\mathsf{BS}}(k_t^1,k_{t-1}^1)\beta_t^{\mathsf{BS}}(k_t^2,k_{t-1}^2)\mathsf{T}^{(1)}_{t-1,b}(k_{t-1}^1,k_{t-1}^2)$$

$$F^N_{3,b} := \sum_{k_t^{1:3}\in[N]^3} \mathbb{1}_{k_t^1\neq k_t^2\neq k_t^3} \sum_{k_{t-1}^{1:3}\in[N]^3} \prod_{\ell=1}^3 \beta_t^{\mathsf{BS}}(k_t^\ell,k_{t-1}^\ell)\mathsf{T}^{(2)}_{t-1,b}(k_{t-1}^1,k_{t-1}^2,k_{t-1}^3)$$

$$F^N_{4,b} := \sum_{k_t^{1:4}\in[N]^4} \mathbb{1}_{k_t^1\neq k_t^2\neq k_t^3\neq k_t^4} \sum_{k_{t-1}^{1:4}\in[N]^4} \prod_{\ell=1}^4 \beta_t^{\mathsf{BS}}(k_t^\ell,k_{t-1}^\ell)\widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^1,k_{t-1}^2)\widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^3,k_{t-1}^4)\,,$$

where $\mathsf{T}^{(1)}_{t-1,b}$ and $\mathsf{T}^{(2)}_{t-1,b}$ are defined in Proposition B.1.10. We now upperbound each $\mathbb{E}[F^N_{1,b}F^N_{i,b}|\mathcal{G}^N_{t-1}]$ for $i\in[2:4]$.

Consider first the case $\mathbb{E}[F^N_{1,b}F^N_{4,b}|\mathcal{G}^N_{t-1}]$. Let $S_6 := \{k^{1:6}\in[N]^6 : k_t^1\neq k_t^2\neq k_t^3\neq k_t^4, k_t^5\neq k_t^6\}$. Then, $S_6\subset(\mathcal{I}_0^3\cap\mathcal{S}_3^4)\sqcup(\mathcal{I}_0^3\cap\mathcal{S}_3^5)\sqcup(\mathcal{I}_0^3\cap\mathcal{S}_3^6)$ and

$$\mathbb{E}[F^N_{1,b}F^N_{4,b}|\mathcal{G}^N_{t-1}] = \sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^3 \widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell}) \sum_{k_t^{1:6}\in\mathsf{S}_6} \mathbb{E}\Big[\prod_{n=1}^6 \beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\Big|\mathcal{G}^N_{t-1}\Big]$$
$$\leq \sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^3 \widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell}) \sum_{p=4}^6 \sum_{k_t^{1:6}\in\mathcal{I}_0^3\cap\mathcal{S}_3^p} \mathbb{E}\Big[\prod_{n=1}^6 \beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\Big|\mathcal{G}^N_{t-1}\Big]$$
$$\lesssim \Big(\sum_{p=4}^6 \frac{N^pG_\infty^p}{\Omega_{t-1}^p}\Big) \sum_{k_{t-1}^{1:6}\in[N]^6}\prod_{\ell=1}^3 \widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^{2\ell-1},k_{t-1}^{2\ell})\,, \tag{B.1.54}$$

by Proposition B.1.6 and Lemma B.1.5.

Consider now the case $\mathbb{E}[F^N_{1,b}F^N_{3,b}|\mathcal{G}^N_{t-1}]$ and define $\mathsf{S}_5 := \{k_t^{1:5}\in[N]^5 : k_t^1\neq k_t^2\neq k_t^3, k_t^4\neq k_t^5\}$. Then,

$$\mathbb{E}[F^N_{1,b}F^N_{3,b}|\mathcal{G}^N_{t-1}]$$
$$= \sum_{k_{t-1}^{1:5}\in[N]^5} \mathsf{T}^{(2)}_{t-1,b}(k_{t-1}^1,k_{t-1}^2,k_{t-1}^3)\widetilde{\mathcal{T}}^b_{t-1}(k_{t-1}^4,k_{t-1}^5) \sum_{k_t^{1:5}\in\mathsf{S}_5} \mathbb{E}\Big[\prod_{n=1}^5 \beta_t^{\mathsf{BS}}(k_t^n,k_{t-1}^n)\Big|\mathcal{G}^N_{t-1}\Big]\,.$$

Proceeding similarly as in Proposition B.1.6 and Lemma B.1.5, it can be shown that

$$\sum_{k_t^{1:5}\in \mathsf{S}_5} \mathbb{E}\Big[\prod_{n=1}^{5} \beta_t^{\mathsf{BS}}(k_t^n, k_{t-1}^n)\Big|\mathcal{G}_{t-1}^N\Big] \le \frac{N^5 G_\infty^5}{\Omega_{t-1}^5} + \frac{2N^4 G_\infty^4}{\Omega_{t-1}^4} + \frac{N^3 G_\infty^3}{\Omega_{t-1}^3}\,.$$

Finally, by noting that

$$\sum_{k_{t-1}^{1:5}\in[N]^5} \mathsf{T}_{t-1,b}^{(2)}(k_{t-1}^1, k_{t-1}^2, k_{t-1}^3)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^4, k_{t-1}^5) \le 4 \sum_{k_{t-1}^{1:6}\in[N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_{t-1}^{2\ell})$$

we get

$$\begin{aligned}
&\mathbb{E}[F_{1,b}^N F_{3,b}^N|\mathcal{G}_{t-1}^N] \\
&\quad\le \Big(\frac{N^5 G_\infty^5}{\Omega_{t-1}^5} + \frac{2N^4 G_\infty^4}{\Omega_{t-1}^4} + \frac{N^3 G_\infty^3}{\Omega_{t-1}^3}\Big) \sum_{k_{t-1}^{1:5}\in[N]^5} \mathsf{T}_{t-1,b}^{(2)}(k_{t-1}^1, k_{t-1}^2, k_{t-1}^3)\widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^4, k_{t-1}^5) \\
&\quad\le 4\Big(\frac{N^5 G_\infty^5}{\Omega_{t-1}^5} + \frac{2N^4 G_\infty^4}{\Omega_{t-1}^4} + \frac{N^3 G_\infty^3}{\Omega_{t-1}^3}\Big) \sum_{k_{t-1}^{1:6}\in[N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_{t-1}^{2\ell})\,. \qquad \text{(B.1.55)}
\end{aligned}$$

Consider now the case $\mathbb{E}[F_{1,b}^N F_{2,b}^N|\mathcal{G}_{t-1}^N]$. In the same way as for the previous case, we obtain

$$\mathbb{E}[F_{1,b}^N F_{2,b}^N|\mathcal{G}_{t-1}^N] \le 2\left(\frac{N^4 G_\infty^4}{\Omega_{t-1}^4} + \frac{2N^3 G_\infty^3}{\Omega_{t-1}^3} + \frac{N^2 G_\infty^2}{\Omega_{t-1}^2}\right) \sum_{k_{t-1}^{1:6}\in[N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_{t-1}^{2\ell})\,. \qquad \text{(B.1.56)}$$

Thus, combining (B.1.56), (B.1.55) and (B.1.54), we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{C}_{N,b,t}^3 \sum_{k_t^{1:6}\in[N]^6} \sum_{i^{1:3}\in[M]^3} \mathbb{1}_{i^1=i^2,i^3\notin\{i^1,i^2\}} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^{2\ell-1},t-1}^{i^\ell}, J_{k_t^{2\ell},t-1}^{i^\ell})\Big] \\
&\qquad\lesssim \mathbb{E}\Big[\Big(\sum_{p=2}^{6} \frac{\Omega_{t-1}^6 N^p G_\infty^p}{\Omega_{t-1}^p N^3 (N-1)^3}\Big)\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})^3\Big] \lesssim \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3\,.
\end{aligned}$$

The other triplets (i^1, i^2, i^3) for which $\mathrm{Card}(\{i^1, i^2, i^3\}) = 2$ are handled similarly and are bounded by $\|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3$ up to some multiplicative constant independent of N. We finally obtain

$$\mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3} \sum_{k_t^{1:6}\in[N]^6} \sum_{i^{1:3}\in[M]^3} \mathbb{1}_{\mathrm{Card}(\{i^1,i^2,i^3\})=2} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J_{k_t^{2\ell-1},t-1}^{i^\ell}, J_{k_t^{2\ell},t-1}^{i^\ell})\Big] \lesssim \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3\,. \qquad \text{(B.1.57)}$$

It now remains to treat the case $\mathrm{Card}(\{i^1, i^2, i^3\}) = 1$. Let $(k^1, \cdots, k^d) \in [N]^d$. Denote by $\mathsf{Pos}_d(k^{1:d})$ the set of elements in $[N]^d$ with positions of equal elements similar to those of the equal elements in $k^{1:d}$. For example,

$$\begin{aligned}
\mathsf{Pos}_3((2,1,1)) &= \{(j,i,i) \mid (i,j) \in [N]^2, i \neq j\}\,, \\
\mathsf{Pos}_3((1,1,2)) &= \{(i,i,j) \mid (i,j) \in [N]^2, i \neq j\}\,, \\
\mathsf{Pos}_3((1,2,3)) &= \{(i,j,k) \mid (i,j,k) \in [N]^3, \mathrm{Card}(\{i,j,k\}) = 3\}\,.
\end{aligned}$$

Let $p \in [2:6]$ and $(k_t^1, \cdots, k_t^6) \in \mathcal{I}_0^3 \cap \mathcal{S}_3^p$. Without loss of generality, assume that the first p elements of $k_t^{1:6}$ are all different.

$$\begin{aligned}
&\mathbb{E}\Big[\prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1}) \Big| \mathcal{G}_{t-1}^N\Big] \\
&\qquad = \sum_{k_{t-1}^{1:6} \in \mathsf{Pos}_6(k_t^{1:6})} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_t^{2\ell}) \mathbb{E}\Big[\prod_{n=1}^{p} \beta_t^{\mathsf{BS}}(k_t^n, k_{t-1}^n) \Big| \mathcal{G}_{t-1}^N\Big] \\
&\qquad \leq \frac{G_\infty^p}{\Omega_{t-1}^p} \sum_{k_{t-1}^{1:6} \in \mathsf{Pos}_6(k_t^{1:6})} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_t^{2\ell}) \\
&\qquad \leq \frac{G_\infty^p}{\Omega_{t-1}^p} \sum_{k_{t-1}^{1:6} \in [N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_t^{2\ell}) \,.
\end{aligned}$$

Consequently, by Lemma B.1.5,

$$\begin{aligned}
&\sum_{k_t^{1:6} \in \mathcal{I}_0^3 \cap \mathcal{S}_3^p} \mathbb{E}\Big[\prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1}) \Big| \mathcal{G}_{t-1}^N\Big] \\
&\qquad \leq \sum_{k_t^{1:6} \in \mathcal{I}_0^3 \cap \mathcal{S}_3^p} \frac{G_\infty^p}{\Omega_{t-1}^p} \sum_{k_{t-1}^{1:6} \in [N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_t^{2\ell}) \\
&\qquad \lesssim \frac{N^p G_\infty^p}{\Omega_{t-1}^p} \sum_{k_{t-1}^{1:6} \in [N]^6} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(k_{t-1}^{2\ell-1}, k_t^{2\ell}) \,,
\end{aligned}$$

and,

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3} \sum_{k_t^{1:6} \in [N]^6} \sum_{i^{1:3} \in [M]^3} \mathbb{1}_{\mathrm{Card}(\{i^1,i^2,i^3\})=1} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big] \\
&\qquad \lesssim \mathbb{E}\Big[\Big(\sum_{p=2}^{6} \frac{\Omega_{t-1}^6 N^p G_\infty^p}{\Omega_{t-1}^p N^3 (N-1)^3}\Big) \widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})^3\Big] \lesssim \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3 \,. \quad \text{(B.1.58)}
\end{aligned}$$

Finally, combining (B.1.52), (B.1.57) and (B.1.58) we get

$$\|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(\mathbf{1})\|_3^3 = \mathbb{E}\Big[\frac{\mathcal{C}_{N,b,t}^3}{M^3} \sum_{k_t^{1:6} \in [N]^6} \sum_{i^{1:3} \in [M]^3} \prod_{\ell=1}^{3} \widetilde{\mathcal{T}}_{t-1}^b(J^{i^\ell}_{k_t^{2\ell-1},t-1}, J^{i^\ell}_{k_t^{2\ell},t-1})\Big] \lesssim \|\widetilde{\mathcal{Q}}_{b,t-1}^{N,M}(\mathbf{1})\|_3^3 \,,$$

and hence $\sup_{N \geq 6} \|\widetilde{\mathcal{Q}}_{b,t}^{N,M}(\mathbf{1})\|_3 < \infty$ by the induction hypothesis. This ends the proof for the case $b_t = 0$.

If $M = 2$, then (B.1.52) is equal to 0 and (B.1.57), (B.1.58) remain the same. The result then follows. If $N < 6$ then it suffices to truncate the sums over p to obtain the result. □

B.2 Further algorithmic details

B.2.1 Alternative expression of the genealogy tracing variance estimator

The expression of the CLE estimator (4.3.11) provided in the main chapter is different from the expression of the estimator appearing in Olsson and Douc (2019). We show here that these are

two expressions of the same quantity. Note first that

$$\left(\frac{1}{N}\sum_{i=1}^{N} h(\xi_t^i) - \eta_t^N(h)\right)^2 = 0$$
$$= N^{-2} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i = \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}$$
$$+ N^{-2} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i \neq \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}.$$

On the other hand,

$$\sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i = \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}$$
$$= \sum_{k=1}^{N} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i = \mathrm{E}_{t,0}^j = k} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}$$
$$= \sum_{k=1}^{N} \left(\sum_{i=1}^{N} \mathbb{1}_{\mathrm{E}_{t,0}^i = k} \{h(\xi_t^i) - \eta_t^N(h)\}\right)^2 .$$

Thus,

$$\mathcal{V}_{\eta,t}^N(h) = -N^{-1} \sum_{i,j\in[N]^2} \mathbb{1}_{\mathrm{E}_{t,0}^i \neq \mathrm{E}_{t,0}^j} \{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}$$
$$= N^{-1} \sum_{k=1}^{N} \left(\sum_{i=1}^{N} \mathbb{1}_{\mathrm{E}_{t,0}^i = k} \{h(\xi_t^i) - \eta_t^N(h)\}\right)^2 .$$

where the expression in the second line is that of Olsson and Douc (2019). By a similar reasoning, (4.3.12) is also equivalent to their estimator.

B.2.2 Variance estimators for the predictor and filter

The asymptotic variances of the predictor and filter (4.3.8)-(4.3.9) can be expressed using $\mathcal{V}_{\gamma,t}^\infty$. Indeed,

$$\frac{\mathcal{V}_{\gamma,t}^\infty(h - \eta_t(h))}{\gamma_t(\mathbf{1})^2} = \sum_{s=0}^{t} \left\{ \frac{\gamma_s(\mathbf{1})\gamma_s(\overline{\mathbf{Q}}_{s+1:t}[h - \eta_t(h)]^2)}{\gamma_t(\mathbf{1})^2} - \eta_t(h - \eta_t(h))^2 \right\} = \mathcal{V}_{\eta,t}^\infty(h)\,,$$

and using that

$$\frac{\gamma_t(g_t\{h - \phi_t(h)\})}{\gamma_{t+1}(\mathbf{1})} = \frac{\gamma_t(g_t h)}{\gamma_{t+1}(\mathbf{1})} - \phi_t(h) = 0\,,$$

we get

$$\mathcal{V}_{\phi,t}^\infty(h) = \frac{\mathcal{V}_{\gamma,t}^\infty(g_t\{h - \phi_t(h)\})}{\gamma_{t+1}(\mathbf{1})^2}\,. \tag{B.2.1}$$

Then, replacing $\gamma_t(h)$ and $\phi_t(h)$ by their empirical approximations $\gamma_t^N(h)$ and $\phi_t^N(h)$, we obtain

$$\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}(h) := \frac{-N^t}{(N-1)^{t+1}} \sum_{i,j\in[N]^2} \mathcal{T}_t^{\mathbf{0}}(i,j)\{h(\xi_t^i) - \eta_t^N(h)\}\{h(\xi_t^j) - \eta_t^N(h)\}\,, \tag{B.2.2}$$

$$\mathcal{V}_{\phi,t}^{N,\mathsf{BS}}(h) := \frac{-N^{t+2}}{(N-1)^{t+1}} \sum_{i,j\in[N]^2} \omega_t^i\omega_t^j \mathcal{T}_t^{\mathbf{0}}(i,j)\{h(\xi_t^i) - \phi_t^N(h)\}\{h(\xi_t^j) - \phi_t^N(h)\}\,. \tag{B.2.3}$$

As a consequence of Theorem 4.4.7, these estimators are also weakly consistent.

Corollary B.2.1. *Let* (**A4** : **7**) *hold. For any* $h \in \mathsf{F}(\mathcal{X})$, $\mathcal{V}_{\eta,t}^{N,\mathsf{BS}}(h) \xrightarrow{\mathbb{P}} \mathcal{V}_{\eta,t}^{\infty}(h)$ *and* $\mathcal{V}_{\phi,t}^{N,\mathsf{BS}}(h) \xrightarrow{\mathbb{P}} \mathcal{V}_{\phi,t}^{\infty}(h)$.

Proof. It suffices to note that $\mathcal{Q}_{b,t}^{N,\mathsf{BS}}$ and $\mathcal{Q}_{b,t}$ are bilinear, that $\eta_t^N(h) \xrightarrow{\mathbb{P}} \eta_t(h)$ and to apply Theorem 4.4.4 again:

$$\begin{aligned}
&\mathcal{Q}_{b,t}^{N,\mathsf{BS}}\left(\{h-\eta_t^N(h)\}^{\otimes 2}\right) \\
&\quad = \mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h^{\otimes 2}) - \eta_t^N(h)\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h\otimes \mathbf{1}) - \eta_t^N(h)\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1}\otimes h) + \eta_t^N(h)^2\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(\mathbf{1}) \\
&\quad \xrightarrow{\mathbb{P}} \mathcal{Q}_{b,t}\left(\{h-\eta_t(h)\}^{\otimes 2}\right).
\end{aligned}$$

Hence, $\mathcal{V}_{\gamma,t}^{N,\mathsf{BS}}(h-\eta_t^N(h)) \xrightarrow{\mathbb{P}} \mathcal{V}_{\gamma,t}^{\infty}(h-\eta_t(h))$ and using the fact that $\gamma_t^N(\mathbf{1})^2 \xrightarrow{\mathbb{P}} \gamma_t(\mathbf{1})^2$ we get the consistency for the predictive measures. The remaining limit is a straightforward application. □

Algorithm 5: Update at step $t+1$ of the variance estimator for the predictor

Input: $\widetilde{\omega}_t^{1:N}, \xi_t^{1:N}, \xi_{t+1}^{1:N}$ and $\boldsymbol{\mathcal{T}}_t^{\mathbf{0}}$
Output: $-N^t/(N-1)^{t+1}\sum_{i,j\in[N]^2}\boldsymbol{\mathcal{Q}}_{i,j}$, $\boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}}$.

Compute $\boldsymbol{\beta}_{t+1}^{\mathsf{BS}}$
if *PaRIS* **then**
  for $k \in [1:N]$ **do**
    Sample $J_{k,t}^{1:M} \overset{\text{iid}}{\sim} \beta_{t+1}^{\mathsf{BS}}(k,.)$
  for $(k,\ell) \in [1:N]^2$ **do**
    Set $\mathcal{T}_{t+1}^{\mathbf{0}}(k,\ell) = \mathbb{1}_{k\neq\ell}\sum_{i=1}^{M}\mathcal{T}_t^{\mathbf{0}}(J_{k,t}^i, J_{\ell,t}^i)/M$
else
  Compute $\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} = \boldsymbol{\beta}_{t+1}^{\mathsf{BS}}\boldsymbol{\mathcal{T}}_t^{\mathbf{0}}\boldsymbol{\beta}_{t+1}^{\mathsf{BS}\prime}$.
  Set $\boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}} = \overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}} - \mathrm{Diag}(\overline{\boldsymbol{\mathcal{T}}}_{t+1}^{\mathbf{0}})$.
Compute $\boldsymbol{\mathcal{Q}} = \boldsymbol{\mathcal{T}}_{t+1}^{\mathbf{0}} \odot [\{h(\xi_{t+1}^{1:N}) - \eta_{t+1}^N(h)\}\{h(\xi_{t+1}^{1:N}) - \eta_{t+1}^N(h)\}^\top]$.

B.2.3 GT term by term estimator of the asymptotic variance

In this section we derive the GT counterpart of the term by term estimator (4.4.17). Define for all $t > 0$

$$\mathcal{T}_{b,t}^{\mathsf{GT}}(K_t^1, K_t^2) := \mathbb{E}_{\mathsf{GT}}[\mathrm{I}_{b,t}(K_{0:t}^1, K_{0:t}^2)|\mathcal{F}_t^N, K_t^1, K_t^2], \tag{B.2.4}$$

and $\mathcal{T}_{b,0}^{\mathsf{GT}}(K_0^1, K_0^2) := \mathbb{1}_{K_0^1\neq K_0^2, b_0=0} + \mathbb{1}_{K_0^1=K_0^2, b_0=1}$.

By the tower property and the definition of $\mathbb{E}_{\mathsf{GT}}[\cdot|\mathcal{F}_{t-1}^N]$, for all $(k,\ell) \in [N]^2$ and $t > 0$, if $b_t = 0$,

$$\mathcal{T}_{b,t}^{\mathsf{GT}}(k,\ell) = \mathbb{1}_{k\neq\ell}\sum_{i,j\in[N]^2}\mathbb{1}_{A_t^k=i, A_t^\ell=j}\mathcal{T}_{b,t-1}^{\mathsf{GT}}(i,j) = \mathbb{1}_{k\neq\ell}\mathcal{T}_{b,t-1}^{\mathsf{GT}}(A_{t-1}^k, A_{t-1}^\ell),$$

and if $b_t = 1$,

$$\mathcal{T}_{b,t}^{\mathsf{GT}}(k,\ell) = \mathbb{1}_{k=\ell}\sum_{i,j\in[N]^2}\mathbb{1}_{A_{t-1}^k=i}\omega_{t-1}^j\mathcal{T}_{b,t-1}^{\mathsf{GT}}(i,j) = \mathbb{1}_{k=\ell}\sum_{i=1}^{N}\omega_{t-1}^j\mathcal{T}_{b,t-1}^{\mathsf{GT}}(A_{t-1}^k, j).$$

Similarly to BS, we have by the tower property

$$\mathcal{Q}_{b,t}^{N,\mathsf{BS}}(h) = \prod_{s=0}^{t} N^{b_s} \left(\frac{N}{N-1}\right)^{1-b_s} \frac{\gamma_t^N(\mathbf{1})^2}{N^2} \sum_{k,\ell\in[N]^2} \mathcal{T}_{b,t}^{\mathsf{GT}}(k,\ell) h(\xi_t^k, \xi_t^\ell)\,, \tag{B.2.5}$$

and the term by term estimator is thus

$$\overline{\mathcal{V}}_{\gamma,t}^{N,\mathsf{GT}}(h) = \frac{N^{t-1}\gamma_t^N(\mathbf{1})^2}{(N-1)^t} \sum_{k,\ell\in[N]^2} \left\{ S_t^{\mathsf{GT}}(k,\ell) - \frac{t+1}{N-1} \mathcal{T}_{\mathbf{0},t}^{\mathsf{GT}}(k,\ell) \right\} h(\xi_t^k) h(\xi_t^\ell)\,, \tag{B.2.6}$$

where S_t^{GT} is such that for all $(k,\ell) \in [N]^2$,

$$S_t^{\mathsf{GT}}(k,\ell) = \sum_{s=0}^{t} \mathcal{T}_{e_s,t}^{\mathsf{GT}}(k,\ell) = \mathbb{1}_{k=\ell} \sum_{i=1}^{N} \omega_{t-1}^j \mathcal{T}_{\mathbf{0},t-1}^{\mathsf{GT}}(A_{t-1}^k, j) + \mathbb{1}_{k\neq\ell} S_{t-1}^{\mathsf{GT}}(A_{t-1}^k, A_{t-1}^\ell)\,,$$

which shows that (B.2.6) is also updated online in a rather simple way by propagating the matrices $\boldsymbol{S}_t^{\mathsf{GT}}$ and $\boldsymbol{\mathcal{T}}_{\mathbf{0},t}^{\mathsf{GT}}$.

B.3 Technical results

Theorem B.3.1 (**Generalized dominated convergence theorem**). *Let $(f_N)_{N\in\mathbb{N}}$ be a sequence of $\mathcal{X}$-measurable functions and $(g_N)_{N\in\mathbb{N}}$ a sequence of non-negative $\mathcal{X}$-measurable functions. Assume that the following assumptions hold.*

(i) There exists $C > 0$ such that $|f_N(x)| \leq C g_N(x)$ for all $N \in \mathbb{N}$ and $x \in \mathsf{X}$.

(ii) $(g_N)_{N\in\mathbb{N}}$ converges pointwise to g and $\lim_{N\to\infty} \int g_N \mathrm{d}\nu = \int g \mathrm{d}\nu < \infty$.

(iii) $(f_N)_{N\in\mathbb{N}}$ converges pointwise to f.

Then, f is ν-integrable and $\lim_{N\to\infty} \int f_N \mathrm{d}\nu = \int f \mathrm{d}\nu$.

Proof. The proof can be found in Royden and Fitzpatrick (1988). □

Theorem B.3.2 and Lemma B.3.3 are borrowed from Olsson and Westerborn (2017) and Douc et al. (2011a) respectively.

Theorem B.3.2. *Assume that (**A5** : **7**) hold. Then, for all $s \in \mathbb{N}$, $h_s \in \mathsf{F}(\mathcal{X}^{\otimes s+1})$ and $(f_s, \tilde{f}_s) \in \mathsf{F}(\mathcal{X})^2$, there exist constants $(C_s, \widetilde{C}_s) \in (\mathbb{R}_+^*)^2$, depending on h_s, f_s, and $\tilde{f}_s$, such that for all $N \in \mathbb{N}_{>0}$ and all $\varepsilon \in \mathbb{R}_+^*$,*

$$\mathbb{P}\left(\left| N^{-1} \sum_{i=1}^{N} \widetilde{\omega}_s^i \{ (\mathbf{T}_s^N[h_s] f_s)(\xi_s^i) + \tilde{f}_s(\xi_s^i) \} - \eta_s(\mathbf{T}_s[h_s] f_s + \tilde{f}_s) \right| \geq \varepsilon \right) \leq C_s \exp\left(-\widetilde{C}_s N \varepsilon^2\right)\,, \tag{B.3.1}$$

$$\mathbb{P}\left(\left| \sum_{i=1}^{N} \omega_s^i \left\{ \left(\mathbf{T}_s^N[h_s] f_s\right)(\xi_s^i) + \tilde{f}_s\left(\xi_s^i\right) \right\} - \phi_s\left(\mathbf{T}_s[h_s] f_s + \tilde{f}_s\right) \right| \geq \varepsilon \right) \leq C_s \exp\left(-\widetilde{C}_s N \varepsilon^2\right)\,. \tag{B.3.2}$$

Lemma B.3.3. *Assume that a_N, b_N and b are random variables defined on the same probability space such that there exist positive constants $\beta, B_1, C_1, B_2, C_2$ and M satisfying the following assumptions.*

- *$|a_N / b_N| \leq M$, $\mathbb{P}-a.s.$ and $b \geq \beta$.*

- *For all* $\varepsilon > 0$ *and all* $N \geq 1, \mathbb{P}\left(|b_N - b| > \varepsilon\right) \leq B_1 \exp(-C_1 N\varepsilon^2)$.
- *For all* $\varepsilon > 0$ *and all* $N \geq 1, \mathbb{P}\left(|a_N| > \varepsilon\right) \leq B_2 \exp(-C_2 N\varepsilon^2)$.

Then, there exist two positive constants B_3, C_3 *such that*

$$\mathbb{P}\left(\left|\frac{a_N}{b_N}\right| > \varepsilon\right) \leq B_3 \exp(-C_3 N\varepsilon^2)\,.$$

B.4 Asymptotic variance of the joint predictive distribution

In this section we provide some intuition on (4.3.7). Let $h \in \mathsf{F}(\mathcal{X})$. By the law of total variance,

$$\mathbb{V}[\gamma_{t+1}^N(h)] = \mathbb{V}[\mathbb{E}[\gamma_{t+1}^N(h) \mid \mathcal{F}_t^N]] + \mathbb{E}[\mathbb{V}[\gamma_{t+1}^N(h) \mid \mathcal{F}_t^N]]\,. \tag{B.4.1}$$

As $\gamma_{t+1}^N(\mathbf{1})$ is $\mathcal{F}_t^N$-measurable and the particles at time $t+1$ are i.i.d conditionally on $\mathcal{F}_t^N$, we have that

$$\begin{aligned}\mathbb{E}[\gamma_{t+1}^N(h) \mid \mathcal{F}_t^N] &= \gamma_{t+1}^N(\mathbf{1}) \sum_{i=1}^N \frac{\widetilde{\omega}_t^i}{\Omega_t} M_{t+1}[h](\xi_t^i) \\ &= \gamma_t^N(\mathbf{1}) N^{-1} \Omega_t \sum_{i=1}^N \frac{\widetilde{\omega}_t^i}{\Omega_t} M_{t+1}[h](\xi_t^i) = \gamma_t^N(\mathbf{Q}_{t+1}[h])\,. \end{aligned} \tag{B.4.2}$$

On the other hand,

$$\mathbb{V}[\gamma_{t+1}^N[h] \mid \mathcal{F}_t^N] = \gamma_{t+1}^N(\mathbf{1})^2 \mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N h(\xi_{t+1}^i) \middle| \mathcal{F}_t^N\right] = N^{-1}\gamma_{t+1}^N(\mathbf{1})^2 \mathbb{V}_{\phi_t^N M_{t+1}}[h(\xi_{t+1})]$$

where

$$\mathbb{V}_{\phi_t^N M_{t+1}}[h(\xi_{t+1})] = \phi_t^N M_{t+1}(\{h - \phi_t^N M_{t+1}(h)\}^2)\,.$$

Therefore,

$$\begin{aligned}\mathbb{V}[\gamma_{t+1}^N(h) \mid \mathcal{F}_t^N] &= N^{-2}\gamma_t^N(\mathbf{1})^2 \Omega_t \eta_t^N(\mathbf{Q}_{t+1}[\{h - \phi_t^N M_{t+1}(h)\}^2]) \\ &= N^{-1}\gamma_{t+1}^N(\mathbf{1})\gamma_t^N(\mathbf{Q}_{t+1}[\{h - \phi_t^N M_{t+1}(h)\}^2])\,. \end{aligned} \tag{B.4.3}$$

Replacing (B.4.2)-(B.4.3) in (B.4.1), we get the recursive formula

$$\begin{aligned} N\mathbb{V}[\gamma_{t+1}^N(h)] &= N\mathbb{V}[\gamma_t^N(\mathbf{Q}_{t+1}[h])] + \mathbb{E}[\gamma_{t+1}^N(\mathbf{1})\gamma_t^N(\mathbf{Q}_{t+1}[\{h - \phi_t^N M_{t+1}(h)\}^2])] \\ &= M_0(\overline{\mathbf{Q}}_{1:t+1}[h]^2) - \gamma_{t+1}(h)^2 \\ &\quad + \sum_{s=1}^{t+1} \mathbb{E}[\gamma_s^N(\mathbf{1})\gamma_{s-1}^N(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}^N(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2])]\,. \end{aligned}$$

With multiple applications of (4.3.5) in the main chapter, we get that

$$\begin{aligned}&\gamma_s^N(\mathbf{1})\gamma_{s-1}^N(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}^N(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2]) \\ &\qquad \xrightarrow{a.s.} \gamma_s(\mathbf{1})\gamma_{s-1}(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2])\,, \end{aligned} \tag{B.4.4}$$

and, using that $\gamma_s(\mathbf{1})\phi_{s-1}(M_s[h]) = \gamma_s(h)$ and $\gamma_s(\overline{\mathbf{Q}}_{s+1:t+1}[h]) = \gamma_{t+1}(h)$ for all h we get

$$\begin{aligned}\gamma_s(\mathbf{1})\gamma_{s-1}(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2]) \\ = \gamma_s(\mathbf{1})\gamma_s(\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2) \\ = \gamma_s(\mathbf{1})\gamma_s[\overline{\mathbf{Q}}_{s+1:t+1}(h)^2] - \gamma_{t+1}(h)^2\,.\end{aligned}$$

Finally, by (**A5**) and the boundedness of h, $|\gamma_s^N(\mathbf{1})| \leq G_\infty^s$, $|\overline{\mathbf{Q}}_{s+1:t}[h]| \leq G_\infty^{t-s}|h|_\infty$, thus

$$|\gamma_s^N(\mathbf{1})\gamma_{s-1}^N(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}^N(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2])| \leq 4G_\infty^{2t}|h|_\infty\,,$$

and by the dominated convergence theorem, for any $s \in [0:t]$,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}[\gamma_s^N(\mathbf{1})\gamma_{s-1}^N(\mathbf{Q}_s[\{\overline{\mathbf{Q}}_{s+1:t+1}[h] - \phi_{s-1}^N(M_s[\overline{\mathbf{Q}}_{s+1:t+1}[h]])\}^2])] \\ = \gamma_s(\mathbf{1})\gamma_s[\overline{\mathbf{Q}}_{s+1:t+1}(h)^2] - \gamma_{t+1}(h)^2\,. \quad \text{(B.4.5)}\end{aligned}$$

and $N\mathbb{V}[\gamma_{t+1}^N[h]] \longrightarrow \sum_{s=0}^{t+1}\{\gamma_s(\mathbf{1})\gamma_s[\overline{\mathbf{Q}}_{s+1:t+1}(h)^2] - \gamma_{t+1}(h)^2\}$. As argued in Section 4.4.3 of the main chapter, $\mathbb{E}[N(\gamma_{t+1}^N(h) - \gamma_{t+1}(h))^2]$ converges to the asymptotic variance when N goes to infinity, and since $\gamma_{t+1}^N(h)$ is an unbiased estimator of $\gamma_{t+1}(h)$, $N\mathbb{V}[\gamma_{t+1}^N[h]] = \mathbb{E}[N(\gamma_{t+1}^N(h) - \gamma_{t+1}(h))^2]$. Therefore,

$$\lim_{N\to\infty} \mathbb{E}[N(\gamma_{t+1}^N(h) - \gamma_{t+1}(h))^2] = \sum_{s=0}^{t+1}\{\gamma_s(\mathbf{1})\gamma_s[\overline{\mathbf{Q}}_{s+1:t+1}(h)^2] - \gamma_{t+1}(h)^2\}\,,$$

which ends the proof.

B.5 Computational time comparison

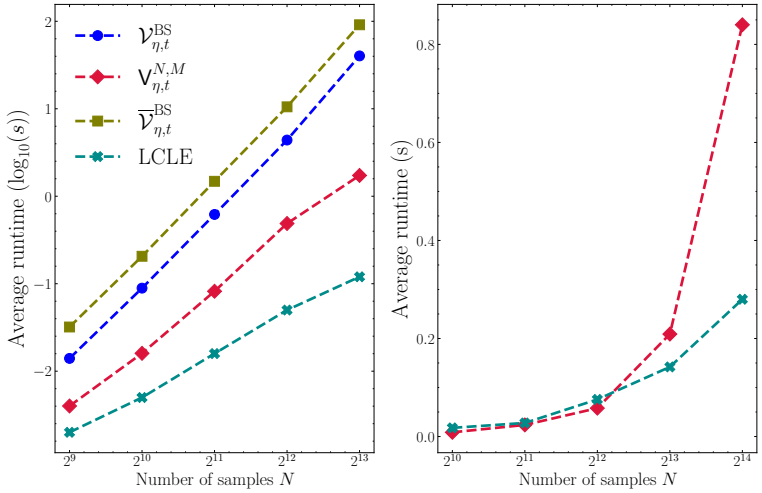

Figure B.1: Comparison of time complexity (left) and runtime (right) for the different estimators, per time step. The runtime on the left plot is on CPU and that on the right plot on GPU, only for our most competitive estimator.

Appendix C

Appendix of Chapter 5

C.1 PPG

In this section, we develop the theoretical framework necessary to establish Theorem 5.3.1. We recall the notions of *Feynman–Kac models*, *many-body Feynman–Kac models*, *backward interpretations*, and *conditional dual processes*. Our presentation follows closely Del Moral et al. (2016) but with a different and hopefully more transparent definition of the many-body extensions. We restate (in Theorem C.1.2 below) a duality formula of Del Moral et al. (2016) relating these concepts. This formula provides a foundation for the *particle Gibbs sampler* described in Algorithm 4.

Notations. Let $(\mathsf{Z}, \mathcal{Z})$ be a measurable space and L another possibly unnormalised transition kernel on $\mathsf{Y} \times \mathcal{Z}$. Define, with K as above,

$$KL : \mathsf{X} \times \mathcal{Z} \ni (x, A) \mapsto \int L(y, A)\, K(x, \mathrm{d}y)$$

and

$$K \otimes L : \mathsf{X} \times (\mathcal{Y} \otimes \mathcal{Z}) \ni (x, A) \mapsto \{\mathbb{1}, \dots, A\}\,(y, z)\, K(x, \mathrm{d}y)\, L(y, \mathrm{d}z),$$

whenever these are well defined. This also defines the $\otimes$ products of a kernel K on $\mathsf{X} \times \mathcal{Y}$ and a measure ν on $\mathcal{X}$ as well as of a kernel L on $\mathsf{Y} \times \mathcal{X}$ and a measure μ on $\mathcal{Y}$ as the measures

$$\nu \otimes K : \mathcal{X} \otimes \mathcal{Y} \ni A \mapsto \{\mathbb{1}, \dots, A\}\,(x, y)\, K(x, \mathrm{d}y)\, \nu(\mathrm{d}x),$$
$$L \otimes \mu : \mathcal{X} \otimes \mathcal{Y} \ni A \mapsto \{\mathbb{1}, \dots, A\}\,(x, y)\, L(y, \mathrm{d}x)\, \mu(\mathrm{d}y).$$

C.1.1 Many-body Feynman–Kac models

In the following, we assume that all random variables are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The distribution flow $\{\eta_m\}_{m \in \mathbb{N}}$ defined in eq. (5.2.3) is intractable in general, but can be approximated by random samples $\boldsymbol{\xi}_m = \{\xi_m^i\}_{i=1}^{\mathbb{N}}$, $m \in \mathbb{N}$, referred to as *particles*, where $\mathbb{N} \in \mathbb{N}_*$ is a fixed Monte Carlo sample size and each particle ξ_m^i is an X_m-valued random variable. Such particle approximation is based on the recursion $\eta_{m+1} = \Phi_m(\eta_m)$, $m \in \mathbb{N}$, where Φ_m denotes the mapping

$$\Phi_m : \mathsf{M}_1(\mathcal{X}_m) \ni \eta \mapsto \frac{\eta Q_m}{\eta g_m} \tag{C.1.1}$$

taking on values in $\mathsf{M}_1(\mathcal{X}_{m+1})$. In order to describe recursively the evolution of the particle population, let $m \in \mathbb{N}$ and assume that the particles $\boldsymbol{\xi}_m$ form a consistent approximation of η_m

in the sense that $\mu(\boldsymbol{\xi}_m)h$, where $\mu(\boldsymbol{\xi}_m) := N^{-1}\sum_{i=1}^{\mathbb{N}}\delta_{\xi_m^i}$, with δ_x denotes the Dirac measure located at x, is the occupation measure formed by $\boldsymbol{\xi}_m$, which serves as a proxy for $\eta_m h$ for all η_m-integrable test functions h. Under general conditions, $\mu(\boldsymbol{\xi}_m)h$ converges in probability to η_m with $N \to \infty$; see Del Moral (2004); Chopin et al. (2020) and references therein. Then, in order to generate an updated particle sample approximating η_{m+1}, new particles $\boldsymbol{\xi}_{m+1} = \{\xi_{m+1}^i\}_{i=1}^{\mathbb{N}}$ are drawn conditionally independently given $\boldsymbol{\xi}_m$ according to

$$\xi_{m+1}^i \sim \Phi_m(\mu(\boldsymbol{\xi}_m)) = \sum_{\ell=1}^{\mathbb{N}} \frac{g_m(\xi_m^\ell)}{\sum_{\ell'=1}^{\mathbb{N}} g_m(\xi_m^{\ell'})} M_m(\xi_m^\ell, \cdot), \quad i \in [1:\mathbb{N}].$$

Since this process of particle updating involves sampling from the mixture distribution $\Phi_m(\mu(\boldsymbol{\xi}_m))$, it can be naturally decomposed into two substeps: *selection* and *mutation*. The selection step consists of randomly choosing the ℓ-th mixture stratum with probability $g_m(\xi_m^\ell)/\sum_{\ell'=1}^{\mathbb{N}} g_m(\xi_m^{\ell'})$ and the mutation step consists of drawing a new particle ξ_{m+1}^i from the selected stratum $M_m(\xi_m^\ell, \cdot)$. In Del Moral et al. (2016), the term *many-body Feynman–Kac models* is related to the law of process $\{\boldsymbol{\xi}_m\}_{m\in\mathbb{N}}$. For all $m \in \mathbb{N}$, let $\mathbf{X}_m := \mathsf{X}_m^{\mathbb{N}}$ and $\boldsymbol{\mathcal{X}}_m := \mathcal{X}_m^{\otimes\mathbb{N}}$; then $\{\boldsymbol{\xi}_m\}_{m\in\mathbb{N}}$ is an inhomogeneous Markov chain on $\{\mathbf{X}_m\}_{m\in\mathbb{N}}$ with transition kernels

$$\boldsymbol{M}_m : \mathbf{X}_m \times \boldsymbol{\mathcal{X}}_{m+1} \ni (\boldsymbol{x}_m, A) \mapsto \Phi_m(\mu(\boldsymbol{x}_m))^{\otimes\mathbb{N}}(A)$$

and initial distribution $\boldsymbol{\eta}_0 = \eta_0^{\otimes\mathbb{N}}$. Now, denote $\mathbf{X}_{0:n} := \prod_{m=0}^n \mathbf{X}_m$ and $\boldsymbol{\mathcal{X}}_{0:n} := \bigotimes_{m=0}^n \boldsymbol{\mathcal{X}}_m$. In the following, we use a bold symbol to stress that a quantity is related to the many-body process. The *many-body Feynman–Kac path model* refers to the flows $\{\boldsymbol{\gamma}_m\}_{m\in\mathbb{N}}$ and $\{\boldsymbol{\eta}_m\}_{m\in\mathbb{N}}$ of the unnormalised and normalised, respectively, probability distributions on $\{\boldsymbol{\mathcal{X}}_{0:m}\}_{m\in\mathbb{N}}$ generated by (5.2.3) and (5.2.2) for the Markov kernels $\{\boldsymbol{M}_m\}_{m\in\mathbb{N}}$, the initial distribution $\boldsymbol{\eta}_0$, the potential functions

$$\boldsymbol{g}_m : \mathbf{X}_m \ni \boldsymbol{x}_m \mapsto \mu(\boldsymbol{x}_m)g_m = \frac{1}{N}\sum_{i=1}^{\mathbb{N}} g_m(x_m^i), \quad m \in \mathbb{N},$$

and the corresponding unnormalised transition kernels

$$\boldsymbol{Q}_m : \mathbf{X}_m \times \boldsymbol{\mathcal{X}}_{m+1} \ni (\boldsymbol{x}_m, A) \mapsto \boldsymbol{g}_m(\boldsymbol{x}_m)\boldsymbol{M}_m(\boldsymbol{x}_m, A), \quad m \in \mathbb{N}.$$

C.1.2 Backward interpretation of Feynman–Kac path flows

Suppose that each kernel Q_n, $n \in \mathbb{N}$, defined in (5.2.1), has a transition density q_n with respect to some dominating measure $\lambda_{n+1} \in \mathsf{M}(\mathcal{X}_{n+1})$. Then for $n \in \mathbb{N}$ and $\eta \in \mathsf{M}_1(\mathcal{X}_n)$ we may define the *backward kernel*

$$\overleftarrow{Q}_{n,\eta} : \mathsf{X}_{n+1} \times \mathcal{X}_n \ni (x_{n+1}, A) \mapsto \frac{\int \mathbb{1}_A(x_n) q_n(x_n, x_{n+1})\,\eta(\mathrm{d}x_n)}{\int q_n(x_n', x_{n+1})\,\eta(\mathrm{d}x_n')}. \tag{C.1.2}$$

Now, denoting, for $n \in \mathbb{N}_*$,

$$B_n : \mathsf{X}_n \times \mathcal{X}_{0:n-1} \ni (x_n, A) \mapsto \int \cdots \int \mathbb{1}_A(x_{0:n-1}) \prod_{m=0}^{n-1} \overleftarrow{Q}_{m,\eta_m}(x_{m+1}, \mathrm{d}x_m), \tag{C.1.3}$$

we may state the following—now classical—*backward decomposition* of the Feynman–Kac path measures, a result that plays a pivotal role in this chapter.

Proposition C.1.1. *For every $n \in \mathbb{N}_*$ it holds that $\gamma_{0:n} = \gamma_n \otimes B_n$ and $\eta_{0:n} = \eta_n \otimes B_n$.*

Although the decomposition in Proposition C.1.1 is well known (see, e.g., Del Moral et al. (2010c, 2016)), we provide a proof in Section C.1.6.1 for completeness. Using the backward decomposition, a particle approximation of a given Feynman–Kac path measure $\eta_{0:n}$ is obtained by first sampling, in an initial forward pass, particle clouds $\{\boldsymbol{\xi}_m\}_{m=0}^n$ from $\boldsymbol{\eta}_0 \otimes \boldsymbol{M}_0 \otimes \cdots \otimes \boldsymbol{M}_{n-1}$ and then sampling, in a subsequent backward pass, for instance N conditionally independent paths $\{\tilde{\xi}_{0:n}^i\}_{i=1}^{\mathbb{N}}$ from $\mathbb{B}_n(\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_n, \cdot)$, where

$$\mathbb{B}_n : \mathbf{X}_{0:n} \times \mathcal{X}_{0:n} \ni (\boldsymbol{x}_{0:n}, A) \mapsto \int \cdots \int \mathbb{1}_A(x_{0:n}) \left(\prod_{m=0}^{n-1} \overleftarrow{Q}_{m,\mu(\boldsymbol{x}_m)}(x_{m+1}, \mathrm{d}x_m) \right) \mu(\boldsymbol{x}_n)(\mathrm{d}x_n) \tag{C.1.4}$$

is a Markov kernel describing the time-reversed dynamics induced by the particle approximations generated in the forward pass. Here and in the following we use blackboard notation to denote kernels related to many-body path spaces. Finally, $\mu(\{\tilde{\xi}_{0:n}^i\}_{i=1}^{\mathbb{N}})h$ is returned as an estimator of $\eta_{0:n}h$ for any $\eta_{0:n}$-integrable test function h. This algorithm is in the literature referred to as the *forward–filtering backward–simulation* (FFBSi) *algorithm* and was introduced in Godsill et al. (2004); see also Cappé et al. (2007); Douc et al. (2011a). More precisely, given the forward particles $\{\boldsymbol{\xi}_m\}_{m=0}^n$, each path $\tilde{\xi}_{0:n}^i$ is generated by first drawing $\tilde{\xi}_n^i$ uniformly among the particles $\boldsymbol{\xi}_n$ in the last generation and then drawing, recursively,

$$\tilde{\xi}_m^i \sim \overleftarrow{Q}_{m,\mu(\boldsymbol{\xi}_m)}(\tilde{\xi}_{m+1}^i, \cdot) = \sum_{j=1}^{\mathbb{N}} \frac{q_m(\xi_m^j, \tilde{\xi}_{m+1}^i)}{\sum_{\ell=1}^N q_m(\xi_m^\ell, \tilde{\xi}_{m+1}^i)} \delta_{\xi_m^j}(\cdot), \tag{C.1.5}$$

i.e., given $\tilde{\xi}_{m+1}^i$, $\tilde{\xi}_m^i$ is picked at random among the $\boldsymbol{\xi}_m$ according to weights proportional to $\{q_m(\xi_m^j, \tilde{\xi}_{m+1}^i)\}_{j=1}^{\mathbb{N}}$. Note that in this basic formulation of the FFBSi algorithm, each backward-sampling operation (C.1.5) requires the computation of the normalising constant $\sum_{\ell=1}^N q_m(\xi_m^\ell, \tilde{\xi}_{m+1}^i)$, which implies an overall quadratic complexity of the algorithm. Still, this heavy computational burden can eased by means of an effective accept–reject technique discussed in Section C.1.4.

C.1.3 Conditional dual processes and particle Gibbs

The *dual process* associated with a given Feynman–Kac model (5.2.3–5.2.2) and a given trajectory $\{z_n\}_{n\in\mathbb{N}}$, where $z_n \in \mathsf{X}_n$ for every $n \in \mathbb{N}$, is defined as the canonical Markov chain with kernels

$$\boldsymbol{M}_n\langle z_{n+1}\rangle : \mathbf{X}_n \times \boldsymbol{\mathcal{X}}_{n+1} \ni (\boldsymbol{x}_n, A) \mapsto \frac{1}{\mathbb{N}} \sum_{i=0}^{\mathbb{N}-1} \left(\Phi_n(\mu(\boldsymbol{x}_n))^{\otimes i} \otimes \delta_{z_{n+1}} \otimes \Phi_n(\mu(\boldsymbol{x}_n))^{\otimes(\mathbb{N}-i-1)} \right)(A), \tag{C.1.6}$$

for $n \in \mathbb{N}$, and initial distribution

$$\boldsymbol{\eta}_0\langle z_0\rangle := \frac{1}{\mathbb{N}} \sum_{i=0}^{\mathbb{N}-1} \left(\eta_0^{\otimes i} \otimes \delta_{z_0} \otimes \eta_0^{\otimes(\mathbb{N}-i-1)} \right). \tag{C.1.7}$$

As clear from (C.1.6–C.1.7), given $\{z_n\}_{n\in\mathbb{N}}$, a realisation $\{\boldsymbol{\xi}_n\}_{n\in\mathbb{N}}$ of the dual process is generated as follows. At time zero, the process is initialised by inserting z_0 at a randomly selected position in the vector $\boldsymbol{\xi}_0$ while drawing independently the remaining components from η_0. Then, given $\boldsymbol{\xi}_n$ at step n, z_{n+1} is inserted at a randomly selected position in $\boldsymbol{\xi}_{n+1}$ while drawing independently the remaining components from $\Phi_n(\mu(\boldsymbol{\xi}_n))$.

In order to describe compactly the law of the conditional dual process, we define the Markov kernel

$$\mathbb{C}_n : \mathsf{X}_{0:n} \times \boldsymbol{\mathcal{X}}_{0:n} \ni (z_{0:n}, A) \mapsto \boldsymbol{\eta}_0\langle z_0\rangle \otimes \boldsymbol{M}_0\langle z_1\rangle \otimes \cdots \otimes \boldsymbol{M}_{n-1}\langle z_n\rangle(A).$$

The following result elegantly combines the underlying model (5.2.3–5.2.2), the many-body Feynman–Kac model, the backward decomposition, and the conditional dual process.

Theorem C.1.2 (Del Moral et al. (2016)). *For all* $n \in \mathbb{N}$,

$$\mathbb{B}_n \otimes \boldsymbol{\gamma}_{0:n} = \gamma_{0:n} \otimes \mathbb{C}_n. \tag{C.1.8}$$

In Del Moral et al. (2016), each state $\boldsymbol{\xi}_n$ of the many-body process maps an outcome ω of the sample space Ω into an *unordered set* of N elements in X_n. However, we have chosen to let each $\boldsymbol{\xi}_n$ take on values in the standard *product space* $\mathsf{X}_n^{\mathbb{N}}$ for two reasons: first, the construction of Del Moral et al. (2016) requires sophisticated measure-theoretic arguments to endow such unordered sets with suitable σ-fields and appropriate measures; second, we see no need to ignore the index order of the particles as long as the Markovian dynamics (C.1.6–C.1.7) of the conditional dual process is symmetrised over the particle cloud. Therefore, in Section C.1.6.2, we include our own proof of duality (C.1.8) for completeness. Note that the measure (C.1.8) on $\mathcal{X}_{0:n} \otimes \boldsymbol{\mathcal{X}}_{0:n}$ is unnormalised, but since the kernels $\mathbb{B}_n$ and $\mathbb{C}_n$ are both Markovian, normalising the identity with $\gamma_{0:n}(\mathsf{X}_{0:n}) = \boldsymbol{\gamma}_{0:n}(\mathbf{X}_{0:n})$ yields immediately

$$\mathbb{B}_n \otimes \boldsymbol{\eta}_{0:n} = \eta_{0:n} \otimes \mathbb{C}_n. \tag{C.1.9}$$

Since the two sides of (C.1.9) provide the full conditionals, it is natural to choose a data-augmentation approach and sample the target (C.1.9) using a two-stage deterministic-scan Gibbs sampler Andrieu et al. (2010); Chopin and Singh (2015b). More specifically, assume that we have generated a state $(\boldsymbol{\xi}_{0:n}[\ell], \zeta_{0:n}[\ell])$ comprising a dual process with associated path on the basis of $\ell \in \mathbb{N}$ iterations of the sampler; then the next state $(\boldsymbol{\xi}_{0:n}[\ell+1], \zeta_{0:n}[\ell+1])$ is generated in a Markovian fashion by sampling first $\boldsymbol{\xi}_{0:n}[\ell+1] \sim \mathbb{C}_n(\zeta_{0:n}[\ell], \cdot)$ and then sampling $\zeta_{0:n}[\ell+1] \sim \mathbb{B}_n(\boldsymbol{\xi}_{0:n}[\ell+1], \cdot)$. After arbitrary initialisation (and the discard of possible burn-in iterations), this procedure produces a Markov trajectory $\{(\boldsymbol{\xi}_{0:n}[\ell], \zeta_{0:n}[\ell])\}_{\ell \in \mathbb{N}}$, and under weak additional technical conditions this Markov chain admits (C.1.9) as its unique invariant distribution. In such a case, the Markov chain is ergodic (Douc et al., 2018b, Chapter 5), and the marginal distribution of the conditioning path $\zeta_{0:n}[\ell]$ converges to the target distribution $\eta_{0:n}$. Therefore, for every $h \in \mathsf{F}(\mathcal{X}_{0:n})$,

$$\lim_{L \to \infty} \frac{1}{L} \sum_{\ell=1}^{L} h(\zeta_{0:n}[\ell]) = \eta_{0:n} h, \quad \mathbb{P}\text{-a.s.}$$

C.1.4 The PARIS algorithm

In the following, we assume that we are given a sequence $\{h_n\}_{n \in \mathbb{N}}$ of *additive state functionals* as in (5.2.5). This problem is particularly relevant in the context of maximum-likelihood-based parameter estimation in general state-space models, *e.g.*, when computing the *score-function*, i.e. the gradient of the log-likelihood function, via the Fisher identity or when computing the intermediate quantity of the *Expectation Maximization* (EM) *algorithm*, in which case $\eta_{0:n}$ and h_n correspond to the joint state posterior and an element of some sufficient statistic, respectively; see Cappé and Moulines (2005); Douc et al. (2011a); Del Moral et al. (2010c); Poyiadjis et al. (2011); Olsson and Westerborn (2017) and the references therein. Interestingly, as noted in Cappé (2011); Del Moral et al. (2010c), the backward decomposition allows, when applied to additive state functionals, a forward recursion for the expectations $\{\eta_{0:n} h_n\}_{n \in \mathbb{N}}$. More specifically, using the forward decomposition $h_{n+1}(x_{0:n+1}) = h_n(x_{0:n}) + \tilde{h}_n(x_n, x_{n+1})$ and the backward

kernel B_{n+1} defined in (C.1.3), we may write, for $x_{n+1} \in \mathsf{X}_{n+1}$,

$$\begin{aligned} B_{n+1}h_{n+1}(x_{n+1}) &= \int \overleftarrow{Q}_{n,\eta_n}(x_{n+1}, \mathrm{d}x_n) \int \Big(h_n(x_{0:n}) + \tilde{h}_n(x_n, x_{n+1})\Big) B_n(x_n, \mathrm{d}x_{0:n-1}) \\ &= \overleftarrow{Q}_{n,\eta_n}(B_n h_n + \tilde{h}_n)(x_{n+1}), \end{aligned} \tag{C.1.10}$$

which by Proposition C.1.1 implies that

$$\eta_{0:n+1}h_{n+1} = \eta_{n+1}\overleftarrow{Q}_{n,\eta_n}(B_n h_n + \tilde{h}_n). \tag{C.1.11}$$

Since the marginal flow $\{\eta_n\}_{n\in\mathbb{N}}$ can be expressed recursively via the mappings $\{\Phi_n\}_{n\in\mathbb{N}}$, (C.1.11) provides, in principle, a basis for online computation of $\{\eta_{0:n}h_n\}_{n\in\mathbb{N}}$. To handle the fact that the marginals are generally intractable we may, following Del Moral et al. (2010c), plug particle approximations $\mu(\boldsymbol{\xi}_{n+1})$ and $\overleftarrow{Q}_{n,\mu(\boldsymbol{\xi}_n)}$ (see (C.1.5)) of η_{n+1} and $\overleftarrow{Q}_{n,\mu(\eta_n)}$, respectively, into the recursion (C.1.11). More precisely, we proceed recursively and assume that at time n we have at hand a sample $\{(\xi_n^i, \beta_n^i)\}_{i=1}^{\mathbb{N}}$ of particles with associated statistics, where each statistic β_n^i serves as an approximation of $B_n h_n(\xi_n^i)$; then evolving the particle cloud according to $\boldsymbol{\xi}_{n+1} \sim \boldsymbol{M}_n(\boldsymbol{\xi}_n, \cdot)$ and updating the statistics using (C.1.10), with $\overleftarrow{Q}_{n,\eta_n}$ replaced by $\overleftarrow{Q}_{n,\mu(\boldsymbol{\xi}_n)}$, yields the particle-wise recursion

$$\beta_{n+1}^i = \sum_{\ell=1}^{\mathbb{N}} \frac{q_n(\xi_n^\ell, \xi_{n+1}^i)}{\sum_{\ell'=1}^{\mathbb{N}} q_n(\xi_n^{\ell'}, \xi_{n+1}^i)} \Big(\beta_n^\ell + \tilde{h}_n(\xi_n^\ell, \xi_{n+1}^i)\Big), \quad i \in [1:\mathbb{N}], \tag{C.1.12}$$

and, finally, the estimator

$$\mu(\boldsymbol{\beta}_n)(\mathrm{id}) = \frac{1}{N}\sum_{i=1}^{\mathbb{N}} \beta_n^i \tag{C.1.13}$$

of $\eta_{0:n}h_n$, where $\boldsymbol{\beta}_n := (\beta_n^1, \ldots, \beta_n^{\mathbb{N}})$, $i \in [1:\mathbb{N}]$. The procedure is initialised by simply letting $\beta_0^i = 0$ for all $i \in [1:N]$. Note that (C.1.13) provides a particle interpretation of the backward decomposition in Proposition C.1.1. This algorithm is a special case of the *forward–filtering backward–smoothing* (FFBSm) *algorithm* (see Andrieu and Doucet (2003); Godsill et al. (2004); Douc et al. (2011a); Särkkä (2013)) for additive functionals satisfying (5.2.5). It allows for online processing of the sequence $\{\eta_{0:n}h_n\}_{n\in\mathbb{N}}$, but has also the appealing property that only the current particles $\boldsymbol{\xi}_n$ and statistics $\boldsymbol{\beta}_n$ need to be stored. However, since each update (C.1.12) requires the summation of $\mathbb{N}$ terms, the scheme has an overall *quadratic* complexity in the number of particles, leading to a computational bottleneck in applications to complex models that require large particle sample sizes $\mathbb{N}$.

In order to detour the computational burden of this forward-only implementation of FFBSm, the `PARIS` algorithm Olsson and Westerborn (2017) updates the statistics $\boldsymbol{\beta}_n$ by replacing each sum (C.1.12) by a Monte Carlo estimate

$$\beta_{n+1}^i = \frac{1}{M}\sum_{j=1}^{M} \Big(\tilde{\beta}_n^{i,j} + \tilde{h}_n(\tilde{\xi}_n^{i,j}, \xi_{n+1}^i)\Big), \quad i \in [1:N], \tag{C.1.14}$$

where $\{(\tilde{\xi}_n^{i,j}, \tilde{\beta}_n^{i,j})\}_{j=1}^M$ are drawn randomly among $\{(\xi_n^i, \beta_n^i)\}_{i=1}^{\mathbb{N}}$ with replacement, by assigning $(\tilde{\xi}_n^{i,j}, \tilde{\beta}_n^{i,j})$ the value of $(\xi_n^\ell, \beta_n^\ell)$ with probability $q_n(\xi_n^\ell, \xi_{n+1}^i)/\sum_{\ell'=1}^{\mathbb{N}} q_n(\xi_n^{\ell'}, \xi_{n+1}^i)$, and the Monte Carlo sample size $M \in \mathbb{N}_*$ is supposed to be much smaller than $\mathbb{N}$ (say, less than 5). Formally,

$$\{(\tilde{\xi}_n^{i,j}, \tilde{\beta}_n^{i,j})\}_{j=1}^M \sim \left(\sum_{\ell=1}^{\mathbb{N}} \frac{q_n(\xi_n^\ell, \xi_{n+1}^i)}{\sum_{\ell'=1}^{\mathbb{N}} q_n(\xi_n^{\ell'}, \xi_{n+1}^i)} \delta_{(\xi_n^\ell, \beta_n^\ell)}\right)^{\otimes M}, \quad i \in [1:\mathbb{N}].$$

The resulting procedure, summarised in Algorithm 3, allows for online processing with constant memory requirements, since it only needs to store the current particle cloud and the estimated auxiliary statistics at each iteration. Moreover, in the case where the Markov transition densities of the model can be uniformly bounded, *i.e.* when there exists, for every $n \in \mathbb{N}$, an upper bound $\bar{\sigma}_n > 0$ such that for all $(x_n, x_{n+1}) \in \mathsf{X}_n \times \mathsf{X}_{n+1}$, $m_n(x_n, x_{n+1}) \leq \bar{\sigma}_n$ (a weak assumption satisfied for most models of interest), a sample $(\tilde{\xi}_n^{i,j}, \beta_n^{i,j})$ can be generated by drawing, with replacement and until acceptance, candidates $(\tilde{\xi}_n^{i,*}, \tilde{\beta}_n^{i,*})$ from $\{(\xi_n^i, \beta_n^i)\}_{i=1}^{\mathbb{N}}$ according to the normalised particle weights $\{g_n(\xi_n^\ell)/\sum_{\ell'} g_n(\xi_n^{\ell'})\}_{\ell=1}^{\mathbb{N}}$, obtained as a by-product in the generation of $\boldsymbol{\xi}_{n+1}$, and accepting the same with probability $m_n(\tilde{\xi}_n^{i,*}, \xi_{n+1}^i)/\bar{\sigma}_n$. As this sampling procedure bypasses completely the calculation of the normalising constant $\sum_{\ell'=1}^{\mathbb{N}} q_n(\xi_n^{\ell'}, \xi_{n+1}^i)$ of the targeted categorical distribution, it yields an overall $\mathcal{O}(MN)$ complexity of the algorithm as a whole; see Douc et al. (2011a) for details.

Increasing M improves the accuracy of the algorithm at the cost of additional computational complexity. As shown in Olsson and Westerborn (2017), there is a qualitative difference between the cases $M = 1$ and $M \geq 2$, and it turns out that the latter is required to keep `PARIS` numerically stable. More precisely, in the latter case, it can be shown that the `PARIS` estimator $\mu(\boldsymbol{\beta}_n)$ satisfies, as $\mathbb{N}$ tends to infinity while M is held fixed, a central limit theorem (CLT) at the rate $\sqrt{N}$ and with an n-normalised asymptotic variance of order $\mathcal{O}(1 - 1/(M-1))$. As clear from this bound, using a large M only yields a waste of computational work, and setting M to 2 or 3 typically works well in practice.

We now introduce the *Parisian particle Gibbs* (`PPG`) *algorithm*. For all $t \in \mathbb{N}_*$, let $\mathsf{Y}_t := \mathsf{X}_{0:t} \times \mathbb{R}$ and $\mathcal{Y}_t := \mathcal{X}_{0:t} \otimes \mathcal{B}(\mathbb{R})$. Moreover, let $\mathsf{Y}_0 := \mathsf{X}_0 \times \{0\}$ and $\mathcal{Y}_0 := \mathcal{X}_0 \otimes \{\{0\}, \emptyset\}$. An element of Y_t will always be denoted by $y_t = (x_{0:t|t}, b_t)$. The Parisian particle Gibbs sampler comprises, as a key ingredient, a *conditional* `PARIS` *step*, which updates recursively a set of Y_t-valued random variables $\upsilon_t^i := (\xi_{0:t|t}^i, \beta_t^i)$, $i \in [1 : \mathbb{N}]$. Let $(\boldsymbol{\upsilon}_t)_{t \in \mathbb{N}}$ denote the corresponding many-body process, each $\boldsymbol{\upsilon}_t := \{(\xi_{0:t|t}^i, \beta_t^i)\}_{i=1}^{\mathbb{N}}$ taking on values in the space $\mathbf{Y}_t := \mathsf{Y}_t^{\mathbb{N}}$, which we furnish with a σ-field $\boldsymbol{\mathcal{Y}}_t := \mathcal{Y}_t^{\otimes \mathbb{N}}$. The space $\mathbf{Y}_0$ and the corresponding σ-field $\boldsymbol{\mathcal{Y}}_0$ are defined accordingly. For every $t \in \mathbb{N}$, we write $\boldsymbol{\xi}_{0:t|t}$ for the collection $\{\xi_{0:t|t}^i\}_{i=1}^{\mathbb{N}}$ of paths in $\boldsymbol{\upsilon}_t$, and $\boldsymbol{\xi}_{t|t}$ for the collection $\{\xi_{t|t}^i\}_{i=1}^{\mathbb{N}}$ of end points of the same.

In the following, we let $t \in \mathbb{N}$ be a fixed time horizon, and describe in detail how the `PPG` approximates $\eta_{0:t} h_t$ iteratively. In short, at each iteration ℓ, the `PPG` produces, given an input conditional path $\zeta_{0:t}[\ell]$, a many-body system $\boldsymbol{\upsilon}_t[\ell+1]$ by means of a series of conditional `PARIS` operations; then, an updated path $\zeta_{0:t}[\ell+1]$, serving as input at the next iteration, is generated by picking one of the paths $\boldsymbol{\xi}_{0:t|t}[\ell+1]$ in $\boldsymbol{\upsilon}_t[\ell+1]$ at random. At each iteration, the produced statistics $\boldsymbol{\beta}_t$ in $\boldsymbol{\upsilon}_t$ provides an approximation of $\eta_{0:t} h_t$ according to (C.1.13).

More precisely, given the path $\zeta_{0:t}[\ell]$, the conditional `PARIS` operations are executed as follows. In the initial step, $\boldsymbol{\xi}_{0|0}[\ell+1]$ are drawn from $\boldsymbol{\eta}_0\langle\zeta_0[\ell]\rangle$ defined in (C.1.7) and $\upsilon_0^i[\ell+1] \leftarrow (\xi_{0|0}^i[\ell+1], 0)$ for all $i \in [1 : \mathbb{N}]$; then, recursively for $m \in [0 : t]$, assuming access to $\boldsymbol{\upsilon}_m[\ell+1]$,

(1) we generate an updated particle cloud $\boldsymbol{\xi}_{m+1}[\ell+1] \sim \boldsymbol{M}_m\langle\zeta_{m+1}[\ell]\rangle(\boldsymbol{\xi}_{m|m}[\ell+1], \cdot)$,

(2) we pick at random, for each $i \in [1 : N]$, an ancestor path with associated statistics $(\tilde{\xi}_{0:m}^{i,1}[\ell+1], \tilde{\beta}_m^{i,1}[\ell+1])$ among $\boldsymbol{\upsilon}_m[\ell+1]$ by drawing

$$(\tilde{\xi}_{0:m}^{i,1}[\ell+1], \tilde{\beta}_m^{i,1}[\ell+1]) \sim \sum_{s=1}^{\mathbb{N}} \frac{q_m(\xi_{m|m}^s[\ell+1], \xi_{m+1}^i[\ell+1])}{\sum_{s'=1}^{\mathbb{N}} q_m(\xi_{m|m}^{s'}[\ell+1], \xi_{m+1}^i[\ell+1])} \delta_{\upsilon_m^s[\ell+1]}, \quad i \in [1 : N],$$

(3) we draw, with replacement, $M-1$ ancestor particles and associated statistics $\{(\tilde{\xi}_m^{i,j}[\ell+$

$1], \tilde{\beta}_m^{i,j}[\ell+1])\}_{j=2}^M$ at random from $\{(\xi_{m|m}^s[\ell+1], \beta_m^s)[\ell+1]\}_{s=1}^{\mathbb{N}}$ according to

$$\{(\tilde{\xi}_m^{i,j}[\ell+1], \tilde{\beta}_m^{i,j}[\ell+1])\}_{j=2}^M \sim \left(\sum_{s=1}^{\mathbb{N}} \frac{q_m(\xi_{m|m}^s[\ell+1], \xi_{m+1}^i[\ell+1])}{\sum_{s'=1}^{\mathbb{N}} q_m(\xi_{m|m}^{s'}[\ell+1], \xi_{m+1}^i[\ell+1])} \delta_{(\xi_{m|m}^s[\ell+1], \beta_m^s[\ell+1])} \right)^{\otimes(M-1)},$$

(4) we set, for all $i \in [1:\mathbb{N}]$, $\xi_{0:m+1|m+1}^i[\ell+1] \leftarrow (\tilde{\xi}_{0:m}^{i,1}[\ell+1], \xi_{m+1}^i[\ell+1])$ and $\upsilon_{m+1}^i[\ell+1] \leftarrow (\xi_{0:m+1|m+1}^i[\ell+1], \beta_{m+1}^i[\ell+1])$, where

$$\beta_{m+1}^i[\ell+1] \leftarrow M^{-1} \sum_{j=1}^M \left(\tilde{\beta}_m^{i,j}[\ell+1] + \tilde{h}_m(\tilde{\xi}_m^{i,j}[\ell+1], \xi_{m+1}^i[\ell+1]) \right).$$

This conditional `PARIS` procedure is summarised in Algorithm 3 and step (**1**) is summarized in Algorithm 8 below.

Algorithm 8 One conditional particle filter step CPF_{s+1}

Input: ζ_{s+1}
Result: $\boldsymbol{\xi}_{s+1} = (\xi_{s+1}^1, \ldots, \xi_{s+1}^N)$
draw $I \sim \text{Uniform}(1/N)$
set $\xi_{s+1}^I = \zeta_{s+1}$
for $i \leftarrow 1$ **to** $\mathbb{N}$ **do**
  if $i \neq I$ **then**
    draw $\alpha_s^i \sim \text{Categorical}(\{\omega_i^s\}_{i=1}^N)$
    draw $\xi_{s+1}^i \sim M_s(\xi_s^{\alpha_s^i}, \cdot)$

Once the set of trajectories and associated statistics $\boldsymbol{\upsilon}_t[\ell+1]$ is formed by means of n recursive conditional `PARIS` updates, an updated path $\zeta_{0:t}[\ell+1]$ is drawn from $\mu(\boldsymbol{\xi}_{0:t|t}[\ell+1])$. A full sweep of the `PPG` is summarised in Algorithm 4.

The following Markov kernels will play an instrumental role in the following. For a given path $\{z_m\}_{m\in\mathbb{N}}$, the conditional `PARIS` update in Algorithm 3 defines an inhomogeneous Markov chain on the spaces $\{(\mathbf{Y}_m, \boldsymbol{\mathcal{Y}}_m)\}_{m\in\mathbb{N}}$ with kernels

$$\mathbf{Y}_m \times \boldsymbol{\mathcal{Y}}_{m+1} \ni (\boldsymbol{y}_m, A) \mapsto \int \boldsymbol{M}_m\langle z_{m+1}\rangle(\boldsymbol{x}_{m|m}, \mathrm{d}\boldsymbol{x}_{m+1})\, \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, A), \quad m \in \mathbb{N},$$

where

$$\begin{aligned}
&\boldsymbol{S}_m : \mathbf{Y}_m \times \mathbf{X}_{m+1} \times \boldsymbol{\mathcal{Y}}_{m+1} \ni (\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, A) \qquad \text{(C.1.15)}\\
&\mapsto \int \cdots \int \prod_{i=1}^{\mathbb{N}} \mathbb{1}_A \left(\left\{ \left((\tilde{x}_{0:m}^{i,1}, x_{m+1}^i), \frac{1}{M} \sum_{j=1}^M \left(\tilde{b}_m^{i,j} + \tilde{h}_m(\tilde{x}_m^{i,j}, x_{m+1}^i) \right) \right) \right\}_{i=1}^{\mathbb{N}} \right)\\
&\times \left(\sum_{\ell=1}^{\mathbb{N}} \frac{q_m(x_{m|m}^\ell, x_{m+1}^i)}{\sum_{\ell'=1}^{\mathbb{N}} q_m(x_{m|m}^{\ell'}, x_{m+1}^i)} \delta_{y_m^\ell}(\mathrm{d}(\tilde{x}_{0:m}^{i,1}, \tilde{b}_m^{i,1})) \right)\\
&\times \left(\sum_{\ell=1}^{\mathbb{N}} \frac{q_m(x_{m|m}^\ell, x_{m+1}^i)}{\sum_{\ell'=1}^{\mathbb{N}} q_m(x_{m|m}^{\ell'}, x_{m+1}^i)} \delta_{(x_{m|m}^\ell, b_m^\ell)} \right)^{\otimes(M-1)} (\mathrm{d}(\tilde{x}_m^{i,2}, \tilde{b}_m^{i,2}, \ldots, \tilde{x}_m^{i,M}, \tilde{b}_m^{i,M})) \Bigg).
\end{aligned}$$

In addition, we introduce the joint law

$$\mathbb{S}_t : \mathbf{X}_{0:t} \times \boldsymbol{\mathcal{Y}}_t \ni (\boldsymbol{x}_{0:t}, A) \mapsto \int \cdots \int \mathbb{1}_A(\boldsymbol{y}_t)\, \boldsymbol{S}_0(\mathbf{J}\boldsymbol{x}_0, \boldsymbol{x}_1, \mathrm{d}\boldsymbol{y}_1) \prod_{m=1}^{t-1} \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, \mathrm{d}\boldsymbol{y}_{m+1}), \tag{C.1.16}$$

where we have defined $\mathbf{J} := \mathrm{I}_{\mathbb{N}} \otimes (0,1)^\intercal$.

The kernel $\mathbb{S}_t$ can be viewed as a *superincumbent sampling kernel* describing the distribution of the output $\boldsymbol{\upsilon}_t$ generated by a sequence of `PARIS` iterates when the many-body process $\{\boldsymbol{\xi}_m\}_{m=0}^t$ associated with the underlying SMC algorithm is given. This allows us to describe alternatively the `PPG` as follows: given $\zeta_{0:t}[\ell]$, draw $\boldsymbol{\xi}_{0:t}[\ell+1] \sim \mathbb{C}_t(\zeta_{0:t}[\ell], \cdot)$; then, draw $\boldsymbol{\upsilon}_t[\ell+1] \sim \mathbb{S}_t(\boldsymbol{\xi}_{0:t}[\ell+1], \cdot)$ and pick a trajectory $\zeta_{0:t}[\ell+1]$ from $\boldsymbol{\xi}_{0:t|t}[\ell+1]$ at random. The following proposition, which will be instrumental in the coming developments, establishes that the conditional distribution of $\zeta_{0:t}[\ell+1]$ given $\boldsymbol{\xi}_{0:t}[\ell+1]$ coincides, as expected, with the particle-induced backward dynamics $\mathbb{B}_t$.

Proposition C.1.3. *For all $t \in \mathbb{N}_*$, $\mathbb{N} \in \mathbb{N}_*$, $\boldsymbol{x}_{0:t} \in \mathbf{X}_{0:t}$, and $h \in \mathsf{F}(\mathcal{X}_{0:t})$,*

$$\int \mathbb{S}_t(\boldsymbol{x}_{0:t}, \mathrm{d}\boldsymbol{y}_t)\, \mu(\boldsymbol{x}_{0:t|t}) h = \mathbb{B}_t h(\boldsymbol{x}_{0:t}).$$

Finally, we define the Markov kernel induced by the `PPG` as well as the extended probability distribution targeted by the same. For this purpose, we introduce the extended measurable space $(\mathbf{E}_t, \boldsymbol{\mathcal{E}}_t)$ with

$$\mathbf{E}_t := \mathbf{Y}_t \times \mathsf{X}_{0:t}, \quad \boldsymbol{\mathcal{E}}_t := \boldsymbol{\mathcal{Y}}_t \otimes \mathcal{X}_{0:t}.$$

The `PPG` described in Algorithm 4 defines a Markov chain on $(\mathbf{E}_t, \boldsymbol{\mathcal{E}}_t)$ with Markov transition kernel

$$\mathbb{K}_t : \mathbf{E}_t \times \boldsymbol{\mathcal{E}}_t \ni (\boldsymbol{y}_t, z_{0:t}, A) \mapsto \iiint \mathbb{1}_A(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t})\, \mathbb{C}_t(z_{0:t}, \mathrm{d}\tilde{\boldsymbol{x}}_{0:t})\, \mathbb{S}_t(\tilde{\boldsymbol{x}}_{0:t}, \mathrm{d}\tilde{\boldsymbol{y}}_t)\, \mu(\tilde{\boldsymbol{x}}_{0:t|t})(\mathrm{d}\tilde{z}_{0:t}). \tag{C.1.17}$$

Note that the values of $\mathbb{K}_t$ defined above do not depend on $\boldsymbol{y}_t$, but only on $(z_{0:t}, A)$. For any given initial distribution $\xi \in \mathsf{M}_1(\mathcal{X}_{0:t})$, let $\mathbb{P}_\xi$ be the distribution of the canonical Markov chain induced by the kernel $\mathbb{K}_t$ and the initial distribution ξ. In the special case where $\xi = \delta_{z_{0:t}}$ for some given path $z_{0:t} \in \mathsf{X}_{0:t}$, we use the short-hand notation $\mathbb{P}_{\delta_{z_{0:t}}} = \mathbb{P}_{z_{0:t}}$. In addition, denote by

$$K_t : \mathsf{X}_{0:t} \times \mathcal{X}_{0:t} \ni (z_{0:t}, A) \mapsto \iiint \mathbb{1}_A(\tilde{z}_{0:t})\, \mathbb{C}_t(z_{0:t}, \mathrm{d}\tilde{\boldsymbol{x}}_{0:t})\, \mathbb{S}_t(\tilde{\boldsymbol{x}}_{0:t}, \mathrm{d}\tilde{\boldsymbol{y}}_t)\, \mu(\tilde{\boldsymbol{x}}_{0:t|t})(\mathrm{d}\tilde{z}_{0:t}) \tag{C.1.18}$$

the path-marginalised version of $\mathbb{K}_t$. By Proposition C.1.3 it holds that $K_t = \mathbb{C}_t\mathbb{B}_t$, which shows that K_t coincides with the Markov transition kernel of the backward-sampling-based particle Gibbs sampler discussed in Section C.1.3. It is also possible to specify the invariant distribution of $\mathbb{K}_t$.

Proposition C.1.4. *For all $t \in \mathbb{N}_*$, it holds that*

$$\eta_{0:t}\mathbb{C}_t\mathbb{S}_t\mathbb{K}_t = \eta_{0:t}\mathbb{C}_t\mathbb{S}_t\,. \tag{C.1.19}$$

Proof. Let $f \in \mathsf{M}(\mathsf{E}_t^{\otimes(k-k_0)})$.

$$\begin{aligned}
&\int f(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t}) \eta_{0:t}(\mathrm{d}z_{0:t}) \mathbb{C}_t \mathbb{S}_t(z_{0:t}, \mathrm{d}(\boldsymbol{y}_t, z'_{0:t})) \mathbb{K}_t(z'_{0:t}, \boldsymbol{y}_t, \mathrm{d}(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t})) \\
&= \int f(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t}) \eta_{0:t}(\mathrm{d}z_{0:t}) \mathbb{C}_t \mathbb{S}_t(z_{0:t}, \mathrm{d}(\boldsymbol{y}_t, z'_{0:t})) \mathbb{C}_t \mathbb{S}_t(z'_{0:t}, \mathrm{d}(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t})) \\
&= \int f(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t}) \eta_{0:t}(\mathrm{d}z_{0:t}) K_t(z_{0:t}, \mathrm{d}z'_{0:t}) \mathbb{C}_t \mathbb{S}_t(z'_{0:t}, \mathrm{d}(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t})) \\
&= \int f(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t}) \eta_{0:t}(\mathrm{d}z'_{0:t}) \mathbb{C}_t \mathbb{S}_t(z'_{0:t}, \mathrm{d}(\tilde{\boldsymbol{y}}_t, \tilde{z}_{0:t})) \,.
\end{aligned}$$

□

Finally, in order prepare for the statement of our theoretical results on the PPG we need to introduce the following Feynman–Kac path model *with a frozen path*. More precisely, for a given path $z_{0:t} \in \mathsf{X}_{0:t}$, define, for every $m \in [0 : t-1]$, the unnormalised kernel

$$Q_m\langle z_{m+1}\rangle : \mathsf{X}_m \times \mathcal{X}_{m+1} \ni (x_m, A) \mapsto \left(1 - \frac{1}{\mathbb{N}}\right) Q_m(x_m, A) + \frac{1}{\mathbb{N}} g_m(x_m)\, \delta_{z_{m+1}}(A)$$

and the initial distribution $\eta_0\langle z_0\rangle : \mathcal{X}_0 \ni A \mapsto (1 - 1/\mathbb{N})\eta_0(A) + \delta_{z_0}(A)/\mathbb{N}$. Given these quantities, define, for $m \in [0 : t]$, $\gamma_m\langle z_{0:m}\rangle := \eta_0\langle z_0\rangle Q_0\langle z_1\rangle \cdots Q_{m-1}\langle z_m\rangle$ along with the normalised counterpart $\eta_m\langle z_{0:m}\rangle := \gamma_m\langle z_{0:m}\rangle / \gamma_m\langle z_{0:m}\rangle \mathbb{1}_{\mathsf{X}_{0:m}}$. Finally, we introduce, for $m \in [0 : t]$, the kernels

$$B_m\langle z_{0:m-1}\rangle : \mathsf{X}_m \times \mathcal{X}_{0:m-1} \ni (x_m, A) \mapsto \int \cdots \int \mathbb{1}_A(x_{0:m-1}) \prod_{m=0}^{t-1} \overleftarrow{Q}_{m, \eta_m\langle z_{0:m}\rangle}(x_{m+1}, \mathrm{d}x_m),$$

as well as the path model $\eta_{0:m}\langle z_{0:m}\rangle := B_m\langle z_{0:m-1}\rangle \otimes \eta_m\langle z_{0:m}\rangle$.

C.1.5 Proof of Theorem 5.3.1

We start by establishing bias, MSE and covariance bounds for a fixed iteration of the PPG estimator.

Theorem C.1.5. *Assume (**A9**). Then for every $t \in \mathbb{N}$ there exist c_t^{bias}, c_t^{mse}, and c_t^{cov} in $\mathbb{R}_+^*$ such that for every $M \in \mathbb{N}_*$, $\xi \in \mathsf{M}_1(\mathcal{X}_{0:t})$, $\ell \in \mathbb{N}_*$, $s \in \mathbb{N}_*$, and $\mathbb{N} \in \mathbb{N}_*$ such that $\mathbb{N} > \mathbb{N}_t$,*

$$|\mathbb{E}_\xi\left[\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id})\right] - \eta_{0:t} h_t| \leq \mathsf{c}_t^{bias} \left(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty\right) \mathbb{N}^{-1} \kappa_{\mathbb{N},t}^{\ell}, \quad \text{(C.1.20)}$$

$$\mathbb{E}_\xi\left[\left(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t} h_t\right)^2\right] \leq \mathsf{c}_t^{mse} \left(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty\right)^2 \mathbb{N}^{-1}, \quad \text{(C.1.21)}$$

$$|\mathbb{E}_\xi\left[\left(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t} h_t\right)\left(\mu(\boldsymbol{\beta}_t[\ell+s])(\mathrm{id}) - \eta_{0:t} h_t\right)\right]| \leq \mathsf{c}_t^{cov} \left(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty\right)^2 \mathbb{N}^{-3/2} \kappa_{\mathbb{N},t}^{s}. \quad \text{(C.1.22)}$$

The constants c_t^{bias}, c_t^{mse}, and c_t^{cov} are explicitly given in the proof. Since the focus of this chapter is on the dependence on $\mathbb{N}$ and the index ℓ, we have made no attempt to optimise the dependence of these constants on t in our proofs; still, we believe that it is possible to prove, under the stated assumptions, that this dependence is linear. The proof of the bound in Theorem C.1.5 is based on four key ingredients. The first is the following unbiasedness property of the PARIS under the many-body Feynman–Kac path model.

Theorem C.1.6. *For every $t \in \mathbb{N}$, $\mathbb{N} \in \mathbb{N}_*$, and $\ell \in \mathbb{N}_*$,*

$$\mathbb{E}_{\eta_{0:t}}\left[\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id})\right] = \int \eta_{0:t}\mathbb{C}_t\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}) = \int \boldsymbol{\eta}_{0:t}\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}) = \eta_{0:t}h_t.$$

The proof of Theorem C.1.6 is postponed to Section C.1.6.3. The second ingredient of the proof of Theorem C.1.5 is the uniform geometric ergodicity of the particle Gibbs with backward sampling established in Del Moral and Jasra (2018).

Theorem C.1.7. *Assume (**A9**). Then, for every $t \in \mathbb{N}$, $(\mu,\nu) \in \mathsf{M}_1(\mathcal{X}_{0:t})^2$, $\ell \in \mathbb{N}_*$, and $\mathbb{N} \in \mathbb{N}_*$ such that $N > 1 + 5\rho_t^2 t/2$, $\|\mu K_t^\ell - \nu K_t^\ell\|_{\mathrm{TV}} \leq \kappa_{\mathbb{N},t}^\ell$, where $\kappa_{\mathbb{N},t}$ is defined in (5.3.2).*

As a third ingredient, we require the following uniform exponential concentration inequality of the conditional `PARIS` with respect to the frozen-path Feynman–Kac model defined in the previous section.

Theorem C.1.8. *For every $t \in \mathbb{N}$ there exist $\mathsf{c}_t > 0$ and $\mathsf{d}_t > 0$ such that for every $M \in \mathbb{N}_*$, $z_{0:t} \in \mathsf{X}_{0:t}$, $\mathbb{N} \in \mathbb{N}_*$, and $\varepsilon > 0$,*

$$\int \mathbb{C}_t\mathbb{S}_t(z_{0:t},\mathrm{d}\boldsymbol{b}_t)\mathbb{1}\left\{|\mu(\boldsymbol{b}_t)(\mathrm{id}) - \eta_{0:t}\langle z_{0:t}\rangle h_t| \geq \varepsilon\right\} \leq \mathsf{c}_t \exp\left(-\frac{\mathsf{d}_t\mathbb{N}\varepsilon^2}{2(\sum_{m=0}^{t-1}\|\tilde{h}_m\|_\infty)^2}\right).$$

Theorem C.1.8, whose proof is postponed to Section C.1.6.5, implies, in turn, the following conditional variance bound.

Proposition C.1.9. *For every $t \in \mathbb{N}$, $M \in \mathbb{N}^*$, $z_{0:t} \in \mathsf{X}_{0:t}$, and $\mathbb{N} \in \mathbb{N}_*$,*

$$\int \mathbb{C}_t\mathbb{S}_t(z_{0:t},\mathrm{d}\boldsymbol{b}_t)\,|\mu(\boldsymbol{b}_t)(\mathrm{id}) - \eta_{0:t}\langle z_{0:t}\rangle h_t|^2 \leq \frac{\mathsf{c}_t}{\mathsf{d}_t}\left(\sum_{m=0}^{t-1}\|\tilde{h}_m\|_\infty\right)^2 \mathbb{N}^{-1}.$$

Using Proposition C.1.9, we deduce, in turn, the following bias bound, whose proof is postponed to Section C.1.6.7.

Proposition C.1.10. *For every $t \in \mathbb{N}$ there exists $\bar{\mathsf{c}}_t^{bias} > 0$ such that for every $M \in \mathbb{N}_*$, $z_{0:t} \in \mathsf{X}_{0:t}$, and $\mathbb{N} \in \mathbb{N}_*$,*

$$\left|\int \mathbb{C}_t\mathbb{S}_t(z_{0:t},\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}) - \eta_{0:t}\langle z_{0:t}\rangle h_t\right| \leq \bar{\mathsf{c}}_t^{bias}\mathbb{N}^{-1}\left(\sum_{m=0}^{t-1}\|\tilde{h}_m\|_\infty\right).$$

A fourth and last ingredient in the proof of Theorem C.1.5 is the following bound on the discrepancy between additive expectations under the original and frozen-path Feynman–Kac models. This bound is established using novel results in Gloaguen et al. (2022). More precisely, since for every $m \in \mathbb{N}$, $(x,z) \in \mathsf{X}_m^2$, $\mathbb{N} \in \mathbb{N}_*$, and $h \in \mathsf{F}(\mathcal{X}_{m+1})$, using (**A9**),

$$|Q_m\langle z\rangle h(x) - Q_m h(x)| \leq \frac{1}{\mathbb{N}}\|g_m^{\|}(\,)\infty\|h\|_\infty \leq \frac{1}{\mathbb{N}}\bar{\tau}_m\|h\|_\infty,$$

applying (Gloaguen et al., 2022, Theorem 4.3) yields the following.

Proposition C.1.11. *Assume (**A9**). Then there exists $\mathsf{c} > 0$ such that for every $t \in \mathbb{N}$, $\mathbb{N} \in \mathbb{N}$, and $z_{0:t} \in \mathsf{X}_{0:t}$,*

$$|\eta_{0:t}\langle z_{0:t}\rangle h_t - \eta_{0:t}h_t| \leq \mathsf{c}\mathbb{N}^{-1}\sum_{m=0}^{t-1}\|\tilde{h}_m\|_\infty.$$

Note that assuming, in addition, that $\sup_{t\in\mathbb{N}} \|\tilde{h}_t\|_\infty < \infty$ yields an $\mathcal{O}(n/\mathbb{N})$ bound in Proposition C.1.11.

Finally, by combining these ingredients we are now ready to present a proof of Theorem C.1.5.

Proof of Theorem C.1.5. Write, using the tower property,

$$\mathbb{E}_\xi\left[\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id})\right] = \mathbb{E}_\xi\left[\mathbb{E}_{\zeta_{0:t}[\ell]}\left[\mu(\boldsymbol{\beta}_t[0])(\mathrm{id})\right]\right] = \int \xi K_t^\ell \mathbb{C}_t\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}).$$

Thus, by the unbiasedness property in Theorem C.1.6,

$$\begin{aligned}\left|\mathbb{E}_\xi\left[\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id})\right] - \eta_{0:t}h_t\right| &= \left|\int \xi K_t^\ell \mathbb{C}_t\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}) - \int \eta_{0:t}\mathbb{C}_t\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})\right| \\ &\leq \left\|\xi K_t^\ell - \eta_{0:t}\right\|_{\mathrm{TV}} \operatorname{osc}\left(\int \mathbb{C}_t\mathbb{S}_t(\cdot,\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})\right),\end{aligned}$$

where, by Theorem C.1.7, $\|\xi K_t^\ell - \eta_{0:t}\|_{\mathrm{TV}} \leq \kappa_{\mathbb{N},t}^\ell$. Moreover, to derive an upper bound on the oscillation, we consider the decomposition

$$\operatorname{osc}\left(\int \mathbb{C}_t\mathbb{S}_t(\cdot,\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})\right) \leq 2\left(\left\|\int \mathbb{C}_t\mathbb{S}_t(\cdot,\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}) - \eta_{0:t}\langle\cdot\rangle h_t\right\|_\infty + \left\|\eta_{0:t}\langle\cdot\rangle h_t - \eta_{0:t}h_t\right\|_\infty\right),$$

where the two terms on the right-hand side can be bounded using Proposition C.1.11 and Proposition C.1.10, respectively. This completes the proof of (C.1.20). We now consider the proof of (C.1.21). Writing

$$\mathbb{E}_\xi\left[\left(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) \;-\eta_{0:t}h_t\right)^2\right] = \int \xi K_t^\ell(\mathrm{d}z_{0:t})\mathbb{C}_t\mathbb{S}_t(z_{0:t},\mathrm{d}\boldsymbol{b}_t)\left(\mu(\boldsymbol{b}_t)(\mathrm{id}) - \eta_{0:t}h_t\right)^2,$$

we may establish (C.1.21) using Proposition C.1.9 and Proposition C.1.11. We finally consider (C.1.22). Using the Markov property we obtain

$$\begin{aligned}&\mathbb{E}_\xi\left[\left(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) \;-\eta_{0:t}h_t\right)\left(\mu(\boldsymbol{\beta}_t[\ell+s])(\mathrm{id}) \;-\eta_{0:t}h_t\right)\right] \\ &\qquad = \mathbb{E}_\xi\left[\left(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) \;-\eta_{0:t}h_t\right)\left(\mathbb{E}_{\zeta_{0:t}[\ell]}[\mu(\boldsymbol{\beta}_t[s])(\mathrm{id})] \;-\eta_{0:t}h_t\right)\right],\end{aligned}$$

from which (C.1.22) follows by (C.1.20) and (C.1.21). □

We are finally equipped to prove Theorem 5.3.1.

Proof of Theorem 5.3.1. We first consider the bias, which can be bounded according to

$$\begin{aligned}\left|\mathbb{E}_\xi[\Pi_{(k_0,k),N}(f)] - \eta_{0:t}h_t\right| &\leq (k-k_0)^{-1}\sum_{\ell=k_0+1}^{k}\left|\mathbb{E}_\xi\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t}h_t\right| \\ &\leq (k-k_0)^{-1}N^{-1}\mathsf{c}_t^{bias}\left(\sum_{m=0}^{t-1}\|\tilde{h}_m\|_\infty\right)\sum_{\ell=k_0+1}^{k}\kappa_{\mathbb{N},t}^\ell,\end{aligned}$$

from which the bound (5.3.3) follows immediately.

We turn to the MSE. Using the decomposition

$$\begin{aligned}\mathbb{E}_\xi[(\Pi_{(k_0,k),N}(f) - \eta_{0:t}h_t)^2] \leq (k-k_0)^{-2}\Bigg\{&\sum_{\ell=k_0+1}^{k}\mathbb{E}_\xi[(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t}h_t)^2] \\ &+ 2\sum_{\ell=k_0+1}^{k}\sum_{j=\ell+1}^{k}\mathbb{E}_\xi[(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t}h_t)(\mu(\boldsymbol{\beta}_t[j])(\mathrm{id}) - \eta_{0:t}h_t)]\Bigg\},\end{aligned}$$

the MSE bound in Theorem C.1.5 implies that

$$\sum_{\ell=k_0+1}^{k} \mathbb{E}_\xi[(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t} h_t)^2] \le \mathsf{c}_t^{mse} \left(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty \right)^2 \mathbb{N}^{-1}(k - k_0).$$

Moreover, using the covariance bound in Theorem C.1.5, we deduce that

$$\sum_{\ell=k_0+1}^{k} \sum_{j=\ell+1}^{k} \mathbb{E}_\xi[(\mu(\boldsymbol{\beta}_t[\ell])(\mathrm{id}) - \eta_{0:t} h_t)(\mu(\boldsymbol{\beta}_t[j])(\mathrm{id}) - \eta_{0:t} h_t)] \le \mathsf{c}_t^{cov} \left(\sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty \right)^2 \mathbb{N}^{-3/2} \left(\sum_{\ell=k_0+1}^{k} \sum_{j=\ell+1}^{k} \kappa_{\mathbb{N},t}^{(j-\ell)} \right).$$

Thus, the proof is concluded by noting that $\sum_{\ell=k_0+1}^{k} \sum_{j=\ell+1}^{k} \kappa_{\mathbb{N},t}^{(j-\ell)} \le (k - k_0)/(1 - \kappa_{\mathbb{N},t})$. □

C.1.6 Proofs of intermediate results

C.1.6.1 Proof of Proposition C.1.1

Using the identity

$$\eta_0 Q_0 \cdots Q_{t-1} \mathbb{1}_{\mathsf{X}_t} = \prod_{m=0}^{t-1} \eta_m Q_m \mathbb{1}_{\mathsf{X}_{m+1}}$$

and the fact that each kernel Q_m has a transition density, write, for $h \in \mathsf{F}(\mathcal{X}_{0:t})$,

$$\begin{aligned}
\eta_{0:t} h &= \int \cdots \int h(x_{0:t})\, \eta_0(\mathrm{d}x_0) \prod_{m=0}^{t-1} \left(\frac{\eta_m[q_m(\cdot, x_{m+1})]\, \lambda_{m+1}(\mathrm{d}x_{m+1})}{\eta_m Q_m \mathbb{1}_{\mathsf{X}_{m+1}}} \right) \left(\frac{q_m(x_m, x_{m+1})}{\eta_m[q_m(\cdot, x_{m+1})]} \right) \\
&= \int \cdots \int h(x_{0:t})\, \eta_t(\mathrm{d}x_t) \prod_{m=0}^{t-1} \frac{\eta_m(\mathrm{d}x_m)\, q_m(x_m, x_{m+1})}{\eta_m[q_m(\cdot, x_{m+1})]} \qquad \text{(C.1.23)} \\
&= \left(\overleftarrow{Q}_{0,\eta_0} \otimes \cdots \otimes \overleftarrow{Q}_{n-1,\eta_{t-1}} \otimes \eta_t \right) h,
\end{aligned}$$

which was to be established.

C.1.6.2 Proof of Theorem C.1.2

Lemma C.1.12. *For all $t \in \mathbb{N}$, $\boldsymbol{x}_t \in \mathbf{X}_t$, and $h \in \mathsf{F}(\boldsymbol{\mathcal{X}}_{t+1} \otimes \mathcal{X}_{t+1})$,*

$$\{h, \ldots, (\}\, \boldsymbol{x}_{t+1}, z_{t+1})\, \boldsymbol{Q}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1})\, \mu(\boldsymbol{x}_{t+1})(\mathrm{d}z_{t+1}) = \{h, \ldots, (\}\, \boldsymbol{x}_{t+1}, z_{t+1})\, \mu(\boldsymbol{x}_t) Q_t(\mathrm{d}z_{t+1})\, \boldsymbol{M}_t\langle z_{t+1}\rangle(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}). \qquad \text{(C.1.24)}$$

In addition, for all $h \in \mathsf{F}(\boldsymbol{\mathcal{X}}_0 \otimes \mathcal{X}_0)$,

$$\{h, \ldots, (\}\, \boldsymbol{x}_0, z_0)\, \boldsymbol{\eta}_0(\mathrm{d}\boldsymbol{x}_0)\, \mu(\boldsymbol{x}_0)(\mathrm{d}z_0) = \{h, \ldots, (\}\, \boldsymbol{x}_0, z_0)\, \boldsymbol{\eta}_0\langle z_0\rangle(\mathrm{d}\boldsymbol{x}_0)\, \eta_0(\mathrm{d}z_0). \qquad \text{(C.1.25)}$$

Proof. Since $\mu(\boldsymbol{x}_t)\, Q_t(\mathrm{d}z_{t+1}) = \boldsymbol{g}_t(\boldsymbol{x}_t)\, \Phi_t(\mu(\boldsymbol{x}_t))(\mathrm{d}z_{t+1})$, we may rewrite the right-hand side of

(C.1.24) according to

$$\begin{aligned}
&\{h, \ldots, (\,\}\, \boldsymbol{x}_{t+1}, z_{t+1})\, \mu(\boldsymbol{x}_t) Q_t(\mathrm{d}z_{t+1})\, \boldsymbol{M}_t\langle z_{t+1}\rangle(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}) \\
&\quad = \boldsymbol{g}_t(\boldsymbol{x}_t) \frac{1}{\mathbb{N}} \sum_{i=0}^{\mathbb{N}-1} \{h, \ldots, (\,\}\, \boldsymbol{x}_{t+1}, z_{t+1})\, \Phi_t(\mu(\boldsymbol{x}_t))(\mathrm{d}z_{t+1}) \\
&\qquad\quad \times \left(\Phi_t(\mu(\boldsymbol{x}_t))^{\otimes i} \otimes \delta_{z_{t+1}} \otimes \Phi_t(\mu(\boldsymbol{x}_t))^{\otimes(\mathbb{N}-i-1)}\right)(\mathrm{d}\boldsymbol{x}_{t+1}) \\
&\quad = \boldsymbol{g}_t(\boldsymbol{x}_t) \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \int \cdots \int h((x_{t+1}^1, \ldots, x_{t+1}^{i-1}, z_{t+1}, x_{t+1}^{i+1}, \ldots, x_{t+1}^{\mathbb{N}}), z_{t+1}) \\
&\qquad\quad \times \Phi_t(\mu(\boldsymbol{x}_t))(\mathrm{d}z_{t+1}) \prod_{\ell \neq i} \Phi_t(\mu(\boldsymbol{x}_t))(\mathrm{d}x_{t+1}^{\ell}) \\
&\quad = \boldsymbol{g}_t(\boldsymbol{x}_t) \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \int h(\boldsymbol{x}_{t+1}, x_{t+1}^i)\, \boldsymbol{M}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}).
\end{aligned}$$

On the other hand, note that the left-hand side of (C.1.24) can be expressed as

$$\{h, \ldots, (\,\}\, \boldsymbol{x}_{t+1}, z_{t+1})\, \boldsymbol{Q}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1})\, \mu(\boldsymbol{x}_{t+1})(\mathrm{d}z_{t+1}) = \boldsymbol{g}_t(\boldsymbol{x}_t) \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \int h(\boldsymbol{x}_{t+1}, x_{t+1}^i)\, \boldsymbol{M}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}),$$

which establishes the identity. The identity (C.1.25) is established along similar lines. □

We establish Theorem C.1.2 by induction; thus, assume that the claim holds true for n and show that for all $h \in \mathsf{F}(\boldsymbol{\mathcal{X}}_{0:t+1} \otimes \mathcal{X}_{0:t+1})$,

$$\begin{aligned}
&\{h, \ldots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \boldsymbol{\gamma}_{0:t+1}(\mathrm{d}\boldsymbol{x}_{0:t+1})\, \mathbb{B}_{t+1}(\boldsymbol{x}_{0:t+1}, \mathrm{d}z_{0:t+1}) \\
&\qquad = \{h, \ldots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \gamma_{0:t+1}(\mathrm{d}z_{0:t+1})\, \mathbb{C}_{t+1}(z_{0:t+1}, \mathrm{d}\boldsymbol{x}_{0:t+1}). \quad \text{(C.1.26)}
\end{aligned}$$

To prove this, we process, using definition (C.3.4), the left-hand side of (C.1.26) according to

$$\begin{aligned}
&\{h, \ldots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \boldsymbol{\gamma}_{0:t+1}(\mathrm{d}\boldsymbol{x}_{0:t+1})\, \mathbb{B}_{t+1}(\boldsymbol{x}_{0:t+1}, \mathrm{d}z_{0:t+1}) \\
&= \left\{\boldsymbol{\gamma}_{,}, \ldots, 0 : t\right\}(\mathrm{d}\boldsymbol{x}_{0:t})\, \mathbb{B}_t(\boldsymbol{x}_{0:t}, \mathrm{d}z_{0:t}) \qquad \text{(C.1.27)} \\
&\qquad \times \{\bar{,}, \ldots, h\}\, (\boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \boldsymbol{Q}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1})\, \mu(\boldsymbol{x}_{t+1})(\mathrm{d}z_{t+1}),
\end{aligned}$$

where we have defined the function

$$\bar{h}(\boldsymbol{x}_{0:t+1}, z_{0:t+1}) := \frac{q_t(z_t, z_{t+1}) h(\boldsymbol{x}_{0:t+1}, z_{0:t+1})}{\mu(\boldsymbol{x}_t)[q_t(\cdot, z_{t+1})]}.$$

Now, applying Lemma C.1.12 to the inner integral and using that

$$\mu(\boldsymbol{x}_t) Q_t(\mathrm{d}z_{t+1}) = \mu(\boldsymbol{x}_t)[q_t(\cdot, z_{t+1})]\, \lambda_{t+1}(\mathrm{d}z_{t+1})$$

yields, for every $\boldsymbol{x}_{0:t}$ and $z_{0:t}$,

$$\begin{aligned}
&\{\bar{,}, \ldots, h\}\, (\boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \boldsymbol{Q}_t(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1})\, \mu(\boldsymbol{x}_{t+1})(\mathrm{d}z_{t+1}) \\
&= \{\bar{,}, \ldots, h\}\, (\boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \mu(\boldsymbol{x}_t) Q_t(\mathrm{d}z_{t+1})\, \boldsymbol{M}_t\langle z_{t+1}\rangle(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}) \\
&= \{h, \ldots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, Q_t(z_t, \mathrm{d}z_{t+1})\, \boldsymbol{M}_t\langle z_{t+1}\rangle(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}).
\end{aligned}$$

Inserting the previous identity into (C.1.27) and using the induction hypothesis provides

$$\begin{aligned}
&\{h, \dots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \boldsymbol{\gamma}_{0:t+1}(\mathrm{d}\boldsymbol{x}_{0:t+1})\, \mathbb{B}_{t+1}(\boldsymbol{x}_{0:t+1}, \mathrm{d}z_{0:t+1}) \\
&= \{\gamma, \dots, 0 : t\}\, (\mathrm{d}z_{0:t})\, \mathbb{C}_t(z_{0:t}, \mathrm{d}\boldsymbol{x}_{0:t}) \\
&\qquad\qquad \times \{h, \dots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, Q_t(z_t, \mathrm{d}z_{t+1})\, \boldsymbol{M}_t\langle z_{t+1}\rangle(\boldsymbol{x}_t, \mathrm{d}\boldsymbol{x}_{t+1}) \\
&= \{h, \dots, (\,\}\, \boldsymbol{x}_{0:t+1}, z_{0:t+1})\, \gamma_{0:t+1}(\mathrm{d}z_{0:t+1})\, \mathbb{C}_{t+1}(z_{0:t+1}, \mathrm{d}\boldsymbol{x}_{0:t+1}),
\end{aligned}$$

which establishes (C.1.26).

C.1.6.3 Proof of Theorem C.1.6

First, define, for $m \in \mathbb{N}$,

$$\boldsymbol{P}_m : \mathbf{Y}_m \times \boldsymbol{\mathcal{Y}}_{m+1} \ni (\boldsymbol{y}_m, A) \mapsto \int \boldsymbol{M}_m(\boldsymbol{x}_{m|m}, \mathrm{d}\boldsymbol{x}_{m+1})\, \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, A). \tag{C.1.28}$$

For any given initial distribution $\boldsymbol{\psi}_0 \in \mathsf{M}_1(\boldsymbol{\mathcal{Y}}_0)$, let $\mathbb{P}^{\boldsymbol{P}}_{\boldsymbol{\psi}_0}$ be the distribution of the canonical Markov chain induced by the Markov kernels $\{\boldsymbol{P}_m\}_{m\in\mathbb{N}}$ and the initial distribution $\boldsymbol{\psi}_0$. By abuse of notation we write, for $\boldsymbol{\eta}_0 \in \mathsf{M}_1(\boldsymbol{\mathcal{X}}_0)$, $\mathbb{P}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}$ instead of $\mathbb{P}^{\boldsymbol{P}}_{\boldsymbol{\psi}_0[\boldsymbol{\eta}_0]}$, where we have defined the extension $\boldsymbol{\psi}_0[\boldsymbol{\eta}_0](A) = \int \mathbb{1}_A(\mathbf{J}\boldsymbol{x}_0)\, \boldsymbol{\eta}_0(\mathrm{d}\boldsymbol{x}_0)$, $A \in \boldsymbol{\mathcal{Y}}_0$. We preface the proof of Theorem C.1.6 by some technical lemmas and a proposition.

Lemma C.1.13. *For all $t \in \mathbb{N}$ and $(f_{t+1}, \tilde{f}_{t+1}) \in \mathsf{F}(\mathcal{X}_{t+1})^2$,*

$$\gamma_{t+1}(f_{t+1}B_{t+1}h_{t+1} + \tilde{f}_{t+1}) = \gamma_t\{Q_t f_{t+1} B_t h_t + Q_t(\tilde{h}_t f_{t+1} + \tilde{f}_{t+1})\}.$$

Proof. Pick arbitrarily $\varphi \in \mathsf{F}(\mathcal{X}_{t:t+1})$ and write, using definition (C.1.3) and the fact that Q_t has a transition density,

$$\begin{aligned}
&\{\varphi, \dots, (\,\}\, x_{t:t+1})\, \gamma_t(\mathrm{d}x_t)\, Q_t(x_t, \mathrm{d}x_{t+1}) \\
&\qquad = \{\varphi, \dots, (\,\}\, x_{t:t+1}) \gamma_t[q_t(\cdot, x_{t+1})]\, \lambda_{t+1}(\mathrm{d}x_{t+1})\, \frac{\gamma_t(\mathrm{d}x_t) q_t(x_t, x_{t+1})}{\gamma_t[q_t(\cdot, x_{t+1})]} \\
&\qquad = \{\varphi, \dots, (\,\}\, x_{t:t+1})\, \gamma_{t+1}(\mathrm{d}x_{t+1})\, \overleftarrow{Q}_{n,\eta_t}(x_{t+1}, \mathrm{d}x_t).
\end{aligned} \tag{C.1.29}$$

Now, by (C.1.10) it holds that

$$B_{t+1}h_{t+1}(x_{t+1}) = \int \overleftarrow{Q}_{n,\eta_t}(x_{t+1}, \mathrm{d}x_t) \left(\tilde{h}_t(x_{t:t+1}) + \int h_t(x_{0:t})\, B_t(x_t, \mathrm{d}x_{0:t-1})\right);$$

therefore, by applying (C.1.29) with

$$\varphi(x_{t:t+1}) := f_{t+1}(x_{t+1}) \left(\tilde{h}_t(x_{t:t+1}) + \int h_t(x_{0:t})\, B_t(x_t, \mathrm{d}x_{0:t-1})\right)$$

we obtain that

$$\begin{aligned}
\gamma_{t+1}(f_{t+1}B_{t+1}h_{t+1}) &= \{\varphi, \dots, (\,\}\, x_{t:t+1})\, \gamma_{t+1}(\mathrm{d}x_{t+1})\, \overleftarrow{Q}_{n,\eta_t}(x_{t+1}, \mathrm{d}x_t) \\
&= \{\varphi, \dots, (\,\}\, x_{t:t+1})\, \gamma_t(\mathrm{d}x_t)\, Q_t(x_t, \mathrm{d}x_{t+1}) \\
&= \gamma_t(Q_t f_{t+1} B_t h_t + Q_t \tilde{h}_t f_{t+1}).
\end{aligned}$$

Now the proof is concluded by noting that since $\gamma_{t+1} = \gamma_t Q_t$, $\gamma_{t+1}\tilde{f}_{t+1} = \gamma_t Q_t \tilde{f}_{t+1}$. □

Lemma C.1.14. *For every $t \in \mathbb{N}_*$, $h_t \in \mathsf{F}(\mathcal{Y}_t)$, and $\boldsymbol{\eta}_0 \in \mathsf{M}_1(\boldsymbol{\mathcal{X}}_0)$ it holds that*

$$\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}[h_t(\boldsymbol{\upsilon}_t) \mid \boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t}] = \mathbb{S}_t h_t(\boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t}), \quad \mathbb{P}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\text{-a.s.}$$

Proof. Pick arbitrarily $v_t \in \mathsf{F}(\mathcal{X}_{0:t})$. We show that

$$\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}[v_t(\boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t}) h_t(\boldsymbol{\upsilon}_t)] = \mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}[v_t(\boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t}) \mathbb{S}_t h_t(\boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t})], \tag{C.1.30}$$

from which the claim follows. Using the definition (C.1.28), the left-hand side of the previous identity may be rewritten as

$$\begin{aligned}
&\int \cdots \int \boldsymbol{\psi}_0[\boldsymbol{\eta}_0](\mathrm{d}\boldsymbol{y}_0) \prod_{m=0}^{t-1} \boldsymbol{P}_m(\boldsymbol{y}_m, \mathrm{d}\boldsymbol{y}_{m+1})\, h_t(\boldsymbol{y}_t) v_t(\boldsymbol{x}_{0|0}, \dots, \boldsymbol{x}_{t|t}) \\
&\quad = \int \cdots \int \boldsymbol{\eta}_0(\mathrm{d}\boldsymbol{x}_{0|0}) \prod_{m=0}^{t-1} \boldsymbol{M}_m(\boldsymbol{x}_{m|m}, \mathrm{d}\boldsymbol{x}_{m+1})\, \boldsymbol{S}_0(\mathbf{J}\boldsymbol{x}_{0|0}, \boldsymbol{x}_1, \mathrm{d}\boldsymbol{y}_1) \\
&\qquad \times \prod_{m=0}^{t-1} \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, \mathrm{d}\boldsymbol{y}_{m+1})\, h_t(\boldsymbol{y}_t) v_t(\boldsymbol{x}_{0|0}, \dots, \boldsymbol{x}_{t|t}) \\
&\quad = \int \cdots \int \boldsymbol{\eta}_0(\mathrm{d}\boldsymbol{x}_0) \prod_{m=0}^{t-1} \boldsymbol{M}_m(\boldsymbol{x}_m, \mathrm{d}\boldsymbol{x}_{m+1})\, \boldsymbol{S}_0(\mathbf{J}\boldsymbol{x}_0, \boldsymbol{x}_1, \mathrm{d}\boldsymbol{y}_1) \\
&\qquad \times \prod_{m=0}^{t-1} \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, \mathrm{d}\boldsymbol{y}_{m+1})\, h_t(\boldsymbol{y}_t) v_t(\boldsymbol{x}_0, \dots, \boldsymbol{x}_t).
\end{aligned}$$

Thus, we may conclude the proof by using the definition (C.1.16) of $\mathbb{S}_t$ together with Fubini's theorem. □

Lemma C.1.15. *For every $t \in \mathbb{N}_*$ and $h_t \in \mathsf{F}(\mathcal{Y}_t)$,*

$$\mathbb{E}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t-1} \boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right) h_t(\boldsymbol{\upsilon}_t)\right] = \int \boldsymbol{\gamma}_{0:t} \mathbb{S}_t(\mathrm{d}\boldsymbol{y}_t)\, h_t(\boldsymbol{y}_t).$$

Proof. The claim of the lemma is a direct implication of Lemma C.1.14; indeed, by applying the tower property and the latter we obtain

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t-1} \boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right) h_t(\boldsymbol{\upsilon}_t)\right] \\
&\qquad = \mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t-1} \boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right) \mathbb{S}_t h_t(\boldsymbol{\xi}_{0|0}, \dots, \boldsymbol{\xi}_{t|t})\right] \\
&\qquad = \int \cdots \int \boldsymbol{\eta}_0(\mathrm{d}\boldsymbol{x}_0) \prod_{m=0}^{t-1} \boldsymbol{g}_m(\boldsymbol{x}_m)\, \boldsymbol{M}_m(\boldsymbol{x}_m, \mathrm{d}\boldsymbol{x}_{m+1})\, \mathbb{S}_t h_t(\boldsymbol{x}_{0:t}) \\
&\qquad = \int \boldsymbol{\gamma}_{0:t} \mathbb{S}_t(\mathrm{d}\boldsymbol{y}_t)\, h_t(\boldsymbol{y}_t).
\end{aligned}$$

□

Proposition C.1.16. *For all $t \in \mathbb{N}_*$, $(\mathbb{N}, M) \in (\mathbb{N}_*)^2$, and $(f_t, \tilde{f}_t) \in \mathsf{F}(\mathcal{X}_t)^2$,*

$$\int \boldsymbol{\gamma}_{0:t} \mathbb{S}_t(\mathrm{d}\boldsymbol{y}_t) \left(\frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \{b_t^i f_t(x_{t|t}^i) + \tilde{f}_t(x_{t|t}^i)\}\right) = \gamma_t(f_t B_t h_t + \tilde{f}_t).$$

Proof. Applying Lemma C.1.15 yields

$$\int \boldsymbol{\gamma}_{0:t}\mathbb{S}_t(\mathrm{d}\boldsymbol{y}_t)\left(\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{b_t^i f_t(x_{t|t}^i)+\tilde{f}_t(x_{t|t}^i)\}\right)=\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}\left[\left(\prod_{m=0}^{t-1}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta_t^i f_t(\xi_{t|t}^i)+\tilde{f}_t(\xi_{t|t}^i)\}\right]. \tag{C.1.31}$$

In the following we will use repeatedly the following filtrations. Let $\tilde{\mathcal{F}}_t := \sigma(\{\boldsymbol{\upsilon}_m\}_{m=0}^t)$ be the σ-field generated by the output of the `PARIS` (Algorithm 3) during the first t iterations. In addition, let $\mathcal{F}_t := \tilde{\mathcal{F}}_{t-1} \vee \sigma(\boldsymbol{\xi}_{t|t})$.

We proceed by induction. Thus, assume that the statement of the proposition holds true for a given $t \in \mathbb{N}_*$ and consider, for arbitrarily chosen $(f_{t+1}, \tilde{f}_{t+1}) \in \mathsf{F}(\mathcal{X}_{t+1})^2$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}&\left[\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta_{t+1}^i f_{t+1}(\xi_{t+1|t+1}^i)+\tilde{f}_{t+1}(\xi_{t+1|t+1}^i)\}\mid\tilde{\mathcal{F}}_t\right]\\&=\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}[\beta_{t+1}^1 f_{t+1}(\xi_{t+1|t+1}^1)+\tilde{f}_{t+1}(\xi_{t+1|t+1}^1)\mid\tilde{\mathcal{F}}_t],\end{aligned}$$

where we used that the variables $\{\beta_{t+1}^i f_{t+1}(\xi_{t+1|t+1}^i)+\tilde{f}_{t+1}(\xi_{t+1|t+1}^i)\}_{i=1}^{\mathbb{N}}$ are conditionally i.i.d. given $\tilde{\mathcal{F}}_t$. Note that, by symmetry,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}\left[\beta_{t+1}^1\mid\mathcal{F}_{t+1}\right]&=\int\boldsymbol{S}_t(\boldsymbol{\upsilon}_t,\boldsymbol{\xi}_{t+1|t+1},\mathrm{d}\boldsymbol{y}_{t+1})\,b_{t+1}^1\\&=\int\cdots\int\left(\prod_{j=1}^{M}\sum_{\ell=1}^{\mathbb{N}}\frac{q_t(\xi_{t|t}^\ell,\xi_{t+1|t+1}^1)}{\sum_{\ell'=1}^{\mathbb{N}}q_t(\xi_{t|t}^{\ell'},\xi_{t+1|t+1}^1)}\delta_{(\xi_{t|t}^\ell,\beta_t^\ell)}(\mathrm{d}\tilde{x}_t^{1,j},\mathrm{d}\tilde{b}_t^{1,j})\right)\\&\qquad\times\frac{1}{M}\sum_{j=1}^{M}\left(\tilde{b}_t^{1,j}+\tilde{h}_t(\tilde{x}_t^{1,j},\xi_{t+1|t+1}^1)\right)\\&=\sum_{\ell=1}^{\mathbb{N}}\frac{q_t(\xi_{t|t}^\ell,\xi_{t+1|t+1}^1)}{\sum_{\ell'=1}^{\mathbb{N}}q_t(\xi_{t|t}^{\ell'},\xi_{t+1|t+1}^1)}\left(\beta_t^\ell+\tilde{h}_t(\xi_{t|t}^\ell,\xi_{t+1|t+1}^1)\right).\end{aligned}\tag{C.1.32}$$

Thus, using the tower property,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}\left[\beta_{t+1}^1 f_{t+1}(\xi_{t+1|t+1}^1)\mid\tilde{\mathcal{F}}_t\right]\\&\qquad=\int\Phi_t(\mu(\boldsymbol{\xi}_{t|t}))(\mathrm{d}x_{t+1})\,f_{t+1}(x_{t+1})\sum_{\ell=1}^{\mathbb{N}}\frac{q_t(\xi_{t|t}^\ell,x_{t+1})}{\sum_{\ell'=1}^{\mathbb{N}}q_t(\xi_{t|t}^{\ell'},x_{t+1})}\left(\beta_t^\ell+\tilde{h}_t(\xi_{t|t}^\ell,x_{t+1})\right),\end{aligned}$$

and consequently, using definition (C.1.1),

$$\begin{aligned}&\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P}}\left[\beta_{t+1}^1 f_{t+1}(\xi_{t+1|t+1}^1)\mid\tilde{\mathcal{F}}_t\right]\\&=\left(\prod_{m=0}^{t-1}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\int\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}q_t(\xi_{t|t}^i,x_{t+1})\\&\quad\times f_{t+1}(x_{t+1})\sum_{\ell=1}^{\mathbb{N}}\frac{q_t(\xi_{t|t}^\ell,x_{t+1})}{\sum_{\ell'=1}^{\mathbb{N}}q_t(\xi_{t|t}^{\ell'},x_{t+1})}\left(\beta_t^\ell+\tilde{h}_t(\xi_{t|t}^\ell,x_{t+1})\right)\lambda_{t+1}(\mathrm{d}x_{t+1})\\&=\left(\prod_{m=0}^{t-1}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}\left(\beta_t^\ell Q_t f_{t+1}(\xi_{t|t}^\ell)+Q_t(\tilde{h}_t f_{t+1})(\xi_{t|t}^\ell)\right).\end{aligned}$$

Thus, applying the induction hypothesis,

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\beta^i_{t+1}f_{t+1}(\xi^i_{t+1|t+1})\right]\\
&=\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t-1}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}\left(\beta^\ell_t Q_t f_{t+1}(\xi^\ell_{t|t})+Q_t(\tilde{h}_t f_{t+1})(\xi^\ell_{t|t})\right)\right]\\
&=\gamma_t\left(Q_t f_{t+1}B_t h_t+Q_t(\tilde{h}_t f_{t+1})\right). \qquad \text{(C.1.33)}
\end{aligned}$$

In the same manner, it can be shown that

$$\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\tilde{f}_{t+1}(\xi^i_{t+1|t+1})\right]=\gamma_t Q_t\tilde{f}_{t+1}. \qquad \text{(C.1.34)}$$

Now, by (C.1.33–C.1.34) and Lemma C.1.13,

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{P}}_{\boldsymbol{\eta}_0}\left[\left(\prod_{m=0}^{t}\boldsymbol{g}_m(\boldsymbol{\xi}_{m|m})\right)\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta^i_{t+1}f_{t+1}(\xi^i_{t+1|t+1})+\tilde{f}_{t+1}(\xi^i_{t+1|t+1})\}\right]\\
&=\gamma_t\left(Q_t f_{t+1}B_t h_t+Q_t(\tilde{h}_t f_{t+1}+Q_t\tilde{f}_{t+1})\right)\\
&=\gamma_{t+1}(f_{t+1}B_{t+1}h_{t+1}+\tilde{f}_{t+1}),
\end{aligned}$$

which shows that the claim of the proposition holds at time $n+1$.

It remains to check the base case $n=0$, which holds trivially true as $\boldsymbol{\beta}_0=\mathbf{0}$, $B_0h_0=0$ by convention, and the initial particles $\boldsymbol{\xi}_{0|0}$ are drawn from $\boldsymbol{\eta}_0$. This completes the proof. □

Proof of Theorem C.1.6. The identity $\int\boldsymbol{\eta}_{0:t}(\mathrm{d}\boldsymbol{x}_{0:t})\,\mathbb{S}_t(\boldsymbol{x}_{0:t},\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})=\eta_{0:t}h_t$ follows immediately by letting $f_t\equiv 1$ and $\tilde{f}_t\equiv 0$ in Proposition C.1.16 and using that $\boldsymbol{\gamma}_{0:t}(\mathbf{X}_{0:t})=\gamma_{0:t}(\mathsf{X}_{0:t})$. Moreover, applying Theorem C.1.2 yields

$$\begin{aligned}
\int\eta_{0:t}\mathbb{C}_t\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})&=\{\eta,\ldots,0:t\}\,(\mathrm{d}z_{0:t})\,\mathbb{C}_t(z_{0:t},\mathrm{d}\boldsymbol{x}_{0:t})\int\mathbb{S}_t(\boldsymbol{x}_{0:t},\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})\\
&=\left\{\boldsymbol{\eta},\ldots,0:t\right\}(\mathrm{d}\boldsymbol{x}_{0:t})\,\mathbb{B}_t(\boldsymbol{x}_{0:t},\mathrm{d}z_{0:t})\int\mathbb{S}_t(\boldsymbol{x}_{0:t},\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id})\\
&=\int\boldsymbol{\eta}_{0:t}\mathbb{S}_t(\mathrm{d}\boldsymbol{b}_t)\,\mu(\boldsymbol{b}_t)(\mathrm{id}).
\end{aligned}$$

Finally, the first identity holds true since K_t leaves $\eta_{0:t}$ invariant. □

C.1.6.4 Proof of Proposition C.1.3

First, note that, by definitions (C.1.15) and (C.1.16),

$$\begin{aligned}
H_t(\boldsymbol{x}_{0:t})&:=\int\mathbb{S}_t(\boldsymbol{x}_{0:t},\mathrm{d}\boldsymbol{y}_t)\,\mu(\boldsymbol{x}_{[}0:n|n])h\\
&=\int\cdots\int\left(\frac{1}{\mathbb{N}}\sum_{j_t=1}^{\mathbb{N}}h(x^{j_t}_{0:t-1|t},x^{j_t}_t)\right)\\
&\quad\times\prod_{m=0}^{t-1}\prod_{i_{m+1}=1}^{\mathbb{N}}\int\sum_{j_m=1}^{\mathbb{N}}\frac{q_m(x^{j_m}_m,x^{i_{m+1}}_{m+1})}{\sum_{j'_m=1}^{\mathbb{N}}q_m(x^{j'_m}_m,x^{i_{m+1}}_{m+1})}\delta_{x^{j_m}_{0:m|m}}(\mathrm{d}x^{i_{m+1}}_{0:m|m+1}),
\end{aligned}$$

where $x^i_{0:-1|0} = \emptyset$ for all $i \in [1:\mathbb{N}]$ by convention. We will show that for every $k \in [0:t]$, $H_{k,t} \equiv H_t$, where

$$H_{k,n}(\boldsymbol{x}_{0:t}) := \frac{1}{\mathbb{N}} \sum_{j_t=1}^{\mathbb{N}} \cdots \sum_{j_k=1}^{\mathbb{N}} \prod_{\ell=k}^{t-1} \frac{q_\ell(x_\ell^{j_\ell}, x_{\ell+1}^{j_{\ell+1}})}{\sum_{j'_\ell=1}^{\mathbb{N}} q_\ell(x_\ell^{j'_\ell}, x_{\ell+1}^{j_{\ell+1}})} a_{k,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}, x_k^{j_k}, \ldots, x_t^{j_t})$$

with

$$\begin{aligned} &a_{k,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}, x_k^{j_k}, \ldots, x_t^{j_t}) \\ &\quad = \int \prod_{m=0}^{k-1} \prod_{i_{m+1}=1}^{\mathbb{N}} \sum_{j_m=1}^{\mathbb{N}} \frac{q_m(x_m^{j_m}, x_{m+1}^{i_{m+1}})}{\sum_{j'_m=1}^{\mathbb{N}} q_m(x_m^{j'_m}, x_{m+1}^{i_{m+1}})} \delta_{x_{0:m|m}^{j_m}}(\mathrm{d}x_{0:m|m+1}^{i_{m+1}}) h(x_{0:k-1|k}^{j_k}, x_k^{j_k}, \ldots, x_t^{j_t}). \end{aligned}$$

Since, by convention, $\prod_{\ell=n}^{t-1} \ldots = 1$, $H_{n,n}(\boldsymbol{x}_{0:t}) = \mathbb{N}^{-1} \sum_{j_t=1}^{\mathbb{N}} a_{n,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{[}n-1], x_t^{j_t})$, and we note that $H_t \equiv H_{n,n}$. We now show that $H_{k,n} \equiv H_{k-1,n}$ for every $k \in [1:t]$; for this purpose, note that

$$\begin{aligned} &a_{k,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}, x_k^{j_k}, \ldots, x_t^{j_t}) \\ &\qquad = \int \prod_{m=0}^{k-2} \prod_{i_{m+1}=1}^{\mathbb{N}} \sum_{j_m=1}^{\mathbb{N}} \frac{q_m(x_m^{j_m}, x_{m+1}^{i_{m+1}})}{\sum_{j'_m=1}^{\mathbb{N}} q_m(x_m^{j'_m}, x_{m+1}^{i_{m+1}})} \delta_{x_{0:m|m}^{j_m}}(\mathrm{d}x_{0:m|m+1}^{i_{m+1}}) \\ &\quad \times \int \prod_{i_k=1}^{\mathbb{N}} \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{i_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{i_k})} \delta_{x_{0:k-1|k-1}^{j_{k-1}}}(\mathrm{d}x_{0:k-1|k}^{i_k})\, h(x_{0:k-1|k}^{j_k}, x_k^{j_k}, \ldots, x_t^{j_t}), \end{aligned}$$

and since $x_{0:k-1|k-1}^{j_{k-1}} = (x_{0:k-2|k-1}^{j_{k-1}}, x_{k-1}^{j_{k-1}})$, it holds that

$$\begin{aligned} &\int \prod_{i_k=1}^{\mathbb{N}} \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{i_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{i_k})} \delta_{x_{0:k-1|k-1}^{j_{k-1}}}(\mathrm{d}x_{0:k-1|k}^{i_k})\, h(x_{0:k-1|k}^{j_k}, x_k^{j_k}, \ldots, x_t^{j_t}) \\ &\qquad = \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{j_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{j_k})} h(x_{0:k-2|k-1}^{j_{k-1}}, x_{k-1}^{j_{k-1}}, x_k^{j_k}, \ldots, x_t^{j_t}). \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} &a_{k,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}, x_k^{j_k}, \ldots, x_t^{j_t}) \\ &\qquad = \int \prod_{m=0}^{k-2} \prod_{i_{m+1}=1}^{\mathbb{N}} \sum_{j_m=1}^{\mathbb{N}} \frac{q_m(x_m^{j_m}, x_{m+1}^{i_{m+1}})}{\sum_{j'_m=1}^{\mathbb{N}} q_m(x_m^{j'_m}, x_{m+1}^{i_{m+1}})} \delta_{x_{0:m|m}^{j_m}}(\mathrm{d}x_{0:m|m+1}^{i_{m+1}}) \\ &\qquad\quad \times \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{j_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{j_k})} h(x_{0:k-2|k-1}^{j_{k-1}}, x_{k-1}^{j_{k-1}}, x_k^{j_k}, \ldots, x_t^{j_t}). \end{aligned}$$

Now, changing the order of summation with respect to j_{k-1} and integration on the right hand side of the previous display yields

$$\begin{aligned} &a_{k,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-1}, x_k^{j_k}, \ldots, x_t^{j_t}) \\ &\qquad = \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{j_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{j_k})} a_{k-1,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-2}, x_{k-1}^{j_{k-1}}, \ldots, x_t^{j_t}). \end{aligned}$$

Thus,

$$
\begin{aligned}
&H_{k,n}(\boldsymbol{x}_{0:t}) \\
&= \frac{1}{\mathbb{N}} \sum_{j_t=1}^{\mathbb{N}} \cdots \sum_{j_k=1}^{\mathbb{N}} \prod_{\ell=k}^{t-1} \frac{q_\ell(x_\ell^{j_\ell}, x_{\ell+1}^{j_{\ell+1}})}{\sum_{j'_\ell=1}^{\mathbb{N}} q_\ell(x_\ell^{j'_\ell}, x_{\ell+1}^{j_{\ell+1}})} \\
&\quad \times \sum_{j_{k-1}=1}^{\mathbb{N}} \frac{q_{k-1}(x_{k-1}^{j_{k-1}}, x_k^{j_k})}{\sum_{j'_{k-1}=1}^{\mathbb{N}} q_{k-1}(x_{k-1}^{j'_{k-1}}, x_k^{j_k})} a_{k-1,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-2}, x_{k-1}^{j_{k-1}}, \ldots, x_t^{j_t}) \\
&= \frac{1}{\mathbb{N}} \sum_{j_t=1}^{\mathbb{N}} \cdots \sum_{j_{k-1}=1}^{\mathbb{N}} \prod_{\ell=k-1}^{t-1} \frac{q_\ell(x_\ell^{j_\ell}, x_{\ell+1}^{j_{\ell+1}})}{\sum_{j'_\ell=1}^{\mathbb{N}} q_\ell(x_\ell^{j'_\ell}, x_{\ell+1}^{j_{\ell+1}})} a_{k-1,n}(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k-2}, x_{k-1}^{j_{k-1}}, \ldots, x_t^{j_t}) \\
&= H_{k-1,n}(\boldsymbol{x}_{0:t}),
\end{aligned}
$$

which establishes the recursion. Therefore, $H_t \equiv H_{0,n}$ and we may now conclude the proof by noting that $\mathbb{B}_t h \equiv H_{0,n}$.

C.1.6.5 Proof of Theorem C.1.8

In order to establish Theorem C.1.8 we will prove the following more general result, of which Theorem C.1.8 is a direct consequence.

Proposition C.1.17. *For every $t \in \mathbb{N}$ and $M \in \mathbb{N}_*$ there exist $\mathsf{c}_t > 0$ and $\mathsf{d}_t > 0$ such that for every $\mathbb{N} \in \mathbb{N}_*$, $z_{0:t} \in \mathsf{X}_{0:t}$, $(f_t, \tilde{f}_t) \in \mathsf{F}(\mathcal{X}_t)^2$, and $\varepsilon > 0$,*

$$
\int \mathbb{C}_t \mathbb{S}_t(z_{0:t}, \mathrm{d}\boldsymbol{b}_t) \mathbb{1}\left\{ \left| \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \{b_t^i f_t(x_{t|t}^i) + \tilde{f}_t(x_{t|t}^i)\} - \eta_t\langle z_{0:t}\rangle (f_t B_t \langle z_{0:t-1}\rangle h_t + \tilde{f}_t) \right| \geq \varepsilon \right\}
$$
$$
\leq \mathsf{c}_t \exp\left(-\frac{\mathsf{d}_t \mathbb{N} \varepsilon^2}{2\kappa_t^2} \right),
$$

where

$$
\kappa_t := \|f_t\|_\infty \sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty + \|\tilde{f}_t\|_\infty. \tag{C.1.35}
$$

To prove Proposition C.1.17 we need the following technical lemma.

Lemma C.1.18. *For every $t \in \mathbb{N}$, $(f_{t+1}, \tilde{f}_{t+1}) \in \mathsf{F}(\mathcal{X}_{t+1})^2$, $z_{0:t+1} \in \mathsf{X}_{0:t+1}$, and $\mathbb{N} \in \mathbb{N}_*$,*

$$
\begin{aligned}
&\gamma_{t+1}\langle z_{0:t+1}\rangle (f_{t+1} B_{t+1} \langle z_{0:t}\rangle h_{t+1} + \tilde{f}_{t+1}) \\
&\qquad = \left(1 - \frac{1}{\mathbb{N}}\right) \gamma_t \langle z_{0:t}\rangle \{Q_t f_{t+1} B_t \langle z_{0:t-1}\rangle h_t + Q_t(\tilde{h}_t f_{t+1} + \tilde{f}_{t+1})\} \\
&\qquad\qquad + \frac{1}{\mathbb{N}} \gamma_t \langle z_{0:t}\rangle g_t \left(f_{t+1}(z_{t+1}) B_{t+1} \langle z_{0:t}\rangle h_{t+1}(z_{t+1}) + \tilde{f}_{t+1}(z_{t+1}) \right).
\end{aligned}
$$

Proof. Since Lemma C.1.13 holds also for the Feynman–Kac model with a frozen path, we obtain

$$
\gamma_{t+1}\langle z_{0:t+1}\rangle (f_{t+1} B_{t+1} \langle z_{0:t}\rangle h_{t+1} + \tilde{f}_{t+1}) = \gamma_t \langle z_{0:t}\rangle \{Q_t \langle z_{t+1}\rangle f_{t+1} B_t \langle z_{0:t}\rangle h_t + Q_t \langle z_{t+1}\rangle (\tilde{h}_t f_{t+1} + \tilde{f}_{t+1})\}.
$$

Thus, the proof is concluded by noting that for every $x_t \in \mathsf{X}_t$ and $h \in \mathsf{F}(\mathcal{X}_{t:t+1})$,

$$
Q_t\langle z_{t+1}\rangle h(x_t) = \left(1 - \frac{1}{\mathbb{N}}\right) Q_t h(x_t) + \frac{1}{\mathbb{N}} g(x_t) h(x_t, z_{t+1}).
$$

□

Finally, before proceeding to the proof of Proposition C.1.17, we introduce the law of the PARIS evolving conditionally on a frozen path $z = \{z_m\}_{m\in\mathbb{N}}$. Define, for $m \in \mathbb{N}$ and $z_{m+1} \in \mathsf{X}_{m+1}$,

$$\boldsymbol{P}_m\langle z_{m+1}\rangle : \mathbf{Y}_m \times \boldsymbol{\mathcal{Y}}_{m+1} \ni (\boldsymbol{y}_m, A) \mapsto \int \boldsymbol{M}_m\langle z_{m+1}\rangle(\boldsymbol{x}_{m|m}, \mathrm{d}\boldsymbol{x}_{m+1})\, \boldsymbol{S}_m(\boldsymbol{y}_m, \boldsymbol{x}_{m+1}, A).$$

For any given initial distribution $\boldsymbol{\psi}_0 \in \mathsf{M}_1(\boldsymbol{\mathcal{Y}}_0)$, let $\mathbb{P}^{\boldsymbol{P},z}_{\boldsymbol{\psi}_0}$ be the distribution of the canonical Markov chain induced by the Markov kernels $\{\boldsymbol{P}_m\langle z_{m+1}\rangle\}_{m\in\mathbb{N}}$ and the initial distribution $\boldsymbol{\psi}_0$. By abuse of notation we write $\mathbb{P}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}$ instead of $\mathbb{P}^{\boldsymbol{P},z}_{\boldsymbol{\psi}_0[\boldsymbol{\eta}_0\langle z_0\rangle]}$, where the extension $\boldsymbol{\psi}_0[\boldsymbol{\eta}_0]$ is defined in Section C.1.6.3.

Proof of Proposition C.1.17. We proceed by forward induction over t. Let the σ-fields $\tilde{\mathcal{F}}_t$ and $\mathcal{F}_t$ be defined as in the proof of Theorem C.1.6, but for the conditional PARIS dual process. Then, under the law $\mathbb{P}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}$, reusing (C.1.32),

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}\left[\ \beta^1_t f_t(\xi^1_t) + \tilde{f}_t(\xi^1_t) \mid \tilde{\mathcal{F}}_{t-1}\right] \\
&= \mathbb{E}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}\left[\mathbb{E}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}\left[\ \beta^1_t \mid \mathcal{F}_t\right]\ f_t(\xi^1_t) + \tilde{f}_t(\xi^1_t) \mid \tilde{\mathcal{F}}_{t-1}\right] \\
&= \mathbb{E}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}\left[\ f_t(\xi^1_t)\sum_{\ell=1}^{\mathbb{N}} \frac{q_{t-1}(\xi^\ell_{t-1}, \xi^1_t)}{\sum_{\ell'=1}^{\mathbb{N}} q_{t-1}(\xi^{\ell'}_{t-1}, \xi^1_t)}\left(\beta^\ell_{t-1} + \tilde{h}_{t-1}(\xi^\ell_{t-1}, \xi^1_t)\right) + \tilde{f}_t(\xi^1_t) \mid \tilde{\mathcal{F}}_{t-1}\right].
\end{aligned}$$

Using (C.1.6), we get

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{P},z}_{\boldsymbol{\eta}_0}\left[\beta^1_t f_t(\xi^1_t) + \tilde{f}_t(\xi^1_t) \mid \tilde{\mathcal{F}}_{t-1}\right] \\
&\qquad = \left(1 - \frac{1}{\mathbb{N}}\right)\frac{\sum_{\ell=1}^{\mathbb{N}}\{\beta^\ell_{t-1} Q_{t-1} f_t(\xi^\ell_{t-1}) + Q_{t-1}(\tilde{h}_{t-1} f_t + \tilde{f}_t)(\xi^\ell_{t-1})\}}{\sum_{\ell'=1}^{\mathbb{N}} g_{t-1}(\xi^{\ell'}_{t-1})} \\
&\qquad\quad + \frac{1}{\mathbb{N}}\left(f_t(z_t)\sum_{\ell=1}^{\mathbb{N}} \frac{q_{t-1}(\xi^\ell_{t-1}, z_t)}{\sum_{\ell'=1}^{\mathbb{N}} q_{t-1}(\xi^{\ell'}_{t-1}, z_t)}\left(\beta^\ell_{t-1} + \tilde{h}_t(\xi^\ell_{t-1}, z_t)\right) + \tilde{f}_t(z_t)\right). \qquad \text{(C.1.36)}
\end{aligned}$$

In order to apply the induction hypothesis to each term on the right-hand side of the previous identity, note that

$$B_t\langle z_{0:t-1}\rangle h_t(z_t) = \frac{\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot, z_t)\{B_{t-1}\langle z_{0:t-2}\rangle h_{t-1}(\cdot) + \tilde{h}_{t-1}(\cdot, z_t)\}]}{\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot, z_t)]}.$$

Therefore, using Lemma C.1.18 and noting that $\gamma_t\langle z_{0:t}\rangle \mathbb{1}_{\mathsf{X}_t}/\gamma_{t-1}\langle z_{0:t}\rangle \mathbb{1}_{\mathsf{X}_{t-1}} = \eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}$ yields

$$\begin{aligned}
\eta_t\langle z_{0:t}\rangle(f_t B_t\langle z_{0:t-1}\rangle h_t + \tilde{f}_t) &= \frac{1}{\mathbb{N}}\left(f_t(z_t) B_t\langle z_{0:t-1}\rangle h_t(z_t) + \tilde{f}_t(z_t)\right) \\
&\quad + \left(1 - \frac{1}{\mathbb{N}}\right)\frac{\eta_{t-1}\langle z_{0:t-1}\rangle\{Q_{t-1} f_t B_{t-1}\langle z_{0:t-2}\rangle h_t + Q_{t-1}(\tilde{h}_{t-1} f_t + \tilde{f}_t)\}}{\eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}}. \qquad \text{(C.1.37)}
\end{aligned}$$

By combining (C.1.36) with (C.1.37), we decompose the error according to

$$\begin{aligned}
&\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta_t^i f_t(\xi_{t|t}^i)+\tilde{f}_t(\xi_{t|t}^i)\}-\eta_t\langle z_{0:t}\rangle(f_t B_t\langle z_{0:t-1}\rangle h_t+\tilde{f}_t)\\
&\quad=\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta_t^i f_t(\xi_{t|t}^i)+\tilde{f}_t(\xi_{t|t}^i)\}-\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left[\,\beta_t^1 f_t(\xi_t^1)+\tilde{f}_t(\xi_t^1)\mid\tilde{\mathcal{F}}_{t-1}\right]\\
&\qquad+\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left[\beta_t^1 f_t(\xi_t^1)+\tilde{f}_t(\xi_t^1)\mid\tilde{\mathcal{F}}_{t-1}\right]-\eta_t\langle z_{0:t}\rangle(f_t B_t\langle z_{0:t-1}\rangle h_t+\tilde{f}_t)\\
&\quad=\mathrm{I}_{\mathbb{N}}^{(1)}+\left(1-\frac{1}{\mathbb{N}}\right)\mathrm{I}_{\mathbb{N}}^{(2)}+\frac{1}{\mathbb{N}}\mathrm{I}_{\mathbb{N}}^{(3)},
\end{aligned}\tag{C.1.38}$$

where

$$\begin{aligned}
\mathrm{I}_{\mathbb{N}}^{(1)}&:=\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{\beta_t^i f_t(\xi_t^i)+\tilde{f}_t(\xi_t^i)\}-\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left[\beta_t^1 f_t(\xi_t^1)+\tilde{f}_t(\xi_t^1)\mid\tilde{\mathcal{F}}_{t-1}\right],\\
\mathrm{I}_{\mathbb{N}}^{(2)}&:=\frac{\sum_{\ell=1}^{\mathbb{N}}\{\beta_{t-1}^\ell Q_{t-1}f_t(\xi_{t-1}^\ell)+Q_{t-1}(\tilde{h}_{t-1}f_t+\tilde{f}_t)(\xi_{t-1}^\ell)\}}{\sum_{\ell'=1}^{\mathbb{N}}g_{t-1}(\xi_{t-1}^{\ell'})}\\
&\qquad-\frac{\eta_{t-1}\langle z_{0:t-1}\rangle\{Q_{t-1}f_t B_t\langle z_{0:t-1}\rangle h_t+Q_{t-1}(\tilde{h}_{t-1}f_t+\tilde{f}_t)\}}{\eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}},
\end{aligned}\tag{C.1.39}$$

and

$$\begin{aligned}
\mathrm{I}_{\mathbb{N}}^{(3)}&:=f_t(z_t)\sum_{\ell=1}^{\mathbb{N}}\frac{q_{t-1}(\xi_{t-1}^\ell,z_t)}{\sum_{\ell'=1}^{\mathbb{N}}q_{t-1}(\xi_{t-1}^{\ell'},z_t)}\left(\beta_{t-1}^\ell+\tilde{h}_{t-1}(\xi_{t-1}^\ell,z_t)\right)\\
&\qquad-f_t(z_t)\frac{\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot,z_t)\{B_{t-1}\langle z_{0:t-2}\rangle h_{t-1}(\cdot)+\tilde{h}_{t-1}(\cdot,z_t)\}]}{\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot,z_t)]}.
\end{aligned}\tag{C.1.40}$$

The proof is now completed by treating the terms $\mathrm{I}_{\mathbb{N}}^{(1)}$, $\mathrm{I}_{\mathbb{N}}^{(2)}$, and $\mathrm{I}_{\mathbb{N}}^{(3)}$ separately, using Hoeffding's inequality and its generalisation in (Douc et al., 2011a, Lemma 4). Choose $\varepsilon>0$; then, by Hoeffding's inequality,

$$\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(|\,\mathrm{I}_{\mathbb{N}}^{(1)}\,|\geq\varepsilon\right)\leq 2\exp\left(-\frac{1}{2}\frac{\varepsilon^2}{\kappa_t^2}\mathbb{N}\right).\tag{C.1.41}$$

To treat $\mathrm{I}_{\mathbb{N}}^{(2)}$, we apply the induction hypothesis to the numerator and denominator, each normalised by $1/\mathbb{N}$, yielding, since $\|Q_{t-1}h\|_\infty\leq\bar{\tau}_{t-1}\|h\|_\infty$ for all $h\in\mathsf{F}(\mathcal{X}_{t-1}\otimes\mathcal{X}_t)$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(\left|\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}\{\beta_{t-1}^\ell Q_{t-1}f_t(\xi_{t-1}^\ell)+Q_{t-1}(\tilde{h}_{t-1}f_t+\tilde{f}_t)(\xi_{t-1}^\ell)\}\right.\right.\\
&\qquad\left.\left.-\eta_{t-1}\langle z_{0:t-1}\rangle\{Q_{t-1}f_t B_t\langle z_{0:t-1}\rangle h_t+Q_{t-1}(\tilde{h}_{t-1}f_t+\tilde{f}_t)\}\right|\geq\varepsilon\right)\\
&\qquad\leq\mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\frac{\varepsilon^2}{\bar{\tau}_{t-1}^2\kappa_t^2}\mathbb{N}\right)
\end{aligned}$$

and

$$\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(\left|\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}g_{t-1}(\xi_{t-1}^\ell)-\eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}\right|\geq\varepsilon\right)\leq\mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\frac{\varepsilon^2}{\bar{\tau}_{t-1}^2}\mathbb{N}\right).$$

Combining the previous two bounds with the generalised Hoeffding inequality in (Douc et al., 2011a, Lemma 4) yields, using also the bounds

$$\frac{\sum_{\ell=1}^{\mathbb{N}}\{\beta_{t-1}^{\ell}Q_{t-1}f_t(\xi_{t-1}^{\ell})+Q_{t-1}(\tilde{h}_{t-1}f_t+\tilde{f}_t)(\xi_{t-1}^{\ell})\}}{\sum_{\ell'=1}^{\mathbb{N}}g_{t-1}(\xi_{t-1}^{\ell'})}\leq \kappa_t$$

and $\eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}\geq \underline{\tau}_{t-1}$, the inequality

$$\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(|\,\mathrm{I}_{\mathbb{N}}^{(2)}\,|\geq\varepsilon\right)\leq \mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\frac{\underline{\tau}_{t-1}^2\varepsilon^2}{\bar{\tau}_{t-1}^2\kappa_t^2}\mathbb{N}\right). \tag{C.1.42}$$

The last term $\mathrm{I}_{\mathbb{N}}^{(3)}$ is treated along similar lines; indeed, by the induction hypothesis, since $\|q_{t-1}\|_{\infty}\leq\bar{\tau}_{t-1}\bar{\sigma}_{t-1}$,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\Bigg(\Bigg|\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}q_{t-1}(\xi_{t-1}^{\ell},z_t)\left(\beta_{t-1}^{\ell}+\tilde{h}_{t-1}(\xi_{t-1}^{\ell},z_t)\right)\\ -\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot,z_t)\{B_{t-1}\langle z_{0:t-1}\rangle h_{t-1}(\cdot)+\tilde{h}_{t-1}(\cdot,z_t)\}]\Bigg|\geq\varepsilon\Bigg)\\ \leq \mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\left(\frac{\varepsilon}{\bar{\tau}_{t-1}\bar{\sigma}_{t-1}\sum_{m=0}^{t-1}\|\tilde{h}_m\|_{\infty}}\right)^2\mathbb{N}\right)\end{aligned}$$

and

$$\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(\left|\frac{1}{\mathbb{N}}\sum_{\ell=1}^{\mathbb{N}}q_{t-1}(\xi_{t-1}^{\ell},z_t)-\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot,z_t)]\right|\geq\varepsilon\right)\leq \mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\left(\frac{\varepsilon}{\bar{\tau}_{t-1}\bar{\sigma}_{t-1}}\right)^2\mathbb{N}\right).$$

Thus, since

$$\sum_{\ell=1}^{\mathbb{N}}\frac{q_{t-1}(\xi_{t-1}^{\ell},z_t)}{\sum_{\ell'=1}^{\mathbb{N}}q_{t-1}(\xi_{t-1}^{\ell'},z_t)}\left(\beta_{t-1}^{\ell}+\tilde{h}_{t-1}(\xi_{t-1}^{\ell},z_t)\right)\leq\sum_{m=0}^{t-1}\|\tilde{h}_m\|_{\infty}$$

and $\eta_{t-1}\langle z_{0:t-1}\rangle[q_{t-1}(\cdot,z_t)]\geq\underline{\tau}_{t-1}$, the generalised Hoeffding inequality provides

$$\mathbb{P}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}\left(|\,\mathrm{I}_{\mathbb{N}}^{(3)}\,|\geq\varepsilon\right)\leq \mathsf{c}_{t-1}\exp\left(-\mathsf{d}_{t-1}\left(\frac{\underline{\tau}_{t-1}\varepsilon}{2\bar{\tau}_{t-1}\bar{\sigma}_{t-1}\|f_t\|_{\infty}\sum_{m=0}^{t-1}\|\tilde{h}_m\|_{\infty}}\right)^2\mathbb{N}\right). \tag{C.1.43}$$

Finally, combining the bounds (C.1.41–C.1.43) completes the proof. □

C.1.6.6 Proof of Proposition C.1.9

The statement of Proposition C.1.9 is implied by the following more general result, which we will prove below.

Proposition C.1.19. *For every $t\in\mathbb{N}$, $M\in\mathbb{N}^*$, $\mathbb{N}\in\mathbb{N}_*$, $z_{0:t}\in\mathsf{X}_{0:t}$, $(f_t,\tilde{f}_t)\in\mathsf{F}(\mathcal{X}_t)^2$, and $p\geq 2$, it holds that*

$$\int\mathbb{C}_t\mathbb{S}_t(z_{0:t},\mathrm{d}\boldsymbol{b}_t)\left|\frac{1}{\mathbb{N}}\sum_{i=1}^{\mathbb{N}}\{b_t^i f_t(x_{t|t}^i)+\tilde{f}_t(x_{t|t}^i)\}-\eta_t\langle z_{0:t}\rangle(f_tB_t\langle z_{0:t-1}\rangle h_t+\tilde{f}_t)\right|^p\leq \mathsf{c}_t(p/\mathsf{d}_t)^{p/2}\mathbb{N}^{-p/2}\kappa_t^p,$$

where $\mathsf{c}_t>0$, $\mathsf{d}_t>0$ and κ_t are defined in Proposition C.1.17 and (C.1.35), respectively.

Before proving Proposition C.1.19, we establish the following result.

Lemma C.1.20. *Let X be an $\mathbb{R}^d$-valued random variable, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, satisfying $\mathbb{P}(|X| \geq t) \leq c \exp(-t^2/(2\sigma^2))$ for every $t \geq 0$ and some $c > 0$ and $\sigma > 0$. Then for every $p \geq 2$ it holds that $\mathbb{E}[|X|^p] \leq c p^{p/2} \sigma^p$.*

Proof. Using Fubini's theorem and the change of variable formula,

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty p t^{p-1} \mathbb{P}(|X| \geq t) \, \mathrm{d}t = c p 2^{p/2-1} \sigma^p \Gamma(p/2),$$

where Γ is the Gamma function. It remains to apply the bound $\Gamma(p/2) \leq (p/2)^{p/2-1}$ (see Anderson and Qiu (1997)), which holds for $p \geq 2$ by [2, Theorem 1.5]. □

Proof of Proposition C.1.19. By combining Proposition C.1.17 and Lemma C.1.20 we obtain

$$\mathbb{N} \int \mathbb{C}_t \mathbb{S}_t(z_{0:t}, \mathrm{d}\boldsymbol{b}_t) \left| \frac{1}{\mathbb{N}} \sum\nolimits_{i=1}^{\mathbb{N}} \{ b_t^i f_t(x_{t|t}^i) + \tilde{f}_t(x_{t|t}^i) \} - \eta_t \langle z_{0:t} \rangle (f_t B_t \langle z_{0:t-1} \rangle h_t + \tilde{f}_t) \right|^2$$
$$\leq \mathsf{c}_t (p/\mathsf{d}_t)^{p/2} \mathbb{N}^{-p/2} \left(\|f_t\|_\infty \sum_{m=0}^{t-1} \|\tilde{h}_m\|_\infty + \|\tilde{f}_t\|_\infty \right)^p,$$

which was to be established. □

C.1.6.7 Proof of Proposition C.1.10

Like previously, we establish Proposition C.1.10 via a more general result, namely the following.

Proposition C.1.21. *For every $t \in \mathbb{N}$, the exists $\bar{\mathsf{c}}_t^{bias} < \infty$ such that for every $M \in \mathbb{N}^*$, $\mathbb{N} \in \mathbb{N}_*$, $z_{0:t} \in \mathsf{X}_{0:t}$, and $(f_t, \tilde{f}_t) \in \mathsf{F}(\mathcal{X}_t)^2$,*

$$\left| \int \mathbb{C}_t \mathbb{S}_t(z_{0:t}, \mathrm{d}\boldsymbol{b}_t) \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \{ b_t^i f_t(x_{t|t}^i) + \tilde{f}_t(x_{t|t}^i) \} - \eta_t \langle z_{0:t} \rangle (f_t B_t \langle z_{0:t-1} \rangle h_t + \tilde{f}_t) \right| \leq \bar{\mathsf{c}}_t^{bias} \kappa_t \mathbb{N}^{-1},$$

where κ_t is defined in (C.1.35).

We preface the proof of Proposition C.1.21 by a technical lemma providing a bound on the bias of ratios of random variables.

Lemma C.1.22. *Let α and β be (possibly dependent) random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and such that $\mathbb{E}[\alpha^2] < \infty$ and $\mathbb{E}[\beta^2] < \infty$. Moreover, assume that there exist $c > 0$ and $d > 0$ such that $|\alpha/\beta| \leq c$, $\mathbb{P}$-a.s., $|a/b| \leq c$, $\mathbb{E}[(\alpha - a)^2] \leq c^2 d^2$, and $\mathbb{E}[(\beta - b)^2] \leq d^2$. Then*

$$|\mathbb{E}[\alpha/\beta] - a/b| \leq 2c(d/b)^2 + c|\mathbb{E}[\beta - b]|/|b| + |\mathbb{E}[\alpha - a]|/|b|. \tag{C.1.44}$$

Proof. Using the identity

$$\mathbb{E}[\alpha/\beta] - a/b = \mathbb{E}[(\alpha/\beta)(b - \beta)^2]/b^2 + \mathbb{E}[(\alpha - a)(b - \beta)]/b^2 + a\mathbb{E}[b - \beta]/b^2 + \mathbb{E}[\alpha - a]/b,$$

the claim is established by applying the Cauchy–Schwarz inequality and the assumptions of the lemma according to

$$\begin{aligned}
&|\mathbb{E}[\alpha/\beta] - a/b| \\
&\quad \leq c\mathbb{E}[(\beta - b)^2]/b^2 + \{\mathbb{E}[(\alpha - a)^2]\mathbb{E}[(\beta - b)^2]\}^{1/2}/b^2 + |a||\mathbb{E}[\beta - b]|/b^2 + |\mathbb{E}[\alpha - a]|/b^2 \\
&\quad \leq 2c(d/b)^2 + c|\mathbb{E}[\beta - b]|/|b| + |\mathbb{E}[\alpha - a]|/|b|.
\end{aligned}$$

□

Proof of Proposition C.1.10. We proceed by induction and assume that the claim holds true for $n-1$. Reusing the error decomposition (C.1.38), it is enough to bound the expectations of the terms $\mathrm{I}_{\mathbb{N}}^{(2)}$ and $\mathrm{I}_{\mathbb{N}}^{(3)}$ given in (C.1.39) and (C.1.40), respectively (since $\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\mathrm{I}_{\mathbb{N}}^{(1)}] = 0$). This will be done using the induction hypothesis, Lemma C.1.22, and Proposition C.1.19. More precisely, to bound the expectation of $\mathrm{I}_{\mathbb{N}}^{(2)}$, we use Lemma C.1.22 with $\alpha \leftarrow \alpha_t$, $\beta \leftarrow \beta_t$, $a \leftarrow a_t$, and $b \leftarrow b_t$, where

$$\alpha_t := \frac{1}{\mathbb{N}} \sum_{\ell=1}^{\mathbb{N}} \{\beta_{t-1}^{\ell} Q_{t-1} f_t(\xi_{t-1}^{\ell}) + Q_{t-1}(\tilde{h}_{t-1} f_t + \tilde{f}_t)(\xi_{t-1}^{\ell})\}, \quad \beta_t := \frac{1}{\mathbb{N}} \sum_{\ell=1}^{\mathbb{N}} g_{t-1}(\xi_{t-1}^{\ell}),$$
$$a_t := \eta_{t-1}\langle z_{0:t-1}\rangle \{Q_{t-1} f_t B_t \langle z_{0:t-1}\rangle h_t + Q_{t-1}(\tilde{h}_{t-1} f_t + \tilde{f}_t)\}, \quad b_t := \eta_{t-1}\langle z_{0:t-1}\rangle g_{t-1}.$$

For this purpose, note that $|\alpha_t/\beta_t| \le \kappa_t$ and $|a_t/b_t| \le \kappa_t$, where κ_t is defined in (C.1.35). On the other hand, using Proposition C.1.19 (applied with $p = 2$), we obtain

$$\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[(\alpha_t - a_t)^2] \le d_t^2 \kappa_t^2 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[(\beta_t - b_t)^2] \le d_t^2,$$

where $d_t^2 := \mathsf{c}_t \bar{\tau}_{t-1}^2 / (\mathsf{d}_t \mathbb{N})$. Using the induction assumption, we get

$$|\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\alpha_t] - a_t| \le \bar{\mathsf{c}}_{t-1}^{bias} \mathbb{N}^{-1} \bar{\tau}_{t-1} \kappa_t \quad \text{and} \quad |\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\beta_t] - b_t| \le \bar{\mathsf{c}}_{t-1}^{bias} \mathbb{N}^{-1} \bar{\tau}_{t-1}.$$

Hence, the conditions of Lemma C.1.22 are satisfied and we deduce that

$$|\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\mathrm{I}_{\mathbb{N}}^{(2)}]| = |\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\alpha_t/\beta_t] - a_t/b_t| \le 2\kappa_t \frac{\mathsf{c}_t}{\mathsf{d}_t \mathbb{N}} \frac{\bar{\tau}_{t-1}^2}{\underline{\tau}_{t-1}^2} + 2\bar{\mathsf{c}}_{t-1}^{bias} \kappa_t \frac{\bar{\tau}_{t-1}}{\underline{\tau}_{t-1} \mathbb{N}}.$$

The bound on $|\mathbb{E}_{\boldsymbol{\eta}_0}^{\boldsymbol{P},z}[\mathrm{I}_{\mathbb{N}}^{(2)}]|$ is obtained along the same lines. □

C.2 Learning with PPG

This section is divided into three subsections. Section C.2.1 establishes, following closely Karimi et al. (2019), a non-asymptotic bound for stochastic approximation schemes under general assumptions. Section C.2.2 shows how assumptions (**A9-10**) imply the assumptions provided in C.2.1 and therefore allow to establish Theorem 5.4.1. Finally, Section C.2.3 provides sufficient assumptions on the model ensuring that (**A10**) holds.

C.2.1 Non-asymptotic bound

We follow closely Karimi et al. (2019). Consider the recursion

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H_{\theta_n}(X_{n+1}), \quad n \in \mathbb{N},$$

where $\theta_n \in \Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}_*$ and $\{X_n\}_{n\in\mathbb{N}}$ is a *state-dependent* Markov chain on some measurable space $(\mathsf{X}, \mathcal{X})$ in the sense that $X_{n+1} \sim \mathbb{P}_{\theta_n}(X_n, \cdot)$ with $\mathbb{P}_\theta$ being some Markov kernel on $(\mathsf{X}, \mathcal{X})$. Let $h(\theta) = \int H_\theta(x)\, \pi_\theta(\mathrm{d}x)$, where π_θ is the invariant measure of $\mathbb{P}_\theta$ and $e_{n+1} := H_{\theta_n}(X_{n+1}) - h(\theta_n)$. As all norms are equivalent in finite dimensional vector spaces, we use $\|\cdot\|$ to denote a generic norm. We denote by $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ the natural filtration of the Markov chain $\{X_n\}_{n\in\mathbb{N}}$.

(**A17**) There exists a Borel measurable function $V : \Theta \to \mathbb{R}$ such that for every $\theta \in \Theta$, $\nabla V(\theta) = h(\theta)$.

(**A18**) There exists $L^V \in \mathbb{R}_{\geq 0}$ such that for every $(\theta, \theta') \in \Theta^2$,

$$\|\nabla V(\theta) - \nabla V(\theta')\| \leq L^V \|\theta - \theta'\|.$$

(**A19**) There exists a Borel measurable function $\widehat{H} : \Theta \times \mathsf{X} \to \Theta$ such that for every $\theta \in \Theta$ and $x \in \mathsf{X}$,

$$\widehat{H}_\theta(x) - \mathbb{P}_\theta \widehat{H}_\theta(x) = H_\theta(x) - h(\theta)\,.$$

(**A20**) There exists $L^{\mathbb{P}\widehat{H}} \in \mathbb{R}_{\geq 0}$ such that for every $(\theta_0, \theta_1) \in \Theta^2$,

$$\sup_{x \in \mathsf{X}} \|\mathbb{P}_{\theta_0} \widehat{H}_{\theta_0}(x) - \mathbb{P}_{\theta_0} \widehat{H}_{\theta_1}(x)\| \leq L^{\mathbb{P}\widehat{H}} \|\theta_0 - \theta_1\|\,.$$

(**A21**) There exists $L_0^{\mathbb{P}\widehat{H}} \in \mathbb{R}_{\geq 0}$ such that

$$\sup_{\theta \in \Theta} \|\mathbb{P}_\theta \widehat{H}_\theta\| \leq L_0^{\mathbb{P}\widehat{H}}\,.$$

(**A22**) There exists $\sigma_{mse} \in \mathbb{R}_{\geq 0}$ such that for every $x \in \mathsf{X}$ and $\theta \in \Theta$,

$$\int \|H_\theta(x') - h(\theta)\|^2 \, \mathbb{P}_\theta(x, \mathrm{d}x') \leq \sigma_{mse}^2\,.$$

(**A23**) There exists $L^{\widehat{H}} \in \mathbb{R}_{\geq 0}$ such that for every $x \in \mathsf{X}$,

$$\sup_{\theta \in \Theta} \int \|\widehat{H}_\theta\| \, \mathbb{P}_\theta(x, \mathrm{d}x') \leq L^{\widehat{H}}\,.$$

Theorem C.2.1. *Assume that (**A17-23**) hold. In addition, assume that there exist $a > 0$ and $a' > 0$ such that for all $n \in \mathbb{N}$,*

$$\gamma_{n+1} \leq \gamma_n \leq a\gamma_{n+1}\,, \quad \gamma_n - \gamma_{n+1} \leq a'\gamma_n^2\,, \quad \gamma_1 \leq (L^V + C_h)^{-1}/2\,.$$

Moreover, for any $n \in \mathbb{N}_*$, let ϖ be a $[0:n]$-valued random variable, independent of $\{\mathcal{F}_\ell\}_{\ell \geq 0}$ and such that $\mathbb{P}(\varpi = k) = \gamma_{k+1} / \sum_{\ell=0}^n \gamma_{\ell+1}$ for $k \in [0:n]$. Then,

$$\mathbb{E}\left[\|h(\theta_\varpi)\|^2\right] \leq 2 \frac{V_{0,n} + C_{0,n} + (\sigma_{mse}^2 L^V + C_\gamma) \sum_{k=0}^n \gamma_{k+1}^2}{\sum_{k=0}^n \gamma_{k+1}}\,,$$

where $V_{0,n} := \mathbb{E}\left[V(\theta) - V(\theta_n)\right]$ and

$$C_{0,n} := \gamma_1 h(\theta_0) L^{\widehat{H}} + L_0^{\mathbb{P}\widehat{H}}(\gamma_1 - \gamma_{n+1} + 1)\,, \tag{C.2.1}$$

$$C_\gamma := \sigma_{mse} L^{\mathbb{P}\widehat{H}} + (1 + \sigma_{mse}) L^V L_0^{\mathbb{P}\widehat{H}}\,, \tag{C.2.2}$$

$$C_h := L^{\mathbb{P}\widehat{H}}\left((a+1)/2 + a\sigma_{mse}\right) + (L^V + a' + 1) L_0^{\mathbb{P}\widehat{H}}\,. \tag{C.2.3}$$

Proof. We follow closely the proof of (Karimi et al., 2019, Theorem 2) and adapt it to our setting. First, note that by (**A17**), assumptions **A1** and **A2** of (Karimi et al., 2019, Theorem 2) hold with $c_0 = d_0 = 0$ and $c_1 = d_1 = 1$. In addition, the claim in (Karimi et al., 2019, Lemma 1) holds true since by (**A18**), their **A3** holds. Moreover, (Karimi et al., 2019, Equation 17) can also be established under (**A22**), as we may rewrite it as

$$\sum_{\ell=0}^n \gamma_{\ell+1}^2 \mathbb{E}\left[\|e_{\ell+1}\|^2\right] = \sum_{\ell=0}^n \gamma_{\ell+1}^2 \mathbb{E}\left[\mathbb{E}\left[\|e_{\ell+1}\|^2 \mid \mathcal{F}_\ell\right]\right] \leq \sigma_{mse}^2 \sum_{\ell=0}^n \gamma_{\ell+1}^2\,.$$

Following the proof of (Karimi et al., 2019, Lemma 2), consider the decomposition

$$\mathbb{E}\left[-\sum_{\ell=0}^{n}\gamma_{\ell+1}\left\langle\nabla V(\theta_\ell), e_{\ell+1}\right\rangle\right] = \mathbb{E}\left[A_1 + A_2 + A_3 + A_4 + A_5\right],$$

where

$$\begin{aligned}
A_1 &:= -\sum_{\ell=1}^{n}\gamma_{\ell+1}\left\langle\nabla V(\theta_\ell), \widehat{H}_{\theta_\ell}(X_{\ell+1}) - \mathbb{P}_{\theta_\ell}\widehat{H}_{\theta_\ell}(X_\ell)\right\rangle,\\
A_2 &:= -\sum_{\ell=1}^{n}\gamma_{\ell+1}\left\langle\nabla V(\theta_\ell), \mathbb{P}_{\theta_\ell}\widehat{H}_{\theta_\ell}(X_\ell) - \mathbb{P}_{\theta_{\ell-1}}\widehat{H}_{\theta_{\ell-1}}(X_\ell)\right\rangle,\\
A_3 &:= -\sum_{\ell=1}^{n}\gamma_{\ell+1}\left\langle\nabla V(\theta_\ell) - \nabla V(\theta_{\ell-1}), \mathbb{P}_{\theta_{\ell-1}}\widehat{H}_{\theta_{\ell-1}}(X_\ell)\right\rangle,\\
A_4 &:= -\sum_{\ell=1}^{n}(\gamma_{\ell+1} - \gamma_\ell)\left\langle\nabla V(\theta_{\ell-1}), \mathbb{P}_{\theta_{\ell-1}}\widehat{H}_{\theta_{\ell-1}}(X_\ell)\right\rangle,\\
A_5 &:= -\gamma_1\left\langle\nabla V(\theta_0), \widehat{H}_{\theta_0}(X_1)\right\rangle + \gamma_{n+1}\left\langle\nabla V(\theta_n), \mathbb{P}_{\theta_n}\widehat{H}_{\theta_n}(X_{n+1})\right\rangle.
\end{aligned}$$

As $\widehat{H}_{\theta_\ell}(X_{\ell+1}) - \mathbb{P}_{\theta_\ell}\widehat{H}_{\theta_\ell}(X_\ell)$ is a martingale difference, it holds that $\mathbb{E}\left[A_1\right] = 0$. The upper bounds on the expectations of A_2, A_3 and A_4 are obtained similarly as in Karimi et al. (2019). Using (**A20**),

$$A_2 \le L^{\mathbb{P}\widehat{H}}\left(\sigma_{mse}\sum_{k=1}^{n}\gamma_k^2 + \frac{1}{2}\left(1 + 2a\sigma_{mse} + a\right)\sum_{k=0}^{n}\gamma_{k+1}^2\|h(\theta_k)\|^2\right).$$

By (**A18-21**),

$$A_3 \le L^V L_0^{\mathbb{P}\widehat{H}}\left(\left(1 + \sigma_{mse}\right)\sum_{k=1}^{n}\gamma_k^2 + \sum_{k=1}^{n}\gamma_k^2\|h(\theta_k)\|^2\right).$$

On the other hand,

$$A_4 \le L_0^{\mathbb{P}\widehat{H}}\left(\gamma_1 - \gamma_{n+1} + a'\sum_{k=1}^{n}\gamma_k^2\|h(\theta_{k-1})\|^2\right).$$

We now focus on A_5. As in the proof of (Karimi et al., 2019, Lemma 2), the expectation of the first term can be straightforwardly bounded by $\gamma_1\|h(\theta_0)\|L^{\widehat{H}}$ using the Cauchy–Schwarz inequality and (**A23**). The second term can, using (**A21**) and $\gamma_{n+1}\|h(\theta_n)\| \le 1 + \gamma_{n+1}^2\|h(\theta_n)\|^2$, be bounded in the same way according to

$$\begin{aligned}
\gamma_{n+1}\left\langle\nabla V(\theta_n), \mathbb{P}_{\theta_n}\widehat{H}_{\theta_n}(X_{n+1})\right\rangle \le L_0^{\mathbb{P}\widehat{H}}\gamma_{n+1}\|h(\theta_n)\| &\le L_0^{\mathbb{P}\widehat{H}}\left(1 + \gamma_{n+1}^2\|h(\theta_n)\|^2\right)\\
&\le L_0^{\mathbb{P}\widehat{H}}\left(1 + \sum_{\ell=0}^{n}\gamma_{\ell+1}^2\|h(\theta_\ell)\|^2\right).
\end{aligned}$$

The rest of the proof follows that of (Karimi et al., 2019, Theorem 2). □

C.2.2 Application to Theorem 5.4.1

The goal of this section is to establish that the assumptions of Theorem 5.4.1 ensure all the assumptions in section C.2.1, which in turn allows Theorem C.2.1 to be applied. First, we start by explicitly defining the kernel $\mathbb{P}_\theta$ and the function h in terms of the kernels presented in section C.1. We write $\mathbb{P}_{\theta,t}$ instead of $\mathbb{P}_\theta$ to explicit the dependence of the kernel on the *fixed* number of observations t.

C.2.2.1 Verification of the assumptions of Theorem C.2.1

For $(k_0, k) \in (\mathbb{N}_*)^2$ such that $k_0 < k$, define

$$\mathbb{P}_{\theta,t} : \mathsf{E}_t^{k-k_0} \times \boldsymbol{\mathcal{E}}_t^{\otimes(k-k_0)} \ni (\boldsymbol{y}_t[k_0:k], z_{0:t}[k_0:k], A) \mapsto \mathbb{K}_{\theta,t}^{k_0} \otimes \mathbb{K}_{\theta,t}^{\otimes(k-k_0)}(z_{0:t}[k], A), \tag{C.2.4}$$

where $\mathbb{K}_{\theta,t}$ is the PPG kernel defined in (C.1.17). Note that $\mathbb{P}_{\theta,t}$ depends only on the last frozen path, namely $z_{0:t}[k]$. Note also that, since $\mathbb{K}_{\theta,t}$ depends only on the paths, there is no dependence between $\boldsymbol{y}_{t,\ell}[k_0:k]$ and $\boldsymbol{y}_{t,\ell+1}[k_0:k]$. The score ascent algorithm (Algorithm 6) can be formulated as follows.

1. Sample $(z_{0:t,\ell}[k_0:k], \boldsymbol{y}_{t,\ell}[k_0:k]) \sim \mathbb{P}_{\theta_\ell,t}((z_{0:t,\ell-1}[k_0:k], \boldsymbol{y}_{t,\ell-1}[k_0:k]), \cdot)$.
2. Update the parameter according to $\eta_{\ell+1} = \eta_\ell + \gamma_{\ell+1} H(z_{0:t,\ell}[k_0:k], \boldsymbol{y}_{t,\ell}[k_0:k])$, where

$$H(z_{0:t,\ell}[k_0:k], \boldsymbol{y}_{t,\ell}[k_0:k]) = \frac{1}{k-k_0+1} \sum_{i=k_0}^{k} \mu(\boldsymbol{\beta}_{t,\ell}[i])(\mathrm{id}) = \Pi_{(k_0-1,k),N}(h_t),$$

where $\Pi_{(k_0-1,k),N}(h_t)$ is defined in (5.3.1). We denote by $\pi_{\theta,t}$ the invariant distribution of $\mathbb{P}_{\theta,t}$, which, by Proposition C.1.4, is given by $\pi_{\theta,t} = (\eta_{0:t} \otimes \mathbb{C}_t \mathbb{S}_t)^{\otimes(k-k_0)}$.

We also require the strong mixing assumption to hold uniformly in θ.

(**A24**) [Strong mixing uniformly in θ] For every $s \in \mathbb{N}$ there exist $\underline{\tau}_s$, $\bar{\tau}_s$, $\underline{\sigma}_s$, and $\bar{\sigma}_s$ in $\mathbb{R}_+^*$ such that for all $\theta \in \Theta$,

(i) $\underline{\tau}_s \le g_{s,\theta}(x_s) \le \bar{\tau}_s$ for every $x_s \in \mathsf{X}_s$,
(ii) $\underline{\sigma}_s \le m_{s,\theta}(x_s, x_{s+1}) \le \bar{\sigma}_s$ for every $(x_s, x_{s+1}) \in \mathsf{X}_{s:s+1}$.

Note that the assumption above implies that $\kappa_{\mathbb{N},t}$ is also uniform in θ.

Proof that (A17) holds.

Proposition C.2.2. *For all $\theta \in \Theta$, $h(\theta) = \nabla V(\theta)$, where $V(\theta) = \log \gamma_{0:t,\theta}(\mathsf{X}_{0:t})$ is the log-likelihood function.*

Proof. By Theorem C.1.6,

$$\begin{aligned} h(\theta) &= \int H(\tilde{\boldsymbol{y}}_t[k_0:k], \tilde{x}_{0:t}[k_0:k])\, \pi_{\theta,t}(\mathrm{d}(\tilde{\boldsymbol{y}}_t[k_0:k], \tilde{x}_{0:t}[k_0:k])) \\ &= \frac{1}{k-k_0+1} \sum_{i=k_0}^{k} \int [\eta_{0:t,\theta} \otimes \mathbb{C}_{t,\theta} \mathbb{S}_{t,\theta}]\, (\mathrm{d}(\tilde{\boldsymbol{y}}_t[i], \tilde{x}_{0:t}[i])) \mu(\tilde{\boldsymbol{\beta}}_{t,\ell}[i])(\mathrm{id}) \\ &= \eta_{0:t,\theta}\, (s_{0:t,\theta}) = \nabla V(\theta). \end{aligned}$$

□

Proof that (A18) holds. (**A18**) is trivially implied by (**A10**)(i).

Proof that (A19-21) hold. Let $\widehat{H}_\theta$ be given by

$$\widehat{H}_\theta : \mathsf{E}_t^{k-k_0} \ni (\boldsymbol{y}_t[k_0:k], z_{0:t}[k_0:k]) \mapsto \sum_{r=0}^{\infty} \{\mathbb{P}_{\theta,t}^r H(\boldsymbol{y}_t[k_0:k], z_{0:t}[k_0:k]) - h(\theta)\}. \tag{C.2.5}$$

Then the following holds true.

Lemma C.2.3. *Assume (**A24**). Then for all $\theta \in \Theta$ and $t \in \mathbb{N}_*$,*

$$\|\mathbb{P}_{\theta,t}\widehat{H}_\theta\|_\infty \leq \sigma_{bias}(1-\kappa_{N,t}^k)^{-1}\,.$$

Proof. By Theorem 5.3.1, we have for any $r > 0$

$$\left|\mathbb{P}_{\theta,t}^r H(\boldsymbol{y}_t[k_0:k], z_{0:t}[k_0:k]) - h(\theta)\right| \leq \sigma_{bias}\kappa_{N,t}^{(r-1)k}$$

and thus

$$\|\mathbb{P}_{\theta,t}\widehat{H}_\theta\|_\infty \leq \sum_{r=1}^{\infty}\left\|\mathbb{P}_{\theta,t}^r H - h(\theta)\right\|_\infty \leq \sigma_{bias}\sum_{r=0}^{\infty}\kappa_{N,t}^{rk} \leq \sigma_{bias}(1-\kappa_{N,t}^k)^{-1}\,,$$

where $\kappa_{N,t} \in (0,1)$. □

Lemma C.2.3 proves (**A19-21**) with $L_0^{\mathbb{P}\widehat{H}} := \sigma_{bias}(1-\kappa_{N,t}^k)^{-1}$.

Proof that (A20) holds.

Theorem C.2.4. *Assume (**A24**) and (**A10**). Then for every $t \in \mathbb{N}$, $\theta \in \Theta$ and $\mathbb{N} \in \mathbb{N}_*$ such that $N > 1 + 5\rho_t^2 t/2$,*

$$\left\|\mathbb{P}_{\theta_1,t}\widehat{H}_{\theta_1} - \mathbb{P}_{\theta_2,t}\widehat{H}_{\theta_2}\right\|_\infty \leq L^{\mathbb{P}\widehat{H}}\|\theta_1 - \theta_2\|\,,$$

where

$$L^{\mathbb{P}\widehat{H}} := \|L_2^P\|_\infty\left[1+\kappa_{\mathbb{N},t}^k(1-\kappa_{\mathbb{N},t}^k)\right] + L^V + \sigma_{bias}(1-\kappa_{\mathbb{N},t})^{-1}(1-\kappa_{\mathbb{N},t}^k)^{-1}\left[\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t}^k)^{-1} + L^\eta\kappa_{\mathbb{N},t}^k\right]. \quad \text{(C.2.6)}$$

Proof. We establish the claim by adapting the proof of (Karimi et al., 2019, Lemma 7). First, recall that the kernel $K_{\theta,t}$ defined in (C.1.18) is the path marginalized version of $\mathbb{K}_{\theta,t}$ given in (C.1.17). Note that for every $x \in \mathsf{E}_t^{k-k_0}$,

$$\mathbb{P}_{\theta_1,t}\widehat{H}_{\theta_1}(x) = \sum_{n=0}^{\infty}\delta_x\mathbb{P}_{\theta_1,t}\left\{\mathbb{P}_{\theta_1,t}^n H - h(\theta_1)\right\} = \sum_{n=0}^{\infty}\delta_x K_{\theta_1,t}^{kn}\left\{\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H\right\}\,,$$

where we have used (i) the fact that the backward statistics output by $\mathbb{P}_{\theta,t}$ are independent of the input backward statistics and (ii) the penultimate line in the computation of $h(\theta)$ above. We follow the proof of (Fort et al., 2011, Lemma 4.2) and consider the following decomposition: for $n \in \mathbb{N}_*$,

$$\begin{aligned}
&\delta_x K_{\theta_1,t}^{kn}\left(\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H\right) - \delta_x K_{\theta_2,t}^{kn}\left(\mathbb{P}_{\theta_2,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right) \qquad \text{(C.2.7)}\\
&\quad= \sum_{j=0}^{n-1}\left(\delta_x K_{\theta_1,t}^{kj} - \eta_{0:t,\theta_1}\right)\left(K_{\theta_1,t}^{kj} - K_{\theta_2,t}^{kj}\right)\left(K_{\theta_2,t}^{k(n-j-1)}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right)\\
&\quad\quad - \left(\delta_x K_{\theta_2,t}^{kn}\mathbb{P}_{\theta_2,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right) + \left(\delta_x K_{\theta_2,t}^{kn}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right)\\
&\quad\quad - \eta_{0:t,\theta_1}\left(K_{\theta_2,t}^{kn}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right)\,.
\end{aligned}$$

Applying Theorem C.1.7 with $\mu = \delta_x$ and $\nu = \eta_{0:t,\theta}$ and using the fact that $\eta_{0:t,\theta}K_{\theta,t}^\ell = \eta_{0:t,\theta}$ for all $\ell \in \mathbb{N}$, we obtain that for all $\ell \in \mathbb{N}$ and all $\theta \in \Theta$, $\left\|\delta_x K_{\theta,t}^\ell - \eta_{0:t,\theta}\right\|_{\mathrm{TV}} \leq \kappa_{N,t}^\ell$. Note that

by (**A10**)(iii), $K_{\theta,t}$ is Lipschitz; therefore, for all $r \in \mathbb{N}_*$, by Lemma C.3.10, $K^r_{\theta,t}$ is Lipschitz with constant $\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}$. Combining all this together, we obtain

$$\begin{aligned}
&\left|\left(\delta_x K^{kj}_{\theta_1,t} - \eta_{0:t,\theta_1}\right)\left(K^{kj}_{\theta_1,t} - K^{kj}_{\theta_2,t}\right)\left(K^{k(n-j-1)}_{\theta_2,t}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right)\right| \\
&\quad = \left|\left(\delta_x K^{kj}_{\theta_1,t} - \eta_{0:t,\theta_1}\right)\left(K^{kj}_{\theta_1,t} - K^{kj}_{\theta_2,t}\right)\left\{K^{k(n-j-1)}_{\theta_2,t}\left[\mathbb{P}_{\theta_1,t}H - h(\theta_1)\right] - \eta_{0:t,\theta_2}\left[\mathbb{P}_{\theta_1,t}H - h(\theta_1)\right]\right\}\right| \\
&\quad \leq \|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}\kappa^{kj}_{N,t}\kappa^{k(n-j-1)}_{N,t}\|\mathbb{P}_{\theta_1,t}H - h(\theta_1)\|_\infty\|\theta_1-\theta_2\| \\
&\quad \leq \sigma_{bias}\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}\kappa^{k(n-1)}_{N,t}\|\theta_1-\theta_2\|\,,
\end{aligned}$$

where the last inequality is due to Theorem 5.3.1. Therefore, the first term of the right side of (C.2.7) is upper bounded by $\sigma_{bias}\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}n\kappa^{k(n-1)}_{N,t}\|\theta_1-\theta_2\|$. The second term of (C.2.7) can be written

$$\begin{aligned}
&-\left(\delta_x K^{kn}_{\theta_2,t}\mathbb{P}_{\theta_2,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right) + \left(\delta_x K^{kn}_{\theta_2,t}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right) \\
&\qquad\qquad = \left(\delta_x K^{kn}_{\theta_2,t} - \eta_{0:t,\theta_2}\right)\left(\mathbb{P}_{\theta_1,t}H - \mathbb{P}_{\theta_2,t}H\right)\,,
\end{aligned}$$

and using again the ergodicity of $K_{\theta,t}$ and the fact that $\theta \mapsto \mathbb{P}_{\theta,t}H$ is uniformly Lipschitz by (**A10**)(iv), we may conclude that it is upper bounded by $\|L_2^P\|_\infty\kappa^{kn}_{N,t}\|\theta_1-\theta_2\|$. Finally, for the last term, using the facts that $K^k_{\theta,t}$ is $\eta_{0:t,\theta}$-invariant and geometrically ergodic and that $\theta \mapsto \eta_{0:t,\theta}$ is Lipschitz by (**A10**)(iv) yields

$$\begin{aligned}
&\left|\eta_{0:t,\theta_1}\left(K^{kn}_{\theta_2,t}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_1,t}H\right)\right| \\
&\quad = \left|\left(\eta_{0:t,\theta_1} - \eta_{0:t,\theta_2}\right)\left\{K^{kn}_{\theta_2,t}\left[\mathbb{P}_{\theta_1,t}H - h(\theta_1)\right] - \eta_{0:t,\theta_2}\left[\mathbb{P}_{\theta_1,t}H - h(\theta_1)\right]\right\}\right| \\
&\quad \leq L^\eta\kappa^{kn}_{N,t}\|\mathbb{P}_{\theta_1,t}H - h(\theta_1)\|_\infty\|\theta_1-\theta_2\| \\
&\quad \leq L^\eta\sigma_{bias}(1-\kappa_{N,t})^{-1}\kappa^{kn}_{N,t}\|\theta_1-\theta_2\|\,.
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
&\delta_x K^{kn}_{\theta_1,t}\left(\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H\right) - \delta_x K^{kn}_{\theta_2,t}\left(\mathbb{P}_{\theta_2,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right) \\
&\quad \leq \left\{\sigma_{bias}\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}n\kappa^{k(n-1)}_{N,t} + \left[\|L_2^P\|_\infty + L^\eta\sigma_{bias}(1-\kappa_{N,t})^{-1}\right]\kappa^{kn}_{N,t}\right\}\|\theta_1-\theta_2\|\,.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\left|\mathbb{P}_{\theta_1,t}\widehat{H}_{\theta_1}(x) - \mathbb{P}_{\theta_2,t}\widehat{H}_{\theta_2}(x)\right| \\
&\leq \left|\delta_x\mathbb{P}_{\theta_1,t}H - \delta_x\mathbb{P}_{\theta_2,t}H\right| + \left|\eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right| \\
&\qquad + \left|\sum_{n=1}^\infty \delta_x K^{kn}_{\theta_1,t}\left(\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H\right) - \delta_x K^{kn}_{\theta_2,t}\left(\mathbb{P}_{\theta_2,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right)\right| \\
&\leq \left|\delta_x\mathbb{P}_{\theta_1,t}H - \delta_x\mathbb{P}_{\theta_2,t}H\right| + \left|\eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\right| \\
&\quad + \left\{\sigma_{bias}\|L_1^P\|_\infty(1-\kappa_{\mathbb{N},t})^{-1}(1-\kappa^k_{\mathbb{N},t})^{-2}\right. \\
&\quad \left. + \left[\|L_2^P\|_\infty + L^\eta\sigma_{bias}(1-\kappa_{N,t})^{-1}\right]\kappa^k_{\mathbb{N},t}(1-\kappa^k_{\mathbb{N},t})^{-1}\right\}\|\theta_1-\theta_2\|\,.
\end{aligned}$$

To conclude, note that by (**A10**)(iv), $\|\delta_x\mathbb{P}_{\theta_1,t}H - \delta_x\mathbb{P}_{\theta_2,t}H\| \leq \|L_2^P\|_\infty\|\theta_1-\theta_2\|$. Furthermore, note that by Theorem C.1.6 we obtain that for all $\theta \in \Theta$, $\eta_{0:t,\theta}\mathbb{P}_{\theta,t}H = \eta_{0:t,\theta}s_{0:t,\theta} = \nabla V(\theta)$. Therefore, by (**A10**)(i) we obtain that $\|\eta_{0:t,\theta_1}\mathbb{P}_{\theta_1,t}H - \eta_{0:t,\theta_2}\mathbb{P}_{\theta_2,t}H\| \leq L^V\|\theta_1-\theta_2\|$, concluding the proof. □

Proof that (A22) holds. (**A22**) is simply a bound on the MSE of the roll-out `PPG` estimator, given by Theorem 5.3.1.

Proof that (A23) holds.

Proposition C.2.5. *For all $\theta \in \Theta$ and all $\ell \in [1 : t-1]$*

$$\mathbb{E}\left[\|\widehat{H}_\theta\| \mid \mathcal{F}_\ell\right] \leq 2\|s_{0:t,\theta}\|_\infty + \sigma_{bias}(1-\kappa_{\mathbb{N},t}^k)^{-1}\,.$$

Proof. Note that for all $x \in \mathsf{E}_t^{k-k_0}$ and all $\theta \in \Theta$,

$$\widehat{H}_\theta(x) = H(x) - h(\theta) + \mathbb{P}_{\theta,t}\widehat{H}_\theta(x)\,. \tag{C.2.8}$$

Lemma C.2.3 shows that $\|\mathbb{P}_{\theta,t}\widehat{H}_\theta\|_\infty \leq \sigma_{bias}(1-\kappa_{\mathbb{N},t}^k)^{-1}$. Note that $h(\theta) \leq \|s_{0:t,\theta}\|_\infty$ We write

$$\mathbb{E}\left[\|H\| \mid \mathcal{F}_\ell\right] \leq \frac{1}{(k-k_0+1)N}\sum_{i=k_0}^{k}\sum_{j=1}^{N}\mathbb{E}\left[\|\beta_{t,\ell}^j[i]\| \mid \mathcal{F}_\ell\right]\,.$$

By Proposition C.3.9, $\mathbb{E}\left[\|\beta_{t,\ell}^j[i]\| \mid \mathcal{F}_\ell\right] \leq \|s_{0:t,\theta}\|_\infty$, concluding the proof. □

(**A23**) follows directly by Proposition C.2.5 and by considering $\sup_{\theta\in\Theta}\|s_{0:t,\theta}\|_\infty$.

C.2.2.2 Proof of Theorem 5.4.1

We have shown in Section C.2.2.1 that under (**A10-24**), it is possible to apply Theorem C.2.1. To conclude the proof of Theorem 5.4.1 we just have to rearrange the constants. We start by rewriting the constant in Theorem C.2.4

$$L^{\mathbb{P}\widehat{H}} = C_1 + \sigma_{bias}(1-\kappa_{\mathbb{N},t})^{-1}(1-\kappa_{\mathbb{N},t}^k)^{-1}C_2,$$

with

$$\begin{aligned}
C_1 &= \left\|L_2^P\right\|_\infty\left[1+\kappa_{\mathbb{N},t}^k(1-\kappa_{\mathbb{N},t}^k)^{-1}\right] + L^V\\
C_2 &= \left\|L_1^P\right\|_\infty(1-\kappa_{\mathbb{N},t}^k)^{-1} + L^\eta\kappa_{\mathbb{N},t}^k\,.
\end{aligned}$$

By (C.2.2) and Lemma C.2.3,

$$\begin{aligned}
C_\gamma &= \sigma_{mse}L^{\mathbb{P}\widehat{H}} + (1+\sigma_{mse})L^V L_0^{\mathbb{P}\widehat{H}}\\
&= \sigma_{mse}\left[C_1 + \sigma_{bias}(1-\kappa_{\mathbb{N},t})^{-1}(1-\kappa_{\mathbb{N},t}^k)^{-1}C_2\right] + (1+\sigma_{mse})L^V\sigma_{bias}(1-\kappa_{N,t}^k)^{-1}\\
&= \sigma_{mse}C_1 + \sigma_{mse}\sigma_{bias}(1-\kappa_{N,t}^k)^{-1}\left[L^V + (1-\kappa_{\mathbb{N},t})^{-1}C_2\right] + \sigma_{bias}L^V(1-\kappa_{N,t}^k)^{-1}\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
C_{0,\gamma} &:= \sigma_{mse}^2L^V + C_\gamma\\
&= \sigma_{mse}^2L^V + \sigma_{mse}C_1 + \sigma_{mse}\sigma_{bias}(1-\kappa_{N,t}^k)^{-1}\left[L^V + (1-\kappa_{\mathbb{N},t})^{-1}C_2\right] + \sigma_{bias}L^V(1-\kappa_{N,t}^k)^{-1}\,.
\end{aligned}$$

In the same way, we can rewrite (C.2.3) as

$$\begin{aligned}
C_h &= L^{\mathbb{P}\widehat{H}}\left[(a+1)/2 + a\sigma_{mse}\right] + (L^V + a' + 1)L_0^{\mathbb{P}\widehat{H}}\\
&= \left[C_1 + \sigma_{bias}(1-\kappa_{\mathbb{N},t})^{-1}(1-\kappa_{\mathbb{N},t}^k)^{-1}C_2\right]\left[(a+1)/2 + a\sigma_{mse}\right] + (L^V + a' + 1)\sigma_{bias}(1-\kappa_{\mathbb{N},t}^k)^{-1}\,.
\end{aligned}$$

The constant C_0 from Theorem 5.4.1 is $L^{\widehat{H}} = 2\sup_{\theta\in\Theta}\|s_{0:t,\theta}\|_\infty + \sigma_{bias}(1-\kappa_{N,t}^k)^{-1}$ which completes the proof.

C.2.3 Conditions on the model to verify (A10)

In our specific application to score ascent, we work with the following assumptions.

(**A25**) (i) For all $t \in \mathbb{N}$, there exists $L_t^s \in \mathsf{M}(\mathsf{X}_{t:t+1})$ such that for all $(x_t, x_{t+1}) \in \mathsf{X}_{t:t+1}$, the function $\theta \mapsto s_{t,\theta}(x_t, x_{t+1})$ is $L_t^s(x_t, x_{t+1})$-Lipschitz and $\mathsf{X}_{t:t+1} \ni (x_t, x_{t+1}) \mapsto s_{t,\theta}(x_t, x_{t+1})$ is bounded by $\|s_t(\theta)\|_\infty$ for all $\theta \in \Theta$. Furthermore, $\|L_k^s\|_\infty < \infty$.

(ii) For all $t \in \mathbb{N}$, there exists $L_t^q \in \mathsf{X}_{\mathsf{t:t+1}}$ such that $\|L_t^q\|_\infty < \infty$ and that for all $(x_t, x_{t+1}) \in \mathsf{X}_{t:t+1}$, $\theta \mapsto q_{t,\theta}(x_t, x_{t+1})$ is $L_t^q(x_t, x_{t+1})$-Lipschitz.

Lemma C.2.6 ((**A18**)(i) holds)**.** *Assume (**A24**) and (**A10**). There exists a constant L^V such that the Lyapunov function V satisfies, for all $(\theta_1, \theta_2) \in \Theta^2$,*

$$\|\nabla V(\theta_1) - \nabla V(\theta_2)\| \leq L^V \|\theta_1 - \theta_2\|.$$

Proof. For all θ_1, θ_2,

$$\begin{aligned}\|\nabla V(\theta_1) - \nabla V(\theta_2)\| &= \|\eta_{0:t,\theta_1}(s_{0:t,\theta_1}) - \eta_{0:t,\theta_2}(s_{0:t,\theta_2})\| \\ &\leq \|\eta_{0:t,\theta_1}(s_{0:t,\theta_1}) - \eta_{0:t,\theta_1}(s_{0:t,\theta_2})\| + \|\eta_{0:t,\theta_1}(s_{0:t,\theta_2}) - \eta_{0:t,\theta_2}(s_{0:t,\theta_2})\|\,.\end{aligned}$$

By (9) and by (Gloaguen et al., 2022, Theorem 4.10) there exists a constant c such that

$$\|\eta_{0:t,\theta_1}(s_{0:t,\theta_2}) - \eta_{0:t,\theta_2}(s_{0:t,\theta_2})\| \leq ct\|\theta_1 - \theta_2\| \sup_\theta \sup_k \|s_k(\theta)\|_\infty\,,$$

Using (**A9**) and (**A10**)[i], we can write:

$$\begin{aligned}\|\eta_{0:t,\theta_1}(s_{0:t,\theta_1}) - \eta_{0:t,\theta_1}(s_{0:t,\theta_2})\| &\leq \sum_{u=0}^{t-1} \eta_{0:t,\theta_1}\left[\|s_{u,\theta_1}(x_{u:u+1}) - s_{u,\theta_2}(x_{u:u+1})\|\right], \\ &\leq \sum_{u=0}^{t-1} \eta_{0:t,\theta_1}\left[L_u^s(x_{u:u+1})\right] \|\theta_1 - \theta_2\|, \\ &\leq \frac{\sigma_+}{\sigma_-} \sup_{u\in[0:t-1]}\left[L_u^s\right] \|\theta_1 - \theta_2\| t.\end{aligned}$$

□

Theorem C.2.7 (Lipschitz continuity of Particle Gibbs with Backward Sampling)**.** *Assume (**A25**). For every $t \in \mathbb{N}$, $\theta \in \Theta$ and $\mathbb{N} \in \mathbb{N}_*$*

$$\sup_{x_{0:t}\in\mathsf{X}_{0:t}} \|K_{\theta_1,t}(x_{0:t}, .) - K_{\theta_2,t}(x_{0:t}, .)\|_{\mathrm{TV}} \leq L_{t,N}^K \|\theta_1 - \theta_2\|\,,$$

where

$$L_{t,N}^K := \sum_{\ell=0}^{t-1} \bar{\tau}_\ell^{-1}\left[\bar{\sigma}_\ell^{-1} + (N-1)\right] \|L_\ell^q\|_\infty\,. \tag{C.2.9}$$

Proof. We know that $K_{\theta,t} = \mathbb{C}_{m,\theta}\mathbb{B}_{t,\theta}$. Therefore, by Lemmas C.3.5, C.3.7 and C.3.11, we have that $K_{\theta,t}$ is Lipschitz with constant equals $L_t^{\mathbb{C}} + \sup_\theta \mathbb{C}_{t,\theta} L_t^{\mathbb{B}}$. □

Corollary C.2.8 ((**A10**)(iii) holds.)**.** *Assume (**A25**). For every $t \in \mathbb{N}$, $\theta \in \Theta$, $r \in \mathbb{N}_*$ and $\mathbb{N} \in \mathbb{N}_*$ such that $N > 1 + 5\rho_t^2 t/2$*

$$\sup_{x_{0:t}\in\mathsf{X}_{0:t}} \left\|K_{\theta_1,t}^r(x_{0:t}, .) - K_{\theta_2,t}^r(x_{0:t}, .)\right\|_{\mathrm{TV}} \leq L_{t,N}^P \|\theta_1 - \theta_2\|$$

where

$$L_{t,N}^P := (1 - \kappa_{t,N})^{-1} \|L_{t,N}^K\|_\infty \tag{C.2.10}$$

where $L_{t,N}^K$ is defined in (C.2.9).

Proof. Under 24, the Particle Gibbs with backward sampling is geometrically ergodic with contraction rate $\kappa_{t,N}$ and thus $L^K_{t,N}$ is bounded and the result follows from Lemma C.3.10 □

Corollary C.2.9 ((**A10**)(i))**.** *Assume (**A24**) and (**A25**). For all $t \in \mathbb{N}_*$, $(\theta_0, \theta_1) \in \Theta^2$,*

$$\|\eta_{0:t,\theta_0} - \eta_{0:t,\theta_1}\|_{\mathrm{TV}} \leq L^\eta \|\theta_0 - \theta_1\|,$$

where

$$L^\eta := L^P_{t,N^*}\,, \tag{C.2.11}$$

and $L^P_{t,N}$ is defined in (C.2.10) and $N^* = \lceil 1 + 5\rho_t^2/2 \rceil$.

Proof. Consider the following decomposition, valid for all $k \in \mathbb{N}^*$ and $N \geq 1 + 5\rho_t^2/2$, and all $x_{0:t} \in \mathsf{X}_{0:t}$,

$$\begin{aligned}\|\eta_{0:t,\theta_1} - \eta_{0:t,\theta_2}\|_{\mathrm{TV}} &\leq \left\|\eta_{0:t,\theta_1} - K^k_{\theta_1,t}(x_{0:t},\cdot)\right\|_{\mathrm{TV}} + \left\|\eta_{0:t,\theta_2} - K^k_{\theta_2,t}(x_{0:t},\cdot)\right\|_{\mathrm{TV}} + \left\|K^k_{\theta_1,t}(x_{0:t},\cdot) - K^k_{\theta_2,t}(x_{0:t},\cdot)\right\|_{\mathrm{TV}} \\ &\leq \left\|\eta_{0:t,\theta_1} - K^k_{\theta_1,t}(x_{0:t},\cdot)\right\|_{\mathrm{TV}} + \left\|\eta_{0:t,\theta_2} - K^k_{\theta_2,t}(x_{0:t},\cdot)\right\|_{\mathrm{TV}} + L^P_{t,N}\|\theta_1 - \theta_2\|\,,\end{aligned}$$

where we applied Corollary C.2.8. Since the Lipschitz constant of $K_{\theta,t}$ is independent of k, and $K_{\theta,t}$ is geometrically ergodic for all θ, we obtain by taking the limit when k goes to infinity with N fixed,

$$\|\eta_{0:t,\theta_1} - \eta_{0:t,\theta_2}\|_{\mathrm{TV}} \leq \frac{\|L^K_{t,N}\|_\infty}{1 - \kappa_{t,N}}\|\theta_1 - \theta_2\,\|\,,$$

for all $N \geq 1 + 5\rho_t^2/2$, where the dependence in N is hidden in $L^P_{t,N}$. The result follows by choosing $N = \lceil 1 + 5\rho_t^2/2 \rceil$. □

Remark C.2.10. *As noted by Lindholm and Lindsten (2018), the Lipschitz constant appearing in Corollary C.2.8 possesses an unexpected dependence on $N - 1$. One would expect it not to be true, in that we know that $\mathbb{K}_{\theta,t}$ converges geometrically fast and uniformly to $\eta_{0:t}$ and this is faster as N gets bigger. Therefore, for large N the Lipschitz constant is expected to converge to that of $\eta_{0:t}$ whose Lipschitz constant is independent of N.*

Proposition C.2.11 (Lipschitz continuity of $\theta \mapsto \mathbb{K}_{\theta,t}\mu(\boldsymbol{\beta}_t)(\mathrm{id})$)**.** *Assume (**A25**). For every $t \in \mathbb{N}$, $\theta \in \Theta$ and $\mathbb{N} \in \mathbb{N}_*$,*

$$\|\mathbb{K}_{\theta_1,t}\mu(\boldsymbol{\beta}_t)(\mathrm{id}) - \mathbb{K}_{\theta_2,t}\mu(\boldsymbol{\beta}_t)(\mathrm{id})\|_\infty \leq L^{\mathbb{K}}_t\|\theta_1 - \theta_2\|\,,$$

where

$$L^{\mathbb{K}}_t := (N-1)\sum_{\ell=0}^{t-1}\bar{\tau}_\ell\|L^q_\ell\|_\infty + \sum_{j=1}^{m}\|L^{\overleftarrow{Q}}_j\|_\infty\left[\sum_{\ell=0}^{m-1}s^\infty_\ell\right] + \sum_{j=1}^{m}\|L^s_j\|_\infty\,. \tag{C.2.12}$$

Proof. Consider $e = (x_{0:t}, \boldsymbol{y}_{0:t}) \in \mathsf{E}_t$ and $f_\theta(e) := \int \mathbb{S}_{m,\theta}(x_{0:t}, \mathrm{d}\tilde{\boldsymbol{y}}_t)\mu(\boldsymbol{b}_t)(id)$. Then $\mathbb{K}_{\theta,t}\mu(\boldsymbol{b}_t)(id) = \mathbb{C}_{m,\theta}f_\theta(x_{0:t})$ is a composition of a Markov kernel and a Lipschitz function, therefore Lipschitz. □

Corollary C.2.12 ((**A10**)(iv) holds.)**.** *Assume (**A25**). For every $t \in \mathbb{N}$, $\theta \in \Theta$ and $\mathbb{N} \in \mathbb{N}_*$*

$$\sup_{x_{0:t}\in\mathsf{X}_{0:t}}\|\mathbb{P}_{\theta_1,t}H - \mathbb{P}_{\theta_2,t}H\| \leq L^P_2\|\theta_1 - \theta_2\|\,,$$

where

$$L^P_2 = L^P_{t,N} + L^{\mathbb{K}}_t\,, \tag{C.2.13}$$

with L^P and $L^{\mathbb{K}}_t$ are defined in (C.2.12) and (C.2.10).

Proof. Let $\tilde{f} : \mathsf{E}^{k-k_0} \ni (x_{0:t}[k_0:k], \boldsymbol{x}_{0:t|t}[k_0:k], \boldsymbol{b}_t[k_0:k]) \mapsto (k-k_0)^{-1} \sum_{\ell=k_0+1}^{k} \mu(\boldsymbol{b}_t[\ell])(\mathrm{id})$. As $\mathbb{K}_{\theta,t}$ depends only on the path, with a slight abuse of notation, we can define $f_\theta(x_{0:t}) := \mathbb{K}_{\theta,t}^{\otimes k-k_0}(\tilde{f})(x_{0:t})$. By proposition C.2.11, we have that f_θ is Lipschitz with $L^f = L_t^{\mathbb{K}}$. Note that $\mathbb{P}_{\theta,t} H(x_{0:t}, \boldsymbol{y}_t) = K_{\theta,t}^{k_0} f_\theta(x_{0:t})$, therefore, by lemma C.3.11 Lipschitz with constant $L^P + L_t^{\mathbb{K}}$. □

C.3 Lipschitz properties

C.3.1 Lipschitz continuity of $\mathbb{P}_{\theta,}$

In this section we prove the following items:

- $\mathbb{C}_{m,\theta}(z_{0:m}, \cdot)$ is Lipschitz, see Section C.3.1.1
- $\mathbb{B}_{m,\theta}(\boldsymbol{x}_{0:m}, \cdot)$ is Lipschitz, see Section C.3.1.2
- $\int \mathbb{S}_{m,\theta}(\boldsymbol{x}_{0:m}, \mathrm{d}\boldsymbol{b}_m)\mu(\boldsymbol{b}_m)(\mathrm{Id})$ is Lipschitz, see Section C.3.1.3

The following technical lemma will be useful.

Lemma C.3.1. *Let $\alpha \in]0,1]$, $x \in \mathbb{R}_{\geq 0}$ and $\ell \in \mathbb{N}$. Then for all $\lambda_i \in \mathbb{R}_{\geq 0}$, $i \in [0:\ell]$, such that $\alpha \geq \prod_{i=0}^{\ell}(1-\lambda_i x)$ it holds that $\alpha \geq 1 - x\sum_{i=0}^{\ell}\lambda_i$.*

Proof. Consider first the case where $x\lambda_i \leq 1$ for all $i \in [0:\ell]$. We prove the result by induction. The case $\ell = 0$ is straightforward. Assume now that the result holds for some $r \in [0:\ell-1]$. Then,

$$\prod_{i=0}^{r+1}(1-\lambda_i x) = (1-\lambda_{r+1}x)\prod_{i=0}^{r}(1-\lambda_i x) \geq (1-\lambda_{r+1}x)(1-x\sum_{i=0}^{r}\lambda_i)$$
$$= 1 - x\sum_{i=0}^{r+1}\lambda_i + x^2\sum_{i=0}^{r}\lambda_i\lambda_{r+1} \geq 1 - x\sum_{i=0}^{r+1}\lambda_i\,.$$

Consider now the case where there is a index $j \in [0:\ell]$ such that $x\lambda_j \geq 1$. Then $\alpha \geq 0 \geq 1 - (\sum_{i=0}^{\ell}\lambda_i)x$. □

We begin with some important definitions. Let P and Q be probability distributions on some common measurable space $(\mathsf{X}, \mathcal{X})$, and assume that these distributions admit densities p and q w.r.t some common reference measure λ. Let $\mathbb{M}[P,Q]$ denote a maximal coupling between P and Q. As in (Lindholm and Lindsten, 2018, Theorem 2), it is possible to explicitly construct one such maximal coupling by

$$\mathbb{M}[P,Q](\mathrm{d}(x,y)) := \min\{p(x), g(x)\}\lambda(\mathrm{d}x)\delta_x(\mathrm{d}y) + \frac{[P(\mathrm{d}x) - \min\{p(x),g(x)\}\lambda(\mathrm{d}x)]\,[Q(\mathrm{d}y) - \min\{p(y),g(y)\}\lambda(\mathrm{d}y)]}{1 - \lambda(\min\{p,q\})}\,. \quad \text{(C.3.1)}$$

From this definition it follows that for continuous and discrete dominating measures λ,

$$\int \mathbb{1}_{\{x=y\}}\mathbb{M}[P,Q]\,\mathrm{d}(x,y) = \int \min\{p(x),g(x)\}\lambda(\mathrm{d}x)\,.$$

Moreover, for two Markov transition kernels K_1 and K_2 on $(\mathsf{X}, \mathcal{X})$, which are assumed to admit transition densities with respect to some common dominating measure, we let, for $(x_1,x_2) \in \mathsf{X}^2$, $\mathbb{M}[K_1,K_2]((x_1,x_2),\cdot)$ denote the maximal coupling between the measures $K_1(x_1,\cdot)$ and

$K_2(x_2, \cdot)$. Defined in this way, $\mathbb{M}[K_1, K_2]$ defines a Markov transition kernel on the product space $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$

The following Lemma will be crucial in what follows.

Lemma C.3.2. *(i) Let (μ_1, μ_2) be two probability measures admitting a density with respect to a common dominating measure and let (K_1, K_2) two Markov transition kernels also admitting transition densities with respect to some dominating measure. Then the probability measure*

$$\mathbb{M}[\mu_1, \mu_2]\mathbb{M}[K_1, K_2](\mathrm{d}(x_1, x_2)) = \int \mathbb{M}[\mu_1, \mu_2](\mathrm{d}(z_1, z_2))\mathbb{M}[K_1, K_2]((z_1, z_2), \mathrm{d}(x_1, x_2)),$$

is a coupling of $(\mu_1 K_1, \mu_2 K_2)$, and it holds that

$$\int \mathbb{1}_{x_1=x_2}\mathbb{M}[\mu_1 K_1, \mu_2 K_2](\mathrm{d}(x_1, x_2)) \geq \int\int \mathbb{1}_{z_1=z_2}\mathbb{1}_{x_1=x_2}\mathbb{M}[\mu_1, \mu_2](\mathrm{d}(z_1, z_2))\mathbb{M}[K_1, K_2]((z_1, z_2), \mathrm{d}(x_1, x_2)).$$

(ii) Let $(\mu_1, \cdots, \mu_n)$ and $(\nu_1, \cdots, \nu_n)$ be probability measures such that for all $i \in [1:n]$, μ_i and ν_i admit densities with respect to the same dominating measure. Then $\bigotimes_{i=1}^n \mathbb{M}[\mu_i, \nu_i]$ is a coupling of $\bigotimes_{i=1}^n \mu_i$ and $\bigotimes_{i=1}^n \nu_i$, and thus

$$\int \prod_{i=1}^n \mathbb{1}_{x_i=y_i}\mathbb{M}\left[\bigotimes_{i=1}^n \mu_i, \bigotimes_{i=1}^n \nu_i\right](\mathrm{d}(x_1, \ldots, x_n, y_1, \ldots, y_n)) \geq \int \prod_{i=1}^n \mathbb{1}_{x_i=y_i} \bigotimes_{i=1}^n \mathbb{M}[\mu_i, \nu_i](\mathrm{d}(x_1, \ldots, x_n, y_1, \ldots, y_n)).$$

Proof. It is enough to show that $\mathbb{M}[\mu_1, \mu_2]\mathbb{M}[K_1, K_2]$ admits $\mu_1 K_1$ and $\mu_2 K_2$ as marginal distributions. This follows immediately from the fact that $\mathbb{M}[\mu_1, \mu_1]$ and $\mathbb{M}[K_1, K_2]$ admit the right marginal distributions; indeed,

$$\begin{aligned}
&\mathbb{M}[\mu_1, \mu_2]\mathbb{M}[K_1, K_2](\mathsf{X} \times A) \\
&\quad = \int \mathbb{M}[\mu_1, \mu_2](\mathrm{d}z_1, \mathrm{d}_2)\,\mathbb{M}[K_1, K_2](z_1, z_2, \mathrm{d}(x_1, x_2))\,\mathbb{1}_{\mathsf{X}\times A}(x_1, x_2)\mathbb{1}_{\mathsf{X}^2}(z_1, z_2) \\
&\quad = \int \mathbb{M}[\mu_1, \mu_2](\mathrm{d}z_1, \mathrm{d}_2)K_2(z_2, A) \\
&\quad = \int \mu_2(\mathrm{d}z_2)K_2(z_2, A) \\
&\quad = \mu_2 K_2(A).
\end{aligned}$$

The derivation for the first marginal distribution follows similarly. For the second point, since $\mathbb{M}[\mu_1, \mu_2]\mathbb{M}[K_1, K_2]$ is a coupling of $(\mu_1 K_1, \mu_2 K_2)$ and $\mathbb{M}[\mu_1 K_1, \mu_2 K_2]$ is the maximal coupling, we have that

$$\begin{aligned}
&\int \mathbb{1}_{x_1=x_2}\mathbb{M}[\mu_1 K_1, \mu_2 K_2](\mathrm{d}(x_1, x_2)) \\
&\quad \geq \int \mathbb{1}_{x_1=x_2}\mathbb{M}[\mu_1, \mu_2](\mathrm{d}(z_1, z_2))\,\mathbb{M}[K_1, K_2](z_1, z_2; \mathrm{d}(x_1, x_2)) \\
&\quad \geq \int \mathbb{1}_{x_1=x_2}\mathbb{1}_{z_1=z_2}\mathbb{M}[\mu_1, \mu_2](\mathrm{d}(z_1, z_2))\,\mathbb{M}[K_1, K_2](z_1, z_2; \mathrm{d}(x_1, x_2)).
\end{aligned}$$

The proof of the second item follows similarly. □

C.3.1.1 $\theta \mapsto \mathbb{C}_{m,\theta}$ is Lipschitz.

We proceed by a coupling method that is inspired by (Lindholm and Lindsten, 2018, Theorem 2). The coupling we consider is that where the *selection* and *mutation* steps of the particle filter are respectively coupled maximally.

Algorithm 9 Coupling $\mathbb{C}_{m,\theta}$

Data: θ_1, θ_2, $\zeta_{0:m}$
Result: $\boldsymbol{x}_{0:m,1}$, $\boldsymbol{x}_{0:m,1}$
draw $\boldsymbol{x}_{0,1}, \boldsymbol{x}_{0,2} \sim \mathbb{M}\left[\boldsymbol{\eta}_0\langle\zeta_0\rangle, \boldsymbol{\eta}_0\langle\zeta_0\rangle\right]$
for $s \leftarrow 1$ **to** t **do**
  draw $(\boldsymbol{x}_{s,1}, \boldsymbol{x}_{s,2}) \sim \mathbb{M}\left[\boldsymbol{M}_{s-1,\theta_1}\langle\zeta_s\rangle(\boldsymbol{x}_{s-1,1}, \cdot), \boldsymbol{M}_{s-1,\theta_2}\langle\zeta_s\rangle(\boldsymbol{x}_{s-1,2}, \cdot)\right]$

First, let us prove that the one step *selection–mutation* kernel is Lipschitz.

Lemma C.3.3. *For all $t \in \mathbb{N}$, $\boldsymbol{x}_{t-1} \in \mathbf{X}_{t-1}$ and $(\theta_1, \theta_2) \in \Theta^2$,*

$$\int \mathbb{1}_{\{x_1=x_2\}} \mathbb{M}\left[\Phi_{t-1,\theta_1}(\mu(\boldsymbol{x}_{t-1})), \Phi_{t-1,\theta_2}(\mu(\boldsymbol{x}_{t-1}))\right](\mathrm{d}(x_1, x_2)) \geq 1 - \frac{\sum_{i=1}^N \lambda_t(L_{t-1}^q(x_{t-1}^i, \cdot))}{\mathbb{N}\bar{\tau}_n}\|\theta_1 - \theta_2\|. \tag{C.3.2}$$

Proof. By (**A9**)(i) and (**A10**)(iii),

$$\begin{aligned}
&\int \mathbb{1}_{\{x_1=x_2\}} \mathbb{M}\left[\Phi_{t-1,\theta_1}(\mu(\boldsymbol{x}_{t-1})), \Phi_{t-1,\theta_2}(\mu(\boldsymbol{x}_{t-1}))\right](\mathrm{d}(x_1, x_2)) \\
&\quad = \int \min\left(\sum_{i=1}^N \frac{q_{t-1,\theta_1}(x_{t-1}^i, x)}{\sum_{j=1}^N g_{t-1,\theta_1}(x_{t-1}^j)}, \sum_{i=1}^N \frac{q_{t-1,\theta_2}(x_{t-1}^i, x)}{\sum_{j=1}^N g_{t-1,\theta_2}(x_{t-1}^j)}\right) \lambda_t(\mathrm{d}x) \\
&\quad \geq \sum_{j=1}^N \int \min\left(\frac{q_{t-1,\theta_1}(x_{t-1}^i, x)}{\sum_{j=1}^N g_{t-1,\theta_1}(x_{t-1}^j)}, \frac{q_{t-1,\theta_2}(x_{t-1}^i, x)}{\sum_{j=1}^N g_{t-1,\theta_2}(x_{t-1}^j)}\right) \lambda_t(\mathrm{d}x) \\
&\quad \geq \frac{1}{\sum_{j=1}^N \max\left(g_{t-1,\theta_1}(x_{t-1}^j), g_{t-1,\theta_2}(x_{t-1}^j)\right)} \sum_{j=1}^N \int \min\left(q_{t-1,\theta_1}(x_{t-1}^j, x), q_{t-1,\theta_2}(x_{t-1}^j, x)\right) \lambda_t(\mathrm{d}x) \\
&\quad \geq \frac{\sum_{j=1}^N \max\left(g_{t-1,\theta_1}(x_{t-1}^j), g_{t-1,\theta_2}(x_{t-1}^j)\right) - \sum_{i=1}^N \lambda_t\left(L_{t-1}^q(x_{t-1}^i, \cdot)\right)\|\theta_1 - \theta_2\|}{\sum_{j=1}^N \max\left(g_{t-1,\theta_1}(x_{t-1}^j), g_{t-1,\theta_2}(x_{t-1}^j)\right)} \\
&\quad \geq 1 - \frac{\sum_{i=1}^N \lambda_t(L_{t-1}^q(x_{t-1}^i, \cdot))}{\mathbb{N}\bar{\tau}_n}\|\theta_1 - \theta_2\|,
\end{aligned}$$

where we have used that

$$\begin{aligned}
\int \max(q_{t-1,\theta_1}(x_{t-1}^i, x), q_{t-1,\theta_2}(x_{t-1}^i, x))\lambda_t(\mathrm{d}x) &\geq \max\left(\int q_{t-1,\theta_1}(x_{t-1}^i, x)\lambda_t(\mathrm{d}x), \int q_{t-1,\theta_2}(x_{t-1}^i, x)\lambda_t(\mathrm{d}x)\right) \\
&\geq \max(g_{t-1,\theta_1}(x_{t-1}^i), g_{t-1,\theta_2}(x_{t-1}^i)).
\end{aligned}$$

□

Lemma C.3.4. *For all $t \in \mathbb{N}$, $\boldsymbol{x}_{t-1} \in \mathbf{X}_{t-1}$, $z \in \mathsf{X}_t$ and $(\theta_1, \theta_2) \in \Theta^2$,*

$$\|\boldsymbol{M}_{t-1,\theta_1}\langle z\rangle(\boldsymbol{x}_{t-1}, \cdot) - \boldsymbol{M}_{t-1,\theta_2}\langle z\rangle(\boldsymbol{x}_{t-1}, \cdot)\|_{\mathrm{TV}} \leq L_{t-1}^{\boldsymbol{M}}(\boldsymbol{x}_{t-1})\|\theta_1 - \theta_2\|$$

where $L_{t-1}^{\boldsymbol{M}}(\boldsymbol{x}_{t-1}) = (1 - N^{-1})\bar{\tau}_{t-1}^{-1} \sum_{i=1}^N \lambda_t\left(L_{t-1}^q(x_{t-1}^i, \cdot)\right)$.

Proof. Let us denote by $\mathrm{U}[1:n]$ the uniform distribution on $[1:n]$. By definition of the kernel $\boldsymbol{M}_{t-1,\theta}\langle z\rangle$, we have that

$$\boldsymbol{M}_{t-1,\theta}\langle z\rangle(\boldsymbol{x}_{t-1},\mathrm{d}\boldsymbol{x}_t) = \int \mathrm{U}[1:n](\mathrm{d}j)\{\Phi_{t-1}(\mu(\boldsymbol{x}_{t-1}))^{\otimes j}\otimes\delta_z\otimes\Phi_{t-1}(\mu(\boldsymbol{x}_{t-1}))^{\otimes(\mathbb{N}-j-1)}\}(\mathrm{d}\boldsymbol{x}_t)$$

and thus, applying the two items of Lemma C.3.2 combined with the fact that $\mathbb{M}\left[\mu,\mu\right]\left(\mathrm{d}(x_1,x_2)\right) = \mu(\mathrm{d}x_1)\delta_{x_1}(\mathrm{d}x_2)$ for any probability measure μ, we get that

$$\begin{aligned}
&\int \mathbb{1}_{\{\boldsymbol{x}_{t,1}=\boldsymbol{x}_{t,2}\}}\mathbb{M}\left[\boldsymbol{M}_{t-1,\theta_1}\langle z\rangle(\boldsymbol{x}_{t-1},\cdot),\boldsymbol{M}_{t-1,\theta_2}\langle z\rangle(\boldsymbol{x}_{t-1},\cdot)\right]\mathrm{d}(\boldsymbol{x}_{t,1},\boldsymbol{x}_{t,2})\\
&\qquad\geq \int \mathbb{1}_{\boldsymbol{x}_{t,1}=\boldsymbol{x}_{t,2},i_1=i_2}\mathbb{M}\left[\mathrm{U}[1:n],\mathrm{U}[1:n]\right]\left(\mathrm{d}(i_1,i_2)\right)\\
&\qquad\qquad\times\mathbb{M}\left[\Phi_{t-1,\theta_1}(\mu(\boldsymbol{x}_{t-1})),\Phi_{t-1,\theta_2}(\mu(\boldsymbol{x}_{t-1}))\right]^{\otimes i_1}\otimes\mathbb{M}\left[\delta_z,\delta_z\right]\\
&\qquad\qquad\otimes\mathbb{M}\left[\Phi_{t-1,\theta_1}(\mu(\boldsymbol{x}_{t-1})),\Phi_{t-1,\theta_2}(\mu(\boldsymbol{x}_{t-1}))\right]^{\otimes N-i_1-1}\mathrm{d}(\boldsymbol{x}_{t,1},\boldsymbol{x}_{t,2})\\
&\qquad= \frac{1}{N}\sum_{i=1}^{N}\int\prod_{k=1,k\neq i}^{n}\mathbb{1}_{x^i_{t,1}=x^i_{t,2}}\mathbb{M}\left[\Phi_{t-1,\theta_1}(\mu(\boldsymbol{x}_{t-1})),\Phi_{t-1,\theta_2}(\mu(\boldsymbol{x}_{t-1}))\right]\left(\mathrm{d}(x^i_{t,1},x^i_{t,2})\right)\\
&\qquad\geq \left(1-\frac{\sum_{i=1}^{N}\lambda_t(L^q_{t-1}(x^i_{t-1},\cdot))}{\mathbb{N}\bar{\tau}_{t-1}}\|\theta_1-\theta_2\|\right)^{N-1}\\
&\qquad\geq 1-\frac{N-1}{\bar{\tau}_{t-1}\mathbb{N}}\sum_{i=1}^{N}\lambda_t(L^q_{t-1}(x^i_{t-1},\cdot))\|\theta_1-\theta_2\|\,.
\end{aligned}$$

where we have applied Lemma C.3.3 in the penultimate line and Lemma C.3.1 in the last one. □

Lemma C.3.5. *For every $t\in\mathbb{N}_*$, there exists $L^{\mathbb{C}}_t\in\mathsf{M}(\mathcal{X}_{0:t})$ such that*

$$\|\mathbb{C}_{t,\theta_1}(z_{0:t})-\mathbb{C}_{t,\theta_2}(z_{0:t})\|_{\mathrm{TV}}\leq L^{\mathbb{C}}_t(z_{0:t})\|\theta_1-\theta_2\|\,, \tag{C.3.3}$$

*where $L^{\mathbb{C}}_t(z_{0:t}) = \sup_\theta \mathbb{C}_{t,\theta}\left[\sum_{i=0}^{t-1}L^{\boldsymbol{M}}_i\right](z_{0:t})$. Under (**A25**)(i), we obtain that $\|L^{\mathbb{C}}_t\|_\infty\leq(N-1)\sum_{\ell=0}^{t-1}\bar{\tau}_\ell\|L^q_\ell\|_\infty$.*

Proof. This is a direct application of lemma C.3.13. □

C.3.1.2 $\theta\mapsto\mathbb{B}_{t,\theta}(\boldsymbol{x}_{0:t},\cdot)$ is Lipschitz

We start by recalling the definition of $\mathbb{B}_m$

$$\mathbb{B}_{t,\theta}:\mathbf{X}_{0:t}\times\mathcal{X}_{0:t}\ni(\boldsymbol{x}_{0:t},A)\mapsto\int\cdots\int\mathbb{1}_A(x_{0:t})\left(\prod_{s=0}^{t-1}\overleftarrow{Q}_{s,\mu(\boldsymbol{x}_s)}(x_{s+1},\mathrm{d}x_s)\right)\mu(\boldsymbol{x}_t)(\mathrm{d}x_t)\,. \tag{C.3.4}$$

Lemma C.3.6. *For all $s\in[0:t]$, $x_{t+1}\in\mathsf{X}_{t+1}$, $\boldsymbol{x}_t\in\mathbf{X}_t$ and $(\theta_1,\theta_2)\in\Theta^2$*

$$\left\|\overleftarrow{Q}_{s,\mu(\boldsymbol{x}_s),\theta_1}(x_{s+1},\cdot)-\overleftarrow{Q}_{s,\mu(\boldsymbol{x}_s),\theta_2}(x_{s+1},\cdot)\right\|_{\mathrm{TV}}\leq L^{\overleftarrow{Q}}_s(x_{s+1},\boldsymbol{x}_s)\|\theta_1-\theta_2\|\,. \tag{C.3.5}$$

*with $L^{\overleftarrow{Q}}_s(x_{s+1},\boldsymbol{x}_s) = (\mathbb{N}\bar{\tau}_t\bar{\sigma}_s)^{-1}\sum_{i=1}^{N}L^q_s(x^i_s,x_{s+1})$. Under (**A25**)(i), we have $\|L^{\overleftarrow{Q}}_m\|_\infty = (\bar{\tau}_m\bar{\sigma}_m)^{-1}\|L^q_m\|_\infty$.*

Proof. Note that $\overleftarrow{Q}_{t,\mu(\boldsymbol{x}_t)}(x_{t+1},\cdot) = \sum_{\ell=1}^{\mathbb{N}} \frac{q_t(x_t^\ell,x_{t+1})}{\sum_{\ell'=1}^{\mathbb{N}} q_t(x_t^{\ell'},x_{t+1})}\delta_{x_t^\ell}$. Therefore, similarly to the proof of Lemma C.3.3,

$$\begin{aligned}
&\int \mathbb{1}_{\{x_{t,1}=x_{t,2}\}}\mathbb{M}\left[\overleftarrow{Q}_{t,\mu(\boldsymbol{x}_t),\theta_1}(x_{t+1},\cdot), \overleftarrow{Q}_{t,\mu(\boldsymbol{x}_t),\theta_2}(x_{t+1},\cdot)\right]\mathrm{d}(x_{t,1},x_{t,2}) \\
&\qquad \geq \frac{\sum_{\ell=1}^N \max(q_{t,\theta_1}(x_t^\ell,x_{t+1}), q_{t,\theta_2}(x_t^\ell,x_{t+1})) - L_t^q(x_t^\ell,x_{t+1})\|\theta_1-\theta_2\|}{\sum_{\ell=1}^N \max(q_{t,\theta_1}(x_t^\ell,x_{t+1}), q_{t,\theta_2}(x_t^\ell,x_{t+1}))} \\
&\qquad \geq 1 - \frac{\sum_{\ell=1}^N L_t^q(x_t^\ell,x_{t+1})}{\mathbb{N}\bar{\tau}_t\bar{\sigma}_t}\|\theta_1-\theta_2\|\,.
\end{aligned}$$

□

Lemma C.3.7. *For all $t\in\mathbb{N}$, $\boldsymbol{x}_{0:t}\in\mathbf{X}_{0:t}$ and $(\theta_1,\theta_2)\in\Theta^2$*

$$\|\mathbb{B}_{t,\theta_1}(\boldsymbol{x}_{0:t},\cdot) - \mathbb{B}_{t,\theta_2}(\boldsymbol{x}_{0:t},\cdot)\|_{\mathrm{TV}} \leq L_t^{\mathbb{B}}(\boldsymbol{x}_{0:t})\|\theta_1-\theta_2\| \tag{C.3.6}$$

*where $L_t^{\mathbb{B}}(\boldsymbol{x}_{0:t}) = \sup_\theta \mathbb{B}_t\left[\sum_{i=0}^{t-1} L_i^{\overleftarrow{Q}}\right](\boldsymbol{x}_{0:t})$. Under (**A25**)(i), we have that $\|L_t^{\mathbb{B}}\|_\infty = \sum_{i=0}^{t-1}(\bar{\tau}_i\bar{\sigma}_i)^{-1}\|L_i^q\|_\infty$.*

Proof. Apply lemma C.3.11 and lemma C.3.6. □

C.3.1.3 $\theta\mapsto\int\mathbb{S}_{t,\theta}(\boldsymbol{x}_{0:t},\mathrm{d}\boldsymbol{b}_t)\mu(\boldsymbol{b}_t)(\mathrm{id})$ is Lipschitz

Define the backward ancestors kernel

$$\mathcal{B}_{\theta,t} : \mathsf{X}_{t+1}\times\mathbf{X}_t\times\sigma([1:N]) \mapsto \int \mathbb{1}_A(\tilde{j})\left(\sum_{\ell=1}^{\mathbb{N}}\frac{q_t(x_t^\ell,x_{t+1})}{\sum_{\ell'=1}^{\mathbb{N}} q_t(x_t^{\ell'},x_{t+1})}\delta_\ell(\mathrm{d}\tilde{j})\right).$$

Lemma C.3.8. *($\mathcal{B}_{\theta,t}$ is Lipschitz) For every $m\in[0:t]$, there exists $L_m^{BK}\in\mathsf{M}(\boldsymbol{\mathcal{X}}_{m:m+1})$ such that*

$$\|\mathcal{B}_{\theta_1,m}(x_{m+1},\boldsymbol{x}_m) - \mathcal{B}_{\theta_2,m}(x_{m+1},\boldsymbol{x}_m)\|_{\mathrm{TV}} \leq L_m^{\overleftarrow{Q}}(x_{m+1},\boldsymbol{x}_m)\|\theta_1-\theta_2\|\,, \tag{C.3.7}$$

where $L_s^{\overleftarrow{Q}}$ is defined in Lemma C.3.6

Proof. $\mathcal{B}_{\theta,s}$ is the index version of the kernel (C.3.4) and thus it is Lipschitz with the same constant. □

Proposition C.3.9. *For every $m\in[0:t]$, we have that*

$$\left|\int \mathbb{C}_m\mathbb{S}_{m,\theta}(z_{0:m},\mathrm{d}\boldsymbol{b}_m)\mu(\boldsymbol{b}_m)(\mathrm{Id})\right| \leq \sum_{\ell=0}^{m-1} s_\ell^\infty \tag{C.3.8}$$

and

$$\left|\int \mathbb{S}_{m,\theta_1}(\boldsymbol{x}_{0:m},\mathrm{d}\boldsymbol{b}_m)\mu(\boldsymbol{b}_m)(\mathrm{Id}) - \int \mathbb{S}_{m,\theta_2}(\boldsymbol{x}_{0:m},\mathrm{d}\boldsymbol{b}_m)\mu(\boldsymbol{b}_m)(\mathrm{Id})\right| \leq L_m^{\mathbb{S}\mu}(\boldsymbol{x}_{0:m})\|\theta_1-\theta_2\|\,. \tag{C.3.9}$$

where $L_m^{\mathbb{S}\mu}(\boldsymbol{x}_{0:m}) = N^{-1}\sum_{i=1}^N L_m^B(x_m^k,\boldsymbol{x}_{0:m})$ and L_m^B is defined recursively as

$$L_{m+1}^B(x_{m+1}^k,\boldsymbol{x}_{0:m}) = L_m^{\overleftarrow{Q}}(x_{m+1}^k,\boldsymbol{x}_m)\sum_{\ell=0}^m s_\ell^\infty + \int \mathcal{B}_{\theta,m}(x_{m+1}^k,\boldsymbol{x}_m,\mathrm{dJ})\left\{L_m^s(x_m^{\mathrm{J}},x_{m+1}^k) + L_m^B(x_m^{\mathrm{J}},\boldsymbol{x}_{0:m-1})\right\}. \tag{C.3.10}$$

*In particular, under (**A25**), we have that $L_m^B \leq \sum_{j=1}^m \|L_j^{\overleftarrow{Q}}\|_\infty\left[\sum_{\ell=0}^{m-1} s_\ell^\infty\right] + \sum_{j=1}^m \|L_j^s\|_\infty$.*

Proof. Consider the following kernels,

$$\widetilde{\mathbb{S}}_{m,\theta}(\boldsymbol{x}_{0:m+1}, \mathrm{d}(\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M}) := \prod_{\ell=0}^{m} \prod_{k=1}^{N} \widetilde{\boldsymbol{S}}_{\ell,\theta}(x_{\ell+1}^k, \boldsymbol{x}_\ell, \mathrm{d}(\mathrm{J}_\ell^{k,j})_{j=1}^{M}), \tag{C.3.11}$$

$$\widetilde{\boldsymbol{S}}_{\ell,\theta}(x_{\ell+1}^k, \boldsymbol{x}_\ell, \mathrm{d}(\mathrm{J}_\ell^{k,j})_{j=1}^{M}) := \prod_{j=1}^{M} \mathcal{B}_{\theta,\ell}(x_{\ell+1}^k, \boldsymbol{x}_\ell, \mathrm{dJ}_\ell^{k,j}). \tag{C.3.12}$$

Define for all $k \in [1:N]$, $m \in \mathbb{N}_{>0}$,

$$B_{m+1,k} : \theta \mapsto \int \widetilde{\mathbb{S}}_{m,\theta}(\boldsymbol{x}_{0:m+1}, \mathrm{d}(\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M}) b_{m+1}^k(\boldsymbol{x}_{0:m+1}, (\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M}),$$

where $b_{m+1}^k(\boldsymbol{x}_{0:m+1}, (\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M})$ is defined recursively as

$$b_{m+1}^k(\boldsymbol{x}_{0:m+1}, (\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M}) = M^{-1} \sum_{\ell=1}^{M} b_m^{\mathrm{J}_m^{k,\ell}}(\boldsymbol{x}_{0:m}, (\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_{m-1}^{i,j})_{i=1,j=1}^{N,M}) + s_{m,\theta}(x_m^{\mathrm{J}_m^{k,\ell}}, x_{m+1}^k).$$

For notational convenience, we henceforth drop the arguments and simply write b_{m+1}^k.

We herebelow show that $B_{m+1,k}$ is Lipschitz with constant $L_m^B(x_{m+1}^k, \boldsymbol{x}_m)$ and bounded by $\sum_{\ell=0}^{m-1} s_\ell^\infty$. For $m > 2$ and $k \in [1:N]$,

$$\begin{aligned}
B_{m+1,k}(\theta) &= \int \widetilde{\mathbb{S}}_{m,\theta}(\boldsymbol{x}_{0:m+1}, \mathrm{d}(\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_m^{i,j})_{i=1,j=1}^{N,M}) b_{m+1}^k \\
&= \int \cdots \int \widetilde{\mathbb{S}}_{m-1,\theta}(\boldsymbol{x}_{0:m}, \mathrm{d}(\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_{m-1}^{i,j})_{i=1,j=1}^{N,M}) \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m+1}^k, \boldsymbol{x}_m, \mathrm{d}(\mathrm{J}_m^{k,j})_{j=1}^{M}) \\
&\qquad \times \left\{ M^{-1} \sum_{\ell=1}^{M} b_m^{\mathrm{J}_m^{k,\ell}} + s_{m,\theta}(x_m^{\mathrm{J}_m^{k,\ell}}, x_{m+1}^k) \right\} \\
&= \int \cdots \int \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m+1}^k, \boldsymbol{x}_m, \mathrm{d}\{\mathrm{J}_m^{k,j}\}_{j=1}^{M}) \Big[M^{-1} \sum_{\ell=1}^{M} \Big\{ s_{m,\theta}(x_m^{\mathrm{J}_m^{k,\ell}}, x_{m+1}^k) \\
&\qquad + \int \widetilde{\mathbb{S}}_{m-1,\theta}(\boldsymbol{x}_{0:m}, \mathrm{d}(\mathrm{J}_0^{i,j}, \ldots, \mathrm{J}_{m-1}^{i,j})_{i=1,j=1}^{N,M}) b_m^{\mathrm{J}_m^{k,\ell}} \Big\} \Big] \\
&= \int \cdots \int \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m+1}^k, \boldsymbol{x}_m, \mathrm{d}(\mathrm{J}_m^{k,j})_{j=1}^{M}) \left[M^{-1} \sum_{\ell=1}^{M} \left\{ s_{m,\theta}(x_m^{\mathrm{J}_m^{k,\ell}}, x_{m+1}^k) + B_{m,\mathrm{J}_m^{k,\ell}}(\theta) \right\} \right] \\
&= \int \mathcal{B}_{\theta,m}(x_{m+1}^k, \boldsymbol{x}_m, \mathrm{dJ}) \left\{ s_{m,\theta}(x_m^{\mathrm{J}}, x_{m+1}^k) + B_{m,\mathrm{J}}(\theta) \right\}
\end{aligned}$$

Applying the induction hypothesis conditionally on $\mathrm{J}_m^{k,\ell}$, $B_{m,\mathrm{J}_m^{k,\ell}}$ is Lipschitz with constant $L_m^B(x_m^{\mathrm{J}_m^{k,\ell}}, \boldsymbol{x}_{0:m-1})$ and thus the Lipschitz constant of $B_{m+1,k}$ is

$$L_{m+1}^B(x_{m+1}^k, \boldsymbol{x}_{0:m}) = L_m^{\overleftarrow{Q}}(x_{m+1}^k, \boldsymbol{x}_m) \sum_{\ell=0}^{m} s_\ell^\infty + \int \mathcal{B}_{\theta,m}(x_{m+1}^k, \boldsymbol{x}_m, \mathrm{dJ}) \left\{ L_m^s(x_m^{\mathrm{J}}, x_{m+1}^k) + L_m^B(x_m^{\mathrm{J}}, \boldsymbol{x}_{0:m-1}) \right\}. \tag{C.3.13}$$

where we have used the fact that $\mathcal{B}_{\theta,m}$ and $s_{m,\theta}$ are also Lipschitz. Again by induction $B_{m+1,k}$ is bounded uniformly by $\sum_{\ell=0}^{m} s_\ell^\infty$. The induction is concluded by noting that for the base case $m = 0$, $\beta_m^k = 0$ for all $k \in \mathbb{N}$ and thus the result holds.

It now remains to check that for all $\theta \in \Theta$, $m \in [0:t]$ and $k \in [1:N]$,

$$B_{m,k}(\theta) = \int \mathbb{S}_m(\boldsymbol{x}_{0:m}, \mathrm{d}\boldsymbol{b}_m) b_m^k.$$

Again, we proceed by induction.

$$
\begin{aligned}
&\int \mathbb{S}_m(\boldsymbol{x}_{0:m}, \mathrm{d}\boldsymbol{b}_m) b_m^k \\
&= \int \cdots \int \mathbb{S}_{m-1}(\boldsymbol{x}_{0:m-1}, \mathrm{d}\boldsymbol{b}_{m-1}) \boldsymbol{S}_m(\boldsymbol{b}_{m-1}, \boldsymbol{x}_{m-1:m}, \mathrm{d}\boldsymbol{b}_m) b_m^k \\
&= \int \cdots \int \mathbb{S}_{m-1}(\boldsymbol{x}_{0:m-1}, \mathrm{d}\boldsymbol{b}_{m-1}) \\
&\qquad \times \prod_{j=1}^{M} \left(\sum_{p=1}^{N} \frac{q_{m-1}(x_{m-1}^p, x_m^k)}{\sum_{\ell=1}^{N} q_{m-1}(x_{m-1}^\ell, x_m^k)} \delta_{x_{m-1}^p, b_{m-1}^p}(\mathrm{d}(\tilde{x}_{m-1}^{k,j}, \tilde{b}_{m-1}^{k,j})) \right) \\
&\qquad \times \left[M^{-1} \sum_{n=1}^{M} \left\{ \tilde{b}_{m-1}^{k,n} + s_{m,\theta}(\tilde{x}_{m-1}^{k,n}, x_m^k) \right\} \right] \\
&= \int \cdots \int \mathbb{S}_{m-1}(\boldsymbol{x}_{0:m-1}, \mathrm{d}\boldsymbol{b}_{m-1}) \\
&\qquad \times \prod_{j=1}^{M} \left(\sum_{p=1}^{N} \frac{q_{m-1}(x_{m-1}^p, x_m^k)}{\sum_{\ell=1}^{N} q_{m-1}(x_{m-1}^\ell, x_m^k)} \delta_p(\mathrm{dJ}_{m-1}^{k,j}) \right) \left[M^{-1} \sum_{n=1}^{M} \left\{ b_{m-1}^{\mathrm{J}_{m-1}^{k,n}} + s_{m,\theta}(x_{m-1}^{\mathrm{J}_{m-1}^{k,n}}, x_m^k) \right\} \right] \\
&= \int \cdots \int \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m-1}^k, \boldsymbol{x}_{\ell-1}, \mathrm{d}(\mathrm{J}_{\ell-1}^{k,j})_{j=1}^{M}) \\
&\quad \times \left[M^{-1} \sum_{\ell=1}^{M} \left\{ s_{m,\theta}(x_{m-1}^{\mathrm{J}_{m-1}^{k,\ell}}, x_m^k) + \mathbb{S}_{m-1}(\boldsymbol{x}_{0:m-1}, \mathrm{d}\boldsymbol{b}_{m-1}) b_{m-1}^{\mathrm{J}_{m-1}^{k,\ell}} \right\} \right] \\
&= \int \cdots \int \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m-1}^k, \boldsymbol{x}_{\ell-1}, \mathrm{d}(\mathrm{J}_{\ell-1}^{k,j})_{j=1}^{M}) \\
&\quad \times \left[M^{-1} \sum_{\ell=1}^{M} \left\{ s_{m,\theta}(x_{m-1}^{\mathrm{J}_{m-1}^{k,\ell}}, x_m^k) + \int \mathbb{S}_{m-1}(\boldsymbol{x}_{0:m-1}, \mathrm{d}\boldsymbol{b}_{m-1}) b_{m-1}^{\mathrm{J}_{m-1}^{k,\ell}} \right\} \right] \\
&= \int \cdots \int \widetilde{\boldsymbol{S}}_{m,\theta}(x_{m-1}^k, \boldsymbol{x}_{\ell-1}, \mathrm{d}(\mathrm{J}_{\ell-1}^{k,j})_{j=1}^{M}) \left[M^{-1} \sum_{\ell=1}^{M} \left\{ s_{m,\theta}(x_{m-1}^{\mathrm{J}_{m-1}^{k,\ell}}, x_m^k) + B_{m-1,\mathrm{J}_{m-1}^{k,\ell}}(\theta) \right\} \right] \\
&= B_{m,k}(\theta)
\end{aligned}
$$

The proof is finalized by noting that

$$
\int \mathbb{S}_m(\boldsymbol{x}_{0:m}, \mathrm{d}\boldsymbol{b}_m) \mu(\boldsymbol{b}_m)(\mathrm{Id}) = N^{-1} \sum_{k=1}^{N} B_{m,k}(\theta)
$$

and thus it is Lipschitz with constant $L_m^{\mathbb{S}\mu}(\boldsymbol{x}_{0:m}) = N^{-1} \sum_{i=1}^{N} L_m^B(x_m^k, \boldsymbol{x}_{m-1})$. □

C.3.2 Lipschitz properties of Markov Kernels

Lemma C.3.10 (Composition of ergodic Lipschitz kernels is lipschitz)**.** *Let P_θ be a Markov kernel over $X \times \mathcal{Y}$ that is uniformly π-geometrically ergodic for any θ with contraction constant ρ independent of θ and such that there exists $L_p > 0$ such that for every $x \in \mathsf{X}$*

$$
\|P_{\theta_0}(x, \cdot) - P_{\theta_1}(x, \cdot)\|_{\mathrm{TV}} \leq L_P \|\theta_0 - \theta_1\|.
$$

Then, for all $k > 0$

$$
\left\| P_{\theta_0}^k(x, \cdot) - P_{\theta_1}^k(x, \cdot) \right\|_{\mathrm{TV}} \leq \frac{L_P}{1 - \rho} \|\theta_0 - \theta_1\|.
$$

Proof. We use the following decomposition borrowed from Fort et al. (2011). For any $k \geq 1$,

$$P_{\theta_0}^k f - P_{\theta_1}^k f = \sum_{j=0}^{k-1} P_{\theta_0}^j (P_{\theta_0} - P_{\theta_1})(P_{\theta_1}^{k-j-1} f - \pi f).$$

Then, for any f s.t. $\|f\|_\infty \leq 1$ and $x \in \mathsf{X}$,

$$\begin{aligned}
|P_{\theta_0}^k f(x) - P_{\theta_1}^k f(x)| &\leq \sum_{j=0}^{k-1} \left| \int P_{\theta_0}^j (x, \mathrm{d}y) \sup_{z \in \mathsf{X}} |P_{\theta_1}^{k-j-1} f(z) - \pi f| \right| L_P \|\theta_0 - \theta_1\| \\
&\leq L_P \Big(\sum_{j=0}^{k-1} \rho^{k-j-1} \Big) \|\theta_0 - \theta_1\| \\
&\leq \frac{L_P}{1-\rho} \|\theta_0 - \theta_1\|.
\end{aligned}$$

□

Lemma C.3.11 (Composition of Lipschitz kernels is lipschitz)**.** *Let P_θ, Q_θ be two kernels defined over $X \times \mathcal{Y}$ and $Y \times \mathcal{Z}$ such that for ever $x \in X$, $y \in Y$ there are $L_p \in \mathsf{M}(X)$, $L_q \in \mathsf{M}(Y)$ that satisfy*

$$\|P_{\theta_0}(x, \cdot) - P_{\theta_1}(x, \cdot)\|_{\mathrm{TV}} \leq L_p(x) \|\theta_0 - \theta_1\|$$

and

$$\|Q_{\theta_0}(y, \cdot) - Q_{\theta_1}(y, \cdot)\|_{\mathrm{TV}} \leq L_q(y) \|\theta_0 - \theta_1\|.$$

Then

$$\|P_{\theta_0} Q_{\theta_0}(x, \cdot) - P_{\theta_1} Q_{\theta_1}(x, \cdot)\|_{\mathrm{TV}} \leq L_{pq}(x) \|\theta_0 - \theta_1\|,$$

where $L_{pq}(x) = (\sup_\theta P_\theta L_q(x) + L_p(x) \sup_y \sup_\theta Q_\theta(y, Z))$.

Proof. Let $f \in \mathsf{M}$ such that $\|f\|_\infty \leq 1$.

$$\begin{aligned}
\|P_{\theta_1} Q_{\theta_1} f - P_{\theta_2} Q_{\theta_2} f\| &\leq \|P_{\theta_1} [Q_{\theta_1} f - Q_{\theta_2} f]\| + \|(P_{\theta_1} - P_{\theta_2}) Q_{\theta_2} f\| \\
&\leq (P_{\theta_1} L_q(x) + L_p(x) \|Q_{\theta_2} f\|_\infty) \|\theta_1 - \theta_2\|.
\end{aligned}$$

□

Corollary C.3.12. *Let P_θ, Q_θ be two Markov kernels defined over $X \times \mathcal{Y}$ and $Y \times \mathcal{Z}$ such that for ever $x \in X$, $y \in Y$ there are $L_p \in \mathsf{M}(X)$, $L_q \in \mathsf{M}(Y)$ that satisfy*

$$\|P_{\theta_0}(x, \cdot) - P_{\theta_1}(x, \cdot)\|_{\mathrm{TV}} \leq L_p(x) \|\theta_0 - \theta_1\|$$

and

$$\|Q_{\theta_0}(y, \cdot) - Q_{\theta_1}(y, \cdot)\|_{\mathrm{TV}} \leq L_q(y) \|\theta_0 - \theta_1\|.$$

Then

$$\|P_{\theta_0} Q_{\theta_0}(x, \cdot) - P_{\theta_1} Q_{\theta_1}(x, \cdot)\|_{\mathrm{TV}} \leq L_{pq}(x) \|\theta_0 - \theta_1\|,$$

where $L_{pq}(x) = (\sup_\theta P_\theta L_q(x) + L_p(x))$.

Lemma C.3.13 (Product of Lipschitz kernels is lipschitz)**.** *Let P_θ, Q_θ be two Markov kernels that are uniformly Lipschitz with constants L_P, L_Q. Then $P_\theta \otimes Q_\theta$ is uniformly Lipschitz with constant $L_P + L_Q$.*

Proof. Let $h_\theta : y \mapsto \int Q_\theta(y, \mathrm{d}z) f(y, z)$. Then $(P_{\theta_i} \otimes Q_{\theta_i})(f) = P_{\theta_i}(h_{\theta_i})$ and the proof is similar to that of the previous Lemma since h_θ is Lipschitz with constant L_Q and $\|h_\theta\|_\infty \leq 1$. □

C.3.3 PPG

All the experiments were performed on a server equipped with 7 A40 Nvidia GPUs. The algorithms were implemented in Python with the JAX Python package Bradbury et al. (2018) and run on GPU.

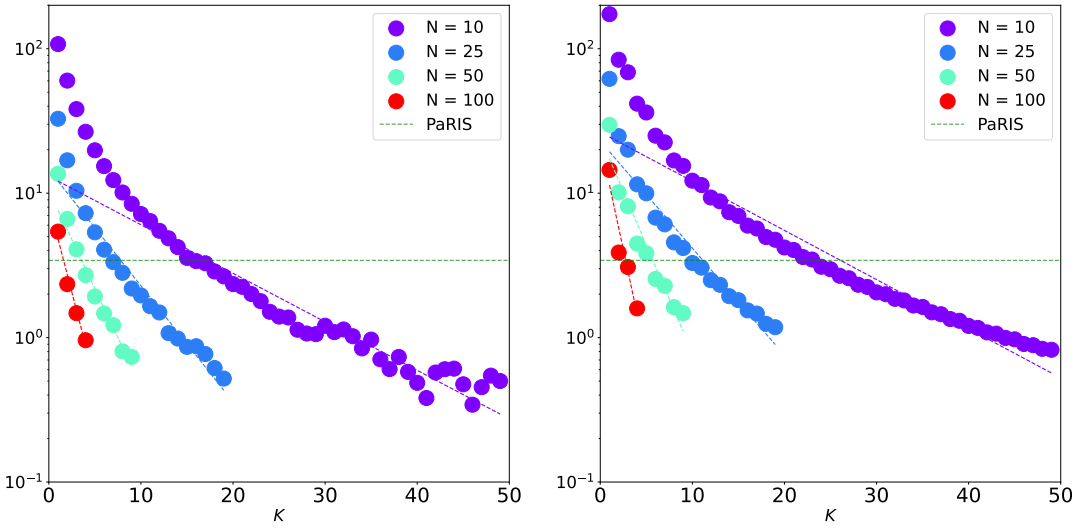

Figure C.1: Output of the `PPG` *roll-out estimator for the LGSSM. The curves describe the evolution of the bias with increasing k for different particle sample sizes $\mathbb{N}$. The left and right panels correspond to $k_0 = k - 1$ and $k_0 = \lfloor k/2 \rfloor$, respectively.*

C.3.4 Learning

For both experiments, all the parameters were initialized by sampling from a centered multivariate gaussian distribution with covariance matrix of $0.01I$. We have used the ADAM optimizer with a learning rate decay of $1/\sqrt{\ell}$ where ℓ is the iteration index, with a starting learning rate of 0.2. We rescale the gradients by T.

LGSSM For LGSSM we evaluated for fixed number of particles ($N = 64$) and number of gibbs iterations ($k = 8$) the influence of the burn-in phase (k_0) over the final distance obtained to the MLE estimator. Table C.1 indicates that configurations with smaller k_0 perform better. A possible interpretation of this phenomenon is that, since between two gradient ascent iterates the conditioning path is being passed on, this conditioning path from a moment on makes the estimates less biased, so the importance of having k_0 high to have less bias vanishes, but the effect of augmenting the variance with k_0 is still shown, since the fact of having a conditioning particle from the right marginal does not affect the variance of the estimator, only it's bias.

Algorithm	N	k_0	k	D_{mle}
PPG	64	0	8	0.205 ± 0.013
PPG	64	1	8	0.213 ± 0.016
PPG	64	2	8	0.201 ± 0.010
PPG	64	3	8	0.201 ± 0.010
PPG	64	4	8	0.207 ± 0.012
PPG	64	5	8	0.212 ± 0.015
PPG	64	6	8	0.210 ± 0.017
PPG	64	7	8	0.211 ± 0.018

Table C.1: Distance to θ_{MLE} for each configuration in the LGSSM case.

Appendix D

Appendix of Chapter 6

D.1 `SMCdiff` extension

The identity (6.2.14) allows us to extend `SMCdiff` Trippe et al. (2023) to handle noisy inverse problems as we now show. We have that

$$\begin{aligned}\phi_\tau^{\tilde{y}_\tau}(\underline{x}_\tau) &= \frac{\int p_\tau(\tilde{y}_\tau \frown \underline{x}_\tau | x_{\tau+1}) \left\{\prod_{s=\tau+1}^{n-1} p_s(\mathrm{d}x_s|x_{s+1})\right\} \mathsf{p}_n(\mathrm{d}x_n)}{\int \mathsf{p}_\tau(\tilde{y}_\tau \frown \underline{z}_\tau)\mathrm{d}\underline{z}_\tau} \\ &= \int b_{\tau:n}^{\tilde{y}_\tau}(\underline{x}_{\tau:n}|\overline{x}_{\tau+1:n}) f_{\tau+1:n}^{\tilde{y}_\tau}(\mathrm{d}\overline{x}_{\tau+1:n})\mathrm{d}\underline{x}_{\tau+1:n}\,,\end{aligned}$$

where

$$\begin{aligned}b_{\tau:n}(\underline{x}_{\tau:n}|\overline{x}_{\tau+1:n}) &= \frac{p_\tau(\tilde{y}_\tau \frown \underline{x}_\tau|x_{\tau+1})\left\{\prod_{s=\tau+1}^{n-1} \underline{p}_s(\underline{x}_s|x_{s+1})\overline{p}_s(\overline{x}_s|x_{s+1})\right\}\underline{\mathsf{p}}_n(\underline{x}_n)}{\mathsf{L}_{\tau:n}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n})}\,, \\ f_{\tau+1:n}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n}) &= \frac{\mathsf{L}_{\tau:n}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n})}{\int \mathsf{p}_\tau(\tilde{y}_\tau \frown \underline{z}_\tau)\mathrm{d}\underline{z}_\tau}\,,\end{aligned}$$

and

$$\mathsf{L}_{\tau:n}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n}) = \int p_\tau(\tilde{y}_\tau \frown \underline{z}_\tau | \overline{x}_{\tau+1} \frown \underline{z}_{\tau+1}) \left\{\prod_{s=\tau+1}^{n-1} \underline{p}_s(\mathrm{d}\underline{z}_s|\overline{x}_{s+1} \frown \underline{z}_{s+1})\overline{p}_s(\overline{x}_s|\overline{x}_{s+1} \frown \underline{z}_{s+1})\right\} \underline{\mathsf{p}}_n(\mathrm{d}\underline{z}_n)\,.$$

Next, (6.2.13) implies that

$$\begin{aligned}\int \mathsf{p}_{s+1}(\overline{x}_{s+1} \frown \underline{z}_{s+1})\underline{p}_s(\mathrm{d}\underline{z}_s|\overline{x}_{s+1} \frown \underline{z}_{s+1})\overline{p}_s(\overline{x}_s|\overline{x}_{s+1} \frown \underline{z}_{s+1})\mathrm{d}\underline{z}_{s:s+1} = \\ \int \mathsf{p}_s(\overline{x}_s \frown \underline{z}_s)\overline{q}_{s+1}(\overline{x}_{s+1}|\overline{x}_s)\underline{q}_{s+1}(\underline{z}_{s+1}|\underline{z}_s)\mathrm{d}\underline{z}_{s:s+1}\,,\end{aligned}$$

and applied repeatedly, we find that

$$\mathsf{L}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n}) = \int \mathsf{p}_\tau(\tilde{y}_\tau \frown \underline{x}_\tau)\mathrm{d}\underline{x}_\tau \cdot \int \delta_{\tilde{y}_\tau}(\mathrm{d}\overline{x}_\tau) \prod_{s=\tau+1}^{n} \overline{q}_s(\overline{x}_s|\overline{x}_{s-1})\,.$$

and thus, $f_{\tau:n}^{\tilde{y}_\tau}(\overline{x}_{\tau+1:n}) = \int \delta_{\tilde{y}_\tau}(\mathrm{d}\overline{x}_\tau)\prod_{s=\tau+1}^{n}\overline{q}_s(\overline{x}_s|\overline{x}_{s-1})$. In order to approximate $\phi_\tau^{\tilde{y}_\tau}$ we first diffuse the noised observation up to time n, resulting in $\overline{x}_{\tau+1:n}$, and then estimate $b_{\tau+1:n}^{\tilde{y}_\tau}(\cdot|\overline{x}_{\tau+1:n})$ using a particle filter with $\underline{p}_s(\underline{x}_s|x_{s+1})$ as transition kernel at step $s \in [\tau+1:n]$ and $g_s : \underline{z}_s \mapsto \overline{p}_{s-1}(\overline{x}_{s-1}|\overline{x}_s \frown \underline{z}_s)$ as potential, similarly to `SMCdiff`.

D.2 Proofs

D.2.1 Proof of Proposition 6.2.2

Preliminary definitions.

We preface the proof with notations and definitions of a few quantities that will be used throughout.

For a probability measure μ and f a bounded measurable function, we write $\mu(f) := \int f(x)\mu(\mathrm{d}x)$ the expectation of f under μ and if $K(\mathrm{d}x|z)$ is a transition kernel we write $K(f)(z) := \int f(x)K(\mathrm{d}x|z)$.

Define the *smoothing* distribution

$$\phi^y_{0:n}(\mathrm{d}x_{0:n}) \propto \delta_y(\mathrm{d}\overline{x}_0)\mathsf{p}_{0:n}(x_{0:n})\mathrm{d}\underline{x}_0\mathrm{d}x_{1:n}\,, \tag{D.2.1}$$

which admits the posterior ϕ^y_0 as time 0 marginal. Its particle estimate known as the *poor man smoother* is given by

$$\phi^N_{0:n}(\mathrm{d}x_{0:n}) = N^{-1}\sum_{k_{0:n}\in[1:N]^{n+1}} \delta_{y^\frown \underline{\xi}^{k_0}_0}(\mathrm{d}x_0)\prod_{s=1}^n \mathbb{1}\{k_s = A^{k_{s-1}}_s\}\delta_{\xi^{k_s}_s}(\mathrm{d}x_s)\,. \tag{D.2.2}$$

We also let $\Phi^N_{0:n}$ be the probability measure defined for any $B \in \mathcal{B}(\mathbb{R}^{\mathsf{d}_x})^{\otimes n+1}$ by

$$\Phi^N_{0:n}(B) = \mathbb{E}[\phi^N_{0:n}(B)]\,,$$

where the expectation is with respect to the probability measure

$$\begin{aligned} P^N_{0:n}(\mathrm{d}(x^{1:N}_{0:n}, a^{1:N}_{1:n})) = \prod_{i=1}^N p^y_n(\mathrm{d}x^i_n)\prod_{\ell=2}^n\left\{\prod_{j=1}^N\sum_{k=1}^N \omega^k_{\ell-1}\delta_k(\mathrm{d}a^j_\ell)p^y_{\ell-1}(\mathrm{d}x^j_{\ell-1}|x^{a^j_\ell}_\ell)\right\} \\ \times\prod_{j=1}^N\sum_{k=1}^N \omega^k_0\delta_k(\mathrm{d}a^j_1)\underline{p}^y_0(\mathrm{d}\underline{x}^j_0|x^{a^j_1}_1)\delta_y(\mathrm{d}\overline{x}^j_0)\,, \end{aligned} \tag{D.2.3}$$

where $\omega^i_t := \widetilde{\omega}_t(\xi^i_{t+1})/\sum_{j=1}^N \widetilde{\omega}_t(\xi^j_{t+1})$ and which corresponds to the joint law of all the random variables generated by Algorithm 4. It then follows by definition that for any $C \in \mathcal{B}(\mathbb{R}^{\mathsf{d}_x})$,

$$\int \Phi^N_{0:n}(\mathrm{d}z_{0:n})\mathbb{1}_C(z_0) = \mathbb{E}\left[\int \phi^N_{0:n}(\mathrm{d}z_{0:n})\mathbb{1}_C(z_0)\right] = \mathbb{E}[\phi^N_0(C)] = \Phi^N_0(C)\,.$$

Define also the law of the *conditional* particle cloud

$$\begin{aligned} \mathbf{P}^N(\mathrm{d}(x^{1:N}_{0:n}, a^{1:N}_{1:n})|z_{0:n}) &= \delta_{z_n}(\mathrm{d}x^N_n)\prod_{i=1}^{N-1} p^y_n(\mathrm{d}x^i_n) \\ &\times\prod_{\ell=2}^n \delta_{z_{\ell-1}}(\mathrm{d}x^N_{\ell-1})\delta_N(\mathrm{d}a^N_{\ell-1})\prod_{j=1}^{N-1}\sum_{k=1}^N \omega^k_{\ell-1}\delta_k(\mathrm{d}a^j_\ell)p^y_{\ell-1}(\mathrm{d}x^j_{\ell-1}|x^{a^j_\ell}_\ell) \\ &\times\delta_{z_0}(\mathrm{d}x^N_0)\delta_N(\mathrm{d}a^N_1)\prod_{j=1}^{N-1}\sum_{k=1}^N \omega^k_0\delta_k(\mathrm{d}a^j_1)\underline{p}^y_0(\mathrm{d}\underline{x}^j_0|x^{a^j_1}_1)\delta_y(\mathrm{d}\overline{x}^j_0)\,. \end{aligned} \tag{D.2.4}$$

In what follows $\mathbb{E}_{z_{0:n}}$ refers to expectation with respect to $\mathbf{P}^N(\cdot|z_{0:n})$. Finally, for $s \in [0:n-1]$ we let Ω^N_s denote the sum of the filtering weights at step s, i.e. $\Omega^N_s = \sum_{i=1}^N \widetilde{\omega}_s(\xi^i_{s+1})$. We also write $\mathcal{Z}_0 = \int \mathsf{p}_0(x_0)\delta_y(\mathrm{d}\overline{x}_0)\mathrm{d}\underline{x}_0$ and for all $\ell \in [1:n]$, $\mathcal{Z}_\ell = \int \overline{q}_{\ell|0}(\overline{x}_\ell|y)\mathsf{p}_\ell(\mathrm{d}x_\ell)$.

The proof of Proposition 6.2.2 relies on two Lemmata stated below and proved in Section D.2.1; in Lemma D.2.1 we provide an expression for the Radon-Nikodym derivative $\mathrm{d}\phi^y_{0:n}/\mathrm{d}\Phi^y_{0:n}$ and in Lemma D.2.2 we explicit its leading term.

Lemma D.2.1. *$\phi^y_{0:n}$ and $\Phi^N_{0:n}$ are equivalent and we have that*

$$\Phi^N_{0:n}(\mathrm{d}z_{0:n}) = \mathbb{E}_{z_{0:n}}\left[\frac{N^n \mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega^N_s}\right]\phi^y_{0:n}(\mathrm{d}z_{0:n})\,. \tag{D.2.5}$$

Lemma D.2.2. *It holds that*

$$\begin{aligned}\frac{\mathcal{Z}_n}{\mathcal{Z}_0}\mathbb{E}_{z_{0:n}}\left[\prod_{s=0}^{n-1} N^{-1}\Omega^N_s\right] &= \left(\frac{N-1}{N}\right)^n \\ &+ \frac{(N-1)^{n-1}}{N^n}\sum_{s=1}^{n}\frac{\mathcal{Z}_s/\mathcal{Z}_0}{\overline{q}_{s|0}(\overline{z}_s|y)}\int p_{0|s}(x_0|z_s)\delta_y(\mathrm{d}\overline{x}_0)\mathrm{d}\underline{x}_0 + \frac{\mathsf{D}^y_{0:n}}{N^2}\,.\end{aligned} \tag{D.2.6}$$

where $\mathsf{D}^y_{0:n}$ is a positive constant.

Before proceeding with the proof of Proposition 6.2.2, let us note that having $z \mapsto \widetilde{\omega}_\ell(z)$ bounded on $\mathbb{R}^{\mathsf{d}_x}$ for all $\ell \in [0:n-1]$ is sufficient to guarantee that $\mathsf{C}^y_{0:n}$ and $\mathsf{D}^y_{0:n}$ are finite since in this case it follows immediately that $\mathbb{E}_{z_{0:n}}\left[\prod_{s=0}^{n-1} N^{-1}\Omega^N_s\right]$ is bounded and so is the right hand side of (D.2.6). This can be achieved with a slight modification of (6.2.5) and (6.2.6). Indeed, consider instead the following recursion for $s \in [0:n]$ where $\delta > 0$,

$$\begin{aligned}\phi^y_n(x_n) &\propto (\overline{q}_{n|0}(\overline{x}_n|y) + \delta)\mathsf{p}_n(x_n)\,,\\ \phi^y_s(x_s) &\propto \int \phi^y_{s+1}(x_{s+1})p_s(\mathrm{d}x_s|x_{s+1})\frac{\overline{q}_s(\overline{x}_s|y)+\delta}{\overline{q}_{s+1}(\overline{x}_{s+1}|y)+\delta}\mathrm{d}x_{s+1}\,.\end{aligned}$$

Then we have that

$$\phi^y_0(\underline{x}_0) \propto \int \phi^y_1(x_1)\underline{p}_0(\underline{x}_0|x_1)\frac{\overline{p}_0(y|x_1)}{\overline{q}_{1|0}(\overline{x}_1|y)+\delta}\mathrm{d}x_1\,.$$

We can then use Algorithm 4 to produce a particle approximation of ϕ^y_0 using the following transition and weight function,

$$\begin{aligned}p^{y,\delta}_s(x_s|x_{s+1}) &= \frac{\gamma_s(y|x_{s+1})}{\gamma_s(y|x_{s+1})+\delta}p^y_s(x_s|x_{s+1}) + \frac{\delta}{\gamma_s(y|x_{s+1})+\delta}p_s(x_s|x_{s+1})\,,\\ \widetilde{\omega}_s(x_{s+1}) &= (\gamma_s(y|x_{s+1})+\delta)/(\overline{q}_{s+1|0}(\overline{x}_{s+1}|y)+\delta)\,,\end{aligned}$$

where $\gamma_s(y|x_{s+1}) = \int \overline{q}_{s|0}(\overline{x}_s|y)p_s(x_s|x_{s+1})\mathrm{d}x_s$ is available in closed form and p^y_s is defined in (6.2.3). $\widetilde{\omega}_s$ is thus clearly bounded for all $s \in [0:n-1]$ and it is still possible to sample from $p^{y,\delta}_s$ since it is simply a mixture between the transition (6.2.3) and the "prior" transition.

Proof of Proposition 6.2.2. Consider the *forward* Markov kernel

$$\overrightarrow{\mathbf{B}}_{1:n}(z_0, \mathrm{d}z_{1:n}) = \frac{\mathsf{p}_{1:n}(\mathrm{d}z_{1:n})p_0(z_0|z_1)}{\int \mathsf{p}_{1:n}(\mathrm{d}\tilde{z}_{1:n})p_0(\tilde{z}_0|\tilde{z}_1)}\,, \tag{D.2.7}$$

which satisfies

$$\phi^y_{0:n}(\mathrm{d}z_{0:n}) = \phi^y_0(\mathrm{d}z_0)\overrightarrow{\mathbf{B}}_{1:n}(z_0, \mathrm{d}z_{1:n})\,.$$

By Lemma D.2.1 we have for any $C \in \mathcal{B}(\mathbb{R}^{\mathsf{d}_x})$ that

$$\begin{aligned}\Phi_0^N(C) &= \int \Phi_{0:n}^N(\mathrm{d}z_{0:n})\mathbb{1}_C(z_0) \\ &= \int \mathbb{1}_C(z_0)\mathbb{E}_{z_{0:n}}\left[\frac{N^n\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega_s^N}\right]\phi_{0:n}^y(\mathrm{d}z_{0:n}) \\ &= \int \mathbb{1}_C(z_0)\int \overrightarrow{\mathbf{B}}_{1:n}(z_0,\mathrm{d}z_{1:n})\mathbb{E}_{z_{0:n}}\left[\frac{N^n\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega_s^N}\right]\phi_0^y(\mathrm{d}z_0)\,,\end{aligned}$$

which shows that the Radon-Nikodym derivative $\mathrm{d}\Phi_0^N/\mathrm{d}\phi_0^y$ is,

$$\frac{\mathrm{d}\Phi_0^N}{\mathrm{d}\phi_0^y}(z_0) = \int \overrightarrow{\mathbf{B}}_{1:n}(z_0,\mathrm{d}z_{1:n})\mathbb{E}_{z_{0:n}}\left[\frac{N^n\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega_s^N}\right]\,.$$

Applying Jensen's inequality twice yields

$$\frac{\mathrm{d}\Phi_0^N}{\mathrm{d}\phi_0^y}(z_0) \geq \frac{N^n\mathcal{Z}_0/\mathcal{Z}_n}{\int \overrightarrow{\mathbf{B}}_{1:n}(z_0,\mathrm{d}z_{1:n})\mathbb{E}_{z_{0:n}}\left[\prod_{s=0}^{n-1}\Omega_s^N\right]}\,,$$

and it then follows that

$$\mathsf{KL}(\phi_0^y \parallel \Phi_0^N) \leq \int \log\left(\frac{\mathcal{Z}_n}{\mathcal{Z}_0}\int \overrightarrow{\mathbf{B}}_{1:n}(z_0,\mathrm{d}z_{1:n})\mathbb{E}_{z_{0:n}}\left[\prod_{s=0}^{n-1}N^{-1}\Omega_s^N\right]\right)\phi_0^y(\mathrm{d}z_0)\,.$$

Finally, using Lemma D.2.2 and the fact that $\log(1+x) < x$ for $x > 0$ we get

$$\mathsf{KL}(\phi_0^y \parallel \Phi_0^N) \leq \frac{\mathsf{C}_{0:n}^y}{N-1} + \frac{\mathsf{D}_{0:n}^y}{N^2}$$

where

$$\mathsf{C}_{0:n}^y := \sum_{s=1}^{n}\int \frac{\mathcal{Z}_s/\mathcal{Z}_0}{\overline{q}_{s|0}(\overline{z}_s|y)}\left(p_{0|s}(x_0|z_s)\delta_y(\mathrm{d}\overline{x}_0)\mathrm{d}\underline{x}_0\right)\phi_s^y(\mathrm{d}z_s)\,,$$

and $\phi_s^y(z_s) \propto \mathsf{p}_s(z_s)\int p_{0|s}(z_0|z_s)\delta_y(\mathrm{d}\underline{z}_0)\mathrm{d}\overline{z}_0$. □

Proof of Lemma D.2.1 and Lemma D.2.2

Proof of Lemma D.2.1. We have that

$$
\begin{aligned}
&\Phi^N_{0:n}(\mathrm{d}z_{0:n}) \\
&= N^{-1} \int P^N_{0:n}(\mathrm{d}x^{1:N}_{0:n}, \mathrm{d}a^{1:N}_{1:n}) \sum_{k_{0:n} \in [1:N]^{n+1}} \delta_{y^\frown \underline{x}^{k_0}_0}(\mathrm{d}z_0) \prod_{s=1}^n \mathbb{1}\{k_s = a^{k_{s-1}}_s\} \delta_{x^{k_s}_s}(\mathrm{d}z_s) \\
&= N^{-1} \int \sum_{k_{0:n}} \sum_{a^{1:N}_{1:n}} \delta_{y^\frown \underline{x}^{k_0}_0}(\mathrm{d}z_0) \prod_{s=1}^n \mathbb{1}\{k_s = a^{k_{s-1}}_s\} \delta_{x^{k_s}_s}(\mathrm{d}z_s) \\
&\quad \times \prod_{j=1}^N p^y_n(\mathrm{d}x^j_n) \left\{ \prod_{\ell=2}^n \prod_{i=1}^N \omega^{a^i_\ell}_{\ell-1} p^y_{\ell-1}(\mathrm{d}x^i_{\ell-1} | x^{a^i_\ell}_\ell) \right\} \prod_{r=1}^N \omega^{a^r_1}_0 \underline{p}^y_{\ell-1}(\mathrm{d}\underline{x}^r_0 | x^{a^r_1}_1) \delta_y(\overline{x}^r_0) \\
&= N^{-1} \int \sum_{k_{0:n}} \sum_{a^{1:N}_{1:n}} p^y_n(\mathrm{d}x^{k_n}_n) \delta_{x^{k_n}_n}(\mathrm{d}z_n) \prod_{j \neq k_n} p^y_n(\mathrm{d}x^j_n) \prod_{\ell=2}^n \Big\{ \prod_{i \neq k_{\ell-1}} \omega^{a^i_\ell}_{\ell-1} p^y_{\ell-1}(\mathrm{d}x^i_{\ell-1} | x^{a^i_\ell}_\ell) \\
&\quad \times \mathbb{1}\{a^{k_{\ell-1}}_\ell = k_\ell\} \frac{\tilde{\omega}_{\ell-1}(x^{a^{k_{\ell-1}}_\ell}_\ell)}{\Omega^N_{\ell-1}} p^y_{\ell-1}(\mathrm{d}x^{k_{\ell-1}}_\ell | x^{a^{k_{\ell-1}}_\ell}_\ell) \delta_{x^{k_{\ell-1}}_{\ell-1}}(\mathrm{d}z_{\ell-1}) \Big\} \\
&\quad \times \Big\{ \prod_{r \neq k_0} \omega^{a^r_1}_0 \underline{p}^y_0(\mathrm{d}\underline{x}^r_0 | x^{a^r_1}_1) \delta_y(\mathrm{d}\overline{x}^r_0) \Big\} \mathbb{1}\{a^{k_0}_1 = k_1\} \frac{\tilde{\omega}_0(x^{a^{k_0}_1}_1)}{\Omega^N_0} p^y_0(\mathrm{d}\underline{x}^{k_0}_0 | x^{a^{k_0}_1}_0) \delta_{y^\frown \underline{x}^{k_0}_0}(\mathrm{d}z_0) \,.
\end{aligned}
$$

Then, using that for all $s \in [2:n]$

$$
\tilde{\omega}_{s-1}(x^{k_s}_s) p^y_{s-1}(\mathrm{d}x^{k_{s-1}}_{s-1} | x^{k_s}_s) = \frac{\overline{q}_{s-1|0}(\overline{x}^{k_{s-1}}_{s-1} | y)}{\overline{q}_{s|0}(\overline{x}^{k_s}_s | y)} p_s(\mathrm{d}x^{k_{s-1}}_{s-1} | x^{k_s}_s) \,,
$$

we recursively get that

$$
\begin{aligned}
&p^y_n(\mathrm{d}x^{k_n}_n) \delta_{x^{k_n}_n}(\mathrm{d}z_n) \prod_{s=2}^n \mathbb{1}\{a^{k_{s-1}}_s = k_s\} \frac{\tilde{\omega}_{s-1}(x^{a^{k_{s-1}}_s}_s)}{\Omega^N_{s-1}} p^y_{s-1}(\mathrm{d}x^{k_{s-1}}_{s-1} | x^{a^{k_{s-1}}_s}_s) \delta_{x^{k_{s-1}}_{s-1}}(\mathrm{d}z_{s-1}) \\
&\qquad \times \mathbb{1}\{a^{k_0}_1 = k_1\} \frac{\tilde{\omega}_0(x^{a^{k_0}_1}_1)}{\Omega^N_0} p^y_0(\mathrm{d}\underline{x}^{k_0}_0 | x^{a^{k_0}_1}_1) \delta_{y^\frown \underline{x}^{k_0}_0}(\mathrm{d}z_0) \\
&= \frac{\overline{q}_{n|0}(z_n | y) \mathsf{p}_n(\mathrm{d}z_n)}{\mathcal{Z}_n} \delta_{z_n}(\mathrm{d}x^{k_n}_n) \prod_{s=2}^n \mathbb{1}\{a^{k_{s-1}}_s = k_s\} \frac{\overline{q}_{s-1|0}(\overline{z}_{s-1} | y)}{\Omega^N_{s-1} \overline{q}_{s|0}(\overline{z}_s | y)} p_{s-1}(\mathrm{d}z_{s-1} | z_s) \delta_{z_{s-1}}(\mathrm{d}x^{k_{s-1}}_{s-1}) \\
&\qquad \times \mathbb{1}\{a^{k_0}_1 = k_1\} \frac{\overline{p}_0(y | z_1)}{\Omega^N_0 \overline{q}_{1|0}(\overline{z}_1 | y)} p_0(\mathrm{d}\underline{z}_0 | z_1) \delta_y(\mathrm{d}\overline{z}_0) \delta_{z_0}(\mathrm{d}x^{k_0}_0) \\
&= \frac{\mathcal{Z}_0}{\mathcal{Z}_n} \phi^y_{0:n}(\mathrm{d}z_{0:n}) \delta_{z_n}(\mathrm{d}x^{k_n}_n) \prod_{s=1}^n \mathbb{1}\{a^{k_{s-1}}_s = k_s\} \frac{1}{\Omega^N_{s-1}} \delta_{z_{s-1}}(\mathrm{d}x^{k_{s-1}}_{s-1}) \,.
\end{aligned}
$$

Thus, we obtain

$$\begin{aligned}\Phi^N_{0:n}(\mathrm{d}z_{0:n}) &= N^{-1}\int \sum_{k_{0:n}}\sum_{a^{1:N}_{1:n}} \phi^y_{0:n}(\mathrm{d}z_{0:n})\frac{\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega^N_s}\delta_{z_n}(\mathrm{d}x_n^{k_n})\prod_{j\neq k_n} p^y_n(\mathrm{d}x_n^j)\\ &\quad\times\prod_{\ell=2}^{n}\mathbb{1}\{a_\ell^{k_{\ell-1}}=k_\ell\}\delta_{z_{\ell-1}}(\mathrm{d}x_{\ell-1}^{k_{\ell-1}})\prod_{i\neq k_{\ell-1}}\omega^{a^i_\ell}_{\ell-1}p^y_{\ell-1}(\mathrm{d}x^i_{\ell-1}|x^{a^i_\ell}_\ell)\\ &\quad\times\mathbb{1}\{a_1^{k_0}=k_1\}\delta_{z_0}(\mathrm{d}x_0^{k_0})\prod_{i\neq k_0}\omega_0^{a^i_1}\underline{p}_0(\underline{x}^i_0|x_1^{a^i_1})\delta_y(\mathrm{d}\overline{x}^i_0)\\ &= N^{-1}\sum_{k_{0:n}}\phi^y_{0:n}(\mathrm{d}z_{0:n})\mathbb{E}^{k_{0:n}}_{z_{0:n}}\left[\frac{\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega^N_s}\right],\end{aligned}$$

where for all $k_{0:n}\in[1:N]^{n+1}$ $\mathbb{E}^{k_{0:n}}_{z_{0:n}}$ denotes the expectation under the Markov kernel

$$\begin{aligned}\mathbf{P}^N_{k_{0:n}}(\mathrm{d}(x^{1:N}_{0:n},a^{1:N}_{1:n})|z_{0:n}) &= \delta_{z_n}(\mathrm{d}x_n^{k_n})\prod_{i\neq k_n}p^y_n(\mathrm{d}x^i_n)\\ &\quad\times\prod_{\ell=2}^{n}\delta_{z_{\ell-1}}(\mathrm{d}x_{\ell-1}^{k_{\ell-1}})\delta_{k_\ell}(\mathrm{d}a_\ell^{k_{\ell-1}})\prod_{j\neq k_{\ell-1}}\sum_{k=1}^N\omega^k_{\ell-1}\delta_k(\mathrm{d}a^j_\ell)p^y_{\ell-1}(\mathrm{d}x^j_{\ell-1}|x^{a^j_\ell}_\ell)\\ &\quad\times\delta_{z_0}(\mathrm{d}x_0^{k_0})\delta_{k_1}(\mathrm{d}a_1^{k_0})\prod_{j\neq k_0}\sum_{k=1}^N\omega_0^k\delta_k(\mathrm{d}a_1^j)\underline{p}^y_0(\mathrm{d}\underline{x}^j_0|x_1^{a^j_1})\delta_y(\mathrm{d}\overline{x}_0)\,.\end{aligned}$$

Note however that for all $(k_{0:n},\ell_{0:n})\in([1:N]^{n+1})^2$,

$$\mathbb{E}^{k_{0:n}}_{z_{0:n}}\left[\frac{1}{\prod_{s=0}^{n-1}\Omega^N_s}\right]=\mathbb{E}^{\ell_{0:n}}_{z_{0:n}}\left[\frac{1}{\prod_{s=0}^{n-1}\Omega^N_s}\right]$$

and thus it follows that

$$\Phi^N_{0:n}(\mathrm{d}z_{0:n})=\mathbb{E}_{z_{0:n}}\left[\frac{N^n\mathcal{Z}_0/\mathcal{Z}_n}{\prod_{s=0}^{n-1}\Omega^N_s}\right]\phi^y_{0:n}(\mathrm{d}z_{0:n})\,. \tag{D.2.8}$$

□

Denote by $\{\mathcal{F}_s\}_{s=0}^n$ the filtration generated by a conditional particle cloud sampled from the kernel $\mathbf{P}^N$ (D.2.4), i.e. for all $\ell\in[0:n-1]$

$$\mathcal{F}_s=\sigma(\xi^{1:N}_{s:n},A^{1:N}_{s+1:n})\,.$$

and $\mathcal{F}_n=\sigma(\xi^{1:N}_n)$. Define for all bounded f and $\ell\in[0:n-1]$

$$\gamma^N_{\ell:n}(f)=\left\{\prod_{s=\ell+1}^{n-1}N^{-1}\Omega^N_s\right\}N^{-1}\sum_{k=1}^N\widetilde{\omega}_\ell(\xi^k_{\ell+1})f(\xi^k_{\ell+1})\,, \tag{D.2.9}$$

with the convention $\gamma^N_{\ell:n}(f)=1$ if $\ell\geq n$. Define also the transition Kernel

$$Q^y_{\ell-1|\ell+1}:\mathbb{R}^{\mathsf{d}_x}\times\mathcal{B}(\mathbb{R}^{\mathsf{d}_x})\ni(x_{\ell+1},A)\mapsto\int\mathbb{1}_A(x_\ell)\widetilde{\omega}_{\ell-1}(x_\ell)p^y_\ell(\mathrm{d}x_\ell|x_{\ell+1})\,. \tag{D.2.10}$$

Using eqs. (6.2.3) and (6.2.4), it is easily seen that for all $\ell\in[0:n-1]$,

$$\widetilde{\omega}_\ell(x_{\ell+1})Q^y_{\ell-1|\ell+1}(f)(x_{\ell+1})=\frac{1}{\overline{q}_{\ell+1|0}(\overline{x}_{\ell+1}|y)}\int\overline{q}_{\ell|0}(\overline{x}_s|y)\widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)p_\ell(\mathrm{d}x_\ell|x_{\ell+1})\,. \tag{D.2.11}$$

Define $\mathbf{1}:x\in\mathbb{R}^{\mathsf{d}_x}\mapsto 1$. We may thus write that $\gamma^N_{\ell:n}(f)=N^{-1}\gamma^N_{\ell+1:n}(\mathbf{1})\sum_{k=1}^N\widetilde{\omega}_\ell(\xi^k_{\ell+1})f(\xi^k_{\ell+1})$.

Lemma D.2.3. *For all $\ell \in [0:n-1]$ it holds that*

$$\mathbb{E}_{z_{0:n}}[\gamma^N_{\ell-1:n}(f)] = \frac{N-1}{N}\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}\left(Q^y_{\ell-1|\ell+1}(f)\right)\right] + \frac{1}{N}\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}(\mathbf{1})\right]\widetilde{\omega}_{\ell-1}(z_\ell)f(z_\ell)\,.$$

Proof. By the tower property and the fact that $\gamma^N_{\ell:n}(f)$ is $\mathcal{F}_{\ell+1}$-measurable, we have that

$$\mathbb{E}_{z_{0:n}}[\gamma^N_{\ell-1:n}(f)] = \mathbb{E}_{z_{0:n}}\left[N^{-1}\gamma^N_{\ell+1:n}(\mathbf{1})\Omega^N_\ell\mathbb{E}_{z_{0:n}}\left[N^{-1}\sum_{k=1}^N \widetilde{\omega}_{\ell-1}(\xi^k_\ell)f(\xi^k_\ell)\middle|\mathcal{F}_{\ell+1}\right]\right]\,.$$

Note that for all $\ell \in [0:n-1]$, $(\xi^1_\ell,\dots,\xi^{N-1}_\ell)$ are identically distributed conditionally on $\mathcal{F}_{\ell+1}$ and

$$\mathbb{E}_{z_{0:n}}\left[\widetilde{\omega}_{\ell-1}(\xi^j_\ell)f(\xi^j_\ell)\middle|\mathcal{F}_{\ell+1}\right] = \frac{1}{\Omega^N_\ell}\sum_{k=1}^N \widetilde{\omega}_\ell(\xi^k_{\ell+1})\int \widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)p^y_\ell(\mathrm{d}x_\ell|\xi^k_{\ell+1})\,,$$

leading to

$$\mathbb{E}_{z_{0:n}}\left[N^{-1}\sum_{k=1}^N \widetilde{\omega}_{\ell-1}(\xi^k_\ell)f(\xi^k_\ell)\middle|\mathcal{F}_{\ell+1}\right] = \frac{N-1}{N\Omega^N_\ell}\sum_{k=1}^N \widetilde{\omega}_\ell(\xi^k_{\ell+1})\int \widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)p^y_\ell(\mathrm{d}x_\ell|\xi^k_{\ell+1}) + \frac{1}{N}\widetilde{\omega}_{\ell-1}(z_\ell)f(z_\ell)\,,$$

and the desired recursion follows. □

Proof of Lemma D.2.2. We proceed by induction and show for all $\ell \in [0:n-2]$

$$\begin{aligned}
&\mathbb{E}_{z_{0:n}}[\gamma^N_{\ell:n}(f)] \\
&= \left(\frac{N-1}{N}\right)^{n-\ell}\frac{\int \mathsf{p}_{\ell+1}(\mathrm{d}x_{\ell+1})\overline{q}_{\ell+1|0}(\overline{x}_{\ell+1}|y)\widetilde{\omega}_\ell(x_{\ell+1})f(x_{\ell+1})}{\mathcal{Z}_n} \\
&+ \frac{(N-1)^{n-\ell-1}}{N^{n-\ell}}\Bigg[(\mathcal{Z}_{\ell+1}/\mathcal{Z}_n)f(z_{\ell+1})\widetilde{\omega}_\ell(z_{\ell+1}) \\
&+ \sum_{s=\ell+2}^n \frac{\mathcal{Z}_s/\mathcal{Z}_n}{\overline{q}_{s|0}(\overline{z}_s|y)}\int \widetilde{\omega}_\ell(x_{\ell+1})\overline{q}_{\ell+1|0}(\overline{x}_{\ell+1}|y)f(x_{\ell+1})p_{\ell+1|s}(\mathrm{d}x_{\ell+1}|z_s)\Bigg] + \frac{\mathsf{D}^y_{\ell:n}}{N^2}\,.
\end{aligned} \tag{D.2.12}$$

where f is a bounded function and $\mathsf{D}^y_{\ell:n}$ is a a positive constant. The desired result in Lemma D.2.2 then follows by taking $\ell = 0$ and $f = \mathbf{1}$.

Assume that (D.2.12) holds at step ℓ. To show that it holds at step $\ell-1$ we use Lemma D.2.3 and we compute $\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}\left(Q^y_{\ell-1|\ell+1}(f)\right)\right]$ and $\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}(\mathbf{1})\right]\widetilde{\omega}_{\ell-1}(z_\ell)f(z_\ell)$.

Using the following identities which follow from (D.2.11)

$$\int \overline{q}_{\ell+1|0}(\overline{x}_{\ell+1}|y)\widetilde{\omega}_\ell(x_{\ell+1})Q^y_{\ell-1|\ell+1}(f)(x_{\ell+1})\mathsf{p}_{\ell+1}(\mathrm{d}x_{\ell+1}) = \int \overline{q}_{\ell|0}(\overline{x}_\ell|y)\widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)\mathsf{p}_\ell(\mathrm{d}x_\ell)\,,$$

and

$$\int \widetilde{\omega}_\ell(x_{\ell+1})\overline{q}_{\ell+1|0}(\overline{x}_{\ell+1}|y)Q^y_{\ell-1|\ell+1}(f)(x_{\ell+1})p_{\ell+1|s}(\mathrm{d}x_{\ell+1}|x_s) = \int \widetilde{\omega}_{\ell-1}(x_\ell)\overline{q}_{\ell|0}(\overline{x}_\ell|y)f(x_\ell)p_{\ell|s}(\mathrm{d}x_\ell|x_s)\,,$$

we get by (D.2.12) that

$$
\begin{aligned}
&\frac{N-1}{N}\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}\left(Q^y_{\ell-1|\ell+1}(f)\right)\right] \\
&\quad= \left(\frac{N-1}{N}\right)^{n-\ell+1}\frac{\int \overline{q}_{\ell|0}(\overline{x}_\ell|y)\widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)\mathsf{p}_\ell(\mathrm{d}x_\ell)}{\mathcal{Z}_n} \\
&\qquad+ \frac{(N-1)^{n-\ell}}{N^{n-\ell+1}}\left[\frac{\mathcal{Z}_{\ell+1}/\mathcal{Z}_n}{\overline{q}_{\ell+1|0}(\overline{z}_{\ell+1}|y)}\int \overline{q}_{\ell|0}(\overline{x}_\ell|y)\widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)p_\ell(\mathrm{d}x_\ell|z_{\ell+1})\right. \\
&\qquad\left.+ \sum_{s=\ell+2}^{n}\frac{\mathcal{Z}_s/\mathcal{Z}_n}{\overline{q}_{s|0}(\overline{z}_s|y)}\int \widetilde{\omega}_{\ell-1}(x_\ell)\overline{q}_{\ell|0}(\overline{x}_\ell|y)f(x_\ell)p_{\ell|s}(\mathrm{d}x_\ell|z_s)\right] + \frac{\mathsf{D}^y_{\ell:n}}{N^2} \\
&\quad= \left(\frac{N-1}{N}\right)^{n-\ell+1}\frac{\int \overline{q}_{\ell|0}(\overline{x}_\ell|y)\widetilde{\omega}_{\ell-1}(x_\ell)f(x_\ell)\mathsf{p}_\ell(\mathrm{d}x_\ell)}{\mathcal{Z}_n} \\
&\qquad+ \frac{(N-1)^{n-\ell}}{N^{n-\ell+1}}\sum_{s=\ell+1}^{n}\frac{\mathcal{Z}_s/\mathcal{Z}_n}{\overline{q}_{s|0}(\overline{z}_s|y)}\int \widetilde{\omega}_{\ell-1}(x_\ell)\overline{q}_{\ell|0}(\overline{x}_s|y)f(x_\ell)p_{\ell|s}(\mathrm{d}x_\ell|z_s) + \frac{\mathsf{D}^y_{\ell:n}}{N^2}\,.
\end{aligned}
\tag{D.2.13}
$$

The induction step is finished by using again (D.2.12) and noting that

$$
\frac{1}{N}\mathbb{E}_{z_{0:n}}\left[\gamma^N_{\ell:n}(\mathbf{1})\right]\widetilde{\omega}_{\ell-1}(z_\ell)f(z_\ell) = \frac{(N-1)^{n-\ell}}{N^{n-\ell+1}}(\mathcal{Z}_\ell/\mathcal{Z}_n)\widetilde{\omega}_{\ell-1}(z_\ell)f(z_\ell) + \frac{\widetilde{\mathsf{D}}^y_{\ell:n}}{N^2}\,.
$$

and then setting $\mathsf{D}^y_{\ell-1:n} = \mathsf{D}^y_{\ell:n} + \widetilde{\mathsf{D}}^y_{\ell:n}$.
It remains to compute the initial value at $\ell = n-2$. Note that

$$
\mathbb{E}_{z_{0:n}}\left[\gamma^N_{n-1:n}(f)\right] = \frac{N-1}{N}\int p^y_n(\mathrm{d}x_n)\widetilde{\omega}_{n-1}(x_n)f(x_n) + \frac{1}{N}\widetilde{\omega}_{n-1}(z_n)f(z_n) \tag{D.2.14}
$$

and thus by Lemma D.2.3 and similarly to the previous computations

$$
\begin{aligned}
&\mathbb{E}_{z_{0:n}}\left[\gamma^N_{n-2:n}(f)\right] \\
&= \left(\frac{N-1}{N}\right)^2\int p^y_n(\mathrm{d}x_n)\widetilde{\omega}_{n-1}(x_n)Q^y_{n-2|n}(f)(x_n) + \frac{N-1}{N^2}\left[\widetilde{\omega}_{n-1}(z_n)Q^y_{n-2|n}(f)(z_n)\right. \\
&\left.+ \widetilde{\omega}_{n-2}(z_{n-1})f(z_{n-1})\int p^y_n(\mathrm{d}x_n)\widetilde{\omega}_{n-1|n}(x_n)\right] + \frac{\mathsf{D}^y_{n-2fa:n}}{N^2} \\
&= \left(\frac{N-1}{N}\right)^2\frac{\int \overline{q}_{n-1|0}(x_{n-1}|y)\widetilde{\omega}_{n-2}(x_{n-1})\mathsf{p}_{n-1}(\mathrm{d}x_{n-1})}{\mathcal{Z}_n} \\
&+ \frac{N-1}{N^2}\left[(\mathcal{Z}_{n-1}/\mathcal{Z}_n)\widetilde{\omega}_{n-2}(z_{n-1})f(z_{n-1})\right. \\
&\left.+ \frac{1}{\overline{q}_{n|0}(\overline{x}_n|y)}\int \overline{q}_{n-1|0}(\overline{x}_{n-1}|y)\widetilde{\omega}_{n-2}(x_{n-1})f(x_{n-1})p_{n-1}(\mathrm{d}x_{n-1}|z_n)\right] + \frac{\mathsf{D}^y_{n-2:n}}{N^2}\,.
\end{aligned}
$$

□

D.2.2 Proof of Proposition 6.2.3 and Lemma D.2.4

In this section and only in this section we make the following assumption

(**A26**) For all $s \in [0:n-1]$, $\mathfrak{p}_s(x_s)q_{s+1}(x_{s+1}|x_s) = \mathfrak{p}_{s+1}(x_{s+1})\lambda_s(x_s|x_{s+1})$.

We also consider $\sigma_\delta = 0$. In what follows we let $\tau_{\mathsf{d_y}+1} = n$ and we write $\tau_{1:\mathsf{d_y}} = \{\tau_1, \dots, \tau_{\mathsf{d_y}}\}$ and $\overline{\tau_{1:\mathsf{d_y}}} = [1:n] \setminus \tau_{1:t}$. Define the measure

$$\Gamma^{\mathbf{y}}_{0:n}(\mathrm{d}\mathbf{x}_{0:n}) = \mathfrak{p}_n(\mathrm{d}\mathbf{x}_n) \prod_{s \in \overline{\tau_{1:\mathsf{d_y}}}} \lambda_s(\mathrm{d}\mathbf{x}_s|\mathbf{x}_{s+1}) \prod_{i=1}^{\mathsf{d_y}} \lambda_{\tau_i}(\mathbf{x}_{\tau_i}|\mathbf{x}_{\tau_i+1}) \mathrm{d}\mathbf{x}_{\tau_i}^{\setminus i} \delta_{\mathbf{y}[i]}(\mathrm{d}\mathbf{x}_{\tau_i}[i])\,. \tag{D.2.15}$$

Under (**A26**) it has the following alternative *forward* expression,

$$\Gamma^{\mathbf{y}}_{0:n}(\mathrm{d}\mathbf{x}_{0:n}) = \mathfrak{p}_0(\mathrm{d}\mathbf{x}_0) \prod_{s \in \overline{\tau_{1:\mathsf{d_y}}}} q_{s+1}(\mathrm{d}\mathbf{x}_{s+1}|\mathbf{x}_s) \prod_{i=1}^{\mathsf{d_y}} q_{\tau_i}(\mathbf{x}_{\tau_i}|\mathbf{x}_{\tau_i-1}) \mathrm{d}\mathbf{x}_{\tau_i}^{\setminus i} \delta_{\mathbf{y}[i]}(\mathrm{d}\mathbf{x}_{\tau_i}[i])\,. \tag{D.2.16}$$

Since the forward kernels decompose over the dimensions of the states, i.e.

$$q_{s+1}(\mathbf{x}_{s+1}|\mathbf{x}_s) = \prod_{\ell=1}^{\mathsf{d}_x} q^\ell_{s+1}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_s[\ell])$$

where $q^\ell_{s+1}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_s[\ell]) = \mathcal{N}(\mathbf{x}_{s+1}[\ell]; (\alpha_{s+1}/\alpha_s)^{1/2}\mathbf{x}_s[\ell], 1-(\alpha_{s+1}/\alpha_s))$, we can write

$$\Gamma^{\mathbf{y}}_{0:n}(\mathbf{x}_{0:n}) = \mathfrak{p}_0(\mathbf{x}_0) \prod_{\ell=1}^{\mathsf{d}_x} \Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell]|\mathbf{x}_0[\ell])\,, \tag{D.2.17}$$

where for $\ell \in [1:\mathsf{d_y}]$

$$\Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell]|\mathbf{x}_0[\ell]) = q^\ell_{\tau_\ell}(\mathbf{y}[\ell]|\mathbf{x}_{\tau_\ell-1}[\ell]) \prod_{s \neq \tau_\ell} q^\ell_s(\mathrm{d}\mathbf{x}_s[\ell]|\mathbf{x}_{s-1}[\ell])\,, \tag{D.2.18}$$

and for $\ell \in [\mathsf{d_y}+1:\mathsf{d}_x]$,

$$\Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell]|\mathbf{x}_0[\ell]) = \prod_{s=0}^{n-1} q^\ell_{s+1}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_s[\ell])\,. \tag{D.2.19}$$

With these quantities in hand we can now prove Proposition 6.2.3.

Proof of Proposition 6.2.3. Note that for $\ell \in [1:\mathsf{d_y}]$,

$$\begin{aligned}\mathcal{N}(\mathbf{y}[\ell]; \alpha_{\tau_\ell}\mathbf{x}_0[\ell], 1-\alpha_{\tau_\ell}) = q^\ell_{\tau_\ell|0}(\mathbf{y}[\ell]|\mathbf{x}_0[\ell]) &= \int q^\ell_{\tau_\ell}(\mathbf{y}[\ell]|\mathbf{x}_{\tau_\ell-1}[\ell]) \prod_{s\neq\tau_\ell} q^\ell_s(\mathrm{d}\mathbf{x}_s[\ell]|\mathbf{x}_{s-1}[\ell]) \\ &= \int \Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathrm{d}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell])|\mathbf{x}_0[\ell])\end{aligned}$$

and thus by (**??**) we have that

$$\begin{aligned}\mathfrak{p}_0(\mathbf{x}_0) g^y_0(\mathbf{x}_0) &\propto \mathfrak{p}_0(\mathbf{x}_0) \prod_{\ell=1}^{\mathsf{d_y}} \mathcal{N}(\mathbf{y}[\ell]; \alpha_{\tau_\ell}\mathbf{x}_0[\ell], 1-\alpha_{\tau_\ell}) \\ &= \mathfrak{p}_0(\mathbf{x}_0) \prod_{\ell=1}^{\mathsf{d_y}} \int \Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathrm{d}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell])|\mathbf{x}_0[\ell]) \\ &= \mathfrak{p}_0(\mathbf{x}_0) \prod_{\ell=1}^{\mathsf{d}_x} \int \Gamma^{\mathbf{y}}_{1:n|0,\ell}(\mathrm{d}(\mathbf{x}_1[\ell], \dots, \mathbf{x}_n[\ell])|\mathbf{x}_0[\ell])\,.\end{aligned}$$

By (D.2.16) it follows that

$$\phi_0^{\mathbf{y}}(\mathbf{x}_0) = \frac{1}{\int \Gamma_{0:n}^{\mathbf{y}}(\tilde{\mathbf{x}}_{0:n})\mathrm{d}\tilde{\mathbf{x}}_{0:n}} \int \Gamma_{0:n}^{\mathbf{y}}(\mathbf{x}_{0:n})\mathrm{d}\mathbf{x}_{1:n}\,,$$

and hence by (D.2.16) and (D.2.15) we get

$$\phi_0^{\mathbf{y}}(\mathbf{x}_0) \propto \int \mathfrak{p}_{\tau_{\mathsf{d}_\mathbf{y}}}(\mathbf{x}_{\tau_{\mathsf{d}_\mathbf{y}}})\delta_{\mathbf{y}[\mathsf{d}_\mathbf{y}]}(\mathrm{d}\mathbf{x}_{\tau_{\mathsf{d}_\mathbf{y}}}[\mathsf{d}_\mathbf{y}])\mathrm{d}\mathbf{x}_{\tau_{\mathsf{d}_\mathbf{y}}}^{\setminus \mathsf{d}_\mathbf{y}} \left\{ \prod_{i=1}^{\mathsf{d}_\mathbf{y}-1} \lambda_{\tau_i|\tau_{i+1}}(\mathbf{x}_{\tau_i}|\mathbf{x}_{\tau_{i+1}})\delta_{\mathbf{y}[i]}(\mathrm{d}\mathbf{x}_{\tau_i}[i])\mathrm{d}\mathbf{x}_{\tau_i}^{\setminus i} \right\} \lambda_{0|\tau_1}(\mathbf{x}_0|\mathbf{x}_{\tau_1})\,.$$

This completes the proof. □

Let $\gamma_{0,s}^{\mathbf{y}}$ denote the joint time 0 and s marginal of the measure (D.2.15), i.e.

$$\gamma_{0,s}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_s) = \int \Gamma_{0:n}^{\mathbf{y}}(\mathbf{x}_{0:n})\mathrm{d}\mathbf{x}_{1:s-1}\mathrm{d}\mathbf{x}_{s+1:n} \tag{D.2.20}$$

We now prove the following result.

Lemma D.2.4. *Assume (**A26**) and let $\tau_0 := 0$, $\tau_{\mathsf{d}_\mathbf{y}+1} := n$. For all $k \in [1 : \mathsf{d}_\mathbf{y}]$,*

(i) If $s \in [\tau_k + 1 : \tau_{k+1}]$,

$$\gamma_{0,s}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_s) = \int \gamma_{0,s+1}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_{s+1}) q_{\underline{s}|s+1,0}^{\sigma}(\underline{x}_s|\underline{x}_{s+1}, \underline{x}_0) g_s^{\mathbf{y}}(\overline{x}_s) \prod_{\ell=k+1}^{\mathsf{d}_\mathbf{y}} q_{s|s+1,0}^{\sigma,\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell], \mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}\,.$$

(ii) If $s = \tau_k$,

$$\gamma_{0,s}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_s) = \int \gamma_{0,s+1}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_{s+1}) q_{\underline{s}|s+1,0}^{\sigma}(\underline{x}_s|\underline{x}_{s+1}, \underline{x}_0) \times \prod_{i=1}^{k-1} g_{s,i}^{\mathbf{y}}(\overline{x}_s[i]) \prod_{\ell=k+1}^{\mathsf{d}_\mathbf{y}} q_{s|s+1,0}^{\sigma,\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell], \mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}\,.$$

Proof of Lemma D.2.4. Let $k \in [1 : \mathsf{d}_\mathbf{y}]$ and assume that $s \in [\tau_k + 1 : \tau_{k+1} - 2]$. By (**A26**), (D.2.16), (D.2.18) and (D.2.19) we have that

$$\gamma_{0,s}^{\mathbf{y}}(\mathbf{x}_0, \mathbf{x}_s) = \mathfrak{p}_0(\mathbf{x}_0) q_{\underline{s}|0}(\underline{x}_s|\underline{x}_0) \prod_{i=1}^{k} q_{\tau_i|0}^{i}(\mathbf{y}[i]|\mathbf{x}_0[i]) q_{s|\tau_i}^{i}(\mathbf{x}_s[i]|\mathbf{y}[i]) \times \prod_{\ell=k+1}^{\mathsf{d}_\mathbf{y}} q_{s|0}^{\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_0[\ell]) q_{\tau_\ell|s}^{\ell}(\mathbf{y}[\ell]|\mathbf{x}_s[\ell])\,,$$

and thus, using the following identity valid for $\ell \in [k+1 : \mathsf{d}_\mathbf{y}]$

$$\begin{aligned} &q_{s|0}^{\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_0[\ell]) q_{\tau_\ell|s}^{\ell}(\mathbf{y}[\ell]|\mathbf{x}_s[\ell]) \\ &\quad = q_{s|0}^{\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_0[\ell]) \int q_{\tau_\ell|s+1}^{\ell}(\mathbf{y}[\ell]|\mathbf{x}_{s+1}[\ell]) q_{s+1}^{\ell}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_s[\ell])\mathrm{d}\mathbf{x}_{s+1}[\ell] \\ &\quad = \int q_{s|s+1,0}^{\sigma,\ell}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell], \mathbf{x}_0[\ell]) q_{\tau_\ell|s+1}^{\ell}(\mathbf{y}[\ell]|\mathbf{x}_{s+1}[\ell]) q_{s+1|0}^{\ell}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}[\ell]\,, \end{aligned}$$

and that $\underline{q}_{s|0}(\underline{x}_s|\underline{x}_0)\underline{q}_{s+1}(\underline{x}_{s+1}|\underline{x}_s) = \underline{q}^{\sigma}_{s|s+1,0}(\underline{x}_s|\underline{x}_{s+1},\underline{x}_0)\underline{q}_{s+1|0}(\underline{x}_{s+1}|\underline{x}_0)$ we get that

$$\begin{aligned}
&\gamma^{\mathbf{y}}_{0,s}(\mathbf{x}_0,\mathbf{x}_s)\\
&= \int \mathfrak{p}_0(\mathbf{x}_0)\underline{q}_{s|0}(\underline{x}_s|\underline{x}_0)\underline{q}_{s+1}(\mathrm{d}\underline{x}_{s+1}|\underline{x}_s)\\
&\quad\times \prod_{i=1}^{k} q^i_{\tau_i|0}(\mathbf{y}[i]|\mathbf{x}_0[i])q^i_{s|\tau_i}(\mathbf{x}_s[i]|\mathbf{y}[i])q^i_{s+1|\tau_i}(\mathrm{d}\mathbf{x}_{s+1}[i]|\mathbf{y}[i])\\
&\quad\times \prod_{\ell=k+1}^{\mathsf{d}_{\mathbf{y}}} q^{\sigma,\ell}_{s|s+1,0}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell],\mathbf{x}_0[\ell])q^{\ell}_{\tau_\ell|s+1}(\mathbf{y}[\ell]|\mathbf{x}_{s+1}[\ell])q^{\ell}_{s+1|0}(\mathbf{x}_{s+1}[\ell]|\mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}[\ell]\\
&= \int \gamma^{\mathbf{y}}_{0,s+1}(\mathbf{x}_0,\mathbf{x}_{s+1})\underline{q}^{\sigma}_{s|s+1,0}(\underline{x}_s|\underline{x}_{s+1},\underline{x}_0)g^{\mathbf{y}}_s(\overline{x}_s)\prod_{\ell=k+1}^{\mathsf{d}_{\mathbf{y}}} q^{\sigma,\ell}_{s|s+1,0}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell],\mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}\,.
\end{aligned}$$

If $s = \tau_{k+1}$ then

$$\begin{aligned}
\gamma^{\mathbf{y}}_{0,s}(\mathbf{x}_0,\mathbf{x}_s) = \mathfrak{p}_0(\mathbf{x}_0)\underline{q}_{s|0}(\underline{x}_s|\underline{x}_0)\prod_{i=1}^{k} q^i_{\tau_i|0}(\mathbf{y}[i]|\mathbf{x}_0[i])q^i_{s|\tau_i}(\mathbf{x}_s[i]|\mathbf{y}[i])\\
\times q^{k+1}_{\tau_{k+1}|0}(\mathbf{y}[k+1]|\mathbf{x}_0[k+1])\prod_{\ell=k+2}^{\mathsf{d}_{\mathbf{y}}} q^{\ell}_{s|0}(\mathbf{x}_s[\ell]|\mathbf{x}_0[\ell])q^{\ell}_{\tau_\ell|s}(\mathbf{y}[\ell]|\mathbf{x}_s[\ell])\,,
\end{aligned} \tag{D.2.21}$$

and similarly to the previous case we get

$$\begin{aligned}
&\gamma_{0,s}(\mathbf{x}_0,\mathbf{x}_s)\\
&\quad= \int \gamma^{\mathbf{y}}_{0,s+1}(\mathbf{x}_0,\mathbf{x}_{s+1})\underline{q}^{\sigma}_{s|s+1,0}(\underline{x}_s|\underline{x}_{s+1},\underline{x}_0)g^{\mathbf{y}}_s(\overline{x}_s)\prod_{\ell=k+2}^{\mathsf{d}_{\mathbf{y}}} q^{\sigma,\ell}_{s|s+1,0}(\mathbf{x}_s[\ell]|\mathbf{x}_{s+1}[\ell],\mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{s+1}\,.
\end{aligned}$$

Finally, if $s = \tau_{k+1} - 1$, then

$$\begin{aligned}
&\gamma^{\mathbf{y}}_{0,s}(\mathbf{x}_0,\mathbf{x}_s) = \mathfrak{p}_0(\mathbf{x}_0)\underline{q}_{s|0}(\underline{x}_s|\underline{x}_0)\prod_{i=1}^{k} q^i_{\tau_i|0}(\mathbf{y}[i]|\mathbf{x}_0[i])q^i_{s|\tau_i}(\mathbf{x}_s[i]|\mathbf{y}[i])\\
&\times q^{k+1}_{s|0}(\mathbf{x}_s[k+1]|\mathbf{x}_0[k+1])q^{k+1}_{\tau_{k+1}|s}(\mathbf{y}[k+1]|\mathbf{x}_s[k+1])\prod_{\ell=k+2}^{\mathsf{d}_{\mathbf{y}}} q^{\ell}_{s|0}(\mathbf{x}_s[\ell]|\mathbf{x}_0[\ell])q^{\ell}_{\tau_\ell|s}(\mathbf{y}[\ell]|\mathbf{x}_s[\ell])\,,
\end{aligned}$$

and using

$$\begin{aligned}
&q^{k+1}_{s|0}(\mathbf{x}_s[k+1]|\mathbf{x}_0[k+1])q^{k+1}_{\tau_{k+1}|s}(\mathbf{y}[k+1]|\mathbf{x}_s[k+1])\\
&\qquad= q^{\sigma,k+1}_{s|\tau_{k+1},0}(\mathbf{x}_s[k+1]|\mathbf{x}_{\tau_{k+1}}[k+1],\mathbf{x}_0[k+1])q^{k+1}_{\tau_{k+1}|0}(\mathbf{y}[k+1]|\mathbf{x}_0[k+1])
\end{aligned}$$

we find that

$$\begin{aligned}
&\gamma_{0,s}(\mathbf{x}_0,\mathbf{x}_s)\\
&\quad= \int \gamma^{\mathbf{y}}_{0,\tau_{k+1}}(\mathbf{x}_0,\mathbf{x}_{\tau_{k+1}})\underline{q}^{\sigma}_{s|\tau_{k+1},0}(\underline{x}_s|\underline{x}_{\tau_{k+1}},\underline{x}_0)g^{\mathbf{y}}_s(\overline{x}_s)\prod_{\ell=k+1}^{\mathsf{d}_{\mathbf{y}}} q^{\sigma,\ell}_{s|s+1,0}(\mathbf{x}_s[\ell]|\mathbf{x}_{\tau_{k+1}}[\ell],\mathbf{x}_0[\ell])\mathrm{d}\mathbf{x}_{\tau_{k+1}}\,.
\end{aligned}$$

□

D.2.3 Algorithmic details and numerics

The code for both experiments is available at https://anonymous.4open.science/r/mcgdiff/README.md.

D.2.3.1 Transition kernels and weights.

In this section we give explicit formulas for the kernel and weights used in Algorithm 4 and for the noisy case. We first give the formulas for the noisy case since those used in Algorithm 4 are only a special case.

Consider $\ell \in [1 : \mathsf{d}_{\mathbf{y}}]$, $s \in [\tau_\ell : n]$ and $\sigma_\delta > 0$. Define $\sigma^2_{s|\tau_\ell} = 1 - (1 - \sigma^2_\delta)\alpha_s/\alpha_{\tau_\ell}$ and write $\overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}) := \boldsymbol{\mu}_{s+1}(\overline{x}_{s+1}, \overline{\boldsymbol{\chi}}_{0|s+1}(\mathrm{V}\mathbf{x}_{s+1}))$, $\underline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}) := \boldsymbol{\mu}_{s+1}(\underline{x}_{s+1}, \underline{\boldsymbol{\chi}}_{0|s+1}(\mathrm{V}\mathbf{x}_{s+1}))$ and so $\overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}) \frown \underline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1})$ is the mean of the backward kernel p_s.

The density of the ℓ-th coordinate of the proposal kernel is

$$p_s^{\mathbf{y},\ell}(\overline{x}_s[\ell]|\mathbf{x}_{s+1}) \propto \mathcal{N}\left(\alpha_s^{1/2}\mathbf{y}[\ell]; \overline{x}_s[\ell], \sigma^2_{s|\tau_\ell}\right) \mathcal{N}\left(\overline{x}_s[\ell]; \overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1})[\ell], \sigma^2_{s+1}\right),$$

and the weight function is $\widetilde{\omega}_s(\mathbf{x}_{s+1}) = \prod_{\ell=\tau(s)}^{1} \widetilde{\omega}_s^\ell(\mathbf{x}_{s+1})$ (we recall that $\tau(s) = \max\{k \in [1 : \mathsf{d}_{\mathbf{y}}] : s - \tau_k \geq 0\}$) where

$$\widetilde{\omega}_s^\ell(\mathbf{x}_{s+1}) \propto \frac{\int \mathcal{N}\left(\alpha_s^{1/2}\mathbf{y}[\ell]; \overline{x}_s[\ell], \sigma^2_{s|\tau_\ell}\right) \mathcal{N}\left(\overline{x}_s[\ell]; \overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1})[\ell], \sigma^2_{s+1}\right)}{\mathcal{N}\left(\alpha_{s+1}^{1/2}\mathbf{y}[\ell]; \overline{x}_{s+1}[\ell], \sigma^2_{s|\tau_\ell}\right)}.$$

Therefore, letting $\mathsf{k}_{s,\ell} := \sigma^2_{s+1}/(\sigma^2_{s+1} + \sigma^2_{s|\tau_\ell})$ and $\tilde{\sigma}^2_{s,\ell} := \sigma^2_{s|\tau_\ell}\mathsf{k}_{s,\ell}$, we have

$$p_s^{\mathbf{y},\ell}(\overline{x}_s[\ell]|\mathbf{x}_{s+1}) = \mathcal{N}\left(\overline{x}_s[\ell]; \mathsf{k}_{s,\ell}\alpha_s^{1/2}\mathbf{y}[\ell] + (1 - \mathsf{k}_{s,\ell})\overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1})[\ell], \tilde{\sigma}^2_{s,\ell}\right),$$
$$\widetilde{\omega}_s^\ell(\mathbf{x}_{s+1}) = \frac{\mathcal{N}\left(\alpha_s^{1/2}\mathbf{y}[\ell]; \overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1})[\ell], \sigma^2_{s+1} + \sigma^2_{s|\tau_\ell}\right)}{\mathcal{N}\left(\alpha_{s+1}^{1/2}\mathbf{y}[\ell]; \overline{x}_{s+1}[\ell], \sigma^2_{s+1|\tau_\ell}\right)}.$$

The proposal kernel is thus

$$p_s^{\mathbf{y}}(\mathbf{x}_s|\mathbf{x}_{s+1}) = p_s(\underline{x}_s|\mathbf{x}_{s+1}) \prod_{k=\tau(s)+1}^{\mathsf{d}_{\mathbf{y}}} p_s^k(\mathbf{x}_s[k]|\mathbf{x}_{s+1}) \prod_{\ell=1}^{\tau(s)} p_s^{\mathbf{y},\ell}(\overline{x}_s[\ell]|\mathbf{x}_{s+1}).$$

Define $\mathsf{k}_s := \sigma^2_{s+1}/(\sigma^2_{s+1} + 1 - \alpha_s)$ and $\tilde{\sigma}^2_s = (1 - \alpha_s)\mathsf{k}_s$. The kernel and weight function used in Algorithm 4 correspond to the case $\tau_\ell = 0$ for all $\ell \in [1 : \mathsf{d}_{\mathbf{y}}]$, $\sigma_\delta = 0$ are given by

$$p_s^{\mathbf{y}}(\mathbf{x}_s|\mathbf{x}_{s+1}) = \mathcal{N}\left(\underline{x}_s; \underline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}), \sigma^2_{s+1}\right) \mathcal{N}\left(\overline{x}_s; \mathsf{k}_s\alpha_s^{1/2}\mathbf{y} + (1 - \mathsf{k}_s)\overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}), \tilde{\sigma}^2_s\right),$$
$$\widetilde{\omega}_s(\mathbf{x}_{s+1}) = \frac{\mathcal{N}\left(\alpha_s^{1/2}\mathbf{y}; \overline{\boldsymbol{\mu}}_{s|s+1}(\mathbf{x}_{s+1}), \sigma^2_{s+1} + 1 - \alpha_s\right)}{\mathcal{N}\left(\alpha_{s+1}^{1/2}\mathbf{y}; \overline{x}_{s+1}, 1 - \alpha_{s+1}\right)}.$$

D.2.3.2 GMM

For a given dimension d_x, we consider $\mathsf{q}_{\mathrm{data}}$ a mixture of 25 Gaussian random variables. The components have mean $\boldsymbol{\mu}_{i,j} := (8i, 8j, \cdots, 8i, 8j) \in \mathbb{R}^{\mathsf{d}_x}$ for $(i, j) \in \{-2, -1, 0, 1, 2\}^2$ and unit variance. The associated unnormalized weights $\omega_{i,j}$ are independently drawn according to a χ^2 distribution. We have set $\sigma^2_\delta = 10^{-4}$.

Score Note that $\mathsf{q}_s(x_s) = \int q_{s|0}(x_s|x_0)\mathsf{q}_{\text{data}}(x_0)\mathrm{d}x_0$. As q_{data} is a mixture of Gaussians, $\mathsf{q}_s(x_s)$ is also a mixture of Gaussians with means $\alpha_s^{1/2}\boldsymbol{\mu_{i,j}}$ and unitary variances. Therefore, using automatic differentiation libraries, we can calculate $\nabla \log \mathsf{q}_s(x_s)$. Setting $\mathbf{e}(x_s, s) = -(1-\alpha_s)^{1/2}\nabla \log \mathsf{q}_s(x_s)$ leads to the optimum of (6.1.9).

Forward process scaling We chose the sequence of $\{\beta_s\}_{s=1}^{1000}$ as a linearly decreasing sequence between $\beta_1 = 0.2$ and $\beta_{1000} = 10^{-4}$.

Measurement model For a pair of dimensions $(\mathsf{d}_x, \mathsf{d_y})$ the measurement model $(y, \mathrm{A}, \sigma_y)$ is drawn as follows:

- A: We first draw $\tilde{\mathrm{A}} \sim \mathcal{N}(0_{\mathsf{d_y}\times\mathsf{d}_x}, \mathrm{I}_{\mathsf{d_y}\times\mathsf{d}_x})$ and compute the SVD decomposition of $\tilde{A} = \mathrm{USV}^T$. Then, we sample for $(i,j) \in \{-2,-1,0,1,2\}^2$, $s_{i,j}$ according to a uniform in $[0,1]$. Finally, we set $\mathrm{A} = \mathrm{U}\,\mathrm{Diag}(\{s_{i,j}\}_{(i,j)\in\{-2,-1,0,1,2\}^2})\mathrm{V}^T$.
- σ_y: We draw σ_y uniformly in the interval $[0, \max(s_1, \cdots, s_{\mathsf{d_y}})]$.
- y: We then draw $x_* \sim \mathsf{q}_{\text{data}}$ and set $y := \mathrm{A}x_* + \sigma_y\epsilon$ where $\epsilon \sim \mathcal{N}(0_{\mathsf{d_y}}, \mathrm{I}_{\mathsf{d_y}})$.

Posterior Once we have drawn both q_{data} and $(y, \mathrm{A}, \sigma_y)$, the posterior can be exactly calculated using Bayes formula and gives a mixture of Gaussians with mixture components $c_{i,j}$ and associated weights $\tilde{\omega}_{i,j}$

$$
\begin{aligned}
c_{i,j} &:= \mathcal{N}(\Sigma\left(\mathrm{A}^T y/\sigma_y^2 + \boldsymbol{\mu}_{i,j}\right), \Sigma)\,, \\
\tilde{\omega}_i &:= \omega_i \mathcal{N}(y; \mathrm{A}\boldsymbol{\mu}_{i,j}, \sigma^2\,\mathrm{I}_{\mathsf{d}_x} + \mathrm{A}\mathrm{A}^T)\,,
\end{aligned}
$$

where $\Sigma := \left(\mathrm{I}_{\mathsf{d}_x} + \sigma_y^{-2}\mathrm{A}^T\mathrm{A}\right)^{-1}$.

Choosing `DDIM` timesteps for a given measurement model. Given a number of `DDIM` samples R, we choose the timesteps $1 = t_1 < \cdots < t_R = 1000 \in [1:1000]$ as to try to satisfy the two following constraints:

- For all $i \in [1:\mathsf{d_y}]$ there exists a t_j such that $\sigma_y\alpha_{t_j}^{1/2} \approx (1-\alpha_{t_j})^{1/2}s_i$,
- For all $i \in [1:R-1]$, $\alpha_{t_i}^{1/2} - \alpha_{t_{i+1}}^{1/2} \approx \delta$ for some $\delta > 0$.

The first constraint comes naturally from the definition of τ_i. Since the potentials have mean $\alpha_{t_i}^{1/2}y$, the second condition constrains the intermediate laws remain "close". An algorithm that approximately satisfies both constraints is given below.

Additional plots We now proceed to illustrate the first 2 components for one of the measurement models for all the different combinations of `DDIM` steps and $(\mathsf{d}_x, \mathsf{d_y})$ combinations used in table 6.1. Figures D.1 to D.3 are grouped by $\mathsf{d_y} = 1, 2, 4$ respectively.

Algorithm 6: Timesteps choice

Input: Number of `DDIM` steps R, σ_y, $\{s_i\}_{i=1}^{\mathsf{d_y}}$, $\{\alpha_i\}_{i=1}^{1000}$
Output: $\{t_j\}_{j=1}^{R}$

Set $S_\tau = \{\}$.
for $j \leftarrow [1 : \mathsf{d_y}]$ **do**
    Set $\tilde{\tau}_j = \text{argmin}_{\ell \in [1:1000]} |\sigma_y \alpha_\ell^{1/2} - (1 - \alpha_\ell)^{1/2}) s_j|$.
    Add $\tilde{\tau}_j$ to S_τ if $\tilde{\tau}_j \notin S_\tau$.
Set $n_m = R - \#S_\tau - 1$ and $\delta = (\alpha_1^{1/2} - \alpha_{1000}^{1/2})/n_m$.
Set $t_1 = 1$, $e = 1$ and $i_e = 1$. **for** $\ell \leftarrow [2 : 1000]$ **do**
    if $\alpha_e^{1/2} - \alpha_\ell^{1/2} > \delta$ *or* $\ell \in S_\tau$ **then**
        Set $e = \ell$, $i_e = i_e + 1$ and $\tau_{i_e} = \ell$.
Set $\tau_R = 1000$.

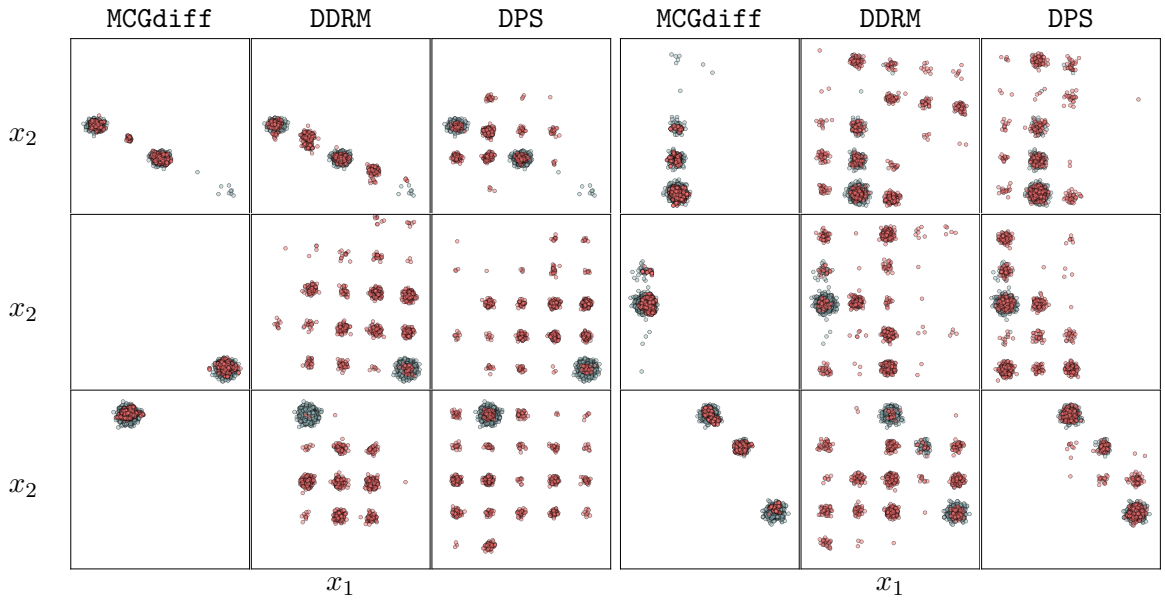

Figure D.1: We display the first two dimensions of the GMM inverse problem for one of the measurement models tested. The blue dots represent samples from the exact posterior, while the red dots correspond to samples generated by each of the algorithms used (the names of the algorithms are given at the top of each column). The first three columns correspond to 20 *`DDIM` steps and the last three to* 100 *`DDIM` steps.* $\mathsf{d_y} = 1$ *and* $\mathsf{d}_x = (8, 80, 800)$ *from top to bottom.*

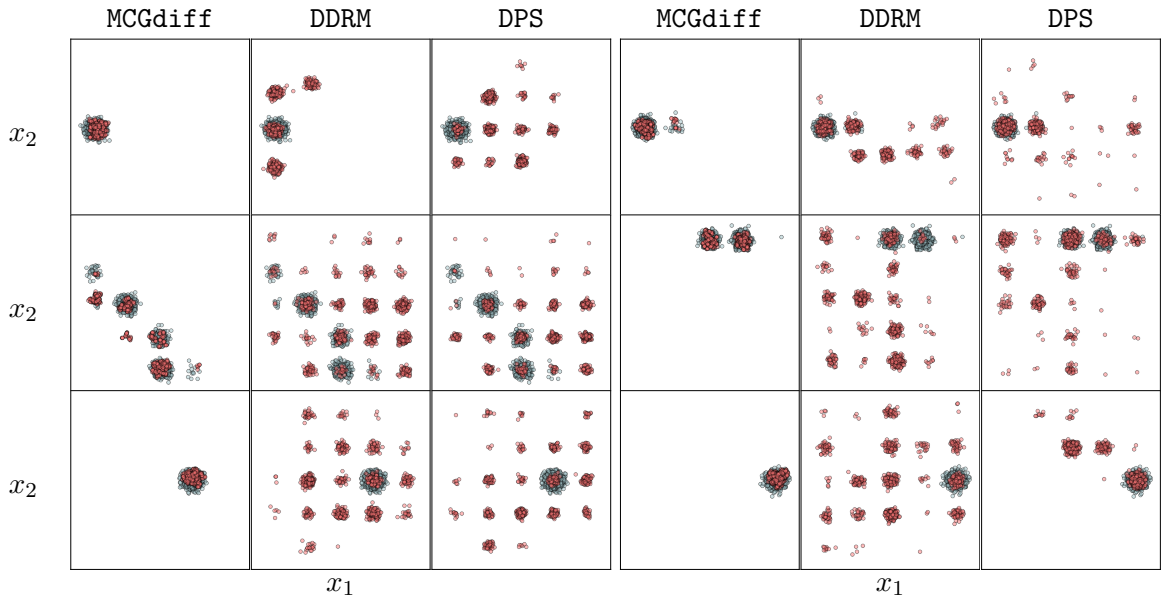

Figure D.2: We display the first two dimensions of the GMM inverse problem for one of the measurement models tested. The blue dots represent samples from the exact posterior, while the red dots correspond to samples generated by each of the algorithms used (the names of the algorithms are given at the top of each column). The first three columns correspond to 20 *DDIM steps and the last three to* 100 *DDIM steps.* $\mathsf{d_y} = 2$ *and* $\mathsf{d}_x = (8, 80, 800)$ *from top to bottom.*

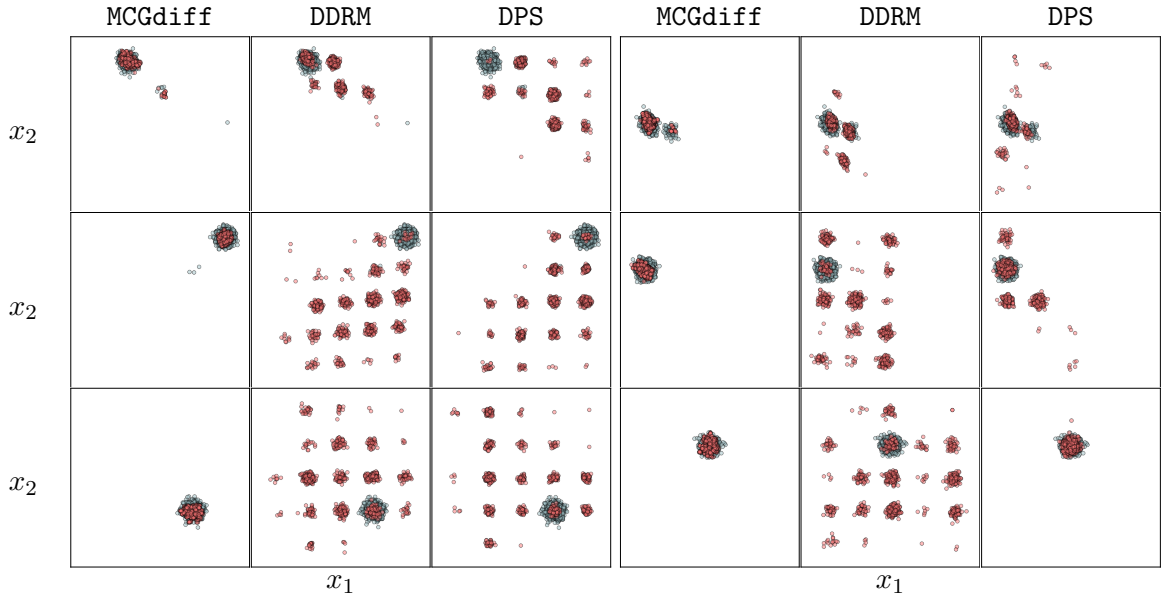

Figure D.3: We display the first two dimensions of the GMM inverse problem for one of the measurement models tested. The blue dots represent samples from the exact posterior, while the red dots correspond to samples generated by each of the algorithms used (the names of the algorithms are given at the top of each column). The first three columns correspond to 20 *DDIM steps and the last three to* 100 *DDIM steps.* $\mathsf{d_y} = 4$ *and* $\mathsf{d}_x = (8, 80, 800)$ *from top to bottom.*

We also show in fig. D.4 the evolution of each observed coordinate in the noise case with $\mathsf{d_y} = 4$. We can see that it follows closely the forward path of the diffused observations indicated by the

blue line.

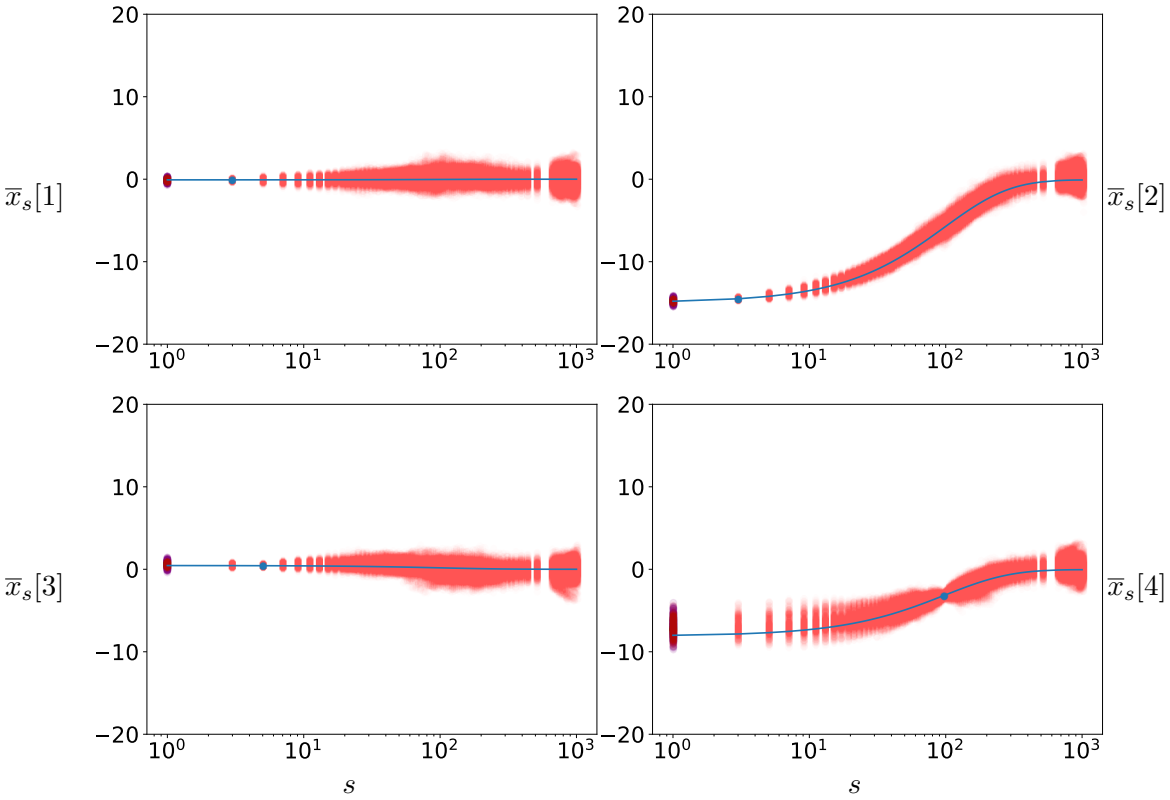

Figure D.4: Illustration of the particle cloud of the 4 first observed coordinate in the case $(\mathsf{d}_{\mathbf{y}}, \mathsf{d}_x) = (4, 800)$ *with* 100 *DDIM steps. The red points represent the particle cloud, while the purple points at the origin represent the posterior distribution. The blue curve corresponds to the curve* $s \to \alpha_s^{1/2}\mathbf{y}[\ell]$ *and the blue dot on the curve to* $\alpha_{\tau_\ell}^{1/2}\mathbf{y}[\ell]$.

D.2.3.3 CelebA

We show in fig. D.5 the evolution of the particle cloud with s.

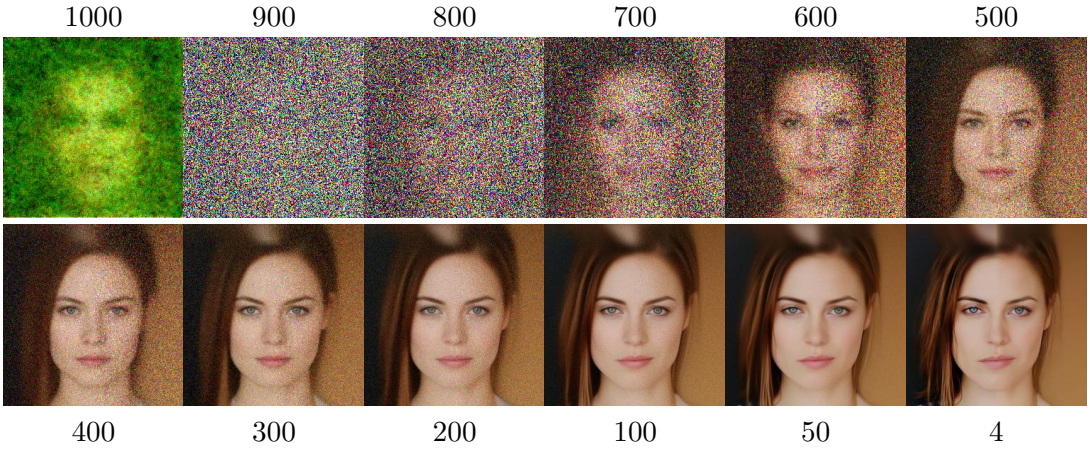

Figure D.5: Evolution of the particle cloud for one of the masks. The numbers on top and bottom indicate the step s of the approximation.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Méthodes Monte Carlo pour l'apprentissage machine: contributions pratiques et théoriques pour l'échantillonnage préferentiel et les méthodes séquentielles

Mots clés : Méthodes Monte Carlo, Echantillonnage préferentiel, Monte Carlo séquentiel, Apprentissage profond

Résumé : Cette thèse contribue au vaste domaine des méthodes de Monte Carlo avec de nouveaux algorithmes visant à traiter l'inférence en grande dimension et la quantification de l'incertitude. Dans une première partie, nous développons deux nouvelles méthodes pour l'échantillonnage d'importance. Le premier algorithme est une nouvelle loi de proposition, basée sur sur des étapes d'optimisation et de coût de calcul faible, pour le calcul des constantes de normalisation. L'algorithme résultant est ensuite étendu en un nouvel algorithme MCMC. Le deuxième algorithme est un nouveau schéma pour l'apprentissage de propositions d'importance adaptées aux cibles complexes et multimodales. Dans une deuxième partie, nous nous concentrons sur les méthodes de Monte Carlo séquentielles. Nous développons de nouveaux estimateurs de la variance asymptotique du filtre à particules et fournissons le premier estimateur de la variance asymptotique d'un lisseur à particules. Ensuite, nous proposons une procédure d'apprentissage des paramètres dans les modèles de Markov cachés en utilisant un lisseur à particules dont le biais est réduit par rapport aux méthodes existantes. Enfin, nous concevons un algorithme de Monte Carlo séquentiel pour résoudre des problèmes inverses linéaires bayésiens avec des lois a priori obtenues par modèles génératifs.

Title : Monte Carlo Methods for Machine Learning: Practical and Theoretical Contributions for Importance Sampling and Sequential Methods

Keywords : Monte Carlo methods, Deep learning, Importance sampling, Sequential Monte Carlo

Abstract : This thesis contributes to the vast domain of Monte Carlo methods with novel algorithms that aim at adressing high dimensional inference and uncertainty quantification. In a first part, we develop two novel methods for Importance Sampling. The first algorithm is a lightweight optimization based proposal for computing normalizing constants and which extends into a novel MCMC algorithm. The second one is a new scheme for learning sharp importance proposals. In a second part, we focus on Sequential Monte Carlo methods. We develop new estimators for the asymptotic variance of the particle filter and provide the first estimator of the asymptotic variance of a particle smoother. Next, we derive a procedure for parameter learning within hidden Markov models using a particle smoother with provably reduced bias. Finally, we devise a Sequential Monte Carlo algorithm for solving Bayesian linear inverse problems with generative model priors.

Institut Polytechnique de Paris
91120 Palaiseau, France