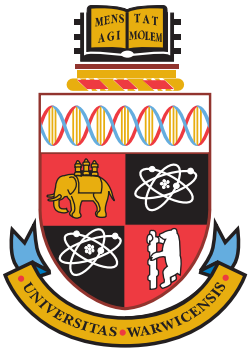


Some Interacting Particle Methods with Non-Standard Interactions

Francesca Romana Crucinio

Thesis submitted for the degree of *Doctor of Philosophy*

University of Warwick
Department of Statistics

May 2021

"What is not surrounded by uncertainty cannot be the truth."

—Perfectly Reasonable Deviations from the Beaten Track
Richard P. Feynman

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

There are many people without whom I could not have written this thesis.

First, I would like to thank my supervisors, Adam M. Johansen and Arnaud Doucet, I am extremely grateful for their help, encouragement and guidance over the past three-and-a-half years. I consider myself extremely lucky to have been given the chance to work with them.

Next, I am extremely grateful to my parents for their support and patience, for being by my side in all ups and downs.

I would also like to thank Juan Kuntz and Valentin De Bortoli for stimulating discussions around the work in this thesis and more. I am thankful to the members of the Department of Statistics at Warwick for making it such a nice place to study, they have all, in some way, contributed to this thesis.

Special thanks go to the bridge club: Jack Carter, Giulio Morina, Marco Palma, William Thomas, and to Suzie Brown, for introducing us to the game that kept us entertained for three years. I would also like to thank Virginia Aglietti, Nayia Constantinou, Pier Guarniero, Ana Ignatieva, Jierong Luo, Francesca Panero, Mattia Tosi, Panayiota Touloupou and Juan Ungredda, I am very grateful to have met all of you.

I would also like to acknowledge the Engineering and Physical Sciences Research Council (grant number EP/L016710/1) for financial support during my PhD.

Francesca R Crucinio, May 2021

Addendum: I am grateful to my examiners Nicolas Chopin and Gareth Peters for taking the time to read this thesis and for your helpful comments and suggestions, and to Richard Everitt, Simon French and Gareth Roberts, for taking the time to read my biannual reports which lead to this thesis.

Francesca R Crucinio, August 2021

Declaration

This thesis is the result of my own work and research, that I carried out under the supervision of Professor Adam M Johansen and Professor Arnaud Doucet, for the degree of Doctor of Philosophy in Statistics.

A condensed version of Chapters 3–5 is contained in the manuscript

> F. R. Crucinio, A. Doucet, and A. M. Johansen. A particle method for solving Fredholm equations of the first kind. *Journal of the American Statistical Association*, In press, 2021b

A manuscript partially based on Chapters 6–7 is currently being finalised in collaboration with Dr Valentin De Bortoli (Crucinio et al., 2021a).

The work in Chapter 8 is an extract from a longer manuscript investigating the theoretical properties of Divide and Conquer SMC currently in preparation with Dr Juan Kuntz Nussio (Kuntz et al., 2021a).

I have not used sources or means without declaration in the text. This thesis has not been submitted for examination to any other institution than the University of Warwick.

Francesca R Crucinio

Abstract

Interacting particle methods are widely used to perform inference in complex models, with applications ranging from Bayesian statistics to applied sciences. This thesis is concerned with the study of families of interacting particles which present non-standard interactions. The non-standard interactions that we study arise from the particular class of problems we are interested in, Fredholm integral equations of the first kind or from algorithmic design, as in the case of the Divide and Conquer sequential Monte Carlo algorithm.

Fredholm integral equations of the first kind are a class of inverse ill-posed problems for which finding numerical solutions remains challenging. These equations are ubiquitous in applied sciences and engineering, with applications in epidemiology, medical imaging, nonlinear regression settings and partial differential equations. We develop two interacting particle methods which provide an adaptive stochastic discretisation and do not require strong assumptions on the solution. While similar to well-studied families of interacting particle methods the two algorithms that we develop present non-standard elements and require a novel theoretical analysis. We study the theoretical properties of the two proposed algorithms, establishing a strong law of large numbers and $\mathbb{L}_p$ error estimates, and compare their performances with alternatives on a suite of examples, including simulated data and realistic systems.

The Divide and Conquer sequential Monte Carlo algorithm is an interacting particle method in which different sequential Monte Carlo approximations are merged together according to the topology of a given tree. We study the effect of the additional interactions due to the merging operations on the theoretical properties of the algorithm. Specifically, we show that the approximation error decays at rate $N^{-1/2}$ in the number of particles, establish a strong law of large numbers and show that the approximations of the normalising constant are unbiased.

1

Introduction

1.1 Context and Motivation

Interacting particle methods refer to Monte Carlo methods which approximate non-linear processes evolving in time. The non-linearity of these processes induces a natural interaction in the evolution of the particle approximation models which interact in a *mean-field* sense, i.e. the particle system acts over one fixed particle through the empirical measure of the system only (Méléard, 1996).

The study of interacting particle systems originated in the field of statistical mechanics with the work of McKean (1966); finding its main application in the study of the dynamics of gases. In the mid-1990s interacting particle methods were introduced in the filtering literature as a powerful tool to perform inference in non-linear settings (Del Moral, 1996; Gordon et al., 1993). Their applicability widened over the subsequent years to include general inference for state space models (Chopin and Papaspiliopoulos, 2020, Chapters 10, 12), approximation of Bayesian posteriors (Chopin, 2002; Del Moral et al., 2006a), rare event simulation (Kroese et al., 2013, Chapter 10), optimisation (Borovykh et al. (2021); Finke (2015, Chapter 15)), inference for differential equations (Wang and Wang, 2020). In recent years families of interacting particles have also been used to describe interactions between neurons (Baladron et al., 2012), social and financial interactions (Carmona, 2016) and game theory (Lasry and Lions, 2007).

The main objective of this work is to study families of interacting particles which present non-standard interactions. The particular interacting particle methods that we study in this work arise from two different perspectives: the need for stable numerical methods to solve Fredholm integral equations of the first kind and the merging of different particle populations which characterises Divide and Conquer

sequential Monte Carlo (Lindsten et al., 2017).

Fredholm integral equations of the first kind are the prototypical example of ill-posed linear inverse problems in the sense that the solution is often non-unique and unstable to small changes is the observed data (Kress, 2014). These integral equations are ubiquitous in applied sciences and engineering; they model, among other things, density deconvolution (Delaigle, 2008; Ma, 2011; Pensky et al., 2017; Yang et al., 2020) and image reconstruction (Aster et al., 2018; Clason et al., 2020; Zhang et al., 2019), find applications in epidemiology (Goldstein et al., 2009; Gostic et al., 2020; Marschner, 2020), statistics (Hall et al., 2005; Miao et al., 2018) and naturally arise as a counterpart to partial differential equations (Colton and Kress, 2012; Tanana et al., 2016).

A common technique to overcome non-uniqueness and instability is to consider regularisation. However, solving the regularised problem remains computationally very challenging. In most cases, an approximate solution is obtained iteratively using a fixed discretisation of its domain (Burger et al., 2019; Chae et al., 2018a; Green, 1990; Yang et al., 2020) or by assuming that the solution can be expressed as a linear combination of basis functions (Islam and Smith, 2020; Jin and Ding, 2016; Kopeć, 1993; Mead, 1986). While these assumptions are common, they are rarely satisfied in practice.

In this work we show that interacting particle methods offer a valuable alternative to standard discretisation schemes and basis expansion for the solution of Fredholm integral equations. On one hand, Monte Carlo methods require less stringent assumptions on the solution of the integral equation, on the other hand the convergence rate of Monte Carlo does not depend on the dimension of the domain of the solution and is therefore suitable to tackle problems in higher dimension than those normally dealt by deterministic discretisation.

The Divide and Conquer sequential Monte Carlo algorithm introduced in Lindsten et al. (2017) is a natural extension of a particular class of particle methods, sequential Monte Carlo methods, in which several particle populations are merged sequentially. While some preliminary work was done in Lindsten et al. (2017), the algorithm's theoretical properties remain underexplored. We extend some of the results of standard sequential Monte Carlo to its Divide and Conquer version.

1.2 Overview and Contributions

This thesis is divided into two parts: in Part I we explore the use of interacting methods to solve Fredholm integral equations of the first kind, while in Part II we

focus on the theoretical characterisation of the the Divide and Conquer sequential Monte Carlo (DaC-SMC) algorithm proposed in Lindsten et al. (2017).

In Part I the non-standard interactions arise from the particular class of problems we are interested in, Fredholm integral equations of the first kind. We propose two interacting particle methods to solve Fredholm integral equations: the first one is a sequential Monte Carlo (SMC) algorithm whose weights are intractable and are approximated using the particle population itself, giving rise to an additional layer of interaction with respect to standard SMC algorithms. In the second case we describe the solution of the Fredholm integral equation as the law of a process satisfying a a particular mean-field stochastic differential equation (SDE), by discretising this SDE we obtain a family of interacting particles.

In Part II the non-standard interactions arise from algorithmic design: Divide and Conquer SMC evolves populations of particles on a tree, with independent populations merging whenever two branches of the tree coalesce. This coalescence step introduces extra-interactions which are not found in standard SMC algorithms.

Chapter 2 is foremost an introductory chapter to two families of interacting particle methods: sequential Monte Carlo (SMC) and interacting particle systems for mean-field stochastic differential equations (SDEs). Both families can be thought of Monte Carlo approximations of a limiting object, Feynman-Kac measure flows and mean-field SDEs respectively. We discuss the numerical implementation and theoretical properties of both families.

Part I In the first part of this thesis we introduce and study the properties of two interacting particle methods for solving Fredholm integral equations and we demonstrate their performances on a number of examples. The main contributions are in Chapters 4–7.

Chapter 3 is a brief introduction to Fredholm integral equations of the first kind and serves as motivation for the following chapters. Fredholm integral equations are ubiquitous in science and engineering and in this chapter we list some of their applications and review standard regularisation techniques for their solution. Particular emphasis is given to maximum likelihood methods and Expectation Maximisation (EM) algorithms.

In Chapter 4 we introduce a continuous version of the Expectation Maximisation Smoothing (EMS) of Silverman et al. (1990) and discuss its properties and limitations. In particular, we show that the continuous EMS map admits a fixed point in the space of probability measures under mild assumptions and extend the results on uniqueness of the fixed point for the discretised EMS to smoothing matrices with

positive entries.

Chapter 5 introduces a sequential Monte Carlo (SMC) algorithm to approximate the continuous EMS recursion. Our contributions are three-fold.

- We show that the EMS recursion can be interpreted as a Feynman-Kac measure flow and use this connection to build a novel particle version of EMS which does not suffer from the limitations of the original scheme;
- Since standard SMC convergence results do not apply, we provide an original theoretical analysis of the algorithm and show that this Monte Carlo scheme enjoys the usual $1/\sqrt{N}$ rate of convergence;
- We demonstrate the algorithm on toy models as well as realistic image reconstruction problems.

Chapter 6 introduces a penalised maximum likelihood estimator for the solution of the integral equation and shows how to use a gradient flow construction to obtain a partial differential equation (PDE) whose solution maximises the penalised likelihood. In particular, we establish that the gradient flow PDE admits a (unique) solution under the conditions we consider.

In Chapter 7 we connect the gradient flow PDE with its corresponding stochastic differential equation (SDE), discuss the numerical implementation and present results on both toy models and real data.

Part II The second part of this thesis is concerned with the study of Divide and Conquer SMC (Lindsten et al., 2017). In Chapter 8 we describe the algorithm and extend some of the standard results for SMC to this algorithm. In particular, we show that DaC-SMC estimators of both the unnormalised and normalised targets are consistent and satisfy $\mathbb{L}_p$ inequalities. In the case of the unnormalised target, we also show that the estimators are unbiased, while in that of the normalised one, we show that the estimators' bias decays linearly with the number of particles. The study of DaC-SMC, and in particular of the merging operations necessary to combine two particle populations into one, lead us to study a class of Monte Carlo estimators for multi-dimensional targets that we named product-form estimators (Kuntz et al., 2021b). While product for estimators heavily feature in DaC-SMC their use is not limited to this setting, and we anticipate that they could be embedded within more complicated Monte Carlo routines to tackle the aspects of the problem exhibiting product structure.

1.3 Notation

We collect in this section the notation which will be used throughout this work. For easy reference, we also provide a list of frequently used symbols on page 192 along with a list of abbreviations on page 193.

Sets and Vectors Let us denote by $\mathbb{N}$ the set of positive integers, and by $\mathbb{R}$ the set of real numbers. We endow any subset $\mathbb{H}$ of $\mathbb{R}^d$ with the Borel σ-algebra $B(\mathbb{H})$ with respect to the Euclidean norm, and we endow any product space with the product σ-algebra. For a vector $x \in \mathbb{H} \subseteq \mathbb{R}^d$ we denote by

$$\|x\|_1 := \sum_{i=1}^{d} |x_i| \qquad \text{and} \qquad \|x\|_2 := \left(\sum_{i=1}^{d} |x_i|^2 \right)^{1/2}$$

the l_1 and l_2 norm, respectively. We denote by $x^i_{1:n} := (x^i_1, \ldots, x^i_n)$.

Functions Let the Banach space of real-valued bounded measurable functions on $\mathbb{H}$, endowed with the supremum norm, $\|\varphi\|_\infty = \sup_{u \in \mathbb{H}} |\varphi(u)|$, be denoted by $\mathcal{B}_b(\mathbb{H})$. We denote by $C_b(\mathbb{H}) \subset \mathcal{B}_b(\mathbb{H})$ the subset of bounded continuous functions and by $\mathrm{BL}(\mathbb{H}) \subset C_b(\mathbb{H})$ the subset of bounded Lipschitz continuous functions endowed with the bounded-Lipschitz norm

$$\|\varphi\|_{BL} := \|\varphi\|_\infty + \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{\|x - y\|_2}.$$

For a set $A \in B(\mathbb{H})$ we denote by $\mathbb{I}(A)$ the indicator function taking value 1 on A and 0 otherwise.

Measures Let $\mathcal{M}(\mathbb{H})$ be the Banach space of signed finite measures on $(\mathbb{H}, B(\mathbb{H}))$ endowed with the bounded Lipschitz norm (e.g. Dudley (2002, page 394))

$$\beta(\eta) := \sup_{\|\varphi\|_{BL} \leq 1} \left| \int_{\mathbb{H}} \eta(\mathrm{d}x) \varphi(x) \right|. \tag{1.1}$$

For ease of notation, for every measure $\nu \in \mathcal{M}(\mathbb{H})$ and every $\varphi \in \mathcal{B}_b(\mathbb{H})$ we denote the integral of φ with respect to ν by $\nu(\varphi) := \int_{\mathbb{H}} \nu(\mathrm{d}u) \varphi(u)$.

We denote by $\mathcal{M}^+(\mathbb{H}) \subset \mathcal{M}(\mathbb{H})$ the set of (unsigned) measures of nonzero mass and by $\mathcal{P}(\mathbb{H}) \subset \mathcal{M}^+(\mathbb{H})$ the set of all probability measures on $(\mathbb{H}, B(\mathbb{H}))$. For every

$\eta \in \mathcal{P}(\mathbb{H})$ we have

$$\beta(\eta) = \sup_{\|\varphi\|_{BL}\leq 1} \left| \int_{\mathbb{H}} \eta(\mathrm{d}x)\varphi(x) \right| \leq \sup_{\|\varphi\|_{BL}\leq 1} \|\varphi\|_{\infty}\eta(\mathbb{H}) \leq 1.$$

The β norm induces a metric on $\mathcal{M}(\mathbb{H})$

$$\beta(\mu,\eta) \equiv \beta(\mu-\eta) = \sup_{\|\varphi\|_{BL}\leq 1} \left| \int_{\mathbb{H}} \mu(\mathrm{d}x)\varphi(x) - \int_{\mathbb{H}} \eta(\mathrm{d}x)\varphi(x) \right|$$

for $\mu, \eta \in \mathcal{M}(\mathbb{H})$. This metric metrises weak convergence (Dudley, 2002, Theorem 11.3.3) in $\mathcal{M}(\mathbb{X})$: for every $\mu \in \mathcal{M}(\mathbb{H})$ and sequence $\{\mu_n\}_{n\geq 1}$ taking values in $\mathcal{M}(\mathbb{H})$, $\beta(\mu_n, \mu) \to 0$ is equivalent to $\mu_n(\varphi) \to \mu(\varphi)$ for all continuous bounded functions $\varphi \in C_b(\mathbb{H})$.

Let us denote the set of probability measures with finite second moment on $\mathbb{H}$ by

$$\mathcal{P}_2(\mathbb{H}) = \left\{ \mu \in \mathcal{P}(\mathbb{H}) : \int \mu(\mathrm{d}x)\|x\|_2^2 < \infty \right\}$$

and we define the 2-Wasserstein distance on this set

$$W_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \pi\left(\mathrm{d}(x,y)\right) \right)^{1/2}$$

where $\Pi(\mu,\nu)$ is the set of all possible couplings between μ and ν. We denote by $\mathcal{P}_2^{ac}(\mathbb{H}) \subset \mathcal{P}_2(\mathbb{H})$ the subset of these measures which is absolutely continuous w.r.t. the appropriate Lebesgue measure.

For any $\nu \in \mathcal{M}^+(\mathbb{H})$ and any positive function G which is integrable with respect to ν we denote by $\Psi_G(\nu)(\mathrm{d}x)$ the Boltzmann-Gibbs transform

$$\Psi_G(\nu)(\mathrm{d}x) = \frac{1}{\nu(G)} G(x)\nu(\mathrm{d}x).$$

For each $\omega \in \Omega$, we obtain a realisation of the particle system with N particles at time n and a corresponding random measure denoted by $\eta_n^N : \omega \in \Omega \mapsto \eta_n^N(\omega) \in \mathcal{P}(\mathbb{H})$

$$\eta_n^N(\omega)(\cdot) = \frac{1}{N}\sum_{i=1}^{N} \delta_{(X_n^i(\omega), Y_n^i(\omega))}(\cdot),$$

where we suppress from the notation the dependence of $X_n^i(\omega)$ and $Y_n^i(\omega)$ upon N,

as we shall throughout in the interest of readability.

Operators For every operator O acting from $\mathcal{M}(\mathbb{X})$ into itself we define the operator norm

$$\| \mathrm{O} \|_{op} = \sup \left\{ \frac{\beta(\mathrm{O}\,\nu)}{\beta(\nu)} : \nu \in \mathcal{M}(\mathbb{X}),\ \beta(\nu) \neq 0 \right\}.$$

For bounded linear operators, we have the equivalent definition

$$\| \mathrm{O} \|_{op} = \sup \left\{ \beta(\mathrm{O}\,\nu) : \nu \in \mathcal{M}(\mathbb{X}),\ \beta(\nu) = 1 \right\}.$$

A Markov kernel M from $\mathbb{H}$ to $\mathbb{H}$ induces two operators. One acts upon measures in $\mathcal{M}(\mathbb{H})$ and takes values in $\mathcal{M}(\mathbb{H})$ and is defined by

$$\forall \nu \in \mathcal{M}(\mathbb{H}) \quad \nu M(\cdot) = \int_{\mathbb{H}} \nu(\mathrm{d}u) M(u, \cdot)$$

and the other acts upon functions in $\mathcal{B}_b(\mathbb{H})$ and takes values in $\mathcal{B}_b(\mathbb{H})$ and may be defined as

$$\forall u \in \mathbb{H} \quad \forall \varphi \in \mathcal{B}_b(\mathbb{H}) \quad M(\varphi)(u) = \int_{\mathbb{H}} M(u, \mathrm{d}v) \varphi(v).$$

2

Interacting Particle Methods

2.1 Introduction

Interacting particle methods are a family of Monte Carlo methods which approximate a probability distribution through an interacting population of (weighted) samples evolving over time. We focus here on particle systems interacting in a *mean-field* sense, that is systems in which each sample interacts with the rest of the population through the empirical measure of the system (Méléard, 1996).

We introduce two classes of interacting particle methods, sequential Monte Carlo (SMC) and interacting particle systems for mean-field stochastic differential equations (SDEs). Broadly speaking, both classes are Monte Carlo approximations of limiting processes, sequential Monte Carlo algorithms can be seen as mean field approximations of an appropriate Feynman-Kac measure flow while mean-field SDEs are the population limit of a family of interacting particles.

Sequential Monte Carlo (SMC) methods allow to approximate distributions and integrals with respect to those distributions through a population of weighted samples (or particles). After their introduction in the context of filtering for state-space models in the seminal works of Gordon et al. (1993) and Del Moral (1996), SMC methods have been growing in popularity and have found application in a wide range of fields, including posterior approximation (e.g. Chopin (2002)), smoothing for state-space models (e.g. Del Moral et al. (2009), Briers et al. (2010)), parameter estimation (e.g. Kantas et al. (2015)), Bayesian model comparison (e.g. Zhou et al. (2016)), maximum likelihood estimation (e.g. Finke (2015, Chapter 6) and references therein). A book-length introductory treatment is given in Chopin and Papaspiliopoulos (2020).

Mean-field SDEs are stochastic differential equations whose coefficients depend

not only on the process itself but also on its distribution. The evolution of the process solving the SDE is approximated by considering several copies of the SDE which interact through their drift and diffusion coefficients where the empirical distribution of the copies approximates the distribution of the solution. Originally introduced to study the dynamics of gases (McKean, 1966; Méléard, 1996), mean-field SDEs also find wide applications in the theory of mean-field games (Lasry and Lions, 2007) describing large scale social and financial interactions (Carmona, 2016; Carmona et al., 2016), in neuroscience, to describe the interactions between neurons (Baladron et al., 2012) and in the analysis of the limiting behaviour of neural networks (e.g. De Bortoli et al. (2020) and references therein).

In the first part of this chapter, we briefly review the fundamental ideas of Feynman-Kac measure flows and their approximation via sequential Monte Carlo. The connection with Feynman-Kac measure flows is particularly convenient, as it allows to study the properties of a wide range of SMC algorithms (Del Moral, 2004, 2013). We present selected results on the convergence of these Monte Carlo algorithms, and a number of algorithms which fall into the broader class of sequential Monte Carlo methods, with particular focus on works related to the topic of this thesis. In the second part of this chapter we introduce mean-field SDEs, their numerical implementation and corresponding theoretical properties with particular focus on the results which will be relevant in the development of this work.

2.2 Feynman-Kac Measure Flows

Feynman-Kac measure flows are a class of mathematical models used to describe the evolution of a sequence of measures through time. These models can be defined both in continuous and in discrete time, but, for the sake of brevity, we will only discuss the discrete time formulation as this is the only one needed in the development of this thesis. A comprehensive review of the discrete case can be found in Del Moral (2004, 2013), while the continuous case is studied in e.g. Del Moral and Miclo (2000).

Consider a collection of measurable spaces $(\mathbb{H}^n, \mathcal{H}_n)$, $n \in \mathbb{N}$. For simplicity, we assume that $\mathbb{H}^n$ is endowed with a separable topology, so that we can select $\mathcal{H}_n$ to be the Borel σ-algebras $B(\mathbb{H}^n)$, $n \in \mathbb{N}$. Let us denote the set of signed finite measures on $(\mathbb{H}^n, B(\mathbb{H}^n))$ by $\mathcal{M}(\mathbb{H}^n)$ and the set of all probability measures on $(\mathbb{H}^n, B(\mathbb{H}^n))$ by $\mathcal{P}(\mathbb{H}^n) \subset \mathcal{M}(\mathbb{H}^n)$ for all $n \in \mathbb{N}$. Let the Banach space of real-valued bounded measurable functions on $\mathbb{H}$, endowed with the supremum norm, $\|\varphi\|_\infty = \sup_{u \in \mathbb{H}} |\varphi(u)|$, be denoted by $\mathcal{B}_b(\mathbb{H})$. For ease of notation, for every measure

$\eta \in \mathcal{M}(\mathbb{H})$ and every $\varphi \in \mathcal{B}_b(\mathbb{H})$ we denote the integral of φ with respect to η by $\eta(\varphi) := \int_{\mathbb{H}} \eta(\mathrm{d}u)\varphi(u)$.

A sequence of probability measures $\{\hat{\eta}_n(x_{1:n})\}_{n\geq 1}$ of increasing dimension defined on the measurable spaces $(\mathbb{H}^n, B(\mathbb{H}^n))_{n\geq 1}$ follows a Feynman-Kac evolution flow if

$$\hat{\eta}_n(\mathrm{d}x_{1:n}) \propto \hat{\eta}_{n-1}(\mathrm{d}x_{1:n-1})M_n(x_{n-1}, \mathrm{d}x_n)G_n(x_n) \tag{2.1}$$

where M_n are Markov mutation kernels

$$M_n : \mathbb{H} \times B(\mathbb{H}) \to [0, 1], \qquad M_n : (x_{1:n-1}, A) \mapsto M_n(x_{1:n-1}, A)$$

and $G_n(x_n)$ are non-negative potential functions which are integrable with respect to $\hat{\eta}_n$ for all $n \geq 1$ (Del Moral, 2004, 2013).

Recursion (2.1) can be decomposed into two steps. In the mutation step, a new state is proposed according to M_n, the resulting distribution is the predictive distribution. In the selection step, the proposed state is weighted according to the potential function G_n, the weighted distribution is the updated distribution:

$$\begin{aligned} &\text{Prediction:} && \eta_n(\mathrm{d}x_{1:n}) \propto \hat{\eta}_{n-1}(\mathrm{d}x_{1:n-1})M_n(x_{1:n-1}, dx_n) \\ &\text{Update:} && \hat{\eta}_n(\mathrm{d}x_{1:n}) \propto \eta_n(\mathrm{d}x_{1:n})G_n(x_n). \end{aligned}$$

It follows that the predictive distributions $\{\eta_n(\mathrm{d}x_{1:n})\}_{n\geq 1}$ satisfy the recursion

$$\eta_{n+1}(\mathrm{d}x_{1:n+1}) \propto \eta_n(\mathrm{d}x_{1:n})G_n(x_n)M_{n+1}(x_{1:n}, \mathrm{d}x_{n+1}). \tag{2.2}$$

It is easy to show that (2.1) and (2.2) describe the same evolution model, and that analysing (2.1) is equivalent to analysing (2.2) (Del Moral, 2004, Proposition 2.4.1).

Of complementary interest is the unnormalised flow of the updated distribution, $\hat{\gamma}_n \in \mathcal{M}^+(\mathbb{H}^n)$

$$\hat{\gamma}_n(\mathrm{d}x_{1:n}) = \hat{\gamma}_{n-1}(\mathrm{d}x_{1:n-1})M_n(x_{1:n-1}, \mathrm{d}x_n)G_n(x_n) \tag{2.3}$$

with $\hat{\eta}_n(\mathrm{d}x_{1:n}) = \hat{\gamma}_n(\mathrm{d}x_{1:n})/\hat{\mathcal{Z}}_n$ and $\hat{\mathcal{Z}}_n := \hat{\gamma}_n(\mathbb{H}^n)$ denoting the normalising constant of $\hat{\eta}_n(\mathrm{d}x_{1:n})$. In particular, the integral of $\varphi(x_{1:n}) \equiv 1$ with respect to the unnormalised flow provides the normalising constant $\hat{\mathcal{Z}}_n = \hat{\gamma}_n(1)$.

Remark 2.1. Our notation slightly differs from that of Del Moral (2004, 2013) as we denote the joint path-space distribution, i.e. the distribution of $x_{1:n}$, by η_n and its unnormalised version by γ_n. On one hand, this choice reduces the amount of

notation introduced, on the other hand, in the remainder of this work we are mostly interested in the marginal distribution at time n that we denote, with a slight abuse of notation, by $\eta_n(\mathrm{d}x_n)$.

A wide class of models satisfy (2.1). We present here two classical examples, one from statistical physics and the well-known filtering problem for state-space models.

Example 2.1 (Particle absorption models)**.** Consider a particle evolving on the measurable space $(\mathbb{H}^n, B(\mathbb{H}^n))$ to which a cemetery state $\{c\}$ is added, so that when the particle reaches the cemetery state it is killed. The evolution of the particle is described by a Markov chain $\{X_n\}_{n\geq 1}$ with transition kernel M_n and a killing probability $1 - G_n(x)$ with $G_n(x) \in [0, 1]$ for all $x \in \mathbb{H}$. Assume the particle is at location X_n at time n, then with probability $1 - G_n(X_n)$ it is killed, otherwise the particles moves to $X_{n+1} \sim M_{n+1}(X_n, \cdot)$. If we denote by T the first time at which the particle enters the cemetery state, then the law of $(X_{1:n})$ conditioned on $T \geq n$ is described by the Feynman-Kac flow (2.1) while the normalising constant $\hat{\mathcal{Z}}_n$ corresponds to $\mathbb{P}(T \geq n)$, the probability of the particle surviving up to time n (Del Moral, 2013, Section 7.1.1).

Example 2.2 (Filtering for State Space Models)**.** Consider a discrete-time Markov process $\{X_n\}_{n\geq 1}$ such that

$$X_1 \sim \eta_1(x_1) \qquad \text{and} \qquad X_n \mid (X_{n-1} = x_{n-1}) \sim f(x_{n-1}, x_n),$$

where $f(x', x)$ is the probability density associated with moving from x' to x, and a sequence of conditionally independent observations $\{Y_n\}_{n\geq 1}$

$$Y_n \mid (X_n = x_n) \sim g(y_n \mid x_n),$$

where g is the likelihood of observation y_n given the latent state x_n (e.g. Doucet and Johansen (2011)). The distributions of interest are the filtering distributions, i.e. the laws of the hidden state, X_n at time n, given observations $y_{1:n}$, $\hat{\eta}_n(\mathrm{d}x_{1:n}) = p(\mathrm{d}x_{1:n} \mid y_{1:n})$ for each $n \geq 1$. These distributions satisfy the Feynman-Kac flow in (2.1) with $M_n \equiv f$ and $G_n \equiv g$. The unnormalised flow (2.3) describes the evolution of the marginal likelihood $\hat{\mathcal{Z}}_n = p(y_{1:n})$.

2.3 Sequential Monte Carlo Methods

Sequential Monte Carlo (SMC) methods approximate $\hat{\eta}_n$ and η_n for $n \geq 1$ through a particle population. The population consists of a set of N weighted parti-

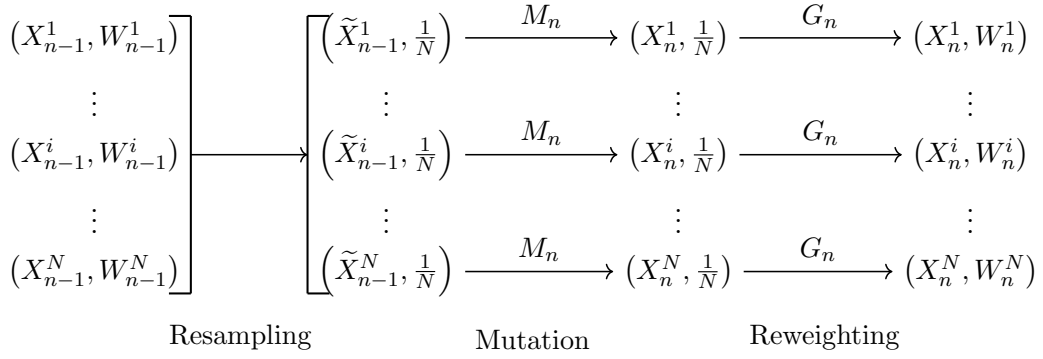


Figure 1: Evolution of the particle population in SMC.

cles $\{X^i_n, W^i_n\}_{i=1}^N$ evolved in time according to the dynamic in (2.2). The evolution has two fundamental steps: mutation according to M_n and selection via the potential function G_n and a resampling mechanism. In the mutation step, the equally weighted population at time $n-1$, $\{\widetilde{X}^i_{n-1}, \frac{1}{N}\}_{i=1}^N$, evolves into the population at time n, $\{X^i_n, \frac{1}{N}\}_{i=1}^N$ where $M_n(\widetilde{X}^i_{n-1}, \mathrm{d}x_n)$ is a Markov kernel. In the selection step, the fitness of the new particles is measured through G_n, which gives the weights

$$W^i_n = \frac{G_n(X^i_n)}{\sum_{j=1}^N G_n(X^j_n)}. \tag{2.4}$$

The new particles are then replicated or discarded according to the Boltzmann-Gibbs transform associated with the potential G_n

$$\Psi_{G_n}(\eta_n)(\mathrm{d}x_{1:n}) := \frac{1}{\eta_n(G_n)} G_n(x_n)\eta_n(\mathrm{d}x_{1:n}).$$

The equally weighted population at time n, $\{\widetilde{X}^i_n, \frac{1}{N}\}_{i=1}^N$, is obtained by resampling from $\{X^i_n, W^i_n\}_{i=1}^N$. Resampling consists of selecting with replacement N particles from $\{X^i_n\}_{i=1}^N$ to get N equally weighted new positions $\{\widetilde{X}^i_n\}_{i=1}^N$. There are a number of options for the selection mechanism. We review some of them in Section 2.3.1.

Figure 1 gives a schematic representation of the evolution of the particle population from time $n-1$ to time n. Iterating this simple mutation-reweighting-resampling scheme gives Algorithm 1, often called sequential importance resampling (SIR). The SIR strategy is the basis of every SMC algorithm.

At each time step n, the particle population provides estimators of both the predictive distribution η_n and the updated distribution $\hat{\eta}_n$. The predictive distribution η_n is approximated through the distribution of the particle population after the

Algorithm 1: Sequential Importance Resampling (SIR)

At time $n = 1$

1 Given η_1: Sample $X_1^i \sim \eta_1$ for $i = 1, \ldots, N$

2 Compute the the normalised weights (2.4) for $i = 1, \ldots, N$

At time $n > 1$

3 (Re)Sample $\left\{X_{1:n-1}^i, W_n^i\right\}$ to get $\left\{\widetilde{X}_{1:n-1}^i, \frac{1}{N}\right\}$ for $i = 1, \ldots, N$

4 Sample $X_n^i \sim M_n(\widetilde{X}_{n-1}^i, \cdot)$ and set $X_{1:n}^i = (\widetilde{X}_{1:n-1}^i, X_n^i)$ for $i = 1, \ldots, N$

5 Compute the the normalised weights (2.4) for $i = 1, \ldots, N$

mutation step

$$\eta_n^N(\mathrm{d}x_{1:n}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_{1:n}^i}(\mathrm{d}x_{1:n}),$$

while the updated distribution $\hat{\eta}_n$ is approximated by both the distribution after the reweighting step

$$\Psi_{G_n}(\eta_n^N)(\mathrm{d}x_{1:n}) = \frac{\sum_{i=1}^{N} G_n(X_n^i)\delta_{X_{1:n}^i}(\mathrm{d}x_{1:n})}{\sum_{j=1}^{N} G_n(X_n^j)}, \tag{2.5}$$

and the distribution after the resampling step

$$\hat{\eta}_n^N(\mathrm{d}x_{1:n}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widetilde{X}_{1:n}^i}(\mathrm{d}x_{1:n}). \tag{2.6}$$

The approximation (2.5) is usually preferred to (2.6), as the variance of (2.5) is smaller than that of (2.6) as a consequence of the Rao-Blackwell Theorem (Blackwell, 1947; Rao, 1992). Algorithm 1 also provides estimates of the normalising constant $\hat{\mathcal{Z}}_n$

$$\hat{\mathcal{Z}}_n^N = \prod_{p=1}^{n} \frac{1}{N}\sum_{i=1}^{N} G_p(X_p^i)$$

and of the corresponding unnormalised flow $\hat{\gamma}_n^N(\mathrm{d}x_{1:n}) = \hat{\mathcal{Z}}_n^N \hat{\eta}_n^N(\mathrm{d}x_{1:n})$.

2.3.1 Resampling

Resampling is a key idea in SMC, since the introduction of a resampling step, despite increasing the variance at time n, leads to more stable approximations in the long run and avoids path degeneracy (Douc et al., 2005). Broadly speaking, a resampling scheme is a selection mechanism which given a set of weighted samples

$\{X^i, W^i\}_{i=1}^N$ outputs a sequence of equally weighted samples $\{\widetilde{X}^i, 1/N\}_{i=1}^N$ in which for all $i = 1, \ldots, N$ $\widetilde{X}^i = X^j$ for some j.

Several resampling mechanisms have been considered in the literature. We consider the most widely used class of schemes, those in which the selection mechanism is random and the expected number of copies of X^i is proportional to W_i for each $i = 1, \ldots, N$ (Douc et al., 2005). In particular, we focus on multinomial resampling (Gordon et al., 1993) and adaptive multinomial resampling (Liu, 2008, page 35). This choice is motivated by the following observations. First, many theoretical results (e.g. the central limit theorem of Chopin (2004)) are obtained for multinomial resampling, as lower variance resampling strategies (e.g. stratified/systematic resampling) can substantially complicate the theoretical analysis. Secondly, lower variance resampling schemes could easily replace the multinomial scheme and would be expected to improve performances. Douc et al. (2005) provides a comparative survey of multinomial, residual, stratified and systematic resampling and Gerber et al. (2019) analyses the consistency properties of resampling schemes.

In multinomial resampling (Gordon et al., 1993) the number of copies of particle X_n^i is a multinomial random variable with N trials and probabilities given by W_n^i for all particles $i = 1, \ldots, N$.

In adaptive multinomial resampling the resampling step is performed only if a given condition is satisfied. Following Liu (2008), we consider the effective sample size (ESS)

$$\text{ESS} = \left(\sum_{j=1}^{N} W_n^j \right)^2 \left(\sum_{i=1}^{N} (W_n^i)^2 \right)^{-1}$$

and we perform a multinomial resampling step only if ESS $< N/2$. Recently, Del Moral et al. (2012) showed that this approach shares many favourable convergence properties with its non-adaptive counterpart.

2.3.2 Convergence Properties of SMC

The particle approximations provided by SMC possess various convergence properties. A complete survey of the convergence results available in the literature is beyond the scope of this work, for the convenience of the reader we present here selected results and refer to e.g. Del Moral (2004, 2013) for a deeper discussion.

Our aim in presenting these results is twofold. On one hand, they provide formal justification for the use of SMC methods, on the other hand the SMC algorithms that we will study in the following are non-standard, and we will show how to extend

some of these results to the particular non-standard cases at hand.

In the literature, two approaches have been widely used to establish convergence results. As SMC methods are interpreted as mean field approximations of Feynman-Kac measure flows (Del Moral, 2004, 2013), it is natural to use this connection to analyse the particle approximations provided by SMC. Secondly, one could directly analyse Algorithm 1 as in, e.g., Crisan and Doucet (2000), Crisan and Doucet (2002), Míguez et al. (2013), Chopin (2004), Andrieu et al. (2010).

We only present results on the final time marginals

$$\hat{\eta}_n^N(\mathrm{d}x_n) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widetilde{X}_n^i}(\mathrm{d}x_n),$$

of the updated distribution (2.6) because the final time marginals are the quantities we will focus on the remainder of this work. In addition, when the potential functions G_n only depend on the terminal point x_n and not on $x_{1:n-1}$ (as in the SIR algorithm described above) the path measure $\hat{\eta}_n^N(\mathrm{d}x_{1:n})$ and its marginal $\hat{\eta}_n^N(\mathrm{d}x_n)$ have the same structure (Del Moral, 2004, Section 2.4.1) and the results below (with the exception of Proposition 2.2) hold for $\hat{\eta}_n^N(\mathrm{d}x_{1:n})$ too. Equivalent results hold for the predictive distribution η_n and its particle approximation (5) (Del Moral, 2004, Section 2.4.1).

We begin by presenting those results concerning expectations with respect to $\hat{\eta}_n^N$, $\hat{\eta}_n^N(\varphi) := N^{-1}\sum_{i=1}^N \varphi(\widetilde{X}_n^i)$, and its unnormalised version $\hat{\gamma}_n^N$, $\hat{\gamma}_n^N(\varphi) := \hat{\mathcal{Z}}_n^N \hat{\eta}_n^N(\varphi)$, for measurable bounded test functions $\varphi \in \mathcal{B}_b(\mathbb{H})$. These results are presented under fairly strong assumptions:

Assumption 2.1. The potentials G_n are bounded for all $n \geq 1$ and positive everywhere, $G_n(x_n) > 0$ for every $x_n \in \mathbb{H}$.

This assumption is common in the SMC literature (Del Moral, 2004, 2013), in particular $G_n(x_n) > 0$ for every $x_n \in \mathbb{H}$ ensures that the system does not become extinct (i.e. the weights have never all simultaneously taken the value zero), and can be relaxed introducing stopping times (Del Moral, 2004) or considering local boundedness (Whiteley, 2013).

The first result we present is the $\mathbb{L}_p$-inequality,

Proposition 2.1 ($\mathbb{L}_p$-inequality)**.** Under Assumption 2.1, for every time $n \geq 1$ and every $p \geq 1$ there exists a finite constant $C_{p,n}$ such that for every measurable

bounded function $\varphi \in \mathcal{B}_b(\mathbb{H})$

$$\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq C_{p,n} \frac{\|\varphi\|_\infty}{\sqrt{N}},$$

where the expectations are taken with respect to the law of all random variables generated within the SMC algorithm.

This result was established for $p = 2$ using the direct approach in Crisan and Doucet (2000, 2002) and generalised to any p in Míguez et al. (2013) (under slightly different assumptions). The mean field approximation approach is detailed in Theorem 7.4.3 of Del Moral (2004).

This statement can be strengthened to a time uniform bound (Del Moral, 2004, Theorem 7.4.4) under an additional mixing assumption on the Markov kernels M_n detailed in Del Moral (2004, page 116):

Proposition 2.2 (Time uniform bounds)**.** Under Assumption 2.1 and additional mixing assumptions on the Markov kernels M_n, for every $p \geq 1$

$$\sup_{n \geq 1} \sup_{\varphi \in \mathcal{B}_b(\mathbb{H})} \mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq C_p \frac{\|\varphi\|_\infty}{\sqrt{N}},$$

where the expectations are taken with respect to the law of all random variables generated within the SMC algorithm and the finite constant C_p depends on the mixing conditions on M_n and on the boundedness conditions on G_n.

Closely related to the $\mathbb{L}_p$ inequality is the strong law of large numbers (SLLN)

Proposition 2.3 (Strong law of large numbers)**.** Under Assumption 2.1, for all $n \geq 1$ and for every $\varphi \in \mathcal{B}_b(\mathbb{H})$, we have $\hat{\eta}_n^N(\varphi) \overset{\text{a.s.}}{\to} \hat{\eta}_n(\varphi)$.

The SLLN can be obtained using a direct inductive argument as in Crisan and Doucet (2000, 2002), the mean field approximation approach as in Del Moral (2004, Corollary 7.4.2) or from the $\mathbb{L}_p$ inequality using Markov's inequality within a Borel-Cantelli argument as shown in e.g. Boustati et al. (2020, Appendix D).

We then focus on the bias of the approximations. The approximations of the unnormalised flow γ_n are unbiased:

Proposition 2.4 (Unbiasedness of unnormalised flow)**.** Under Assumption 2.1, for any $N \geq 1$, $n \geq 1$ and any $\varphi \in \mathcal{B}_b(\mathbb{H})$ we have

$$\mathbb{E}\left[\hat{\gamma}_n^N(\varphi)\right] = \hat{\gamma}_n(\varphi).$$

In particular, the normalising constant estimate is unbiased: $\mathbb{E}\left[\hat{\mathcal{Z}}_n^N\right] = \hat{\mathcal{Z}}_n$.

The unbiasedness result is given in Del Moral (2004, Theorem 7.4.2), while Andrieu et al. (2010) obtain the same result by directly analysing Algorithm 1. Under the additional assumption that the potential functions are bounded below for all n, i.e. $G_n(\cdot) \geq \beta > 0$ it is possible to obtain the following bias estimates (e.g. Del Moral (2013, Proposition 9.5.6) and Olsson and Rydén (2004))

Proposition 2.5 (Bias estimate)**.** Under Assumption 2.1, if the potential functions are bounded below $G_n(\cdot) \geq \beta > 0$, $n \geq 1$ and any $\varphi \in \mathcal{B}_b(\mathbb{H})$ we have

$$\left|\mathbb{E}\left[\hat{\eta}_n^N(\varphi)\right] - \hat{\eta}_n(\varphi)\right| \leq \frac{C_n \|\varphi\|_\infty}{N} \tag{2.7}$$

for some finite C_n. The expectations are taken with respect to the law of all random variables generated within the SMC algorithm.

Similar bias estimates can be obtained without the additional assumption on the lower bound of the potentials by considering stopping times as in Del Moral (2004, Theorem 7.4.3) with the addition on the right hand-side of (2.7) of a term depending on the properties of M_n which decays exponentially fast with N.

The last result on convergence of integrals that we present is the central limit theorem

Proposition 2.6 (CLT)**.** Under Assumption 2.1, for every time $n \geq 1$ and every continuous bounded function $\varphi \in C_b(\mathbb{H})$

$$\sqrt{N}\left[\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)\right] \xrightarrow{d} \mathcal{N}\left(0, \sigma_n^2(\varphi)\right)$$

for some finite positive $\sigma_n^2(\varphi)$.

The CLT is proven in Del Moral (2004, Chapter 9) using the mean field approximation approach and in Chopin (2004) using the direct approach (for a wider class of test functions). In the first case the asymptotic variance $\sigma_n^2(\varphi)$ is given as sum of integral expressions, while in the second case the $\sigma_n^2(\varphi)$ is obtained recursively; the two expression can be shown to be equal (see Chopin (2004) and Johansen and Doucet (2007, Proposition A.1.1))

Among the results which do not involve expectations of test functions φ, we present the almost sure convergence in the weak topology of measures:

Proposition 2.7 (Almost sure convergence in the weak topology)**.** Under Assumption 2.1, $\hat{\eta}_n^N$ converges almost surely in the weak topology to $\hat{\eta}_n$, $\hat{\eta}_n^N \rightharpoonup \hat{\eta}_n$ as $N \to \infty$ for all $n \geq 1$.

This result is a corollary of the SLLN, as shown in Crisan and Doucet (2002) and more recently in Schmon et al. (2021).

2.3.3 Some Examples of SMC Algorithms

This section provides some popular examples of SMC algorithms, making the connection with Feynman-Kac flows explicit.

Consider the filtering problem for state space models described in Example 2.2 and the corresponding Feynman-Kac flow (2.1) with $M_n \equiv f$ and $G_n \equiv g$. Then, the SMC algorithm approximating this flow is the bootstrap particle filter of Gordon et al. (1993).

The filtering problem has an intrinsic sequential structure. When this is not the case, it is often possible to give a sequential structure to the problem at hand. An excellent example is posterior estimation. In this context, one could sequentially introduce the available data $y_{1:n}$ in the likelihood evaluation (Chopin, 2002) or gradually increase the influence of the likelihood over the prior (using for example the geometric construction of Neal (2001)). SMC algorithms for these scenarios have been considered in e.g. Chopin (2002); Del Moral et al. (2006a).

Often, in order to introduce a sequential structure in the problem, new parameters influencing M_n and G_n have to be introduced. To improve algorithmic efficiency, it is natural to choose the values of these parameters adaptively. This approach is named adaptive SMC and has been recently shown to satisfy some of the convergence properties of its non-adaptive counterpart, specifically, a weak law of large numbers and a central limit theorem (Beskos et al., 2016).

Recently, SMC methods have also been implied to find eigenvalues and eigenfunctions of a given integral kernel (Whiteley and Kantas, 2017). This algorithm is somewhat similar in spirit to the algorithms we will introduce in this thesis, as both aim at finding a fixed point of a particular expression.

SMC algorithms for Fredholm Integral Equations

The use of SMC methods to solve Fredholm integral equations is not new. In particular, sequential importance sampling (SIS), a version of the SIR algorithm without resampling, has been employed to solve Fredholm integral equations of the second kind:

$$f(x_1) = \int_{\mathbb{X}} g(x_1, x_2) f(x_2) \mathrm{d}x_2 + h(x_1), \tag{2.8}$$

where $h : \mathbb{X} \to \mathbb{R}$ and $g : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ are known functions on some space $\mathbb{X}$ and $f : \mathbb{X} \to \mathbb{R}$ is unknown (Halton, 1994; Spanier and Gelbard, 2008, Chapter 2).

When the operator $f \mapsto \int_{\mathbb{X}} g(\cdot, x_2) f(x_2) \mathrm{d}x_2$ is a contraction (i.e. if $\max_{y \in \mathbb{Y}} \int_{\mathbb{X}} |g(y \mid x) \mathrm{d}x| < 1$), the solution of (2.8) admits the following Von-Neumann series representation (see Kress (2014, Section 2.4) and Doucet et al. (2010) for details):

$$f(x_1) = h(x_1) + \sum_{n=2}^{\infty} \int_{\mathbb{X}^n} \prod_{k=2}^{n} g(x_{k-1}, x_k) h(x_n) \mathrm{d}x_{1:n} \tag{2.9}$$

and then use SIS to approximate the right-hand-side of (2.9).

We detail here the approach adopted in Spanier and Gelbard (2008, Chapter 2); see also Doucet et al. (2010). The Feynman-Kac flow (2.1) is obtained by selecting an initial probability distribution μ such that $\mu(\mathrm{d}x) > 0$ on $\mathbb{X}$ and a Markov kernel M on $\mathbb{X} \times \mathbb{X}$ such that $M(x_1, x_2) > 0$ if $g(x_1, x_2) \neq 0$ and with an absorbing state x^a such that $M(x, x^a) = p$ for all $x \in \mathbb{X}$. At the first time step the mutation kernel M_n is given by μ and the potential function G_n is $h(x_1)/(p\mu(x_1))$ so that the product of the two (conditional on no-absorption) gives the first term in (2.9).

Similarly, at following time steps, $M_n \equiv M$ and the potential functions are

$$G_n(x_n) = h(x_n) \frac{g(x_{n-1}, x_n)}{M(x_{n-1}, x_n)}.$$

Then, the normalising constant of $\hat{\eta}_n$ in (2.1) corresponds to the integral in (2.9) at given n.

2.4 Mean-Field Stochastic Differential Equations

Mean-field stochastic differential equations (SDEs), also known as McKean-Vlasov SDEs, are stochastic differential equations whose coefficients depend on the law ρ_t of the process itself

$$\mathrm{d}X_t = b(X_t, \rho_t)\mathrm{d}t + s(X_t, \rho_t)\mathrm{d}W_t, \tag{2.10}$$

with X_t a d-dimensional process, W_t a d-dimensional Brownian motion, $X_0 \sim \rho_0$ square integrable, $b : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ and $b : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^{d \times d}$ where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of probability measures on $\mathbb{R}^d$.

These equations provide a probabilistic representation of non-linear partial differential equations (PDEs) describing the limiting behaviour of a particle evolving within a large system of particles interacting in a *mean-field* sense and where first

studied by McKean (1966).

Example 2.3 (Burgers' Equation)**.** Burgers' partial differential equation describes the speed $u(t,x)$ of a fluid with viscosity σ^2

$$\begin{cases} \frac{\partial u}{\partial t}(t,x) = \frac{1}{2}\sigma^2 \frac{\partial^2 u}{\partial x^2}(t,x) - u(t,x)\frac{\partial u}{\partial x}(t,x), & \text{in } [0,T]\times\mathbb{R}, \\ u(0,x) = u_0(x) \end{cases} \tag{2.11}$$

As a consequence of Itô's Lemma (Itô, 1951), the process X_t with cumulative distribution function $u(t,x)$ satisfies

$$\mathrm{d}X_t = \left[\int_{\mathbb{R}} \rho_t(\mathrm{d}y) H(X_t - y)\right]\mathrm{d}t + \sigma \mathrm{d}W_t,$$

where H is the Heaviside function and ρ_t is the law of X_t.

2.4.1 Numerical Implementation

Since mean-field SDEs have coefficients which depend on the law ρ_t of the process, the first step towards numerically approximating (2.10) is the introduction of a space discretisation. This can be achieved by considering N copies of (2.10) and replacing the non linearity of the coefficients by interaction (Bossy and Talay, 1997). Given N copies $(X_t^{1,N}, ..., X_t^{N,N})$, at $t = 0$ sample i.i.d. $X_0^{i,N} \sim \rho_0$ and then evolve each particle according to the non-linear SDE

$$\mathrm{d}X_t^{i,N} = b(X_t^{i,N}, \rho_t^N)\mathrm{d}t + s(X_t^{i,N}, \rho_t^N)\mathrm{d}W_t^i, \tag{2.12}$$

where W_t^i for $i = 1, \ldots, N$ are N independent d-dimensional standard Brownian motions and ρ_t^N is the empirical measure given by the N particles

$$\rho_t^N(\mathrm{d}x) := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_t^{i,N}}(\mathrm{d}x).$$

The limit behaviour of (2.12) as $N \to \infty$ depends on the continuity properties of the coefficients b, s, in particular we distinguish between the case in which b, s are Lipschitz continuous which is widely studied and the general case for which fewer results exist (see Section 2.4.2 below). Generally, we are interested in the *propagation of chaos* result:

Proposition 2.8 (Propagation of Chaos)**.** Under continuity assumptions on b, s which will be made explicit in Section 2.4.2, for any $T \geq 0$ there exist $C_1 < \infty$ and

$\gamma > 0$ such that for any $N \in \mathbb{N}$

$$\sup_{1\leq i\leq N} \mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N} - X_t^i\right|^2\right] \leq C_1 N^{-\gamma}. \tag{2.13}$$

Intuitively, the results above tells us that if at time $t = 0$ the N particles are i.i.d. from ρ_0, then the law of N fixed particles tends to the distribution of N independent particles with same law ρ_t when the size of the system goes to infinity (Méléard, 1996). From a practical point of view, (2.13) justifies the use of Monte Carlo methods to approximate the process X_t solving (2.10).

Numerical implementation of (2.12) requires a time discretisation scheme. We focus here on Euler type discretisations; schemes with higher order of convergence exist (Bao et al., 2020; Kumar and Neelima, 2021) but require Lions derivatives on measure spaces (Lasry and Lions, 2007) and are therefore not considered here. The explicit Euler scheme for (2.12) with discretisation step Δt is

$$Y_{k+1}^{i,N} = Y_k^{i,N} + b(Y_k^{i,N}, \rho_k^N)\ \Delta t + s(Y_k^{i,N}, \rho_k^N)\Delta W_k^i, \tag{2.14}$$

where ΔW_k^i are independent centred Gaussian random variables with variance Δt and ρ_k^N is the empirical measure corresponding to the time-discretised N-particle system

$$\rho_k^N(\mathrm{d}y) := \frac{1}{N}\sum_{i=1}^N \delta_{Y_k^{i,N}}(\mathrm{d}y). \tag{2.15}$$

As for standard SDEs the time continuous interpolation of the Euler scheme with time step Δt

$$\mathrm{d}Y_t^{i,N} = b\left(Y_{t_k}^{i,N}, \rho_{t_k}^N\right)\mathrm{d}t + s\left(Y_{t_k}^{i,N}, \rho_{t_k}^N\right)\mathrm{d}W_t^i, \qquad \rho_{t_k}^N(\mathrm{d}y) := \frac{1}{N}\sum_{i=1}^N \delta_{Y_{t_k}^{i,N}}(\mathrm{d}y),$$

where W_t^i for $i = 1, \ldots, N$ are N independent d-dimensional standard Brownian motions and $t_k := \max\{k : k \leq t\}$, is then compared to the N-particle system (2.12) to check its convergence properties:

Proposition 2.9 (Time Discretisation)**.** Under continuity assumptions on b, s which will be made explicit in Section 2.4.2, for any $T \geq 0$ there exist $\delta > 0$ and a constant

$C_2 > 0$ independent on N and δ such that

$$\sup_{1\leq i\leq N} \mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N} - Y_t^{i,N}\right|^2\right] \leq C_2(\Delta t)^\delta.$$

Combining the two results above we obtain the rate of convergence of (2.14) with respect to the time discretisation Δt and the number of particles N:

Proposition 2.10. Under continuity assumptions on b, s which will be made explicit in Section 2.4.2, for any $T \geq 0$ there exist $\delta, \gamma > 0$ and a constant $C > 0$ independent on N and δ such that

$$\sup_{1\leq i\leq N} \mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i} - Y_t^{i,N}\right|^2\right] \leq C\left(N^{-\gamma} + (\Delta t)^\delta\right).$$

While results for general b, s require some continuity assumptions, it is possible to obtain estimates like that in Proposition 2.10 also for models with discontinuous b, s as the Burgers' equation in Example 2.3 (Bossy and Talay, 1997):

Example 2.4 (Burgers' equation)**.** consider the model for fluid diffusion given by the Burgers' equation in Example 2.3. Burgers' PDE can be solved numerically through (2.14), Bossy and Talay (1997) show that the empirical distribution function of the particles (2.14) satisfies an estimate like that in Proposition 2.10 with $\gamma = \delta = 1/2$.

While in Example 2.4 we are interested in letting $N \to \infty, \Delta t \to 0$ to obtain a approximate solution for (2.11), in the theory of mean-field games the behaviour of (2.12) for finite N is often more interesting (Carmona and Delarue, 2018):

Example 2.5 (A Toy Model of Systemic Risk; Carmona et al. (2015))**.** Consider a toy example in which each particle X_t^i represents the logarithm of the cash reserves of bank $i = 1, \ldots, N$ at time t. The following SDE describes a simple model for borrowing and lending between banks

$$\mathrm{d}X_t^{i,N} = a(\bar{X}_t - X_t^{i,N})\mathrm{d}t + \sigma \mathrm{d}B_t^i, \qquad i = 1, \ldots, N$$

where for $\rho \in [-1, 1]$

$$\mathrm{d}B_t^i = \sqrt{1-\rho^2}\mathrm{d}W_t^i + \rho \mathrm{d}W_t^0$$

with W_t^i for $0 = 1, \ldots, N$ are independent one-dimensional Brownian motions, models the correlations between banks, $a, \sigma > 0$ and $\bar{X}_t^N$ denotes the sample mean at

time t. The sample mean in the drift coefficient represents the interaction between banks.

This model can be solved explicitly observing that the process $\bar{X}_t$ is a Brownian motion with diffusion coefficient $\sigma/\sqrt{N}$ and thus in the limit $N \to \infty$ each $X_t^{i,N}$ converge to independents Ornstein-Uhlenbeck processes.

2.4.2 Properties of mean-field SDEs

We collect here some results on mean-field SDEs (2.10) and their particle implementation (2.12). We focus on existence and uniqueness results and convergence results like those in Propositions 2.8– 2.9. First, we consider the case in which b, s are Lipschitz continuous as this is the case for which the majority of the results hold, then we move onto some results which require less stringent assumptions on b, s.

The Lipschitz Case

The case in which both b and s are Lipschitz continuous is well studied, and one can establish existence and uniqueness for both the mean-field SDE (2.10) and the particle system (2.12). Existence and uniqueness of a strong solution of (2.10) can be obtained, under the assumption that X_0 is square integrable, exploiting fixed point arguments (e.g. Jourdain et al. (2008); McKean (1967); Sznitman (1991)) or considering (2.10) as the limit for $N \to \infty$ of the particle system (2.12) and exploiting the propagation of chaos result in Proposition 2.8 (Méléard, 1996).

In the case in which the drift can be decomposed into a term involving the interaction and one which only depends on the process itself

$$b(X_t, \rho_t) = b_1(X_t) + \varepsilon b_2(X_t, \rho_t)$$

for $\varepsilon > 0$ small (i.e. if the interaction term is small), a number of results on the uniform ergodicity of (2.10) have been established (Bogachev et al., 2019; Butkovsky, 2014; Eberle, 2016; Malrieu, 2001). Similar results also exists for mean-field SDEs with unit diffusion coefficient (Benachour et al., 1998; Veretennikov, 2006) The general case has been considered in Song (2020) under smoothness assumptions on b, s involving Lions derivatives in measure spaces.

In the Lipschitz case, existence and uniqueness of a solution for the N-particle system (2.12) carries over from the existence and uniqueness of (2.10) (e.g. Protter

(2005, Theorem 7, page 259)). In the particular case

$$b(X_t, \rho_t) = \int \rho_t(\mathrm{d}y)\beta(X_t, y) \qquad s(x, \rho_t) = \int \rho_t(\mathrm{d}y)\sigma(X_t, y)$$

with β, σ Lipschitz, propagation of chaos results appeared early on in the literature, in this case the rate in Proposition 2.8 is $\gamma = 1$ (McKean, 1967; Sznitman, 1991). The general case has first been addressed in Oelschlager (1984), in which convergence of the left-hand-side of (2.13) is established without providing explicit rates. Under strong integrability assumptions on X_0, Jourdain et al. (2008) and Carmona (2016, Theorem 1.10) show that the rate deteriorates with the dimension d of X_t, $\gamma = 2/(d+4)$, however, sharper results exist in particular contexts (e.g. convergence of neural networks; De Bortoli et al. (2020)).

Given the generator of the N-particle system (2.12) for any twice-differentiable function $V : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}$ with bounded derivatives

$$\mathcal{L}^N V(x^{1:N}) = \sum_{i=1}^{N} L^{i,N}_{\rho^N(x)} V(x^{1,N}, \ldots, x^{i,N}, \ldots, x^{N,N}),$$

where $L^{i,N}_{\rho^N(x)}$ is the generator of the mean-field SDE (2.10) acting over any twice-differentiable function $V : \mathbb{R}^d \to R$ with bounded derivatives

$$L_{\rho(x)} V(x) := \sum_{k=1}^{d} b(x, \rho) \frac{\partial V}{\partial x_k}(x) + \frac{1}{2} \sum_{k=1}^{d} \sum_{l=1}^{d} \left(s(x, \rho) s(x, \rho)^T \right) \frac{\partial^2 V}{\partial x_k \partial x_l}, \tag{2.16}$$

applied to component i of V, the superscript T denotes the matrix transpose and $\rho^N(x)$ is the empirical measure associated with the particle system; the ergodic properties of the particle system (2.12) can be established using standard Foster–Lyapunov conditions (Meyn and Tweedie, 1993).

The Euler scheme (2.14) is analysed in Bossy and Talay (1997) and, in the Lipschitz case, admits strong order of convergence 0.5 and thus $\delta = 1$ in Proposition 2.9. As observed by the authors, this rate is sub-optimal; indeed Antonelli and Kohatsu-Higa (2002) show that the rate can be improved to $\delta = 2$. In the case of mean-field SDEs, convergence of the empirical measure (2.15) to ρ_t is also of interest. In particular, if b, s are bounded and sufficiently smooth, then $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ admits a density with respect to the Lebesgue measure for all t (Antonelli and Kohatsu-Higa, 2002) and smooth approximations of ρ_t can be obtained from the particle system using

standard kernel density estimator (KDE) procedures:

$$\tilde{\rho}_k^N(x) := \frac{1}{N}\sum_{i=1}^{N} |\Sigma|^{-1/2}\phi\left(\Sigma^{-1/2}(X_k^{i,N} - x)\right)$$

where ϕ is the density of d-dimensional standard Gaussian and Σ is a bandwidth matrix. In the one dimensional case, if $\mathbf{H} = \Delta t^2$ we have the following bound on the error of the KDE estimator for any fixed $t \in (0, T]$

$$\int \mathbb{E}\left[\left|\rho_t(x) - \tilde{\rho}_t^N(x)\right|\right] \mathrm{d}x \leq D\left(N^{-1/2} + \Delta t + N^{-1/2}\Delta t^{1/4}\right)$$

where with a slight abuse of notation we denote by ρ_t both the measure and its density and D is a finite constant independent of $N, \Delta t$ (Antonelli and Kohatsu-Higa, 2002, Theorem 3.1). The generalisation for $d > 1$ follows straightforwardly.

Non-Lipschitz Coefficients

The case in which the drift and diffusion coefficients are not Lipschitz continuous is less understood and fewer results exist. A complete review is out of the scope of this thesis, however we collect here some recent results which extend the results above to the non-Lipschitz case.

A case of particular interest is that in which $b(x, \rho)$ is not Lipschitz continuous with respect to the state x but is globally Lipschitz with respect to ρ. In this setting, existence and uniqueness of strong solutions of (2.10) can be obtained if b is locally Lipschitz continuous with polynomial growth in x (Dos Reis et al., 2019). Propagation of chaos results have also been established with rate $\gamma = 1/2$ in the one dimensional case and with rate deteriorating with dimension for $d > 1$ (Dos Reis et al., 2021).

When the coefficients of the SDE are only locally Lipschitz, the explicit Euler scheme does not perform well (Hutzenthaler et al., 2012). Bao et al. (2020) study tamed Euler schemes for (2.12)

$$X_{k+1}^{i,N} = X_k^{i,N} + \frac{b(X_k^{i,N}, \rho_k^N)}{1 + \Delta t\|b(X_k^{i,N}, \rho_k^N)\|_2}\Delta t + \sigma(X_k^{i,N}, \rho_k^N)\Delta W_k^i, \tag{2.17}$$

and show that Proposition 2.9 holds with $\delta = 1$. This modification of the drift guarantees that the norm of $\Delta t b(X_k^{i,N}, \rho_k^N)/(1 + \Delta t\|b(X_k^{i,N}, \rho_k^N)\|_2)$ is bounded by 1 for all $k, \Delta t$ and therefore prevents the drift term from taking extraordinarily large values. Additionally, the tamed Euler scheme (2.17) coincides with the explicit Euler

method (2.14) up to terms of second order (Hutzenthaler et al., 2012).

2.5 Summary

We have introduced two families of interacting particle systems, sequential Monte Carlo and systems approximating mean-field SDEs. Both classes allow us to approximate a limiting object by evolving a population of N particles through time, as the population size goes to infinity a number of convergence properties hold which guarantee the effectiveness of the Monte Carlo methods.

We have collected in this chapter all the definitions and results we rely on in the development of this work. The algorithms proposed in Chapter 5 and 7 are non-standard versions of those introduced in this chapter, thus their theoretical analysis will exploit some of the results introduced here.

Part I

Interacting Particle Methods for Fredholm Integral Equations

3

Fredholm Integral Equations of the First Kind

3.1 Introduction

Integral equations, in contrast to differential equations, is the name given by du Bois-Reymond (1888) to a class of mathematical models in which the unknown function appears within an integral sign. Integral equations gained popularity during the 20th century, thanks to the work of Volterra (1913) and Fredholm (1903) who started studying integral equations systematically. Their studies led to one of the main results in the theory of integral equations, the Fredholm alternative, which addresses the existence of solutions. Integral equations have been the topic of several reviews and monographs over the years: an introduction to the origins of the field is given in Michal (1950); Groetsch (2007) provides a short review of integral equations of the first kind while a complete discussion is given in e.g. Kress (2014). The series of publications *Integral Equations and Operator Theory* provides recent developments in the field.

The particular type of integral equations considered in this work was first introduced by Fredholm (1903) as a generalisation of linear systems of equations to the infinite-dimensional setting. These equations arise in a variety of fields, ranging from the study of partial differential equations (Colton and Kress, 2012; Tanana et al., 2016) to image reconstruction (Aster et al., 2018; Clason et al., 2020; Zhang et al., 2019) and density deconvolution (Delaigle, 2008; Ma, 2011; Pensky et al., 2017; Yang et al., 2020).

In the last century, there has been a great interest in studying the existence of solutions and solution methods for Fredholm integral equations of the first kind. While existence theorems appeared early in the literature (e.g. Fredholm (1903); Picard (1910)), finding good solution methodologies is an ongoing area of research.

In fact, these integral equations are the prototypical example of ill-posed linear inverse problems, as the solution is often non-unique and unstable to changes in g. The lack of stability of (3.1) is the primary concern when attempting to solve Fredholm integral equations: in practical applications we often only have access to (noisy) observations from g and not to its analytic form, and the instability of (3.1) means that small errors in g do not necessarily correspond to small errors in the recovered solution f (Groetsch, 2007).

A common technique to overcome non-uniqueness and instability is to consider regularisation. Regularised solutions can be obtained by projecting the infinite dimensional integral equation onto finite dimensional spaces (Kress, 2014, Chapter 17) or by solving a minimisation problem associated with the integral equation (e.g., Tikhonov's regularisation (Phillips, 1962; Tikhonov, 1963), maximum likelihood estimation (Mülthei et al., 1989) and maximum entropy methods (Kopeć, 1993; Mead, 1986)). However, solving the regularised problem remains computationally very challenging. In most cases, an approximate solution is obtained iteratively using a fixed discretisation of its domain (Burger et al., 2019; Chae et al., 2018a; Green, 1990; Yang et al., 2020) or by assuming that the solution can be expressed as a linear combination of basis functions. Common choices of basis functions are polynomials (Mead, 1986), piecewise linear functions (Jin and Ding, 2016), B-splines (Islam and Smith, 2020); wavelets are also widely used, as they lead to faster algorithms compared to the other methods (Maleknejad and Sohrabi, 2007). While these approaches are common, they have several drawbacks: using a fixed discretisation of the domain leads to approximate solutions which are piecewise constant or linear; on the other hand, in order to implement algorithms based on basis functions expansions, the number of basis functions involved needs to be finite. In practice, the solutions of (non-degenerate) integral equations are rarely given by a finite combination of basis functions, nor can be expressed as piecewise constant or linear functions.

We briefly review Fredholm integral equations of the first kind in Section 3.2 and provide some examples of applications of Fredholm integral equations in Section 3.2.1. A short summary of popular regularisation techniques is given in Section 3.2.2. Section 3.3 focuses on the solution method of Kondor (1983), based on an infinite dimensional expectation maximisation algorithm (Dempster et al., 1977), the issues encountered by this method because of the ill-posedness of Fredholm integral equations of the first kind motivate the following evolution of this thesis.

3.2 Fredholm Integral Equations of the First Kind

Fredholm integral equations of the first kind

$$h(y) = \int_{\mathbb{X}} f(x)g(y \mid x)\mathrm{d}x \qquad \forall y \in \mathbb{Y}, \tag{3.1}$$

with $\mathbb{X}$, $\mathbb{Y}$ suitable subsets of $\mathbb{R}^{d_{\mathbb{X}}}, \mathbb{R}^{d_{\mathbb{Y}}}$ respectively, are inverse problems in which f is the unknown object to be inferred and h, g are known functions.

Example 3.1. (Gaussian integral) Consider two one-dimensional Gaussian distributions on $\mathbb{X} = \mathbb{Y} = \mathbb{R}$

$$h(y) = \mathcal{N}(y; \mu, \sigma_h^2), \qquad g(y \mid x) = \mathcal{N}(y; x, \sigma_g^2),$$

such that $\sigma_h^2 = \sigma_f^2 + \sigma_g^2$. One can easily show that setting $f(x) = \mathcal{N}(x; \mu, \sigma_f^2)$ satisfies equation (3.1) and the Fredholm integral equation can be solved analytically.

We now introduce the framework in which we will work for the remainder of this thesis.

Assumption 3.1. The functions f, g, h and the sets $\mathbb{X}, \mathbb{Y}$ in (3.1) satisfy

(a) $\mathbb{X} \subseteq \mathbb{R}^{d_{\mathbb{X}}}$ and $\mathbb{Y} \subseteq \mathbb{R}^{d_{\mathbb{Y}}}$ are subsets of Euclidean spaces endowed with the Borel σ-algebras $B(\mathbb{X})$ and $B(\mathbb{Y})$ respectively.

(b) f and h are (Lebesgue) probability densities on $\mathbb{X}$ and $\mathbb{Y}$, respectively, and $g(y \mid x)$ is the density of a Markov kernel from $\mathbb{X}$ to $\mathbb{Y}$.

We distinguish two sub-cases of Assumption 3.1-(a), in Chapters 4-5 we take $\mathbb{X} \subset \mathbb{R}^{d_{\mathbb{X}}}$ and $\mathbb{Y} \subset \mathbb{R}^{d_{\mathbb{Y}}}$ compact subsets of Euclidean spaces, while in Chapters 6-7 we take $\mathbb{X} = \mathbb{R}^{d_{\mathbb{X}}}$ and $\mathbb{Y} = \mathbb{R}^{d_{\mathbb{Y}}}$. These two cases cover most applications of Fredholm integral equations of the first kind, we give some examples in Section 3.2.1.

As observed by Chae et al. (2018a, Section 6), Assumption 3.1-(b) is not too restrictive. Many applications of Fredholm integral equations are concerned with the reconstruction of functions f that are a priori known to be non-negative (Clason et al., 2020); in addition, a wide class of Fredholm integral equations of the first kind can be cast into this framework: if f, h and g are positive and appropriately integrable functions, using the following transformations

$$\tilde{h}(y) = \frac{h(y)}{\int_{\mathbb{Y}} h(y')\mathrm{d}y'}, \qquad \tilde{g}(y \mid x) = \frac{g(y \mid x)}{\int_{\mathbb{Y}} g(y' \mid x)\mathrm{d}y'}, \qquad \tilde{f}(x) = \frac{f(x)\int_{\mathbb{Y}} g(y' \mid x)\mathrm{d}y'}{\int_{\mathbb{Y}} h(y')\mathrm{d}y'}$$

we obtain an integral equation

$$\tilde{h}(y) = \int_{\mathbb{X}} \tilde{f}(x)\tilde{g}(y \mid x)\mathrm{d}x \qquad \forall y \in \mathbb{Y},$$

involving densities as in Assumption 3.1-(b). If f, h are bounded below and g is positive, the shifted functions $f(x) + t$, $h(y) + t\int_{\mathbb{X}} g(y \mid x)\mathrm{d}x$ are positive for some appropriate $t > 0$, and, assuming integrability, we can apply the normalisation described above. Finally, if g is not necessarily non-negative, (3.1) can be transformed into a non-negative integral equation by considering the positive and negative parts of g (Chae et al., 2018a, Section 6).

In applications the analytic form of h is often unknown, and the available data arise from discretisation of h over $\mathbb{Y}$ or from sampling. In the first case, the functional representation of h is obtained by assuming that h is piecewise constant, so that h takes constant values on a partition of $\mathbb{Y}$, as in e.g. Vardi and Lee (1993). In the second case, a functional representation of h can be obtained through an histogram or a kernel density estimator, the latter is considered in, e.g. Ma (2011).

In the remainder of this thesis we do not consider degenerate kernels, $g(y \mid x) = \sum_{i=1}^{n} a_n(x)b_n(y)$ for some $n \in \mathbb{N}$, as in this case solving the integral equation (3.1) reduces to solving a finite system of linear equations (Groetsch, 2007). We point out that there is another class of integral equations to which the algorithms developed in this thesis could be applied: Volterra integral equations of the second kind, which can be seen as a special case of Fredholm integral equations of the first kind for which $\mathbb{X} = \mathbb{Y}$ and $g(y \mid x) = 0$ for $x > y$. However, Volterra integral equations are not ill-posed and can be efficiently solved using the Von-Neumann representation discussed in Section 2.3.3.

3.2.1 Applications

Fredholm integral equations of the first kind are often used to model the task of reconstructing a signal f from an observed distorted signal h when the type of distortion g is known. This is a common problem in engineering, science and statistics.

The earliest applications of Fredholm integral equations arise in physics. For instance, in potential theory, (3.1) relates the gravitational potential of an object to its mass density (Joachimsthal, 1861). Often, integral equations arise as a counterpart to partial differential equations (PDEs) with boundary conditions (Green, 1828). In particular, inverse boundary problems are modelled through (3.1) (Tanana et al.,

2016). An example of inverse boundary problem is the reconstruction of the shape of an object from the diffusion of an acoustic/electromagnetic wave which hits the object itself (Colton and Kress, 2012).

In image processing, (3.1) describes the deconvolution problem, i.e. the reconstruction of images from distorted observations. Simple examples are motion deblurring, in which h represents a blurred version of the image f and g describes the type of motion causing the blur (see e.g. Lee and Vardi (1994) for some examples) and positron emission tomography (PET) (Phelps, 2000), in which reconstructions of brain cross-sections are obtained from the radial data provided by the PET scanner (Vaquero and Kinahan, 2015). 3-dimensional examples arise in e.g. electron microscopy (Scheres et al., 2007).

From the statistical point of view, (3.1) describes the task of reconstructing the density f from a set of observations from h (Laird, 1978). This problem is known as density deconvolution or indirect density estimation (Delaigle et al., 2008; Ma, 2011; Pensky et al., 2017; Yang et al., 2020); and appears in e.g., causal inference with missing confounders, where the missing data mechanism has to be inferred from the distribution of the non-missing observed data (Miao et al., 2018; Yang et al., 2019), inference with spatially censored data (Fan et al., 2011) and estimation of nonlinear regression functions (Hall et al., 2005). In epidemiology, these equations link the unknown reproduction number of a disease to the observed number of deaths (Goldstein et al., 2009; Gostic et al., 2020; Marschner, 2020).

Similar indirect density estimation problems also arise in stereology, when the size distribution of a set of spherical particles has to be recovered from the distribution of the diameters of the section profiles (Wicksell's corpuscle problem; Wicksell (1925) and Silverman et al. (1990)).

3.2.2 Regularisation Methods

Since Fredholm integral equations of the first kind can be seen as an infinite dimensional linear system, it is natural to consider least squares solutions

$$f^\star := \arg\min \int_{\mathbb{Y}} \left| h(y) - \int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x \right|^2 \mathrm{d}y$$

over an appropriate set of functions f; Tikhonov's regularisation (Phillips, 1962; Tikhonov, 1963) consists of approximating the solution of the integral equation with

the least square solution of minimum $\mathbb{L}^2$-norm

$$f^\star := \arg\min \int_{\mathbb{Y}} \left| h(y) - \int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x \right|^2 \mathrm{d}y + \alpha \int_{\mathbb{X}} |f(x)|^2 \mathrm{d}x \tag{3.2}$$

for some regularisation parameter $\alpha > 0$. This regularisation is particularly convenient, as the regularised integral equation is a Fredholm equation of the second kind, a well-posed problem for which stable solution methods exist (Kress, 2000, Chapter 12-14). An iterative method to minimise the square loss has been proposed in Landweber (1951). Alternatives to the Tikhonov regularisation for the least squares problem include total variation regularisation (Rudin et al., 1992), sparsity regularisation through the $\mathbb{L}_1$ norm of the solution (Donoho, 1992) and entropy regularisation (Amato and Hughes, 1991).

Maximum entropy methods (Jaynes, 1957a,b) have been widely studied; these methods approximate f by finding a function maximising the differential entropy subject to moment constraints obtained by integrating both sides of (3.1) with respect to a set of basis functions (Beylkin et al., 1991; Islam and Smith, 2020; Jin and Ding, 2016; Kopeć, 1993; Mead, 1986). Methods involving the Kullback–Leibler divergence rather than the $\mathbb{L}_2$ distance have also been considered (Kondor, 1983; Resmerita and Anderssen, 2007). This class of algorithms is particularly interesting due to its link with maximum likelihood estimation and will be described in Section 3.3.

Regularisation can also be obtained by discretisation, i.e. by projecting the integral equation onto a finite dimensional space (e.g. Kress (2014, Chapter 17)). A recent review of regularisation techniques for inverse problems is given in Benning and Burger (2018), while the monograph Engl et al. (1996) focuses on classical regularisation techniques. Yuan and Zhang (2019) provide a recent and extensive review of numerical methods for integral equations.

3.3 Expectation Maximisation Algorithms

If we consider (3.1) as an indirect density estimation problem, it is natural to consider maximum likelihood estimates for f. This can in principle be achieved by maximising an incomplete data likelihood for f through the Expectation Maximisation (EM) algorithm (Dempster et al., 1977). Nevertheless, the maximum likelihood estimator is not consistent, as the parameter to be estimated (i.e. f) is infinite dimensional and the lack of continuous dependence on h of (3.1) discussed in the introduction aggravates this problem (Laird, 1978; Silverman et al., 1990).

To overcome these issues, several regularisation techniques have been proposed in the literature. We briefly review a number of iterative schemes based on the EM algorithm which aim to find approximate solutions of (3.1) through regularisation.

The starting point is the iterative method of Kondor (1983), an infinite dimensional EM algorithm

$$f_{n+1}(x) = f_n(x) \int \frac{g(y \mid x)}{\int f_n(z) g(y \mid z) \mathrm{d}z} h(y) \mathrm{d}y, \tag{3.3}$$

which minimises the Kullback–Leibler divergence

$$\mathrm{KL}\left(h, \int_{\mathbb{X}} f(x) g(\cdot \mid x) \mathrm{d}x\right) = \int_{\mathbb{Y}} h(y) \log\left(\frac{h(y)}{\int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x}\right) \mathrm{d}y, \tag{3.4}$$

with respect to f over the set of probability densities on $\mathbb{X}$ (Mülthei et al., 1989). Minimising (3.4) is equivalent to maximising

$$\Lambda(f) := \int_{\mathbb{Y}} h(y) \log \int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x \mathrm{d}y.$$

The functional Λ can be seen as a continuous or asymptotic version of the incomplete data log-likelihood for the function f given a finite sample from h, $y_1, \ldots, y_M$, in fact, in the case of a finite number of observations the log-likelihood is given by (Chae et al., 2018b)

$$L(f) = \sum_{m=1}^{M} \log \int_{\mathbb{X}} g(y_m \mid x) f(x) \mathrm{d}x,$$

and $\Lambda(f) = \lim_{M \to +\infty} M^{-1} L(f)$.

The EM iteration (3.3) has a number of good properties: iterating (3.3) monotonically decreases (3.4) (Mülthei et al., 1987, Theorem 7) and if the iterative formula converges to a limit, then this is a minimiser of (3.4) (Mülthei et al., 1987, Theorem 8). However, the minimiser need not to be unique. Convergence of the EM iteration (3.3) to a fixed point has recently been proved under the existence of a sequence $(f_s^\star)_{s \geq 1}$ with $h_s^\star(y) = \int_{\mathbb{X}} f_s^\star(x) g(y \mid x) \mathrm{d}x$, such that $\mathrm{KL}(h, h_s^\star)$ converges to the infimum of (3.4) and some additional integrability conditions (Chae et al., 2018a).

In general, implementing the recursive formula (3.3) analytically is not possible and approximations are needed. Under the assumption of piecewise constant signals f, h and g, we can discretise the EM recursion (3.3) over a discretisation grid

obtained by dividing $\mathbb{X}$ and $\mathbb{Y}$ into bins $\mathbb{X}_1, \ldots, \mathbb{X}_B$ and $\mathbb{Y}_1, \ldots, \mathbb{Y}_D$

$$f_b = \frac{1}{|\mathbb{X}_b|}\int_{\mathbb{X}_b} f(x)\mathrm{d}x, \quad h_d = \frac{1}{|\mathbb{Y}_d|}\int_{\mathbb{Y}_d} h(y)\mathrm{d}y, \quad g_{bd} = \frac{1}{|\mathbb{X}_b||\mathbb{Y}_d|}\int_{\mathbb{Y}_d}\int_{\mathbb{X}_b} g(y \mid x)\mathrm{d}y,$$

where $|A|$ denotes the Lebesgue measure of set A, for $b = 1, \ldots, B$ and $d = 1, \ldots, D$. In practice, for a large enough number of bins B, D and if the densities f, h, g are continuous, we can approximate the value of each density on a given bin by its value at the bin centre, e.g. $f_b \approx f(x_b)$, where x_b is the centre of bin b. With this discretisation, the EM recursion (3.3) reduces to the EM algorithm for Poisson data (Vardi and Lee, 1993), known as the Richardson-Lucy (RL) algorithm in the image processing field (Lucy, 1974; Richardson, 1972), where the intensities of pixels are modelled as Poisson counts,

$$f_b^{(n+1)} = f_b^{(n)} \sum_{d=1}^{D} \left(\frac{h_d g_{bd}}{\sum_{k=1}^{B} f_k^{(n)} g_{kd}} \right). \tag{3.5}$$

The Iterative Bayes (IB) algorithm of Ma (2011) considers the case in which only samples from h are available. These samples are used to build a kernel density estimator (KDE) for h, which is then plugged into the discretised EM iteration (3.5).

Example 3.2. (Gaussian integral) Consider the Fredholm integral equation defined in Example 3.1 with $\sigma_f^2 = 0.043^2, \sigma_g^2 = 0.045^2$ and $\mathbb{X} = \mathbb{Y} = \mathbb{R}$ (although note that $|1 - \int_0^1 f(x)dx| < 10^{-30}$ and restricting out attention to $[0, 1]$ would not significantly alter the results). We apply the EM iteration (3.5) for 100 steps with initial distribution f_1 Uniform on $[0, 1]$ and discretisation grid given by $B = 100$ equally spaced intervals in $[0, 1]$, noting that discretisation schemes essentially require known compact support and this interval contains almost all of the probability mass (Figure 2). Even for this simple one-dimensional example, the reconstructions provided by EM are non-smooth and fail to recover the shape of f.

As discussed earlier, despite being popular and easy to implement, the EM algorithm (3.5) has a number of drawbacks. In particular, after a certain number of iterations the EM approximations deteriorate resulting in high variance estimates that lack smoothness and give non-regular estimates of f (Nychka, 1990; Silverman et al., 1990). Byrne and Eggermont (2015) emphasize that (3.4) does not deal with the ill-posedness of the problem and further regularisation is needed.

A natural way to introduce regularisation is considering maximum penalised

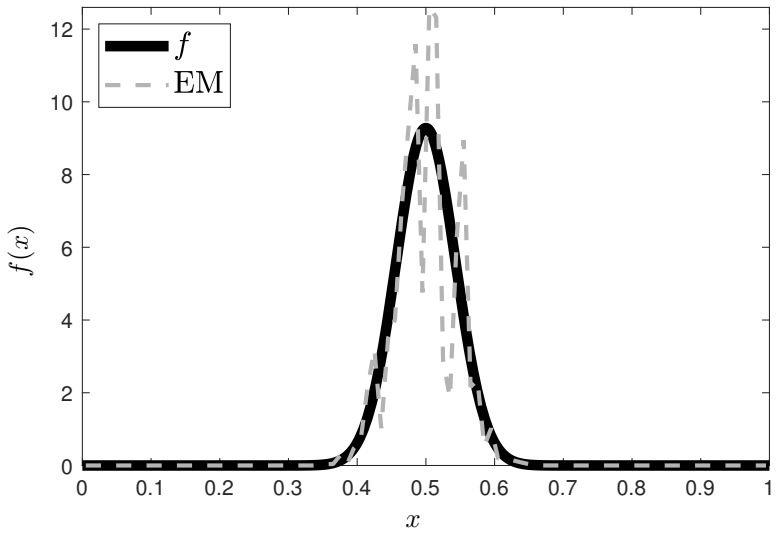


Figure 2: Reconstruction given by the EM algorithm. The target density is $\mathcal{N}(x; 0.5, 0.043^2)$.

likelihood estimation (MPLE)

$$\Lambda'(f) := \int_{\mathbb{Y}} h(y) \log \int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x \mathrm{d}y - P(f),$$

where P is a penalty term (e.g. Green (1990)). Popular choices of P are Gibbs smoothing, which penalises large differences in neighbouring values of f (e.g. Lange (1990)), $\mathbb{L}_1$ penalty of the first order derivative of f, leading to Total Variation regularisation (Dey et al., 2004) and general roughness penalties (e.g. Good (1971)).

In most cases, an updating formula like (3.5) cannot be obtained straightforwardly for MPLE because the derivative of $P(f)$ usually involves several derivatives of f (Green, 1990). A possible solution is to update the estimate of f from iteration f_n to f_{n+1} evaluating the penalty term at f_n, rather than at the new value f_{n+1}. This is known as the one-step late (OSL-EM) algorithm (Green, 1990). The resulting update formula is usually easier to compute but there is no guarantee that each iteration will increase the penalised log-likelihood. However, if convergence occurs, the OSL-EM algorithm converges more quickly that the corresponding EM for penalised likelihood.

3.4 Connections with Other Works

Among the numerous solution methodologies for Fredholm integral equations the two presented below share some connections with those that will be explored in this work. Section 3.4.1 explores the use of divergences other than the Kullback–Leibler divergence while Section 3.4.2 presents a particle method which performs Tikhonov regularisation.

3.4.1 Alternative Divergences

In recent years, there has been an increasing interest in considering families of divergences between probability measures which are more flexible than the Kullback–Leibler divergence, examples of such divergences are α-, β- and γ- divergences (Cichocki and Amari, 2010). This direction has been explored in e.g., Bayesian inference (Bissiri et al., 2016; Jewson et al., 2018), variational inference (Blei et al. (2017, Section 5.4)), Knoblauch et al. (2019) and sequential Monte Carlo methodology (Boustati et al., 2020). These families of divergences are very flexible, and, with careful choice of the parameters α, β and γ, allow to enforce desirable properties such as robustness to outliers, ability to capture tail behaviour or to target the mode and many others (Regli and Silva, 2018).

In the context of Fredholm integral equations, distances other than the Kullback-Leibler divergence (leading to maximum likelihood estimation) and the Euclidean distance (leading to Tikhonov's regularisation) have been little explored. We summarise here two examples of use of β- and α-divergences; β-divergences guarantee robustness to outliers, while α-divergences guarantee better coverage of regions of small probability than the Kullback–Leibler divergence (Regli and Silva, 2018).

Pouchol and Verdier (2020, Section 4.2) consider β-divergences between two density functions p_1, p_2

$$D_\beta(p_1, p_2) := \int \frac{1}{\beta(\beta-1)} \left(p_1(t)^\beta + (\beta-1)p_2(t)^\beta - \beta p_1(t)p_2(t)^{\beta-1} \right) \mathrm{d}t,$$

for $\beta \in [0, 2]$, and propose to solve (3.1) by minimising $D_\beta\left(h, \int_{\mathbb{X}} f(x)g(\cdot \mid x)\mathrm{d}x\right)$. Pouchol and Verdier (2020, Theorem 1.1) shows that in the high-noise regime (i.e. when we only have access to a noisy version of h), which corresponds to the presence of outliers in the standard statistical setting, the solution f obtained by minimising the β-divergence will be sparse, an undesirable property for f which is usually expected to be smooth. This behaviour is closely related to the inconsistency of the

maximum likelihood estimator minimising (3.4) and could be, in principle, addressed by considering penalties such as those for maximum penalised likelihood estimation.

As in the case of the Kullback-Leibler divergence, it is possible to obtain an iterative scheme which monotonically decreases D_β (Pouchol and Verdier, 2020, Proposition C.1); this scheme admits the EM iteration as a special case, obtained with $\beta \to 1$. However, the iterative scheme of Pouchol and Verdier (2020) requires the computation of the adjoint of the operator $f \mapsto \int_{\mathbb{X}} f(x)g(y \mid x)\mathrm{d}x$, which can be rarely obtained in the continuous setting. The discretised version of the iterative scheme for β-divergences only requires the adjoint of the matrix discretisation of $f \mapsto \int_{\mathbb{X}} f(x)g(y \mid x)\mathrm{d}x$ and is therefore easier to implement, but requires knowledge of a functional form of h, a condition which is not generally satisfied (e.g. Delaigle (2008); Hall et al. (2005); Marschner (2020)). In fact, most of this thesis is concerned with algorithms which can be implemented when h is known only through a sample.

An alternative to β-divergences is given by α-divergences

$$D_\alpha(p_1, p_2) := \int \frac{1}{\alpha(\alpha-1)} \left[\left(\frac{p_1(t)}{p_2(t)} \right)^\alpha - 1 \right] p_2(t)\mathrm{d}t, \tag{3.6}$$

with $\alpha \in \mathbb{R} \setminus \{0, 1\}$. These divergences provide generalisations of the KL divergence which are more robust with respect to outliers (Cichocki and Amari, 2010) and admit the KL divergence and the reverse KL divergence as special cases, obtained for $\alpha \to 1$ and $\alpha \to 0$ respectively.

In the context of variational inference, Daudel et al. (2020) develop an iterative algorithm, named (α, Γ)-descent, which minimises (3.6) when p_1 is a mixture distribution

$$p_1(t) = \int \mu(\mathrm{d}\theta) q(\theta, t),$$

where μ is a probability measure and q is the density of a Markov kernel. The (α, Γ)-descent algorithm can be seen as a generalisation of the EM algorithm (3.3) for α-divergences.

We now derive the corresponding (α, Γ)-descent algorithm for the Fredholm integral equation (3.1). The (α, Γ)-descent algorithm for $D_\alpha\left(h, \int_{\mathbb{X}} f(x)g(\cdot \mid x)\mathrm{d}x\right)$ consists of an expectation step

$$\tilde{f}_n(x) = \frac{1}{(\alpha-1)} \int_{\mathbb{Y}} g(y \mid x) \left(\frac{\int_{\mathbb{X}} f_n(z)g(y \mid z)\mathrm{d}z}{h(y)} \right)^{\alpha-1} \mathrm{d}y, \tag{3.7}$$

and an iteration step

$$f_{n+1}(x) = f_n(x)\frac{\Gamma\left(\tilde{f}_n(x)\right)}{Z}, \tag{3.8}$$

where $\Gamma : \mathbb{R} \to (0, \infty)$ is a decreasing, continuously differentiable function which satisfies a particular set of inequalities and Z is the appropriate normalising constant. This scheme shares a number of properties with the EM algorithm: it monotonically decreases D_α (Daudel et al., 2020, Theorem 1) and, for a particular choice of Γ, if a limit exists then this is a fixed point of the iterative map and a minimiser of D_α (Daudel et al., 2020, Theorem 4).

As for the EM scheme of Kondor (1983), the (α, Γ)-descent iteration cannot be implemented analytically. Daudel et al. (2020) propose a stochastic version of the (α, Γ)-descent which requires sampling from $\int_{\mathbb{X}} f_n(x)g(y \mid x)\mathrm{d}x$ at each iteration. To simplify the algorithm, one can select the initial distribution f_1 to be a weighted sum of Dirac δs, in this case the (α, Γ)-descent corresponds to an update of the weights of the mixture and the approximation f_n is a mixture of Dirac delta at each n. In the case of Fredholm integral equations, a smooth reconstruction could be obtained by, e.g., standard kernel density estimation procedures (Silverman, 1986).

The (α, Γ)-descent algorithm, although introduced in a different context, is similar in spirit to the algorithms we study in this thesis. However, the EM algorithm which minimises the Kullback–Leibler divergence does not require the specification of the function Γ (which guarantees that the iterates are non-negative and influences the resulting reconstructions; see Daudel et al. (2020, Section 5)) and the additional normalisation in (3.8), as the iterates provided by (3.3) and its discrete counterpart (3.5) are already normalised.

Since the EM algorithm seems the most appropriate in the context of Fredholm equations, alternative divergences are not further considered in the remainder of this work.

Remark 3.1. The β-divergence considered in this section addresses the noise in the observations from h; however, there is another form of robustness which could be considered when solving (3.1): the misspecification of the kernel g. In practical applications the analytic form of g will often be unknown, and it is necessary to model g to fit the available information on the distortion process and taking into account computational considerations as well (see, e.g, 7.4.2). To the best of our knowledge, a systematic study of the effect of misspecification on g as so far not been conducted; nevertheless, we believe it would be of great interest for the broader

community working on integral equations.

3.4.2 A Particle Method for Inverse Problems

Fredholm integral equations (3.1) are one of the most popular examples of linear inverse problems; in this section we make the connection with an interacting particle method for generic (possibly nonlinear) inverse problems which is similar to the particle methods that we will study in this thesis.

Consider the inverse problem of finding $\mathbf{f} \in \mathbb{R}^B$ from $\mathbf{h} \in \mathbb{R}^D$ where

$$\mathbf{h} = G(\mathbf{f}) + \epsilon, \tag{3.9}$$

and $G : \mathbb{R}^B \to \mathbb{R}^D$ is a known non-linear forward operator. Under the assumption of piecewise constant signals over a fixed grid with B equally spaced bins in $\mathbb{X}$ and D equally spaced bins in $\mathbb{Y}$, the linear Fredholm integral equation (3.1) is a special case of (3.9) with G a $D \times B$ matrix

$$\mathbf{h} = G\mathbf{f},$$

where $\mathbf{f} = (f_b)_{b=1}^B$, $\mathbf{h} = (h_d)_{d=1}^D$ and G is the matrix with entries g_{bd} for $b = 1, \ldots, B$ and $d = 1, \ldots, D$.

Garbuno-Inigo et al. (2020) propose an interacting particle method, called ensemble Kalman sampler (EKS), to provide approximate solutions of (3.9) without requiring the use of derivatives or adjoints of the forward model G. Consider the following noisy version of (3.9) with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \Gamma)$ with covariance matrix Γ,

$$\mathbf{h} = G(\mathbf{f}) + \epsilon,$$

and adopt a Bayesian approach: given a Gaussian prior on $\mathbf{f}$ with covariance matrix $\Gamma_\alpha := \Gamma_0/\alpha^2$ for some $\alpha > 0$ and Γ_0 a positive-definite symmetric matrix, and using the l_2 norm as loss function, one can obtain the posterior for $\mathbf{f}$

$$\pi(\mathbf{f}) \propto \exp\left(-\frac{1}{2}\left(\|\mathbf{h} - G(\mathbf{f})\|_\Gamma^2 + \|\mathbf{f}\|_{\Gamma_\alpha}^2\right)\right) \tag{3.10}$$

where for any positive-definite symmetric Γ

$$\|\mathbf{f}\|_\Gamma^2 := \left\langle \mathbf{f}, \mathbf{f}^T \right\rangle_\Gamma = \left\langle \Gamma^{-1/2}\mathbf{f}, \Gamma^{-1/2}\mathbf{f}^T \right\rangle = \|\Gamma^{-1/2}\mathbf{f}\|_2^2$$

with $\langle \cdot, \cdot \rangle$ denoting the dot-product and superscript T the matrix transpose. The maximum a posteriori given by (3.10) corresponds to the solution of (3.9) obtained by Tikhonov's regularisation, i.e. minimising the discretised version of (3.2)

$$\Phi(\mathbf{f}) := \frac{1}{2}\|\mathbf{h} - G(\mathbf{f})\|_\Gamma^2 + \frac{\alpha}{2}\|\mathbf{f}\|_{\Gamma_0}^2$$

To sample from (3.10), Garbuno-Inigo et al. (2020) propose an ensemble Kalman sampler (EKS): take N particles $\mathbf{f}^{1,N}, \dots, \mathbf{f}^{N,N} \in \mathbb{R}^D$ and evolve them according to a non-linear SDE with drift

$$b\left(\mathbf{f}_t^{i,N}, \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{f}_t^{i,N}}\right) = \frac{1}{N}\sum_{k=1}^N \left\langle G(\mathbf{f}_t^{k,N}) - \bar{G}), G\mathbf{f}_t^{i,N} - \mathbf{h}\right\rangle_\Gamma \mathbf{f}_t^{k,N} - C(\mathbf{f}_t^{1:N})\Gamma_\alpha^{-1}\mathbf{f}_t^{i,N}$$

involving the covariance matrix of the particles and the average $\bar{G}$

$$\bar{G} := \frac{1}{N}\sum_{i=1}^N G(\mathbf{f}_t^{i,N}), \qquad C(\mathbf{f}_t^{1:N}) := \frac{1}{N}\sum_{i=1}^N (\mathbf{f}_t^{i,N} - \bar{\mathbf{f}}_t) \otimes (\mathbf{f}_t^{i,N} - \bar{\mathbf{f}}_t)$$

and diffusion coefficient $\sigma\left(\mathbf{f}_t^{i,N}, \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{f}_t^{i,N}}\right) = \sqrt{2C(\mathbf{f}_t^{1:N})}$. The term $G(\mathbf{f}_t^{k,N}) - \bar{G}$ in the drift b is an approximation of the derivative of G, which in the linear case coincide with the derivative itself.

In the limit $N \to \infty$ the particle system converges to a mean-field SDE whose stationary states minimise a functional involving $\Phi(\mathbf{f})$ (Garbuno-Inigo et al., 2020).

The EKS described above is similar in spirit to the mean-field SDE approach we will introduce in Chapters 6-7, with two key differences. We are concerned with linear operators for which computation of the derivative is straightforward and no approximations are needed. This allows us to avoid the discretisation step in (3.9) and to work directly with the densities f, h. To give some intuition on the difference between the two approaches consider again the toy model in Example 3.1. As we did for the EM algorithm discussed in Example 3.2 we fix the discretisation grid over $[0, 1]$ and set $D = B = 100$, that is each particle $\mathbf{f}^{i,N}$ is a 100-dimensional vector. On the contrary, in Chapter 7 we do not introduce any discretisation of the integral equation (3.1) and consider particles which are samples drawn from f. Thus, while in the EKS approach the particle population at any given time can be stored in a $D \times N$ matrix, in the case of the approach described in Chapter 7 this reduces to a N dimensional vector. Since the memory cost of EKS is considerably higher than that of the particle methods developed in this thesis, this approach will not be further discussed.

3.5 Summary

In this chapter we introduced Fredholm integral equations of the first kind and the framework in which we will work (i.e. that of probability densities in Euclidean spaces). Among the many regularisation techniques available in the literature, our focus is on maximum-likelihood based methods; in the following development of this work we will study two maximum penalised likelihood methods which aim at constructing smooth reconstructions of f. In Chapters 4-5 the penalty term is implicit, as we directly modify the iteration (3.1) introducing a smoothing component; in Chapters 6-7 we consider a maximum penalised likelihood with an entropic penalty.

4

Expectation Maximisation Smoothing for Fredholm Equations of The First Kind

A short version of this chapter is presented in (Crucinio et al., 2021b).

4.1 Introduction

As is the case for many inverse problems, Fredholm integral equations of the first kind are ill-posed and some type of regularisation is needed. Maximum penalised likelihood estimation (MPLE) is a natural way to introduce regularisation, but very often leads to an iterative scheme which cannot be implemented in practice. An easy-to-implement regularised version of the EM recursion (3.5) is the expectation maximization smoothing (EMS) algorithm of Silverman et al. (1990), an EM-like algorithm in which a smoothing matrix K is applied to the EM estimates at each iteration

$$f_b^{(n+1)} = \sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa^{(n)} \sum_{d=1}^{D} \left(\frac{h_d g_{\kappa d}}{\sum_{k=1}^{B} f_k^{(n)} g_{kd}} \right). \tag{4.1}$$

This algorithm has long been attractive from a practical point of view as the addition of the smoothing step to the EM recursion (3.5) gives good empirical results, with convergence occurring in a relatively small number of iterations (e.g. Li et al. (2017, 2020); Silverman et al. (1990)).

Example 4.1. (Gaussian integral) For the Gaussian integral Example 3.1 with $\sigma_f^2 = 0.043^2, \sigma_g^2 = 0.045^2$ and $\mathbb{X} = \mathbb{Y} = \mathbb{R}$ the EMS iteration for 100 steps with initial distribution f_1 Uniform on $[0, 1]$ and discretisation grid given by $B = 100$ equally spaced intervals in $[0, 1]$ gives smooth reconstructions (Figure 3).

The aim of this chapter is to introduce a continuous version of the EMS of

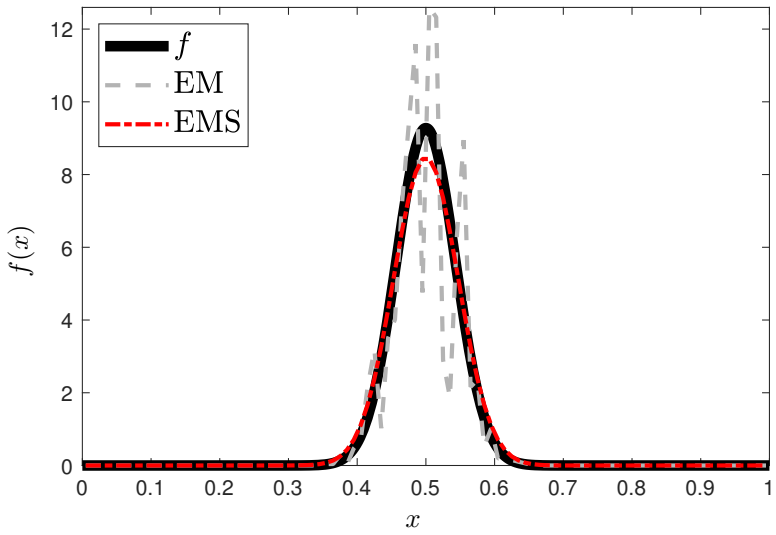


Figure 3: Reconstructions given by the EM algorithm without and with smoothing step. The target density is $\mathcal{N}(x; 0.5, 0.043^2)$, which solves the toy Fredholm integral equation in Example 3.1.

Silverman et al. (1990), to study its properties and establish existence of a fixed point of the corresponding iterative scheme.

4.2 Continuous EMS

In order to introduce a continuous version of the EMS iteration (4.1) we consider the EM iteration (3.3) and smoothing operators $\mathrm{O_K} : f \mapsto \mathrm{O_K} f$, with $\mathrm{O_K} f(\cdot) := \int_{\mathbb{X}} K(u, \cdot) f(u) \mathrm{d}u$, rather than smoothing matrices. The resulting iterative scheme is

$$\begin{aligned} f_{n+1}(x) &= \int_{\mathbb{X}} K(x', x) f_n(x') \int_{\mathbb{Y}} \frac{g(y \mid x') h(y)}{\int_{\mathbb{X}} f_n(z) g(y \mid z) \mathrm{d}z} \mathrm{d}y \mathrm{d}x' \\ &= \int_{\mathbb{X}} K(x', x) f_n(x') \int_{\mathbb{Y}} \frac{g(y \mid x') h(y)}{h_n(y)} \mathrm{d}y \mathrm{d}x', \end{aligned} \tag{4.2}$$

where $h_n(y) := \int_{\mathbb{X}} f_n(z) g(y \mid z) \mathrm{d}z$.

A similar approach in which a non linear smoothing operator

$$\widetilde{\mathrm{O_K}} f := \exp\left(\mathrm{O_K}(\log f)\right)$$

is applied to f_n both in the numerator and in the denominator of (3.3) has been considered by Eggermont and LaRiccia (1995). In this case the fixed point of the

iteration maximises the modified log-likelihood

$$\int_{\mathbb{Y}} h(y) \log \int_{\mathbb{X}} (\widetilde{\mathrm{O_K}} f)(x) g(y \mid x) \mathrm{d}x \mathrm{d}y - \int_{\mathbb{X}} f(x) \mathrm{d}x.$$

However, because of the non-linearity of the smoothing operator $\widetilde{\mathrm{O_K}}$, the algorithm of Eggermont and LaRiccia (1995) cannot easily be implemented without further approximation and requires the knowledge of the analytic form of h.

In order to study the properties of the EMS recursion (4.2), we consider the corresponding map from the set of unsigned measures of nonzero mass $\mathcal{M}^+(\mathbb{X})$ to the set of probability measures $\mathcal{P}(\mathbb{X})$.

Hence, we consider the EMS map, $\mathrm{F_{EMS}} : \mathcal{M}^+(\mathbb{X}) \to \mathcal{P}(\mathbb{X})$ such that

$$\mathrm{F_{EMS}} : \eta \mapsto \mathrm{F_{EMS}}\,\eta := \int_{\mathbb{X}} \eta(\mathrm{d}x') K(x', \cdot) \int_{\mathbb{Y}} \frac{g(y \mid x') h(\mathrm{d}y)}{\int_{\mathbb{X}} \eta(\mathrm{d}z) g(y \mid z)}. \tag{4.3}$$

This map is the composition of the linear smoothing map $\mathrm{O_K}$ and the non-linear Boltzmann-Gibbs transform corresponding to the EM iteration $\mathrm{F_{EM}} : \mathcal{M}^+(\mathbb{X}) \to \mathcal{P}(\mathbb{X})$,

$$\mathrm{F_{EM}}(\eta)(\mathrm{d}x) = \frac{1}{\eta(\bar{G}_\eta)} \bar{G}_\eta(x) \eta(\mathrm{d}x) \tag{4.4}$$

associated with the potential function $\bar{G}_\eta$

$$\bar{G}_\eta(\cdot) := \int_{\mathbb{Y}} \frac{g(y \mid \cdot) h(\mathrm{d}y)}{\int_{\mathbb{X}} \eta(\mathrm{d}z) g(y \mid z)},$$

where the normalising constant $\eta(\bar{G}_\eta) \equiv 1$ is introduced to highlight the connection with the SMC methods we will introduce in Chapter 5; so that $\mathrm{F_{EMS}}\,\eta = \mathrm{O_K}\left(\mathrm{F_{EM}}(\eta)\right) = \left(\mathrm{F_{EM}}\,\eta\right) K$.

In order to prove that the $\mathrm{F_{EMS}}$ admits a fixed point, a number of properties of the EM map, $\mathrm{F_{EM}}$, of the smoothing operator $\mathrm{O_K}$ and of the EMS map itself must be established. To do so, we work with the following set of assumptions:

Assumption 4.1. In addition to Assumption 3.1-(b), the following hold:

(a) $\mathbb{X} \subset \mathbb{R}^{d_{\mathbb{X}}}$ and $\mathbb{Y} \subset \mathbb{R}^{d_{\mathbb{Y}}}$ are compact subsets of Euclidean spaces endowed with the Borel σ-algebras $B(\mathbb{X})$ and $B(\mathbb{Y})$ respectively.

(b) The density of the positive kernel $g(y \mid x)$ is continuous, bounded and bounded

away from 0

$$\exists m_g > 0 \text{ such that } \quad 0 < m_g^{-1} \leq g(y \mid x) \leq m_g < \infty \quad \forall (x, y) \in \mathbb{H} = \mathbb{X} \times \mathbb{Y}.$$

(c) The smoothing operator

$$\mathrm{O_K} : \eta \mapsto \eta K := \int_{\mathbb{X}} \eta(\mathrm{d}v) K(v, \cdot)$$

is such that, for any $A \in B(\mathbb{X})$,

$$K(v, A) = \frac{\int_{A \cap \mathbb{X}} T(u - v) \mathrm{d}u}{\int_{\mathbb{X}} T(u' - v) \mathrm{d}u'}$$

where T is a bounded density in $\mathbb{R}^{d_{\mathbb{X}}}$, such that $\inf_{v \in \mathbb{X}} \int_{\mathbb{X}} T(u - v) \mathrm{d}u > 0$.

It is easy to see that Assumption 4.1-(a) is satisfied by most of the examples in Section 3.2.1. For instance, in image processing both $\mathbb{X}$ and $\mathbb{Y}$ are typically of the form $[-a, a] \times [-b, b]$ for $a, b > 0$, f and h are continuous densities on $\mathbb{X}$ and $\mathbb{Y}$ respectively, and the available data are the values of h over the discretisation of $\mathbb{Y}$ induced by the pixels of the image (e.g. an image with 10×10 pixels induces a discretisation on $\mathbb{Y}$ in which the intervals $[-a, a]$ and $[-b, b]$ are each divided into 10 bins). Similarly, when integral equations arise in connection to PDEs with boundary conditions (e.g. the Dirichlet problem for the Laplace's equation first described by Green (1828)) the domain on which the PDE is defined is bounded.

Assumption 4.1-(b) is a common assumption in the literature on Fredholm integral equations. In particular, if $\mathbb{X}$ is compact and g is continuous, then the integral operator is compact in $\mathbb{L}_2(\mathbb{H})$ (Kress, 2014, Theorem 2.28), ruling out degenerate integral equations which require special treatment (Kress, 2014, Chapter 5). The boundedness condition on g gives compactness of $\{h_f : h_f(y) = \int_{\mathbb{X}} f(x) g(y \mid x) \mathrm{d}x\}$, for f a probability density, in the set of continuous functions endowed with the supremum norm (Mülthei, 1992) and ensures the existence of a minimizer of (3.4) (Mülthei, 1992, Theorem 1).

Under Assumption 4.1-(c), K is a Markov kernel on $\mathbb{X}$ which admits the following density with respect to the Lebesgue measure

$$\frac{dK(v, \cdot)}{d\lambda}(u) = \frac{T(u - v) \mathbb{I}_{\mathbb{X}}(u)}{\int_{\mathbb{X}} T(u' - v) \mathrm{d}u'}. \tag{4.5}$$

Assumption 4.1-(c) on T is mild and is satisfied by most commonly used kernels for

density estimation (Silverman, 1986) and implies that $K(v, \cdot)$ is a density over $\mathbb{X}$ for any fixed v. We can draw samples from $K(v, \cdot)$, e.g. by rejection sampling whenever T is proportional to a density from which sampling is feasible.

Remark 4.1. The choice of the smoothing kernel K is crucial for the EMS iteration. A natural choice is to consider kernels K depending on a parameter ε which controls the level of smoothing; a classic example is Gaussian smoothing, i.e. $T(u - v) = \mathcal{N}(u - v; 0, \varepsilon^2)$ with variance ε^2. As a rule of thumb, ε should not be too small, as arbitrarily small levels of smoothing would make the EMS recursion (4.2) collapse onto the EM one (3.3). On the other hand, ε should not excessively large, as over-smoothing results in flatter reconstructions of f which do not allow to recover e.g. local maxima. A common approach to the selection of regularisation parameters for Fredholm integral equations is cross validation (see, e.g. Amato and Hughes (1991); Wahba (1977)). In the remainder of this chapter, we will consider K to be given as we are mostly interested in the theoretical properties of the continuous EMS recursion.

Remark 4.2. The theoretical results established in this chapter hold regardless of the particular form of K (as long as Assumption 4.1-(c) is satisfied). In the following chapters we will always consider K to be an isotropic Gaussian kernel with marginal variance ε^2; however, non-isotropic kernels might be expected to lead to better results when the smoothness properties of f vary between dimensions (Genton, 2001).

4.2.1 Properties of the EMS map

This section summarises a number of properties of the EM map, F_{EM}, of the smoothing kernel K and of the EMS map, $\mathrm{F}_{\mathrm{EMS}}$, which will be used to establish existence of a fixed point. The results are obtained with respect to the the bounded Lipschitz norm (e.g. Dudley (2002, page 394))

$$\beta(\eta) := \sup_{\|\varphi\|_{BL} \leq 1} \left| \int_{\mathbb{H}} \eta(\mathrm{d}x)\varphi(x) \right|, \tag{4.6}$$

where $\|\cdot\|_{BL}$ denotes the bounded Lipschitz norm for bounded Lipschitz functions φ

$$\|\varphi\|_{BL} := \|\varphi\|_\infty + \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{\|x - y\|_2};$$

this norm metrises weak convergence of measures and we will make often use of the following equivalence (Dudley, 2002, Theorem 11.3.3): for every $\mu \in \mathcal{M}(\mathbb{H})$, and sequence $\{\mu_n\}_{n\geq 1}$ taking values in $\mathcal{M}(\mathbb{H})$, $\beta(\mu_n, \mu) \to 0$ is equivalent to $\mu_n(\varphi) \to \mu(\varphi)$ for all continuous bounded functions $\varphi \in C_b(\mathbb{H})$.

The following properties of $\mathrm{F_{EM}}$ and $\mathrm{O_K}$ lead to compactness of $\mathrm{F_{EMS}}$:

Proposition 4.1. Under Assumption 4.1-(a),(b), the Boltzmann-Gibbs transform $\mathrm{F_{EM}}$ in (4.4) is a continuous and bounded operator on $\mathcal{M}^+(\mathbb{X})$ endowed with the weak topology.

Proof. Let $\eta \in \mathcal{M}^+(\mathbb{X})$ and $\{\eta_n\}_{n\geq 1}$ be a sequence of measures in $\mathcal{M}^+(\mathbb{X})$ converging to η in the weak topology as $n \to \infty$. For any $\varphi \in C_b(\mathbb{X})$ consider

$$
\begin{aligned}
&\left|\int_{\mathbb{X}} \mathrm{F_{EM}}(\eta_n)(\mathrm{d}x)\varphi(x) - \int_{\mathbb{X}} \mathrm{F_{EM}}(\eta)(\mathrm{d}x)\varphi(x)\right| \\
&= \left|\int_{\mathbb{X}} \eta_n(\mathrm{d}x)\varphi(x) \int_{\mathbb{Y}} \frac{g(y|x)h(\mathrm{d}y)}{\eta_n\left(g(y|\cdot)\right)} - \int_{\mathbb{X}} \eta(\mathrm{d}x)\varphi(x) \int_{\mathbb{Y}} \frac{g(y|x)h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\right| \\
&= \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \varphi(x)g(y|x)h(\mathrm{d}y)\left[\frac{\eta_n(\mathrm{d}x)}{\eta_n\left(g(y|\cdot)\right)} - \frac{\eta(\mathrm{d}x)}{\eta\left(g(y|\cdot)\right)}\right]\right| \\
&\leq \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \frac{\eta_n(\mathrm{d}x)\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta_n\left(g(y|\cdot)\right)\eta\left(g(y|\cdot)\right)}\left[\eta\left(g(y|\cdot)\right) - \eta_n\left(g(y|\cdot)\right)\right]\right| \\
&+ \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \left(\eta_n(\mathrm{d}x) - \eta(\mathrm{d}x)\right)\frac{\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\right|,
\end{aligned}
$$

where the second equality follows from Fubini's Theorem since g, φ are bounded functions.

The first term can be bounded by

$$
\begin{aligned}
&\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \frac{\eta_n(\mathrm{d}x)\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta_n\left(g(y|\cdot)\right)\eta\left(g(y|\cdot)\right)}\left[\eta\left(g(y|\cdot)\right) - \eta_n\left(g(y|\cdot)\right)\right]\right| \\
&\leq \|\varphi\|_\infty \int_{\mathbb{Y}} \frac{h(\mathrm{d}y)\int_{\mathbb{X}} \eta_n(\mathrm{d}x)g(y|x)}{\eta_n\left(g(y|\cdot)\right)\eta\left(g(y|\cdot)\right)}\left|\eta\left(g(y|\cdot)\right) - \eta_n\left(g(y|\cdot)\right)\right| \\
&\leq \|\varphi\|_\infty \int_{\mathbb{Y}} \frac{h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\left|\eta\left(g(y|\cdot)\right) - \eta_n\left(g(y|\cdot)\right)\right|.
\end{aligned}
$$

Under Assumption 4.1-(b), g is bounded below by $1/m_g$ and we have

$$
\eta\left(g(y|\cdot)\right) = \int_{\mathbb{X}} \eta(\mathrm{d}x)g(y|x) \geq \frac{1}{m_g}\int_{\mathbb{X}} \eta(\mathrm{d}x) = \frac{1}{m_g}\eta(\mathbb{X}) > 0
$$

since $\eta \in \mathcal{M}^+(\mathbb{X})$ is an unsigned measure with nonzero mass. Therefore we obtain

$$\begin{aligned}\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \frac{\eta_n(\mathrm{d}x)\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta_n\left(g(y|\cdot)\right)\eta\left(g(y|\cdot)\right)}\left[\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right]\right| \\ \|\varphi\|_\infty \int_{\mathbb{Y}} \frac{h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\left|\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right| \\ \leq \|\varphi\|_\infty \frac{m_g}{\eta(\mathbb{X})}\int_{\mathbb{Y}} h(\mathrm{d}y)\left|\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right|\end{aligned}$$

For fixed y, $g(y|\cdot) \in C_b(\mathbb{X})$, and we have that

$$\left|\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right| \to 0$$

as $n \to \infty$ since η_n converges to η in the weak topology. Since g is uniformly bounded by m_g, the Dominated Convergence Theorem then gives

$$\int_{\mathbb{Y}} h(\mathrm{d}y)\left|\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right| \to 0$$

as $n \to \infty$, from which we obtain

$$\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \frac{\eta_n(\mathrm{d}x)\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta_n\left(g(y|\cdot)\right)\eta\left(g(y|\cdot)\right)}\left[\eta\left(g(y|\cdot)\right)-\eta_n\left(g(y|\cdot)\right)\right]\right| \to 0 \tag{4.7}$$

as $n \to \infty$.

For the second term, consider the function

$$x \mapsto \int_{\mathbb{Y}} \frac{\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}. \tag{4.8}$$

This function is bounded by $m_g^2\|\varphi\|_\infty/\eta(\mathbb{X})$; to see that it is also continuous, recall that φ, g are continuous functions while the continuity of $y \mapsto \eta\left(g(y|\cdot)\right)$ follows from the continuity of g and the Dominated Convergence Theorem. The Dominated Convergence theorem and the fact that g is continuous, bounded above and below give continuity of (4.8).

Using Fubini's Theorem, whose applicability is granted by the boundedness of g, φ, we obtain

$$\begin{aligned}&\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} \left(\eta_n(\mathrm{d}x)-\eta(\mathrm{d}x)\right)\frac{\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\right| \\ &\qquad = \left|\int_{\mathbb{X}} \left(\eta_n(\mathrm{d}x)-\eta(\mathrm{d}x)\right)\int_{\mathbb{Y}} \frac{\varphi(x)g(y|x)h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)}\right| \to 0\end{aligned} \tag{4.9}$$

as $n \to \infty$.

Combining (4.7) and (4.9) we obtain convergence of $\mathrm{F}_{\mathrm{EM}}\,\eta_n(\varphi)$ to $\mathrm{F}_{\mathrm{EM}}\,\eta(\varphi)$ for every $\varphi \in C_b(\mathbb{X})$, and thus convergence in the weak topology of $\mathrm{F}_{\mathrm{EM}}\,\eta_n$ to $\mathrm{F}_{\mathrm{EM}}\,\eta$ (Dudley, 2002, Theorem 11.3.3) whenever η_n converges weakly to η, proving that the EM map is continuous in $\mathcal{M}^+(\mathbb{X})$.

Finally, consider boundedness. A non-linear operator is bounded if and only if it maps bounded sets into bounded sets (e.g. Zeidler (1985, page 757)). The EM map maps the space of unsigned measures of nonzero mass $\mathcal{M}^+(\mathbb{X})$ into the space of probability measures $\mathcal{P}(\mathbb{X})$, whose elements have β norm uniformly bounded by 1; in particular F_{EM} maps any bounded subset of $\mathcal{M}^+(\mathbb{X})$ into a uniformly bounded subset of $\mathcal{P}(\mathbb{X})$, showing that F_{EM} is a bounded operator. □

Proposition 4.2. Under Assumption 4.1-(a),(c), the smoothing operator O_{K} is compact on $\mathcal{P}(\mathbb{X})$ endowed with the weak topology.

Proof. To prove that O_{K} is compact we need to prove that it maps bounded subsets into relatively compact subsets (Kress, 2014, Definition 2.17). It is sufficient to observe that $\mathbb{X}$ is a complete subset of $\mathbb{R}^{d_{\mathbb{X}}}$ (as it is a compact subset of a metric space) from which it follows that $\mathcal{P}(\mathbb{X})$ is complete by Prokhorov's Theorem (e.g. Dudley (2002, Corollary 11.5.5)). Completeness of $\mathcal{P}(\mathbb{X})$ implies that every sequence in $\mathcal{P}(\mathbb{X})$ admits a convergence subsequence, using the fact that the β norm metrises weak convergence and the equivalence of compactness and existence of convergent subsequences in metric spaces we obtain that $\mathcal{P}(\mathbb{X})$ is relatively compact (Dudley, 2002, Theorem 11.5.4).

□

Given the two results above, compactness of $\mathrm{F}_{\mathrm{EMS}}$ follows straightforwardly:

Proposition 4.3 (Compactness of $\mathrm{F}_{\mathrm{EMS}}$)**.** Under Assumption 4.1, the EMS map, $\mathrm{F}_{\mathrm{EMS}}$, is compact on $\mathcal{M}^+(\mathbb{X})$ endowed with the weak topology.

Proof. The EMS map is the composition of the continuous and bounded operator F_{EM} (by Proposition 4.1) which maps bounded sets into bounded sets, with the compact smoothing operator O_{K} (by Proposition 4.2) which maps bounded sets into relatively compact sets. It follows that $\mathrm{F}_{\mathrm{EMS}}$ is continuous and maps bounded sets into relatively compact sets, hence $\mathrm{F}_{\mathrm{EMS}}$ is compact (e.g. Zeidler (1985, page 54)). □

4.2.2 Existence of the Fixed Point

Thanks to the properties of the EMS map established in the previous section, we can show that the continuous EMS map admits a fixed point in the set of probability measures:

Corollary 4.1 (Existence of Fixed Point)**.** Under Assumption 4.1, the EMS map $\mathrm{F}_{\mathrm{EMS}}$ admits a fixed point in $\mathcal{P}(\mathbb{X})$ endowed with the weak topology.

Proof. Since $\mathbb{X}$ is a compact metric space (and therefore complete), the set of probability measures $\mathcal{P}(\mathbb{X}) \subset \mathcal{M}(\mathbb{X})$ is complete by Prokhorov's Theorem (e.g. Dudley (2002, Corollary 11.5.5)) and therefore $\mathcal{P}(\mathbb{X})$ is closed. Moreover, $\mathcal{P}(\mathbb{X})$ is non-empty, bounded (since all of its elements have β norm bounded by 1) and convex: take $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and $t \in [0,1]$, then for every $A \in B(\mathbb{X})$

$$t\mu(A) + (1-t)\nu(A) \geq 0 \qquad\qquad t\mu(\mathbb{X}) + (1-t)\nu(\mathbb{X}) = 1$$

as $t, 1-t \geq 0$ and $\mathcal{P}(\mathbb{X}) \subset \mathcal{M}^+(\mathbb{X})$, showing that $t\mu+(1-t)\nu \in \mathcal{P}(\mathbb{X})$ for all $t \in [0,1]$ and all $\mu, \nu \in \mathcal{P}(\mathbb{X})$.

These properties and the compactness of the EMS map (Proposition 4.3) give the existence of a fixed point by Schauder's fixed point theorem (e.g. Zeidler (1985, Theorem 2.A)). □

4.2.3 Towards Uniqueness of the Fixed Point

It is well known that the EM iteration generally does not admit a unique solution (Dempster et al., 1977), the discretised EM for Fredholm integral equations (3.5) is no exception (Latham, 1995, Remark 4). Silverman et al. (1990, Section 5.4) and Latham (1995, Section 5.3) conjecture that the introduction of a smoothing step, leading to EMS, guarantees existence of a unique solution of the EMS iteration.

Uniqueness of the fixed point can be obtained for analytic maps through the following result (Zeidler, 1985, Theorem 14.B):

Theorem 4.1. Suppose that the compact mapping $m : \bar{G} \to \mathbb{X}$ is analytic on the non-empty, open, bounded, convex subset G of the complex Banach space S, and that $m(\partial G) \subseteq G$. Then m has exactly one fixed point on G.

In this section, we establish that the $\mathrm{F}_{\mathrm{EMS}}$ is analytic in $\mathcal{M}^+(\mathbb{X})$. However, we have so far been unable to identify an appropriate open set G containing $\mathcal{P}(\mathbb{X})$ which satisfies the conditions of Theorem 4.1 and over which the continuous EMS

map is defined. Showing that the EMS is analytic over $\mathcal{M}^+(\mathbb{X})$ is a first step towards establishing the uniqueness of the fixed point. We study the uniqueness of the fixed point for the discretised EMS scheme (4.1) in Section 4.3.1.

Before showing that $\mathrm{F}_{\mathrm{EMS}}$ is analytic we prove the following two Lemmata, giving a closed form expression its the Fréchet derivative of order k. Since the computation of the derivatives is rather involved, we only give their expression here and postpone the proofs to Appendix A.

Lemma 4.1 (First order Fréchet derivative)**.** The first order Fréchet derivative of $\mathrm{F}_{\mathrm{EMS}}$ at $\eta \in \mathcal{M}^+(\mathbb{X})$ is the bounded linear operator $D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}} : \mathcal{M}(\mathbb{X}) \to \mathcal{M}(\mathbb{X})$ which maps $\nu \in \mathcal{M}(\mathbb{X})$ to

$$\begin{aligned} D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu(\mathrm{d}x') := &\int_{\mathbb{X}} \nu(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \mathrm{d}x' \\ &- \int_{\mathbb{X}} \eta(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^2} \nu\left(g(y \mid \cdot)\right) \mathrm{d}x'. \end{aligned} \tag{4.10}$$

In particular,

$$\beta\left(D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu\right) \leq 2 \frac{m_g^2}{\eta(\mathbb{X})} \beta(\nu). \tag{4.11}$$

Proof. See Appendix A. □

A similar result for the EM map, F_{EM}, is proved in Mülthei et al. (1987, Lemma 1); the first order Fréchet derivative is derived with respect to $(\mathbb{L}_1, \|\cdot\|_1)$ for positive continuous functions. An analogous result for positive measures is obtained by taking the smoothing operator O_{K} equal to the identity in (4.10).

As a consequence of Lemma 4.1 we can define the derived mapping $D\,\mathrm{F}_{\mathrm{EMS}} : \mathcal{M}^+(\mathbb{X}) \to \mathcal{L}(\mathcal{M}(\mathbb{X}), \mathcal{M}(\mathbb{X}))$ which for each $\eta \in \mathcal{M}^+(\mathbb{X})$ gives the corresponding bounded linear operator $D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}$. In order to show that the EMS map is analytic we need to compute the derivatives of order $k \geq 2$, $D^{(k)} \mathrm{F}_{\mathrm{EMS}}$. These maps are defined inductively: $\mathrm{F}_{\mathrm{EMS}}$ is k times differentiable at $\eta \in \mathcal{M}^+(\mathbb{X})$ if $D^{(k-1)} \mathrm{F}_{\mathrm{EMS}}$ is differentiable at η (Cartan, 1971, page 58). The map $D^{(k)} \mathrm{F}_{\mathrm{EMS}}$ maps each $\eta \in \mathcal{M}^+(\mathbb{X})$ to the k-linear map $D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}$ (Cartan, 1971, page 63). Fortunately, to show that $\mathrm{F}_{\mathrm{EMS}}$ is analytic we only need to compute $D_\eta^{(k)} \nu^k = D_\eta^{(k)} \nu \cdots \nu$ (Zeidler, 1985, page 362):

Lemma 4.2 (Higher order Fréchet derivatives of $\mathrm{F}_{\mathrm{EMS}}$)**.** The Fréchet derivative of

$\mathrm{F}_{\mathrm{EMS}}$ at $\eta \in \mathcal{M}^+(\mathbb{X})$ of order $k \geq 2$ satisfies

$$D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k = (-1)^{k+1} k! \int_{\mathbb{X}} \nu(\mathrm{d}x) K(x, \cdot) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^k} \nu\left(g(y \mid \cdot)\right)^{k-1} \tag{4.12}$$
$$+ (-1)^k k! \int_{\mathbb{X}} \left[\eta + (k-1)\nu\right](\mathrm{d}x) K(x, \cdot) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^k .$$

Moreover, $D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k$ is bounded for all $k \geq 2$

$$\beta\left(D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k\right) \leq (k+1)! m_g^{2k+2} \frac{\beta(\nu)^k}{\eta(\mathbb{X})^{k+1}} \left(\eta(\mathbb{X}) + |\nu(\mathbb{X})|\right). \tag{4.13}$$

Proof. See Appendix A. □

The analyticity of $\mathrm{F}_{\mathrm{EMS}}$ is obtained showing that the Taylor expansion of $\mathrm{F}_{\mathrm{EMS}}$ at each $\eta \in \mathcal{M}^+(\mathbb{X})$ is converging in a neighbourhood of η.

Proposition 4.4 (Analyticity of $\mathrm{F}_{\mathrm{EMS}}$)**.** Under Assumption 4.1, the EMS map, $\mathrm{F}_{\mathrm{EMS}}$, is analytic on $\mathcal{M}^+(\mathbb{X})$ endowed with the weak topology.

Proof. The EMS map $\mathrm{F}_{\mathrm{EMS}}$ is analytic at $\eta \in \mathcal{M}^+(\mathbb{X})$ if the series

$$\mathrm{F}_{\mathrm{EMS}}(\nu - \eta) + \sum_{k=1}^{\infty} \frac{1}{k!} D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}(\nu - \eta)^k \tag{4.14}$$

is converging in a neighbourhood of η, $\{\nu : \beta(\nu, \eta) < r\}$, with $r > 0$ and $\nu \in \mathcal{M}^+(\mathbb{X})$ (Zeidler, 1985, page 362).

To show that the series above is converging we use

$$\mathrm{F}_{\mathrm{EMS}}(\nu - \eta) + \sum_{k=1}^{\infty} \frac{1}{k!} D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}(\nu - \eta)^k$$
$$\leq \beta\left(\mathrm{F}_{\mathrm{EMS}}(\nu - \eta)\right) + \sum_{k=1}^{\infty} \frac{1}{k!} \beta\left(D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}(\nu - \eta)^k\right).$$

The definition of $\mathrm{F}_{\mathrm{EMS}}$ gives $\beta\left(\mathrm{F}_{\mathrm{EMS}}(\nu - \eta)\right) \leq 1$. Then, we bound the derivative of order 1 with (4.11) and that of order k with (4.13)

$$\beta\left(D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}(\nu - \eta)\right) \leq 2 \frac{m_g^2}{\eta(\mathbb{X})} \beta(\nu - \eta)$$
$$\beta\left(D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}(\nu - \eta)^k\right) \leq 2(k+1)! m_g^{2k+2} \frac{\beta(\nu - \eta)^k}{\eta(\mathbb{X})^{k+1}} \left(|\nu(\mathbb{X})| + |\eta(\mathbb{X})|\right).$$

Therefore, the series (4.14) is bounded by

$$\begin{aligned}
&\mathrm{F}_{\mathrm{EMS}}(\nu-\eta) + D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\nu-\eta) + \sum_{k=2}^{\infty}\frac{1}{k!}D_\eta^{(k)}\,\mathrm{F}_{\mathrm{EMS}}(\nu-\eta)^k \\
&\leq \beta\left(\mathrm{F}_{\mathrm{EMS}}(\nu-\eta)\right) + \beta\left(D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\nu-\eta)\right) + \sum_{k=2}^{\infty}\frac{1}{k!}\beta\left(D_\eta^{(k)}\,\mathrm{F}_{\mathrm{EMS}}(\nu-\eta)^k\right) \\
&\leq 1 + 2\frac{m_g^2}{\eta(\mathbb{X})}\beta(\nu-\eta) + 2\frac{m_g^2\left(|\nu(\mathbb{X})| + |\eta(\mathbb{X})|\right)}{\eta(\mathbb{X})}\sum_{k=2}^{\infty}(k+1)\left(\frac{m_g^2\beta(\nu-\eta)}{\beta(\eta)}\right)^k.
\end{aligned}$$

The series on the right-hand side is a power series of the form $\sum_{k=2}^{\infty}(k+1)x^k$, which, by the ratio test, converges when $|x|<1$, thus

$$\frac{m_g^2\beta(\nu-\eta)}{\beta(\eta)} < 1 \qquad \Rightarrow \qquad \beta(\nu-\eta) < \frac{\beta(\eta)}{m_g^2}.$$

Hence, since $\beta(\nu-\eta)=\beta(\nu,\eta)$, for any $r \in \left(0, \beta(\eta)/m_g^2\right)$ and for any ν such that $\beta(\nu,\eta)<r$ the series

$$\mathrm{F}_{\mathrm{EMS}}(\nu-\eta) + \sum_{k=1}^{\infty}\frac{1}{k!}D_\eta^{(k)}\,\mathrm{F}_{\mathrm{EMS}}(\nu-\eta)^k$$

is converging and the $\mathrm{F}_{\mathrm{EMS}}$ is analytic at $\eta \in \mathcal{M}^+(\mathbb{X})$ (Zeidler, 1985, page 362). The same argument applies to all $\eta \in \mathcal{M}^+(\mathbb{X})$, hence $\mathrm{F}_{\mathrm{EMS}}$ is analytic on $\mathcal{M}^+(\mathbb{X})$. □

4.3 Additional Results for the Discretised EMS

After its introduction in Silverman et al. (1990), a number of properties of the discretised EMS iteration

$$f_b^{(n+1)} = \sum_{\kappa=1}^{B}\mathrm{O}_{\mathrm{K}b\kappa}\,f_\kappa^{(n)}\sum_{d=1}^{D}\left(\frac{h_d g_{\kappa d}}{\sum_{k=1}^{B}f_k^{(n)}g_{kd}}\right), \qquad \mathbf{f}^{(n+1)} := (f_1^{(n+1)},\ldots,f_B^{(n+1)})$$

have been established. In the discrete setting, the set of finite measures with nonzero mass coincides with the set of positive vectors with at least one non-zero entry

$$C := \left\{\mathbf{f}\in\mathbb{R}^B : f_b \geq 0,\ b=1,\ldots,B\right\}\setminus\{\mathbf{0}\} \tag{4.15}$$

and the set of probability measures with the set of positive vectors with sum equal to 1, $\mathcal{P} := \mathcal{H} \cap C$ where $\mathcal{H}$ is the hyperplane

$$\mathcal{H} = \left\{ \mathbf{f} \in \mathbb{R}^B : \|\mathbf{h}\|_1^{-1} \sum_{b=1}^{B} f_b = 1 \right\}, \tag{4.16}$$

where $\mathbf{h} := (h_1, \ldots, h_D)$ is the vector obtained by approximating h with its value at the mid-point of each of the D bins of the discretisation of $\mathbb{Y}$ and $\|\mathbf{h}\|_1 := \sum_{d=1}^{D} |h_d|$ (Latham and Anderssen, 1992). Existence of a fixed point for the discrete EMS recursion (4.1) has been proved under mild assumptions on the smoothing matrix $\mathrm{O_K}$ (Latham and Anderssen, 1992).

Nychka (1990) shows that with a particular choice of smoothing matrix, the fixed point of the EMS recursion minimises a penalised likelihood with a particular roughness penalty. With this choice of penalty, the OSL-EM and the EMS recursion have the same fixed point (Green, 1990). The proof of Nychka (1990) does not extend straightforwardly to the continuous case, since the smoothing operator $\mathrm{O_K}$ is not invertible, a key feature of the smoothing matrix considered in Nychka (1990).

Fan et al. (2011) establish convergence of the discrete EMS recursion to local-EM, an expectation maximisation algorithm for maximum local-likelihood estimation, when the smoothing kernel K is a symmetric, positive convolution kernel with compact support and with positive bandwidth ε. Under the same assumptions on the kernel K, Fan et al. (2011) show that the spectral radius of the discrete EMS mapping at the fixed point monotonically decreases below 1 as the bandwidth ε increases, giving local convergence to a fixed point. Moreover, if the space on which the EMS mapping is defined is bounded, the discrete EMS mapping is globally convergent when the bandwidth is sufficiently large. Unfortunately, the result of Fan et al. (2011) does not give an estimate of the order of magnitude of the bandwidths for which convergence occurs, clearly, if $\varepsilon \to \infty$ and the space is bounded, then the EMS recursion will convergence to a Uniform distribution, regardless of h and g (see Remark 4.1). If the discrete EMS iteration (4.1) converges, the convergence occurs at a geometric rate, improving on the notoriously slow convergence of the EM algorithm (Nychka, 1990). Convergence results for the continuous EMS iteration (4.2) are not yet available.

4.3.1 Uniqueness of the Fixed Point

Silverman et al. (1990, Section 5.4) and Latham (1995, Section 5.3) conjecture that the introduction of a smoothing step, leading to EMS, leads to existence of a

unique solution of the discretised EMS iteration.

We showed in Proposition 4.4 that the EMS map is analytic, we can now apply the result above to obtain uniqueness of the fixed point for the discretised EMS (4.1). In the discrete setting, the EMS map is a map from C in (4.15) into the set of positive vectors with sum equal to 1, $\mathcal{P}$; however, the set C on which the EMS is defined is not bounded. To circumvent this issue we restrict the EMS map to an appropriate subset G of C and assume that we are able to initialise the discrete EMS iteration (4.1) at a vector $\mathbf{f}^{(1)} = (f_1^{(1)}, \ldots, f_B^{(1)}) \in G$.

To show that the fixed point of the discrete EMS map is unique, consider the open (as intersection of open sets) set

$$G := \left\{ \mathbf{f} \in \mathbb{R}^B : 1 - \varepsilon < \frac{\|\mathbf{f}\|_1}{\|\mathbf{h}\|_1} < 1 + \varepsilon \right\} \bigcap \left\{ \mathbf{f} \in \mathbb{R}^B : f_b > 0,\ b = 1, \ldots, B \right\}$$

whose closure

$$\bar{G} \subseteq \left\{ \mathbf{f} \in \mathbb{R}^B : 1 - \varepsilon \leq \frac{\|\mathbf{f}\|_1}{\|\mathbf{h}\|_1} \leq 1 + \varepsilon \right\} \bigcap (C \cup \{\mathbf{0}\}) \subset C$$

is a subset of C, over which the EMS map is defined. We can then show the following result using Theorem 4.1

Proposition 4.5. If $\mathrm{O_K}$ is a stochastic matrix with positive entries, and g is as in Assumption 4.1-(b), then the discretised EMS map (4.1) has a unique fixed point in

$$\mathcal{P} \cap \left\{ \mathbf{f} \in \mathbb{R}^B : f_b > 0,\ b = 1, \ldots, B \right\}.$$

Proof. Clearly, G is non-empty, bounded, and convex: take $\mathbf{v}, \mathbf{w}$ in G and $t \in [0, 1]$, then

$$\begin{aligned} \|t\mathbf{v} + (1-t)\mathbf{w}\|_1 &= \sum_{b=1}^{B} |tv_b + (1-t)w_b| \\ &= \sum_{b=1}^{B} tv_b + (1-t)w_b = t\|\mathbf{v}\|_1 + (1-t)\|\mathbf{w}\|_1 \end{aligned}$$

since $t \geq 0$ and $v_b, w_b \geq 0$ for all $b = 1, \ldots, B$. To show that for every $\mathbf{f} \in \bar{G}$,

$\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f} \in G$ consider (4.1). For every $\mathbf{f} \in \bar{G}$, $\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f} \in \mathcal{H}$:

$$\begin{aligned}\|\mathbf{h}\|_1^{-1} \sum_{b=1}^{B} (\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f})_b &= \|\mathbf{h}\|_1^{-1} \sum_{b=1}^{B}\sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa \sum_{d=1}^{D} \left(\frac{h_d g_{\kappa d}}{\sum_{k=1}^{B} f_k g_{kd}} \right) \\ &= \|\mathbf{h}\|_1^{-1} \sum_{\kappa=1}^{B} f_\kappa \sum_{d=1}^{D} \left(\frac{h_d g_{\kappa d}}{\sum_{k=1}^{B} f_k g_{kd}} \right) = \|\mathbf{h}\|_1^{-1} \sum_{d=1}^{D} h_d = 1\end{aligned}$$

and, since under Assumption 4.1-(b) each g_{kd} is bounded below, each element $\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f}$ is strictly positive

$$\begin{aligned}(\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f})_b &= \sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa \sum_{d=1}^{D} \left(\frac{h_d g_{\kappa d}}{\sum_{k=1}^{B} f_k g_{kd}} \right) \\ &\geq \frac{1}{m_g^2} \sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa \sum_{d=1}^{D} \left(\frac{h_d}{\sum_{k=1}^{B} f_k} \right) = \frac{1}{m_g^2} \sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa \frac{\|\mathbf{h}\|_1}{\|\mathbf{f}\|_1} > 0,\end{aligned}$$

where the last inequality follows from the fact that $\sum_{\kappa=1}^{B} \mathrm{K}_{b\kappa}\, f_\kappa > 0$ since K is a positive matrix and at least one of the entries of $\mathbf{f}$ is strictly positive. Therefore $(\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f})_b > 0$ for all $b = 1, \ldots, B$ and

$$\mathrm{F}_{\mathrm{EMS}}\,\mathbf{f} \in \mathcal{P} \cap \left\{ \mathbf{f} \in \mathbb{R}^B : f_b > 0,\ b = 1, \ldots, B \right\} \subset G.$$

Since the EMS map is analytic by Proposition 4.4, the discretised EMS map (4.1) has a unique fixed point by Theorem 4.1. □

As discussed in Section 4.2.3, this result does not straightforwardly extend to the continuous EMS iteration, as the space of probability measures $\mathcal{P}(\mathbb{X})$ is a manifold in the Banach space of finite measures $\mathcal{M}(\mathbb{X})$ which is not contained in an open set on which $\mathrm{F}_{\mathrm{EMS}}$ is defined. However, in our experiments (see Section 5.5.1) we observed that convergence occurs to a unique fixed point regardless of the initial distribution f_1, giving further evidence that the continuous EMS mapping (4.2) indeed admits a unique fixed point. We conjecture that, as in the case of Proposition 4.5, uniqueness of the fixed point is guaranteed under some assumptions on the smoothing kernel K, in particular, it is well known that if the smoothing kernel K is close to a Dirac δ (e.g., if the bandwidth $\varepsilon \to 0$) the EMS iteration collapses onto the EM one, ruling out the existence of a unique fixed point.

4.4 Summary

We introduced an EMS recursion on continuous spaces and studied its theoretical properties showing the existence of a fixed point in the set of probability measures. In the following chapter we will present a way to implement the continuous EMS recursion which provides an adaptive stochastic discretisation of the domain of f, does not require strong assumptions on the regularity of f and can be implemented when only samples from h are available.

5

Particle Implementation of Expectation Maximisation Smoothing

A short version of this chapter is presented in (Crucinio et al., 2021b).

5.1 Introduction

The Expectation Maximisation Smoothing (EMS) algorithm provides an attractive approach to deal with the ill-posedness of Fredholm integral equations of the first kind, unfortunately, the iterative scheme can rarely be implemented analytically and some kind of discretisation is needed. The standard approach in the literature requires discretisation of the domain, $\mathbb{X}$, and the assumption that the signals h,f are piecewise constant, which restricts their applications to low-dimensional scenarios (Li et al., 2020; Silverman et al., 1990). Under this assumption, the support of h, f is divided into bins and the value of the signal in each of those bins is obtained by iterating the discretised version of the EMS scheme (4.1). This scheme has two main drawbacks: the assumption of piecewise constant signals is rarely satisfied in practice and a functional form of the observed signal h is needed for the implementation. If only observations from h are available, a functional representation can be obtained by e.g., kernel density estimation, but this is not fully satisfactory.

In the previous chapter we introduced a continuous version of the EMS iteration, and showed that a fixed point of the iterative scheme exists also in this setting. We propose a stochastic discretisation of the continuous EMS recursion through sequential Monte Carlo (SMC) which does not need the assumption of piecewise constant signals, provides an adaptive stochastic discretisation of the domain and outputs a sample approximation of f through which a smooth approximation can be obtained via a natural kernel density estimation procedure. This approach naturally

deals with observations from h, therefore we assume that

Assumption 5.1. The function g can be evaluated pointwise and a sample $\mathbf{Y}$ from h is available.

The existence of a fixed point of the continuous EMS recursion (Corollary 4.1) is established taking h to be any probability distribution over $\mathbb{Y}$, and shows that a fixed point exists both in the case in which h admits a density and that in which h is the empirical distribution of a sample $\mathbf{Y}$

$$m(\mathbf{Y}) := \frac{1}{|\mathbf{Y}|}\sum_{j=1}^{|\mathbf{Y}|}\delta_{Y_j} \tag{5.1}$$

where $|\mathbf{Y}|$ denotes the number of samples from h available. This setting is the most common in applications, and is the setting we are concerned with.

The construction of this novel SMC scheme is based on the connection between the EMS recursion and Feynman-Kac measure flows, which we make explicit in this chapter. The resulting SMC scheme is not standard, thus we extend some of the results known for standard SMC to this case (e.g. $\mathbb{L}_p$ inequality, strong law of large numbers, almost sure convergence in the weak topology and bias estimate). Finally, we compare the novel method with alternatives using a simulation study and present results for realistic systems, including motion deblurring and reconstruction of cross-section images of the brain from positron emission tomography.

5.2 A Stochastic Interpretation of EMS

The EMS recursion

$$f_{n+1}(x) = \int_{\mathbb{X}} K(x', x) f_n(x') \int_{\mathbb{Y}} \frac{g(y \mid x')h(y)}{\int_{\mathbb{X}} g(y \mid z) f_n(z)\ \mathrm{d}z}\mathrm{d}y\mathrm{d}x',$$

can be modelled as a Feynman-Kac measure flow by considering an extended state space. Denote by $\eta_n(x, y)$ the joint density at $(x, y) \in \mathbb{H}$ defined by $\eta_n(x, y) = f_n(x)h(y)$ so that $f_n(x) = \eta_n|_{\mathbb{X}}(x) = \int_{\mathbb{Y}} \eta_n(x, y)\ \mathrm{d}y$. This density satisfies a recursion similar to that above

$$\eta_{n+1}(x, y) = \int_{\mathbb{X}}\int_{\mathbb{Y}} \eta_n(x', y')K(x', x)h(y)\frac{g(y' \mid x')}{h_n(y')}\ \mathrm{d}y'\ \mathrm{d}x', \tag{5.2}$$

where $h_n(y') := \int g(y' \mid z)f_n(z)\mathrm{d}z$. With a slight abuse of notation, we also denote by η_n the joint density of $(x_{1:n}, y_{1:n}) \in \mathbb{H}^n$ obtained by iterative application of (5.2)

with the integrals removed and consider the corresponding measure (also denoted by η_n).

The following result connects the augmented EMS recursion to the Feynman-Kac measure flows introduced in Chapter 2:

Proposition 5.1. The sequence of measures $\{\eta_n\}_{n\geq 1}$ defined over the product spaces $\mathbb{H}^n$ by (2.2) with

$$M_{n+1}\left((x_n, y_n), \mathrm{d}(x_{n+1}, y_{n+1})\right) = K(x_n, \mathrm{d}x_{n+1})h(\mathrm{d}y_{n+1}) \tag{5.3}$$

and

$$G_n(x_n, y_n) = \frac{g(y_n \mid x_n)}{\int_{\mathbb{X}} \eta_n|_{\mathbb{X}}(\mathrm{d}z)g(y_n \mid z)} \tag{5.4}$$

satisfies, marginally, recursion (5.2). In particular, the marginal distribution over x_n of η_n,

$$\eta_n|_{\mathbb{X}}(A) = \eta_n\left((\mathbb{X}^{n-1} \times A) \times \mathbb{Y}^n\right) = \eta_n\left(A \times \mathbb{Y}\right), \tag{5.5}$$

for all $A \subset \mathbb{X}$, satisfies the EMS recursion if we make the identification $f_n(x) = \eta_n|_{\mathbb{X}}(\mathrm{d}x)$.

Proof. Starting from (2.2) with M_{n+1} and G_n as in (5.3)-(5.4)

$$\eta_{n+1}\left(\mathrm{d}(x_{n+1}, y_{n+1})\right) = \frac{\eta_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) G_n(x_n, y_n)}{\eta_n(G_n)} M_{n+1}\left((x_n, y_n), \mathrm{d}(x_{n+1}, y_{n+1})\right), \tag{5.6}$$

where $\eta_n(G_n) := \int_{\mathbb{H}} \eta_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) G_n(x_n, y_n) = 1$, and integrating out $(x_{1:n}, y_{1:n})$

$$\begin{aligned}
&\eta_{n+1}\left(\mathrm{d}(x_{n+1}, y_{n+1})\right) \\
&\quad = \int_{\mathbb{H}^n} \frac{\eta_n\left(\mathrm{d}(x_{1:n}, \mathrm{d}y_{1:n})\right) G_n(x_n, y_n)}{\eta_n(G_n)} M_{n+1}\left((x_n, y_n), \mathrm{d}(x_{n+1}, y_{n+1})\right) \\
&\quad = \int_{\mathbb{H}} \int_{\mathbb{H}^{n-1}} \Big\{ \eta_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) \\
&\qquad\qquad \times \frac{g(y_n \mid x_n)}{\int \eta_n|_{\mathbb{X}}(\mathrm{d}z)g(y_n|z)} K(x_n, \mathrm{d}x_{n+1})h(\mathrm{d}y_{n+1}) \Big\} \\
&\quad = \int_{\mathbb{H}} \eta_n\left(\mathrm{d}(x_n, y_n)\right) \frac{g(y_n \mid x_n)}{\int \eta_n|_{\mathbb{X}}(\mathrm{d}z)g(y_n|z)} K(x_n, \mathrm{d}x_{n+1})h(\mathrm{d}y_{n+1}).
\end{aligned}$$

The above shows that the marginals of (5.6) at time $n+1$ satisfy the augmented EMS recursion (5.2).

We can then compute the marginal over $\mathbb{X}$, $\eta_{n+1}|_{\mathbb{X}}$

$$\begin{aligned}\eta_{n+1}|_{\mathbb{X}}(\mathrm{d}x_{n+1}) &= \int_{\mathbb{Y}} \eta_{n+1}\left(\mathrm{d}(x_{n+1}, y_{n+1})\right) \\ &= \int_{\mathbb{Y}} h(\mathrm{d}y_{n+1}) \int_{\mathbb{H}} \eta_n\left(\mathrm{d}(x_n, y_n)\right) \frac{g(y_n \mid x_n)}{\int \eta_n|_{\mathbb{X}}(dz) g(y_n|z)} K(x_n, dx_{n+1}) \\ &= \int_{\mathbb{X}} \eta_n|_{\mathbb{X}}(\mathrm{d}x_n) K(x_n, \mathrm{d}x_{n+1}) \int_{\mathbb{Y}} h(\mathrm{d}y_n) \frac{g(y_n \mid x_n)}{\int \eta_n|_{\mathbb{X}}(\mathrm{d}z) g(y_n|z)}\end{aligned}$$

which, with the given identifications, satisfies the EMS recursion (4.2). □

To facilitate the theoretical analysis we separate the contribution of the mutation kernels (5.3) and of the potential functions (5.4), in particular, we denote the weighted distribution obtained from η_n by

$$\Psi_{G_n}(\eta_n)\left(\mathrm{d}(x_n, y_n)\right) := \frac{1}{\eta_n(G_n)} \eta_n\left(\mathrm{d}(x_n, y_n)\right) G_n(x_n, y_n).$$

5.3 SMC Implementation of the EMS Recursion

Having shown that the EMS recursion describes a sequence of densities satisfying a Feynman-Kac flow, it is possible to use SMC techniques to approximate this recursion. This involves replacing the true density at each step with a sample approximation obtained at the previous iteration, giving rise to Algorithm 2, which describes the case in which only a fixed number of samples from h are available and in line 1-2 we draw Y_n^i from their empirical distribution; when sampling freely from h is feasible one could instead draw these samples from it.

Algorithm 2: Sequential Monte Carlo for Expectation Maximisation Smoothing (SMC-EMS)

At time $n = 1$

1 Sample $\widetilde{X}_1^i \sim f_1$, $\widetilde{Y}_1^i$ from $m(\mathbf{Y})$ for $i = 1, \ldots, N$ and set $W_1^i = \frac{1}{N}$

At time $n > 1$

2 Sample $X_n^i \sim K(\widetilde{X}_{n-1}^i, \cdot)$ and Y_n^i from $m(\mathbf{Y})$ for $i = 1, \ldots, N$

3 Compute the approximated potentials $G_n^N(X_n^i, Y_n^i)$ in (5.8) and obtain the normalized weights $W_n^i = G_n^N(X_n^i, Y_n^i) / \sum_{j=1}^N G_n^N(X_n^j, Y_n^j)$ for $i = 1, \ldots, N$

4 (Re)Sample $\left\{(X_n^i, Y_n^i), W_n^i\right\}$ to get $\left\{(\widetilde{X}_n^i, \widetilde{Y}_n^i), \frac{1}{N}\right\}$ for $i = 1, \ldots, N$

5 Estimate $f_{n+1}(x)$ as in (5.10)

The resulting SMC scheme is *not* a standard particle approximation of Feynman-Kac measure flows, because of the definition of the potential (5.4). Indeed, G_n

cannot be computed exactly, because $f_n(z)$ is not known. However, the SMC scheme provides an approximation for f_n at time n. Let us denote by $\eta_n^N|_{\mathbb{X}}$ the particle approximation of the marginal $\eta_n|_{\mathbb{X}} = f_n$ in (5.5)

$$\eta_n^N|_{\mathbb{X}}(A) := \int_{A\times\mathbb{Y}} \eta_n^N\left(\mathrm{d}(x,y)\right)$$

for all $A \subset \mathbb{X}$. We can approximate

$$G_n(x_n, y_n) = \frac{g(y_n \mid x_n)}{h_n(y_n)} = \frac{g(y_n \mid x_n)}{\int_{\mathbb{X}} \eta_n|_{\mathbb{X}}(\mathrm{d}z) g(y_n \mid z)}$$

by using the particle approximation of the denominator $h_n(y_n)$

$$h_n^N(y_n) := \int_{\mathbb{X}} \eta_n^N|_{\mathbb{X}}(\mathrm{d}z) g(y_n \mid z) = \eta_n^N|_{\mathbb{X}}\left(g(y_n \mid \cdot)\right) \tag{5.7}$$

to get the approximate potentials

$$G_n^N(x_n, y_n) := \frac{g(y_n \mid x_n)}{h_n^N(y_n)}. \tag{5.8}$$

The use of G_n^N within the importance weighting corresponds to an additional approximation which is not found in standard SMC algorithms. In particular, (5.8) are biased estimators of the true potentials (5.4). As a consequence, it is not possible to use arguments based on extensions of the state space (as in particle filters using unbiased estimates of the potentials (Del Moral et al., 2006b; Fearnhead et al., 2008; Liu and Chen, 1998)) to provide theoretical guarantees for this SMC scheme. If G_n itself were available then it would be preferable to make use of it; in practice this will never be the case but the idealized algorithm which employs such a strategy is of use for theoretical analysis.

At time $n+1$, we estimate $f_{n+1}(x)$ by computing a kernel density estimate (KDE) of

$$\Psi_{G_n^N}(\eta_n^N)|_{\mathbb{X}}(\mathrm{d}x) = \sum_{i=1}^{N} \frac{G_n^N(X_n^i, Y_n^i)}{\sum_{j=1}^{N} G_n^N(X_n^j, Y_n^j)} \delta_{(X_n^i)}(\mathrm{d}x),$$

where $\Psi_{G_n^N}$ is the Boltzmann-Gibbs transform associated with the approximated potentials G_n^N, and *then* applying the EMS smoothing kernel K. This approach may seem counter-intuitive but the KDE kernel and the EMS kernel are fulfilling different roles. The KDE gives a good smooth approximation of the density asso-

ciated with the EMS recursion at a point in that recursion which we expect to be under-smoothed and is driven by the usual considerations of KDE when obtaining a smooth density approximation from an empirical distribution; going on to apply the EMS smoothing kernel is simply part of the EMS regularisation procedure. One could instead apply kernel density estimation after step 2 of the subsequent iteration of the algorithm but this would simply introduce additional Monte Carlo variance, with the described approach corresponding to a Rao-Blackwellisation of that slightly simpler strategy. Using the kernel of Fredholm equations of the *second* kind to extract smooth approximations of their solution from Monte Carlo samples has also been found empirically to perform well (Doucet et al., 2010). Depending on the intended use of the approximation, the KDE step can be omitted entirely; the empirical distribution provides a good (in the sense of Proposition 5.3) approximation to that given by the EMS recursion but one which does not admit a density.

We consider standard $d_{\mathbb{X}}$-dimensional kernels for KDE with smoothing bandwidth s_N and bounded kernel S (Silverman, 1986), $u \mapsto s_N^{-d_{\mathbb{X}}}|\Sigma|^{-1/2}S\left(\left(s_N^2\Sigma\right)^{-1/2}u\right)$. To account for the dependence between samples, when computing the bandwidth, s_N, instead of N we use the effective sample size (Kong et al., 1994)

$$\mathrm{ESS} = \left(\sum_{i=1}^{N} G_n^N(X_n^i, Y_n^i)^2\right)^{-1}\left(\sum_{j=1}^{N} G_n^N(X_n^j, Y_n^j)\right)^2. \tag{5.9}$$

The resulting estimator,

$$\begin{aligned} f_{n+1}^N(x) = &\sum_{i=1}^{N} \frac{G_n^N(X_n^i, Y_n^i)}{\sum_{j=1}^{N} G_n^N(X_n^j, Y_n^j)} s_N^{-d_{\mathbb{X}}}|\Sigma|^{-1/2} \\ &\times \int_{\mathbb{X}} K(x', x) S\left(\left(s_N^2\Sigma\right)^{-1/2}(X_n^i - x')\right) \mathrm{d}x' \end{aligned} \tag{5.10}$$

satisfies the standard KDE convergence results in $\mathbb{L}_1$ (Devroye and Wagner, 1979) and in $\mathbb{L}_2$ (Silverman, 1986) (see Section 5.4.3).

As the EMS recursion (4.2) aims at finding a fixed point, after a certain number of iterations the approximation of f provided by the SMC scheme stabilises. We could therefore average over approximations obtained at different iterations to get more stable reconstructions. When the storage cost is prohibitive, a thinned set of iterations could be used.

In principle, one could reduce the variance of associated estimators by using a different proposal distribution within Algorithm 2 just as in standard particle

methods (see, e.g., Del Moral (2004, Section 2.4.2), Doucet and Johansen (2011, Section 25.4.1)) but this proved unnecessary in all of the examples which we explored as we obtained good performance with this simple generic scheme (the effective sample size was above 70% in all the examples considered).

5.3.1 Algorithmic Setting

There are a number of algorithmic parameters which must be specified. The initial density, f_1, must be specified but we did not find performance to be sensitive to this choice (see Appendix 5.5.1). We advocate choosing f_1 to be a diffuse distribution with support intended to include that of f because the resampling step allows SMC to more quickly forget overly diffuse initializations than overly concentrated ones. For problems with bounded domains, choosing f_1 to be uniform over $\mathbb{X}$ is a sensible default choice and the one which we use in Section 5.5.

We propose to stop the iteration in Algorithm 2 when the difference between successive approximations, measured through the $\mathbb{L}^2$ norm of the reconstruction of h obtained by convolution of f_n^N with g

$$\hat{h}_n^N(y) := \int_{\mathbb{X}} f_n^N(x) g(y \mid x) \mathrm{d}x,$$

is smaller than the variability due to the Monte Carlo approximation of (4.2)

$$\int_{\mathbb{Y}} \left| \hat{h}_{n+1}^N(y) - \hat{h}_n^N(y) \right|^2 \mathrm{d}y < \mathrm{var}\left(\zeta(f_k^N); k = n+1-m, \ldots n+1 \right), \tag{5.11}$$

where ζ is some function of the estimator f_{n+1}^N and we consider its variance over the last m iterations. The term on the left-hand side is an indicator of whether the EMS recursion (4.2) has reached a fixed point, while the variance takes into account the error introduced by approximating (4.2) through Monte Carlo. For given N there is a point at which further increasing n does not improve the estimate because Monte Carlo variability dominates. We employ this stopping rule in the PET example in Section 5.5.4.

The amount of regularisation introduced by the smoothing step is controlled by the smoothing kernel K. In principle, any density T can be used to specify K as in Assumption 4.1-(c); we opted for isotropic Gaussian kernels since in this case the integral in (5.10) can be computed analytically with an appropriate choice of S. In the case of isotropic Gaussian kernels the amount of smoothing is controlled by the variance ε^2, if the expected smoothness of the fixed point of the EMS recursion (4.2)

is known, ε should be chosen so that (5.10) matches this knowledge. If no information is known on the expected smoothness, the level of smoothing introduced could be picked by cross validation, comparing, e.g., the reconstruction accuracy or smoothness. In addition, one could allow extra flexibility by letting K change at each iteration. For instance, allowing for larger moves in early iterations proves to be beneficial in standard SMC settings to improve stability and ergodicity. Another option is to choose the smoothing parameter adaptively using information on the smoothness of the current estimate.

We end this section by identifying a further degree of freedom which can be exploited to improve performance. A variance reduction can be achieved by averaging over several Y_n^i when computing the approximated potentials G_n^N. At time n, draw M samples Y_n^{ij} for $j = 1, \dots, M$ without replacement for each particle $i = 1, \dots, N$ and compute the approximated potentials by averaging over the M replicates Y_n^{ij}

$$G_n^{N,M}(X_n^i, Y_n^i) = \frac{1}{M} \sum_{j=1}^{M} \frac{g(Y_n^{ij} \mid X_n^i)}{h_n^N(Y_n^{ij})}.$$

This incurs an $O(MN)$ computational cost and can be justified by further extending the state space to $\mathbb{X} \times \mathbb{Y}^m$. Unfortunately, the results on the optimal choice of M obtained for pseudo-marginal methods (e.g. Pitt et al. (2012)) cannot be applied here, as the estimates of G_n given by (5.8) are *not* unbiased. In the examples shown in Section 5.5 we resample without replacement M samples from $\mathbf{Y}$ where M is the smallest between N and the size of $\mathbf{Y}$, but smaller values of M could be considered (see Section 5.5.1). This strategy yields an $O(|\mathbf{Y}|N) = O(N)$ cost for a fixed sample $\mathbf{Y}$ from h, in contrast to the naive strategy of drawing N times with replacement which would yield an $O(N^2)$ cost.

5.3.2 Comparison of EMS and SMC-EMS

The discretised EMS (4.1) and Algorithm 2 both approximate the EMS recursion (4.2). There are two main aspects under which the SMC implementation of EMS is an improvement with respect to the one obtained by brute-force discretisation: the information on h which is needed to run the algorithm and the scaling with the dimensionality of the domain of f.

The discretised EMS (4.1) requires the value of h on each of the D bins of the space discretisation of $\mathbb{Y}$, when we only have a sample $\mathbf{Y}$ from h, as it is the case in most applications (Delaigle, 2008; Goldstein et al., 2009; Gostic et al., 2020; Hall et al., 2005; Marschner, 2020; Miao et al., 2018), h should then be approximated

through a histogram or a kernel density estimator as in the Iterative Bayes algorithm (Ma, 2011). On the contrary, Algorithm 2 does not require this additional approximation and naturally deals with samples from h. In Section 5.5.2 we show on a one dimensional example that the brute-force discretisation (4.1) struggles at recovering the shape of a bimodal distribution while the SMC implementation achieves much better performances in terms of accuracy. In addition, increasing the number of bins for EMS has a milder effect on the accuracy than increasing the number of particles in the SMC implementation.

Similar considerations apply when $\mathbb{X}, \mathbb{Y}$ are higher dimensional (i.e. $d_{\mathbb{X}} \geq 2$). The number of bins B in the EMS recursion (4.1) necessary to achieve reasonable accuracy increases exponentially with $d_{\mathbb{X}}$, resulting in higher runtimes which quickly exceed those needed to run Algorithm 2. On the contrary the convergence rate for SMC remains $N^{-1/2}$, and although the associated constants may grow with $d_{\mathbb{X}}$, its performance is shown empirically to scale better with dimension than EMS in Section 5.5.6.

5.4 Convergence Properties of the SMC Scheme

As the potentials (5.4) cannot be computed exactly but need to be estimated, the convergence results for standard SMC (e.g., Chopin and Papaspiliopoulos (2020); Del Moral (2004, 2013)) do not hold. For simplicity, we only consider multinomial resampling (Gordon et al., 1993). Lower variance resampling schemes can be employed but considerably complicate the theoretical analysis (Douc et al., 2005; Gerber et al., 2019). We extend the $\mathbb{L}_p$ inequality by analysing the contribution of the additional approximation introduced by using G_n^N instead of G_n and then combining the results with existing arguments for standard SMC (Crisan and Doucet, 2000, 2002; Míguez et al., 2013). As a consequence of the $\mathbb{L}_p$ inequality for η_n^N, it is possible to show that the approximated potentials G_n^N converge almost surely to the exact ones G_n (Corollary 5.2).

The strong law of large numbers (SLLN) follows from the $\mathbb{L}_p$ inequality and leads to the almost sure convergence of η_n^N to η_n as $N \to \infty$ in the weak topology for all $n \geq 1$ (Proposition 5.3). This result is particularly interesting, because it shows that the empirical measures obtained with the SMC scheme described in Section 5.3 converge (almost surely in the weak topology) to those obtained by the measure flow (5.6), whose marginal over x satisfies the EMS recursion (4.2).

Additionally, we extend the bias estimate for standard SMC algorithms showing that the bias of $\eta_n^N(\varphi)$ decreases with N at rate $1/N$ (see, e.g., Del Moral (2013,

Proposition 9.5.6), Del Moral et al. (2007)). Finally, we provide theoretical guarantees for the estimator (5.10).

5.4.1 $\mathbb{L}_p$ Inequality, Law of Large Numbers and Bias Estimates

The SLLN is stated in Corollary 5.1. This result follow from the $\mathbb{L}_p$ inequality in Proposition 5.2, the proof of which is given in Appendix B and follows the inductive argument of Crisan and Doucet (2000, 2002); Míguez et al. (2013). Both results are proved for test functions φ in the Banach space of measurable bounded real-valued functions on $\mathbb{H}$, denoted by $\mathcal{B}_b(\mathbb{H})$, endowed with the supremum norm, $\|\varphi\|_\infty = \sup_{u\in\mathbb{H}} |\varphi(u)|$.

As a consequence of Assumption 4.1-(b), the potentials G_n and G_n^N are bounded and bounded away from 0 (see Lemma B.1), a strong mixing condition that is common in the SMC literature and is satisfied in most of the applications which we have considered.

Proposition 5.2 ($\mathbb{L}_p$-inequality)**.** Under Assumptions 4.1 and 5.1, for every time $n \geq 1$ and every $p \geq 1$ there exist finite constants $\widetilde{C}_{p,n}, \widehat{C}_{p,n}$ such that for every measurable bounded function $\varphi \in \mathcal{B}_b(\mathbb{H})$

$$\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n} \frac{\|\varphi\|_\infty}{\sqrt{N}} \tag{5.12}$$

and

$$\mathbb{E}\left[|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n)(\varphi)|^p\right]^{1/p} \leq \widehat{C}_{p,n} \frac{\|\varphi\|_\infty}{\sqrt{N}} \tag{5.13}$$

where the expectations are taken with respect to the law of all random variables generated within the SMC algorithm.

Proof. See Section 5.4.2 and Appendix B. □

The first inequality controls the error between the exact evolution of the measure flow (5.6) at iteration n and the evolution given by the particle population with the approximated potential G_n^N, while the second inequality controls the error before the resampling step at time n. The SLLN is a corollary to the $\mathbb{L}_p$-inequality:

Corollary 5.1 (Strong law of large numbers)**.** Under Assumptions 4.1 and 5.1, for all $n \geq 1$ and for every $\varphi \in \mathcal{B}_b(\mathbb{H})$, we have almost surely as $N \to \infty$:

$$\eta_n^N(\varphi) \to \eta_n(\varphi) \qquad \text{and} \qquad \Psi_{G_n^N}(\eta_n^N)(\varphi) \to \Psi_{G_n}(\eta_n)(\varphi).$$

Proof. For any continuous bounded function φ and $\forall \varepsilon > 0$ consider the probability $\mathbb{P}\left(|\eta_n^N(\varphi) - \eta_n(\varphi)|^p \geq \varepsilon\right)$. By Markov's inequality and Proposition 5.2 with $p > 2$:

$$\begin{aligned}\sum_{N=0}^{\infty} \mathbb{P}\left(|\eta_n^N(\varphi) - \eta_n(\varphi)|^p \geq \varepsilon\right) &\leq \sum_{N=0}^{\infty} \frac{\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]}{\varepsilon} \\ &\leq \sum_{N=0}^{\infty} \frac{\tilde{C}_{p,n}^p \|\varphi\|_\infty^p}{N^{p/2}\varepsilon} < \infty.\end{aligned}$$

Fix $\varepsilon = 1/m$ for some $m \in \mathbb{N}$ and apply the Borel-Cantelli lemma

$$\mathbb{P}\left(\limsup_N \left\{|\eta_n^N(\varphi) - \eta_n(\varphi)|^p \geq \frac{1}{m}\right\}\right) = 0.$$

It follows that $\limsup_N |\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)| < 1/m$ almost surely.

Define $B_m = \{\limsup_N |\eta_n^N(\varphi) - \eta_n(\varphi)| < 1/m\}$ and $B = \bigcap_{m\in\mathbb{N}} B_m$. The previous statement implies $\mathbb{P}(B_m) = 1$ and hence $\mathbb{P}(B) = 1$. This is equivalent to

$$\mathbb{P}\left(\limsup_N |\eta_n^N(\varphi) - \eta_n(\varphi)| < 1/m \ \forall m \in \mathbb{N}\right) = 1,$$

which gives almost sure convergence of $\eta_n^N(\varphi)$ to $\eta_n(\varphi)$. An analogous argument shows almost sure convergence of $\Psi_{G_n^N}(\eta_n^N)(\varphi)$ to $\Psi_{G_n}(\eta_n)(\varphi)$. □

The following properties of the approximated potentials G_n^N are a direct consequence of Proposition 5.2 and Corollary 5.1:

Corollary 5.2 (Properties of the approximated potential)**.** Under Assumptions 4.1 and 5.1, for fixed (x, y) and for all $n \geq 1$, the potentials $G_n^N(x, y)$ and $G_n(x, y)$ satisfy

1. For all $N \geq 1$, $n \geq 1$ and every $p \geq 1$

$$\mathbb{E}\left[|G_n^N(x, y) - G_n(x, y)|^p\right]^{1/p} \leq m_g^4 \widetilde{C}_{p,n} \frac{1}{\sqrt{N}},$$

   with $\widetilde{C}_{p,n}$ finite.

2. $G_n^N(x, y) \to G_n(x, y)$ almost surely as $N \to \infty$.

3. If $\mathbb{H}$ is compact, then $\|G_n^N - G_n\|_\infty \to 0$ almost surely as $N \to \infty$.

Proof. The first and second statements follow straightforwardly from Proposition 5.2 and Corollary 5.1 respectively. The pointwise almost sure convergence can be

strengthened to uniform almost sure convergence exploiting the continuity of g and the compactness of $\mathbb{H}$ as shown in Appendix B. □

Using standard techniques following Dudley (2002, Chapter 11, Theorem 11.4.1) and Berti et al. (2006) and given in detail for the context of interest by Schmon et al. (2021, Theorem 4), the result of Corollary 5.1 can be strengthened to the convergence of the measures in the weak topology:

Proposition 5.3 (Almost sure weak convergence)**.** Under Assumptions 4.1 and 5.1, for all $n \geq 1$, η_n^N converges almost surely in the weak topology to η_n, $\eta_n^N \rightharpoonup \eta_n$ as $N \to \infty$.

Proof. Consider $\mathrm{BL}(\mathbb{H}) \subset \mathcal{B}_b(\mathbb{H})$, the Banach space of bounded Lipschitz functions. As shown in Schmon et al. (2021, Supplementary Material, Proposition 2) and Dudley (2002, Theorem 11.4.1) $\mathrm{BL}(\mathbb{H})$ admits a countable dense subclass $C \subset \mathrm{BL}(\mathbb{H})$.

For every $\varphi \in C$ define $A_\varphi = \{\omega \in \Omega : \eta_n^N(\omega)(\varphi) \to \eta_n(\varphi)\ N \to \infty\}$. Then $\mathbb{P}(A_\varphi) = 1\ \forall \varphi \in C$ by Corollary 5.1 and

$$\mathbb{P}\left(\{\omega \in \Omega : \eta_n^N(\omega)(\varphi) \to \eta_n(\varphi)\ N \to \infty\ \forall \varphi \in C\}\right) = \mathbb{P}\left(\bigcap_{\varphi \in C} A_\varphi\right) = 1.$$

The result follows from the fact that C is dense in $\mathrm{BL}(\mathbb{H})$ and the Portmanteau Theorem (e.g. Billingsley (1995, Theorem 29.1), see also Schmon et al. (2021, Supplementary Material, Theorem 1)). □

The following bias estimates shows that the use of approximated potentials rather than exact ones does not increase the rate of convergence of the bias:

Proposition 5.4 (Bias estimate)**.** Under Assumptions 4.1 and 5.1, for any $N \geq 1$, $n \geq 1$ and any $\varphi \in \mathcal{B}_b(\mathbb{H})$ we have

$$\left|\mathbb{E}\left[\eta_n^N(\varphi)\right] - \eta_n(\varphi)\right| \leq \frac{C_n \|\varphi\|_\infty}{N}$$

and

$$\left|\mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right] - \eta_n(G_n\varphi)\right| \leq \frac{D_n \|\varphi\|_\infty}{N}$$

for some finite C_n and D_n.

Proof. See Section 5.4.2 and Appendix B. □

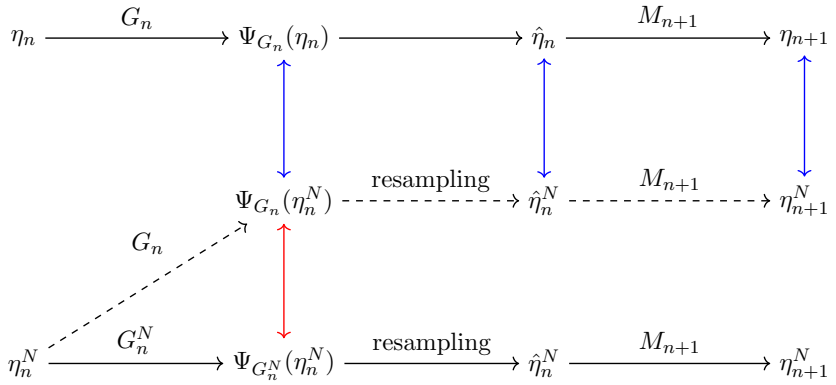


Figure 4: Schematic representation of the non-implementable standard particle approximation of the Feynman-Kac flow (5.6) (dashed) and of the SMC scheme in Algorithm (2) (solid). Standard results in the SMC literature control the approximation errors for the dashed approximations (in blue), we provide a bound on the error of the additional approximation of the weights (in red).

5.4.2 Method of Proof

The theoretical characterisation of the particle method approximating the EMS recursion is carried out by decomposing Algorithm 2 into three steps: mutation, reweighting and resampling. This decomposition is standard in the study of SMC algorithms (Chopin, 2004; Crisan and Doucet, 2000, 2002; Míguez et al., 2013) and allows us to examine the novelty of the particle approximation introduced in Section 5.3 by directly considering the contribution to the overall approximation error of the use of approximate weights G_n^N.

First, consider the following decomposition of the dynamics in (5.6) with potentials (5.4) and Markov kernels (5.3). In the selection step, the current state is weighted according to the potential function G_n

$$\hat{\eta}_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) \equiv \Psi_{G_n}(\eta_n)\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) = \frac{1}{\eta_n(G_n)} G_n(x_n, y_n)\eta_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right);$$

then, in the mutation step, a new state is proposed according to M_{n+1}

$$\eta_{n+1}\left(\mathrm{d}(x_{1:n+1}, y_{1:n+1})\right) \propto \hat{\eta}_n\left(\mathrm{d}(x_{1:n}, y_{1:n})\right) M_{n+1}(x_n, \mathrm{d}x_{n+1}).$$

Each step of the evolution above is compared to its particle approximation counterpart: the weighted distribution $\Psi_{G_n}(\eta_n^N)$ is compared with $\Psi_{G_n}(\eta_n) \equiv \hat{\eta}_n$, the resampled distribution $\hat{\eta}_n^N$ is compared with $\hat{\eta}_n$ and finally η_n^N is compared with η_n (see Figure 4).

As the resampling step is not different from that of standard SMC implementation, we only consider the multinomial resampling scheme described by Gordon et al. (1993). The convergence properties of alternative resampling schemes have been considered in e.g., Gerber et al. (2019).

$\mathbb{L}_p$ Inequality

The proof of the $\mathbb{L}_p$-inequality in Proposition 5.2 follows the inductive approach of Crisan and Doucet (2000, 2002); Míguez et al. (2013) and consists of 4 Lemmata. Lemmata B.4, B.2 and B.3 in Appendix B are due to Crisan and Doucet (2000, 2002); Míguez et al. (2013) and establish $\mathbb{L}_p$-error estimates for the mutation step, the reweighting step performed with the exact potential G_n (exact reweighting) and the multinomial resampling step. Lemma 5.1 compares the exact reweighting with the reweighting obtained by using the approximated potentials G_n^N and is the main element of novelty in the proof:

Lemma 5.1 (Approximate reweighting)**.** Assume that for any $\varphi \in \mathcal{B}_b(\mathbb{H})$ and some finite constants $\widetilde{C}_{p,n}$

$$\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}},$$

for every $p \geq 1$, then

$$\mathbb{E}\left[|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n^N)(\varphi)|^p\right]^{1/p} \leq \ddot{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for any $\varphi \in \mathcal{B}_b(\mathbb{H})$ and for some finite constant $\ddot{C}_{p,n}$.

Proof. Apply the definition of Ψ_{G_n} and $\Psi_{G_n^N}$ and consider the following decomposition

$$\begin{aligned}
|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n^N)(\varphi)| &= \left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n^N)} - \frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)}\right| \\
&\leq \left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n^N)} - \frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n)}\right| \\
&+ \left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n)} - \frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)}\right|.
\end{aligned}$$

Then, for the first term

$$\left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n^N)}-\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n)}\right| = \left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n^N)}\right|\left|\frac{\eta_n^N(G_n)-\eta_n^N(G_n^N)}{\eta_n^N(G_n)}\right| \leq \frac{\|\varphi\|_\infty}{|\eta_n^N(G_n)|}\eta_n^N(|G_n-G_n^N|).$$

For the second term

$$\left|\frac{\eta_n^N(G_n^N\varphi)}{\eta_n^N(G_n)}-\frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)}\right| = \frac{1}{|\eta_n^N(G_n)|}|\eta_n^N(G_n^N\varphi)-\eta_n^N(G_n\varphi)| \leq \frac{\|\varphi\|_\infty}{|\eta_n^N(G_n)|}\eta_n^N(|G_n^N-G_n|).$$

Hence,

$$\begin{aligned}|\Psi_{G_n^N}(\eta_n^N)(\varphi)-\Psi_{G_n}(\eta_n^N)(\varphi)| &\leq 2\frac{\|\varphi\|_\infty}{|\eta_n^N(G_n)|}\eta_n^N(|G_n^N-G_n|)\\ &\leq 2m_g^2\|\varphi\|_\infty\eta_n^N(|G_n^N-G_n|).\end{aligned}$$

By applying Minkowski's inequality and the decomposition of the potentials in Lemma B.1

$$\begin{aligned}&\mathbb{E}\left[\left|\eta_n^N(|G_n^N-G_n|)\right|^p\right]^{1/p}\\ &=\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N\left|G_n^N(X_n^i,Y_n^i)-G_n(X_n^i,Y_n^i)\right|\right|^p\right]^{1/p}\\ &\leq\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[\left|G_n^N(X_n^i,Y_n^i)-G_n(X_n^i,Y_n^i)\right|^p\right]^{1/p}\\ &\leq\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[\left|\frac{G_n^N(X_n^i,Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)}\right|^p|\eta_n|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)-\eta_n^N|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)|^p\right]^{1/p}\\ &\leq\frac{1}{N}\sum_{i=1}^N m_g^3\,\mathbb{E}\left[|\eta_n|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)-\eta_n^N|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)|^p\right]^{1/p}.\end{aligned}$$

Then, consider $\mathcal{S}_n^N := \sigma\left(Y_n^i : i\in\{1,\dots,N\}\right)$, the σ-field generated by all the Y_n^i at time n. By construction, the evolution of X_n^i for $i=1,\dots,N$ is independent of $\mathcal{S}_n^N$ (this is due to the definition of the mutation kernel (5.3)). Conditionally on $\mathcal{S}_n^N$, the Y_n^i are fixed for $i=1,\dots,N$ and we can use the fact that, for fixed y, η_n and $\eta_n|_{\mathbb{X}}$

coincide and so do η_n^N and $\eta_n^N|_{\mathbb{X}}$:

$$\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(y\mid\cdot)\right)-\eta_n^N|_{\mathbb{X}}\left(g(y\mid\cdot)\right)|^p\right]^{1/p}=\mathbb{E}\left[\left|\eta_n\left(g(y\mid\cdot)\right)-\eta_n^N\left(g(y\mid\cdot)\right)|^p\right]^{1/p}\leq\frac{m_g\widetilde{C}_{p,n}}{\sqrt{N}}$$

where the last inequality follows the hypothesis of the Lemma because $g(y\mid\cdot)$ is a bounded and measurable function for all fixed $y\in\mathbb{Y}$.

Hence, since Y_n^i is $\mathcal{S}_n^N$-measurable and independent of $\eta_n^N|_{\mathbb{X}}$, we have

$$\begin{aligned}\mathbb{E}\left[\left|\eta_n^N(|G_n^N-G_n|)\right|^p\right]^{1/p}&\leq m_g^3\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)-\eta_n^N|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)|^p\right]^{1/p}\\&\leq m_g^3\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)-\eta_n^N|_{\mathbb{X}}\left(g(Y_n^i\mid\cdot)\right)|^p\mid\mathcal{S}_n^N\right]\right]^{1/p}\\&\leq\frac{m_g^4\widetilde{C}_{p,n}}{\sqrt{N}}.\end{aligned}$$

Therefore, setting $\ddot{C}_{p,n}=2\widetilde{C}_{p,n}m_g^6$ we obtain

$$\mathbb{E}\left[\left|\Psi_{G_n^N}(\eta_n^N)(\varphi)-\Psi_{G_n}(\eta_n^N)(\varphi)|^p\right]^{1/p}\leq\ddot{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}.$$

□

Bias Estimate

The proof of the bias estimate is given in Appendix B makes use of Lemma 5.2, which controls the bias introduced by the approximated potentials.

Lemma 5.2. Assume that for any $\varphi\in\mathcal{B}_b(\mathbb{H})$ and some finite constant C_n

$$\left|\mathbb{E}\left[\eta_n^N(\varphi)\right]-\eta_n(\varphi)\right|\leq\frac{C_n\|\varphi\|_\infty}{N},\tag{5.14}$$

then for any $\varphi\in\mathcal{B}_b(\mathbb{H})$ and for some finite constant $\widetilde{D}_n$

$$\left|\mathbb{E}\left[\eta_n^N(G_n^N\varphi)-\eta_n^N(G_n\varphi)\right]\right|\leq\frac{\widetilde{D}_n}{N}.$$

Proof. Take

$$\begin{aligned}
\left|\mathbb{E}\left[\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right]\right| &= \left|\mathbb{E}\left[\eta_n^N\left((G_n^N - G_n)\varphi\right)\right]\right| \qquad (5.15)\\
&= \left|\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(G_n^N(X_n^i, Y_n^i) - G_n(X_n^i, Y_n^i)\right)\varphi(X_n^i, Y_n^i)\right]\right|\\
&= \left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\frac{\varphi(X_n^i, Y_n^i)G_n^N(X_n^i, Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right)\right]\right|\\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{E}\left[\frac{\varphi(X_n^i, Y_n^i)G_n^N(X_n^i, Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right)\right]\right|,
\end{aligned}$$

where we have used the decomposition of the potentials in Lemma B.1.

Let us denote by $\mathcal{S}_n^N := \sigma\left(Y_n^i : i \in \{1, \ldots, N\}\right)$, the σ-field generated by Y_n^i for all $i = 1, \ldots, N$, by $\sigma\left(X_n^i\right)$, the σ-field generated by the single particle X_n^i and by $\mathcal{H}_n^{N,i} := \mathcal{S}_n^N \vee \sigma\left(X_n^i\right)$.

By conditioning we can write the expectation above as

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{\varphi(X_n^i, Y_n^i)G_n^N(X_n^i, Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right)\right]\right| \qquad (5.16)\\
&\leq \mathbb{E}\left[\left|\frac{\varphi(X_n^i, Y_n^i)G_n^N(X_n^i, Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}\right|\left|\mathbb{E}\left[(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right|\right]\\
&\leq m_g^3\|\varphi\|_\infty\,\mathbb{E}\left[\left|\mathbb{E}\left[(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right|\right].
\end{aligned}$$

The conditional expectation can be decomposed as follows

$$\begin{aligned}
&\left|\mathbb{E}\left[(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right| \qquad (5.17)\\
&= \left|\mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N} g(Y_n^i \mid X_n^j) \mid \mathcal{H}_n^{N,i}\right] - \eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|\\
&= \left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) \mid \mathcal{H}_n^{N,i}\right] + \frac{1}{N}g(Y_n^i \mid X_n^i) - \eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|\\
&\leq \left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) \mid \mathcal{H}_n^{N,i}\right] - \eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right| + \frac{1}{N}g(Y_n^i \mid X_n^i).
\end{aligned}$$

The triangle inequality gives

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) \mid \mathcal{H}_n^{N,i}\right] - \eta_n|_{\mathbb{X}}\left(M_n\left(g(Y_n^i \mid \cdot)\right)\right)\right| \\
&\leq \left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right| \\
&+ \left|\mathbb{E}\left[\eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right] - \eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|.
\end{aligned}$$

The first term of this bound is equal to

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{1}{N}\sum_{j\neq i} g(Y_n^i \mid X_n^j) - \frac{1}{N}\sum_{j=1}^{N} g(Y_n^i \mid X_n^j) \mid \mathcal{H}_n^{N,i}\right]\right| \\
&= \left|\frac{1}{N} g(Y_n^i \mid X_n^i)\right|.
\end{aligned}$$

The second one is bounded using the hypothesis (5.14) and observing that for every fixed $y \in \mathbb{Y}$

$$\begin{aligned}
\left|\mathbb{E}\left[\eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot)\right)\right] - \eta_n|_{\mathbb{X}}\left(g(y \mid \cdot)\right)\right| &= \left|\mathbb{E}\left[\eta_n^N\left(g(y \mid \cdot)\right)\right] - \eta_n\left(g(y \mid \cdot)\right)\right| \\
&\leq \frac{C_n m_g}{N}
\end{aligned}$$

since $g(y \mid \cdot)$ is a bounded measurable function for all $y \in \mathbb{Y}$.

Therefore the conditional expectation (5.17) is bounded by

$$\begin{aligned}
\left|\mathbb{E}\left[\left(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}}\right)\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right| &\leq \frac{C_n m_g}{N} + \frac{2}{N} g(Y_n^i \mid X_n^i) \\
&\leq \frac{C_n m_g}{N} + \frac{2m_g}{N}.
\end{aligned} \tag{5.18}$$

Hence, using (5.15), (5.16) and (5.18) we have

$$
\begin{aligned}
&\left|\mathbb{E}\left[\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right]\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{E}\left[\frac{\varphi(X_n^i, Y_n^i)G_n^N(X_n^i, Y_n^i)}{\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right)\right]\right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N} m_g^3\|\varphi\|_\infty \mathbb{E}\left[\left|\mathbb{E}\left[(\eta_n^N|_{\mathbb{X}} - \eta_n|_{\mathbb{X}})\left(g(Y_n^i \mid \cdot)\right) \mid \mathcal{H}_n^{N,i}\right]\right|\right] \\
&\leq \frac{1}{N}\sum_{i=1}^{N} m_g^3\|\varphi\|_\infty\left(\frac{C_n m_g}{N} + \frac{2m_g}{N}\right) \\
&\leq m_g^4\|\varphi\|_\infty\frac{(C_n+2)}{N}
\end{aligned}
$$

with $\widetilde{D}_n = m_g^4\|\varphi\|_\infty(C_n+2)$. □

5.4.3 Convergence of Density Estimator

Using a version of the Dominated Convergence Theorem for weakly converging measures (Feinberg et al., 2020; Serfozo, 1982), standard result on kernel density estimation (e.g. Cacoullos (1966); Parzen (1962)) and an argument based on compactness as in Newey (1991) we can establish the following result

Proposition 5.5. Under Assumptions 4.1 and 5.1, if $s_N \to 0$ as $N \to \infty$, the estimator $f_{n+1}^N(x)$ in (5.10) converges uniformly to $f_{n+1}(x)$ with probability 1 for all $n \geq 1$.

Proof. Let us define for $N \in \mathbb{N}$

$$
\varphi^N(t,x) := \int_{\mathbb{X}} K(x',x)s_N^{-d_{\mathbb{X}}}|\Sigma|^{-1/2}S\left((s_N^2\Sigma)^{-1/2}(t-x')\right)\mathrm{d}x',
$$

and note that the estimator (5.10) is given by $f_{n+1}^N(x) = \Psi_{G_n^N}(\eta_n^N)\left(\varphi^N(\cdot,x)\right)$ for any fixed $x \in \mathbb{X}$. Standard results in the literature on kernel density estimation show that $\varphi^N(\cdot,x)$ converges to $K(\cdot,x)$ pointwise for all $x \in \mathbb{X}$ (e.g. Cacoullos (1966, Theorem 2.1)). Because $\mathbb{X}$ is compact, Assumption 4.1-(c) ensures that K is *uniformly* continuous on $\mathbb{X}$ (e.g. Rudin (1964, Theorem 4.19)), then, as argued in Parzen (1962, Theorem 3.A), the sequence $\varphi^N(\cdot,x)$ converges uniformly to $K(\cdot,x)$ in $\mathbb{X}$ (see also Cacoullos (1966, Theorem 3.3)). As a consequence, the sequence $\{\varphi^N(\cdot,x)\}_{N\in\mathbb{N}}$ is uniformly equicontinuous and uniformly bounded (e.g. Rudin (1964, Theorem 7.25)). It follows that $\{\varphi^N(\cdot,x)\}_{N\in\mathbb{N}}$ is (asymptotically) uniformly integrable in the sense of Feinberg et al. (2020, Definition 2.6).

Using an argument analogous to that in Proposition 5.3 we can establish that $\Psi_{G_n^N}(\eta_n^N)$ converges to $\Psi_{G_n}(\eta_n)$ almost surely in the weak topology, then using the fact that the sequence $\{\varphi^N(\cdot, x)\}_{N\in\mathbb{N}}$ is asymptotically uniformly integrable and equicontinuous with continuous limit $K(\cdot, x)$, the Dominated Convergence theorem for weakly converging measures (Feinberg et al. (2020, Corollary 5.2); see also Serfozo (1982, Theorem 3.3)) implies that

$$f_{n+1}^N(x) = \Psi_{G_n^N}(\eta_n^N)\left(\varphi^N(\cdot, x)\right) \to \Psi_{G_n}(\eta_n)\left(K(\cdot, x)\right) = f_{n+1}(x) \tag{5.19}$$

almost surely as $N \to \infty$ for any fixed $x \in \mathbb{X}$.

To turn the result above into almost sure uniform convergence, i.e.

$$\mathbb{P}\left(\limsup_{N\to\infty}\left\{\sup_{x\in\mathbb{X}}|f_{n+1}^N(x) - f_{n+1}(x)| > \varepsilon\right\}\right) = 0$$

for every $\varepsilon > 0$, we exploit Assumption 4.1-(a) and the resulting continuity properties of K.

Under Assumption 4.1, K is uniformly continuous and we have that for any $\varepsilon > 0$ there exists some $\delta_\varepsilon > 0$ such that

$$\begin{aligned}|f_{n+1}(x) - f_{n+1}(x')| &= |\Psi_{G_n}(\eta_n)\left(K(\cdot, x) - K(\cdot, x')\right)| \\ &\leq \sup_{z\in\mathbb{X}}|K(z, x) - K(z, x')| \leq \frac{\varepsilon}{3}\end{aligned}$$

whenever $\|x - x'\|_2 < \delta_\varepsilon$. Using the definition of φ^N and exploiting again the uniform continuity of K we also have that for every $\varepsilon > 0$

$$\begin{aligned}|\varphi^N(t, x) - \varphi^N(t, x')| &\leq \int_{\mathbb{X}} |K(u, x) - K(u, x')| s_N^{-d_{\mathbb{X}}} |\Sigma|^{-1/2} S\left((s_N^2\Sigma)^{-1/2}(t-u)\right) \mathrm{d}u \\ &\leq \frac{\varepsilon}{3}\int_{\mathbb{X}} s_N^{-d_{\mathbb{X}}} |\Sigma|^{-1/2} S\left((s_N^2\Sigma)^{-1/2}(t-u)\right) \mathrm{d}u \leq \frac{\varepsilon}{3}\end{aligned}$$

if $\|x - x'\|_2 < \delta_\varepsilon$. It follows that f_{n+1}^N is uniformly continuous: for any $\varepsilon > 0$

$$\begin{aligned}|f_{n+1}^N(x) - f_{n+1}^N(x')| &= |\Psi_{G_n^N}(\eta_n^N)\left(\varphi^N(\cdot, x) - \varphi^N(\cdot, x')\right)| \\ &\leq \sup_{z\in\mathbb{X}}|\varphi^N(z, x) - \varphi^N(z, x')| \leq \frac{\varepsilon}{3}\end{aligned}$$

whenever $\|x - x'\|_2 < \delta_\varepsilon$.

Let $B(x, \delta_\varepsilon) := \{x' \in \mathbb{X} : \|x - x'\|_2 < \delta_\varepsilon\}$ denote the ball in $\mathbb{X}$ centred around x of radius δ_ε. Under Assumption 4.1-(a), $\mathbb{X}$ is compact and therefore there exists

a finite subcover $\{B(x^j)\}_{j=1}^J$ of $\{B(x,\delta_\varepsilon)\}_{x\in\mathbb{X}}$. Using the uniform continuity above and the following decomposition, we obtain for all $x \in B(x^j)$, $j = 1,\dots,J$ and for all N,

$$\begin{aligned}|f_{n+1}^N(x) - f_{n+1}(x)| &\leq |f_{n+1}^N(x) - f_{n+1}^N(x^j)| + |f_{n+1}^N(x^j) - f_{n+1}(x^j)| \\ &\quad + |f_{n+1}(x^j) - f_{n+1}(x)| \\ &\leq \frac{\varepsilon}{3} + |f_{n+1}^N(x^j) - f_{n+1}(x^j)| + \frac{\varepsilon}{3} \\ &\leq \frac{2}{3}\varepsilon + \max_{j=1,\dots,J} |f_{n+1}^N(x^j) - f_{n+1}(x^j)|,\end{aligned}$$

from which follows

$$\sup_{x\in\mathbb{X}} |f_{n+1}^N(x) - f_{n+1}(x)| \leq \frac{2}{3}\varepsilon + \max_{j=1,\dots,J} |f_{n+1}^N(x^j) - f_{n+1}(x^j)|.$$

Therefore, to obtain almost sure uniform convergence, it is sufficient to show that

$$\mathbb{P}\left(\left\{\omega\in\Omega : \max_{j=1,\dots,J} |f_{n+1}^N(\omega)(x^j) - f_{n+1}(x^j)| \to 0 \ N\to\infty\right\}\right) = 1.$$

Let us define $A_j := \{\omega \in \Omega : f_{n+1}^N(\omega)(x^j) \to f_{n+1}(x^j) \ N \to \infty\}$. As a consequence of (5.19) we have $\mathbb{P}(A_j) = 1$ for all $j = 1,\dots,J$ and

$$\begin{aligned}&\mathbb{P}\left(\left\{\omega\in\Omega : \max_{j=1,\dots,J} |f_{n+1}^N(\omega)(x^j) - f_{n+1}(x^j)| \to 0 \ N\to\infty\right\}\right) \\ &\qquad = \mathbb{P}\left(\bigcap_{j=1,\dots,J} A_j\right) = 1,\end{aligned}$$

which gives the result. □

Proposition 5.5 implies convergence in $\mathbb{L}_1$ and of the MISE for $f_{n+1}^N(x)$:

Corollary 5.3. Under Assumption 4.1 and 5.1, if $s_N \to 0$ as $N \to \infty$, f_{n+1}^N converges almost surely to f_{n+1} in $\mathbb{L}_1$ for every $n \geq 1$:

$$\lim_{N\to\infty} \int_{\mathbb{X}} |f_{n+1}^N(x) - f_{n+1}(x)| \mathrm{d}x \overset{\text{a.s.}}{=} 0; \tag{5.20}$$

and the MISE satisfies

$$\lim_{N\to\infty} \mathrm{MISE}(f_{n+1}^N) \equiv \lim_{N\to\infty} \mathbb{E}\left[\int_{\mathbb{X}} |f_{n+1}^N(x) - f_{n+1}(x)|^2 \, \mathrm{d}x\right] = 0. \tag{5.21}$$

Proof. A direct consequence of Proposition 5.5 is the almost sure pointwise convergence of $f_{n+1}^N(x)$ to $f_{n+1}(x)$. As both $f_{n+1}^N(x)$ and $f_{n+1}(x)$ are probability densities on $\mathbb{X}$, we can extend them to $\mathbb{R}^{d_\mathbb{X}}$ by taking

$$\psi_{n+1}^N(x) := \begin{cases} f_{n+1}^N(x) & x \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \psi_{n+1}(x) := \begin{cases} f_{n+1}(x) & x \in \mathbb{X} \\ 0 & \text{otherwise} \end{cases}$$

respectively. Both $\psi_{n+1}^N(x)$ and $\psi_{n+1}(x)$ are probability densities on $\mathbb{R}^{d_\mathbb{X}}$ and are measurable functions. Moreover, $\psi_{n+1}^N(x)$ converges almost surely to $\psi_{n+1}(x)$ for all $x \in \mathbb{R}^{d_\mathbb{X}}$. Hence, we can apply Glick's extension to Scheffé's Lemma (e.g. Devroye and Wagner (1979)) to obtain

$$\int_{\mathbb{R}^d} |\psi_{n+1}^N(x) - \psi_{n+1}(x)| \, \mathrm{d}x \overset{\text{a.s.}}{\to} 0$$

from which we can conclude

$$\begin{aligned} \int_{\mathbb{X}} |f_{n+1}^N(x) - f_{n+1}(x)| \, \mathrm{d}x &= \int_{\mathbb{X}} |\psi_{n+1}^N(x) - \psi_{n+1}(x)| \, \mathrm{d}x \\ &= \int_{\mathbb{R}^{d_\mathbb{X}}} |\psi_{n+1}^N(x) - \psi_{n+1}(x)| \, \mathrm{d}x \to 0 \end{aligned}$$

almost surely as $N \to \infty$.

Convergence of the MISE is a consequence of Proposition 5.5, Assumption 4.1-(a) and the Dominated Convergence Theorem

$$\mathbb{E}\left[\int_{\mathbb{X}} |f_{n+1}^N(x) - f_{n+1}(x)|^2 \mathrm{d}x\right] \leq |\mathbb{X}| \, \mathbb{E}\left[\|f_{n+1}^N - f_{n+1}\|_\infty^2\right] \mathrm{d}x \to 0$$

as $N \to \infty$, where $|\mathbb{X}| < \infty$ denotes the Lebesgue measure of $\mathbb{X}$. □

5.5 Examples

This section shows the results obtained using the SMC implementation of the recursive formula (4.2) on some common examples. First, we consider a 1-dimensional toy example in which the analytic form of both f and h is known and the EMS recursion (4.2) can be solved analytically. We use this example to get some insights on the convergence speed of the algorithm and the dependence on the smoothing kernel K. Then, we consider a simple density estimation problem and two realistic examples of image restoration problems, motion deblurring and positron emission

tomography (Hohage and Werner, 2016). In the first example, the analytic form of h is known and is used to implement the discretised EM and EMS. IB and SMC are implemented using a fixed sample $\mathbf{Y}$ drawn from h. In the case of motion deblurring, the RL algorithm (i.e. EM for Poisson counts) is implemented by considering the data image as a discretisation of the unknown density h into bins. For the SMC implementation we consider the observed distorted image as the empirical distribution of a sample $\mathbf{Y}$ from h and resample from it at each iteration of line 2 in Algorithm 2. In the positron emission tomography example we use synthetic data obtained from the Shepp-Logan phantom (e.g. Shepp and Vardi (1982)), a simplified imitation of the brain's metabolic activity. We choose to work with synthetic data as a database containing real data is not yet available (Whiteley et al., 2020).

A number of parameters have to be set in order to run EM, EMS, IB and SMC implementation of EMS. For all algorithms we need to specify an initial density f_1 and the number of iterations n. Unless otherwise stated, the number of iterations is $n = 100$ (we observed that convergence to a fixed point occurs in a smaller number of steps for all algorithms; see Section 5.5.1) and the initial distribution f_1 is uniform over $\mathbb{X}$.

For the EMS recursion, we use isotropic Gaussian kernels with marginal variance ε^2. The bandwidth s_N is the plug-in optimal bandwidth for Gaussian distributions where the effective sample size (5.9) is used instead of the sample size N (Silverman, 1986, page 45).

The deterministic discretisation of EM and EMS ((3.5) and (4.1) respectively) is obtained by considering B equally spaced bins for $\mathbb{X}$ and D for $\mathbb{Y}$ and approximating each density with its values at the centre of the bins. The number of bins, as well as the number of particles, N, for SMC varies depending on the example considered. In the first example, the choice of D, B and N is motivated by a comparison of error and runtime. For the image restoration problems, D, B are the number of pixels in each image. The number of particles N is chosen to achieve a good trade-off between reconstruction accuracy and runtime.

For the resampling step in Algorithm 2, we use the adaptive multinomial resampling scheme described in Liu (2008, page 35). At each iteration the effective sample size (5.9) is evaluated and multinomial resampling is performed if ESS $< N/2$. This choice is motivated by the fact that up to adaptivity (which we anticipate could be addressed by the approach of Del Moral et al. (2012)) this is the setting considered in the theoretical analysis of Section 5.4 and we observed only modest improvements when using lower variance resampling schemes (e.g. residual resampling, see Liu (2008)) instead of multinomial resampling.

The accuracy of the reconstructions is measured through the integrated square error

$$\mathrm{ISE}(f) = \int_{\mathbb{X}} \left(f(x) - f_{n+1}^N(x)\right)^2 \, \mathrm{d}x. \tag{5.22}$$

To characterize the roughness of f_{n+1}^N, we evaluate both f_{n+1}^N and f at the bin centers x_c and for each bin center we approximate (with 1,000 replicates) the mean squared error (MSE)

$$\mathrm{MSE}(x_c) = \mathbb{E}\left[\left(f(x_c) - f_{n+1}^N(x_c)\right)^2\right]. \tag{5.23}$$

Although the density estimation example of Section 5.5.2 and the example considered in Section 5.5.1 do not satisfy the Assumption 4.1-(a) which has been assumed for convenience throughout this chapter, or Assumption 4.1-(b) under which our theoretical guarantees hold; we nonetheless observe good results in terms of reconstruction accuracy and smoothness, demonstrating that Assumption 4.1-(b) is not necessary and could be relaxed. We study the influence on the lower bound on g in Section 5.5.5. The subsequent examples *do* satisfy all of our theoretical assumptions.

MATLAB (2018) code to reproduce all examples is available online[1].

5.5.1 Analytically Tractable Example

Here we consider a toy example involving Gaussian densities for which both the EM recursion (3.3) and the EMS recursion (4.2) can be solved at least implicitly. The Fredholm integral equation we consider is

$$\mathcal{N}(y; \mu, \sigma_f^2 + \sigma_g^2) = \int_{\mathbb{X}} \mathcal{N}(x; \mu, \sigma_f^2)\, \mathcal{N}(y; x, \sigma_g^2) \mathrm{d}x, \qquad y \in \mathbb{Y} \tag{5.24}$$

where $\mathbb{X} = \mathbb{Y} = \mathbb{R}$. The initial distribution $f_1(x)$ is $\mathcal{N}(x; \mu, \sigma_{EMS,1}^2)$ for some $\sigma_{EMS,1}^2 > 0$.

The fixed point f_{EMS} of the EMS recursion (4.2) with Gaussian smoothing kernel $K_n(x', x) = \mathcal{N}(x; x', \varepsilon^2)$ is a Gaussian with mean μ and variance σ_{EMS}^2 solving

$$\sigma_{EMS}^6 + \sigma_{EMS}^4(\sigma_g^2 - \sigma_h^2) - 2\sigma_{EMS}^2\varepsilon^2\sigma_g^2 - 2\varepsilon^2\sigma_g^2 = 0. \tag{5.25}$$

[1] Link: `https://github.com/FrancescaCrucinio/smcems`

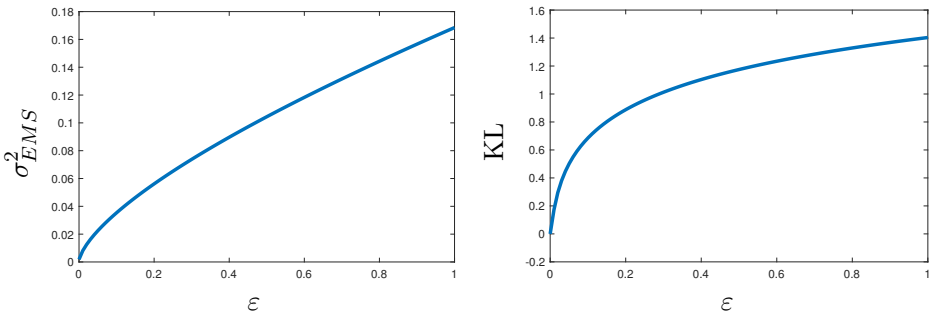


Figure 5: Functional dependence of the variance of the resulting approximation σ^2_{EMS} (left) and the Kullback Leibler divergence (5.26) (right) on the smoothing parameter ε.

We can compute the Kullback–Leibler divergence achieved by f_{EMS}:

$$\mathrm{KL}\left(h, \int_{\mathbb{X}} f_{EMS}(x) g(y \mid \cdot)\mathrm{d}x\right) = \frac{1}{2}\log\frac{\sigma^2_{EMS}+\sigma^2_g}{\sigma^2_h} + \frac{\sigma^2_h}{2(\sigma^2_{EMS}+\sigma^2_g)} - \frac{1}{2}, \quad (5.26)$$

as $\int_{\mathbb{X}} f_{EMS}(x) g(y \mid \cdot)\mathrm{d}x$ is the Gaussian density $\mathcal{N}(y; \mu, \sigma^2_{EMS}+\sigma^2_g)$. The fixed point for the EM recursion (3.3) is obtained setting $\varepsilon = 0$. The corresponding value of the Kullback–Leibler divergence is 0. Figure 5 shows the dependence of σ^2_{EMS} and of the KL divergence on ε.

The conjugacy properties of this model allow us to obtain an exact form for the potential (5.4)

$$G_n(x_n, y_n) = \frac{g(y_n \mid x_n)}{h_n(y_n)} = \frac{\mathcal{N}(y_n; x_n, \sigma^2_g)}{\mathcal{N}(y_n; \mu, \sigma^2_g + \sigma^2_{EMS,n})} \quad (5.27)$$

where $\sigma^2_{EMS,n}$ is the variance of $f_n(x)$.

We use this example to show that the maximum likelihood estimator obtained with the EM iteration (3.5) does not enjoy good properties, and to motivate the addition of a smoothing step in the iterative process (Figure 6). The number of iterations $n = 100$ is fixed for EM, EMS and SMC. The number of bins for both EM and EMS is 100 while the number of particles for SMC is $N = 10^4$ and $\varepsilon = 10^{-2}$. The smoothing matrix for EMS is obtained by discretisation of the smoothing kernel $K_n(x', x) = \mathcal{N}(x; x', \varepsilon^2)$.

We do not include the approximation obtained with IB as this coincides with that of EM. Figure 6 clearly shows that the EM estimate, despite identifying the correct support of the solution, cannot recover the correct shape and is not smooth.

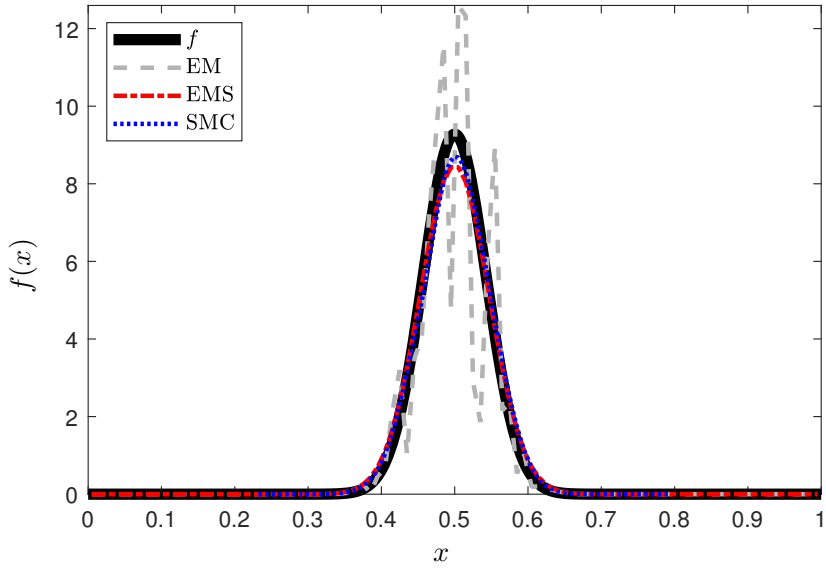


Figure 6: Comparison of EM, EMS and SMC with exact potential G_n for the analytically tractable example.

On the contrary, both EMS and SMC give good reconstruction of f while preserving smoothness.

Then we compare the deterministic discretisation (4.1) of the EMS recursion (4.2) with the stochastic one given by SMC with the exact potential (5.27). To do so, we consider the variance of the obtained reconstructions, their mean integrated square error (5.22), the integrated square error for between h and

$$\hat{h}^N_{n+1}(y) = \int_{\mathbb{X}} f^N_{n+1}(x) g(y \mid x) \mathrm{d}x$$

and the Kullback Leibler divergence $\mathrm{KL}(h, \hat{h}^N_{n+1})$ as the value of the smoothing parameter ε increases (Figure 7). We consider one run of discretised EMS and compare it with 1000 repetitions of SMC for each value of ε (this choice follows from the fact that discretised EMS is a deterministic algorithm). The number of particles for SMC is $N = 10^3$ and for each run we draw a sample $\mathbf{Y}$ of size 10^4 from h and resample from it $M = 10^3$ particles in line 2 of Algorithm 2. Both algorithms correctly identify the mean for every value of ε while the estimated variances increases from that obtained with the EM algorithm ($\varepsilon = 0$) to the variance of a Uniform distribution over $[0, 1]$ (Figure 7 top left). Unsurprisingly, the ISE for both f and h increases with ε (Figure 7 top right and bottom left), showing that an excessive amount of smoothing leads to poor reconstructions. In particular for

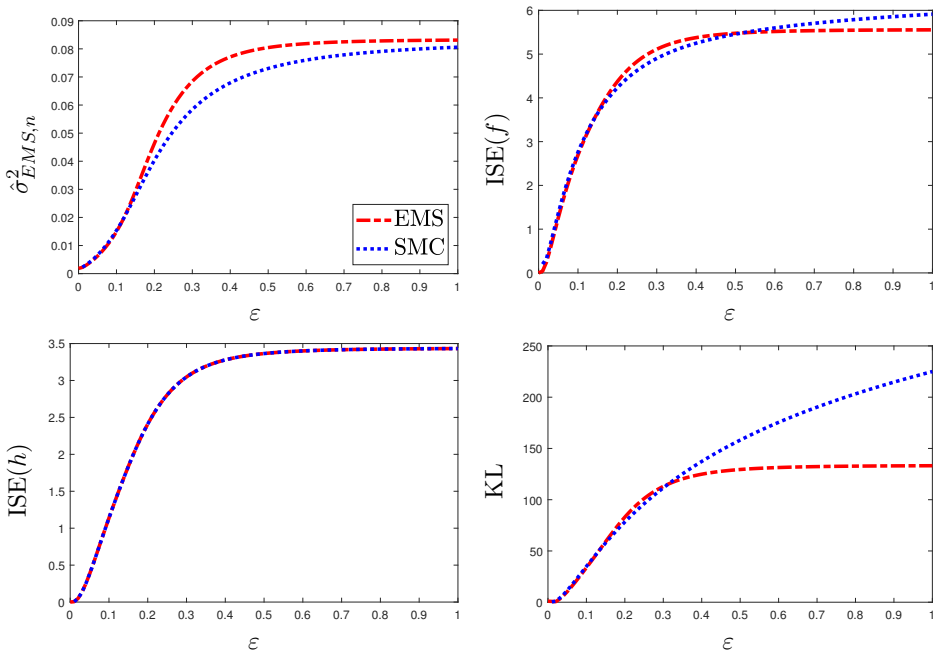


Figure 7: Estimated variance (top left), ISE(f) (top right), ISE(h) (bottom left) and Kullback Leibler divergence (bottom right) as functions of the smoothing parameter ε for the analytically tractable example. The deterministic discretisation (4.1) (red) and the stochastic discretisation via SMC with the exact potentials (5.27) (blue) are compared.

values of $\varepsilon \geq 0.5$ the reconstructions of f become flatter and tend to coincide with a Uniform distribution in the case of EMS and with a normal distribution centred at μ and with high variance (≥ 0.08) in the case of SMC. This difference is reflected in the behaviour of the Kullback Leibler divergence, which stabilises around 133 for EMS while it keeps increasing for SMC (Figure 7 bottom right).

We now consider the effect of the use of the approximated potentials G_n^N in place of the exact ones G_n in the SMC scheme. We compare the ISE for f given by the SMC scheme with exact and approximated potentials for values of the number M of samples Y_n^{ij} drawn from h at each time step between 1 and 1000 with 1000 repetitions for each M. Through this comparison we also address the computational complexity $O(MN)$ of the algorithm, with focus on the choice of the value of M. Figure 8 shows the results for $N = 10^3$ and $\varepsilon = 10^{-2}$. The behaviour for different values of N and ε is similar. The plot of ISE(f) shows a significant improvement when $M > 1$ but little further improvement for $M > 10$.

To further investigate the choice of M we compare the reconstructions obtained using the exact and the approximated potentials for $M = 10$, $M = 10^2$ and

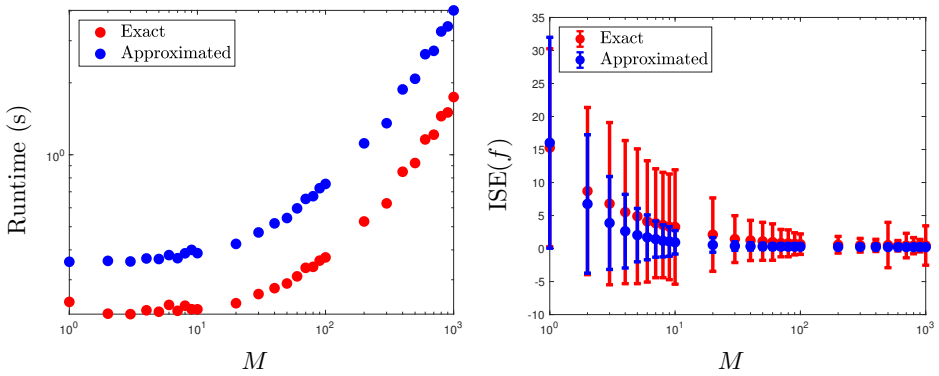


Figure 8: Dependence of runtime and ISE(f) on the value of M, the number of samples drawn from h at each iteration, for the SMC scheme run with the exact potential (blue) and the approximated potential (red). The error bars represent twice the standard deviation of ISE(f).

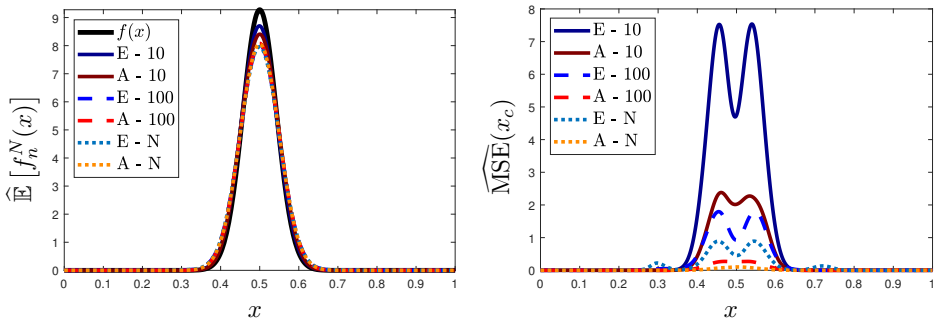


Figure 9: Reconstruction of $f(x) = \mathcal{N}(x; 0.5, 0.043^2)$ from data distribution $h(y) = \mathcal{N}(y; 0.5, 0.043^2 + 0.045^2)$. The number of particles N is 10^3 and the smoothing parameter $\varepsilon = 10^{-3}$. $M = 10$, $M = 100$ and $M = N$ are compared through the pointwise means of the reconstructions and the pointwise mean squared error (MSE).

$M = N = 10^3$. Figure 9 shows pointwise means and pointwise MSE (5.23) for 1,000 reconstructions. The means of the reconstructions with the exact potentials (blue) coincide for the three values of M, the means of the reconstructions with the approximated potentials (red) also coincide but have heavier tails than those obtained with the exact potentials. The MSE is similar for reconstructions with exact and approximated potentials with the same value of M. In particular, the little improvement of the MSE from $M = 10^2$ to $M = 10^3$ suggests that $M = 10^2$ could be used instead of $M = N = 10^3$ if the computational resources are limited. Using $M = 10^2$ instead of $M = 10^3$ reduces the average runtime by $\approx 80\%$ for both the algorithm using the exact potentials and that using the approximated potentials.

Silverman et al. (1990, Section 5.4) conjectured that under suitable assumptions

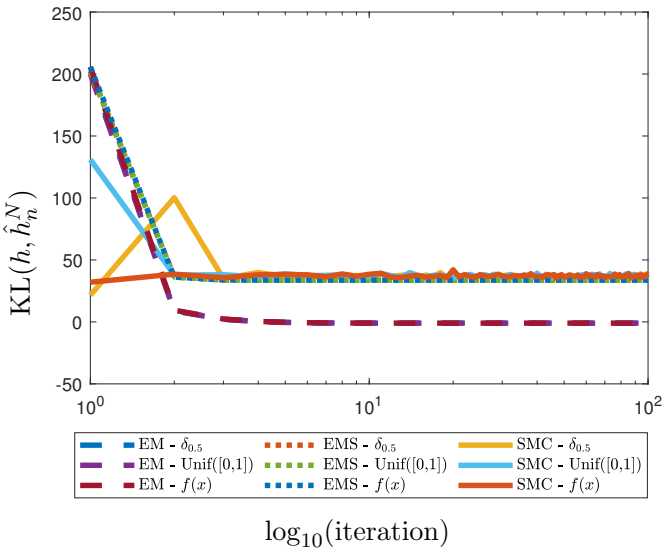


Figure 10: Kullback Leibler divergence between h and $\hat{h}_n^N$ as function of the number of iterations. Three starting distribution are considered: Uniform($[0,1]$), $\delta_{0.5}$, $\mathcal{N}(x;\mu,\sigma_f^2)$. The behaviour of EM (dashed lines), EMS (dotted lines) and SMC (solid lines) is compared.

the EMS map (4.2) has a unique fixed point. This conjecture is empirically confirmed by the results in Figure 10. We run EM, EMS and SMC with approximated potentials for $n = 100$ iterations starting from three initial distributions $f_1(x)$: a Uniform on $[0,1]$, a Dirac δ centred at 0.5 and the solution $\mathcal{N}(x;\mu,\sigma_f^2)$. The number of particles is set to $N = 10^3$ and the smoothing parameter $\varepsilon = 10^{-1}$; this value is chosen for visualisation purposes as for smaller values of ε the discretised EMS gives values of KL divergence which are indistinguishable from those of the EM iteration. Both EMS and SMC converge to the same value of the Kullback Leibler divergence regardless of the starting distribution. The speed of convergence of the three algorithms is similar, in each case little further change is observed once 4 iterations have occurred.

5.5.2 Indirect Density Estimation

The first example is the Gaussian mixture model used in Ma (2011) to compare the Iterative Bayes (IB) algorithm with EM. Take $\mathbb{X} = \mathbb{Y} = \mathbb{R}$ (although note that $|1 - \int_0^1 f(x)\mathrm{d}x| < 10^{-30}$ and restricting out attention to $[0,1]$ would not significantly

alter the results) and

$$
\begin{aligned}
f(x) &= \frac{1}{3}\mathcal{N}(0.3, 0.015^2) + \frac{2}{3}\mathcal{N}(0.5, 0.043^2),\\
g(y \mid x) &= \mathcal{N}(x, 0.045^2),\\
h(y) &= \frac{1}{3}\mathcal{N}(0.3, 0.045^2 + 0.015^2) + \frac{2}{3}\mathcal{N}(0.5, 0.045^2 + 0.043^2).
\end{aligned}
$$

The initial distribution f_1 is Uniform on $[0, 1]$ and the bins for the discretised EMS are B equally spaced intervals in $[0, 1]$, noting that discretisation schemes essentially require known compact support and this interval contains almost all of the probability mass. We run Algorithm 2 assuming that we have a sample $\mathbf{Y}$ of size 10^3 from h from which we re-sample $M = \min(N, 10^3)$ times without replacement at each iteration of line 2.

First, we analyse the influence of the number of bins B and of the number of particles N on the integrated square error of the reconstructions, and on the runtime of the deterministic discretisation of EMS (4.1) and on the SMC implementation of EMS (Figure 11). We compare the two implementations of EMS with a class of estimators specialised for deconvolution problems, deconvolution kernel density estimators with plug-in bandwidth (DKDE-pi; Delaigle and Gijbels (2004)) and cross validated bandwidth (DKDE-cv; Stefanski and Carroll (1990))[2]. These estimators take as input a sample $\mathbf{Y}$ of size N from h and output a kernel density estimator for f.

The discretised EMS has the lowest runtimes for fixed N, however $\mathrm{ISE}(f_{n+1}^N)$ is the highest and finer discretisations for EMS do not significantly improve accuracy. The runtimes of DKDE are closer to those of the SMC implementation, however, the SMC implementation gives better results in terms of $\mathrm{ISE}(f_{n+1}^N)$ for any particle size and, indeed, for given computational cost. We set $\varepsilon = 10^{-3}$, for both EMS and SMC, somewhat arbitrarily, based on the support of the target in this example; where that is not possible cross validation could be used — and might be expected to provide better reconstructions — at the expense of some additional computational cost. We did not find solutions overly sensitive to the precise value of ε (see Appendix 5.5.1). A significant portion of the runtime of DKDE-cv is needed to obtain the bandwidth through cross validation and in this sense the comparison may not be quite fair, but the use of the much cheaper plug-in estimates of bandwidth within DKDE-pi also led to poorer estimates at given cost than those provided by the SMC-EMS

[2]MATLAB code is available on the web page of one of the authors: `https://researchers.ms.unimelb.edu.au/~aurored/links.html#Code`

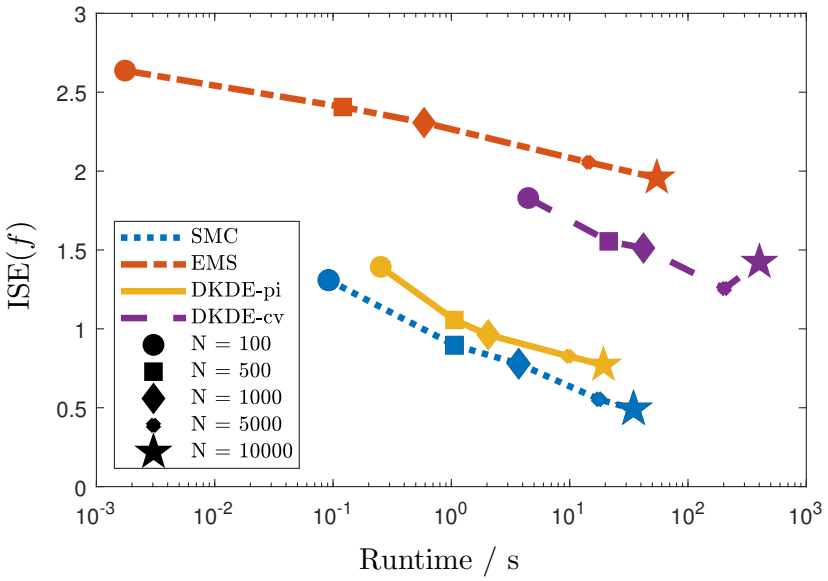


Figure 11: Average ISE(f_{n+1}^N) as function of the runtime for 1,000 repetitions of discretised EMS, SMC and DKDE. The number of bins for EMS and the number of particles/samples N for SMC and DKDE range between 100 and 10,000.

algorithm.

Next, we compare the reconstructions provided by the proposed SMC scheme with those given by deterministic discretisation of the EM iteration (3.5), deterministic discretisation of the EMS iteration (4.1) and deterministic discretisation of the EM iteration when only samples from h are available (IB).

Having observed a small decrease in ISE(f_{n+1}^N) for large B, we fix the number of bins $B = D = 100$. For the SMC scheme, we compare $N = 500$, $N = 1,000$ and $N = 5,000$. We discard $N = 10,000$, as it shows little improvement in ISE(f_{n+1}^N) with respect to $N = 5,000$, and $N = 100$, because of the higher ISE(f_{n+1}^N). We draw a sample $\mathbf{Y}$ from h of size 10^3 and we use this sample to get a kernel density estimator for the IB algorithm, compute the DKDE and (re)sample points at line 2 of Algorithm 2.

We set $\varepsilon = 10^{-3}$ and we compare the smoothing matrix obtained by discretisation of the Gaussian kernel (EMS (K)) with the three-point smoothing proposed in Silverman et al. (1990, Section 3.2.2), where the value $f_b^{(n+1)}$ is obtained by a weighted average over the values $f_\kappa^{(n)}$ of the two nearest neighbors (the third point

	Mean	Variance	ISE(f_{n+1}^N)	MSE(x_c) (95th)	KL	$\log_{10}$ Runtime / s
EM	0.36667	**0.010**	3.26	16.32	**2299**	**-6.01**
EMS (K)	0.36646	0.012	2.41	8.20	2355	-5.90
EMS (3-point)	0.3668	0.011	1.58	13.04	2303	-5.88
IB	**0.43304**	0.011	1.71	10.17	2489	-5.29
SMC (500)	0.43303	0.011	0.90	3.42	2484	0.63
SMC (1000)	0.43302	0.011	0.78	3.33	2483	1.87
SMC (5000)	0.43302	0.011	**0.55**	**2.17**	2485	3.47
DKDE-pi	0.43288	0.012	0.96	3.38	2483	0.81
DKDE-cv	0.43287	0.014	1.50	4.76	2503	4.15

Table 1: Estimates of mean, variance, ISE, 95th-percentile of MSE, KL-divergence and runtime for 1,000 repetitions of EM, EMS, IB, SMC and DKDE with $N = 500, 1,000, 5,000$ for the Gaussian mixture example. The mean of f is 0.43333, the variance is 0.010196. **Bold** indicates best values.

is $f_b^{(n)}$), with weights proportional to the distance $|\kappa - b|$

$$\mathrm{K}_{b\kappa} = 2^{-2}\binom{3-1}{\kappa - b + (3-1)/2}.$$

The reconstruction process is repeated 1,000 times and the reconstructions are compared by computing their means and variances, the integrated squared error (5.22) and the Kullback–Leibler divergence between h and the reconstruction of h obtained by convolution of f_{n+1}^N with g, $\int_{\mathbb{X}} f_{n+1}^N(x) g(y \mid x)\, \mathrm{d}x$, (Table 1). To characterize the roughness of f_{n+1}^N, we use the MSE (5.23). Table 1 shows the 95th percentile w.r.t. the 100 bin centers x_c.

The discretised EM (3.5) gives the best results in terms of Kullback–Leibler divergence (restricting to the $[0, 1]$ interval and computing by numerical integration). This is not surprising, as IB is an approximation of EM when the analytic form of h is not known, and the EMS algorithms (both those with the deterministic discretisation (4.1) and those with the stochastic one given by the SMC scheme) do not seek to minimize the KL divergence, but to provide a more regular solution. However, the solutions recovered by EM have considerably higher ISE than that given by the other algorithms and are considerably worse than the other algorithms at recovering the smoothness of the solution. The reconstructions given by IB are also not smooth, this is not surprising as no smoothing step is present in the IB algorithm.

SMC is generally better at recovering the mean of the solution $\mu = 0.43333$, the global shape of the solution (ISE is at least two times smaller than EM and EMS

(K) and about half than EMS (3-point) and IB) and the smoothness of the solution (the 95th-percentile for MSE(x_c) is at least two times smaller). For the discretised EMS (4.1) and the SMC implementation the estimates of the variance are higher than those of EM, this is a consequence of the addition of the smoothing step and can be controlled by selecting smaller values of ε. DKDE behave similarly to SMC, however their reconstruction accuracy and smoothness are slightly worse than those of SMC (even when both algorithms use the same sample size $N = 1,000$). In particular, DKDE-cv has runtimes of the same order of that of SMC but achieves considerably worse results.

IB, SMC and DKDE give similar values for the KL divergence. The slight increase observed for the SMC scheme with $N = 5,000$ is apparently due to the sensitivity of this divergence to tail behaviors; taking a bandwidth independent of N eliminated this effect.

5.5.3 Motion Deblurring

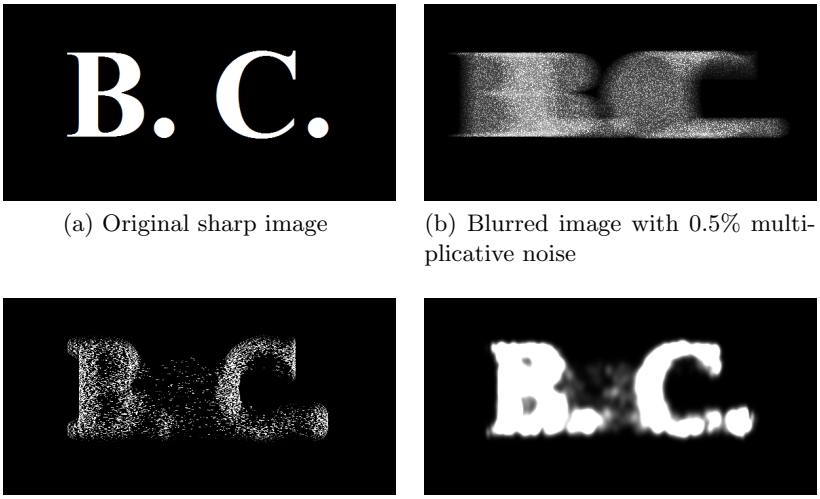
Consider a simple example of motion deblurring where the observed picture h is obtained while the object of interest is moving with constant speed b in the horizontal direction (Lee and Vardi, 1994; Vardi and Lee, 1993). The constant motion in the horizontal direction is modelled by multiplying the density of a uniform random variable on $[-b/2, b/2]$ describing the motion in the horizontal direction and a gaussian $\mathcal{N}(v; y, \sigma^2)$ with small variance (e.g. $\sigma^2 = 0.02^2$), describing the relative lack of motion in the vertical direction

$$g(u, v \mid x, y) = \mathcal{N}(v; y, \sigma^2)\text{Uniform}_{[x-b/2, x+b/2]}(u).$$

We obtain the corrupted image in Figure 12b from the reference image in Figure 12a using the model above with constant speed $b = 128$ pixels and adding multiplicative noise as in Lee and Vardi (1994, Section 6.2). Figure 12b is a noisy discretisation of the unknown $h(u, v)$ on a 300×600 grid. The addition of multiplicative noise makes the model (3.1) misspecified, but still suitable to describe the deconvolution problem when the amount of noise is low. For higher levels of noise, the noise itself should be taken into account when modelling the generation of the data corresponding to h.

We compare the reconstruction obtained using the SMC scheme with that given by the `deconvlucy` function in MATLAB© (The MathWorks Inc., 1993), an efficient implementation of the Richardson-Lucy (RL) algorithm for image processing.

The smoothing parameter is $\varepsilon = 10^{-3}$, and the number of particles is $N = 5000$.



(a) Original sharp image

(b) Blurred image with 0.5% multiplicative noise

(c) Reconstruction with RL

(d) Reconstruction with SMC

Figure 12: Reference image, blurred noisy data distribution and reconstructions for the motion deblurring example. Each scheme used 100 iterations; the SMC scheme used $N = 5000$ particles.

These values are chosen to achieve a trade-off between smoothing and accuracy of the reconstruction and to keep the runtime under 2 minutes on a standard laptop.

The distance between the reconstructions and the original image is evaluated using both the ISE (5.22) and the match distance, i.e. the $\mathbb{L}_1$ norm of the cumulative histogram of the image, a special case of the Earth Mover's Distance for grey-scale images (Rubner et al., 2000). SMC gives visibly smoother images and is better at recovering the shape of the original image (ISE(f) is 1.4617 for SMC and 2.0863 for RL). In contrast, the RL algorithm performs better in terms of match distance (0.0054 for RL and 0.0346 for SMC).

5.5.4 Positron Emission Tomography

Positron Emission Tomography (PET) is a medical diagnosis technique used to analyse internal biological processes from radial projections outside in order to detect medical conditions such as schizophrenia, cancer, Alzheimer's disease and coronary artery disease (Phelps, 2000). The data distribution of the radial projections $h(\phi, \xi)$ is defined on $\mathbb{Y} = [0, 2\pi] \times [-R, R]$ for $R > 0$ and is linked to the cross-section image of the organ of interest $f(x, y)$ defined on the 2D square $\mathbb{X} = [-r, r]^2$ for $r > 0$

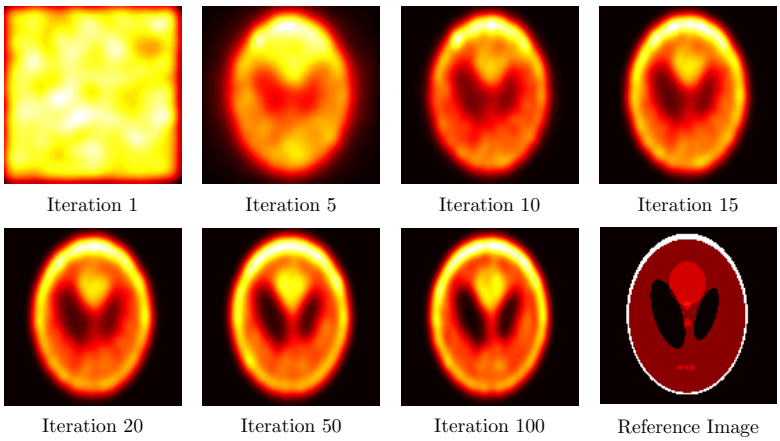


Figure 13: Reconstruction of the Shepp-Logan phantom. The number of particles N is 20,000, the smoothing parameter ε is 0.001. The stopping criterion (5.11) with $m = 15$ is satisfied at iteration 15.

through the Radon transform (Radon, 1986)

$$h(\phi, \xi) = \int_{-\infty}^{+\infty} f(\xi \cos\phi - t \sin\phi, \xi \sin\phi + t \cos\phi) \mathrm{d}t, \tag{5.28}$$

where the right hand side is the line integral along the line with equation $x \cos\phi + y \sin\phi = \xi$. We rewrite (5.28) as a Fredholm integral equation (3.1) modelling the alignment between the projections onto (ϕ, ξ) and the corresponding location (x, y) in the reference image using a Gaussian distribution with small variance

$$h(\phi, \xi) = \int_{\mathbb{X}} \mathcal{N}\left(x \cos\phi + y \sin\phi - \xi; 0, \sigma^2\right) f(x, y) \mathrm{d}x \mathrm{d}y.$$

The kernel $g(\phi, \xi \mid x, y) = \mathcal{N}\left(x \cos\phi + y \sin\phi - \xi; 0, \sigma^2\right)$ satisfies Assumption 4.1-(b) but is not a Markov kernel (in the sense that it does not integrate to 1 for fixed (x, y)), however, we can use the re-normalisation described in Section 3.2 to obtain the Markov kernel

$$\tilde{g}(\phi, \xi \mid x, y) = \frac{g(\phi, \xi \mid x, y)}{C(x, y, \sigma^2)}$$

where $C(x, y, \sigma^2)$ is the normalising constant for each fixed $(x, y) \in \mathbb{X}$

$$C(x, y, \sigma^2) = \int_{\mathbb{Y}} g(\phi, \xi \mid x, y) \mathrm{d}\phi \, \mathrm{d}\xi$$
$$= \int_0^{2\pi} \frac{1}{2} \left[\operatorname{erf}\left(\frac{R - x\cos\phi - y\sin\phi}{\sqrt{2}\sigma} \right) + \operatorname{erf}\left(\frac{R + x\cos\phi + y\sin\phi}{\sqrt{2}\sigma} \right) \right] \mathrm{d}\phi$$

with erf the error function.

The data used in this work are obtained from the reference image in the final panel of Figure 13, a simplified imitation of the brain's metabolic activity (e.g. Vardi and Lee (1993)). The collected data are the values of h at 128 evenly spaced projections over 360° and 185 values of ξ in $[-92, 92]$ to which Poisson noise is added. In this case, $R = 92$ and $(x, y) \in [-64, 64]^2$ (i.e. we want to reconstruct a 128×128 pixels image) and selecting $\sigma = 0.02$ gives

$$\left| \frac{1}{2} \left[\operatorname{erf}\left(\frac{R - x\cos\phi - y\sin\phi}{\sqrt{2}\sigma} \right) + \operatorname{erf}\left(\frac{R + x\cos\phi + y\sin\phi}{\sqrt{2}\sigma} \right) \right] - 1 \right| < 10^{-17}$$

for all $\phi \in [0, 2\pi]$ and $(x, y) \in [-64, 64]$. The above shows that, for σ^2 sufficiently small (e.g. $\sigma^2 = 0.02^2$ as we use in our experiments), i.e. if the Gaussian distribution appropriately describes the alignment onto $x\cos\phi + y\sin\phi = \xi$,

$$\left| C(x, y, \sigma^2) - 2\pi \right| < 10^{-17}$$

for all $(x, y) \in \mathbb{X}$. Therefore we obtain, up to a negligible approximation, an integral equation satisfying Assumption 4.1 dividing h by 2π:

$$\frac{h(\phi, \xi)}{2\pi} = \int_{\mathbb{X}} \frac{\mathcal{N}\left(x\cos(\phi) + y\sin(\phi) - \xi; 0, \sigma^2\right)}{2\pi} f(x, y) \mathrm{d}x \mathrm{d}y.$$

Figure 13 shows the reconstructions obtained with the SMC scheme with smoothing parameter $\varepsilon = 10^{-3}$ and number of particles $N = 20,000$. Convergence to a fixed point occurs empirically in less than 100 iterations, in fact the criterion (5.11) with $\zeta(f_n^N) = \int_{\mathbb{X}} |f_n^N(x)|^2 \mathrm{d}x$ and $m = 15$ stops the iteration at $n = 15$. The ISE between the original image and the reconstructions at Iteration 50 to 100 stabilises below 0.08. Figure 14 shows relative error and ISE for the reconstructions in Figure 13.

The stopping criterion (5.11) is a trade-off between Monte Carlo error and convergence to a fixed point. In particular, when ζ is the variance of the reconstructions, larger values of N will make the r.h.s. of (5.11) which corresponds to a smaller tolerance to assess the convergence to the fixed point. On the other hand, small values

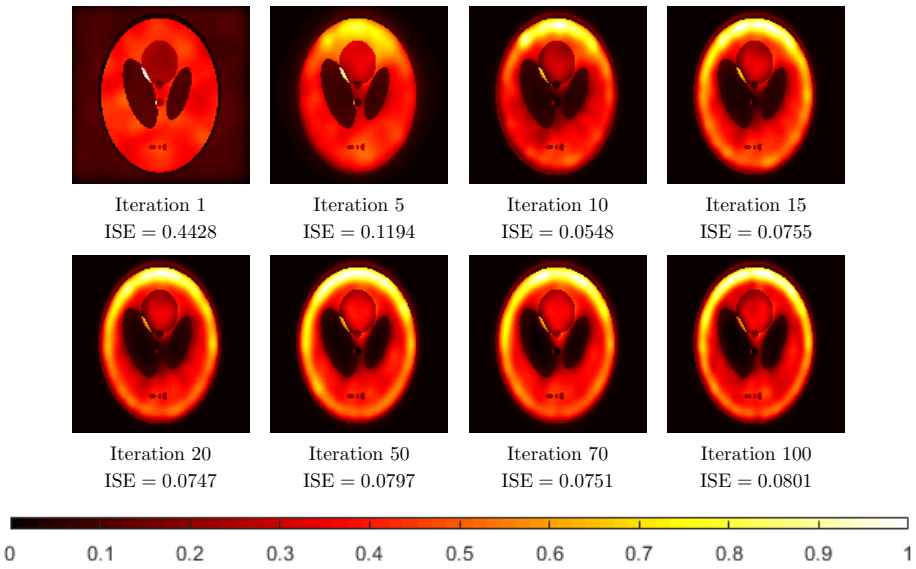


Figure 14: Relative error for the reconstructions in Figure 13. The ISE at each iteration is given in the captions and stabilises below 0.08.

of N will give poorer reconstructions and might require more iterations n to satisfy the stopping criterion (5.11). For instance, for $N = 1,000$ the stopping criterion is not satisfied in 100 iterations despite the r.h.s. of (5.11) being of order 10^{-3} against the $5 \cdot 10^{-5}$ order of the r.h.s. of (5.11) when $N = 20,000$.

The results above show that the SMC implementation of the EMS recursion achieves convergence in a small number of steps ($\approx$ 12 minutes on a standard laptop) and that, contrary to EM (Silverman et al., 1990, Section 4.2), these reconstructions are smooth and do not deteriorate with the number of iterations. In addition, contrary to standard reconstruction methods, e.g. filtered back-projection, ordered-subset EM, Tikhonov regularisation (see, e.g., Tong et al. (2010)) the SMC implementation does not require that a discretisation grid is fixed in advance.

5.5.5 Influence of the Lower Bound on g

Assumption 4.1-(b) and in particular the lower bound on g might seem very restrictive. However, considering lower bounded gs is common in the literature on Fredholm integral equations and is often assumed to establish theoretical results on the EM iteration (Clason et al., 2020; Resmerita et al., 2007) We explore here the influence of this assumption for the Gaussian mixture example in Section 5.5.2 in which g is not lower bounded on $\mathbb{X} = \mathbb{Y} = \mathbb{R}$.

Instead of defining the integrals on $\mathbb{X} = \mathbb{Y} = \mathbb{R}$ take $\mathbb{X} = \mathbb{Y} = [0.4 - a, 0.4 + a]$ with $a \to \infty$ so that we obtain the integral equation

$$\tilde{h}(y) = \int \tilde{f}(x)\tilde{g}(y \mid x)\mathrm{d}x$$

with

$$\begin{aligned}
\tilde{h}(y) &= \frac{h(y)}{\frac{1}{3}C(a, 0.3, 0.045^2 + 0.015^2) + \frac{2}{3}C(a, 0.5, 0.045^2 + 0.043^2)} \\
\tilde{g}(y \mid x) &= \frac{g(y \mid x)}{C(a, x, 0.045^2)} \\
\tilde{f}(x) &= \frac{f(x)C(a, x, 0.045^2)}{\frac{1}{3}C(a, 0.3, 0.045^2 + 0.015^2) + \frac{2}{3}C(a, 0.5, 0.045^2 + 0.043^2)}
\end{aligned}$$

where

$$\begin{aligned}
C(a, \mu, \sigma) &:= \int_{0.4-a}^{0.4+a} \mathcal{N}(x; \mu, \sigma^2)\mathrm{d}x \\
&= \frac{1}{2}\left(\mathrm{erf}\left(\frac{(a + 0.4 - \mu)}{\sqrt{2\sigma^2}}\right) - \mathrm{erf}\left(\frac{(0.4 - a - \mu)}{\sqrt{2\sigma^2}}\right)\right).
\end{aligned}$$

In any of the intervals $[0.4-a, 0.4+a]$ Assumption 4.1-(b) is satisfied, in particular $\tilde{g}$ is bounded below. We study the behaviour of the reconstructions as $a \to \infty$ to check the influence of the lower bound on g on the accuracy of the reconstructions measured through the average ISE in (5.22) over 100 repetitions.The algorithmic set up is the same of Section 5.5.2. Figure 15 show that for $a \in [0.2, 1]$ (which corresponds to $\mathbb{X} = \mathbb{Y} = [0.2, 0.6]$ to $\mathbb{X} = \mathbb{Y} = [-0.6, 1.4]$) the average reconstruction error is not influenced by the lower bound on g, the behaviour for larger values of a is equivalent since $|1 - \int_{-0.6}^{1.4} f(x)\mathrm{d}x| < 10^{-30}$.

5.5.6 Scaling with Dimension

To explore the scaling with the dimension $d_{\mathbb{X}}$ of the domain of f of the discretised EMS (4.1) and the SMC implementation of EMS we revisit the Gaussian mixture model in Section 5.5.2 and extend it to higher dimension

$$\begin{aligned}
f(x) =& \frac{1}{3}\mathcal{N}(x; 0.3 \cdot \mathbf{1}_{d_{\mathbb{X}}}, 0.07^2 I_{d_{\mathbb{X}}}) + \frac{2}{3}\mathcal{N}(x; 0.7 \cdot \mathbf{1}_{d_{\mathbb{X}}}, 0.1^2 I_{d_{\mathbb{X}}}), \\
g(y \mid x) =& \mathcal{N}(y; x, 0.15^2 I_{d_{\mathbb{X}}}), \\
h(y) =& \frac{1}{3}\mathcal{N}(y; 0.3 \cdot \mathbf{1}_{d_{\mathbb{X}}}, (0.07^2 + 0.15^2) I_{d_{\mathbb{X}}}) + \frac{2}{3}\mathcal{N}(y; 0.7 \cdot \mathbf{1}_{d_{\mathbb{X}}}, (0.1^2 + 0.15^2) I_{d_{\mathbb{X}}}),
\end{aligned}$$

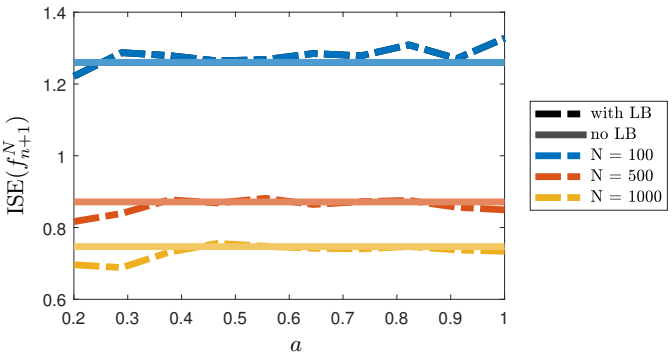


Figure 15: Influence of the lower bound (LB) on g on average reconstruction accuracy over 100 repetitions. The solid lines represent the average $\mathrm{ISE}(f_{n+1}^N)$ for $a = \infty$ while the dashed lines the average $\mathrm{ISE}(f_{n+1}^N)$ for finite a.

where $\mathbb{X} = \mathbb{Y} = \mathbb{R}^{d_{\mathbb{X}}}$ and $\mathbf{1}_{d_{\mathbb{X}}}$, $I_{d_{\mathbb{X}}}$ denote the unit function in $\mathbb{R}^{d_{\mathbb{X}}}$ and the $d_{\mathbb{X}} \times d_{\mathbb{X}}$ identity matrix, respectively. In particular, note that for $d_{\mathbb{X}}$ up to 5 at least 97% of the mass of f is contained in $[0,1]^{d_{\mathbb{X}}}$.

As a first comparison we take $d_{\mathbb{X}} = 2$ and investigate the minimum number of bins/particles necessary to achieve reasonably good reconstructions. We consider three particle sizes $N = 10^2, 50^2, 100^2$ and set the total number of bins $B \approx N$ so that we obtain $\approx N^{1/2}$ equally spaced bins for each dimension. We stop iterating after 30 steps since we observed that convergence occurs within 30 iterations, the value of $\varepsilon = 10^{-3}$ is fixed and used for both the smoothing kernel and the smoothing matrix. The initial distribution is a uniform over $[0,1]^2$ and we assume we have a sample $\mathbf{Y}$ of size 10^6 from h, so that $M = N$. This corresponds to the highest computational cost for Algorithm 2 but as observed in Section 5.5.1 smaller values of M could be considered and would reduce the computational cost of running the SMC implementation of EMS. For small values of B the runtimes of EMS and SMC are similar, however the reconstructions obtained with EMS are poor and low resolution due to the very coarse discretisation (Figure 16-left panel); on the contrary, the presence the kernel density estimator (5.10) guarantees smooth reconstructions even when the particle size is small (Figure 16-middle panel). In addition, as the number of particles N increases the accuracy of the reconstructions provided by SMC keeps increasing, while the EMS reconstructions do not improve as quickly, a phenomenon we already observed for the one-dimensional example in Section 5.5.2. For $N = B \geq 100$ the runtime of SMC is roughly 30% less than that of EMS with the accuracy of SMC being always larger than that of EMS.

Since the accuracy of kernel density estimators decreases when the dimension

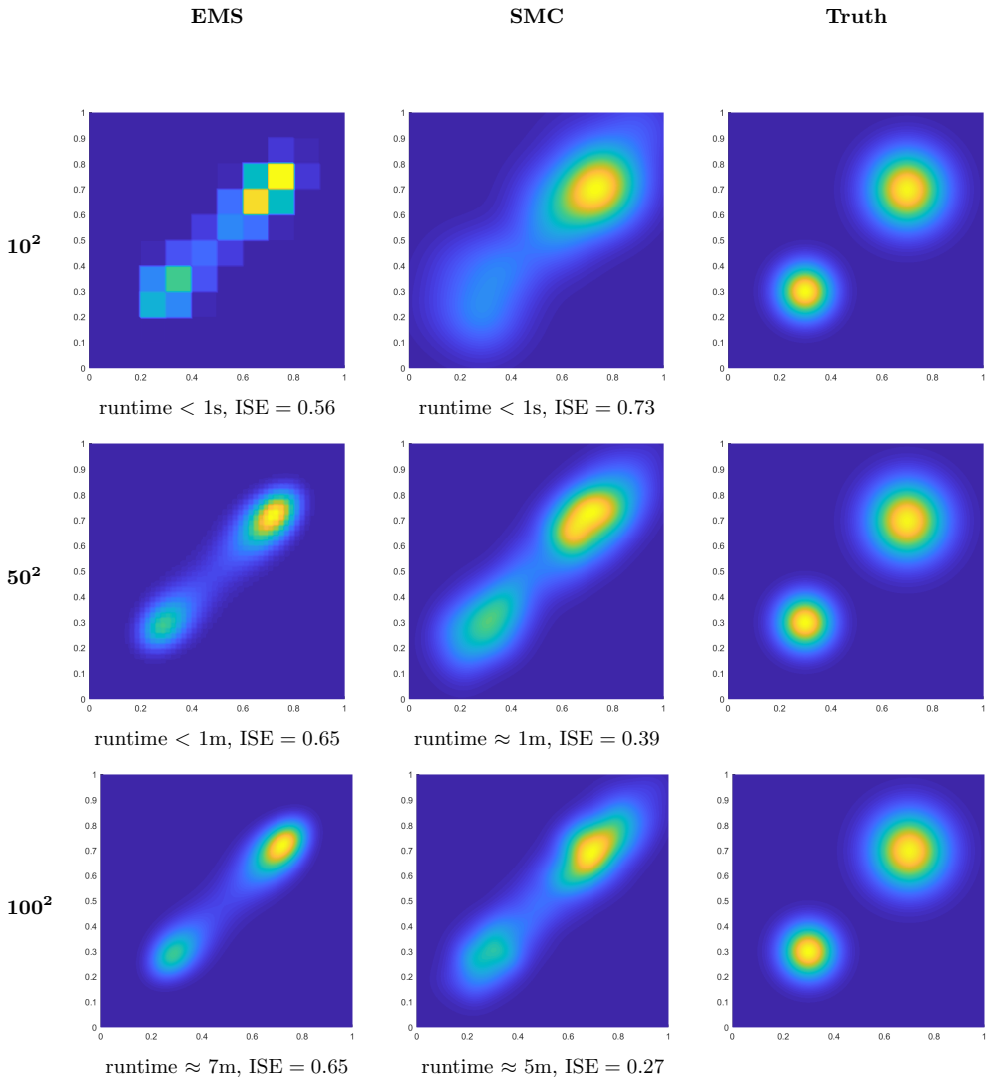


Figure 16: Reconstructions of a 2-dimensional mixture of Gaussian obtained with EMS and SMC. The number of bins/particles increases from 10^2 to 100^2. Runtime and accuracy are reported too.

increases (Silverman, 1986) and is primarily used in this work for visualisation and human interpretation (which becomes less informative in higher dimension, with the exception of low dimensional projections), to compare the performances of EMS and SMC in dimension $d_{\mathbb{X}} \geq 2$ we focus on approximating expectations w.r.t. η_{n+1} of appropriate test functions φ, in this case, in fact, Proposition 5.2 gives the rate of convergence in terms of the number of particles N. In particular, we consider mean, variance, the probability of the region $[0, 0.5]^{d_{\mathbb{X}}}$ and the probability of a hyper-sphere of radius 0.3 around the mode at $(0.3, \ldots, 0.3)$. We compare three particle sizes $N = 10^2, 10^3, 10^4$ and obtain the number of bins for each dimension as $\lceil N^{1/d_{\mathbb{X}}} \rceil$ so that the total number of bins, $B = \lceil N^{1/d_{\mathbb{X}}} \rceil^{d_{\mathbb{X}}}$, where $\lceil \cdot \rceil$ denotes the ceiling function, roughly matches N. This choice allows us to compare EMS and SMC reconstructions which require roughly the same runtime (Table 2).

The SMC implementation in generally better at recovering the variance and the probability of the region $[0, 0.5]^{d_{\mathbb{X}}}$. For small values of N, B, both SMC and EMS have larger errors with discretised EMS achieving better crude estimates. However, as N, B increase SMC is consistently better at approximating the four quantities considered, in particular, in the case of mean and variance the estimates are at least one order of magnitude more accurate. This is achieved at a computational cost which is always smaller than that of EMS and that could be in principle reduced by considering smaller values of M.

	mean	variance	$\mathbb{P}(\square)$	$\mathbb{P}(\bigcirc)$	$\log_{10}$(runtime / s)
$d_{\mathbb{X}} = 2$					
EMS - $B = 10^2$	**1.38e-04**	4.96e-05	5.30e-02	**7.04e-03**	-1.71
SMC - $N = 10^2$	3.87e-04	**1.26e-05**	**4.90e-02**	1.02e-02	**-2.02**
EMS - $B = 32^2$	1.42e-04	5.31e-05	5.17e-02	**5.86e-03**	1.28
SMC - $N = 10^3$	**4.29e-05**	**5.81e-06**	**4.69e-02**	6.14e-03	**0.94**
EMS - $B = 100^2$	1.42e-04	5.38e-05	5.15e-02	6.11e-03	5.31
SMC - $N = 10^4$	**3.84e-06**	**4.51e-06**	**4.67e-02**	**5.68e-03**	**5.11**
$d_{\mathbb{X}} = 3$					
EMS - $B = 5^3$	**2.53e-04**	1.26e-04	1.46e-01	8.59e-03	-1.47
SMC - $N = 10^2$	3.76e-04	**3.23e-05**	**7.19e-02**	**3.92e-03**	**-2.06**
EMS - $B = 10^3$	2.00e-04	5.75e-05	9.00e-02	2.42e-03	1.40
SMC - $N = 10^3$	**4.62e-05**	**8.50e-06**	**8.24e-02**	**1.36e-03**	**1.08**
EMS - $B = 22^3$	2.04e-04	6.12e-05	8.83e-02	1.64e-03	5.66
SMC - $N = 10^4$	**3.53e-0**6	**6.68e-06**	**8.23e-02**	**9.24e-04**	**5.30**
$d_{\mathbb{X}} = 4$					
EMS - $B = 4^4$	**1.98e-04**	**1.55e-05**	1.22e-01	**1.16e-03**	-0.65

continued from previous page					
	mean	variance	$\mathbb{P}(\square)$	$\mathbb{P}(\bigcirc)$	$\log_{10}$(runtime / s)
SMC - $N = 10^2$	4.77e-04	9.77e-05	**7.35e-02**	4.41e-03	**-2.08**
EMS - $B = 6^4$	2.43e-04	4.02e-05	1.09e-01	7.80e-04	1.70
SMC - $N = 10^3$	**3.45e-05**	**1.80e-05**	**9.13e-02**	**5.47e-04**	**0.95**
EMS - $B = 10^4$	2.60e-04	6.59e-05	1.03e-01	5.54e-04	5.32
SMC - $N = 10^4$	**4.10e-06**	**8.58e-06**	**9.45e-02**	**2.38e-04**	**5.12**
$d_{\mathbb{X}} = 5$					
EMS - $B = 3^5$	**5.66e-05**	2.67e-04	2.12e-01	**1.27e-02**	-0.56
SMC - $N = 10^2$	6.59e-04	**1.34e-04**	**5.10e-02**	1.10e-02	**-1.96**
EMS - $B = 4^5$	2.42e-04	**2.08e-05**	1.29e-01	7.59e-04	1.51
SMC - $N = 10^3$	**5.57e-05**	4.54e-05	**7.84e-02**	**4.49e-04**	**1.14**
EMS - $B = 7^5$	2.82e-04	5.71e-05	1.36e-01	2.09e-04	6.63
SMC - $N = 10^4$	**3.39e-06**	**1.27e-05**	**9.57e-02**	**7.43e-05**	**5.36**

Table 2: Runtime and mean squared error over 100 repetitions for mean, variance, probability of the lower quadrant and probability of a circle around the mode for the $d_{\mathbb{X}}$-dimensional Gaussian mixture model. Best values are in **bold**.

5.6 Summary

We have identified a close connection between the continuous EMS recursion and a Feynman-Kac flow in the space of probability measures. This connection allows us to propose a novel SMC algorithm which approximates this flow and provides a stochastic discretisation of the EMS recursion which can be naturally implemented when only samples from the distorted signal h are available. This stochastic discretisation of the EMS recursion does not require the assumption of piecewise constant signals common to deterministic discretisation schemes.

We studied the asymptotic properties of the proposed SMC scheme, extending the results on $\mathbb{L}_p$ convergence of expectations, the strong law of large numbers and the bias estimate for standard SMC. As a consequence, we establish almost sure weak convergence of the SMC approximations to the EMS recursion as the number of particles N goes to infinity. We also provided theoretical guarantees on the proposed estimator for the solution f of the Fredholm integral equation.

An extensive empirical study has demonstrated the good performance of the proposed method on a suite of illustrative examples, showing that the proposed method is competitive in terms of reconstruction accuracy with state-of-the-art algorithms for image reconstruction.

6

Wasserstein Gradient Flows for Fredholm Equations of the First Kind

An extended version of this chapter is in preparation in collaboration with Dr Valentin De Bortoli (Crucinio et al., 2021a).

6.1 Introduction

The EMS recursion discussed in the previous chapters is a natural way to construct smooth approximations of the solution of Fredholm integral equations, however, in its continuous formulation, the connection with maximum penalised likelihood methods is not clear. In particular, we have so far been unable to show that the fixed point of the EMS iteration minimises a particular functional; as a consequence, our convergence checks in Section 5.5 have been based on the Kullback–Leibler divergence, since we expect that for sufficiently low levels of smoothness the EMS recursion still moves the iterates towards approximate solutions which have small Kullback–Leibler divergence.

In this chapter, we go back to the minimisation problem defined by the Kullback–Leibler divergence (3.4)

$$\mathrm{KL}\left(h, \int_{\mathbb{X}} f(x)g(\cdot \mid x)\mathrm{d}x\right) = \int_{\mathbb{Y}} h(y) \log\left(\frac{h(y)}{\int_{\mathbb{X}} f(x)g(y \mid x)\mathrm{d}x}\right)\mathrm{d}y,$$

and consider explicit penalties which enforce smoothness of the approximate solution. A standard way to measure the smoothness of a function f is through its derivatives, however, this choice requires the approximate solution f to be expressed in parametric form or as a combination of differentiable basis functions (Good, 1971) and is not suitable for approaches based on interacting particle methods.

In the image processing literature, a common alternative to penalties involving derivatives of f is the entropic penalty (e.g., Molina et al. (1992)) given by Shannon's differential entropy (Jaynes, 1957a,b)

$$\operatorname{ent}(f) := -\int_{\mathbb{X}} f(x) \log f(x) \mathrm{d}x. \tag{6.1}$$

The remainder of this chapter focusses on the following minimisation problem

$$E_\alpha(f) = \mathrm{KL}\left(h, \int_{\mathbb{X}} f(x) g(\cdot \mid x) \mathrm{d}x\right) - \alpha \operatorname{ent}(f) \tag{6.2}$$

obtained by adding the entropy penalty (6.1) to the Kullback–Leibler divergence. The choice of an entropic penalty is particularly convenient: on one hand it addresses the inconsistency of the maximum likelihood estimator by imposing a smoothness constraint on f, on the other hand it leads to an easy-to-implement numerical scheme as we will see in Chapter 7.

The use of entropy regularisation is not new in the literature on Fredholm integral equations; Amato and Hughes (1991) propose a variant to Tikhonov regularisation (3.2) in which the $\mathbb{L}_2$ norm penalisation is replaced by an entropic penalty, an iterative method performing Tikhonov regularisation with entropic penalty is presented in Burger et al. (2019). The one-step late expectation maximisation (OSL-EM) of Green (1990) gives an iterative scheme to minimise a generic penalised KL divergence, but it is not suitable for an entropic penalty, since the resulting algorithm might give negative estimates for f. Considering (6.2) as a Lagrangian form of the minimisation problem

$$\max \operatorname{ent}(f) \qquad \text{s.t. } \mathrm{KL}\left(h, \int_{\mathbb{X}} f(x) g(\cdot \mid x) \mathrm{d}x\right) = 0, \tag{6.3}$$

connects (6.2) with maximum entropy methods (Jaynes, 1957a,b). A number of maximum entropy approximations of solutions of Fredholm integral equations have been proposed in the literature, most of which maximise the entropy subject to moment constraints obtained by integrating h and $\int_{\mathbb{X}} f(x) g(\cdot \mid x) \mathrm{d}x$ with respect to a set of basis functions (Islam and Smith, 2020; Jin and Ding, 2016; Kopeć, 1993; Mead, 1986).

The constraint in (6.3) cannot be obtained as integral with respect to basis functions and direct minimisation of (6.2) is usually not feasible; we propose to follow a Wasserstein gradient flow approach. This approach stems from the connection between minimisation of functionals in the space of probability measures and partial

differential equations (PDEs) pointed out in Jordan et al. (1998); Otto (2001) and allows us to transform the minimisation of (6.2) into a Fokker-Plank PDE whose solution is a minimiser of E_α.

After recalling the fundamental ideas of Wasserstein gradient flows, we show how this construction can be applied to E_α and identify conditions under which the gradient flow PDE admits a unique solution.

6.2 Wasserstein Gradient Flows

Gradient flows, or gradient descent, is a classical method to find minima of a function $F : \mathbb{R}^d \to \mathbb{R}$ by looking at curves $x(t)$ which follow the direction of steepest descent of F, given by $-\nabla F$. Given an initial point x_0, the curve satisfies the differential equation

$$x'(t) = -\nabla F(x(t)) \tag{6.4}$$

which, provided F is smooth enough, admits a unique solution that is a minimum of F.

Wasserstein gradient flows are the natural extension of gradient flows in $\mathbb{R}^d$ to the space of probability measures on $\mathbb{R}^d$. This extension has been introduced in Jordan et al. (1998) and Otto (2001) and fully developed in Ambrosio et al. (2008). We briefly recall the key ingredients for the definition of Wasserstein gradient flows below. We denote the set of probability measures with finite second moment on $\mathbb{R}^d$ by

$$\mathcal{P}_2(\mathbb{R}^d) = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int \|x\|_2^2 \mu(\mathrm{d}x) < \infty \right\}$$

and we define the 2-Wasserstein distance on this set

$$W_2(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|_2^2 \pi \left(\mathrm{d}(x, y)\right) \right)^{1/2} \tag{6.5}$$

where $\Pi(\mu, \nu)$ is the set of all possible couplings between μ and ν. A minimiser of (6.5) always exists but might not be unique, each minimiser is called an optimal transport plan (Ambrosio et al., 2008, Theorem 6.2.4). We denote by $\mathcal{P}_2^{ac}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ the subset of these measures which is absolutely continuous with respect to the appropriate Lebesgue measure. For every pair $\mu, \nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ there is a unique π attaining the minimum of (6.5), $\pi = (Id, t_\mu^\nu)_\# \mu$, where $T_\# \mu$ denotes the push-forward measure, $T_\# \mu(A) = \mu(T^{-1}(A))$, and t_μ^ν is the unique transport map

between μ and ν (see, for example, Ambrosio et al. (2008, page 150)). It is easy to check that for all $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ we have $\text{ent}(\nu) < +\infty$.

To define gradient flows on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ we need a notion of curves; among the several definitions proposed (Ambrosio et al., 2008, Chapter 7) we consider constant speed geodesics with respect to W_2, i.e. curves $\mu_s \in \mathcal{P}_2(\mathbb{R}^d)$, $s \in [0,1]$, such that

$$W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1) \qquad \forall\ 0 \leq s \leq t \leq 1.$$

For $\mu, \nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, the constant speed geodesic originating from μ and with endpoint ν is

$$\mu_s : s \in [0,1] \mapsto ((1-s)Id + st_\mu^\nu)_{\#}\mu \tag{6.6}$$

where t_μ^ν is the unique optimal transport map between μ and ν. Constant speed geodesics are usually preferred to other families of geodesics due to their easier formulation as they only require the transport map t_μ^ν (Ambrosio et al., 2008, page 11), in addition, the notion of convexity along constant-speed geodesics introduced at the end of this section is intuitively closer to the notion of convexity in standard Euclidean spaces than that of convexity along generalised geodesics.

We then consider functionals $F : \mathcal{P}_2^{ac}(\mathbb{R}^d) \to \mathbb{R}$ defined on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ and build a gradient flow solving the minimisation problem

$$\min_{\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^d)} F(\rho). \tag{6.7}$$

The construction of a gradient flow equation for F allows us to transform the minimisation problem (6.7) into a PDE whose solution is a density ρ solving (6.7). We will restrict our attention to functionals which are proper (i.e. $F(\rho) < +\infty$ for some $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$), continuous with respect to the W_2 metric (which metrises weak convergence, Santambrogio (2017, Theorem 4.4)) and coercive. Following Ambrosio et al. (2008, Definition 2.1b), a functional F defined on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ is coercive if there exist $\tau > 0$ and $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ such that

$$\inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^d)} \frac{1}{2\tau} W_2^2(\nu, \rho) + F(\rho) > -\infty. \tag{6.8}$$

Intuitively, (6.8) means that the functional F decays more slowly than the squared W_2 distance grows as we move away from the centre of the space.

Since we are dealing with an optimisation problem, it is natural to assume that the functional F is convex in an appropriate sense to guarantee existence (and uniqueness) of the minimisers; in $\mathcal{P}_2^{ac}(\mathbb{R}^d)$ the standard notion of convexity in $\mathbb{R}^d$

corresponds to convexity along geodesics: for all $\nu, \mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ take a geodesic connecting μ to ν, then the functional F is λ-geodesically convex if

$$F\left(((1-s)Id + st_\mu^\nu)_\# \mu\right) \leq (1-s)F(\mu) + sF(\nu) - \frac{\lambda}{2}s(1-s)W_2^2(\nu,\mu)$$

for some $\lambda \in \mathbb{R}$. It is easy to see that when $\lambda = 0$ the above corresponds to standard convexity, in this case the functional is called displacement convex, if $\lambda > 0$ the above is stronger than convexity, and for $\lambda < 0$ is weaker (Ambrosio et al., 2008, page 202). In the case of λ-convex functionals, coercivity (6.8) is equivalent to the existence of some $r > 0$ such that F admits a minimiser on the subset of $\mathcal{P}_2(\mathbb{R}^d)$ in which $\int_{\mathbb{R}^d} \|x\|_2^2 \mu(\mathrm{d}x) \leq r$ (Ambrosio et al., 2008, page 295).

The last ingredient in the definition of a gradient flow for functional F on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ is a notion of gradient for F. As standard definitions of derivatives do not apply outside vector spaces, the gradient in (6.4) is replaced with the sub-differential of F, $\partial F(\mu)$. For any proper, continuous functional F defined on $\mathcal{P}_2^{ac}(\mathbb{R}^d)$, ξ belongs to the sub-differential $\partial F(\mu)$ if

$$\liminf_{W_2(\nu,\rho)\to 0} \frac{F(\nu) - F(\rho) - \left\langle \xi, t_\rho^\nu(x) - x \right\rangle}{W_2(\nu,\rho)} \geq 0 \tag{6.9}$$

for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2008, Definition 10.1.1). Equipped with this notion of gradient, we can now define the gradient flow equation for a functional F on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$: ρ_t is a solution of the gradient flow equation for F if

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t) \tag{6.10}$$

with $v_t \in -\partial F(\rho_t)$ (Ambrosio et al., 2008, Section 11.1.2). Existence of solutions of (6.10) given an initial condition $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ is ensured for sufficiently regular functionals (e.g (Ambrosio et al., 2008, Theorem 11.1.6 and Theorem 11.3.2). The solution of (6.10) with initial condition ρ_0 is unique for λ-convex functionals with $\lambda > 0$ (Ambrosio et al., 2008, Theorem 11.1.4). In particular we have the estimate

$$W_2(\rho_t^1, \rho_t^2) \leq e^{-\lambda t} W_2(\rho_0^1, \rho_0^2) \tag{6.11}$$

for all $t > 0$, for ρ_t^i solution of the gradient flow equation with initial condition ρ_0^i, $i = 1, 2$, which shows that the gradient flow is contractive w.r.t. W_2 when $\lambda > 0$ and non-expansive when $\lambda = 0$.

In this section we reviewed the essential tools to develop our methodology, how-

ever the literature on Wasserstein gradient flows is much vaster and allows for generalisations of most of the ideas introduced above (Ambrosio et al., 2008).

6.3 Wasserstein Gradient Flows for Fredholm Integral Equations

In order to apply the gradient flow construction described above to (6.2) it is necessary to consider the probability measures ρ corresponding to the density f and μ corresponding to h, respectively. Thus, we write (3.1) as

$$\mu(y) = \int_{\mathbb{X}} \rho\,(\mathrm{d}x)\, g(y \mid x), \qquad \forall y \in \mathbb{Y} \tag{6.12}$$

and (6.2) as

$$E_\alpha(\rho) = \mathrm{KL}(\mu, \rho g) - \alpha\,\mathrm{ent}(\rho) \tag{6.13}$$

where $\rho g(y) := \int_{\mathbb{X}} \rho\,(\mathrm{d}x)\, g(y \mid x)$.

For the remainder of this chapter and the following one we assume the following:

Assumption 6.1. μ and ρ are probability measures absolutely continuous with respect to the Lebesgue measure with finite second moment (i.e. $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{Y}})$, $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$). In addition, g is the density of a Markov kernel from $\mathbb{R}^{d_\mathbb{X}}$ to $\mathbb{R}^{d_\mathbb{Y}}$ such that

(a) is bounded above, $g(y \mid x) \leq m_g$ for all x, y, and $\lambda(y)$-concave in x for all $y \in \mathbb{R}^{d_\mathbb{Y}}$

$$g(y \mid sx + (1-s)x') \geq sg(y \mid x) + (1-s)g(y \mid x') + \frac{\lambda(y)}{2} s(1-s)\|x - x'\|_2^2$$

with $s \in [0,1]$ and $\lambda(y) \in \mathbb{R}$;

(b) there exist $\Phi : \mathbb{R}^{d_\mathbb{X}} \to [0, +\infty)$, such that $\Phi(x) \leq a\|x\|_2^2 + b$, and a μ-integrable function $\Psi : \mathbb{R}^{d_\mathbb{Y}} \to [0, +\infty)$ such that

$$\sup_{(x,y) \in \mathbb{R}^{d_\mathbb{X}} \times \mathbb{R}^{d_\mathbb{Y}}} \frac{\exp\left(-\Phi(x) - \Psi(y)\right)}{g(y \mid x)} < +\infty;$$

(c) is differentiable with respect to x with bounded gradient with respect to the first argument $\nabla_x g(y \mid x)$, $\|\nabla_x g(y \mid x)\|_2 \leq B$;

(d) can be evaluated pointwise and a sample $\mathbf{Y}$ from h is available.

Under Assumption 6.1-(a) we can show existence (and uniqueness) of the solution of the gradient flow PDE. The λ-concavity assumption is rather strong and rarely satisfied in practice; however, every twice continuously differentiable function is λ-concave on any bounded set for some $\lambda < 0$ (Santambrogio, 2017, page 91). In the experiments that we will present in Chapter 7 this assumption is not satisfied, and the kernels g that we consider are twice-differentiable and therefore only locally λ-concave; however, we observe that the gradient flow construction provides good results. As this is a very active area of research, there are directions to relax the λ-concavity assumption to weaker moduli of convexity (e.g. Craig (2017)).

Assumption 6.1-(b) controls the behaviour of the tails of g, imposing a constraint to how quickly they can decay to 0. Assumption 6.1-(c) implies that g is Lipschitz continuous in x, uniformly in y, with Lipschitz constant B

$$\|g(y \mid x) - g(y \mid x')\|_2 \leq B\|x - x'\|_2 \qquad \forall (x, y), (x', y) \in \mathbb{R}^{d_{\mathbb{X}}} \times \mathbb{R}^{d_{\mathbb{Y}}}, \tag{6.14}$$

and is one of the key assumptions to ensure that the drift of the SDE corresponding to the gradient flow PDE introduced in Chapter 7 does not explode.

6.3.1 Properties of the Functional E_α

Under the assumptions above, we can show that the functional E_α in (6.13) defined on $\mathcal{P}_2^{ac}(\mathbb{R}^{d_{\mathbb{X}}})$ is proper, coercive, lower semi-continuous and geodesically convex; as a consequence we can build a gradient flow targeting the minimum of (6.13).

Proposition 6.1. Under Assumption 6.1, the functional E_α is proper, lower semi-continuous and coercive in $\left(\mathcal{P}_2^{ac}(\mathbb{R}^{d_{\mathbb{X}}}), W_2\right)$.

Proof. First we show that and that $E_\alpha(\rho) = \mathrm{KL}(\mu, \rho g) - \alpha\, \mathrm{ent}(\rho)$ is proper. Under Assumption 6.1-(b) there exists $C \geq 0$ such that $g(y \mid x) \geq C^{-1} \exp\left(-\Phi(x) - \Psi(y)\right)$ for all $(x, y) \in \mathbb{R}^{d_{\mathbb{X}}} \times \mathbb{R}^{d_{\mathbb{Y}}}$, then, using Jensen's inequality,

$$\begin{aligned}
\log\left(\int \rho(\mathrm{d}x) g(y \mid x)\right) &\geq \log\left(C^{-1} \int \rho(\mathrm{d}x) \exp\left(-\Phi(x) - \Psi(y)\right)\right) \\
&\geq \log C^{-1} - \int \rho(\mathrm{d}x)\Phi(x) - \Psi(y).
\end{aligned}$$

From the above we obtain, as $\Phi(x) \leq a\|x\|_2^2 + b$ and $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_{\mathbb{X}}})$,

$$-\int \mu(\mathrm{d}y) \log \rho g(y) \leq -\log C^{-1} + \rho(\Phi) + \mu(\Psi) < \infty.$$

Because $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$ and $\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{Y}})$ their entropies are finite and therefore $E_\alpha(\rho) < \infty$.

To check continuity, recall that W_2 metrises weak convergence and take a sequence $\{\rho_n\}_{n\geq 1} \subset \mathcal{P}_2^{ac}(\mathbb{R}^n)$ converging weakly to ρ. Lower semi-continuity of $-\operatorname{ent}(\rho)$ follows from standard arguments (e.g. Santambrogio (2015, page 331)).

Under Assumption 6.1, $x \mapsto g(y \mid x)$ is a continuous function for all $y \in \mathbb{R}^{d_\mathbb{Y}}$ and weak convergence implies $\rho_n g(y) \to \rho g(y)$, additionally the functions $y \mapsto \rho_n g(y)$ and $y \mapsto \rho g(y)$ are continuous as a consequence of the continuity of $y \mapsto g(y \mid x)$ and the Dominated Convergence Theorem. The continuity of the logarithm then gives $\log \rho_n g(y) \to \log \rho g(y)$ for each fixed y.

Under Assumption 6.1-(a) and (b) there exists $C \geq 0$ such that $g(y \mid x) \geq C^{-1} \exp\left(-\Phi(x) - \Psi(y)\right)$ for all $(x, y) \in \mathbb{R}^{d_\mathbb{X}} \times \mathbb{R}^{d_\mathbb{Y}}$ and we have that

$$|\log \rho_n g(y)| \leq \max\left(m_g, |\log C^{-1} - \rho_n(\Phi) - \Psi(y)|\right). \tag{6.15}$$

Using the fact that Ψ is μ-integrable and that $\Phi(x) \leq a\|x\|_2^2 + b$ we obtain

$$\int \mu(\mathrm{d}y)|\log C^{-1} - \rho_n(\Phi) - \Psi(y)| \leq |\log C^{-1}| + a\int \rho_n(\mathrm{d}x)\|x\|_2^2 + b + \mu(\Psi) < \infty,$$

for all $n \in \mathbb{N}$ since $\{\rho_n\}_{n\geq 1} \subset \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$, which shows that the right-hand-side of (6.15) is uniformly μ-integrable.

Then, Vitali's Convergence Theorem (e.g., Dudley (2002, Theorem 10.3.5)) ensures that

$$\left|\int \mu(\mathrm{d}y)\left[\log \rho g(y) - \log \rho_n g(y)\right]\right| \to 0,$$

as $n \to \infty$. This gives continuity of $-\int \mu(\mathrm{d}y) \log \rho g(y)$, from which follows continuity of $\mathrm{KL}(\mu, \rho g) = -\operatorname{ent}(\mu) - \int \mu(\mathrm{d}y) \log \rho g(y)$ as the first term does not depend on ρ. Combining this with the lower semi-continuity of the entropy we obtain the result.

Coercivity follows straightforwardly from the definition of E_α: since the KL divergence is non-negative and W_2 is a metric we have that for all $\tau > 0$,

$$\begin{aligned}\frac{1}{2\tau}W_2^2(\nu, \rho) + E_\alpha(\rho) &= \frac{1}{2\tau}W_2^2(\nu, \rho) + \mathrm{KL}(\mu, \rho g) - \alpha \operatorname{ent}(\rho) \\ &\geq -\alpha \operatorname{ent}(\rho);\end{aligned}$$

recalling that $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$, we obtain

$$\inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})} \frac{1}{2\tau} W_2^2(\nu, \rho) + E_\alpha(\rho) \geq -\alpha \inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})} \operatorname{ent}(\rho) > -\infty.$$

□

Finally, we show that when g is $\lambda(y)$-concave with $\lambda(y) \geq 0$, E_α is displacement convex, ensuring that, given an initial condition $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$, the gradient flow equation has a unique solution:

Proposition 6.2. Under Assumption 6.1, if g is $\lambda(y)$-concave with $\lambda(y) \geq 0$ for all $y \in \mathbb{R}^{d_\mathbb{Y}}$, the functional E_α is displacement convex in $\left(\mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}}), W_2\right)$.

Proof. We have

$$\begin{aligned} E_\alpha(\rho) &= \mathrm{KL}(\mu, \rho g) - \alpha \operatorname{ent}(\rho) \\ &= -\int \mu(\mathrm{d}y) \log \rho g(y) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu). \end{aligned}$$

The entropy $\operatorname{ent}(\mu)$ is constant with respect to ρ and $\operatorname{ent}(\rho)$ is displacement convex in ρ (Santambrogio, 2017, page 130).

Let us define the functionals $G_y : \rho \mapsto \rho g(y) = \int \rho(\mathrm{d}x) g(y \mid x)$ and $F(\rho) := -\int \mu(\mathrm{d}y) \log G_y(\rho)$. Under Assumption 6.1-(a), for fixed $y \in \mathbb{R}^{d_\mathbb{Y}}$, g is $\lambda(y)$-concave in x and G_y is $\lambda(y)$-geodesically concave; in particular, we have that for all $s \in [0, 1]$

$$\begin{aligned} G_y\left(((1-s)Id + st_\rho^\nu)_\# \rho\right) &\geq (1-s)G_y(\rho) + sG_y(\nu) + \frac{\lambda(y)}{2} s(1-s) W_2^2(\nu, \rho) \\ &\geq (1-s)G_y(\rho) + sG_y(\nu) + \frac{\inf_{y \in \mathbb{R}^{d_\mathbb{Y}}} \lambda(y)}{2} s(1-s) W_2^2(\nu, \rho) \end{aligned}$$

where t_ρ^ν is the unique transport map between ν and ρ (Carrillo et al., 2006, Lemma 5). Because $\inf_{y \in \mathbb{R}^{d_\mathbb{Y}}} \lambda(y) \geq 0$ we then have

$$G_y\left(((1-s)Id + st_\rho^\nu)_\# \rho\right) \geq (1-s)G_y(\rho) + sG_y(\nu)$$

and, since $-\log x$ is a convex decreasing function,

$$\begin{aligned} -\log\left(G_y\left(((1-s)Id + st_\mu^\nu)_\# \rho\right)\right) &\leq -\log\left((1-s)G_y(\rho) + sG_y(\nu)\right) \\ &\leq -(1-s)\log\left(G_y(\rho)\right) - s\log\left(G_y(\nu)\right), \end{aligned}$$

the same inequality holds after integration w.r.t. μ showing that F is displacement

convex. This result and the convexity of the entropy give convexity of E_α.

□

When the kernel g is only weakly concave (i.e. $\lambda(y) < 0$ for some $y \in \mathbb{R}^{d_\mathbb{Y}}$) the functional $F = -\int \mu(\mathrm{d}y) \log \rho g(y)$ is generally not geodesically convex; in this case the construction of a gradient flow for E_α is still possible and follows the same steps as those for the convex case, however, the solution of the gradient flow equation (6.10) need not be unique.

6.3.2 Gradient Flow for E_α

The last step towards the definition of the gradient flow equation (6.10) for the functional E_α consists of finding the quantity v_t which belongs to its (negative) sub-differential. A good candidate is the gradient of the first variation (or functional derivative) of E_α, $\frac{\delta E_\alpha}{\delta \rho}(x)$, the unique (up to additive constant) function such that

$$\lim_{\epsilon \to 0} \epsilon^{-1} \left(E_\alpha(\rho + \epsilon\chi) - E_\alpha(\rho) \right) = \int \chi(\mathrm{d}x) \frac{\delta E_\alpha}{\delta \rho}(x)$$

for every signed measure χ such that $\rho + \epsilon\chi \in \mathcal{P}_2(\mathbb{R}^{d_\mathbb{X}})$ for some $\epsilon > 0$ (Santambrogio, 2017).

Given the functional E_α

$$E_\alpha(\rho) = \int \mu(\mathrm{d}y) \log \rho g(y) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu),$$

the last term, $\operatorname{ent}(\mu)$, does not depend on ρ and does not contribute to $\frac{\delta E_\alpha}{\delta \rho}$. The first variation of the entropy functional is

$$\frac{-\delta \operatorname{ent}}{\delta \rho}(x) = 1 + \log \rho(x) \tag{6.16}$$

and its gradient with respect to x is a sub-differential for $\operatorname{ent}(\rho)$ (Ambrosio et al.,

2008, Chapter 10). Then, let us define $F(\rho) := -\int \mu(\mathrm{d}y) \log \rho g(y)$, and

$$\begin{aligned}
&\lim_{\epsilon \to 0} \epsilon^{-1} \left(F\left(\rho + \epsilon\chi\right) - F(\rho) \right) \\
&\qquad = \lim_{\epsilon \to 0} \epsilon^{-1} \left(-\int \mu(\mathrm{d}y) \log\left(\left(\rho + \epsilon\chi\right) g(y) \right) + \int \mu(\mathrm{d}y) \log \rho g(y) \right) \\
&\qquad = \lim_{\epsilon \to 0} \epsilon^{-1} \left(-\int \mu(\mathrm{d}y) \log\left(1 + \epsilon \frac{\chi g(y)}{\rho g(y)} \right) \right) \\
&\qquad = \lim_{\epsilon \to 0} \epsilon^{-1} \left(-\int \mu(\mathrm{d}y) \left(\epsilon \frac{\chi g(y)}{\rho g(y)} + o\left(\epsilon \frac{\chi g(y)}{\rho g(y)} \right) \right) \right) \\
&\qquad = -\int \mu(\mathrm{d}y) \frac{\chi g(y)}{\rho g(y)},
\end{aligned}$$

where the third equality follows from the Taylor expansion of the logarithm as $\epsilon \to 0$ and the last inequality from the Monotone Convergence Theorem. Applying Fubini's Theorem gives

$$\lim_{\epsilon \to 0} \epsilon^{-1} \left(F\left(\rho + \epsilon\chi\right) - F(\rho) \right) = -\int \chi(\mathrm{d}x) \int \mu(\mathrm{d}y) \frac{g(y \mid x)}{\rho g(y)}$$

showing that

$$\frac{\delta F}{\delta \rho}(x) = -\int \mu\left(\mathrm{d}y\right) \frac{g(y \mid x)}{\rho g(y)}$$

and

$$\nabla_x \frac{\delta F}{\delta \rho}(x) = -\int \mu\left(\mathrm{d}y\right) \frac{\nabla_x g(y \mid x)}{\rho g(y)}, \tag{6.17}$$

where we can apply the Leibniz integral rule for differentiation under the integral sign (e.g. Billingsley (1995, Theorem 16.8)) because g is bounded and differentiable by Assumption 6.1-(a),(c).

To show that (6.17) is a sub-differential for F, take $\rho, \nu \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_{\mathbb{X}}})$, the unique transport map between ρ and ν, t_ρ^ν, and use the definition of sub-differential in (6.9):

$$\begin{aligned}
&F(\nu) - F(\rho) - \int \rho(\mathrm{d}x) \left\langle \nabla_x \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle \\
&= \int \mu(\mathrm{d}y) \left[-\log \nu g(y) + \log \rho g(y) \right] + \int \rho(\mathrm{d}x) \left\langle \int \mu\left(\mathrm{d}y\right) \frac{\nabla_x g(y \mid x)}{\rho g(y)}, t_\rho^\nu(x) - x \right\rangle \\
&= \int \mu(\mathrm{d}y) \left[-\log \nu g(y) + \log \rho g(y) \right] + \int \frac{\mu\left(\mathrm{d}y\right)}{\rho g(y)} \int \rho(\mathrm{d}x) \left\langle \nabla_x g(y \mid x), t_\rho^\nu(x) - x \right\rangle \\
&= \int \mu(\mathrm{d}y) \left[-\log \frac{\nu g(y)}{\rho g(y)} + \frac{1}{\rho g(y)} \int \rho(\mathrm{d}x) \left\langle \nabla_x g(y \mid x), t_\rho^\nu(x) - x \right\rangle \right].
\end{aligned}$$

Since $g(y \mid \cdot)$ is λ-concave (with λ depending on y), $-g(y \mid \cdot)$ is λ-convex and

Ambrosio et al. (2008, Chapter 10) show that $-\nabla_x g(y \mid \cdot)$ is a sub-differential for $\rho g(y)$ and

$$-\nu g(y) + \rho g(y) + \int \rho(\mathrm{d}x)\langle \nabla_x g(y \mid x), t_\rho^\nu(x) - x\rangle \geq \frac{\lambda(y)}{2} W_2^2(\nu, \rho) \tag{6.18}$$

and

$$\begin{aligned}
&F(\nu) - F(\rho) - \int \rho(\mathrm{d}x) \left\langle \nabla_x \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle \\
&\geq \int \mu(\mathrm{d}y) \left[-\log \frac{\nu g(y)}{\rho g(y)} + \frac{-\rho g(y) + \nu g(y)}{\rho g(y)} \right] + W_2^2(\nu, \rho) \int \mu(\mathrm{d}y) \frac{\lambda(y)}{2} \\
&= \int \mu(\mathrm{d}y) \left[-\log \frac{\nu g(y)}{\rho g(y)} + \frac{\nu g(y)}{\rho g(y)} - 1 \right] + W_2^2(\nu, \rho) \frac{\mu(\lambda)}{2}.
\end{aligned}$$

The first order Taylor expansion with Lagrange remainder $-\log x = -\log 1 - (x - 1) + (x - a)^2/(2a^2)$ with a between x and 1 gives

$$\left[-\log \frac{\nu g(y)}{\rho g(y)} + \frac{\nu g(y)}{\rho g(y)} - 1 \right] = \frac{1}{2a(y)^2} \left(\frac{\nu g(y)}{\rho g(y)} - a(y) \right)^2 \geq 0,$$

with $a(y)$ a value between $\nu g(y)$ and $\rho g(y)$. Therefore

$$F(\nu) - F(\rho) - \int \rho(\mathrm{d}x) \left\langle \nabla_x \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle \geq W_2^2(\nu, \rho) \frac{\mu(\lambda)}{2},$$

from which follows

$$\frac{F(\nu) - F(\rho) - \int \rho(\mathrm{d}x) \left\langle \nabla_x \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle}{W_2(\nu, \rho)} \geq W_2(\nu, \rho) \frac{\mu(\lambda)}{2} \to 0$$

as $W_2^2(\nu, \rho) \to 0$, showing that $\nabla_x \frac{\delta F}{\delta \rho}(x)$ is a sub-differential for F.

Putting (6.16) and (6.17) together we obtain the sub-differential for E_α

$$\begin{aligned}
\nabla_x \frac{\delta E_\alpha}{\delta \rho}(x) &= \nabla_x \left[-\int \mu(\mathrm{d}y) \frac{g(y \mid x)}{\rho g(y)} + \alpha(1 + \log \rho(x)) \right] \\
&= -\int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid x)}{\rho g(y)} + \alpha \nabla_x \log \rho(x).
\end{aligned} \tag{6.19}$$

Proposition 6.3. Under Assumption 6.1, if there exists an absolutely continuous

map $t \mapsto \rho_t \in \mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$ such that

$$\partial_t \rho_t = -\nabla_x \cdot \left(\rho_t \int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid x)}{\rho_t g(y)} \right) + \alpha \triangle_x \rho_t,$$

then $\{\rho_t\}_{t \geq 0}$ is a Wasserstein gradient flow for E_α; in particular, if g is $\lambda(y)$-concave with $\lambda(y) \geq 0$ for all $y \in \mathbb{R}^{d_\mathbb{Y}}$ such gradient flow is unique.

Proof. Observing that

$$\nabla_x \cdot (\rho_t \nabla_x \log \rho_t) = \nabla_x \cdot \nabla_x \rho_t = \triangle_x \rho_t,$$

where $\triangle f = \sum_i \partial_i^2 f_i$ is the Laplacian, the PDE in (6.10) becomes

$$\partial_t \rho_t = -\nabla_x \cdot \left(\rho_t \int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid x)}{\rho_t g(y)} \right) + \alpha \triangle_x \rho_t.$$

Since E_α is proper, continuous and coercive (see Section 6.3.1) the result follows from Ambrosio et al. (2008, Definition 11.1.1 and Theorem 11.1.4). □

The PDE in Proposition 6.3 is a Fokker-Plank equation, thus we can write a corresponding stochastic differential equation (SDE) for a process X_t whose marginal law at time t is exactly ρ_t. This will be the focus of the next chapter.

Remark 6.1. Even if Assumption 6.1 does not rule out kernels g which are constant with respect to x, it is clear that when that is the case the drift coefficient of the PDE in Proposition 6.3 will be 0 and the PDE reduces to the heat equation

$$\partial_t \rho_t = \alpha \triangle_x \rho_t$$

with corresponding SDE $\mathrm{d}X_t = \sqrt{2\alpha}\mathrm{d}W_t$, where W_t is a d-dimensional Brownian motion Thus neither the PDE nor the corresponding SDE will target a minimiser of (6.13). Indeed, if g is constant with respect to x, $\nabla_x g(y \mid x) \equiv 0$ and (6.18) becomes

$$0 \geq \frac{\lambda(y)}{2} W_2^2(\nu, \rho),$$

which is satisfied only when $\lambda(y) = 0$ for all y.

6.4 Summary

Starting from the penalised Kullback–Leibler divergence E_α in (6.2) we have used a Wasserstein gradient flow approach to obtain a PDE whose solution is a minimiser of E_α in the space of absolutely continuous probability measures with finite second moment. The properties of the functional E_α guarantee that the PDE has a solution, which is unique whenever the kernel g is strongly concave. Linking the minimisation of E_α to a Fokker-Plank PDE leads to the particle system described in the following chapter.

7

A Mean-Field SDE Approach to Fredholm Integral Equations

An extended version of this chapter is in preparation in collaboration with Dr Valentin De Bortoli (Crucinio et al., 2021a).

7.1 Introduction

The Wasserstein gradient flow construction discussed in the previous chapter allowed us to express a minimiser of the regularised Kullback–Leibler divergence

$$E_\alpha(\rho) = \mathrm{KL}(\mu, \rho g) - \alpha \operatorname{ent}(\rho)$$

as a solution of a Fokker-Plank PDE

$$\partial_t \rho_t = -\nabla_x \cdot \left(\rho_t \int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid x)}{\rho_t g(y)} \right) + \alpha \triangle_x \rho_t, \tag{7.1}$$

in the space of absolutely continuous probability measures with finite second moment, $\mathcal{P}_2^{ac}(\mathbb{R}^{d_\mathbb{X}})$, endowed with the 2-Wasserstein distance.

While a number of schemes to solve (7.1) have been proposed in the literature (see e.g., Risken (1996) for a book-length treatment and Xu et al. (2020) for recent developments) we focus on the well-known connection between Fokker-Plank PDEs and stochastic differential equations (SDEs) given by Itô's Lemma (Itô, 1951): if X_t is a process satisfying

$$\mathrm{d}X_t = \int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid X_t)}{\rho_t g(y)} \mathrm{d}t + \sqrt{2\alpha} \mathrm{d}W_t, \quad X_0 \sim \rho_0, \tag{7.2}$$

with W_t a standard $d_\mathbb{X}$-dimensional Brownian motion, then the law ρ_t of X_t satis-

fies (7.1).

The SDE (7.2) is a mean-field SDE (also known as McKean-Vlasov; Lasry and Lions (2007); McKean (1966)), since the drift coefficient involves the distribution ρ_t of X_t. This chapter is devoted to the numerical implementation of (7.2), which, as discussed in Section 2.4, requires introducing both a space discretisation and a time discretisation. The space discretisation of (7.2) leads to an interacting particle system whose distribution at the final time step provides an approximation to the law of X_t in (7.2). Since the regularisation imposed by the entropic penalty depends on α, we discuss possible choices of this parameter in Section 7.3. We conclude with a number of applications to both toy models and realistic problems.

7.2 Numerical Implementation

In order to approximate numerically (7.2) we introduce both a space discretisation and a time discretisation. The space discretisation is necessary since McKean-Vlasov SDEs present a dependence on the distribution ρ_t of X_t (Bossy and Talay, 1997; McKean, 1966) and is obtained by considering N copies $(X_t^{1,N}, ..., X_t^{N,N})$ of (7.2) such that, at $t = 0$, we sample i.i.d. particles $X_0^{i,N} \sim \rho_0$ and then evolve them according to the non-linear SDE

$$\mathrm{d}X_t^{i,N} = \int \mu(\mathrm{d}y) \frac{\nabla_x g(y \mid X_t^{i,N})}{\rho_t^N g(y)} \mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t^i \tag{7.3}$$

where W_t^i for $i = 1, \ldots, N$ are N independent $d_{\mathbb{X}}$-dimensional standard Brownian motions and ρ_t^N is the empirical measure given by the N particles

$$\rho_t^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^{i,N}}.$$

In most applications μ is not known, but the available data are samples drawn from it (Delaigle, 2008; Goldstein et al., 2009; Gostic et al., 2020; Hall et al., 2005; Ma, 2011; Marschner, 2020; Miao et al., 2018; Pensky et al., 2017; Yang et al., 2020). Thus the integral with respect to μ in the drift coefficient of (7.3) cannot be computed analytically, but can be approximated through the sample average

$$\int \mu^M(\mathrm{d}y) \frac{\nabla_x g(y \mid X_t^{i,N})}{\rho_t^N g(y)}, \qquad \text{with} \quad \mu^M = \frac{1}{M} \sum_{j=1}^{M} \delta_{Y^{j,M}}$$

where $Y^{j,M}$ for $j = 1, \dots, M$ are i.i.d. samples from μ. The corresponding non-linear SDE is for $i = 1, \dots, N$

$$\mathrm{d}X_t^{i,N,M} = \int \mu^M(\mathrm{d}y)\, \frac{\nabla_x g(y \mid X_t^{i,N,M})}{\rho_t^{N,M} g(y)} \mathrm{d}t + \sqrt{2\alpha}\mathrm{d}W_t^i \tag{7.4}$$

where

$$\rho_t^{N,M} = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_t^{i,N,M}}. \tag{7.5}$$

The space discretisation of (7.2) introduced above leads to an interacting particle system in which each copy i of (7.4) interacts with the other copies through the empirical average (7.5).

The drift coefficient of (7.4) is not globally Lipschitz, in particular the denominator in

$$b(X_t, \rho_t) := \int \mu(\mathrm{d}y)\, \frac{\nabla_x g(y \mid X_t)}{\rho_t g(y)} \tag{7.6}$$

can be arbitrarily small if ρ_t assigns low probability to regions where g is appreciable. In this setting, the standard Euler scheme (2.14) is unstable and lacks convergence (Hutzenthaler et al., 2012), therefore, we consider the tamed Euler scheme described in Bao et al. (2020) which guarantees non-explosion of the drift term,

$$\begin{aligned} X_{k+1}^{i,N,M} &= X_k^{i,N} + \frac{b(X_k^{i,N,M}, \rho_k^{N,M})}{1 + \Delta t \|b(X_k^{i,N,M}, \rho_k^{N,M})\|_2} \Delta t + \sqrt{2\alpha}\Delta W_k^i, \\ \rho_k^{N,M} &= \frac{1}{N} \sum_{i=1}^{N} \delta_{X_k^{i,N,M}}, \end{aligned} \tag{7.7}$$

where ΔW_k^i are centred Gaussian random variables with variance Δt. This scheme coincides with the higher order Milstein scheme since the diffusion coefficient is constant.

A smooth estimator of the density of $\rho(x)$ is obtained by standard kernel density estimation (Silverman, 1986)

$$\rho^N(x) := \frac{1}{N} \sum_{i=1}^{N} |\Sigma|^{-1/2} \phi\left(\Sigma^{-1/2}(X_k^{i,N,M} - x)\right) \tag{7.8}$$

where ϕ is the density of a $d_{\mathbb{X}}$-dimensional standard Gaussian distribution and Σ

is the bandwidth matrix. This approach has been shown to have good convergence properties in the one-dimensional case for McKean-Vlasov SDEs with Lipschitz drift (Antonelli and Kohatsu-Higa, 2002; Bossy and Talay, 1997).

Remark 7.1 (Stability and Convergence Properties)**.** As discussed above, the lack of Lipschitz continuity of the drift of (7.2) prevents us from using the standard results summarised in Chapter 2 to characterise the solutions of (7.2) and of its corresponding particle system (7.3). In particular, the drift coefficient is not Lipschitz continuous with respect to the measure component and therefore the results established in Dos Reis et al. (2019, 2021) do not apply. Under appropriate assumptions on the tail behaviour of g we conjecture that it would be possible to obtain local Lipschitz continuity for the drift and therefore exploit some of the results established in the literature; alternatively, it would be possible to consider a slight modification of the functional E_α which guarantees that the denominator in (7.6) is bounded below and, as a consequence, Lipschitz continuity of the drift. We further discuss these future work directions in the closing remarks.

7.3 Choice of α

As the parameter $\alpha > 0$ controls the amount of regularisation introduced by the entropic penalty, its value should be chosen to give a good trade-off between the distance from the data distribution μ, $\mathrm{KL}(\mu, \rho g)$ and the smoothness of the solution ρ measured by $\mathrm{ent}(\rho)$. In principle, one could use a generalised cross validation approach (Wahba, 1977) as detailed in Amato and Hughes (1991) for Tikhonov regularisation with entropic penalty; however this approach relies on a deterministic discretisation of μ, g, ρ which is not feasible when only samples from μ are available.

Since the case in which μ is not known but a sample drawn from it is available is the most likely in applications, we propose the following approach to cross-validation: to estimate the value of α we divide the original sample into L subsets (which of course could contain one sample only, in a bootstrap-like framework) and find the value of α which minimises

$$CV(\alpha) = \sum_{l=1}^{L} E_\alpha(\rho^{N,l}) \tag{7.9}$$

where $\rho^{N,l}$ is obtained as in (7.8) using the samples from μ which are not in group l. The value of E_α is approximated by numerical integration over the support of μ using the kernel density estimators for μ.

As an alternative to cross validation, one could use an empirical Bayes approach as advocated in Vidal et al. (2020) in the context of image reconstruction; however this usually requires assuming a parametric form for the solution ρ, e.g. through a variational family (Blei et al., 2017) or normalising flows (Papamakarios et al., 2021).

If some prior information on the smoothness of the solution ρ is known (e.g. its variance), one could chose α so that the entropy of the approximate solution matches the expected smoothness of ρ. If no prior information on the solution is available, the value of E_α for the approximate solution gives some intuition on appropriate values of α:

Example 7.1. Consider the toy Fredholm integral equation

$$\mathcal{N}(y; m, \sigma_\mu^2 := \sigma_g^2 + \sigma_\rho^2) = \int \mathcal{N}(x; m, \sigma_\rho^2)\, \mathcal{N}(y; x, \sigma_g^2) \mathrm{d}x$$

where $\mathcal{N}(x; m, \sigma^2)$ is a Gaussian distribution with mean m and variance σ^2.

Under the assumption that the minimiser ρ is a Gaussian distribution, $\rho_\beta(x) := \mathcal{N}(x; m, \beta)$, the functional $E_\alpha(\rho_\beta)$ can be computed exactly

$$\begin{aligned} E_\alpha(\rho_\beta) &= \frac{1}{2} \log \frac{\beta + \sigma_g^2}{\sigma_\mu^2} + \frac{\sigma_\mu^2}{2(\beta + \sigma_g^2)} - \frac{1}{2} - \alpha \operatorname{ent}(\rho_\beta) \\ \operatorname{ent}(\rho_\beta) &= \frac{1}{2} + \frac{1}{2} \log(2\pi\beta). \end{aligned}$$

For this simple toy example we can obtain the β minimising E_α as a function of α,

$$\beta(\alpha) = \frac{-(\sigma_g^2 - \sigma_\mu^2 - 2\alpha\sigma_g^2) + \sqrt{\sigma_g^4 + \sigma_\mu^4 - 2\sigma_g^2\sigma_\mu^2(1 - 2\alpha)}}{2(1 - \alpha)}; \tag{7.10}$$

clearly when $\alpha = 0$ (no entropy constraint), $\beta = \sigma_\rho^2$, while $\alpha > 1$ give negative variance.

Example 7.1 shows that the value of E_α depends continuously on $\alpha \in [0, 1]$ through the variance $\beta(\alpha)$. Continuous dependence of on α is a desirable property for the functional E_α since it guarantees that the cross-validation function (7.9) is continuous too, meaning that we can approximate $CV(\alpha)$ numerically and hope to find a minimiser.

In general, finding an analytic expression depending only on α for E_α is not possible. However, we can check the sensitivity to the choice of α by approximating the value of E_α through numerical integration; when only samples from μ are available

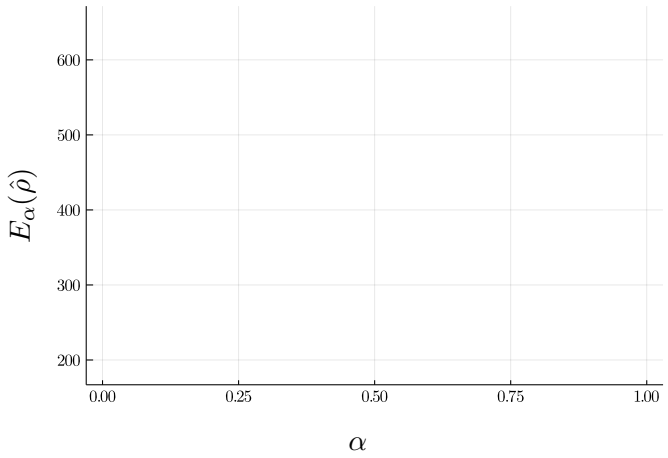


Figure 17: Approximation of E_α (averaged over 100 replications) as a function of α for Example 7.2. The functional depends continuously on α and achieves a minimum at $\alpha \approx 0.08$.

and even an approximate computation of E_α at each iteration is not feasible or too expensive, the entropy can be used as a proxy for E_α, since once a minimiser of E_α has been reached, the entropy will not change.

Example 7.2. Consider the Gaussian mixture model

$$\mu(\mathrm{d}y) = \int \rho(\mathrm{d}x) g(y \mid x)$$

with

$$\begin{aligned} \rho(x) &= \frac{1}{3}\mathcal{N}(x; 0.3, 0.015^2) + \frac{2}{3}\mathcal{N}(x; 0.5, 0.043^2), \\ g(y \mid x) &= \mathcal{N}(y; x, 0.045^2). \\ \mu(y) &= \frac{1}{3}\mathcal{N}(y; 0.3, 0.045^2 + 0.015^2) + \frac{2}{3}\mathcal{N}(y; 0.5, 0.045^2 + 0.043^2) \end{aligned}$$

(where with a slight abuse of notation we denote both a measure and its density with the same symbol) (Ma, 2011). In this case a minimiser of (6.13) cannot be computed analytically, but it is possible to approximate the value of E_α numerically (Figure 17). This shows that E_α is a continuous function of α with a minimum at ≈ 0.08; therefore we expect the value of α chosen by cross validation to be close to this value.

7.4 Examples

We test the performances of the proposed method on a number of examples of Fredholm integral equations (6.12). First, we use the toy Fredholm integral equation in Example 7.1 to explore the influence of the initial distribution ρ_0 on the reconstructions and speed of convergence. This empirical study shows that light tail initial distributions give better performances, as already observed by Antonelli and Kohatsu-Higa (2002); Bossy and Talay (1997) for discretisation of generic McKean-Vlasov SDEs. We also exploit the analytic tractability of this example to compare the reconstructions obtained by implementing (7.7) with the exact minimiser.

We then use the toy model we introduced in Section 5.5.2 and Example 7.2 to compare the method proposed in this chapter with the SMC-EMS algorithm introduced in Chapter 5 and with deconvolution kernel density estimators (DKDE; Delaigle (2008)). We also study two realistic examples of deconvolution: reconstruction of the concentration of the enzyme sucrase in intestinal tissues from noisy measurements and reconstruction of the incidence profile of a disease from the observed number of cases. Finally, we consider applications to medical imaging.

In order to implement (7.7) a number of parameters have to be selected. The initial distribution ρ_0 is chosen in such a way to encapsulate any available information on the solution ρ of (6.12), for example, for deconvolution problems the initial distribution is μ, since we expect the density μ to be a noise-corrupted version of ρ. In most of the examples considered the value of α is selected using the cross validation approach in Section 7.3, occasionally we will use different values to allow comparison with other methods (e.g. Section 7.4.2) or to study the sensitivity of the reconstructions to this parameter (e.g. Section 7.4.1).

As observed in Section 7.2, often μ is only known through a sample and it is necessary to approximate the integral with respect to μ as in (7.4). Under the assumption that a sample from μ is available and following the same considerations in Chapter 5, we propose the following strategy: at each time step approximate the integral with respect to μ using M samples where M is the smallest between the number of particles N chosen for the particle system (7.3) and the total number of samples available from μ; if N is larger than the total number of samples from μ then the whole sample is used at each iteration, if N is smaller we resample without replacement M times from the empirical measure of the sample. This amounts to a maximum cost of $\mathcal{O}(MN)$ per time step. The choice of N largely depends on the dimensionality of the problem, for one-dimensional examples we observe that values between $N = 200$ and $N = 1000$ achieve high enough accuracy to compete with

specialised algorithms (see Section 7.4.2). For image reconstruction problems, the number of particles N should be increased to account for the resolution of the image (see Section 7.4.3).

Similar considerations apply to the choice of the time discretisation step Δt. In particular, this should be chosen taking into account the order of magnitude of the gradient, to give a good trade-off between the Monte Carlo error and the time discretisation error. In practice, we observed good results with Δt between 10^{-1} and 10^{-3}. The number of time steps necessary to give convergence of (7.2) is determined through E_α, once the value of E_α stops decreasing, a minimiser has been reached. When computing E_α is prohibitively expensive we propose to use the entropy of the solution as a proxy.

The kernel density estimator in (7.8) is obtained using isotropic Gaussian kernels and plug-in bandwidths (Wand and Jones, 1994). As in the case of SMC-EMS the particles are *not* independent and therefore the plug-in bandwidth for standard kernel density estimation might provide undersmoothed reconstructions; however, while the SMC-EMS algorithm outputs a weighted particle population from which an indication of the effective sample size is easy to obtain, see Section 5.3, this is not the case for the mean-field SDEs studied in this chapter, the choice of standard bandwidths is not optimal, but common in the literature on approximation of mean-field SDEs (e.g. Bossy and Talay (1997)).

Although the λ-concavity Assumption 6.1-(a) is not satisfied by any of the examples below, we nonetheless observe good results in terms of reconstruction accuracy. We conjecture that local λ-concavity (which is satisfied by all the kernels g considered) is enough to guarantee good performances. Julia code (Bezanson et al., 2017) to reproduce all examples is available online [1].

7.4.1 Analytically Tractable

To analyse the influence of the initial distribution ρ_0 on the reconstructions, we consider the toy Fredholm integral equation

$$\mathcal{N}(y; m, \sigma_\mu^2 := \sigma_g^2 + \sigma_\rho^2) = \int \mathcal{N}(x; m, \sigma_\rho^2)\, \mathcal{N}(y; x, \sigma_g^2) \mathrm{d}x$$

where $\mathcal{N}(x; m, \sigma^2)$ is a Gaussian distribution with mean m and variance σ^2. We set $m = 0.5$, $\sigma_\rho^2 = 0.043^2$, $\sigma_g^2 = 0.045^2$, with this choice $|1 - \int_0^1 \rho(x)\mathrm{d}x| < 10^{-30}$ and take 6 initial distributions (where with a slight abuse of notation we denote both a

[1] Link: `https://github.com/FrancescaCrucinio/entropy_regularisation`

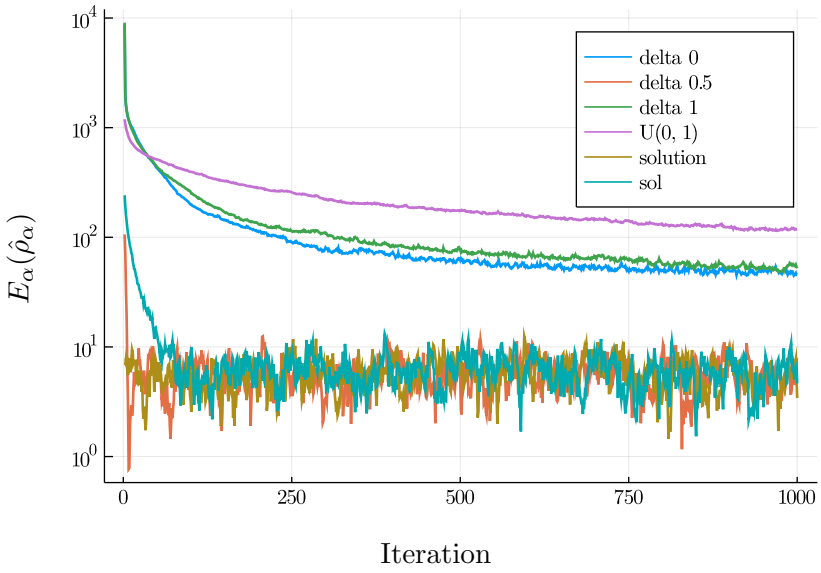


Figure 18: Effect of initial distribution on E_α for the Gaussian toy example with $N = 1000, \Delta t = 10^{-3}, \alpha = 0.05$.

measure and its density w.r.t. Lebesgue with ρ_0) and for each of them we compute the W_2 distance from the solution $\rho_\alpha(x) = \mathcal{N}(x; m, \sigma_\alpha^2)$:

$$W_2(\rho_\alpha, \mathcal{N}(m, \sigma_\alpha^2 + \varepsilon)) \approx \varepsilon \qquad W_2(\rho_\alpha, U(0,1)) = 0.021$$
$$W_2(\rho_\alpha, \delta_{x_0}) = 0.252 \qquad \text{for all } x_0 \in \mathbb{R}$$

Despite having the highest distance from the solution, Figure 18 shows that better results are obtained when ρ_0 is a point mass centred at 0.5 than in the

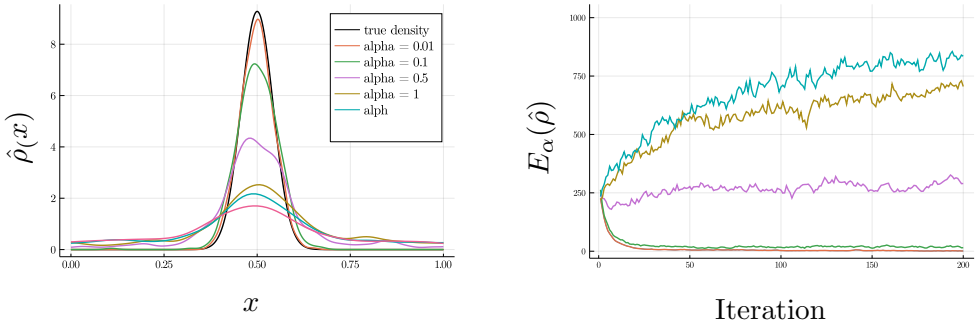


Figure 19: Reconstruction of the solution and corresponding value of E_α for the Gaussian toy example with $\alpha = 0.01, 0.1, 0.5, 1, 1.1.$ and 1.5.

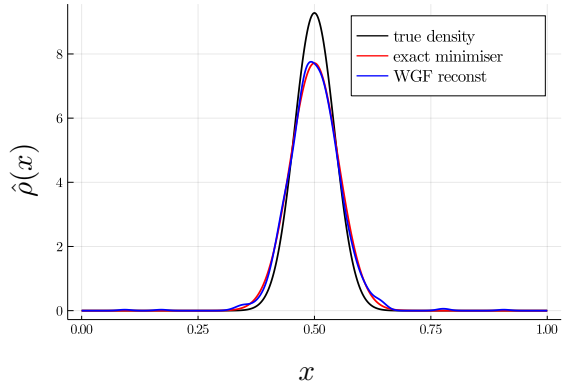


Figure 20: Comparison of the exact minimiser ρ_α for $\alpha = 0.1$ with the corresponding WGF reconstruction with $N = 1000, \Delta t = 10^{-3}$ and 100 iterations.

case $\rho_0 \sim U([0,1])$ and that initial distributions concentrated in the centre of the support of ρ achieve smaller values of E_α. This is more likely an effect of the numerical implementation of McKean-Vlasov SDEs, in fact Antonelli and Kohatsu-Higa (2002); Bossy and Talay (1997) observe that the Euler scheme for McKean-Vlasov SDEs benefits from light tail initial distributions as particles are more easily diffused from the centre of the space rather than moved towards it.

Knowing that for $\alpha > 1$ the functional E_α does not admit a minimiser, we study the behaviour of the numerical reconstruction and of the corresponding value of E_α (Figure 19). While for $\alpha < 1$ the value of E_α decreases with the number of iterations and eventually stabilises, for $\alpha \geq 1$ the value of E_α keeps increasing; this observation can guide the selection of α when exact computation of a minimiser of E_α is not feasible: if for given α the value of E_α increases with the number of iterations, then E_α might not admit a minimiser and smaller values of α should be considered.

For this simple toy model the unique minimiser of E_α can be computed exactly for any given $\alpha \in [0,1]$; by taking $\rho_\alpha(x) = \mathcal{N}(x; m, \beta(\alpha))$ with $\beta(\alpha)$ in (7.10) the functional E_α achieves its minimum. For $\alpha > 1$ the functional does not admit a minimum. We use this toy model to empirically check that once convergence of (7.7) is reached the distribution of the particles approximates the minimiser of E_α (Figure 20).

7.4.2 Indirect Density Estimation

The focus of this section is on deconvolution problems, in this case, $g(y \mid x) \equiv g(y - x)$ and the corresponding integral equation models the task of reconstructing the density of a random variable X from noisy measurements Y. This particu-

lar type of Fredholm integral equations has been widely studied (Delaigle, 2008; Delaigle and Gijbels, 2004; Stefanski and Carroll, 1990) and a specialised class of estimators, deconvolution kernel density estimators (DKDE), achieving optimal rates of convergence exists (Carroll and Hall, 1988).

We compare the proposed method with DKDE on a simulated dataset and on a dataset containing noisy observations of the sucrase concentration in intestinal tissue of 24 patients (see Delaigle (2008) and references therein for model details). Then we consider a particular instance of the deconvolution problem, in which the kernel $g(y-x)$ describes the delay distribution between time of infection and time of death or hospitalisation. Deconvolution techniques have been used to infer ρ in the case of HIV (Becker et al., 1991) and influenza (Goldstein et al., 2009) and have been recently applied to estimate the incidence curve of COVID-19 (Chau et al., 2020; Marschner, 2020; Miller et al., 2020; Wang et al., 2020).

Gaussian Mixture

Consider again Gaussian mixture model described in Example 7.2 and used to compare SMC-EMS with EMS in Chapter 5:

$$\begin{aligned}\rho(x) &= \frac{1}{3}\mathcal{N}(x; 0.3, 0.015^2) + \frac{2}{3}\mathcal{N}(x; 0.5, 0.043^2),\\ g(y \mid x) &= \mathcal{N}(y; x, 0.045^2).\\ \mu(y) &= \frac{1}{3}\mathcal{N}(y; 0.3, 0.045^2 + 0.015^2) + \frac{2}{3}\mathcal{N}(y; 0.5, 0.045^2 + 0.043^2).\end{aligned}$$

We use this example to compare the performances of the mean-field SDE approach studied in this chapter with those of SMC-EMS and of deconvolution kernel density estimators (DKDE). As we did in Chapter 5, we measure the accuracy of the reconstructions through the integrated square error

$$\mathrm{ISE}(\hat{\rho}) = \int (\rho(x) - \hat{\rho}(x))^2 \,\mathrm{d}x,$$

where $\hat{\rho}$ is any estimator of ρ (and we commit the usual abuse of notation of denoting both a measure and its density with the same symbol). Even if μ is known, we assume that we only have available a sample of size 10^3 and we use this sample for both the SMC-EMS estimator and (7.8). The initial distribution is μ for both algorithms and we fix the number of iterations to 100 as we observed convergence in less than 100 iterations for both SMC-EMS and WGF with $\Delta t = 10^{-3}$. The SMC-EMS estimator is obtained using the algorithmic setting described in Chapter 5: the

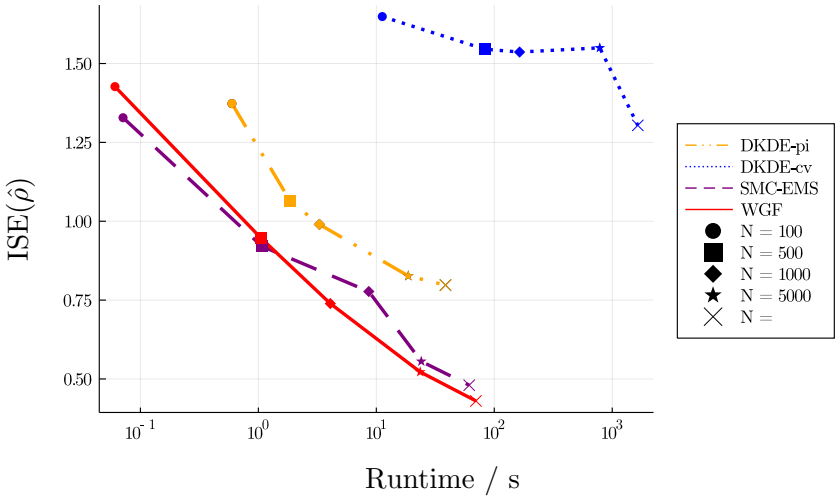


Figure 21: Comparison of WGF with SMC-EMS and DKDE. The number of particles N ranges between 100 and 10,000 and the smoothing/regularisation parameters are chosen so that WGF and SMC-EMS give reconstructions with the same entropy.

smoothing kernels are zero-mean Gaussian distributions with variance ε^2 and the kernel density estimator is obtained using Gaussian kernels with plug-in bandwidth (Silverman, 1986) in which we use the effective sample size instead of the actual number of particles N. For (7.8) we use Gaussian kernels with plug-in bandwidth (Wand and Jones, 1994). To choose the smoothing parameter ε for SMC-EMS and the regularisation parameter α for WGF we fix $\varepsilon = 10^{-3}$ for all N (as this provided good results in the experiments in Chapter 5) and find the corresponding αs giving reconstructions with roughly the same entropy.

We consider different particle sizes (from 100 to 10,000) and compare the reconstruction accuracy with the total runtime of SMC-EMS, WGF and DKDE (Figure 21). In particular, we compare the cost per iteration since we run both algorithms for a fixed number of steps. The computational cost could of course be reduced by considering stopping criteria like that in Section 5.3.1. For small N the SMC-EMS approach performs better than the other algorithms both in terms of runtime and in terms of accuracy. For larger N, accuracy and runtime of SMC-EMS and WGF are similar, with the latter only slightly outperforming the former in terms of accuracy. As already observed in Delaigle and Gijbels (2004) and in Chapter 5, DKDE-cv provides very unstable reconstructions and has the highest runtime for any particle size. The DKDE-pi estimator has runtime comparable to that of SMC-EMS and WGF, but performs worse in terms of accuracy.

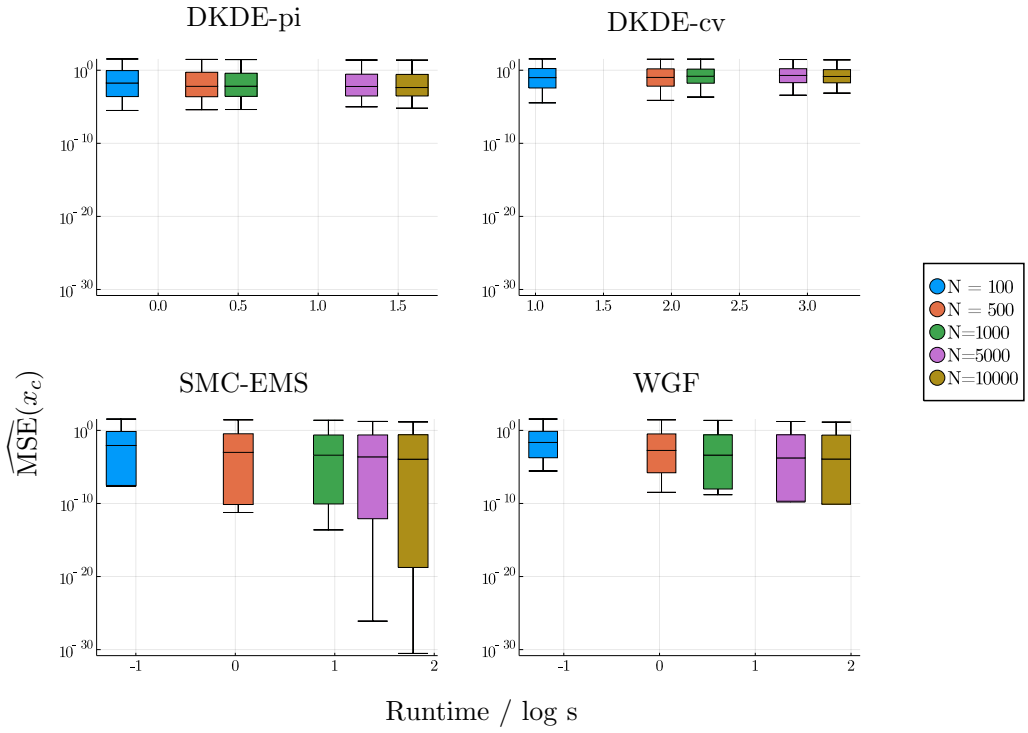


Figure 22: Distribution of MSE as a function of runtime (in log seconds) for WGF, SMC-EMS and DKDE. The number of particles N ranges between 100 and 10,000.

As both SMC-EMS and WGF are regularised versions of the inconsistent maximum likelihood estimator for ρ, it is natural to compare the smoothness of the reconstructions. To characterise the smoothness, we take 100 points x_c in the support of ρ and approximate (with 100 replicates) the mean squared error (MSE)

$$\mathrm{MSE}(x_c) = \mathbb{E}\left[(\rho(x_c) - \hat{\rho}(x_c))^2\right].$$

The distribution of the MSE over the 100 points (Figure 22) shows that DKDE-pi, SMC-EMS and WGF achieve average MSE of the same order (10^{-2}) but the maximum MSE is at least one order of magnitude smaller for SMC and WGF. Generally, SMC-EMS and WGF achieve smaller MSE, showing that the smoothness of the reconstructions provided by these two methods is closer to that of the true density ρ. As the number of particles N increases, the distributions given by SMC-EMS and WGF become left-skewed showing that increasing the number of particles leads to a better exploration of the space. In particular, in the case of SMC-EMS the increase in N forces the particles to explore not only the centre of the distribution but also the tails, which are under-explored for lower particle sizes.

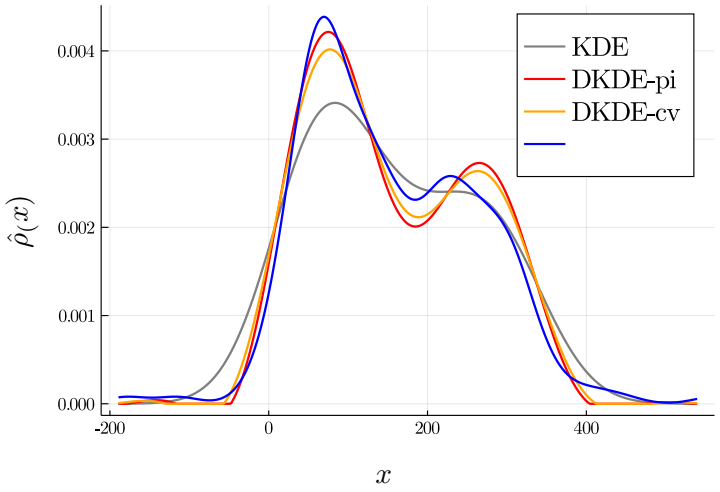


Figure 23: Reconstruction of sucrase concentration for 24 noisy observations. Comparison of kernel density estimator for the noisy observations Y (gray), the deconvolution kernel density estimators for X (red and orange) and WGF reconstruction (blue).

Sucrase Concentration

Next, we consider the one-dimensional density deconvolution problem of Delaigle (2008, Section 4.2) in which the concentration of the enzyme sucrase in intestinal tissues of 24 patients has to be inferred from noisy observations. We reconstruct the density of X from the 24 observations of $Y = X + U$ under the assumption that U follows a Gaussian distribution with variance such that the signal-to-noise ratio is $1/3$ (see Delaigle (2008) and references therein for model details).

We compare the reconstructions obtained by our method with the DKDE and the standard kernel density estimator obtained from the observations of Y. For the WGF implementation we use $N = 200$ particles, $\Delta t = 10^{-1}$, $\alpha = 0.3$ chosen using the cross validation approach in Section 7.3 with $L = 24$ (i.e. leaving one sample out for each repetition) and iterate until convergence ($\approx$ 20,000 iterations). Given the low number of observations from Y (only 24) we set the initial distribution ρ_0 to be a Gaussian density with mean and variance given by the empirical mean and variance of the sample from Y

Figure 23 shows the reconstructed densities for X as well as the kernel density estimator for Y. The three estimators have similar runtimes (between 1 and 10 seconds), and their reconstructions are similar, with DKDEs giving slightly smoother reconstructions; as observed in Delaigle (2008, Section 4.2) the reconstructed den-

sities for X are bimodal and suggest the presence of two groups of patients for which the concentration of sucrase significantly differ. This difference is much less evident in the kernel density estimator for μ, since the convolution with the error distribution g smooths out the two modes.

Epidemiology

To test the mean-field SDE based approach in this context, we use data from the spread of the pandemic influenza in the city of Philadelphia between September and December 1918 (Goldstein et al., 2009). The count of daily deaths and the distribution of delay between infection and death are available through the `R` package `incidental` (Miller et al., 2020).

To obtain a parametric form for g we fit a mixture of Gaussians to the delay data using the expectation maximisation algorithm (`normalmixEM` function in `R`)

$$g(y \mid x) = 0.595\,\mathcal{N}(y - x; 8.63, 2.56^2) + 0.405\,\mathcal{N}(y - x; 15.24, 5.39^2).$$

Although this choice assigns $\approx 10^{-3}$ mass to the negative reals, we found that a mixture of Gaussians fits the observed delay distribution better than other commonly used distributions (e.g. Gamma, log-normal, see Obadia et al. (2012)) and has the additional advantage of having bounded derivative and therefore satisfies Assumption 6.1-(c).

We compare the reconstructions obtained through WGF with the robust incidence deconvolution estimator (RIDE) of Miller et al. (2020) and the Richardson-Lucy (RL) deconvolution described in Goldstein et al. (2009), which corresponds to the expectation maximisation algorithm for Poisson data in (3.5). Figure 24 shows the incidence curves and the reconstructed death counts obtained with the three methods. Following Goldstein et al. (2009), for both the RL algorithm and the WGF approach we set the initial distribution to be the death curve shifted back by nine days. The number of iterations for RL is 200. To implement the reconstruction process via WGF we set $N = M = 500$, $\Delta t = 10^{-1}$ and iterate until convergence (approximately 3,000 steps), the value of $\alpha = 0.009$ is selected using the cross validation approach (7.9) with $L = 5$ groups.

The three methods provide similar shapes for the incidence curve, with the peak of the incidence around day 30 (i.e. early October). The RL algorithm gives the best fit the the observed death counts and has the lowest runtime (< 1 seconds), however the reconstruction of the incidence curve is not smooth and sensitive to noise (Miller et al., 2020). This is a well known weakness of the RL algorithm and

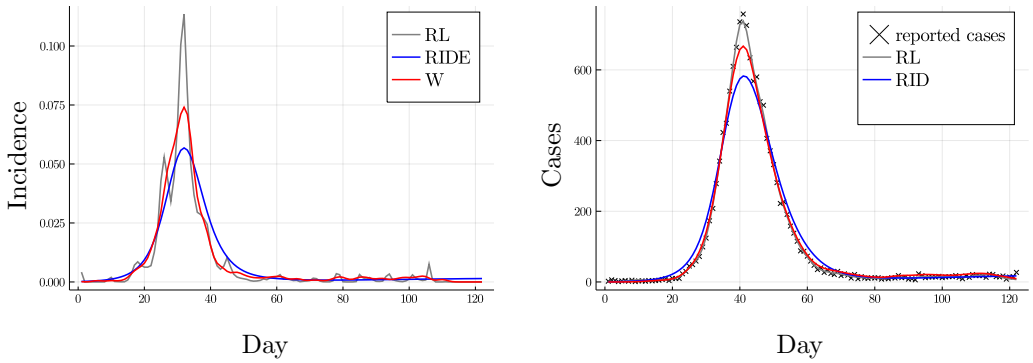


Figure 24: Incidence curve and reconstructed deaths count for the 1918 influenza pandemic in Philadelphia. Comparison of RL (gray), RIDE (red) and WGF (blue).

is related to the inconsistency of the maximum likelihood estimator in the infinite dimensional setting discussed in Chapter 3. Both RIDE and WGF address the lack of smoothness by introducing regularisation, this increases the runtime (≈ 90 seconds for RIDE and 132 seconds for WGF) but provides smoother reconstructions while preserving the shape of the incidence curve. The fit of the reconstructed death counts to the reported one is less accurate, with WGF providing a slightly better fit. This is not unexpected, as the reconstructions provided by WGF are a regularised version of the measure ρ minimising the KL divergence (3.4).

7.4.3 Positron Emission Tomography

As pointed out in Chapter 5, medical imaging is one of the main applications of Fredholm integral equations; in this context the integral equation models the reconstruction of cross-section images of the organ of interest from the noisy measurements provided by PET scanners (Webb, 2017).

PET scanners provide noisy measurements by mapping each point of the organ's cross-section $\rho(x, x)$ onto its radial projection $\mu(\phi, \xi)$ where ϕ denotes an angle between $0°$ and $360°$ and ξ denotes the depth at which the projection is taken (Vardi et al., 1985). We model the alignment between (x, y) and the projections (ϕ, ξ) through a Gaussian kernel with small variance $\sigma^2 \simeq 0.02$

$$g(\phi, \xi \mid x, y) = \frac{\mathcal{N}(x \cos\phi + y \sin\phi - \xi; 0, \sigma^2)}{2\pi}.$$

We consider again the 128-pixels Shepp-Logan phantom reference image (Shepp and Vardi, 1982) and obtain the corresponding data as projections over 128 equally

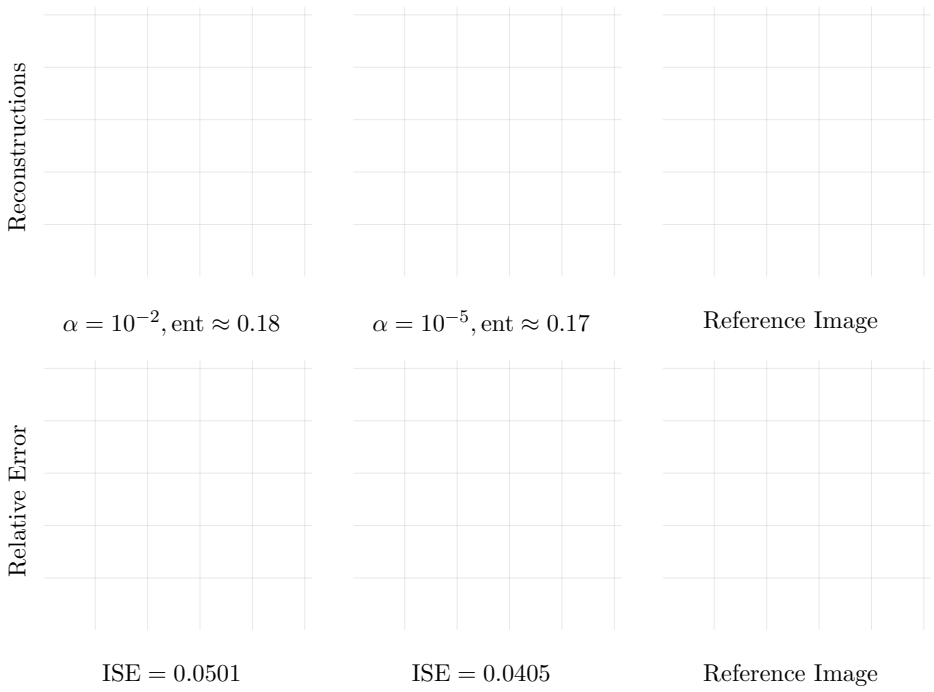


Figure 25: Reconstruction of PET reference image via mean-field SDE with $N = 20,000$ and $\Delta t = 10^{-2}$. Convergence is reached in less than 100 steps.

spaced angles ϕ in $[0, 2\pi]$ and at depths $\xi \in [-92, 92]$.

We implement (7.7) with initial distribution ρ_0 an isotropic Gaussian distribution concentrated at the centre of the image, $N = 20,000$, $\Delta t = 10^{-2}$ and iterate until convergence (measured by approximating E_α and reached after ≈ 100 steps). In Figure 25 we compare the reconstructions obtained by setting $\alpha = 10^{-2}$ to match the entropy of the reconstruction given by SMC-EMS shown in Figure 13 (≈ 0.18) and $\alpha = 10^{-4}$ chosen by cross validation over $L = 5$ repetitions (which gives entropy$\approx$ 0.17). The entropy of the reconstructed images is higher than that of the original image (≈ 0.12); larger values of α result in higher values of the entropy while smaller values of α do not lead to entropy smaller than 0.17. Even if the two reconstructions have similar entropies, the accuracy of the reconstruction obtained with $\alpha = 10^{-4}$ chosen by cross-validation is higher than that obtained with $\alpha = 10^{-2}$. In particular, the former more accurately identifies the profile of the skull (see the relative errors in Figure 13).

7.5 Summary

Starting from the gradient flow PDE obtained in Chapter 6 whose solution is a minimiser of the penalised Kullback–Leibler divergence (6.13), we exploited standard results in stochastic differential calculus to obtain the SDE associated with it. This is SDE is of the mean-field type and its numerical approximation is obtained by considering a population of particles interacting in a mean-field sense.

We tested the performances of the proposed method on a number of toy and real-data examples showing that this method has comparable performances to standard and novel estimators for density deconvolution and density reconstruction from epidemiological data, while enjoying wider applicability. In particular, we showed that the algorithm proposed in this chapter outperforms the SMC-EMS algorithm introduced in Chapter 5; however SMC-EMS requires less stringent assumptions on the function g, and, as mentioned in Remark 6.1 can be implemented also when g admits a Uniform distribution as marginal, a case which is common in deblurring problems.

A

Proofs of the Results for Continuous EMS

A.1 Auxiliary Results for Analyticity

This Appendix is concerned with the derivation of the Fréchet derivatives of order k of the EMS map. First we derive the first order Fréchet derivative in Lemma 4.1 then we obtain the formula for the derivatives of order k in Lemma 4.2 inductively.

In the following arguments, we make extensive use of Fubini's Theorem, whose applicability is granted by the boundedness of all quantities involved. Additionally, we exploit the fact that β-norm defined in Section 4.2.1 metrises weak convergence and that for any bounded measurable function $\varphi \in \mathcal{B}_b(\mathbb{X})$

$$|[\eta_n - \eta](\varphi)| \leq \|\varphi\|_\infty \left| \int [\eta_n - \eta](\mathrm{d}x) \right| \leq \|\varphi\|_\infty \beta(\eta_n, \eta), \tag{A.1}$$

where the last inequality follows from the fact that the unit function $\psi : x \mapsto 1$ is bounded and Lipschitz on $\mathbb{X}$ with Lipschitz norm $\|\psi\|_{BL} = 1$.

Proof of Lemma 4.1. Recall the expression in Lemma 4.1

$$\begin{aligned} D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu(\mathrm{d}x') := & \int_{\mathbb{X}} \nu(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \mathrm{d}x' \\ & - \int_{\mathbb{X}} \eta(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^2} \nu\left(g(y \mid \cdot)\right) \mathrm{d}x', \end{aligned}$$

and consider the definition of Fréchet derivative: $D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}$ is a linear operator such that for a sequence $\{\eta_n\}_{n \geq 1} \subset \mathcal{M}^+(\mathbb{X})$ with $\beta(\eta_n, \eta) \to 0$ as $n \to \infty$

$$\lim_{\beta(\eta_n, \eta) \to 0} \frac{\beta\left(\mathrm{F}_{\mathrm{EMS}}\, \eta_n, \mathrm{F}_{\mathrm{EMS}}\, \eta + D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)\right)}{\beta(\eta_n, \eta)} = 0.$$

Using the definition of β, take $\varphi \in \mathrm{BL}(\mathbb{X})$ with $\|\varphi\|_{BL} \leq 1$ and consider

$$\begin{aligned}
&\left|\mathrm{F}_{\mathrm{EMS}}\,\eta_n(\varphi) - \mathrm{F}_{\mathrm{EMS}}\,\eta(\varphi) - D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)(\varphi)\right| \\
&\quad = \left| \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} \eta_n(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta_n\left(g(y \mid \cdot)\right)} \mathrm{d}x' \right. \\
&\quad - \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} \eta(\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \mathrm{d}x' \\
&\quad - \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} [\eta_n - \eta](\mathrm{d}x) K(x, x') \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \mathrm{d}x' \\
&\quad \left. + \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} K(x, x') \eta(\mathrm{d}x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^2} [\eta_n - \eta]\left(g(y \mid \cdot)\right) \mathrm{d}x' \right|.
\end{aligned}$$

Exploiting the boundedness of all functions involved and Fubini's Theorem we obtain

$$\begin{aligned}
&\left|\mathrm{F}_{\mathrm{EMS}}\,\eta_n(\varphi) - \mathrm{F}_{\mathrm{EMS}}\,\eta(\varphi) - D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)(\varphi)\right| \\
&\quad = \left| \int_{\mathbb{X}} \int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y) \times \right. \\
&\quad \left. \left[\frac{\eta_n(\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{\eta_n(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} + \eta(\mathrm{d}x) \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \right] \right|.
\end{aligned} \tag{A.2}$$

For the expression between square brackets the following identity holds

$$\begin{aligned}
&\frac{\eta_n(\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{\eta_n(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} + \eta(\mathrm{d}x) \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \\
&\quad = [\eta_n - \eta](\mathrm{d}x) \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{1}{\eta\left(g(y \mid \cdot)\right)} \right] \\
&\quad + \eta(\mathrm{d}x) \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{1}{\eta\left(g(y \mid \cdot)\right)} + \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \right].
\end{aligned} \tag{A.3}$$

Then, consider the Taylor expansion of $u \mapsto 1/u$ as $u \to u_0 \neq 0$

$$\frac{1}{u} = \frac{1}{u_0} - \frac{1}{u_0^2}(u - u_0) + \frac{1}{u_0^3}(u - u_0)^2 + o\left((u - u_0)^2\right);$$

if $|[\eta_n - \eta]\left(g(y \mid \cdot)\right)| \to 0$ we have the following approximation of the expression between square brackets

$$\begin{aligned}
&\frac{1}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{1}{\eta\left(g(y \mid \cdot)\right)} \\
&\quad = -\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} + \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2}{\eta\left(g(y \mid \cdot)\right)^3} + o\left(\left([\eta_n - \eta]\left(g(y \mid \cdot)\right)\right)^2\right).
\end{aligned}$$

Because $g(y \mid \cdot) \in C_b(\mathbb{X})$ for all fixed $y \in \mathbb{Y}$, (A.1) gives $|[\eta_n - \eta](g(y \mid \cdot))| \leq m_g \beta(\eta_n, \eta)$, and, if $\beta(\eta_n, \eta) \to 0$, we have

$$\frac{1}{\eta_n(g(y \mid \cdot))} = \frac{1}{\eta(g(y \mid \cdot))} - \frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2} + \frac{[\eta_n - \eta](g(y \mid \cdot))^2}{\eta(g(y \mid \cdot))^3} + o\left(\beta(\eta_n, \eta)^2\right).$$

Plugging the above into (A.3) we obtain

$$\begin{aligned}
&\frac{\eta_n(\mathrm{d}x)}{\eta_n(g(y \mid \cdot))} - \frac{\eta_n(\mathrm{d}x)}{\eta(g(y \mid \cdot))} + \eta(\mathrm{d}x)\frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2} \\
&\qquad = -[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2} + o\left(\beta(\eta_n, \eta)^2\right) \\
&\qquad + \eta_n(\mathrm{d}x)\left[\frac{[\eta_n - \eta](g(y \mid \cdot))^2}{\eta(g(y \mid \cdot))^3} + o\left(\beta(\eta_n, \eta)^2\right)\right].
\end{aligned}$$

Going back to (A.2), using the triangle inequality we find that

$$\begin{aligned}
&\left|\mathrm{F}_{\mathrm{EMS}}\,\eta_n(\varphi) - \mathrm{F}_{\mathrm{EMS}}\,\eta(\varphi) - D^{(1)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)(\varphi)\right| \\
&\quad = \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y)\left[-[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2}\right.\right. \\
&\quad \left.\left. + \eta_n(\mathrm{d}x)\left(\frac{[\eta_n - \eta](g(y \mid \cdot))^2}{\eta(g(y \mid \cdot))^3} + o\left(\beta(\eta_n, \eta)^2\right)\right)\right]\right| \\
&\quad \leq \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y)[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2}\right| \\
&\quad + \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y)\eta_n(\mathrm{d}x)\left(\frac{[\eta_n - \eta](g(y \mid \cdot))^2}{\eta(g(y \mid \cdot))^3} + o\left(\beta(\eta_n, \eta)^2\right)\right)\right|.
\end{aligned}$$

We will consider the two terms separately. For the first one,

$$\begin{aligned}
&\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y)[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2}\right| \\
&\qquad = \left|\int_{\mathbb{Y}} \frac{[\eta_n - \eta](g(y \mid \cdot))}{\eta(g(y \mid \cdot))^2} h(\mathrm{d}y)\int_{\mathbb{X}} K\varphi(x) g(y \mid x)[\eta_n - \eta](\mathrm{d}x)\right| \\
&\qquad \leq \int_{\mathbb{Y}} \frac{|[\eta_n - \eta](g(y \mid \cdot))|}{\eta(g(y \mid \cdot))^2} h(\mathrm{d}y)\left|\int_{\mathbb{X}} K\varphi(x) g(y \mid x)[\eta_n - \eta](\mathrm{d}x)\right| \\
&\qquad \leq \frac{m_g^2}{\eta(\mathbb{X})^2}\int_{\mathbb{Y}} |[\eta_n - \eta](g(y \mid \cdot))|\, h(\mathrm{d}y)\left|\int_{\mathbb{X}} K\varphi(x) g(y \mid x)[\eta_n - \eta](\mathrm{d}x)\right|
\end{aligned}$$

where the first equality is a consequence of the boundedness of φ, g and Fubini's Theorem and $\eta(\mathbb{X}) > 0$ since η has nonzero mass. Using the boundedness of $K\varphi$

and (A.1), we obtain

$$\left|\int_{\mathbb{X}} K\varphi(x)g(y \mid x)[\eta_n - \eta](\mathrm{d}x)\right| \leq \|\varphi\|_\infty m_g \beta(\eta_n, \eta)$$

for every fixed y. Then we have that the first term in bounded by

$$\begin{aligned}\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x)g(y \mid x)h(\mathrm{d}y)[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2}\right| \\ \leq \frac{m_g^4\|\varphi\|_\infty}{\eta(\mathbb{X})^2}\beta(\eta_n, \eta)^2.\end{aligned}$$

Similarly, for the second term, using the boundedness of g and $\|K\varphi\|_\infty \leq \|\varphi\|_\infty \leq 1$ and recalling that η, η_n are finite measures with nonzero mass we obtain

$$\begin{aligned}&\left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x)g(y \mid x)h(\mathrm{d}y)\eta_n(\mathrm{d}x)\left(\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2}{\eta\left(g(y \mid \cdot)\right)^3} + o\left(\beta(\eta_n, \eta)^2\right)\right)\right| \\ &\quad\leq \|K\varphi\|_\infty \int_{\mathbb{X}}\int_{\mathbb{Y}} g(y \mid x)h(\mathrm{d}y)\eta_n(\mathrm{d}x)\left|\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2}{\eta\left(g(y \mid \cdot)\right)^3} + o\left(\beta(\eta_n, \eta)^2\right)\right| \\ &\quad\leq \|\varphi\|_\infty \int_{\mathbb{Y}} h(\mathrm{d}y)\eta_n\left(g(y \mid \cdot)\right)\left|\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2}{\eta\left(g(y \mid \cdot)\right)^3} + o\left(\beta(\eta_n, \eta)^2\right)\right| \\ &\quad\leq \|\varphi\|_\infty m_g^4 \frac{\eta_n(\mathbb{X})}{\eta(\mathbb{X})^3}\int_{\mathbb{Y}} h(\mathrm{d}y)\left|[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2 + o\left(\beta(\eta_n, \eta)^2\right)\right| \\ &\quad\leq \|\varphi\|_\infty m_g^4 \frac{\eta_n(\mathbb{X})}{\eta(\mathbb{X})^3}\left[\beta(\eta_n, \eta)^2 + o\left(\beta(\eta_n, \eta)^2\right)\right].\end{aligned}$$

Putting the two terms together we obtain

$$\begin{aligned}&\frac{\beta\left(\mathrm{F}_{\mathrm{EMS}}\,\eta_n, \mathrm{F}_{\mathrm{EMS}}\,\eta - D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)\right)}{\beta(\eta_n, \eta)} \\ &\quad\leq \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x)g(y \mid x)h(\mathrm{d}y)[\eta_n - \eta](\mathrm{d}x)\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2}\right| \\ &\quad+ \left|\int_{\mathbb{X}}\int_{\mathbb{Y}} K\varphi(x)g(y \mid x)h(\mathrm{d}y)\eta_n(\mathrm{d}x)\left(\frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)^2}{\eta\left(g(y \mid \cdot)\right)^3} + o\left(\beta(\eta_n, \eta)^2\right)\right)\right| \\ &\quad\leq \frac{m_g^4\|\varphi\|_\infty}{\eta(\mathbb{X})^2}\beta(\eta_n, \eta) + \|\varphi\|_\infty m_g^4\frac{\eta_n(\mathbb{X})}{\eta(\mathbb{X})^3}\left[\beta(\eta_n, \eta) + o\left(\beta(\eta_n, \eta)\right)\right]\end{aligned}$$

which tends to 0 as $\beta(\eta_n, \eta) \to 0$, showing that $D_\eta^{(1)}\,\mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)$ is the Fréchet derivative of order one of $\mathrm{F}_{\mathrm{EMS}}$ at η.

To show that, for fixed $\eta \in \mathcal{M}^+(\mathbb{X})$, $D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}$ is bounded, take a bounded and Lipschitz test function $\varphi \in \mathrm{BL}(\mathbb{X})$ with $\|\varphi\|_{BL} \leq 1$ and consider

$$
\begin{aligned}
&|D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu(\varphi)| \\
&\qquad = \left| \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} K(x, x') \nu(\mathrm{d}x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \mathrm{d}x' \right. \\
&\qquad \left. - \int_{\mathbb{X}} \varphi(x') \int_{\mathbb{X}} K(x, x') \eta(\mathrm{d}x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^2} \nu\left(g(y \mid \cdot)\right) \mathrm{d}x' \right| \\
&\qquad \leq \left| \int_{\mathbb{X}} \int_{\mathbb{Y}} K\varphi(x) \nu(\mathrm{d}x) \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)} \right| \\
&\qquad + \left| \int_{\mathbb{X}} \int_{\mathbb{Y}} K\varphi(x) \eta(\mathrm{d}x) \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^2} \nu\left(g(y \mid \cdot)\right) \right|
\end{aligned}
$$

where we have used Fubini's Theorem and the boundedness of g, φ. Using the fact that g is bounded by m_g and $\|K\varphi\|_\infty \leq \|\varphi\|_\infty \leq 1$ we have

$$
\begin{aligned}
|D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu(\varphi)| &\leq \frac{m_g^2}{\eta(\mathbb{X})} \left| \int_{\mathbb{X}} K\varphi(x) \nu(\mathrm{d}x) \right| \\
&\quad + \|K\varphi\|_\infty \int_{\mathbb{Y}} \left| \frac{\int_{\mathbb{X}} \eta(\mathrm{d}x) g(y \mid x)}{\eta\left(g(y \mid \cdot)\right)^2} h(\mathrm{d}y) \nu\left(g(y \mid \cdot)\right) \right| \\
&\leq 2 \frac{m_g^2}{\eta(\mathbb{X})} \|K\varphi\|_\infty \left| \int_{\mathbb{X}} \nu(\mathrm{d}x) \right| \\
&\leq 2 \frac{m_g^2}{\eta(\mathbb{X})} \beta(\nu),
\end{aligned}
$$

where the last inequality follows from (A.1). □

The derivation of the Fréchet derivatives of higher order is carried out by induction and follows the same structure of the proof of Lemma 4.1. In particular, we combine the Taylor expansion of $u \mapsto 1/u^m$ as $u \to u_0 \neq 0$

$$
\frac{1}{u^m} = \frac{1}{u_0^m} - m \frac{1}{u_0^{m+1}} (u - u_0) + o\left((u - u_0)\right)
$$

for all $m \in \mathbb{N}$ and the fact that for fixed y, $|[\eta_n - \eta]\left(g(y \mid \cdot)\right)| \leq m_g \beta(\eta_n, \eta)$ as shown in (A.1), to obtain

$$
\frac{1}{\eta_n\left(g(y \mid \cdot)\right)^m} - \frac{1}{\eta\left(g(y \mid \cdot)\right)^m} + m \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^{m+1}} = o\left(\beta(\eta_n, \eta)\right) \tag{A.4}
$$

for every sequence $\{\eta_n\}$ such that $\beta(\eta_n, \eta) \to 0$ as $n \to \infty$, and use this Taylor

expansion as approximation as we did in Lemma 4.1.

The proof proceeds by induction on the order k of the derivative. In particular, we show that the Fréchet derivative of $\mathrm{F}_{\mathrm{EMS}}$ at $\eta \in \mathcal{M}^+(\mathbb{X})$ of order $k \geq 2$ satisfies

$$\begin{aligned} D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}(\nu, \ldots, \nu, \mu) = (-1)^{k+1}(k-1)!(k-1) \int_{\mathbb{X}} \nu(\mathrm{d}x) K(x, \cdot) \\ \times \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)^k} \mu\left(g(y \mid \cdot)\right) \nu\left(g(y|\cdot)\right)^{k-2} \\ + (-1)^{k+1}(k-1)! \int_{\mathbb{X}} \mu(\mathrm{d}x) K(x, \cdot) \int_{\mathbb{Y}} \frac{g(y|x) h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)^k} \nu\left(g(y|\cdot)\right)^{k-1} \\ + (-1)^k k! \int_{\mathbb{X}} \left[\eta + \mu + (k-2)\nu\right](\mathrm{d}x) K(x, \cdot) \\ \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y|\cdot)\right)^{k+1}} \nu\left(g(y|\cdot)\right)^{k-1} \mu\left(g(y|\cdot)\right), \end{aligned} \tag{A.5}$$

and then obtain the expression in Lemma 4.2 setting $\mu = \nu$ in the above expression.

Proof of Lemma 4.2. Let $k = 2$. For $D_\eta^{(2)} \mathrm{F}_{\mathrm{EMS}}$, we have to show that for $\{\eta_n\}$ with $\beta(\eta_n, \eta) \to 0$ as $n \to \infty$

$$\lim_{\beta(\eta_n, \eta) \to 0} \frac{\| D_{\eta_n}^{(1)} \mathrm{F}_{\mathrm{EMS}} - D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}} - D_\eta^{(2)} \mathrm{F}_{\mathrm{EMS}}(\cdot, \eta_n - \eta) \|_{op}}{\beta(\eta_n, \eta)} = 0. \tag{A.6}$$

Because $D_\eta^{(2)} \mathrm{F}_{\mathrm{EMS}}$ is a bilinear operator, we have that the operator norm in (A.6) is equal to the supremum over $\{v \in \mathcal{M}(\mathbb{X}), \beta(\nu) = 1\}$ of

$$\beta\left(D_{\eta_n}^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu, D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu - D_\eta^{(2)} \mathrm{F}_{\mathrm{EMS}}(\nu, \eta_n - \eta) \right);$$

therefore we will focus on bounding the β norm in the numerator of the fraction above. As for (A.2), take $\varphi \in \mathrm{BL}(\mathbb{X})$ with $\|\varphi\|_{BL} \leq 1$ and consider

$$\begin{aligned} \left| \int_{\mathbb{X}} \varphi(x') \left[D_{\eta_n}^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu - D_\eta^{(1)} \mathrm{F}_{\mathrm{EMS}}\, \nu - D_\eta^{(2)} \mathrm{F}_{\mathrm{EMS}}(\nu, \eta_n - \eta) \right] (\mathrm{d}x') \right| \\ = \left| \int_{\mathbb{X}} \int_{\mathbb{Y}} K\varphi(x) g(y \mid x) h(\mathrm{d}y) \left[\frac{\nu(\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{\eta_n(\mathrm{d}x) \nu\left(g(y \mid x)\right)}{\eta_n\left(g(y \mid \cdot)\right)^2} \right. \right. \\ - \frac{\nu(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} + \frac{\eta(\mathrm{d}x) \nu\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} + \frac{\nu(\mathrm{d}x)[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \\ \left. \left. + \frac{[\eta_n - \eta](\mathrm{d}x) \nu\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} - 2 \frac{\eta_n(\mathrm{d}x)[\eta_n - \eta]\left(g(y \mid \cdot)\right) \nu\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^3} \right] \right|. \end{aligned}$$

Applying the triangle inequality gives

$$\begin{aligned}
&\left|\int_{\mathbb{X}} \varphi(x') \left[D^{(1)}_{\eta_n} \mathrm{F}_{\mathrm{EMS}}\, \nu - D^{(1)}_{\eta} \mathrm{F}_{\mathrm{EMS}}\, \nu - D^{(2)}_{\eta} \mathrm{F}_{\mathrm{EMS}}(\nu, \eta_n - \eta)\right]\ (\mathrm{d}x')\right| \\
&\quad \leq \left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) \nu(\mathrm{d}x) \right. \\
&\quad \left. \times \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{1}{\eta\left(g(y \mid \cdot)\right)} + \frac{\left[\eta_n - \eta\right]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \right] \right| \\
&\quad + \left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) \nu\left(g(y \mid \cdot)\right) \left[-\frac{\eta_n(\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)^2} + \frac{\eta(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^2} \right. \right. \\
&\quad \left. \left. + \frac{[\eta_n - \eta](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^2} - 2\frac{\eta_n(\mathrm{d}x)[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^3} \right] \right| .
\end{aligned}$$

Using (A.4) with $m = 1$ and the boundedness of g and $\|K\varphi\|_\infty \leq \|\varphi\|_\infty \leq 1$ we find that for the first term

$$\begin{aligned}
&\left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) \nu(\mathrm{d}x) \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)} - \frac{1}{\eta\left(g(y \mid \cdot)\right)} + \frac{\left[\eta_n - \eta\right]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^2} \right] \right| \\
&\qquad = o\left(\beta(\eta, \eta_n)\right).
\end{aligned}$$

Similarly, using (A.4) with $m = 2$ for the second term gives

$$\begin{aligned}
&\left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) \nu\left(g(y \mid \cdot)\right) \left[-\frac{\eta_n(\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)^2} + \frac{\eta(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^2} \right. \right. \\
&\qquad \left. \left. + \frac{[\eta_n - \eta](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^2} - 2\frac{\eta_n(\mathrm{d}x)[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^3} \right] \right| \\
&\qquad = \left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) \nu\left(g(y \mid \cdot)\right) \eta_n(\mathrm{d}x) \right. \\
&\qquad \left. \times \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)^2} - \frac{1}{\eta\left(g(y \mid \cdot)\right)^2} + 2\frac{\left[\eta_n - \eta\right]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^3} \right] \right| \\
&\qquad = o\left(\beta(\eta, \eta_n)\right).
\end{aligned}$$

Plugging the above into (A.6) we obtain

$$\frac{\|D^{(1)}_{\eta_n} \mathrm{F}_{\mathrm{EMS}} - D^{(1)}_{\eta} \mathrm{F}_{\mathrm{EMS}} - D^{(2)}_{\eta} \mathrm{F}_{\mathrm{EMS}}(\eta_n - \eta)\|_{op}}{\beta(\eta, \eta_n)} = o(1) \to 0$$

as $n \to \infty$, showing that for $k = 2$ the expression in (A.5) is the Fréchet derivative of $\mathrm{F}_{\mathrm{EMS}}$. To show that $D^{(2)}_{\eta} \mathrm{F}_{\mathrm{EMS}}\, \nu^2$ satisfies (4.12), take $\mu = \nu$ in $D^{(2)}_{\eta} \mathrm{F}_{\mathrm{EMS}}(\nu, \mu)$.

Then, assume (4.12) is true for k. To show that $D^{(k+1)}_{\eta} \mathrm{F}_{\mathrm{EMS}}$ is as in (A.5) take

$\{\eta_n\}_{n\geq 1}$ with $\beta(\eta,\eta_n)\to 0$ as $n\to\infty$ and consider

$$\lim_{\beta(\eta,\eta_n)\to 0}\frac{\|D^{(k)}_{\eta_n}\,\mathrm{F}_{\mathrm{EMS}}-D^{(k)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}-D^{(k+1)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}(\cdots,\eta_n-\eta)\|_{op}}{\beta(\eta,\eta_n)}.$$

Using again the definition of operator norm for linear maps we only need to bound

$$\beta\left(D^{(k)}_{\eta_n}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k, D^{(k)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k - D^{(k+1)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}(\nu^{k-1},\eta_n-\eta)\right)$$

for $\{\nu\in\mathcal{M}(\mathbb{X}),\beta(\nu)=1\}$. Take $\varphi\in\mathrm{BL}(\mathbb{X})$ with $\|\varphi\|_{BL}\leq 1$ and consider

$$\begin{aligned}
&\left|\int_{\mathbb{X}}\varphi(x')\left[D^{(k)}_{\eta_n}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k - D^{(k)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k - D^{(k+1)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}(\nu,\cdots,\nu,\eta_n-\eta)\right](\mathrm{d}x')\right| \\
&\quad=\left|\int_{\mathbb{X}}K\varphi(x)\int_{\mathbb{Y}}g(y\mid x)h(\mathrm{d}y)\right. \\
&\quad\left[(-1)^{k+1}k!\frac{\nu(\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k-1}}{\eta_n\left(g(y\mid\cdot)\right)^{k}}+(-1)^{k}k!\frac{[\eta_n+(k-1)\nu](\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k}}{\eta_n\left(g(y\mid\cdot)\right)^{k+1}}\right. \\
&\quad-(-1)^{k+1}k!\frac{\nu(\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k-1}}{\eta\left(g(y\mid\cdot)\right)^{k}}-(-1)^{k}k!\frac{[\eta+(k-1)\nu](\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k}}{\eta\left(g(y\mid\cdot)\right)^{k+1}} \\
&\quad-(-1)^{k+2}k!\nu\left(g(y\mid\cdot)\right)^{k-1}\left(\frac{k\nu(\mathrm{d}x)[\eta_n-\eta]\left(g(y\mid\cdot)\right)}{\eta\left(g(y\mid\cdot)\right)^{k+1}}-\frac{[\eta_n-\eta](\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)}{\eta\left(g(y\mid\cdot)\right)^{k+1}}\right) \\
&\quad\left.-(-1)^{k+1}(k+1)!\frac{[\eta_n+(k-1)\nu](\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k}[\eta_n-\eta]\left(g(y\mid\cdot)\right)}{\eta\left(g(y\mid\cdot)\right)^{k+2}}\right]\Bigg|.
\end{aligned}$$

Applying the triangle inequality gives

$$\begin{aligned}
&\left|\int_{\mathbb{X}}\varphi(x')\left[D^{(k)}_{\eta_n}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k - D^{(k)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}\,\nu^k - D^{(k+1)}_{\eta}\,\mathrm{F}_{\mathrm{EMS}}(\nu,\cdots,\nu,\eta_n-\eta)\right](\mathrm{d}x')\right| \\
&\quad\leq\left|\int_{\mathbb{X}}K\varphi(x)\int_{\mathbb{Y}}g(y\mid x)h(\mathrm{d}y)(-1)^{k+1}k!\nu(\mathrm{d}x)\nu\left(g(y\mid\cdot)\right)^{k-1}\right. \\
&\quad\left.\left[\frac{1}{\eta_n\left(g(y\mid\cdot)\right)^{k}}-\frac{1}{\eta\left(g(y\mid\cdot)\right)^{k}}+k\frac{[\eta_n-\eta]\left(g(y\mid\cdot)\right)}{\eta\left(g(y\mid\cdot)\right)^{k+1}}\right]\right| \\
&\quad+\left|\int_{\mathbb{X}}K\varphi(x)\int_{\mathbb{Y}}g(y\mid x)h(\mathrm{d}y)(-1)^{k}k!\nu\left(g(y\mid\cdot)\right)^{k}\right. \\
&\quad\left[\frac{[\eta_n+(k-1)\nu](\mathrm{d}x)}{\eta_n\left(g(y\mid\cdot)\right)^{k+1}}-\frac{[\eta+(k-1)\nu](\mathrm{d}x)}{\eta\left(g(y\mid\cdot)\right)^{k+1}}-\frac{[\eta_n-\eta](\mathrm{d}x)}{\eta\left(g(y\mid\cdot)\right)^{k+1}}\right. \\
&\quad\left.\left.+(k+1)\frac{[\eta_n+(k-1)\nu](\mathrm{d}x)[\eta_n-\eta]\left(g(y\mid\cdot)\right)}{\eta\left(g(y\mid\cdot)\right)^{k+2}}\right]\right|.
\end{aligned}$$

We treat the two terms separately as we did for $k = 2$. Using (A.4) with $m = k$ and the boundedness of g and $\|K\varphi\|_\infty \leq \|\varphi\|_\infty \leq 1$ we find that the first term

$$\begin{aligned}
&\left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) (-1)^{k+1} k! \nu(\mathrm{d}x) \nu\left(g(y \mid \cdot)\right)^{k-1} \right. \\
&\qquad \left. \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)^k} - \frac{1}{\eta\left(g(y \mid \cdot)\right)^k} + k \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \right] \right| \\
&\qquad = \left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) k! \nu(\mathrm{d}x) \nu\left(g(y \mid \cdot)\right)^{k-1} \right. \\
&\qquad \left. \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)^k} - \frac{1}{\eta\left(g(y \mid \cdot)\right)^k} + k \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \right] \right| \\
&\qquad = o\left(\beta(\eta, \eta_n)\right).
\end{aligned}$$

Using (A.4) with $m = k + 1$ we obtain the same result for the second term

$$\begin{aligned}
&\left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) (-1)^{k} k! \nu\left(g(y \mid \cdot)\right)^{k} \right. \\
&\qquad \left[\frac{[\eta_n + (k-1)\nu](\mathrm{d}x)}{\eta_n\left(g(y \mid \cdot)\right)^{k+1}} - \frac{[\eta + (k-1)\nu](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} - \frac{[\eta_n - \eta](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \right. \\
&\qquad \left. \left. + (k+1) \frac{[\eta_n + (k-1)\nu](\mathrm{d}x)[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^{k+2}} \right] \right| \\
&\qquad \left| \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} g(y \mid x) h(\mathrm{d}y) k! \nu\left(g(y \mid \cdot)\right)^{k} [\eta_n + (k-1)\nu](\mathrm{d}x) \right. \\
&\qquad \left. \left[\frac{1}{\eta_n\left(g(y \mid \cdot)\right)^{k+1}} - \frac{1}{\eta\left(g(y \mid \cdot)\right)^{k+1}} + (k+1) \frac{[\eta_n - \eta]\left(g(y \mid \cdot)\right)}{\eta\left(g(y \mid \cdot)\right)^{k+2}} \right] \right| \\
&\qquad = o\left(\beta(\eta, \eta_n)\right).
\end{aligned}$$

It follows that

$$\frac{\|D^{(k)}_{\eta_n} \mathrm{F}_{\mathrm{EMS}} - D^{(k)}_{\eta} \mathrm{F}_{\mathrm{EMS}} - D^{(k+1)}_{\eta} \mathrm{F}_{\mathrm{EMS}}(\cdot, \cdot, \eta_n - \eta)\|_{op}}{\beta(\eta, \eta_n)} = o(1) \to 0$$

as $n \to \infty$, proving (A.5). To show (4.12) take $\mu = \nu$ in (A.5)

$$\begin{aligned}
D^{(k)}_{\eta} \mathrm{F}_{\mathrm{EMS}} \nu^k &= (-1)^{k+1} k! \int_{\mathbb{X}} \nu(\mathrm{d}x) K(x, \cdot) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^k} \nu\left(g(y \mid \cdot)\right)^{k-1} \\
&\quad + (-1)^k k! \int_{\mathbb{X}} [\eta + (k-1)\nu](\mathrm{d}x) K(x, \cdot) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^{k}.
\end{aligned}$$

To show that $D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k$ is bounded, take $\varphi \in \mathrm{BL}(\mathbb{X})$ and apply the triangle inequality

$$
\begin{aligned}
&\left| \int_{\mathbb{X}} \varphi(x') D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k(\mathrm{d}x') \right| \\
&= \left| k! \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^k} \nu\left(g(y \mid \cdot)\right)^k \left[-\frac{\nu(\mathrm{d}x)}{\nu\left(g(y \mid \cdot)\right)} + \frac{[\eta + (k-1)\nu](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} \right] \right| \\
&= \Bigg| k! \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^k} \nu\left(g(y \mid \cdot)\right)^k \\
&\quad \left[-\frac{\nu(\mathrm{d}x)}{\nu\left(g(y \mid \cdot)\right)} - \frac{\nu(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} + \frac{\nu(\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} + \frac{[\eta + (k-1)\nu](\mathrm{d}x)}{\eta\left(g(y \mid \cdot)\right)} \right] \Bigg| \\
&= \Bigg| k! \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^k} \nu\left(g(y \mid \cdot)\right)^k \\
&\quad \left[\frac{\nu(\mathrm{d}x)}{\nu\left(g(y \mid \cdot)\right) \eta\left(g(y \mid \cdot)\right)} [\eta + \nu]\left(g(y \mid \cdot)\right) + \frac{1}{\eta\left(g(y \mid \cdot)\right)} [\eta + k\nu](\mathrm{d}x) \right] \Bigg| \\
&= \Bigg| k! \int_{\mathbb{X}} K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^k \\
&\quad \left[\frac{\nu(\mathrm{d}x)}{\nu\left(g(y \mid \cdot)\right)} [\eta + \nu]\left(g(y \mid \cdot)\right) + [\eta + k\nu](\mathrm{d}x) \right] \Bigg| \\
&\leq \left| k! \int_{\mathbb{X}} \nu(\mathrm{d}x) K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^{k-1} [\eta + \nu]\left(g(y \mid \cdot)\right) \right| \\
&+ \left| k! \int_{\mathbb{X}} [\eta + k\nu](\mathrm{d}x) K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^k \right| .
\end{aligned}
$$

We can bound the first term using Assumption 4.1-(b) and $\|K\varphi\|_\infty \leq \|\varphi\|_\infty \leq 1$

$$
\begin{aligned}
&\left| k! \int_{\mathbb{X}} \nu(\mathrm{d}x) K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^{k-1} [\eta + \nu]\left(g(y \mid \cdot)\right) \right| \qquad \text{(A.7)} \\
&\leq k! \frac{m_g^{k+2}}{\eta(\mathbb{X})} \left| \int_{\mathbb{X}} \nu(\mathrm{d}x) K\varphi(x) \right| \left| \int_{\mathbb{Y}} h(\mathrm{d}y) \nu\left(g(y \mid \cdot)\right)^{k-1} [\eta + \nu]\left(g(y \mid \cdot)\right) \right| \\
&\leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^{k-1}}{\eta(\mathbb{X})^{k+1}} \left| [\eta + \nu](\mathbb{X}) \right| \left| \int_{\mathbb{X}} \nu(\mathrm{d}x) K\varphi(x) \right| \\
&\leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left| [\eta + \nu](\mathbb{X}) \right| .
\end{aligned}
$$

Similarly for the second term

$$\begin{aligned}
&\left| k! \int_{\mathbb{X}} [\eta + k\nu](\mathrm{d}x) K\varphi(x) \int_{\mathbb{Y}} \frac{g(y \mid x) h(\mathrm{d}y)}{\eta\left(g(y \mid \cdot)\right)^{k+1}} \nu\left(g(y \mid \cdot)\right)^k \right| \qquad \text{(A.8)} \\
&\leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left| \int_{\mathbb{X}} [\eta + k\nu](\mathrm{d}x) K\varphi(x) \right| \\
&\leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left| [\eta + k\nu](\mathbb{X}) \right|
\end{aligned}$$

Putting (A.7)–(A.8) together gives

$$\begin{aligned}
&\left| \int_{\mathbb{X}} \varphi(x') D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k(\mathrm{d}x') \right| \\
&\qquad \leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left| [\eta + \nu] \right| (\mathbb{X}) + k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left| [\eta + k\nu](\mathbb{X}) \right| \\
&\qquad \leq k! m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left(2\eta(\mathbb{X}) + (k+1)|\nu(\mathbb{X})| \right) \\
&\qquad \leq k!(k+1) m_g^{2k+2} \frac{|\nu(\mathbb{X})|^k}{\eta(\mathbb{X})^{k+1}} \left(\eta(\mathbb{X}) + |\nu(\mathbb{X})| \right),
\end{aligned}$$

recalling that $\mathbb{I}_{\mathbb{X}} \in \mathrm{BL}(\mathbb{X})$ we obtain

$$\beta\left(D_\eta^{(k)} \mathrm{F}_{\mathrm{EMS}}\, \nu^k \right) \leq (k+1)! m_g^{2k+2} \frac{\beta(\nu)^k}{\eta(\mathbb{X})^{k+1}} \left(\eta(\mathbb{X}) + |\nu(\mathbb{X})| \right)$$

for all $\eta \in \mathcal{M}^+(\mathbb{X})$. □

B

Proofs of the Results for SMC-EMS

B.1 Proof of the $\mathbb{L}_p$ Inequality

Before proceeding to the proof of Proposition 5.2 we introduce the following auxiliary Lemma giving some properties of the approximated potentials G_n^N:

Lemma B.1. Under Assumption 4.1-(b), the approximated and exact potentials are positive functions, bounded and bounded away from 0

$$\|G_n\|_\infty \leq m_g^2 < \infty \qquad \text{and} \qquad \inf_{(x,y)} |G_n(x,y)| \geq \frac{1}{m_g^2} > 0$$
$$\|G_n^N\|_\infty \leq m_g^2 < \infty \qquad \text{and} \qquad \inf_{(x,y)} |G_n^N(x,y)| \geq \frac{1}{m_g^2} > 0.$$

We have the following decomposition

$$\begin{aligned} G_n^N(x,y) - G_n(x,y) &= G_n(x,y)\frac{\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}{\eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))} \\ &= G_n^N(x,y)\frac{\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}{\eta_n|_{\mathbb{X}}(g(y \mid \cdot))} \end{aligned}$$

for fixed $(x,y) \in \mathbb{H}$.

Proof. The boundedness of G_n and G_n^N follows from definitions (5.4) and (5.8) and the boundedness of g. The second assertion is proved by considering the relative errors between the exact and the approximated potential, using the definition of G_n and G_n^N and the fact that the denominator of G_n and G_n^N can be seen as the integral

of g with respect to a particular measure:

$$\begin{aligned}\frac{G_n^N(x,y)-G_n(x,y)}{G_n(x,y)} &= \frac{h_n(y)}{g(y \mid x)}\left[\frac{g(y \mid x)}{h_n^N(y)} - \frac{g(y \mid x)}{h_n(y)}\right] \\ &= h_n(y)\left[\frac{1}{h_n^N(y)} - \frac{1}{h_n(y)}\right] \\ &= \frac{h_n(y)-h_n^N(y)}{h_n^N(y)} \\ &= \frac{\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}{\eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}\end{aligned}$$

and

$$\frac{G_n^N(x,y)-G_n(x,y)}{G_n^N(x,y)} = \frac{\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}{\eta_n|_{\mathbb{X}}(g(y \mid \cdot))}$$

respectively.

□

We collect here three auxiliary Lemmata due to Crisan and Doucet (2000, 2002); Míguez et al. (2013) which establish $\mathbb{L}_p$-inequalities for the mutation step, the reweighting step performed with the exact potential G_n (exact reweighting) and the multinomial resampling step.

Lemma B.2 (Exact reweighting)**.** Assume that for any $\varphi \in \mathcal{B}_b(\mathbb{H})$, for some $p \geq 1$ and some finite constants and $\widetilde{C}_{p,n}$

$$\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}},$$

then

$$\mathbb{E}\left[|\Psi_{G_n}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n)(\varphi)|^p\right]^{1/p} \leq \bar{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for any $\varphi \in \mathcal{B}_b(\mathbb{H})$ and for some finite constant $\bar{C}_{p,n}$.

Proof. The proof follows that of Crisan and Doucet (2000, Lemma 2) and Crisan and Doucet (2002, Lemma 4) by exploiting the boundedness of G_n (Lemma B.1).

Apply the definition of Ψ_{G_n} to get the following decomposition

$$\begin{aligned}\Psi_{G_n}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n)(\varphi) &= \frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)} - \frac{\eta_n(G_n\varphi)}{\eta_n(G_n)} \\ &= \frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)} - \frac{\eta_n^N(G_n\varphi)}{\eta_n(G_n)} + \frac{\eta_n^N(G_n\varphi)}{\eta_n(G_n)} - \frac{\eta_n(G_n\varphi)}{\eta_n(G_n)}.\end{aligned}$$

For the first term

$$\left|\frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)} - \frac{\eta_n^N(G_n\varphi)}{\eta_n(G_n)}\right| = \left|\frac{\eta_n^N(G_n\varphi)}{\eta_n^N(G_n)}\right| \left|\frac{\eta_n(G_n) - \eta_n^N(G_n)}{\eta_n(G_n)}\right| \leq \frac{\|\varphi\|_\infty}{|\eta_n(G_n)|}|\eta_n(G_n) - \eta_n^N(G_n)|.$$

For the second term

$$\left|\frac{\eta_n^N(G_n\varphi)}{\eta_n(G_n)} - \frac{\eta_n(G_n\varphi)}{\eta_n(G_n)}\right| = \frac{1}{|\eta_n(G_n)|}|\eta_n^N(G_n\varphi) - \eta_n(G_n\varphi)|.$$

Applying Minkowski's inequality and the hypothesis,

$$\begin{aligned}\mathbb{E}\left[|\Psi_{G_n}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n)(\varphi)|^p\right]^{1/p} &\leq \frac{\|\varphi\|_\infty}{|\eta_n(G_n)|}\,\mathbb{E}\left[|\eta_n(G_n) - \eta_n^N(G_n)|^p\right]^{1/p} \\ &\quad + \frac{1}{|\eta_n(G_n)|}\,\mathbb{E}\left[|\eta_n^N(G_n\varphi) - \eta_n(G_n\varphi)|^p\right]^{1/p} \\ &\leq \frac{2\widetilde{C}_{p,n}m_g^2}{|\eta_n(G_n)|}\frac{\|\varphi\|_\infty}{N^{1/2}}.\end{aligned}$$

Hence, $\bar{C}_{p,n} = 2\widetilde{C}_{p,n}m_g^2/|\eta_n(G_n)| \leq 2\widetilde{C}_{p,n}m_g^4$. □

Lemma B.3 (Multinomial resampling)**.** Assume that for any $\varphi \in \mathcal{B}_b(\mathbb{H})$, for some $p \geq 1$ and some finite constant $\widehat{C}_{p,n}$

$$\mathbb{E}\left[|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq \widehat{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}},$$

then after the resampling step performed through multinomial resampling

$$\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq C_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for any $\varphi \in \mathcal{B}_b(\mathbb{H})$ and for some finite constant $C_{p,n}$.

Proof. The proof follows that of Crisan and Doucet (2000, Lemma 4) and Crisan and Doucet (2002, Lemma 5) using the Marcinkiewicz-Zygmund type inequality in Del Moral (2004, Lemma 7.3.3) and the hypothesis.

Divide into two terms and apply Minkowski's inequality

$$\begin{aligned}\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} &\leq \mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \Psi_{G_n^N}(\eta_n^N)(\varphi)|^p\right]^{1/p} \\ &\quad + \mathbb{E}\left[|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p}.\end{aligned}$$

Denote by $\mathcal{F}_n^N$ the σ-field generated by the weighted samples up to (and including) time n, $\mathcal{F}_n^N := \vee_{p=1}^n \sigma\left(X_p^i, Y_p^i : i \in \{1,\dots,N\}\right)$ and consider the sequence of functions $\Delta_n^i : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{R}$ for $i = 1,\dots,N$

$$\Delta_n^i(x,y) := \varphi(x,y) - \mathbb{E}\left[\varphi(\widetilde{X}_n^i, \widetilde{Y}_n^i) \mid \mathcal{F}_n^N\right].$$

Conditionally on $\mathcal{F}_n^N$, $\Delta_n^i(\widetilde{X}_n^i, \widetilde{Y}_n^i)$ $i = 1,\dots,N$ are independent and have expectation equal to 0, moreover

$$\begin{aligned}\hat{\eta}_n^N(\varphi) - \Psi_{G_n^N}(\eta_n^N)(\varphi) &= \frac{1}{N}\sum_{i=1}^N \left(\varphi(\widetilde{X}_n^i, \widetilde{Y}_n^i) - \mathbb{E}\left[\varphi(\widetilde{X}_n^i, \widetilde{Y}_n^i) \mid \mathcal{F}_n^N\right]\right) \\ &= \frac{1}{N}\sum_{i=1}^N \Delta_n^i(\widetilde{X}_n^i, \widetilde{Y}_n^i).\end{aligned}$$

By the Marcinkiewicz-Zygmund type inequality in Del Moral (2004, Lemma 7.3.3),

$$\begin{aligned}&\sqrt{N}\,\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \Psi_{G_n^N}(\eta_n^N)(\varphi)|^p \mid \mathcal{F}_n^N\right]^{1/p} \\ &\quad\le b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N \left(\sup(\Delta_n^i) - \inf(\Delta_n^i)\right)^2\right)^{1/2} \\ &\quad\le b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N 4\left(\sup|\Delta_n^i|\right)^2\right)^{1/2} \\ &\quad\le b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N 16\|\varphi\|_\infty^2\right)^{1/2} \\ &\quad\le 4b(p)^{1/p}\|\varphi\|_\infty\end{aligned}$$

where

$$\begin{aligned} b(2n) &:= (2n)_n 2^{-n} \\ b(2n+1) &:= \frac{(2n+1)_{(n+1)}}{\sqrt{n+1/2}}2^{-n+1/2} \qquad \text{with} \qquad (m+n)_n := (m+n)!/n!. \end{aligned} \tag{B.1}$$

Since $\hat{\eta}_n(\varphi) \equiv \Psi_{G_n}(\eta_n)(\varphi)$, this result combined with the hypothesis yields

$$\begin{aligned}\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} &\le 2b(p)^{1/p}\frac{\|\varphi\|_\infty}{N^{1/2}} + \widehat{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}} \\ &\le (2b(p)^{1/p} + \widehat{C}_{p,n})\frac{\|\varphi\|_\infty}{N^{1/2}}.\end{aligned}$$

Thus, $C_{p,n} = 2b(p)^{1/p} + \widehat{C}_{p,n}$. □

Lemma B.4 (Mutation)**.** Assume that for any $\varphi \in \mathcal{B}_b(\mathbb{H})$, for some $p \geq 1$ and some finite constant $C_{p,n}$

$$\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq C_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

then, after the mutation step

$$\mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \eta_{n+1}(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n+1}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for any $\varphi \in \mathcal{B}_b(\mathbb{H})$ and for some finite constant $\widetilde{C}_{p,n+1}$.

Proof. The proof follows that of Crisan and Doucet (2000, Lemma 1) and Crisan and Doucet (2002, Lemma 3).

Similarly to Lemma B.3, divide into two terms and apply Minkowski's inequality

$$\begin{aligned}\mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \eta_{n+1}(\varphi)|^p\right]^{1/p} &= \mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \hat{\eta}_n M_{n+1}(\varphi)|^p\right]^{1/p} \\ &\leq \mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \hat{\eta}_n^N M_{n+1}(\varphi)|^p\right]^{1/p} \\ &\quad + \mathbb{E}\left[|\hat{\eta}_n^N M_{n+1}(\varphi) - \hat{\eta}_n M_{n+1}(\varphi)|^p\right]^{1/p}.\end{aligned}$$

Let $\mathcal{G}_n^N$ denote the σ-field generated by the particle system up to (and including) time n before the mutation step at time $n+1$, $\mathcal{G}_n^N = \vee_{p=1}^n \sigma\left(\widetilde{X}_p^i, \widetilde{Y}_p^i : i \in \{1,\ldots,N\}\right)$ and consider the sequence of functions $\Delta_{n+1}^i : \mathbb{X}\times\mathbb{Y} \mapsto \mathbb{R}$ for $i = 1,\ldots,N$

$$\Delta_{n+1}^i(x,y) := \varphi(x,y) - \mathbb{E}\left[\varphi(X_{n+1}^i, Y_{n+1}^i) \mid \mathcal{G}_n^N\right].$$

Conditionally on $\mathcal{G}_n^N$, $\Delta_{n+1}^i(X_{n+1}^i, Y_{n+1}^i)$, $i = 1,\ldots,N$ are independent and have expectation equal to 0, moreover

$$\begin{aligned}\eta_{n+1}^N(\varphi) - \hat{\eta}_n^N M_{n+1}(\varphi) &= \frac{1}{N}\sum_{i=1}^N \left[\varphi(X_{n+1}^i, Y_{n+1}^i) - M_{n+1}\varphi(\widetilde{X}_n^i, \widetilde{Y}_n^i)\right] \\ &= \frac{1}{N}\sum_{i=1}^N \Delta_{n+1}^i(X_{n+1}^i, Y_{n+1}^i).\end{aligned}$$

Conditioning on $\mathcal{G}_n^N$ and applying the Marcinkiewicz-Zygmund type inequality in

Del Moral (2004, Lemma 7.3.3)

$$\begin{aligned}
&\sqrt{N}\,\mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \hat{\eta}_n^N M_{n+1}(\varphi)|^p \mid \mathcal{G}_{n-1}^N\right]^{1/p} \\
&\qquad \leq b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N \left(\sup(\Delta_{n+1}^i) - \inf(\Delta_{n+1}^i)\right)^2\right)^{1/2} \\
&\qquad \leq b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N 4\left(\sup|\Delta_{n+1}^i|\right)^2\right)^{1/2} \\
&\qquad \leq b(p)^{1/p}\frac{1}{\sqrt{N}}\left(\sum_{i=1}^N 16\|\varphi\|_\infty^2\right)^{1/2} \\
&\qquad \leq 4b(p)^{1/p}\|\varphi\|_\infty
\end{aligned}$$

where $b(p)$ are as in (B.1). Combining this result with the hypothesis yields

$$\mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \eta_{n+1}(\varphi)|^p\right]^{1/p} \leq (2b(p)^{1/p} + C_{p,n})\frac{\|\varphi\|_\infty}{N^{1/2}},$$

and the result holds with $\widetilde{C}_{p,n+1} = 2b(p)^{1/p} + C_{p,n}$. □

The proof of the $\mathbb{L}_p$-inequality in Proposition 5.2 is based on an inductive argument which uses Lemmata B.2-B.4 and Lemma 5.1:

Proof of Proposition 5.2. At time $n = 1$, the particles $(X_1^i, Y_1^i)_{i=1}^N$ are sampled i.i.d. from $\eta_1 \equiv \hat{\eta}_1$ thus $\mathbb{E}\left[\varphi(X_1^i, Y_1^i)\right] = \eta_1(\varphi)$ for $i = 1, \ldots, N$. We can define the sequence of functions $\Delta_1^i : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{R}$ for $i = 1, \ldots, N$

$$\Delta_1^i(x, y) := \varphi(x, y) - \mathbb{E}\left[\varphi(X_1^i, Y_1^i)\right]$$

so that,

$$\eta_1^N(\varphi) - \eta_1(\varphi) = \frac{1}{N}\sum_{i=1}^N \Delta_1^i(X_1^i, Y_1^i),$$

and apply Lemma 7.3.3 in Del Moral (2004) to get

$$\mathbb{E}\left[|\eta_1^N(\varphi) - \eta_1(\varphi)|^p\right]^{1/p} \leq 2b(p)^{1/p}\frac{\|\varphi\|_\infty}{N^{1/2}},$$

with $b(p) < \infty$, for every $p \geq 1$ Then, assume that the result holds at time n: for

every $\varphi \in \mathcal{B}_b(\mathbb{H})$ and some finite constant $\widetilde{C}_{p,n}$

$$\mathbb{E}\left[|\eta_n^N(\varphi) - \eta_n(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}.$$

The $\mathbb{L}_p$-inequality in (5.13) is obtained by combining the results of Lemma B.2 and Lemma 5.1 using Minkowski's inequality

$$\mathbb{E}\left[|\Psi_{G_n^N}(\eta_n^N)(\varphi) - \Psi_{G_n}(\eta_n)(\varphi)|^p\right]^{1/p} \leq (\bar{C}_{p,n} + \ddot{C}_{p,n})\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for every $\varphi \in \mathcal{B}_b(\mathbb{H})$ and some finite constants $\bar{C}_{p,n}, \ddot{C}_{p,n}$. Thus, $\widehat{C}_{p,n} = \bar{C}_{p,n} + \ddot{C}_{p,n}$.

Lemma B.3 gives

$$\mathbb{E}\left[|\hat{\eta}_n^N(\varphi) - \hat{\eta}_n(\varphi)|^p\right]^{1/p} \leq C_{p,n}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for every $\varphi \in \mathcal{B}_b(\mathbb{H})$ and some finite constants $C_{p,n}$, and Lemma B.4 gives

$$\mathbb{E}\left[|\eta_{n+1}^N(\varphi) - \eta_{n+1}(\varphi)|^p\right]^{1/p} \leq \widetilde{C}_{p,n+1}\frac{\|\varphi\|_\infty}{N^{1/2}}$$

for every $\varphi \in \mathcal{B}_b(\mathbb{H})$ and some finite constant $\widetilde{C}_{p,n+1}$. The result follows for all $n \in \mathbb{N}$ by induction. □

B.2 Proof of Corollary 5.2

The results in Corollary 5.2 can be straightforwardly obtained from Proposition 5.2, Proposition 5.1 and an argument exploiting the compactness of $\mathbb{H}$ as in the proof of Proposition 5.5:

Proof of Corollary 5.2. 1. Using the definition of the potentials in (5.4)-(5.8) and Proposition 5.2,

$$\begin{aligned}
&\mathbb{E}\left[|G_n^N(x,y) - G_n(x,y)|^p\right]^{1/p} \\
&\quad= \mathbb{E}\left[|g(y \mid x)|^p \left|\frac{\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}{\eta_n|_{\mathbb{X}}(g(y \mid \cdot))\,\eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))}\right|^p\right]^{1/p} \\
&\quad\leq m_g^3\,\mathbb{E}\left[|\eta_n|_{\mathbb{X}}(g(y \mid \cdot)) - \eta_n^N|_{\mathbb{X}}(g(y \mid \cdot))|^p\right]^{1/p} \\
&\quad\leq m_g^3\widetilde{C}_{p,n}\frac{\|g(y \mid \cdot)\|_\infty}{\sqrt{N}} \\
&\quad\leq m_g^4\widetilde{C}_{p,n}\frac{1}{\sqrt{N}}.
\end{aligned}$$

2. For fixed (x, y), $g(y \mid x)$ is constant, thus

$$G_n^N(x,y) = \frac{g(y \mid x)}{\eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot)\right)} \to \frac{g(y \mid x)}{\eta_n|_{\mathbb{X}}\left(g(y \mid \cdot)\right)} = G_n(x,y)$$

almost surely as a consequence of Proposition 5.1, the continuous mapping theorem and the boundedness of g.

3. The previous result can be strengthened to uniform almost sure convergence using the continuity of g and the compactness of $\mathbb{H}$ following the same construction of Newey (1991). Given the following bound on the difference between the potentials

$$\begin{aligned}
\|G_n^N - G_n\|_\infty &= \sup_{(x,y)} \left|G_n^N(x,y) - G_n(x,y)\right| \\
&= \sup_{(x,y)} |g(y \mid x)| \left| \frac{\eta_n|_{\mathbb{X}}\left(g(y \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot)\right)}{\eta_n|_{\mathbb{X}}\left(g(y \mid \cdot)\right)\eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot)\right)} \right| \\
&\leq m_g^3 \sup_y \left|\left(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}}\right)\left(g(y \mid \cdot)\right)\right|
\end{aligned}$$

it is sufficient to prove that $\sup_y \left|\left(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}}\right)\left(g(y \mid \cdot)\right)\right| \to 0$ almost surely.

First, notice that since $\mathbb{H}$ is compact, the Heine-Cantor Theorem (Rudin, 1964, Theorem 4.19) implies that g is uniformly continuous: for any given $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that

$$|g(y \mid x) - g(\tilde{y} \mid \tilde{x})| < \frac{\varepsilon}{4}$$

if $\|y - \tilde{y}\|_2 + \|x - \tilde{x}\|_2 < \delta_\varepsilon$ for all $(x, y), (\tilde{x}, \tilde{y}) \in \mathbb{H}$. As a consequence, the sequence $\left(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}}\right)\left(g(y \mid \cdot)\right)$ is equicontinuous:

$$\begin{aligned}
&\left|\left(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}}\right)\left(g(y \mid \cdot)\right) - \left(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}}\right)\left(g(\tilde{y} \mid \cdot)\right)\right| \\
&\qquad \leq \left|\eta_n|_{\mathbb{X}}\left(g(y \mid \cdot) - g(\tilde{y} \mid \cdot)\right)\right| + \left|\eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot) - g(\tilde{y} \mid \cdot)\right)\right| \\
&\qquad \leq \eta_n|_{\mathbb{X}}\left(\left|g(y \mid \cdot) - g(\tilde{y} \mid \cdot)\right|\right) + \eta_n^N|_{\mathbb{X}}\left(\left|g(y \mid \cdot) - g(\tilde{y} \mid \cdot)\right|\right) \\
&\qquad \leq 2 \sup_{x \in \mathbb{X}} |g(y \mid x) - g(\tilde{y} \mid x)| \leq \frac{\varepsilon}{2}.
\end{aligned}$$

if $\|y - \tilde{y}\|_2 \leq \delta_\varepsilon$ for all N.

Let $B(y) := \{\tilde{y} \in \mathbb{Y} : \|y - \tilde{y}\|_2 < \delta_\varepsilon\}$ denote the ball in $\mathbb{Y}$ centred around y of radius δ_ε and use compactness of $\mathbb{H}$ (and hence of $\mathbb{Y}$) to extract a finite open subcover $\{B(y^j)\}_{j=1}^J$ of $\{B(y)\}_{y \in \mathbb{Y}}$. As a consequence of the equicontinuity of

the sequence we have that, for all $y \in B(y^j)$, $j = 1, \ldots, J$ and for all N,

$$\begin{aligned}
&|(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,(g(y \mid \cdot))\,| \\
&\quad \leq |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y \mid \cdot) - g(y^j \mid \cdot)\big)\,| + |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y^j \mid \cdot)\big)\,| \\
&\quad \leq \frac{\varepsilon}{2} + \max_{j=1,\ldots,J} |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y^j \mid \cdot)\big)\,|,
\end{aligned}$$

from which follows

$$\sup_y \left|(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,(g(y \mid \cdot))\right| \leq \frac{\varepsilon}{2} + \max_{j=1,\ldots,J} |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y^j \mid \cdot)\big)\,|,$$

we find that $\sup_y \left|(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,(g(y \mid \cdot))\right| \to 0$ almost surely if for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}\left(\limsup_{N\to\infty} \left\{\frac{\varepsilon}{2} + \max_{j=1,\ldots,J} |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y^j \mid \cdot)\big)\,| > \varepsilon\right\}\right) \\
&\qquad = \mathbb{P}\left(\limsup_{N\to\infty} \left\{\max_{j=1,\ldots,J} |(\eta_n|_{\mathbb{X}} - \eta_n^N|_{\mathbb{X}})\,\big(g(y^j \mid \cdot)\big)\,| > \frac{\varepsilon}{2}\right\}\right) = 0.
\end{aligned}$$

The results follows using the almost sure pointwise convergence in point 2 since ε is arbitrary.

□

B.3 Proof of the Bias Estimates

Before proceeding to the derivation of the bias estimates, we state and prove the following auxiliary result, which is a direct consequence of the $\mathbb{L}_p$-inequality (5.12).

Lemma B.5. Under Assumption 4.1 and 5.1, for every $n \geq 1$ and every $p \geq 1$ there exists a finite constant $A_{p,n}$ such that, for every measurable bounded function $\varphi \in \mathcal{B}_b(\mathbb{H})$

$$\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^p\right]^{1/p} \leq A_{p,n}\frac{\|\varphi\|_\infty}{\sqrt{N}}.$$

Additionally, for all $\varphi, \psi \in \mathcal{B}_b(\mathbb{H})$ and for integers $0 \leq l, m < \infty$

$$\begin{aligned}
&\left|\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^l \left|\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right|^m\right]\right| \\
&\qquad \leq \|\varphi\|_\infty^l \|\psi\|_\infty^m \frac{A_{p,n}^{l+m}}{N^{(l+m)/2}}.
\end{aligned}$$

Proof. To prove the first assertion, apply Minkowski's inequality

$$\begin{aligned}\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^p\right]^{1/p} &\leq \mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \eta_n(G_n\varphi)\right|^p\right]^{1/p} \\ &\quad + \mathbb{E}\left[\left|\mathbb{E}\left[\eta_n(G_n\varphi) - \eta_n^N(G_n^N\varphi)\right]\right|^p\right]^{1/p} \\ &\leq 2\,\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \eta_n(G_n\varphi)\right|^p\right]^{1/p} \\ &\leq 2\,\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right|^p\right]^{1/p} \\ &\quad + 2\,\mathbb{E}\left[\left|\eta_n^N(G_n\varphi) - \eta_n(G_n\varphi)\right|^p\right]^{1/p}\end{aligned}$$

where the second inequality follows from Jensen's inequality applied to the second expectation in line 1. Since

$$\begin{aligned}\left|\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right| &\leq \|\varphi\|_\infty \eta_n^N(|G_n^N - G_n|) \\ &= \|\varphi\|_\infty \frac{1}{N}\sum_{i=1}^N \left|G_n^N(X_n^i, Y_n^i) - G_n(X_n^i, Y_n^i)\right|,\end{aligned}$$

Minkowski's inequality gives

$$\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right|^p\right]^{1/p} \leq \|\varphi\|_\infty \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left|G_n^N(X_n^i, Y_n^i) - G_n(X_n^i, Y_n^i)\right|^p\right]^{1/p}$$

and using the decomposition of the potentials in Lemma B.1

$$\begin{aligned}&\mathbb{E}\left[\left|G_n^N(X_n^i, Y_n^i) - G_n(X_n^i, Y_n^i)\right|^p\right]^{1/p} \\ &\quad = \mathbb{E}\left[\left|\frac{G_n^N(X_n^i, Y_n^i)}{\eta_{n|\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)}\right|^p \left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|^p\right]^{1/p} \\ &\quad \leq m_g^3\,\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|^p\right]^{1/p}.\end{aligned} \tag{B.2}$$

Then, consider $\mathcal{S}_n^N := \sigma\left(Y_n^i : i \in \{1, \ldots, N\}\right)$. Conditionally on $\mathcal{S}_n^N$ the Y_n^i are fixed for $i = 1, \ldots, N$ and $\eta_n^N|_{\mathbb{X}}$ is independent on $\mathcal{S}_n^N$ (this is due to the definition of the

mutation kernel (5.3)) thus

$$\begin{aligned}
&\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \eta_n^N(G_n\varphi)\right|^p\right]^{1/p} \\
&\quad\leq \|\varphi\|_\infty \frac{1}{N}\sum_{i=1}^N m_g^3\, \mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|^p\right]^{1/p} \\
&\quad\leq m_g^3\|\varphi\|_\infty \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(Y_n^i \mid \cdot)\right)\right|^p \mid \mathcal{S}_n^N\right]\right]^{1/p} \\
&\quad\leq \|\varphi\|_\infty \frac{m_g^4 \widetilde{C}_{p,n}}{\sqrt{N}}
\end{aligned}$$

where the last inequality follows from (5.12) for all $n \geq 1$ since $g(y \mid \cdot)$ is a continuous bounded function for all fixed $y \in \mathbb{Y}$ and

$$\begin{aligned}
\mathbb{E}\left[\left|\eta_n|_{\mathbb{X}}\left(g(y \mid \cdot)\right) - \eta_n^N|_{\mathbb{X}}\left(g(y \mid \cdot)\right)\right|^p\right]^{1/p} &= \mathbb{E}\left[\left|\eta_n\left(g(y \mid \cdot)\right) - \eta_n^N\left(g(y \mid \cdot)\right)\right|^p\right]^{1/p} \\
&\leq \frac{m_g \widetilde{C}_{p,n}}{\sqrt{N}}.
\end{aligned}$$

Combining this result with (5.12) for the bounded measurable function $G_n\varphi$ yields

$$\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^p\right]^{1/p} \leq 2\|\varphi\|_\infty \frac{m_g^4 \widetilde{C}_{p,n} + m_g^2 \widetilde{C}_{p,n}}{\sqrt{N}}$$

giving $A_{p,n} = 2(m_g^4 3\widetilde{C}_{p,n} + m_g^2 \widetilde{C}_{p,n})$ for all $n \geq 1$. This result and the Cauchy-Schwarz inequality give

$$\begin{aligned}
&\left|\mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^l \left|\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right|^m\right]\right| \\
&\quad\leq \mathbb{E}\left[\left|\eta_n^N(G_n^N\varphi) - \mathbb{E}\left[\eta_n^N(G_n^N\varphi)\right]\right|^{2l}\right]^{1/2} \mathbb{E}\left[\left|\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right|^{2m}\right]^{1/2} \\
&\quad\leq \|\varphi\|_\infty^l \|\psi\|_\infty^m \frac{A_{2l,n}^l A_{2m,n}^m}{N^{(l+m)/2}}.
\end{aligned}$$

□

We can now move onto the proof of the bias estimates in Proposition 5.4. Similarly to the proof of the $\mathbb{L}_p$-inequality, the bias estimates are obtained using an inductive argument considering both statements at the same time and starting from $n = 1$.

Proof of Proposition 5.4. At time $n = 1$, the particles $(X_1^i, Y_1^i)_{i=1}^N$ are i.i.d. samples

from $\eta_1 \equiv \hat{\eta}_1$, thus

$$\begin{aligned}\mathbb{E}\left[\eta_1^N(\varphi)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\varphi(X_1^i, Y_1^i)\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\varphi(X_1^i, Y_1^i)\right] = \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbb{H}}\eta_1\left(\mathrm{d}(x,y)\right)\varphi(x,y) = \eta_1(\varphi)\end{aligned}$$

and

$$\left|\mathbb{E}\left[\eta_1^N(\varphi)\right] - \eta_1(\varphi)\right| = 0 \leq \frac{\|\varphi\|_\infty}{N} \tag{B.3}$$

for every $\varphi \in \mathcal{B}_b(\mathbb{H})$ with $C_1 = 0$.

To argue the second statement we observe that at time $n = 1$ the particles $(X_1^i, Y_1^i)_{i=1}^N$ are i.i.d. samples from $\eta_1 \equiv \hat{\eta}_1$, and therefore the potential function at time $n = 1$ is $G_1^N(x,y) \equiv G_1(x,y) \equiv 1$ for all $(x,y) \in \mathbb{H}$:

$$\mathbb{E}\left[\eta_1^N(G_1^N\varphi)\right] - \eta_1(G_1\varphi) = \mathbb{E}\left[\eta_1^N(\varphi)\right] - \eta_1(\varphi) = \frac{\|\varphi\|_\infty}{N}$$

for every $\varphi \in \mathcal{B}_b(\mathbb{H})$ with $D_1 = 0$.

For $n > 1$, assume that both statements hold at time n. Then consider the two statements at time $n+1$ separately.

Statement 1 First, by conditioning on the σ-field generated by the particle system up to (and including) time n, $\mathcal{G}_n^N := \vee_{p=1}^n \sigma\left(\widetilde{X}_p^i, \widetilde{Y}_p^i : i \in \{1,\ldots,N\}\right)$, we have

$$\begin{aligned}\left|\mathbb{E}\left[\eta_{n+1}^N(\varphi)\right] - \eta_{n+1}(\varphi)\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[\eta_{n+1}^N(\varphi) \mid \mathcal{G}_n^N\right]\right] - \eta_{n+1}(\varphi)\right| \\ &= \left|\mathbb{E}\left[\hat{\eta}_n^N\left(M_{n+1}(\varphi)\right)\right] - \hat{\eta}_n\left(M_{n+1}(\varphi)\right)\right|.\end{aligned}$$

Then, the proof of statement 1 follows that of Olsson and Rydén (2004). By conditioning on $\mathcal{F}_n^N$, the σ-field generated by the particle system up to (and including)

time $n-1$ and by (Y_n^i, X_n^i) for $i = 1, \ldots, N$, and defining $\psi := M_{n+1}(\varphi)$ we have

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{\eta}_n^N(\psi)\right] - \hat{\eta}_n(\psi)\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[\hat{\eta}_n^N(\psi) \mid \mathcal{F}_n^N\right]\right] - \hat{\eta}_n(\psi)\right| && \text{(B.4)}\\
&= \left|\mathbb{E}\left[\frac{\eta_n^N(G_n^N\psi)}{\eta_n^N(G_n^N)}\right] - \hat{\eta}_n(\psi)\right| \\
&\leq \left|\mathbb{E}\left[\frac{\eta_n^N(G_n^N\psi)}{\eta_n^N(G_n^N)}\right] - \frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]}\right| && \text{(B.5)}\\
&\quad + \left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\eta_n(G_n)}\right|.
\end{aligned}$$

The second term is bounded as in Lemma 2.3 of Olsson and Rydén (2004):

$$\begin{aligned}
\left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\eta_n(G_n)}\right| &\leq \left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]}\right| \\
&\quad + \left|\frac{\eta_n(G_n\psi)}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\eta_n(G_n)}\right| \\
&= \left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right] - \eta_n(G_n\psi)}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]}\right| \\
&\quad + \left|\eta_n(G_n\psi)\frac{\eta_n(G_n) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]\eta_n(G_n)}\right| \\
&\leq m_g^2\left|\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right] - \eta_n(G_n\psi)\right| \\
&\quad + m_g^2\|\psi\|_\infty\left|\eta_n(G_n) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right|.
\end{aligned}$$

The induction hypothesis gives

$$\left|\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right] - \eta_n(G_n\psi)\right| \leq \frac{D_n\|\psi\|_\infty}{N}$$

and

$$\left|\eta_n(G_n) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right| \leq \frac{D_n}{N}.$$

Therefore,

$$\begin{aligned}
\left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\eta_n(G_n)}\right| &\leq m_g^2\frac{D_n\|\psi\|_\infty}{N} + m_g^2\|\psi\|_\infty\frac{D_n}{N} \\
&= 2m_g^2\frac{D_n\|\psi\|_\infty}{N}.
\end{aligned}$$

For the first term in (B.4), consider a two-dimensional Taylor expansion of the function $(u, v) \mapsto u/v$ around (u_0, v_0) with a second order remainder of Lagrange form

$$\frac{u}{v} = \frac{u_0}{v_0} + \frac{1}{v_0}(u - u_0) - \frac{u_0}{v_0^2}(v - v_0) + \frac{\theta_u}{\theta_v^3}(v - v_0)^2 - \frac{1}{\theta_v^2}(v - v_0)(u - u_0)$$

where (θ_u, θ_v) is a point on the line segment between (u, v) and (u_0, v_0).

Applying the Taylor expansion above to $\eta_n^N(G_n^N\psi)/\eta_n^N(G_n^N)$ around the point $\left(\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right], \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right)$, as in Lemma 2.4 of Olsson and Rydén (2004), gives

$$\begin{aligned}\frac{\eta_n^N(G_n^N\psi)}{\eta_n^N(G_n^N)} = &\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} + \frac{1}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]}\left(\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right) \\ &- \frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]^2}\left(\eta_n^N(G_n^N) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right) + R_n^N(\theta_u, \theta_v)\end{aligned} \tag{B.6}$$

where the remainder is a function of (θ_u, θ_v), a point on the line segment between $(\eta_n^N(G_n^N\psi), \eta_n^N(G_n^N))$ and $\left(\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right], \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right)$

$$\begin{aligned}R_n^N(\theta_u, \theta_v) := &\frac{\theta_u}{\theta_v^3}\left(\eta_n^N(G_n^N) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right)^2 \\ &- \frac{1}{\theta_v^2}\left(\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right)\left(\eta_n^N(G_n^N) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right).\end{aligned}$$

Taking the expectation of both sides of (B.6) yields

$$\mathbb{E}\left[\frac{\eta_n^N(G_n^N\psi)}{\eta_n^N(G_n^N)}\right] = \frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} + \mathbb{E}\left[R_n^N(\theta_u, \theta_v)\right].$$

Since one of the extremal points of the segment is random, (θ_u, θ_v) is random too; because we have $m_g^{-2} \leq G_n^N \leq m_g^2$ (see Lemma B.1) it follows that

$$\eta_n^N(G_n^N\psi) \leq m_g^2\|\psi\|_\infty, \qquad \eta_n^N(G_n^N) \geq m_g^{-2}$$

from which follows $|\theta_u| \leq m_g^2\|\psi\|_\infty$, $|\theta_v^{-1}| \geq m_g^{-2}$ almost surely.

By Lemma B.5 with $l = m = 1$ and $\psi \equiv 1$

$$\begin{aligned}\left|\mathbb{E}\left[R_n^N(\theta_x, \theta_y)\right]\right| &\leq m_g^8\|\psi\|_\infty \mathbb{E}\left[\left|\eta_n^N(G_n^N) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right|^2\right] \\ &\quad + m_g^4\left|\mathbb{E}\left[\left|\eta_n^N(G_n^N\psi) - \mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]\right|\left|\eta_n^N(G_n^N) - \mathbb{E}\left[\eta_n^N(G_n^N)\right]\right|\right]\right| \\ &\leq m_g^8 A_{2,n}^2\frac{\|\psi\|_\infty}{N} + m_g^4 A_{2,n}^2\frac{\|\psi\|_\infty}{N}.\end{aligned}$$

Hence,

$$\begin{aligned}
\left|\mathbb{E}\left[\eta_{n+1}^N(\varphi)\right] - \eta_{n+1}(\varphi)\right| &= \left|\mathbb{E}\left[\hat{\eta}_n^N(\psi)\right] - \hat{\eta}_n(\psi)\right| \\
&\leq \left|\mathbb{E}\left[\frac{\eta_n^N(G_n^N\psi)}{\eta_n^N(G_n^N)}\right] - \frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]}\right| \\
&+ \left|\frac{\mathbb{E}\left[\eta_n^N(G_n^N\psi)\right]}{\mathbb{E}\left[\eta_n^N(G_n^N)\right]} - \frac{\eta_n(G_n\psi)}{\eta_n(G_n)}\right| \\
&\leq m_g^2(2D_n + m_g^2 A_{2,n}^2(m_g^4+1))\frac{\|\psi\|_\infty}{N} \\
&\leq m_g^2(2D_n + m_g^2 A_{2,n}^2(m_g^4+1))\frac{\|\varphi\|_\infty}{N}
\end{aligned}$$

giving $C_n = m_g^2(2D_n + m_g^2 A_{2,n}^2(m_g^4+1))$.

Statement 2 The following decomposition

$$\begin{aligned}
\left|\mathbb{E}\left[\eta_{n+1}^N(G_{n+1}^N\varphi)\right] - \eta_{n+1}(G_{n+1}\varphi)\right| &\leq \left|\mathbb{E}\left[\eta_{n+1}^N(G_{n+1}^N\varphi) - \eta_{n+1}^N(G_{n+1}\varphi)\right]\right| \\
&+ \left|\mathbb{E}\left[\eta_{n+1}^N(G_{n+1}\varphi)\right] - \eta_{n+1}(G_{n+1}\varphi)\right| \\
&\leq \frac{\widetilde{D}_n + C_n}{N},
\end{aligned}$$

where the first term is bounded using Lemma 5.2 and the second term is bounded using Statement 1, shows that the result holds with $D_n = \widetilde{D}_n + C_n$.

□

Part II

Interactions Between Families of Independent Particles

8

Some Properties of Divide and Conquer Sequential Monte Carlo

The work presented in this chapter and particularly Sections 8.2, 8.4.1 and 8.4.2 is the result of a collaboration with Dr Juan Kuntz Nussio. An extended version of this chapter is in preparation (Kuntz et al., 2021a).

8.1 Introduction

In recent years there has been a growing interest in particle methods which allow for parallel and distributed implementation (see, e.g., Chopin and Papaspiliopoulos (2020, Chapter 19) and references therein); most of the effort has been concentrated on parallelising the resampling step, as this is the only step in which the particles interact in the SIR Algorithm 1. Divide and Conquer SMC is an extension of SMC algorithms proposed in Lindsten et al. (2017), which generalises the classical SMC framework from sequences (or chains) to trees. On one hand, the extension of the SMC methodology to trees results in algorithms which and naturally lends themselves to distributed computing allowing to perform inference on bigger models than those usually dealt with standard SMC, on the other hand, many statistical models can be naturally represented by trees rather than sequences (e.g. Bayesian hierarchical models).

Following the divide-and-conquer computational paradigm, divide and conquer SMC (DaC-SMC) decomposes large inferential problems iteratively into smaller, more manageable ones. Each smaller problem is approximately solved using SMC steps, then the sub-problems are merged following the classical fork-and-join approach. These additional merging steps introduce non-standard interactions between families of particles in the DaC-SMC algorithm.

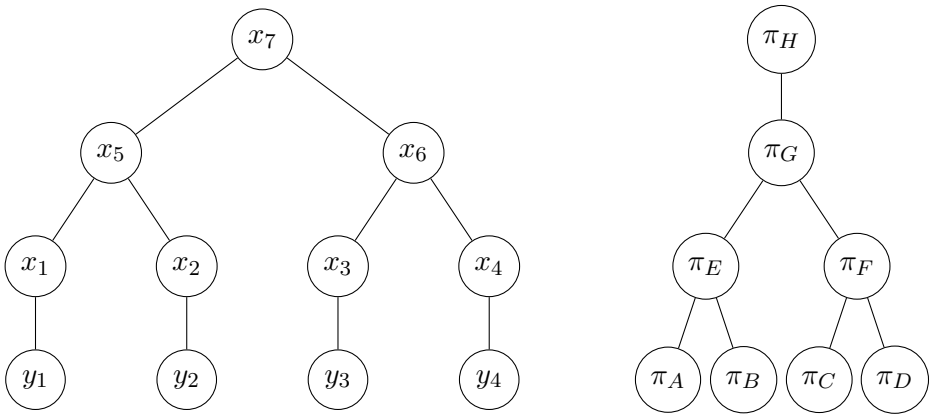


Figure 26: (a) Bayesian hierarchical model of Example 8.1. (b) Corresponding tree decomposition of the posterior π_H.

This chapter is concerned with the theoretical characterisation of DaC-SMC, which is obtained by combining the convergence results for standard SMC summarised in Chapter 2 with results on products of empirical measures which exploit the independence between two families of empirical measures to control the error of their products.

As in the case of SMC, DaC-SMC can be employed to sample from models with an embedded tree structure (Jewell, 2015; Paige and Wood, 2016), alternatively, one could build an artificial tree decomposition to sample from a given target π for computational convenience (Ding and Gandy, 2018) in the spirit of Chopin (2002); Del Moral et al. (2006b) for standard SMC.

Example 8.1 (Bayesian Hierarchical Model)**.** Consider the simple Bayesian hierarchical model in Figure 26 with three observations $y_{1:4}$ and seven latent variables $x_{1:7}$. The target distribution is the posterior $p(x_{1:7} \mid y_{1:4})$, or, in the notation that we will adopt to describe DaC-SMC, $\pi(\boldsymbol{x}_H) = \pi_H(\boldsymbol{x}_H) = p(x_{1:7} \mid y_{1:4})$, where $\boldsymbol{x}_H = x_{1:7}$. We can obtain a tree-like decomposition for this target using the tree structure embedded in this model.

Starting from the first level of the hierarchical model we define $\pi_A(\mathbf{x}_A) = p(x_1 \mid y_1)$, $\pi_B(\mathbf{x}_B) = p(x_2 \mid y_2)$, $\pi_C(\mathbf{x}_C) = p(x_3 \mid y_3)$ and $\pi_D(\mathbf{x}_D) = p(x_4 \mid y_4)$, where $\mathbf{x}_A = x_1, \mathbf{x}_B = x_2, \mathbf{x}_C = x_3$ and $\mathbf{x}_D = 4$. Following the topology of the tree in Figure 26(a) the distributions at the leaves are merged pairwise to obtain $\pi_E(\mathbf{x}_E) = p(x_{1:2} \mid y_{1:2})$ and $\pi_F(\mathbf{x}_F) = p(x_{3:4} \mid y_{3:4})$, where $\mathbf{x}_E = x_{1:2}$ and $\mathbf{x}_F = x_{3:4}$, respectively. Repeating this procedure we obtain $\pi_G(\mathbf{x}_G) = p(x_{1:6} \mid y_{1:4})$, where $\mathbf{x}_G = x_{1:6}$, and, finally,

$\pi_H(\mathbf{x}_H) = p(x_{1:7} \mid y_{1:4})$, where $\mathbf{x}_H = x_{1:7}$.

The collection of target distributions $\{\pi_u(\boldsymbol{x}_u) : u \in \mathbb{T}\}$, where $\mathbb{T}$ is the tree in Figure 26(b), is a tree decomposition of the posterior $p(x_{1:7} \mid y_{1:4})$.

The tree decomposition $\{\pi_u : u \in \mathbb{T}\}$ is often non-unique, for instance, an alternative tree decomposition for this model can be obtained by removing the root of the tree in Figure 26(a) and defining the distributions π_u over each sub-tree recursively as described in (Lindsten et al., 2017, Section 3.1). In this case each distribution of the decomposition is indexed by the tree obtained by removing the leaf nodes (i.e. those corresponding to the observations $y_{1:4}$) from the tree in Figure 26(a).

Example 8.1 shows how to construct tree decomposition of the posterior of interest for a model with an obvious hierarchical structure, however, the tree decomposition might also be completely artificial:

Example 8.2 (Undirected Graphical Models)**.** Consider a simple Ising model defined on a 4×4 lattice. A spin variable $x_k \in \{1, -1\}$ is associated with each node and the target distribution is the probability of a given configuration $\log \pi(\mathbf{x}) = -\beta \sum_{(k,l)\in E} x_k x_l$, where $\beta \geq 0$ is the inverse temperature and E denotes the set of edges of the lattice. To construct the tree decomposition of π we iteratively split the 4×4 lattice into two sub-lattices until each node is isolated.

We briefly recall the generic DaC-SMC algorithm introduced in Lindsten et al. (2017), in particular, the algorithm introduced in Section 8.2 is referred to as 'mixture resampling DaC-SMC' in Lindsten et al. (2017); we choose to work with this formulation since the linear-cost version of the algorithm which was the focus of Lindsten et al. (2017) is a special case of the mixture resampling one. We then move onto the theoretical characterisation of DaC-SMC. We introduce two auxiliary results for independent sequences of random measures in Section 8.3 and then use these results to obtain $\mathbb{L}_p$ error estimates and bias bounds. From the $\mathbb{L}_p$ error estimates we extract a strong law of large numbers and show almost sure convergence in the weak topology. Finally, we show that, equivalently to standard SMC, the estimates of the normalising constants provided by DaC-SMC are unbiased and discuss how they can be employed in particle MCMC methods (Andrieu et al., 2010).

8.2 Divide and Conquer Sequential Monte Carlo

To introduce the algorithm, we consider the case in which we are given a probability measure π defined on a measurable space $(\boldsymbol{E}, \boldsymbol{\mathcal{E}})$ and our goal is to approximate

integrals with respect to π. We assume that $\pi := \gamma/\mathcal{Z}$ arises from normalisation of a (finite) target measure γ which can be computed pointwise, whereas evaluating the normalising constant $\mathcal{Z} := \gamma(\boldsymbol{E})$ is computationally challenging. In the case in which the problem of interest is endowed with a tree topology that allows for model decompositions based on the tree structure (e.g. phylogenetic models; Jewell (2015)) the sequence of distributions building up to π is not an artificial construct introduced for computational convenience, but arises from the particular model at hand.

To approximate π, instead of building a sequence of measures $\{\hat{\pi}_n\}_{n\geq 1}$ evolving on a line, as we would do for standard SMC (Chopin, 2002; Del Moral et al., 2006b), we consider sequences of measures $\{\pi_u\}_{u\in\mathbb{T}^{\blacktriangledown}}$ defined on measurable spaces $(\boldsymbol{E}_u, \boldsymbol{\mathcal{E}}_u)$ of increasing dimension indexed by a set $\mathbb{T}^{\blacktriangledown}$ obtained by removing the root node $\mathfrak{r}$ from a finite tree $\mathbb{T}$:

$$\boldsymbol{E} = \prod_{u\in\mathbb{T}^{\blacktriangledown}} E_u, \quad \boldsymbol{\mathcal{E}} = \prod_{u\in\mathbb{T}^{\blacktriangledown}} \mathcal{E}_u. \tag{8.1}$$

The sequence of distributions $\{\pi_u\}_{u\in\mathbb{T}^{\blacktriangledown}}$ is used to decompose the task of sampling from π into simpler problems, then, similarly to standard SMC, a particle population is evolved from the leaves of the tree to its root using a sequence of mutation, reweighting and resampling steps to which, whenever the branches of the tree merge, a coalescence step is added.

Ideally, one would like to use the marginals

$$\pi^u(A) := \pi(A \times \boldsymbol{E}^{\mathscr{u}}) \quad \forall A \in \boldsymbol{\mathcal{E}}_u, \quad \text{with } \boldsymbol{E}^{\mathscr{u}} := \prod_{v\in\mathbb{T}^{\blacktriangledown}\setminus\mathbb{T}_u^{\mathscr{u}}} E_v, \quad \forall u \in \mathbb{T}, \tag{8.2}$$

where $\mathbb{T}_u$ denotes the sub-tree of $\mathbb{T}$ rooted at u (obtained by removing all nodes from $\mathbb{T}$ except for u and its descendants) and $\mathbb{T}_u^{\mathscr{u}} := \{v \in \mathbb{T}_u : v \neq u\}$, to construct the sequence $\{\pi_u\}_{u\in\mathbb{T}^{\blacktriangledown}}$. However, these are seldom available in practice and, instead, we settle for approximations $(\pi_u)_{u\in\mathbb{T}^{\blacktriangledown}} \approx (\pi^u)_{u\in\mathbb{T}^{\blacktriangledown}}$, with the only constraint that the approximation indexed by the root node coincides with the target (i.e. $\gamma_{\mathfrak{r}} = \gamma$ and $\pi_{\mathfrak{r}} = \pi$). In addition, we assume that each π_u is obtained by normalising a (finite) measure γ_u defined on $(\boldsymbol{E}_u, \boldsymbol{\mathcal{E}}_u)$:

$$\boldsymbol{E}_u := \prod_{v\in\mathbb{T}_u^{\mathscr{u}}} E_v, \quad \boldsymbol{\mathcal{E}}_u := \prod_{v\in\mathbb{T}_u^{\mathscr{u}}} \mathcal{E}_v, \tag{8.3}$$

where

$$\pi_u := \frac{\gamma_u}{\mathcal{Z}_u} \text{ with } \mathcal{Z}_u := \gamma_u(\boldsymbol{E}_u), \quad \forall u \in \mathbb{T}. \tag{8.4}$$

8.2.1 Notation

To facilitate the description of the tree $\mathbb{T}$, we employ the notation summarised here.

Trees We define $\mathbb{T}$ to be a rooted directed tree, a directed acyclic graph in which each node has a parent, except the root $\mathfrak{r}$. We label the root of our tree $\mathbb{T}$ by $\mathfrak{r}$ and we use $\mathbb{T}_u$ to denote the sub-tree of $\mathbb{T}$ rooted at u, i.e. that obtained by removing all nodes from $\mathbb{T}$ expect u and its descendants. We denote the number of children of a node u as c_u, refer to them as $u1, u2, \ldots, uc_u$, and denote the set $\{u1, \ldots, uc_u\}$ by $\mathcal{C}_u$. We use $\mathbb{T}^\partial := \{v \in \mathbb{T} : c_v = 0\}$ to denote the set of $\mathbb{T}$'s leaves and $\mathbb{T}^\partial := \{v \in \mathbb{T} : c_v > 0\}$ that of all other nodes and similarly for any sub-tree $\mathbb{T}_u$. We use $[c_u]$ to denote $\{1, 2, \ldots, c_u\}$, $|A|$ to denote the cardinality of any subset A of $\mathbb{T}$ and A^c to denote its complement, $A^c := \mathbb{T} \setminus A$.

Auxiliary Spaces In addition to the sets $\boldsymbol{E}$ and $\boldsymbol{E}_u$ and their corresponding σ-fields $\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{E}}_u$ defined in (8.1) and (8.3), respectively, we define

$$\bar{\boldsymbol{E}}_u := \boldsymbol{E}_u \cup E_u = \prod_{v \in \mathbb{T}_u} E_v, \qquad \bar{\boldsymbol{\mathcal{E}}}_u := \boldsymbol{\mathcal{E}}_u \cup \mathcal{E}_u = \prod_{v \in \mathbb{T}_u} \mathcal{E}_v,$$

the extensions of $\boldsymbol{E}_u$ and $\boldsymbol{\mathcal{E}}_u$ to include u itself. We denote elements in E_u, $\bar{\boldsymbol{E}}_u$, and $\boldsymbol{E}_u$ by x_u, $\bar{\boldsymbol{x}}_u$, and $\boldsymbol{x}_u$, respectively. Out of notational convenience, we set the empty product of spaces to be the empty set (e.g. $E_\emptyset = \boldsymbol{E}_\emptyset = \mathcal{E}_\emptyset = \boldsymbol{\mathcal{E}}_\emptyset = \emptyset$).

8.2.2 The Algorithm

Similarly to standard SMC, DaC-SMC approximates the sequence $\{\pi_u\}_{u \in \mathbb{T}^\forall}$ through a population of weighted particles which evolve from the leaves of the tree to the root. In doing so the algorithm provides an approximation of $\{\pi_u\}_{u \in \mathbb{T}^\forall}$ and their normalising constants $\{\mathcal{Z}_u\}_{u \in \mathbb{T}^\forall}$.

Algorithm 3 describes the basic instance of DaC-SMC, to make the connection with the divide-and-conquer paradigm evident we describe DaC-SMC as a recursive algorithm, which, when called at the root $\mathfrak{r}$ of the tree, recursively calls itself until it reaches the leaves. As in the case of the SIR algorithm (Algorithm 1) introduced in Chapter 2, Algorithm 3 is a prototypical algorithm which can be modified and

extended to more complicated instances, e.g. including tempering in the spirit of SMC samplers (Del Moral et al., 2006a) or swapping the order of the steps when this is more convenient, e.g., with the alternative decomposition described in Example 8.1 the mutation step naturally occurs between steps 12 and 13 of Algorithm 3.

Algorithm 3: DaC-SMC(u) for u in $\mathbb{T}^{\partial}$.

1: **for** v in $\mathcal{C}_u$ **do**
2: **if** v is a leaf (i.e. $v \in \mathbb{T}^{\partial}$) **then**
3: *Propose:* for all $n \leq N$, draw $X_v^{n,N}$ independently from K_v and set $\bar{\boldsymbol{X}}_v^{n,N} := X_v^{n,N}$ and $\mathcal{Z}_v^N := 1$.
4: **else**
5: *Recurse:* set $(\gamma_v^N, \pi_v^N) :=$ DaC-SMC(v).
6: **for** $n = 1, \ldots, N$ **do**
7: *Resample:* draw $\boldsymbol{X}_v^{n,N}$ independently from π_v^N.
8: *Mutate:* draw $X_v^{n,N}$ independently from $K_v(\boldsymbol{X}_v^{n,N}, \cdot)$ and set $\bar{\boldsymbol{X}}_v^{n,N}$ via (8.6).
9: **end for**
10: **end if**
11: **end for**
12: *Merge:* combine $\bar{\boldsymbol{X}}_v^{n,N}$ for $v \in \mathcal{C}_u$ into the unnormalised product-form estimator (8.10).
13: *Correct:* compute the weights (8.11) and return (γ_u^N, π_u^N) where γ_u^N is as in (8.13) with $\mathcal{Z}_{\mathcal{C}_u}^N$ defined by (8.9), and $\pi_u^N := \gamma_u^N / \gamma_u^N(\boldsymbol{E}_u)$.

As discussed earlier, DaC-SMC operates in a recursive manner, when called at a node $u \in \mathbb{T}$ the algorithm obtains the particle populations at the children of u, $v \in \mathcal{C}_u$, evolves them through the classical SMC steps – resampling and mutation – and then combines them and weights them to obtain the particle population at node u.

At each leaf $v \in \mathbb{T}_u^{\partial}$ a population of N independent samples $X_v^{1,N}, \ldots, X_v^{N,N}$ is obtained from a proposal distribution K_v and the corresponding empirical distribution is computed

$$\pi_v^N = \frac{1}{N} \sum_{n=1}^{N} \delta_{X_v^{n,N}}.$$

If $v \in \mathcal{C}_u$ is not a leaf, the weighted particle population at v, γ_v^N, is obtained recursively by calling DaC-SMC(v). Normalising γ_v^N we obtain an approximation π_v^N of π_v:

$$\pi_v^N := \frac{\gamma_v^N}{\mathcal{Z}_v^N} \approx \frac{\gamma_v}{\mathcal{Z}_v} = \pi_v \text{ where } \mathcal{Z}_v^N := \gamma_v^N(\boldsymbol{E}_v), \tag{8.5}$$

for each non-leaf child v.

Following the SMC paradigm, at each $v \in \mathcal{C}_u$ the particles are resampled, this reduces degeneracy in the weights but also contributes to keeping the computational cost of DaC-SMC under control: at each $v \in \mathcal{C}_u$ the particle population is of size N_v^c, where c_v is the number of children of v; without the resampling step the population size would therefore increase exponentially with the number of nodes in the tree, making the computational cost unmanageable. The empirical distribution of the particles $\boldsymbol{X}_v^{1,N}, \ldots, \boldsymbol{X}_v^{N,N}$ resampled from π_v^N approximates π_v and is denoted by

$$\rho_v^N := \frac{1}{N} \sum_{n=1}^{N} \delta_{\boldsymbol{X}_v^{n,N}} \approx \pi_v$$

We then extend the path of each resampled particle from $\boldsymbol{E}_v$ to $\bar{\boldsymbol{E}}_v$ using a Markov kernel $K_v : \boldsymbol{E}_v \times \mathcal{E}_v \to [0, 1]$:

$$\bar{\boldsymbol{X}}_v^{n,N} := (X_v^{n,N}, \boldsymbol{X}_v^{n,N}) \text{ where } X_v^{n,N} \sim K_v(\boldsymbol{X}_v^{n,N}, \cdot) \tag{8.6}$$

and obtain the empirical distribution of the mutated particles

$$\bar{\lambda}_u^N := \frac{1}{N} \sum_{n=1}^{N} \delta_{\bar{\boldsymbol{X}}_u^{n,N}}$$

and its unnormalised version $\bar{v}_u^N := \mathcal{Z}_u^N \bar{\lambda}_u^N$. Permuting and concatenating these particles across u's children, we obtain an approximation

$$\lambda_u^N := \prod_{v \in \mathcal{C}_u} \bar{\lambda}_v^N = \frac{1}{N^{c_u}} \sum_{n_1=1}^{N} \cdots \sum_{n_{c_u}=1}^{N} \delta_{(\bar{\boldsymbol{X}}_{u1}^{n_1,N}, \ldots, \bar{\boldsymbol{X}}_{uc_u}^{n_{c_u},N})}, \tag{8.7}$$

of the normalised flow $\lambda_u := \bar{\lambda}_{\mathcal{C}_u} = \prod_{v \in \mathcal{C}_u} \bar{\lambda}_v$, where $\bar{\lambda}_v := K_v$ if v is a leaf and, otherwise,

$$\bar{\lambda}_v(\mathrm{d}\bar{\boldsymbol{x}}_v) := (\pi_v \otimes K_v)(\mathrm{d}\bar{\boldsymbol{x}}_v) = \pi_v(\mathrm{d}\boldsymbol{x}_v) K_v(\boldsymbol{x}_v, \mathrm{d}x_v). \tag{8.8}$$

Multiplying λ_u^N by the product $\mathcal{Z}_{\mathcal{C}_u}^N$ of the normalising constant estimates,

$$\mathcal{Z}_{\mathcal{C}_u}^N := \prod_{v \in \mathcal{C}_u} \mathcal{Z}_v^N = \prod_{v \in \mathcal{C}_u} \gamma_v^N(\boldsymbol{E}_v) \tag{8.9}$$

(where $\mathcal{Z}_v^N := 1$ if v is a leaf), we obtain an approximation to the unnormalised flow

$$\upsilon_u^N := \mathcal{Z}_{\mathcal{C}_u}^N \lambda_u^N \approx \mathcal{Z}_{\mathcal{C}_u} \lambda_u = \prod_{v \in \mathcal{C}_u} \mathcal{Z}_v \bar{\lambda}_v = \bar{\upsilon}_{\mathcal{C}_u} =: \upsilon_u, \tag{8.10}$$

where $\bar{\upsilon}_v := K_v$ if v is a leaf and $\bar{\upsilon}_v := \gamma_v \otimes K_v$ otherwise.

As in standard SMC, this mutation step is followed by a reweighting step which corrects the errors introduced by approximating γ_u with υ_u: if

$$w_u := \frac{d\gamma_u}{d\upsilon_u} \tag{8.11}$$

denotes the Radon-Nikodým derivative of γ_u with respect to υ_u, then $\gamma_u(\mathrm{d}\boldsymbol{x}_u) = w_u(\boldsymbol{x}_u)\upsilon_u(\mathrm{d}\boldsymbol{x}_u)$ and we obtain the particle approximation of γ_u as

$$\gamma_u^N(\mathrm{d}\boldsymbol{x}_u) := w_u(\boldsymbol{x}_u)\upsilon_u^N(\mathrm{d}\boldsymbol{x}_u) \approx w_u(\boldsymbol{x}_u)\upsilon_u(\mathrm{d}\boldsymbol{x}_u) = \gamma_u(\mathrm{d}\boldsymbol{x}_u). \tag{8.12}$$

Because γ_u^N may be expressed as

$$\gamma_u^N = \frac{\mathcal{Z}_{\mathcal{C}_u}^N}{N^{c_u}} \sum_{n_1=1}^{N} \cdots \sum_{n_{c_u}=1}^{N} w_u(\bar{\boldsymbol{X}}_{u1}^{n_1,N}, \ldots, \bar{\boldsymbol{X}}_{uc_u}^{n_{c_u},N}) \delta_{(\bar{\boldsymbol{X}}_{u1}^{n_1,N}, \ldots, \bar{\boldsymbol{X}}_{uc_u}^{n_{c_u},N})}, \tag{8.13}$$

we say that γ_u^N is a product-form estimator for γ_u (Kuntz et al., 2021b). Similarly in the cases of λ_u^N, υ_u^N, and $\pi_u^N := \gamma_u^N / \mathcal{Z}_u^N$, where $\mathcal{Z}_u^N := \gamma_u^N(\boldsymbol{E}_u)$.

Of course, for this approach to work, we must choose the unnormalised approximating distributions $(\gamma_u)_{u \in \mathbb{T}^\partial}$ and proposal kernels $(K_u)_{u \in \mathbb{T}^\nu}$ such that the Radon-Nikodým derivatives in (8.11) exist. To simplify the treatment, we further assume that the derivatives are positive everywhere:

Assumption 8.1. For all u in $\mathbb{T}$, γ_u is absolutely continuous w.r.t. υ_u and the weight function w_u in (8.11) is positive everywhere: $w_u(\boldsymbol{x}_u) > 0$ for all $\boldsymbol{x}_u$ in $\boldsymbol{E}_u$.

This assumption is equivalent to those usually found in the SMC literature (e.g. Assumption 2.1) and ensures that the weights never simultaneously take the value zero.

We briefly describe how Algorithm 3 would operate on the simple Bayesian hierarchical model in Example 8.1. To simplify the presentation we describe the DaC-SMC procedure starting at the leaves and progressing towards the root:

Example 8.3 (Bayesian Hierarchical Model)**.** Consider the simple Bayesian hierarchical model in Example 8.1 and assume that we want to approximate the posterior

$\pi(\mathbf{x}_H) = p(x_{1:7} \mid y_{1:4})$. For simplicity we assume that we can sample from the marginals $p(x_5 \mid x_{1:2}, y_{1:2})$, $p(x_6 \mid x_{3:4}, y_{3:4})$ and $p(x_7 \mid x_{1:6}, y_{1:4})$.

At the level of the leaves we obtain particle approximations $\pi_A^N, \pi_B^N, \pi_C^N, \pi_D^N$. Following Algorithm 3 the particle approximations π_A^N, π_B^N are merged to obtain $\upsilon_E^N = \pi_A^N \times \pi_B^N$ and reweighted with weights given by $w_E(\mathbf{x}_E) = p(x_{1:2} \mid y_{1:2})/p(x_1 \mid y_1)p(x_2 \mid y_2)$ to obtain the particle approximation $\pi_E(\mathbf{x}_E)$, and similarly for $\pi_F(\mathbf{x}_F)$.

At the successive level of the tree, resampling is performed separately at node E and F. The states $\mathbf{x}_E, \mathbf{x}_F$ are extended to $(x_5, \mathbf{x}_E), (x_6, \mathbf{x}_F)$ by drawing x_5, x_6 from $p(x_5 \mid x_{1:2}, y_{1:2})$ and $p(x_6 \mid x_{3:4}, y_{3:4})$, respectively. Then, the particle populations at nodes E, F are merged and reweighted to obtain an approximation of $\pi_G(\mathbf{x}_G) = p(x_{1:6} \mid y_{1:4})$. At the root level, after resampling from π_G^N, the new particle population is extended by drawing $x_7 \sim p(\cdot \mid x_{1:6}, y_{1:2})$ and reweighting according to w_H (no merging occurs since H has only one child G).

8.3 Products of Independent Random Measures

As a first step towards providing a theoretical characterisation of DaC-SMC, we consider two sequences of independent random measures $(\eta_1^N)_{N=1}^\infty$ defined over $(S_1, \mathcal{S}_1)$ and $(\eta_2^N)_{N=1}^\infty$ defined over $(S_2, \mathcal{S}_2)$ approximating the probability measures η_1 and η_2, respectively. We will show that, if both $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$ satisfy an $\mathbb{L}_p$ inequality and a bias estimate, then so does the product measure $\eta_1^N \times \eta_2^N$.

We endow the product $S = S_1 \times S_2$ of two measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ with the product sigma algebra $\mathcal{S}_1 \times \mathcal{S}_2$. Given any finite measure η_1 on $(S_1, \mathcal{S}_1)$ and measurable test function $\varphi : S \to \mathbb{R}$ such that $\varphi(\cdot, x_2)$ is η_1-integrable for each x_2 in S_2, we use $\eta_1(\varphi)$ to denote the measurable function on S_2 defined by

$$\eta_1(\varphi)(x_2) := \int \varphi(x_1, x_2)\eta_1(\mathrm{d}x_1) \quad \forall x_2 \in S_2.$$

In the case of the empirical measures η_1^N, η_2^N, we occasionally make the dependence on a particular realisation ω explicit

$$\eta_1^N(\omega, \varphi)(x_2) := \int \varphi(x_1, x_2)\eta_1^N(\omega, \mathrm{d}x_1) \quad \forall x_2 \in S_2.$$

The results for the product measures are obtained exploiting the decomposition

$$\begin{aligned}(\eta_1^N \times \eta_2^N - \eta_1 \times \eta_2)(\varphi) =& (\eta_1^N - \eta_1) \times \eta_2(\varphi) + \eta_1 \times (\eta_2^N - \eta_2)(\varphi) \\ &+ (\eta_1^N - \eta_1) \times (\eta_2^N - \eta_2)(\varphi),\end{aligned} \tag{8.14}$$

whose first two terms can be controlled via the individual results for $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$ and the third one using their independence.

Lemma 8.1. Suppose that $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$ are independent sequences of random probability measures on some measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$ that, for every $N \in \mathbb{N}$, $\varphi \in \mathcal{B}_b(S_k)$, for $k \in \{1, 2\}$, satisfy

$$\mathbb{E}\left[\left|\eta_k^N(\varphi) - \eta_k(\varphi)\right|^p\right]^{1/p} \leq \frac{C_{p,k}\|\varphi\|_\infty}{N^{1/2}} \tag{8.15}$$

for some $p \geq 1$ and $C_{p,k}$ finite constants. Then, for every $N \in \mathbb{N}$, $\varphi \in \mathcal{B}_b(S_1 \times S_2)$.

$$\mathbb{E}\left[\left|(\eta_1^N \times \eta_2^N)(\varphi) - (\eta_1 \times \eta_2)(\varphi)\right|^p\right]^{1/p} \leq \frac{C_p\|\varphi\|_\infty}{N^{1/2}} \tag{8.16}$$

Proof. Fix any $N \in \mathbb{N}$ and φ in $\mathcal{B}_b(S_1 \times S_2)$. Minkowski's inequality and the decomposition in (8.14) tell us that

$$\begin{aligned}\mathbb{E}\left[\left|(\eta_1^N \times \eta_2^N - \eta_1 \times \eta_2)(\varphi)\right|^p\right]^{1/p} \leq\ & \mathbb{E}\left[\left|(\eta_1^N - \eta_1) \times (\eta_2^N - \eta_2)(\varphi)\right|^p\right]^{1/p} \\ &+ \mathbb{E}\left[\left|\eta_1^N(\eta_2(\varphi)) - \eta_1(\eta_2(\varphi))\right|^p\right]^{1/p} \\ &+ \mathbb{E}\left[\left|\eta_2^N(\eta_1(\varphi)) - \eta_2(\eta_1(\varphi))\right|^p\right]^{1/p}.\end{aligned} \tag{8.17}$$

Because φ is bounded and η_1 and η_2 are probability measures, $\eta_1(\varphi)$ and $\eta_2(\varphi)$ are measurable functions bounded above by $\|\varphi\|_\infty$. Hence, we can apply the $\mathbb{L}_p$ inequalities in (8.15) to control the rightmost two terms in (8.17)

$$\mathbb{E}\left[\left|\eta_1^N(\eta_2(\varphi)) - \eta_1(\eta_2(\varphi))\right|^p\right]^{1/p} \leq \frac{C_{p,1}\|\eta_2(\varphi)\|_\infty}{N^{1/2}} \leq \frac{C_{p,1}\|\varphi\|_\infty}{N^{1/2}}, \tag{8.18}$$

$$\mathbb{E}\left[\left|\eta_2^N(\eta_1(\varphi)) - \eta_2(\eta_1(\varphi))\right|^p\right]^{1/p} \leq \frac{C_{p,2}\|\eta_1(\varphi)\|_\infty}{N^{1/2}} \leq \frac{C_{p,2}\|\varphi\|_\infty}{N^{1/2}} \tag{8.19}$$

where $\|\eta_1(\varphi)\|_\infty \leq \|\varphi\|_\infty$ and $\|\eta_2(\varphi)\|_\infty \leq \|\varphi\|_\infty$ because η_1 and η_2 have mass of one. To control the leftmost term in (8.17), let $\mathcal{F}_2 := \sigma(\{\eta_2^N\}_{N=1}^\infty)$ denote the σ-algebra generated by the η_2^Ns. Because, for each ω, $x_1 \mapsto \eta_2^N(\omega, \varphi)(x_1) - \eta_2(\varphi)(x_1)$ is a bounded function on S_1, (8.15) and the independence of $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$

imply that

$$\begin{aligned}
&\mathbb{E}\left[\left|(\eta_1^N - \eta_1) \times (\eta_2^N - \eta_2)(\varphi)\right|^p\right]^{1/p} \qquad (8.20)\\
&= \mathbb{E}\left[\left|\eta_1^N(\eta_2^N(\varphi) - \eta_2(\varphi)) - \eta_1(\eta_2^N(\varphi) - \eta_2(\varphi))\right|^p\right]^{1/p}\\
&= \mathbb{E}\left[\mathbb{E}\left[\left|\eta_1^N(\eta_2^N(\varphi) - \eta_2(\varphi)) - \eta_1(\eta_2^N(\varphi) - \eta_2(\varphi))\right|^p \mid \mathcal{F}_2\right]\right]^{1/p}\\
&\leq \frac{C_{p,1}\,\mathbb{E}\left[\|\eta_2^N(\varphi) - \eta_2(\varphi)\|_\infty^p\right]^{1/p}}{N^{1/2}}\\
&\leq \frac{2C_{p,1}\|\varphi\|_\infty}{N^{1/2}}.
\end{aligned}$$

Putting (8.17)–(8.20) together, we obtain the $\mathbb{L}_p$ inequality (8.16) for the product. □

Similarly, by using the independence of $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$ we can show that the bias of the product measure decays at rate $1/N$:

Lemma 8.2. Suppose that $(\eta_1^N)_{N=1}^\infty$ and $(\eta_2^N)_{N=1}^\infty$ are independent sequences of random probability measures on some measurable spaces $(S_1, \mathcal{S}_1)$ and $(S_2, \mathcal{S}_2)$, respectively, satisfying bias estimates:

$$\left|\mathbb{E}\left[\eta_k^N(\varphi_k)\right] - \eta_k(\varphi_k)\right| \leq \frac{C_k\|\varphi_k\|_\infty}{N} \qquad (8.21)$$

for every $N \in \mathbb{N}$, $\varphi_k \in \mathcal{B}_b(S_k)$ and $k \in \{1, 2\}$, with C_k finite positive constants independent of φ and N. The sequence of products $(\eta_1^N \times \eta_2^N)_{N=1}^\infty$ satisfies a bias estimate:

$$\left|\mathbb{E}\left[(\eta_1^N \times \eta_2^N)(\varphi)\right] - (\eta_1 \times \eta_2)(\varphi)\right| \leq \frac{C\|\varphi\|_\infty}{N} \qquad (8.22)$$

for every $N \in \mathbb{N}$, $\varphi \in \mathcal{B}_b(S_1 \times S_2)$, with C a finite positive constant independent of φ and N.

Proof. Recall the decomposition in (8.14). Because φ is bounded, $\eta_1(\varphi)(x_2) := \int \varphi(x_1, x_2)\eta_1(dx_1)$ and $\eta_2(\varphi)(x_1) := \int \varphi(x_1, x_2)\eta_2(dx_2)$ define bounded measurable functions on S_1 and S_2, respectively. Hence, we can apply the bias estimates in (8.21) to control the rightmost two terms in (8.14):

$$\left|\mathbb{E}\left[\eta_1^N(\eta_2(\varphi))\right] - \eta_1(\eta_2(\varphi))\right| \leq \frac{C_1\|\eta_2(\varphi)\|_\infty}{N} \leq \frac{C_1\|\varphi\|_\infty}{N}, \qquad (8.23)$$

$$\left|\mathbb{E}\left[\eta_2^N(\eta_1(\varphi))\right] - \eta_2(\eta_1(\varphi))\right| \leq \frac{C_2\|\eta_1(\varphi)\|_\infty}{N} \leq \frac{C_2\|\varphi\|_\infty}{N}. \qquad (8.24)$$

To control the leftmost term in (8.14), let $\mathcal{F}_2 := \sigma(\{\eta_2^N\}_{N=1}^{\infty})$, the σ-algebra generated by the η_2^Ns. Because, for each ω, $x_1 \mapsto \eta_2^N(\omega, \varphi)(x_1) - \eta_2(\varphi)(x_1)$ is a bounded function on S_1, (8.21) and the independence of $(\eta_1^N)_{N=1}^{\infty}$ and $(\eta_2^N)_{N=1}^{\infty}$ imply that

$$\begin{aligned}
&\left|\mathbb{E}\left[(\eta_1^N - \eta_1) \times (\eta_2^N - \eta_2)(\varphi)\right]\right| \\
&= \left|\mathbb{E}\left[\eta_1^N(\eta_2^N(\varphi) - \eta_2(\varphi)) - \eta_1(\eta_2^N(\varphi) - \eta_2(\varphi))\right]\right| \\
&= \left|\mathbb{E}\left[\mathbb{E}\left[\eta_1^N(\eta_2^N(\varphi) - \eta_2(\varphi)) - \eta_1(\eta_2^N(\varphi) - \eta_2(\varphi)) \mid \mathcal{F}_2\right]\right]\right| \\
&\leq \mathbb{E}\left[\left|\mathbb{E}\left[\eta_1^N(\eta_2^N(\varphi) - \eta_2(\varphi)) - \eta_1(\eta_2^N(\varphi) - \eta_2(\varphi)) \mid \mathcal{F}_2\right]\right|\right] \\
&\leq \frac{C_1\,\mathbb{E}\left[\|\eta_2^N(\varphi) - \eta_2(\varphi)\|_{\infty}\right]}{N} \\
&\leq \frac{2C_1\|\varphi\|_{\infty}}{N}.
\end{aligned}$$

Putting the above and (8.23)–(8.24) together, we obtain the bias estimate (8.22) for the product. □

8.4 Theoretical Characterisation of DaC-SMC

The theoretical characterisation of DaC-SMC is obtained under the following assumption, which is analogous to those under which the results for standard SMC hold (Assumption 2.1):

Assumption 8.2. For all u in $\mathbb{T}$, w_u in (8.11) is bounded and measurable.

The main difference between DaC-SMC (Algorithm 3) and standard SMC (Algorithm 1) is the merging step: in the former case, the particles on different branches are merged through the product-form estimator (8.7) while, in the latter, the particle population evolves on a line and no merging occurs. Because the particles evolving on different branches at the same depth of the tree $\mathbb{T}$ are independent, the empirical measure (8.7) is a product of c_u independent empirical measures. Thus, we use the results in Section 8.3 to obtain $\mathbb{L}_p$ inequalities and bias estimates for (8.7) and incorporate these results into the well-known methods used to establish $\mathbb{L}_p$ inequalities and bias estimates for standard SMC (Crisan and Doucet, 2000, 2002; Míguez et al., 2013; Olsson and Rydén, 2004).

8.4.1 $\mathbb{L}_p$ Inequality and Strong Law of Large Numbers

To obtain the $\mathbb{L}_p$ inequalities we follow the iterative approach taken in Crisan and Doucet (2000, 2002); Míguez et al. (2013): we first obtain $\mathbb{L}_p$ inequalities for

the empirical distributions of the particles indexed by the tree's leaves, then we inductively show that the steps of Algorithm 3 preserve the inequalities. The resampling and mutation steps follow from arguments analogous to those in Crisan and Doucet (2000, 2002); Míguez et al. (2013). For the merging step, we use the $\mathbb{L}_p$ inequality for product measures in Lemma 8.1 to show that the error introduced by multiplying the approximations over different branches as in (8.7) is also of order $\mathcal{O}(N^{-1/2})$. To complete the argument for the correction step, we then resume with the approach of Crisan and Doucet (2000, 2002); Míguez et al. (2013) and show that the re-weighting in (8.13) preserves the inequalities. Since the proof of the $\mathbb{L}_p$ inequalities stated below requires additional results controlling each one of the steps described above, we postpone the proof to Section 8.4.2.

Proposition 8.1 ($\mathbb{L}_p$ inequalities)**.** If Assumptions 8.1–8.2 hold, then, for each $p \geq 1$, there exist constants $C_{u,p}, \widetilde{C}_{u,p}, \widehat{C}_{u,p}, \ddot{C}_{u,p} < \infty$ such that

$$\mathbb{E}\left[\left|\upsilon_u^N(\varphi) - \upsilon_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\widehat{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}}, \quad \mathbb{E}\left[\left|\lambda_u^N(\varphi) - \lambda_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\ddot{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}},$$
$$\mathbb{E}\left[\left|\gamma_u^N(\varphi) - \gamma_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\widetilde{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}}, \quad \mathbb{E}\left[\left|\pi_u^N(\varphi) - \pi_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{C_{u,p}\|\varphi\|_\infty}{N^{1/2}},$$

for all $N > 0$, all φ in $\mathcal{B}_b(\boldsymbol{E}_u)$ and every $u \in \mathbb{T}$. In particular, $\mathbb{E}\left[\left|\mathcal{Z}_u^N - \mathcal{Z}_u\right|^p\right]^{1/p} \leq \widetilde{C}_{u,p}/N^{1/2}$ for all $N > 0$.

A direct consequence of Proposition 8.1 is the following strong law of large numbers, which shows that Algorithm 3 produces strongly consistent estimates of the unnormalised and normalised targets γ and π: $\gamma^N(\varphi)$ and $\pi^N(\varphi)$ are consistent estimators for $\pi(\varphi)$ and $\gamma(\varphi)$, respectively. This also holds all approximating distributions and their estimators:

Corollary 8.1 (Strong laws of large numbers)**.** If Assumptions 8.1–8.2 are satisfied, u belongs to $\mathbb{T}$, and φ belongs to $\mathcal{B}_b(\boldsymbol{E}_u)$, then

$$\lim_{N\to\infty} \upsilon_u^N(\varphi) = \upsilon_u(\varphi), \qquad \lim_{N\to\infty} \lambda_u^N(\varphi) = \lambda_u(\varphi),$$
$$\lim_{N\to\infty} \gamma_u^N(\varphi) = \gamma_u(\varphi), \qquad \lim_{N\to\infty} \pi_u^N(\varphi) = \pi_u(\varphi),$$

almost surely. In particular, $\mathcal{Z}_u^N \to \mathcal{Z}_u$ as $N \to \infty$ with probability one.

Proof. The result follows from the $\mathbb{L}_p$ inequalities in Proposition 8.1 using a Borel-Cantelli argument as shown in the proof of Proposition 5.1. □

If the underlying spaces possess nice enough topological properties, the above almost sure pointwise convergence can be strengthened to almost sure convergence in the weak topology:

Corollary 8.2 (Almost sure weak convergence)**.** If, in addition to Assumptions 8.1–8.2, the spaces $(E_u)_{u\in\mathbb{T}^\forall}$ are Polish and $(\mathcal{E}_u)_{u\in\mathbb{T}}$ are the corresponding Borel sigma algebras, then

$$\lim_{N\to\infty} \upsilon_u^N(\varphi) = \upsilon_u(\varphi), \qquad \lim_{N\to\infty} \lambda_u^N(\varphi) = \lambda_u(\varphi),$$
$$\lim_{N\to\infty} \gamma_u^N(\varphi) = \gamma_u(\varphi), \qquad \lim_{N\to\infty} \pi_u^N(\varphi) = \pi_u(\varphi),$$

for all bounded continuous functions φ on $\boldsymbol{E}_u$ and u in $\mathbb{T}$, almost surely.

Proof. Because the space of continuous bounded functions on $\boldsymbol{E}_u$ has a countable dense (w.r.t. $\|\cdot\|_\infty$) subset if $\boldsymbol{E}_u$ is Polish (c.f. Berti et al. (2006, Theorem 2.2)) and the product of finitely many Polish spaces is Polish, this follows from Corollary 8.1 and the Portmanteau Theorem (Billingsley, 1995, Theorem 29.1), as shown in the proof of Proposition 5.3. □

8.4.2 Proof of Proposition 8.1

To argue Proposition 8.1, we follow an approach similar to that taken in Crisan and Doucet (2000, 2002); Míguez et al. (2013) for standard SMC. In particular, we derive an $\mathbb{L}_p$ inequality for the empirical distributions of the particles indexed by the tree's leaves and show that each step of the algorithm preserves this inequality. Recall that at the leaves of the tree, $u \in \mathbb{T}^\partial$ the particle are sampled i.i.d. from K_u, then, Del Moral (2004, Lemma 7.3.3) shows that, for any $p \geq 1$, there exist $\bar{C}_{u,p} < \infty$ such that

$$\mathbb{E}\left[\left|\bar{\lambda}_u^N(\varphi) - \bar{\lambda}_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\bar{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}} \quad \forall N > 0,\ \ \varphi \in \mathcal{B}_b(\bar{\boldsymbol{E}}_u). \tag{8.25}$$

Taking the product of $\bar{\lambda}_v^N$ over all children v of a node u to obtain λ_u^N in (8.7) preserves (8.25):

Lemma 8.3 (Product step)**.** If, in addition to Assumptions 8.1–8.2, (8.25) is satisfied for each child v (i.e. it holds with v replacing u therein) of a node u in $\mathbb{T}^\partial$ and some $p \geq 1$, then there exist $\ddot{C}_{u,p} < \infty$ such that

$$\mathbb{E}\left[\left|\lambda_u^N(\varphi) - \lambda_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\ddot{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}} \quad \forall N > 0,\ \ \varphi \in \mathcal{B}_b(\boldsymbol{E}_u). \tag{8.26}$$

Proof. Because $(\bar{\lambda}_{u1}^N)_{N=1}^\infty, \ldots, (\bar{\lambda}_{uc_u}^N)_{N=1}^\infty$ are independent sequences of probability measures by construction, so are $(\bar{\lambda}_{u[k]}^N)_{N=1}^\infty$ and $(\bar{\lambda}_{u(k+1)}^N)_{N=1}^\infty$ for all $k < c_u$, where $u[k] := \{u1, \ldots, uk\}$ denotes the set containing the first k children of u and $\bar{\lambda}_{u[k]}^N := \prod_{v \in u[k]} \bar{\lambda}_v^N$ the corresponding product of $\bar{\lambda}_v^N$s. The Lemma's premise requires that each child satisfies its own $\mathbb{L}_p$ inequality and, consequently, (8.26) follows from repeated applications of Lemma 8.1. □

Emulating the approach of Crisan and Doucet (2000, Lemma 2) and Míguez et al. (2013, Lemma 1), we find that the correction step also respects the inequality:

Lemma 8.4 (Correction step)**.** If, in addition to Assumptions 8.1–8.2, (8.26) is satisfied for some u in $\mathbb{T}^\vartheta$ and $p \geq 1$, then there exist $C_{u,p} < \infty$ such that

$$\mathbb{E}\left[\left|\pi_u^N(\varphi) - \pi_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{C_{u,p}\|\varphi\|_\infty}{N^{1/2}} \quad \forall N > 0, \;\; \varphi \in \mathcal{B}_b(\boldsymbol{E}_u). \tag{8.27}$$

Proof. Fix any $N > 0$ and φ in $\mathcal{B}_b(\boldsymbol{E}_u)$. Recall the definitions in Section 8.2:

$$\upsilon_u^N(\mathrm{d}\boldsymbol{x}_u) = \mathcal{Z}_{\mathcal{C}_u}^N \lambda_u^N(\mathrm{d}\boldsymbol{x}_u), \quad \gamma_u^N(\mathrm{d}\boldsymbol{x}_u) = w_u(\boldsymbol{x}_u)\upsilon_u^N(\mathrm{d}\boldsymbol{x}_u), \quad \pi_u^N(\mathrm{d}\boldsymbol{x}_u) = \frac{\gamma_u^N(\mathrm{d}\boldsymbol{x}_u)}{\gamma_u^N(\boldsymbol{E}_u)}.$$

Hence,

$$\pi_u^N(\varphi) = \frac{\gamma_u^N(\varphi)}{\gamma_u^N(\boldsymbol{E}_u)} = \frac{\upsilon_u^N(w_u\varphi)}{\upsilon_u^N(w_u)} = \frac{\lambda_u^N(w_u\varphi)}{\lambda_u^N(w_u)}.$$

Similarly, because $w_u = d\gamma_u/d\upsilon_u$ and $\upsilon_u = \gamma_{\mathcal{C}_u} \otimes K_{\mathcal{C}_u} = \mathcal{Z}_{\mathcal{C}_u}\lambda_u$,

$$\pi_u(\varphi) = \frac{\gamma_u(\varphi)}{\gamma_u(\boldsymbol{E}_u)} = \frac{\upsilon_u(w_u\varphi)}{\upsilon_u(w_u)} = \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)}.$$

Hence,

$$\left|\pi_u^N(\varphi) - \pi_u(\varphi)\right| \leq \left|\pi_u^N(\varphi) - \frac{\lambda_u^N(w_u\varphi)}{\lambda_u(w_u)}\right| + \left|\frac{\lambda_u^N(w_u\varphi)}{\lambda_u(w_u)} - \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)}\right|. \tag{8.28}$$

To control the first term on the right-hand side, we use

$$\left|\pi_u^N(\varphi) - \frac{\lambda_u^N(w_u\varphi)}{\lambda_u(w_u)}\right| \leq \frac{\left|\pi_u^N(\varphi)\right|\left|\lambda_u(w_u) - \lambda_u^N(w_u)\right|}{\lambda_u(w_u)} \leq \frac{\|\varphi\|_\infty\left|\lambda_u(w_u) - \lambda_u^N(w_u)\right|}{\lambda_u(w_u)}. \tag{8.29}$$

Because, $\lambda_u(w_u) = \mathcal{Z}_{\mathcal{C}_u}^{-1}\upsilon_u(w_u) = \mathcal{Z}_{\mathcal{C}_u}^{-1}\gamma_u(\boldsymbol{E}_u) = \mathcal{Z}_{\mathcal{C}_u}^{-1}\mathcal{Z}_u > 0$, the desired $\mathbb{L}_p$ inequal-

ity (8.27) follows from (8.26),(8.28),(8.29) and Minkowski's inequality:

$$\begin{aligned}
&\mathbb{E}\left[\left|\pi_u^N(\varphi) - \pi_u(\varphi)\right|^p\right]^{\frac{1}{p}} \\
&\qquad \leq \frac{\|\varphi\|_\infty \mathbb{E}\left[\left|\lambda_u(w_u) - \lambda_u^N(w_u)\right|^p\right]^{\frac{1}{p}} + \mathbb{E}\left[\left|\lambda_u(w_u\varphi) - \lambda_u^N(w_u\varphi)\right|^p\right]^{\frac{1}{p}}}{\lambda_u(w_u)} \\
&\qquad \leq \frac{\ddot{C}_u\|\varphi\|_\infty\|w_u\|_\infty + \ddot{C}_u\|w_u\varphi\|_\infty}{N^{1/2}\lambda_u(w_u)} \leq \left(\frac{2\ddot{C}_u\mathcal{Z}_{\mathcal{C}_u}\|w_u\|_\infty}{\mathcal{Z}_u}\right)\frac{\|\varphi\|_\infty}{N^{1/2}}
\end{aligned}$$

with $C_{u,p} = 2\ddot{C}_{u,p}\mathcal{Z}_{\mathcal{C}_u}\|w_u\|_\infty/\mathcal{Z}_u$. □

To show that the resampling step also preserves $\mathbb{L}_p$ inequalities, we use the argument of Crisan and Doucet (2000, Lemma 3) as for the proof of Lemma B.3:

Lemma 8.5 (Resampling step)**.** If, in addition to Assumptions 8.1–8.2, the inequality (8.27) holds for some u in $\mathbb{T}^\partial$ and $p \geq 1$, then there exist $\dot{C}_{u,p} < \infty$

$$\mathbb{E}\left[\left|\rho_u^N(\varphi) - \pi_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\dot{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}} \quad \forall N > 0,\ \ \varphi \in \mathcal{B}_b(\boldsymbol{E}_u). \tag{8.30}$$

Proof. Fix any $N > 0$ and φ in $\mathcal{B}_b(\boldsymbol{E}_u)$. If $\mathcal{F} := \sigma((\bar{\boldsymbol{X}}_{u1}^{n,N}, \ldots, \bar{\boldsymbol{X}}_{uc_u}^{n,N})_{n=1}^N)$ denotes the σ-field generated by the particles prior to resampling, then

$$\mathbb{E}\left[\varphi(\boldsymbol{X}_u^{n,N})|\mathcal{F}\right] = \pi_u^N(\varphi) \quad \text{almost surely, for all } n \leq N.$$

Let us define, for $n = 1, \ldots, N$, the functions $\Delta_u^{n,N} : \boldsymbol{E}_u \mapsto \mathbb{R}$

$$\Delta_u^{n,N}(x) := \varphi(x) - \mathbb{E}\left[\varphi(\boldsymbol{X}_u^{n,N})|\mathcal{F}\right].$$

Conditionally on $\mathcal{F}$, $\Delta_u^{n,N}(\boldsymbol{X}_u^{n,N})$, $n = 1, \ldots, N$ are independent, have expectation equal to 0, and

$$\rho_u^N(\varphi) - \pi_u^N(\varphi) = \frac{1}{N}\sum_{n=1}^N \Delta_u^{n,N}(\boldsymbol{X}_u^{n,N}).$$

Conditioning on $\mathcal{F}$ and applying the Marcinkiewicz-Zygmund type inequality in Del Moral (2004, Lemma 7.3.3) shows that

$$\mathbb{E}\left[\left|\rho_u^N(\varphi) - \pi_u^N(\varphi)\right|^p\right]^{1/p} \leq \frac{4b(p)^{1/p}\|\varphi\|_\infty}{N^{1/2}}$$

with $b(p)$ as in (B.1). Inequality (8.30) then follows from the above, Minkowski's

inequality, and (8.27). □

For the mutation step, we follow Crisan and Doucet (2000, Lemma 1) as in Lemma B.4:

Lemma 8.6 (Mutation step)**.** If, in addition to Assumptions 8.1–8.2, (8.30) is satisfied for some u in $\mathbb{T}^{\partial}$ and $p \geq 1$, then there exist $\bar{C}_{u,p} < \infty$

$$\mathbb{E}\left[\left|\bar{\lambda}_u^N(\varphi) - \bar{\lambda}_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\bar{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}} \quad \forall N > 0, \;\; \varphi \in \mathcal{B}_b(\bar{\boldsymbol{E}}_u).$$

Proof. Fix any $N > 0$ and φ in $\mathcal{B}_b(\boldsymbol{E}_u)$. Minkowski's inequality implies that

$$\mathbb{E}\left[\left|\bar{\lambda}_u^N(\varphi) - \bar{\lambda}_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \mathbb{E}\left[\left|\bar{\lambda}_u^N(\varphi) - \nu_u^N(\varphi)\right|^p\right]^{\frac{1}{p}} + \mathbb{E}\left[\left|\nu_u^N(\varphi) - \bar{\lambda}_u(\varphi)\right|^p\right]^{\frac{1}{p}}, \tag{8.31}$$

where $\nu_u^N := \rho_u^N \otimes K_u$. Because $\bar{\lambda}_u = \pi_u \otimes K_u$ and $\|K_u\varphi\|_\infty \leq \|\varphi\|_\infty$ as K_u is a Markov kernel, (8.30) implies that

$$\mathbb{E}\left[\left|\nu_u^N(\varphi) - \bar{\lambda}_u(\varphi)\right|^p\right]^{\frac{1}{p}} \leq \frac{\dot{C}_{u,p}\|K_u\varphi\|_\infty}{N^{1/2}} \leq \frac{\dot{C}_{u,p}\|\varphi\|_\infty}{N^{1/2}}. \tag{8.32}$$

To control the other term in (8.31), let $\mathcal{F} := \sigma((\boldsymbol{X}_u^{n,N})_{n=1}^N)$ denote the sigma algebra generated by the resampled particles and note that

$$\mathbb{E}\left[\varphi(\bar{\boldsymbol{X}}_u^{n,N})|\mathcal{F}\right] = (K_u\varphi)(\boldsymbol{X}_u^{n,N}) \quad \forall n \leq N.$$

Let us define, for $n = 1, \ldots, N$, the functions $\Delta_u^{n,N} : \bar{\boldsymbol{E}}_u \mapsto \mathbb{R}$

$$\Delta_u^{n,N}(x) := \varphi(x) - \mathbb{E}\left[\varphi(\bar{\boldsymbol{X}}_u^{n,N}) \mid \mathcal{F}\right].$$

Conditionally on $\mathcal{F}$, $\Delta_u^{n,N}(\bar{\boldsymbol{X}}_u^{n,N})$, $i = 1, \ldots, N$ are independent, have expectation equal to 0, and

$$\bar{\lambda}_u^N(\varphi) - \nu_u^N(\varphi) = \frac{1}{N}\sum_{n=1}^N \Delta_u^{n,N}(\bar{\boldsymbol{X}}_u^{n,N}).$$

Conditioning on $\mathcal{F}$ and applying the Marcinkiewicz-Zygmund type inequality in Del Moral (2004, Lemma 7.3.3) shows that

$$\mathbb{E}\left[\left|\bar{\lambda}_u^N(\varphi) - \nu_u^N(\varphi)\right|^p\right]^{1/p} \leq \frac{4b(p)^{1/p}\|\varphi\|_\infty}{N^{1/2}}, \tag{8.33}$$

with $b(p)$ as in (B.1). Combining (8.31)–(8.33) completes the proof. □

The proof of Proposition 8.1 then follows applying Lemmata 8.3, 8.4, 8.5, and 8.6 within an inductive argument:

Proof of Proposition 8.1. Repeatedly applying Lemmata 8.3, 8.4, 8.5, and 8.6 starting from (8.25), we obtain the inequalities for λ_u^N and π_u^N. Suppose that we are able to argue that for some $\widetilde{C}_{u,p} < \infty$

$$\mathbb{E}\left[\left|\mathcal{Z}_{\mathcal{C}_u}^N - \mathcal{Z}_{\mathcal{C}_u}\right|^p\right]^{\frac{1}{p}} \leq \frac{\max_{\emptyset \neq A \subseteq \mathcal{C}_{u,p}} \mathcal{Z}_{A^c} \widetilde{C}_{A,p}}{N^{1/2}} \quad \forall N > 0,\ \ u \in \mathbb{T},\ \ p \geq 1, \tag{8.34}$$

$$\mathbb{E}\left[\left|\mathcal{Z}_u^N - \mathcal{Z}_u\right|^p\right]^{\frac{1}{p}} \leq \frac{\widetilde{C}_{u,p}}{N^{1/2}} \quad \forall N > 0,\ \ u \in \mathbb{T},\ \ p \geq 1, \tag{8.35}$$

where $\widetilde{C}_{A,p} := \prod_{v \in A} \widetilde{C}_{v,p}$ and $\mathcal{Z}_{A^c} := \prod_{v \in \mathcal{C}_u \setminus A} \mathcal{Z}_v$ for all A in $\mathcal{C}_u$. Then, the inequalities for υ_u^N and γ_u^N would follow from those for λ_u^N and π_u^N because

$$\begin{aligned}
\left|\upsilon_u^N(\varphi) - \upsilon_u(\varphi)\right| &= \left|\mathcal{Z}_{\mathcal{C}_u}^N \lambda_u^N(\varphi) - \mathcal{Z}_{\mathcal{C}_u} \lambda_u(\varphi)\right| \\
&\leq \left|\lambda_u^N(\varphi)\right| \left|\mathcal{Z}_{\mathcal{C}_u}^N - \mathcal{Z}_{\mathcal{C}_u}\right| + \mathcal{Z}_{\mathcal{C}_u} \left|\lambda_u^N(\varphi) - \lambda_u(\varphi)\right| \\
&\leq \|\varphi\|_\infty \left|\mathcal{Z}_{\mathcal{C}_u}^N - \mathcal{Z}_{\mathcal{C}_u}\right| + \mathcal{Z}_{\mathcal{C}_u} \left|\lambda_u^N(\varphi) - \lambda_u(\varphi)\right|
\end{aligned}$$

for all $N > 0$, u in $\mathbb{T}^\partial$, and $p \geq 1$, and similarly for $\gamma_u^N, \mathcal{Z}_u^N$, and π_u^N. In the case of a leaf u, (8.35) is trivially satisfied because $\mathcal{Z}_u^N = \mathcal{Z}_u = 1$ by definition. Suppose, instead, that u is not a leaf and that (8.35) holds for each of its children:

$$\mathbb{E}\left[\left|\mathcal{Z}_v^N - \mathcal{Z}_v\right|^p\right]^{\frac{1}{p}} \leq \frac{\widetilde{C}_{v,p}}{N^{1/2}} \quad \forall N > 0,\ \ v \in \mathcal{C}_u,\ \ p \geq 1. \tag{8.36}$$

Using a multinomial expansion of the product of normalising constants we obtain

$$\mathcal{Z}_{\mathcal{C}_u}^N = \prod_{v \in \mathcal{C}_u} \mathcal{Z}_v^N = \prod_{v \in \mathcal{C}_u} \left[(\mathcal{Z}_v^N - \mathcal{Z}_v) + \mathcal{Z}_v\right] = \sum_{A \subseteq \mathcal{C}_u} \left(\prod_{v \in A} (\mathcal{Z}_v^N - \mathcal{Z}_v)\right) \mathcal{Z}_{A^c} \quad \forall N > 0,$$

where $\mathcal{Z}_{A^c} := \prod_{v \in \mathcal{C}_u \setminus A} \mathcal{Z}_v$. The independence of $(\boldsymbol{X}_{u1}^{n,N})_{n=1}^N, \ldots, (\boldsymbol{X}_{uc_u}^{n,N})_{n=1}^N$ and

Minkowski's inequality imply (8.34):

$$\begin{aligned}\mathbb{E}\left[\left|\mathcal{Z}^N_{\mathcal{C}_u}-\mathcal{Z}_{\mathcal{C}_u}\right|^p\right]^{\frac{1}{p}} &\leq \sum_{\emptyset\neq A\subseteq\mathcal{C}_u}\mathbb{E}\left[\left|\prod_{v\in A}(\mathcal{Z}^N_v-\mathcal{Z}_v)\right|^p\right]^{\frac{1}{p}}\mathcal{Z}_{A^c}\\ &= \sum_{\emptyset\neq A\subseteq\mathcal{C}_u}\mathcal{Z}_{A^c}\prod_{v\in A}\mathbb{E}\left[\left|\mathcal{Z}^N_v-\mathcal{Z}_v\right|^p\right]^{\frac{1}{p}} \leq \sum_{\emptyset\neq A\subseteq\mathcal{C}_u}\mathcal{Z}_{A^c}\prod_{v\in A}\frac{\widetilde{C}_{v,p}}{N^{1/2}}\\ &\leq \frac{\max_{\emptyset\neq A\subseteq\mathcal{C}_u}\mathcal{Z}_{A^c}\widetilde{C}_{A,p}}{N^{1/2}} \quad \forall N>0,\end{aligned}$$

where $\widetilde{C}_{A,p} := \prod_{v\in A}\widetilde{C}_{v,p}$ for all A in $\mathcal{C}_u$. Given the above and

$$\begin{aligned}\left|\mathcal{Z}^N_u-\mathcal{Z}_u\right| &= \left|\mathcal{Z}^N_{\mathcal{C}_u}\lambda^N_u(w_u)-\mathcal{Z}_{\mathcal{C}_u}\lambda_u(w_u)\right|\\ &\leq \left|\lambda^N_u(w_u)\right|\left|\mathcal{Z}^N_{\mathcal{C}_u}-\mathcal{Z}_{\mathcal{C}_u}\right|+\mathcal{Z}_{\mathcal{C}_u}\left|\lambda^N_u(w_u)-\lambda_u(w_u)\right|\\ &\leq \|w_u\|_\infty\left|\mathcal{Z}^N_{\mathcal{C}_u}-\mathcal{Z}_{\mathcal{C}_u}\right|+\mathcal{Z}_{\mathcal{C}_u}\left|\lambda^N_u(w_u)-\lambda_u(w_u)\right|,\end{aligned}$$

(8.35) follows from the inequality for λ^N_u and Minkowski's inequality. □

8.4.3 Bias Estimates

The key result to obtain the bias estimate is Lemma 8.2 which controls the bias introduced by the product step, armed with Lemma 8.2 the remainder of the proof follows the approach of Olsson and Rydén (2004): we show that the bias bound holds for the leaves of $\mathbb{T}$ and inductively move towards the root of the tree showing that the bias bound is preserved. In particular, we will make us of the following auxiliary result:

Lemma 8.7. Under Assumptions 8.1-8.2 we have that for every $l,k\in\mathbb{N}$ such that $l+k\geq 1$, there exist finite $A_{u,k}, A_{u,l}>0$

$$\begin{aligned}&\mathbb{E}\left[\left|\lambda^N_u(\varphi)-\mathbb{E}\left[\lambda^N_u(\varphi)\right]\right|^l\left|\lambda^N_u(\psi)-\mathbb{E}\left[\lambda^N_u(\psi)\right]\right|^k\right]\\ &\qquad\leq \frac{2^{k+l}A_{u,k}A_{u,l}\|\varphi\|^l_\infty\|\psi\|^k_\infty}{N^{(l+k)/2}}\end{aligned}$$

for all $N>0$, all $\varphi\in\mathcal{B}_b(\boldsymbol{E}_u)$ and every $u\in\mathbb{T}$.

Proof. As argued in the beginning of Proposition 8.1, the λ^N_u satisfy the $\mathbb{L}_p$ inequality in (8.26). Armed with (8.26), the proof consists of applying Jensen's inequality and the Cauchy-Schwarz inequality as in the proof of (Olsson and Rydén, 2004, Lemma 2.2) and Lemma B.5. □

Combining the above Lemma with the bound on the bias of product of measures in Lemma 8.2 we obtain the following bias estimate:

Proposition 8.2 (Bias estimates for $(\lambda_u^N)_{u\in\mathbb{T}}$ and $(\pi_u^N)_{u\in\mathbb{T}}$)**.** If Assumptions 8.1–8.2 hold, the weight functions are bounded below for every $u \in \mathbb{T}$, $w_u \geq \beta > 0$, then there exist finite constants $C_u, \widetilde{C}_u < \infty$ such that

$$\left|\mathbb{E}\left[\lambda_u^N(\varphi)\right] - \lambda_u(\varphi)\right| \leq \frac{\widetilde{C}_u\|\varphi\|_\infty}{N}, \quad \left|\mathbb{E}\left[\pi_u^N(\varphi)\right] - \pi_u(\varphi)\right| \leq \frac{C_u\|\varphi\|_\infty}{N}$$

for all $N > 0$, all $\varphi \in \mathcal{B}_b(\boldsymbol{E}_u)$ and every $u \in \mathbb{T}$.

Proof. First, let us consider the leaves of $\mathbb{T}$: take any $u \in \mathbb{T}^\partial$, then, since the particles are sampled independently from K_u we have that

$$\left|\mathbb{E}\left[\bar{\lambda}_u^N(\varphi)\right] - \bar{\lambda}_u(\varphi)\right| = 0 \leq \frac{\|\varphi\|_\infty}{N} \quad \forall N > 0, \quad \varphi \in \mathcal{B}_b(\bar{\boldsymbol{E}}_u). \tag{8.37}$$

Then, take π_u^N, $u \in \mathbb{T}$. As shown at the start of Lemma 8.4's proof,

$$\pi_u(\varphi) = \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)}, \quad \pi_u^N(\varphi) = \frac{\lambda_u^N(w_u\varphi)}{\lambda_u^N(w_u)} \quad \forall N > 0,$$

and using the triangle inequality

$$\begin{aligned}\left|\mathbb{E}\left[\pi_u^N(\varphi)\right] - \pi_u(\varphi)\right| \leq &\left|\mathbb{E}\left[\frac{\lambda_u^N(w_u\varphi)}{\lambda_u^N(w_u)}\right] - \frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]}\right| \\ &+ \left|\frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} - \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)}\right|.\end{aligned} \tag{8.38}$$

As we did for the proof of the bias estimates in Proposition 5.4, to control the first term, we consider a two-dimensional Taylor expansion of the function $(u, v) \mapsto u/v$ around (u_0, v_0) with a second order remainder of Lagrange form

$$\frac{u}{v} = \frac{u_0}{v_0} + \frac{1}{v_0}(u - u_0) - \frac{u_0}{v_0^2}(v - v_0) + \frac{\theta_u}{\theta_v^3}(v - v_0)^2 - \frac{1}{\theta_v^2}(v - v_0)(u - u_0),$$

where (θ_u, θ_v) is a point on the line segment between (u, v) and (u_0, v_0).

Applying the Taylor expansion above to $\lambda_u^N(w_u\varphi)/\lambda_u^N(w_u)$ around the point

$(\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right], \mathbb{E}\left[\lambda_u^N(w_u)\right])$ we obtain

$$\begin{aligned}\frac{\lambda_u^N(w_u\varphi)}{\lambda_u^N(w_u)} = \frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} + \frac{\lambda_u^N(w_u\varphi) - \mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} \\ - \frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]^2}\left(\lambda_u^N(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right) + R_u^N(\theta_1,\theta_2)\end{aligned} \tag{8.39}$$

where the remainder is a function of (θ_1, θ_2), a point on the line segment between between $(\lambda_u^N(w_u\varphi), \lambda_u^N(w_u))$ and $\left(\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right], \mathbb{E}\left[\lambda_u^N(w_u)\right]\right)$

$$\begin{aligned}R_u^N(\theta_1,\theta_2) := \frac{\theta_1}{\theta_2^3}\left(\lambda_u^N(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right)^2 \\ - \frac{1}{\theta_2^2}\left(\lambda_u^N(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right)\left(\lambda_u^N(w_u\varphi) - \mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]\right).\end{aligned}$$

Taking expectations of (8.39), we find that

$$\begin{aligned}\left|\mathbb{E}\left[\frac{\lambda_u^N(w_u\varphi)}{\lambda_u^N(w_u)}\right] - \frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]}\right| &= \left|\mathbb{E}\left[R_u^N(\theta_1,\theta_2)\right]\right| \\ &\leq \mathbb{E}\left[\left|\frac{\theta_1}{\theta_2^3}\left(\lambda_u^N(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right)^2\right|\right] \\ &+ \mathbb{E}\left[\left|\frac{1}{\theta_2^2}\left(\lambda_u^N(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right)\left(\lambda_u^N(w_u\varphi) - \mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]\right)\right|\right].\end{aligned}$$

Under Assumption 8.2 we have that $\theta_1 \leq \|\varphi\|_\infty \|w_u\|_\infty$ for all $u \in \mathbb{T}$. Since we assumed that w_u is bounded below, i.e. $w_u \geq \beta > 0$ for all $u \in \mathbb{T}$, we also have that

$$\begin{aligned}\lambda_u^N(w_u) &\geq \beta\lambda_u^N(1) \geq \beta > 0 \\ \mathbb{E}\left[\lambda_u^N(w_u)\right] &\geq \beta\mathbb{E}\left[\lambda_u^N(1)\right] \geq \beta > 0\end{aligned}$$

from which $\theta_2 \geq \beta > 0$ follows straightforwardly. Lemma 8.7 then gives

$$\left|\mathbb{E}\left[R_u^N(\theta_1,\theta_2)\right]\right| \leq \frac{\|\varphi\|_\infty \|w_u\|_\infty^3}{\beta^3}\frac{4A_{u,2}}{N} + \frac{1}{\beta^2}\frac{4A_{u,1}^2\|\varphi\|_\infty\|w_u\|_\infty^2}{N}. \tag{8.40}$$

For the second term in (8.38) apply the triangle inequality:

$$\begin{aligned}
&\left|\frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} - \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)}\right| \\
&\quad\leq \frac{\left|\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right] - \lambda_u(w_u\varphi)\right|}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} + \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)\mathbb{E}\left[\lambda_u^N(w_u)\right]}\left|\lambda_u(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right| \\
&\quad\leq \frac{\left|\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right] - \lambda_u(w_u\varphi)\right|}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} + \frac{\|\varphi\|_\infty}{\mathbb{E}\left[\lambda_u^N(w_u)\right]}\left|\lambda_u(w_u) - \mathbb{E}\left[\lambda_u^N(w_u)\right]\right|.
\end{aligned} \tag{8.41}$$

If u is such that $\mathcal{C}_u \subseteq \mathbb{T}^\partial$, i.e. all the children of u are leaves, then

$$\left|\mathbb{E}\left[\lambda_u^N(\varphi)\right] - \lambda_u(\varphi)\right| \leq \frac{\widetilde{C}_u\|\varphi\|_\infty}{N} \quad \forall N > 0, \quad \varphi \in \mathcal{B}_b(\boldsymbol{E}_u) \tag{8.42}$$

follows from (8.37) and repeated applications of Lemma 8.2, because the sequences $(\bar{\lambda}_{u1}^N)_{N=1}^\infty, \ldots, (\bar{\lambda}_{uc_u}^N)_{N=1}^\infty$ are independent by construction, so are $(\bar{\lambda}_{u[k]}^N)_{N=1}^\infty$ and $(\bar{\lambda}_{u(k+1)}^N)_{N=1}^\infty$ for all $k < c_u$, where $u[k] := \{u1, \ldots, uk\}$ denotes the set containing the first k children of u and $\bar{\lambda}_{u[k]}^N := \prod_{v\in u[k]} \bar{\lambda}_v^N$ the corresponding product of $\bar{\lambda}_v$s.

Otherwise, $\lambda_u^N(\varphi) = \prod_{v\in\mathcal{C}_u} \bar{\lambda}_v^N(\varphi)$ and $\lambda_u(\varphi) = \prod_{v\in\mathcal{C}_u} \bar{\lambda}_v(\varphi)$ with $\bar{\lambda}_v(\varphi) = \pi_v(K_v\varphi)$. Repeated applications of conditioning give

$$\begin{aligned}
\mathbb{E}\left[\bar{\lambda}_v^N(\varphi)\right] &= \mathbb{E}\left[\mathbb{E}\left[\bar{\lambda}_v^N(\varphi) \mid \sigma((\boldsymbol{X}_v^{n,N})_{n=1}^N)\right]\right] \\
&= \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left[(K_v\varphi)(\boldsymbol{X}_v^{n,N})\right] \\
&= \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left[\mathbb{E}\left[(K_v\varphi)(\boldsymbol{X}_v^{n,N}) \mid \sigma(\{(\bar{\boldsymbol{X}}_{v1}^{n,N}, \ldots, \bar{\boldsymbol{X}}_{vc_v}^{n,N})\}_{n=1}^N)\right]\right] \\
&= \mathbb{E}\left[\pi_v^N(K_v\varphi)\right]
\end{aligned}$$

and thus

$$\begin{aligned}
\left|\mathbb{E}\left[\bar{\lambda}_v^N(\varphi)\right] - \bar{\lambda}_v(\varphi)\right| &= \left|\mathbb{E}\left[\pi_v^N(K_v\varphi)\right] - \pi_v(K_v\varphi)\right| \\
&\leq \frac{C_v\|\varphi\|_\infty}{N}
\end{aligned}$$

for all $v \in \mathcal{C}_u$ by the induction hypothesis on the children of u. Repeated applications of Lemma 8.2 give (8.42) for all $u \in \mathbb{T}^\partial$. Putting (8.41) and the above together and

recalling that $\mathbb{E}\left[\lambda_u^N(w_u)\right] \geq \beta > 0$, gives

$$\left| \frac{\mathbb{E}\left[\lambda_u^N(w_u\varphi)\right]}{\mathbb{E}\left[\lambda_u^N(w_u)\right]} - \frac{\lambda_u(w_u\varphi)}{\lambda_u(w_u)} \right| \leq 2\beta\|w_u\|_\infty \frac{\widetilde{C}_u\|\varphi\|_\infty}{N}$$

which combined with (8.38) and (8.40) gives

$$\left|\mathbb{E}\left[\pi_u^N(\varphi)\right] - \pi_u(\varphi)\right| \leq \frac{C_u\|\varphi\|_\infty}{N}$$

with $C_u = 2\beta\|w_u\|_\infty\widetilde{C}_u + 4A_{u,2}\|w_u\|_\infty^3/\beta^3 + 4A_{u,1}^2\|w_u\|_\infty^2/\beta^2$ for all $u \in \mathbb{T}$. □

8.4.4 Unbiasedness of Unnormalised Flow

To prove the unbiasedness of $\upsilon^N(\varphi)$ and $\gamma^N(\varphi)$ (Theorem 8.4), we generalise the argument given in Lindsten et al. (2017, Appendix A.1) for binary balanced trees, which itself builds on the approach taken in Andrieu et al. (2010) to prove the analogous results for standard SMC. In particular, we write the expectations in Theorem 8.4 as integrals with respect to all the variables generated when running Algorithm 3 and then use a change of measure to show that this integral corresponds to an integral with respect to γ_u.

First, emulating the approach of Andrieu et al. (2010) for standard SMC, we introduce additional random variables describing the child/parent relationship between particles at nodes in the same sub-tree, which we call ancestors. Appendix A.1 in Lindsten et al. (2017) provides the construction for binary balanced trees. Through the definition of ancestors we can write the law of all the random variables generated during the running of Algorithm 3 from which the unbiasedness of $\upsilon^N(\varphi)$ and $\gamma^N(\varphi)$ follows using routine arguments (e.g. Andrieu et al. (2010), Ala-Luhtala et al. (2016, Theorem 2) and Naesseth et al. (2019, Appendix 4.A)).

Let us denote the set of all the particles simulated during the running of the algorithm up to node $u \in \mathbb{T}$ by $\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}$, where for each $\boldsymbol{n} = (n_v)_{v\in\mathcal{C}_u}$ in $[N]^{c_u}$, $\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N} := (\bar{\boldsymbol{x}}_v^{n_v,N})_{v\in\mathcal{C}_u}$ denotes the corresponding particle and $\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N} = \bar{\boldsymbol{x}}_v^{n_v,N}$ if v is a leaf. To keep track of the relationship between particles induced by the resampling step, we introduce a set of ancestor variables for all the interior nodes, $\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}$. We denote by $\boldsymbol{a}_{\mathcal{C}_v}^{n,N} := (a_{v1}^{n,N}, \ldots, a_{vc_v}^{n,N})$ the indices of the ancestors of $\boldsymbol{x}_v^{n,N}$, i.e. $\boldsymbol{x}_v^{n,N}$ is a copy of $(\bar{\boldsymbol{x}}_{v1}^{a_{v1}^{n,N}}, \ldots, \bar{\boldsymbol{x}}_{vc_v}^{a_{vc_v}^{n,N}})$; by extension these are also the ancestors of $\bar{\boldsymbol{x}}_v^{n,N}$. In particular, whereas in standard SMC each particle has only one ancestor particle, in DaC-SMC each particle has one ancestor particle for each of its child nodes in

the tree topology. As in Andrieu et al. (2010), for each node $v \in \mathbb{T}_u$, we describe the resampling step through a family of probability distributions on $\{1, \ldots, N\}^N$, $\{r(\cdot \mid \boldsymbol{W}), \boldsymbol{W} \in [0,1]^N\}$ where $\boldsymbol{W}$ is a collection of weights, so that we can describe most standard resampling schemes through r.

With this notation we can write the joint distribution of all the variables simulated during the running of the algorithm on the sub-tree $\mathbb{T}_u$:

Proposition 8.3. The joint law of all the random variables generated during the running of the Algorithm 3 on the sub-tree $\mathbb{T}_u$ is

$$\begin{aligned} &\widetilde{K}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^{\partial}}\right) \\ &= \left(\prod_{v\in\mathbb{T}_u^{\partial}}\prod_{n=1}^{N} K_v(\mathrm{d}\bar{\boldsymbol{x}}_v^{n,N})\right)\left(\prod_{v\in\mathbb{T}_u^{\partial}} r(\mathrm{d}\boldsymbol{a}_{\mathcal{C}_v}^{1:N} \mid \boldsymbol{W}_v^{N^{c_v}})\right)\left(\prod_{v\in\mathbb{T}_u^{\partial}}\prod_{n=1}^{N} K_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{a}_{\mathcal{C}_v}^{n,N}}, \mathrm{d}x_v^{n,N})\right). \end{aligned} \tag{8.43}$$

Proof. Starting from line 2 of Algorithm 3: if v is a leaf, i.e. $v \in \mathbb{T}_u^{\partial}$, then a particle $\bar{\boldsymbol{x}}_v^{n,N} = \boldsymbol{x}_v^{n,N}$ is proposed independently from K_v for each $n \leq N$ and each of the leaves in $\mathbb{T}_u$

$$\prod_{v\in\mathbb{T}_u^{\partial}}\prod_{n=1}^{N} K_v(\mathrm{d}\bar{\boldsymbol{x}}_v^{n,N}). \tag{8.44}$$

If v is not a leaf, i.e. $v \in \mathbb{T}_u^{\partial}$, then the resampling step in line 7 defines the ancestors of each particle, $\boldsymbol{a}_{\mathcal{C}_v}^{1:N} := (\boldsymbol{a}_{\mathcal{C}_v}^{1,N}, \ldots, \boldsymbol{a}_{\mathcal{C}_v}^{N,N})$ through the density r

$$\prod_{v\in\mathbb{T}_u^{\partial}} r(\mathrm{d}\boldsymbol{a}_{\mathcal{C}_v}^{1:N} \mid \boldsymbol{W}_v^{N^{c_v}}). \tag{8.45}$$

The ancestors are obtained given the normalised importance weights $\boldsymbol{W}_v^{N^{c_v}} = \left\{\boldsymbol{W}_v^{\boldsymbol{n},N}, \boldsymbol{n} \in \{1,\ldots,N\}^{N^{c_v}}\right\}$, where

$$\boldsymbol{W}_v^{\boldsymbol{n},N} := \frac{w_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N})}{\sum_{\boldsymbol{m}\in\{1,\ldots,N\}^{N^{c_v}}} w_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{m},N})}. \tag{8.46}$$

Finally, in the mutation step (line 8) the new particles $x_v^{n,N}$ are obtained from K_v

$$\prod_{n=1}^{N} K_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{a}_{\mathcal{C}_v}^{n,N}}, \mathrm{d}x_v^{n,N}) \tag{8.47}$$

and then concatenated with their ancestors as in (8.6)

$$\bar{\boldsymbol{x}}_v^{n,N} = (x_v^{n,N}, \bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{a}_{\mathcal{C}_v}^{n,N}}) \qquad \text{where} \quad \bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{a}_{\mathcal{C}_v}^{n,N}} := (\bar{\boldsymbol{x}}_r^{a_r^{n,N}})_{r\in\mathcal{C}_v}.$$

Putting (8.44), (8.45) and (8.47) together gives (8.43). □

Using the construction of ancestral lineages and the joint law in (8.43) we can establish unbiasedness of integrals w.r.t. the unnormalised measures γ_u^N by showing that the ratio $\mathcal{Z}_u^N/\mathcal{Z}_u$ is the Radon-Nikodým derivative of an appropriate extended target measure (extended to include the ancestral lineages) with respect to the law underlying all sampled variables during the running of the algorithm.

Proposition 8.4 (Unbiasedness of $(\upsilon_u^N)_{u\in\mathbb{T}}$ and $(\gamma_u^N)_{u\in\mathbb{T}}$)**.** If Assumptions 8.1–8.2 hold, then

$$\mathbb{E}\left[\upsilon_u^N(\varphi)\right] = \upsilon_u(\varphi), \qquad \mathbb{E}\left[\gamma_u^N(\varphi)\right] = \gamma_u(\varphi),$$

for all $N > 0$, all $\varphi \in \mathcal{B}_b(\boldsymbol{E}_u)$ and every $u \in \mathbb{T}$. In particular, $\mathbb{E}\left[\mathcal{Z}_u^N\right] = \mathcal{Z}_u$ for all $N > 0$ and u in $\mathbb{T}$.

Proof. We start by considering the integrals w.r.t. γ_u^N in (8.13). Since γ_u^N is a function of $(\bar{\boldsymbol{X}}_{u1}^{n_1,N}, \ldots, \bar{\boldsymbol{X}}_{uc_u}^{n_{c_u},N})$ and given that $\widetilde{K}_u$ is the distribution of all the variables simulated while running Algorithm 3 on the sub-tree rooted at u, we have

$$\begin{aligned}\mathbb{E}\left[\gamma_u^N(\varphi)\right] &= \mathbb{E}_{\widetilde{K}_u}\left[\gamma_u^N(\varphi)\right] \qquad (8.48)\\ &= \frac{1}{N^{c_u}} \sum_{n_1=1}^{N} \cdots \sum_{n_{c_u}=1}^{N} \int \widetilde{K}_u \left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial} \right) \mathcal{Z}_u^N w_u(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}) \varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}),\end{aligned}$$

where $\mathcal{Z}_u^N$ denotes the map from a particular realisation $\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}$ to the corresponding normalising constant

$$\mathcal{Z}_u^N = \prod_{v\in\mathbb{T}_u} \frac{1}{N^{c_v}} \sum_{\boldsymbol{n}\in\{1,\ldots,N\}^{Nc_u}} w_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N})$$

where the weights w_u are the Radon-Nikodým derivative of $\gamma_u = \mathcal{Z}_u\pi_u$ with respect to υ_u

$$w_u(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}) = \mathcal{Z}_u \frac{d\pi_u}{d\upsilon_u}(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}).$$

In particular, it is sufficient to compute the integral in the right-hand-side for a given particle A to compute the expectation.

In order to compute the integral in (8.48), we introduce ancestral lineages for the particles, describing all the ancestors of particle $\boldsymbol{n}$ at the root u of tree $\mathbb{T}_u$. The ancestral lineage is described through a set of variables $\{b_v^{\boldsymbol{n}} : v \in \mathbb{T}_u\}$ giving the indices of the ancestors of particle $\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}$ at each node in $\mathbb{T}_u$:

$$b_u^{\boldsymbol{n}} = \boldsymbol{n} = (n_1, \ldots, n_{c_u}) \qquad \text{and} \qquad b_v^{\boldsymbol{n}} = \boldsymbol{a}_{\mathcal{C}_v}^{b_{p_v}^{\boldsymbol{n}}(v)} \tag{8.49}$$

where p_u denotes the parent of u, $p_u := \{v \in \mathbb{T} : u \in \mathcal{C}_v\}$ and $b_{p_v}^{\boldsymbol{n}}(v)$ is the component of $b_{p_v}^{\boldsymbol{n}}$ corresponding to child v. Each particle $\boldsymbol{n}$ at node u has an ancestral lineage, described by the construction above, associated with it.

Then, we introduce a categorical random variable A taking values on $1, \ldots, N^{c_u}$ such that $A = \boldsymbol{n} = (n_1, \ldots, n_{c_u})$ with probability $w_u(\bar{x}_{\mathcal{C}_u}^{\boldsymbol{n},N})$, i.e. A describes the selection of one of the weighted particles at node u out of the N^{c_u} possibilities. The joint distribution of $\bar{\boldsymbol{x}}_{\mathbb{T}_u}^{N^{c_u}}, \boldsymbol{a}_{\mathbb{T}_u^\partial}^{1:N}$ and A is

$$\begin{aligned} &\bar{K}_u \left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, \mathrm{d}A \right) \\ &= \widetilde{K}_u \left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial} \right) \boldsymbol{W}_u^{A,N} \mathrm{d}A, \end{aligned} \tag{8.50}$$

where with a slight abuse of notation we allow $\mathrm{d}A$ to denote the counting measure on $\{1, \ldots, N\}^{c_u}$. It is straightforward to check that for a given particle A the right-most integral in (8.48) satisfies

$$\mathbb{E}_{\widetilde{K}_u}\left[\mathcal{Z}_u^N w_u(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right] = \mathbb{E}_{\bar{K}_u}\left[\mathcal{Z}_u^N \varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right]. \tag{8.51}$$

Our aim is now to show that for all $\varphi \in \mathcal{B}_b(\boldsymbol{E}_u)$

$$\mathbb{E}_{\bar{K}_u}\left[\mathcal{Z}_u^N \varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right] = \gamma_u(\varphi)$$

or equivalently,

$$\mathbb{E}_{\bar{K}_u}\left[\frac{\mathcal{Z}_u^N}{\mathcal{Z}_u}\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right] = \pi_u(\varphi)$$

To obtain this result, we show that given a particle A with corresponding ancestral lineage described by $\{b_v^A : v \in \mathbb{T}_u\}$ in (8.49), the ratio $\mathcal{Z}_u^N/\mathcal{Z}_u$ corresponds to the Radon-Nikodým derivative of an appropriate distribution $\tilde{\pi}_u$ w.r.t. $\bar{K}_u$ and that $\tilde{\pi}_u$

admits π_u as marginal:

$$\begin{aligned}\frac{\mathcal{Z}_u^N}{\mathcal{Z}_u} &= \left(\prod_{v\in\mathbb{T}_u}\frac{1}{N^{c_v}}\sum_{\boldsymbol{n}\in\{1,\dots,N\}^{Nc_u}} w_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N})\right)\left(\prod_{v\in\mathbb{T}_u} w_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{b_v^A})\right)^{-1}\frac{d\pi_u}{d\upsilon_u}(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}) \\ &= \frac{1}{N^{|\mathbb{T}_u|}}\left(\prod_{v\in\mathbb{T}_u}\boldsymbol{W}_v^{b_v^A}\right)^{-1}\frac{d\pi_u}{d\upsilon_u}(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}) \\ &= \frac{1}{N^{|\mathbb{T}_u|}}\left(\boldsymbol{W}_u^{A,N}\prod_{v\in\mathbb{T}_u^{\mathscr{U}}} r(b_v^A \mid \boldsymbol{W}_v^{Nc_v})\right)^{-1}\frac{d\pi_u}{d\upsilon_u}(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\end{aligned}$$

where the second equality follows from (8.46).

We can then define an extended target measure $\tilde{\pi}_u$ over all the sampled variables which takes into account the ancestral lineage corresponding to A

$$\begin{aligned}\tilde{\pi}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^{\mathscr{J}}}, \mathrm{d}A\right) &= \frac{\pi_u(\mathrm{d}\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\mathrm{d}A}{N^{|\mathbb{T}_u|}} \\ &\times \frac{\widetilde{K}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^{\mathscr{J}}}\right)}{\left(\prod_{v\in\mathbb{T}_u^{\partial}} K_v(\mathrm{d}\bar{\boldsymbol{x}}_v^{b_v^A})\right)\left(\prod_{v\in\mathbb{T}_u^{\mathscr{J}}} r(\mathrm{d}b_v^A \mid \boldsymbol{W}_v^{Nc_v})\right)\left(\prod_{v\in\mathbb{T}_u^{\mathscr{J}}} K_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{b_{\mathcal{C}_v}^A}, \mathrm{d}x_v^{b_v^A})\right)},\end{aligned} \tag{8.52}$$

which can be straightforwardly shown to be a normalised probability measure. Recalling (8.50), it is easy to see that

$$\begin{aligned}&\frac{d\tilde{\pi}_u}{d\bar{K}_u}\left(\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^{\mathscr{J}}}, A\right) \\ &\qquad= \frac{\pi_u(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})}{N^{|\mathbb{T}_u|}}\times\frac{\left(\boldsymbol{W}_u^{A,N}\prod_{v\in\mathbb{T}_u^{\mathscr{J}}} r(b_v^A \mid \boldsymbol{W}_v^{Nc_v})\right)^{-1}}{\left(\prod_{v\in\mathbb{T}_u^{\partial}} K_v(\bar{\boldsymbol{x}}_v^{b_v^A})\right)\left(\prod_{v\in\mathbb{T}_u^{\mathscr{J}}} K_v(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{b_{\mathcal{C}_v}^A}, x_v^{b_v^A})\right)} \\ &\qquad= \frac{1}{N^{|\mathbb{T}_u|}}\left(\boldsymbol{W}_u^{A,N}\prod_{v\in\mathbb{T}_u^{\mathscr{J}}} r(b_v^A \mid \boldsymbol{W}_v^{Nc_v})\right)^{-1}\frac{d\pi_u}{d\upsilon_u}(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}) = \frac{\mathcal{Z}_u^N}{\mathcal{Z}_u}\end{aligned}$$

from which follows

$$\begin{aligned}
&\mathbb{E}_{\bar{K}_u}\left[\frac{\mathcal{Z}_u^N}{\mathcal{Z}_u}\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right] \qquad (8.53)\\
&= \int \bar{K}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, \mathrm{d}A\right)\frac{\mathcal{Z}_u^N}{\mathcal{Z}_u}\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\\
&= \int \bar{K}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, \mathrm{d}A\right)\\
&\qquad\qquad \times \frac{d\tilde{\pi}_u}{d\bar{K}_u}\left(\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, A\right)\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\\
&= \int \tilde{\pi}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, \mathrm{d}A\right)\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}).
\end{aligned}$$

Marginalising over the variables not involved in the ancestral lineage of particle $\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}$

$$\int \tilde{\pi}_u\left(\mathrm{d}\left(\bar{\boldsymbol{x}}_{\mathcal{C}_v}^{\boldsymbol{n},N}\right)_{v\in\mathbb{T}_u}, \mathrm{d}\left(\boldsymbol{a}_{\mathcal{C}_v}^{n,N}\right)_{v\in\mathbb{T}^\partial}, \mathrm{d}A\right)\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N}) = \int \pi_u(\mathrm{d}\bar{\boldsymbol{x}}_{\mathcal{C}_u})\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}) = \pi_u(\varphi).$$

Putting (8.48), (8.51) and (8.53) together shows that

$$\mathbb{E}\left[\gamma_u^N(\varphi)\right] = \mathbb{E}_{\bar{K}_u}\left[\mathcal{Z}_u^N\varphi(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{A,N})\right] = \pi_u(\varphi)\mathcal{Z}_u = \gamma_u(\varphi).$$

Taking $\varphi \equiv 1$ gives $\mathbb{E}\left[\mathcal{Z}_u^N\right] = \mathcal{Z}_u$, while the same argument above with $w_u(\bar{\boldsymbol{x}}_{\mathcal{C}_u}^{\boldsymbol{n},N}) \equiv 1$ in (8.48) gives $\mathbb{E}\left[\upsilon_u^N(\varphi)\right] = \upsilon_u(\varphi)$.

□

Connections with Other Work

The results in this section show that, as its standard counterpart, DaC-SMC can be used to build approximations of proposals for MCMC algorithms (Andrieu et al., 2010). In particular, Theorem 8.4 guarantees that any Metropolis-Hastings algorithm using $\{\pi_u\}_{u\in\mathbb{T}^\vee}$ as proposals is a valid algorithm since the acceptance probability is obtained by taking the ratio of unbiased estimators of $\mathcal{Z}_u$ in a pseudo-marginal fashion (Andrieu and Roberts, 2009). These particle Metropolis-Hastings algorithms can be seen as standard Metropolis-Hastings algorithms targeting the extended distribution (8.52) with proposal (8.50) (Andrieu et al., 2010, Theorem 2, Theorem 4).

The construction of ancestral lineages in (8.49) is key to the definition of conditional SMC, a particular class of SMC algorithms in which the resampling step is performed in such a way that a prespecified path from the leaves to the root of

the tree $\{\boldsymbol{X}_{\mathcal{C}_u}^{\boldsymbol{n},N}\}_{u\in\mathbb{T}^\partial}$ with given ancestral lineage $\{b_u^{\boldsymbol{n}}\}_{u\in\mathbb{T}}$ is ensured to survive all the resampling steps (Andrieu et al., 2010). The conditional SMC update was originally proposed by Andrieu et al. (2010) to perform the updating step within particle Gibbs sampling, where at each iteration one particle and its corresponding ancestral lineage are kept fixed and the remaining particles are updated conditionally. More recently, the conditional SMC update has been employed to approximate smoothing distributions for state space models (Jacob et al., 2020; Karppinen and Vihola, 2021) since the conditional SMC update defined a Markov kernel which leaves the smoothing distribution invariant (Andrieu et al., 2010).

In the standard SMC setting the construction of ancestral lineages has been exploited to build consistent variance estimators which only require one run of the SIR algorithm (Lee and Whiteley, 2018); similar arguments could be explored to build variance estimators for DaC-SMC using the construction of ancestral lineages described above.

8.5 Summary

In this chapter we studied the theoretical properties of the DaC-SMC algorithm introduced in Lindsten et al. (2017) combining the well-known results on standard SMC summarised in Chapter 2 with results on product of empirical measures.

In particular, we showed that the additional interactions due to the merging of different branches of the tree over which DaC-SMC is defined do not influence the rates of convergence of the algorithm, allowing us to extract $\mathbb{L}_p$ inequalities and a strong law of large numbers. These results show that DaC-SMC is a valid algorithm and has some of the standard convergence properties one might expect from a Monte Carlo algorithm, e.g. the mean squared error decays at rate $N^{-1/2}$ as established in Proposition 8.1. In Proposition 8.4 we established that the estimates of the normalising constants given by DaC-SMC are unbiased, showing that DaC-SMC can be used within particle MCMC algorithms as its standard counterpart.

Closing Remarks

This thesis is concerned with the study of families of interacting particles which present non-standard interactions, in particular their numerical implementation and theoretical properties. In Part I the non-standard interactions arise from the particular class of problems we are interested in, Fredholm integral equations of the first kind. In Part II the non-standard interactions arise from algorithmic design.

In Chapters 4-5 we introduced a continuous version of the Expectation Maximisation Algorithm (EMS) for solving Fredholm integral equations and proposed a sequential Monte Carlo (SMC) algorithm which provides an adaptive stochastic discretisation of EMS. This algorithm outputs smooth approximations of the solution of the integral equation and achieves good performances in applications from statistics and image reconstruction. Because of the additional approximation of the weights, the sequential Monte Carlo implementation of EMS (SMC-EMS) does not fall within the standard class of SMC algorithms, therefore we also provide a novel theoretical analysis and prove error estimates (which decay at the standard Monte Carlo rate $N^{-1/2}$), a strong law of large numbers and convergence of the estimator given by SMC-EMS.

In Chapters 6 we studied approximations of the solution of integral equations obtained by minimising a penalised Kullback–Leibler divergence. After introducing a gradient flow construction for minimising functionals over the set of probability measures, we established conditions under which the gradient flow equation admits a solution. In Chapter-7 we used the connection between the gradient flow equation and mean-field stochastic differential equations to introduce an interacting particle system arising to approximate the solution of the minimisation problem and showed the performances of this method on a number of examples.

In Chapter 8 we focused on the Divide and Conquer SMC algorithm and showed that some of the standard theoretical properties of SMC algorithms are preserved despite the additional interactions present in DaC-SMC. In particular, the rate at which the error decays is $N^{-1/2}$ in the number of particles and the estimates of the

normalising constants are unbiased.

We devote the final part of this thesis to suggesting directions for future theoretical investigation and areas in which the proposed algorithms could be employed.

Future Directions

The theoretical results in Chapter 5 show that the SMC-EMS algorithm is well-founded, however they do not provide an easy way to quantify the increase in variance caused by the additional approximation step needed to compute the weights. Obtaining estimates of non-asymptotic variances of SMC is usually quite challenging and the estimators available in the literature (Cérou et al., 2011; Du and Guyader, 2019; Lee and Whiteley, 2018) do not apply to SMC-EMS because of the additional approximation in the weight computation. It would therefore be interesting to establish a central limit theorem for SMC-EMS and compare the asymptotic variance with that of the standard SMC algorithm that we would implement if the weights were known. These results would shed light on the increase in variance caused by the use of the approximated weights instead of the unknown exact weights.

Similar results exist for random weight particle filters, in which the unknown weights are replaced by an unbiased estimator (Fearnhead et al., 2008). However, the case in which the approximate weights are biased but consistent has not been studied yet. There is a number of algorithms using biased estimators of the true unknown weights (Everitt et al., 2017; Klaas et al., 2012; Salomone et al., 2018), establishing a central limit theorem for the general class of random weight particle filters with biased weights (to which SMC-EMS belongs) would therefore be of broader interest.

Since typical applications of Fredholm integral equations are low-dimensional, mostly one to three dimensional with some notable exceptions (Signoroni et al., 2019), an improvement over the proposed SMC-EMS algorithms could be given by employing sequential quasi-Monte Carlo (SQMC) ideas (Gerber and Chopin, 2015). Indeed, SQMC tends to significantly outperform SMC in low-dimensional settings (Chopin and Papaspiliopoulos, 2020).

The mean-field particle system introduced in Chapter 7 aims at approximating the law of the mean-field SDE linked to the gradient flow equation of the regularised Kullback–Leibler divergence studied in Chapter 6. Having obtained the gradient flow equation for the regularised Kullback–Leibler divergence considered, it would be of great interest to check if the corresponding SDE admits a unique solution and to study the convergence properties of the corresponding particle system. Unfortunately, these results cannot be straightforwardly obtained from the available

theory on mean-field SDEs since the drift coefficient of the particle system we considered is not Lipschitz. Recently, we have been looking at approximate solutions minimising a closely related regularised Kullback–Leibler divergence, in which a relative entropy term is used as regulariser instead of Shannon's differential entropy. This minimisation problem can be considered as the probabilistic counterpart of Tikhonov regularisation and preliminary results suggest that the corresponding particle system satisfies most of the theoretical results presented in Chapter 2 (Crucinio et al., 2021a).

In this thesis we focused on providing reconstructions of the solution f of Fredholm integral equations

$$h(y) = \int f(x)g(y \mid x)\mathrm{d}x,$$

however there are scenarios in which we are interested in recovering the function g giving rise to h, the distorted version of f. In particular, in the image processing literature we are often interested in estimating the point spread function (PSF) associated with a specific microscope (Buddha and Boruah, 2020; Li et al., 2018), telescope (Long et al., 2019) or spectrometer (Semenov et al., 2011) from the distorted version h of a reference image f. We believe that the algorithms studied in this thesis could be well-suited to tackle these problems since in the applications listed above $g(y \mid x) \equiv g(y - x)$ and one can easily swap the role of f, g.

Finally, in Chapter 8 we showed that Divide and Conquer SMC possesses some of the theoretical properties of SMC, however there are other results that one might wish to obtain for DaC-SMC, in particular, central limit theorems like those in Chopin (2004); Del Moral (2004, Chapter 9) and summarised in Chapter 2. These results can be obtained by extending the characterisation of products of two families of independent random measures in Chapter 8 to show that the approximation error of the product of k independent families of particles, each one having approximation error of order $\mathcal{O}(N^{-1/2})$, is of order $\mathcal{O}(N^{-k/2})$ (Kuntz et al., 2021a).

List of Symbols

$B(\mathbb{H})$	Borel σ-algebra on $\mathbb{H}$
$\mathcal{B}_b(\mathbb{H})$	set of measurable and bounded real valued functions on $\mathbb{H}$
$C_b(\mathbb{H})$	set of continuous and bounded real valued functions on $\mathbb{H}$
δ_x	Dirac measure/point mass centred at x
$\mathbb{I}(A)$	indicator function of set A
ent	entropy function
ESS	effective sample size
$\mathbb{E}$	expectation under $\mathbb{P}$
ISE	integrated squared error
KL	Kullback–Leibler divergence
$\mathcal{M}(\mathbb{H})$	set of finite measures on $\mathbb{H}$
MISE	mean integrated squared error
MSE	mean squared error
$\mathcal{N}$	Normal distribution
$\mathcal{P}(\mathbb{H})$	set of probability measures on $\mathbb{H}$
$\mathcal{P}_2^{ac}(\mathbb{H})$	set of probability measures on $\mathbb{H}$ with finite second moment and absolutely continuous w.r.t. the Lebesgue measure
$\mathbb{R}$	set of real numbers
$\|\cdot\|_1, \|\cdot\|_2$	l_1, l_2 norm
$\beta(\cdot)$	Bounded–Lipschitz norm for measures
$\|\cdot\|_\infty$	supremum norm
$\|\cdot\|_{BL}$	Bounded–Lipschitz norm for functions
$:=, =:$	'is defined as'
$\sim, \propto$	'is distributed according to', 'is proportional to'

List of Abbreviations

DaC-SMC	Divide and Conquer sequential Monte Carlo
DKDE	Deconvolution kernel density estimator
EM	Expectation Maximisation
EMS	Expectation Maximisation Smoothing
i.i.d.	independent and identically distributed
IB	Iterative Bayes
KDE	Kernel density estimation
MCMC	Markov Chain Monte Carlo
MPLE	Maximum penalised likelihood estimation
OSL-EM	One-step-late expectation mamimisation
PDE	partial differential equation
PET	Positron Emission Tomography
RL	Richardson-Lucy
SDE	stochastic differential equation
SIR	sequential importance resampling
SIS	sequential importance sampling
SLLN	Strong law of large numbers
SMC	sequential Monte Carlo

SMC-EMS	sequential Monte Carlo expectation maximisation smoothing
WGF	Wasserstein gradient flow

Bibliography

J. Ala-Luhtala, N. Whiteley, K. Heine, and R. Piché. An introduction to twisted particle filters and parameter estimation in non-linear state-space models. *IEEE Transactions on Signal Processing*, 64(18):4875–4890, 2016.

U. Amato and W. Hughes. Maximum entropy regularization of Fredholm integral equations of the first kind. *Inverse Problems*, 7(6):793, 1991.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

F. Antonelli and A. Kohatsu-Higa. Rate of convergence of a particle method to the solution of the McKean–Vlasov equation. *The Annals of Applied Probability*, 12 (2):423–476, 2002.

R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. Elsevier, 2018.

J. Baladron, D. Fasoli, O. Faugeras, and J. Touboul. Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons. *The Journal of Mathematical Neuroscience*, 2(1):10, 2012.

J. Bao, C. Reisinger, P. Ren, and W. Stockinger. First order convergence of Milstein schemes for McKean-Vlasov equations and interacting particle systems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2020.

N. G. Becker, L. F. Watson, and J. B. Carlin. A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine*, 10:1527–1542, 1991.

S. Benachour, B. Roynette, D. Talay, and P. Vallois. Nonlinear self-stabilizing processes–I Existence, invariant probability, propagation of chaos. *Stochastic Processes and their Applications*, 75(2):173–201, 1998.

M. Benning and M. Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1–111, 2018.

P. Berti, L. Pratelli, and P. Rigo. Almost sure weak convergence of random probability measures. *Stochastics and Stochastics Reports*, 78(2):91–97, 2006.

A. Beskos, A. Jasra, N. Kantas, and A. Thiery. On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–1146, 2016.

G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Communications on Pure and Applied Mathematics*, 44(2):141–183, 1991.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

P. Billingsley. *Probability and Measure*. John Wiley & Sons., 1995.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103, 2016.

D. Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, pages 105–110, 1947.

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

V. I. Bogachev, M. Röckner, and S. V. Shaposhnikov. Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures. *Journal of Functional Analysis*, 276(12):3681–3713, 2019.

A. Borovykh, N. Kantas, P. Parpas, and G. A. Pavliotis. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *Physica D: Nonlinear Phenomena*, 418:132844, 2021.

M. Bossy and D. Talay. A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Mathematics of Computation*, 66(217):157–192, 1997.

A. Boustati, O. D. Akyildiz, T. Damoulas, and A. Johansen. Generalised Bayesian Filtering via Sequential Monte Carlo. *Adv Neur In*, 33, 2020.

M. Briers, A. Doucet, and S. Maskell. Smoothing algorithms for state–space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–69, 2010.

S. S. G. Buddha and B. R. Boruah. Estimation of point spread function of an optical microscope using stochastic minimization of least square errors. *Journal of Optics*, 22(5):055603, 2020.

M. Burger, E. Resmerita, and M. Benning. An entropic Landweber method for linear ill-posed problems. *Inverse Problems*, 36(1):015009, 2019.

O. Butkovsky. On ergodic properties of nonlinear Markov chains and stochastic McKean–Vlasov equations. *Theory of Probability & Its Applications*, 58(4):661–674, 2014.

C. Byrne and P. P. Eggermont. EM algorithms. In *Handbook of Mathematical Methods in Imaging*, pages 305–388. Springer, 2015.

T. Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966.

R. Carmona. *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.

R. Carmona and F. Delarue. Learning by Examples: What Is a Mean Field Game? In *Probabilistic Theory of Mean Field Games with Applications I*, pages 3–65. Springer, 2018.

R. Carmona, X. Zhu, et al. A probabilistic approach to mean field games with major and minor players. *The Annals of Applied Probability*, 26(3):1535–1580, 2016.

R. A. Carmona, J. P. Fouque, and L. H. Sun. Mean field games and systemic risk. *Communications in Mathematical Sciences*, 13(4):911–933, 2015.

J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Archive for Rational Mechanics and Analysis*, 179(2):217–263, 2006.

R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186, 1988.

H. Cartan. *Differential calculus*, volume 1. Kershaw Publishing, 1971.

F. Cérou, P. Del Moral, and A. Guyader. A nonasymptotic theorem for unnormalized Feynman-Kac particle models. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 47(3):629–649, 2011.

M. Chae, R. Martin, and S. G. Walker. On an algorithm for solving Fredholm integrals of the first kind. *Statistics and Computing*, 29:645–654, 2018a.

M. Chae, R. Martin, and S. G. Walker. Convergence of an iterative algorithm to the nonparametric MLE of a mixing distribution. *Statistics & Probability Letters*, 140:142–146, 2018b.

P. H. Chau, W. Y. Li, and P. S. Yip. Construction of the Infection Curve of Local Cases of COVID-19 in Hong Kong using Back-Projection. *International Journal of Environmental Research and Public Health*, 17(18):6909, 2020.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3): 539–552, 2002.

N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

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

A. Cichocki and S. Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

C. Clason, B. Kaltenbacher, and E. Resmerita. Regularization of ill-posed problems with non-negative solutions. In *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 113–135. Springer, 2020.

D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93. Springer, 2012.

K. Craig. Nonconvex gradient flow in the Wasserstein metric and applications to constrained nonlocal interactions. *Proceedings of the London Mathematical Society*, 114(1):60–102, 2017.

D. Crisan and A. Doucet. Convergence of sequential Monte Carlo methods. *Department of Engineering, University of Cambridge, Technical Report CUED/F-INFENG/TR381*, 2000.

D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal processing*, 50(3):736–746, 2002.

F. R. Crucinio, V. De Bortoli, A. Doucet, and A. M. Johansen. Solving Fredholm integral equations of the first kind via Wasserstein gradient flows. *In preparation.*, 2021a.

F. R. Crucinio, A. Doucet, and A. M. Johansen. A particle method for solving Fredholm equations of the first kind. *Journal of the American Statistical Association*, In press, 2021b.

K. Daudel, R. Douc, and F. Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *arXiv preprint arXiv:2005.10618*, 2020.

V. De Bortoli, A. Durmus, X. Fontaine, and U. Simsekli. Quantitative propagation of chaos for SGD in wide neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

P. Del Moral. Nonlinear filtering: Interacting particle resolution. *Markov Processes and Related Fields*, 2(4):555–580, 1996.

P. Del Moral. *Feynman-Kac formulae*. Springer, 2004.

P. Del Moral. *Mean field simulation for Monte Carlo integration*. Chapman and Hall/CRC, 2013.

P. Del Moral and L. Miclo. Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. In J. Azéma, M. Ledoux, M. Émery, and M. Yor, editors, *Séminaire de probabilités XXXIV*, pages 1–145. Springer, 2000.

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

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo methods for Bayesian Computation. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and W. M., editors, *Bayesian Statistics 8*. Oxford University Press, 2006b.

P. Del Moral, A. Doucet, and G. W. Peters. Sharp propagation of chaos estimates for Feynman–Kac particle models. *Theory of Probability & Its Applications*, 51 (3):459–485, 2007.

P. Del Moral, A. Doucet, and S. Singh. Forward smoothing using sequential Monte Carlo. *arXiv preprint arXiv:1012.5390*, 2009.

P. Del Moral, A. Doucet, and A. Jasra. On adaptive resampling procedures for sequential Monte Carlo methods. *Bernoulli*, 18(1):252–278, 2012.

A. Delaigle. An alternative view of the deconvolution problem. *Statistica Sinica*, pages 1025–1045, 2008.

A. Delaigle and I. Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, 45(2):249–267, 2004.

A. Delaigle, P. Hall, A. Meister, et al. On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2):665–685, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39:2–38, 1977.

L. Devroye and T. Wagner. The L1 convergence of kernel density estimates. *The Annals of Statistics*, 7(5):1136–1139, 1979.

N. Dey, L. Blanc-Féraud, C. Zimmer, P. Roux, Z. Kam, J.-C. Olivo-Marin, and J. Zerubia. 3D Microscopy Deconvolution using Richardson-Lucy Algorithm with Total Variation Regularization. Research Report RR-5272, INRIA, 2004.

D. Ding and A. Gandy. Tree-based Particle Smoothing Algorithms in a Hidden Markov Model. *arXiv preprint arXiv:1808.08400*, 2018.

D. L. Donoho. Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis*, 23(5):1309–1331, 1992.

G. Dos Reis, W. Salkeld, and J. Tugaut. Freidlin–Wentzell LDP in path space for McKean–Vlasov equations and the functional iterated logarithm law. *The Annals of Applied Probability*, 29(3):1487–1540, 2019.

G. Dos Reis, S. Engelhardt, and G. Smith. Simulation of McKean Vlasov SDEs with super linear growth. *IMA Journal of Numerical Analysis*, draa099, 2021.

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

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

A. Doucet, A. M. Johansen, and V. B. Tadić. On solving integral equations using Markov chain Monte Carlo methods. *Applied Mathematics and Computation*, 216 (10):2869–2880, 2010.

Q. Du and A. Guyader. Variance estimation in adaptive sequential Monte Carlo. *arXiv preprint arXiv:1909.13602*, 2019.

P. du Bois-Reymond. Bemerkimgen über $\triangle z = 0$. *Journal für die reine und angewandte Mathematik*, 1888(103):204–229, 1888.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3):851–886, 2016.

P. Eggermont and V. LaRiccia. Maximum smoothed likelihood density estimation for inverse problems. *The Annals of Statistics*, 23(1):199–220, 1995.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

R. G. Everitt, A. M. Johansen, E. Rowing, and M. Evdemon-Hogan. Bayesian model comparison with un-normalised likelihoods. *Statistics and Computing*, 27 (2):403–422, 2017.

C. P. S. Fan, J. Stafford, and P. E. Brown. Local-EM and the EMS Algorithm. *Journal of Computational and Graphical Statistics*, 20(3):750–766, 2011.

P. Fearnhead, O. Papaspiliopoulos, and G. O. Roberts. Particle filters for partially observed diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):755–777, 2008.

E. A. Feinberg, P. O. Kasyanov, and Y. Liang. Fatou's lemma in its classical form and Lebesgue's convergence theorems for varying measures with applications to Markov decision processes. *Theory of Probability & Its Applications*, 65(2):270–291, 2020.

A. Finke. *On Extended State-Space Constructions for Monte Carlo Methods*. PhD thesis, University of Warwick, 2015.

I. Fredholm. Sur une classe d'équations fonctionnelles. *Acta Mathematica*, 27:365–390, 1903.

A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

M. G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

M. Gerber and N. Chopin. Sequential quasi-Monte Carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 3(77):509–579, 2015.

M. Gerber, N. Chopin, and N. Whiteley. Negative association, ordering and convergence of resampling methods. *The Annals of Statistics*, 47(4):2236–2260, 2019.

E. Goldstein, J. Dushoff, J. Ma, J. B. Plotkin, D. J. Earn, and M. Lipsitch. Reconstructing influenza incidence by deconvolution of daily mortality time series. *Proceedings of the National Academy of Sciences*, 106(51):21825–21829, 2009.

I. Good. Non-parametric roughness penalty for probability densities. *Nature physical science*, 229(1):29, 1971.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-Radar and Signal Processing*, volume 140, pages 107–113. IEE, 1993.

K. M. Gostic, L. McGough, E. B. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. A. Hay, P. M. De Salazar, et al. Practical considerations for measuring the effective reproductive number, Rt. *PLoS Computational Biology*, 16(12), 2020.

G. Green. *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism.* Printed for the author, by T. Wheelhouse, 1828.

P. J. Green. On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(3):443–452, 1990.

C. W. Groetsch. Integral equations of the first kind, inverse problems and regularization: a crash course. *Journal of Physics: Conference series*, 73(1):012001, 2007.

P. Hall, J. L. Horowitz, et al. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

J. H. Halton. Sequential Monte Carlo techniques for the solution of linear systems. *Journal of Scientific Computing*, 9(2):213–257, 1994.

T. Hohage and F. Werner. Inverse problems with Poisson data: statistical regularization theory, applications and algorithms. *Inverse Problems*, 32(093001), 2016.

M. Hutzenthaler, A. Jentzen, P. E. Kloeden, et al. Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients. *Annals of Applied Probability*, 22(4):1611–1641, 2012.

M. S. Islam and A. Smith. Approximating solutions of Fredholm integral equations via a general spline maximum entropy method. *International Journal of Applied and Computational Mathematics*, 6(3), 2020.

K. Itô. *On stochastic differential equations.* Number 4 in Memoirs of the American Mathematical Society. American Mathematical Society, 1951.

P. E. Jacob, F. Lindsten, and T. B. Schön. Smoothing with couplings of conditional particle filters. *Journal of the American Statistical Association*, 115(530):721–729, 2020.

E. T. Jaynes. Information theory and statistical mechanics. I. *Physical Review*, 106 (4):620, 1957a.

E. T. Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 108 (2):171, 1957b.

S. W. Jewell. *Divide and conquer sequential Monte Carlo for phylogenetics.* PhD thesis, University of British Columbia, 2015.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

C. Jin and J. Ding. Solving Fredholm integral equations via a piecewise linear maximum entropy method. *Journal of Computational and Applied Mathematics*, 304:130–137, 2016.

F. Joachimsthal. Über ein attractionsproblem. *Journal für die reine und angewandte Mathematik*, 58:135–137, 1861.

A. M. Johansen and A. Doucet. Auxiliary variable sequential Monte Carlo methods. *University of Bristol, Statistics Groups Technical Report 07:09*, 2007.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

B. Jourdain, S. Méléard, and W. A. Woyczynski. Nonlinear SDEs driven by lévy processes and related PDEs. *Alea*, 4:1–29, 2008.

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

S. Karppinen and M. Vihola. Conditional particle filters with diffuse initial distributions. *Statistics and Computing*, 31(3):1–14, 2021.

M. Klaas, N. De Freitas, and A. Doucet. Toward practical N2 Monte Carlo: The marginal particle filter. *arXiv preprint arXiv:1207.1396*, 2012.

J. Knoblauch, J. Jewson, and T. Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

A. Kondor. Method of convergent weights—An iterative procedure for solving Fredholm's integral equations of the first kind. *Nuclear Instruments and Methods in Physics Research*, 216(1-2):177–181, 1983.

A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, March 1994.

S. Kopeć. On application of maxent to solving Fredholm integral equations. In *Maximum Entropy and Bayesian Methods*, pages 63–66. Springer, 1993.

R. Kress. Numerical methods in inverse obstacle scattering. *Anziam Journal*, 42: 44–67, 2000.

R. Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences*. Springer, 2014.

D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*, volume 706. John Wiley & Sons, 2013.

C. Kumar and Neelima. On explicit Milstein-type scheme for McKean–Vlasov stochastic differential equations with super-linear drift coefficient. *Electronic Journal of Probability*, 26:1–32, 2021.

J. Kuntz, F. R. Crucinio, and A. M. Johansen. The divide-and-conquer sequential Monte Carlo algorithm: theoretical properties and limit theorems. *In preparation.*, 2021a.

J. Kuntz, F. R. Crucinio, and A. M. Johansen. Product-form estimators: exploiting independence to scale up Monte Carlo. *arXiv preprint arXiv:2102.11575*, 2021b.

N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

L. Landweber. An iteration formula for Fredholm integral equations of the first kind. *American Journal of Mathematics*, 73(3):615–624, 1951.

K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Transactions on Medical Imaging*, 9(4):439–446, 1990.

J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

G. A. Latham. Existence of EMS solutions and a priori estimates. *SIAM Journal on Matrix Analysis and Applications*, 16(3):943–953, 1995.

G. A. Latham and R. S. Anderssen. A hyperplane approach to the EMS algorithm. *Applied Mathematics Letters*, 5(5):71–74, 1992.

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

D. Lee and Y. Vardi. Experiments with maximum likelihood method for image motion deblurring. *Journal of Applied Statistics*, 21(1-2):355–383, 1994.

B. Li, A. Tambe, S. Aviran, and L. Pachter. PROBer provides a general toolkit for analyzing sequencing-based toeprinting assays. *Cell Systems*, 4(5):568–574, 2017.

J. Li, F. Xue, F. Qu, Y.-P. Ho, and T. Blu. On-the-fly estimation of a microscopy point spread function. *Optics express*, 26(20):26120–26133, 2018.

Z. Li, T. Wang, M. Lopuhaä-Zwakenberg, N. Li, and B. Škoric. Estimating numerical distributions under local differential privacy. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 621–635, 2020.

F. Lindsten, A. M. Johansen, C. A. Naesseth, B. Kirkpatrick, T. B. Schön, J. Aston, and A. Bouchard-Côté. Divide-and-Conquer with sequential Monte Carlo. *Journal of Computational and Graphical Statistics*, 26(2):445–458, 2017.

J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2008.

J. S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

M. Long, Y. Soubo, N. Weiping, X. Feng, and Y. Jun. Point-spread function estimation for adaptive optics imaging of astronomical extended objects. *The Astrophysical Journal*, 888(1):20, 2019.

L. B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745–754, 1974.

J. Ma. Indirect density estimation using the iterative Bayes algorithm. *Computational Statistics & Data Analysis*, 55(3):1180–1195, 2011.

K. Maleknejad and S. Sohrabi. Numerical solution of Fredholm integral equations of the first kind by using Legendre wavelets. *Applied Mathematics and Computation*, 186(1):836–843, 2007.

F. Malrieu. Logarithmic Sobolev inequalities for some nonlinear PDEs. *Stochastic Processes and Their Applications*, 95(1):109–132, 2001.

I. C. Marschner. Back-projection of COVID-19 diagnosis counts to assess infection incidence and control measures: analysis of Australian data. *Epidemiology and Infection*, 148:e97, 2020.

MATLAB. *version 9.5.0 (R2018b)*. The Mathworks, Inc., Natick, Massachusetts, 2018.

H. McKean. A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences of the United States of America*, 56(6):1907–1911, 1966.

H. P. McKean. Propagation of chaos for a class of non-linear parabolic equations. *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967)*, pages 41–57, 1967.

L. R. Mead. Approximate solution of Fredholm integral equations by the maximum-entropy method. *Journal of Mathematical Physics*, 27(12):2903–2907, 1986.

S. Méléard. Asymptotic behaviour of some interacting particle systems; McKean-Vlasov and Boltzmann models. In *Probabilistic models for nonlinear partial differential equations*, pages 42–95. Springer, 1996.

S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

A. D. Michal. Integral equations and functionals. *Mathematics Magazine*, 24(2): 83–95, 1950.

J. Míguez, D. Crisan, and P. M. Djurić. On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.

A. C. Miller, L. Hannah, J. Futoma, N. J. Foti, E. B. Fox, A. D'Amour, M. Sandler, R. A. Saurous, and J. Lewnard. Statistical deconvolution for inference of infection time series. *medRxiv*, 2020.

R. Molina, A. del Olmo, J. Perea, and B. Ripley. Bayesian deconvolution in optical astronomy. *The Astronomical Journal*, 103:666–675, 1992.

H. Mülthei. Iterative continuous maximum-likelihood reconstruction method. *Mathematical Methods in the Applied Sciences*, 15(4):275–286, 1992.

H. Mülthei, B. Schorr, and W. Törnig. On an iterative method for a class of integral equations of the first kind. *Mathematical Methods in the Applied Sciences*, 9(1): 137–168, 1987.

H. Mülthei, B. Schorr, and W. Törnig. On properties of the iterative maximum likelihood reconstruction method. *Mathematical Methods in the Applied Sciences*, 11(3):331–342, 1989.

C. A. Naesseth, F. Lindsten, and T. B. Schon. Elements of Sequential Monte Carlo. *Foundations and Trends in Machine Learning*, 12(3):187–306, 2019.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 1161–1167, 1991.

D. Nychka. Some properties of adding a smoothing step to the EM algorithm. *Statistics & Probability Letters*, 9(2):187–193, 1990.

T. Obadia, R. Haneef, and P.-Y. Boëlle. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC medical informatics and decision making*, 12(1):1–9, 2012.

K. Oelschlager. A martingale approach to the law of large numbers for weakly interacting stochastic processes. *The Annals of Probability*, pages 458–479, 1984.

J. Olsson and T. Rydén. The bootstrap particle filtering bias. *Lund University, Technical Report 929081*, 2004.

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 2001.

B. Paige and F. Wood. Inference networks for sequential Monte Carlo in graphical models. In *International Conference on Machine Learning*, pages 3040–3049, 2016.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

M. Pensky et al. Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity. *The Annals of Statistics*, 45(4):1516–1541, 2017.

M. E. Phelps. Positron emission tomography provides molecular imaging of biological processes. *Proceedings of the National Academy of Sciences*, 97(16):9226–9233, 2000.

D. L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the Association for Computing Machinery*, 9(1):84–97, 1962.

E. Picard. Sur un théorème générale relatif aux équations intégrales de première espèce et sur quelques problèmes de physique mathématique. *Rendiconti del Circolo Matematico di Palermo*, 29:615–629, 1910.

M. K. Pitt, R. dos Santos Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151, 2012.

C. Pouchol and O. Verdier. Linear inverse problems with nonnegativity constraints through the β-divergences: sparsity of optimisers. *arXiv preprint arXiv:2006.15845*, 2020.

P. E. Protter. *Stochastic integration and differential equations*. Springer, 2005.

J. Radon. On the determination of functions from their integral values along certain manifolds. *IEEE T Med Imaging*, 5(4):170–176, 1986.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics*, pages 235–247. Springer, 1992.

J.-B. Regli and R. Silva. Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*, 2018.

E. Resmerita and R. S. Anderssen. Joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems. *Mathematical Methods in the Applied Sciences*, 30(13):1527–1544, 2007.

E. Resmerita, H. W. Engl, and A. N. Iusem. The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. *Inverse Problems*, 23(6):2575–2588, 2007.

W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

H. Risken. *The Fokker-Planck equation*. Springer, 1996.

Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

W. Rudin. *Principles of Mathematical Analysis*, volume 3. McGraw-Hill New York, 1964.

R. Salomone, L. F. South, C. C. Drovandi, and D. P. Kroese. Unbiased and consistent nested sampling via sequential Monte Carlo. *arXiv preprint arXiv:1805.03924*, 2018.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkhäuser, NY*, 55(58-63):94, 2015.

F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.

S. H. Scheres, R. Núñez-Ramírez, Y. Gómez-Llorente, C. San Martín, P. P. Eggermont, and J. M. Carazo. Modeling experimental image formation for likelihood-based classification of electron microscopy data. *Structure*, 15(10):1167–1177, 2007.

S. M. Schmon, G. Deligiannidis, A. Doucet, and M. K. Pitt. Large sample asymptotics of the pseudo-marginal method. *Biometrika*, 108(1):37–51, 2021.

A. A. Semenov, A. V. Moshkov, V. N. Pozhidayev, A. Barducci, P. Marcoionni, and I. Pippi. Estimation of normalized atmospheric point spread function and restoration of remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(7):2623–2634, 2011.

R. Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 380–402, 1982.

L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

A. Signoroni, M. Savardi, A. Baronio, and S. Benini. Deep learning meets hyperspectral image analysis: a multidisciplinary review. *Journal of Imaging*, 5(5):52, 2019.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1986.

B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka. A smoothed EM approach to indirect estimation problems, with particular, reference to stereology and emission tomography. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 52(2):271–324, 1990.

Y. Song. Gradient Estimates and Exponential Ergodicity for Mean-Field SDEs with Jumps. *Journal of Theoretical Probability*, 33(1):201–238, 2020.

J. Spanier and E. M. Gelbard. *Monte Carlo principles and neutron transport problems*. Courier Corporation, 2008.

L. A. Stefanski and R. J. Carroll. Deconvolving kernel density estimators. *Statistics*, 21(2):169–184, 1990.

A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer, Berlin, 1991.

V. P. Tanana, E. Y. Vishnyakov, and A. I. Sidikova. An approximate solution of a Fredholm integral equation of the first kind by the residual method. *Numerical Analysis and Applications*, 9(1):74–81, 2016.

The MathWorks Inc. DECONVLUCY. deblur image using Lucy–Richardson method. `https://uk.mathworks.com/help/images/ref/deconvlucy.html`, 1993. [Online; accessed 25-March-2019].

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

S. Tong, A. M. Alessio, and P. E. Kinahan. Image reconstruction for PET/CT scanners: past achievements and future challenges. *Imaging Med*, 2(5):529, 2010.

J. J. Vaquero and P. Kinahan. Positron emission tomography: current challenges and opportunities for technological advances in clinical and preclinical imaging systems. *Annual Review of Biomedical Engineering*, 17:385–414, 2015.

Y. Vardi and D. Lee. From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(3):569–612, 1993.

Y. Vardi, L. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

A. Y. Veretennikov. On ergodic measures for McKean-Vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 471–486. Springer, 2006.

A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus. Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical Bayesian approach part I: Methodology and experiments. *SIAM Journal on Imaging Sciences*, 13(4):1945–1989, 2020.

V. Volterra. *Leçons sur les équations intégrales et les équations intégro-différentielles*. Gauthier-Villars, 1913.

G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667, 1977.

M. P. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9(2):97–116, 1994.

S. Wang and L. Wang. Adaptive semiparametric Bayesian differential equations via sequential Monte Carlo. *arXiv preprint arXiv:2002.02571*, 2020.

S. Wang, X. Yang, L. Li, P. Nadler, R. Arcucci, Y. Huang, Z. Teng, and Y. Guo. A Bayesian Updating Scheme for Pandemics: Estimating the Infection Dynamics of COVID-19. *IEEE Computational Intelligence Magazine*, 15(4):23–33, 2020.

A. G. Webb. *Introduction to Biomedical Imaging*. John Wiley & Sons, 2017.

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

N. Whiteley and N. Kantas. Calculating principal eigen-functions of non-negative integral kernels: particle approximations and applications. *Mathematics of Operations Research*, 42(4):1007–1034, 2017.

W. Whiteley, W. K. Luk, and J. Gregor. DirectPET: full-size neural network PET reconstruction from sinogram data. *Journal of Medical Imaging*, 7(3):032503, 2020.

S. D. Wicksell. The corpuscle problem: a mathematical study of a biometric problem. *Biometrika*, pages 84–99, 1925.

Y. Xu, H. Zhang, Y. Li, K. Zhou, Q. Liu, and J. Kurths. Solving Fokker-Planck equation using deep learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(1):013133, 2020.

R. Yang, D. W. Apley, J. Staum, and D. Ruppert. Density deconvolution with additive measurement errors using quadratic programming. *Journal of Computational and Graphical Statistics*, pages 1–12, 2020.

S. Yang, L. Wang, and P. Ding. Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888, 2019.

D. Yuan and X. Zhang. An overview of numerical methods for the first kind Fredholm integral equation. *Springer Nature Applied Sciences*, 1(10):1178, 2019.

E. Zeidler. *Nonlinear Functional Analysis and its Applications: Part I: Fixed-Point Theorems*, volume 1. Springer, 1985.

C. Zhang, S. Arridge, and B. Jin. Expectation propagation for Poisson data. *Inverse Problems*, 35(8):085006, 2019.

Y. Zhou, A. M. Johansen, and J. A. Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.