Stochastic Processes and their Applications 173 (2024) 104374

Contents lists available at ScienceDirect

Stochastic Processes and their Applications

journal homepage: www.elsevier.com/locate/spa

Solving a class of Fredholm integral equations of the first kind via Wasserstein gradient flows

Francesca R. Crucinio [a,*], Valentin De Bortoli [b], Arnaud Doucet [c], Adam M. Johansen [d,e]

[a] Department of Mathematics, King's College London, London, WC2R 2LS, UK
[b] Computer Science Department, ENS, CNRS, PSL University, 45 rue d'Ulm, Paris, F-75230 Paris cedex 05, France
[c] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK
[d] Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[e] The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, UK

ARTICLE INFO

MSC:
65Rxx
65C35
65K10
45B05

Keywords:
Interacting particle systems
Inverse problems
McKean–Vlasov SDEs

ABSTRACT

Solving Fredholm equations of the first kind is crucial in many areas of the applied sciences. In this work we consider integral equations featuring kernels which may be expressed as scalar multiples of conservative (i.e. Markov) kernels and we adopt a variational point of view by considering a minimization problem in the space of probability measures with an entropic regularization. Contrary to classical approaches which discretize the domain of the solutions, we introduce an algorithm to asymptotically sample from the unique solution of the regularized minimization problem. As a result our estimators do not depend on any underlying grid and have better scalability properties than most existing methods. Our algorithm is based on a particle approximation of the solution of a McKean–Vlasov stochastic differential equation associated with the Wasserstein gradient flow of our variational formulation. We prove the convergence towards a minimizer and provide practical guidelines for its numerical implementation. Finally, our method is compared with other approaches on several examples including density deconvolution and epidemiology.

1. Introduction

Fredholm integral equations of the first kind are ubiquitous in applied sciences and engineering, they model density deconvolution [31,61,77,95], image reconstruction [7,23,85,96], inverse boundary problems for partial differential equations [24,88] and find applications in epidemiology [39,40,63] and statistics [45,68]. Fredholm integral equations of the first kind are typically defined by

$$g(y) = \int_{\mathbb{R}^d} f(x)\mathrm{k}(x, y)\mathrm{d}x$$

where the kernel $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is known, and $g : \mathbb{R}^p \to \mathbb{R}$ is known but indirectly observed, and the objective is to identify the unknown function $f : \mathbb{R}^d \to \mathbb{R}$. If we interpret this as a Lebesgue integral with $\mathrm{d}x$ denoting the Lebesgue measure then we may identify the functions with measures $\mu(\mathrm{d}y) := g(y)\mathrm{d}y, \pi(\mathrm{d}x) := f(x)\mathrm{d}x$ and k with an integral operator $\mathrm{K}(x, \mathrm{d}y) := \mathrm{k}(x, y)\mathrm{d}y$ which

* Corresponding author.
E-mail address: francesca_romana.crucinio@kcl.ac.uk (F.R. Crucinio).

https://doi.org/10.1016/j.spa.2024.104374
Received 25 November 2022; Received in revised form 2 April 2024; Accepted 28 April 2024
Available online 7 May 2024

leads to the representation that we will consider throughout this paper:

$$\mu(\mathsf{A}) = \pi\mathrm{K}(\mathsf{A}) := \int_{\mathbb{R}^d} \mathrm{K}(x, \mathsf{A})\mathrm{d}\pi(x), \tag{1}$$

for any Borel set $\mathsf{A} \in \mathcal{B}(\mathbb{R}^p)$, with π an unknown measure on $\mathcal{B}(\mathbb{R}^d)$, μ an observed measure on $\mathcal{B}(\mathbb{R}^p)$ and $\mathrm{K} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^p) \to \mathbb{R}$ an integral operator. These equations model the task of reconstructing the signal π from a distorted observed version μ. Since many applications of Fredholm integral equations are concerned with the reconstruction of signals that are *a priori* known to be non-negative [23], we consider the case in which both $\pi \in \mathscr{P}(\mathbb{R}^d)$ and $\mu \in \mathscr{P}(\mathbb{R}^p)$ are probability measures on $\mathbb{R}^d$ and $\mathbb{R}^p$, respectively, and K is a Markov kernel. A great many Fredholm integral equations, particularly those for which the kernel is conservative up to a scalar multiplier in the sense that $\int_{\mathbb{R}^p} \mathrm{K}(x, \mathrm{d}y) = \int_{\mathbb{R}^p} \mathrm{K}(x', \mathrm{d}y)$ for every $x, x' \in \mathbb{R}^d$, can be recast in this framework with appropriate translation and normalization [20, Section 6].

In a probabilistic and variational approach, we seek a solution of (1) by minimizing the Kullback–Leibler divergence between μ and $\pi\mathrm{K}$ [73]:

$$\pi^\star = \arg\min\{\mathrm{KL}(\mu|\pi\mathrm{K}) \ : \ \pi \in \mathscr{P}(\mathbb{R}^d)\}, \tag{2}$$

where $\mathrm{KL}(\mu|\pi\mathrm{K}) = \int_{\mathbb{R}^p} \log((\mathrm{d}\mu/\mathrm{d}\pi\mathrm{K})(y))\mathrm{d}\mu(y)$ if μ admits a density w.r.t. $\pi\mathrm{K}$ and $+\infty$ otherwise. The probability measure $\pi^\star$ corresponds to the maximum likelihood estimator (MLE) for π. Using this formulation, it is possible to define approximate solutions of Fredholm integral equations of the first kind as solutions of a minimization problem on the space of measures. However, in most interesting cases, the integral equation (1) is ill-posed, *i.e.* there might exists more than one solution π to (1), and, even when the solution is unique, it is unstable with respect to changes in μ. This lack of stability is the primary concern when attempting to solve Fredholm integral equations of the first kind: in practical applications we often only have access to (noisy) observations from μ and not to its analytic form, and the instability of (1) means that small errors in μ do not necessarily correspond to small errors in the recovered solution π [57, Chapter 15]. This issue reflects onto the minimization problem (2), which often does not admit a unique minimizer [58].

In order to circumvent this issue, a regularization term is often added to the variational formulation (2) [17,42]. In a similar fashion to the Tikhonov regularization [43], we consider a regularized version of (2) in which a cross-entropic penalty with respect to a reference measure $\pi_0 \in \mathscr{P}(\mathbb{R}^d)$ is introduced as in [37,50,80]

$$\pi^\star = \arg\min\{\mathscr{F}_\alpha(\pi) \ : \ \pi \in \mathscr{P}(\mathbb{R}^d)\}, \qquad \mathscr{F}_\alpha(\pi) = \mathrm{KL}(\mu|\pi\mathrm{K}) + \alpha\mathrm{KL}\left(\pi|\pi_0\right), \tag{3}$$

for a given regularization parameter $\alpha > 0$. In practice, (3) becomes easier to solve for large values of α but the agreement between $\pi^\star\mathrm{K}$ and μ is weak and in the limiting case $\alpha \to +\infty$, $\pi^\star = \pi_0$. For small values of α, $\pi^\star\mathrm{K} \approx \mu$ but the minimization problem is harder to solve and in the limit $\alpha \to 0$, it collapses onto (2), becoming ill-posed. The parameter α should therefore be chosen to achieve a trade-off between minimizing $\mathrm{KL}(\mu|\pi\mathrm{K})$ (i.e. α should be small enough), ensuring that (3) can be solved and that the resulting solution is sufficiently regular (i.e. α should not be zero); this is usually achieved by selecting α through cross-validation, see [1,91].

The use of cross-entropy regularization is not new in the literature on Fredholm integral equations [37,50,80]. In particular, [37] considers a cross-entropy regularization in a least-square setting, [50] studies (3) in the discrete setting, i.e. when π and μ are vectors in the simplex of $\mathbb{R}^d, \mathbb{R}^p$ respectively, and [80] studies (3) when π and μ are integrable densities. In the discrete setting, [50] establishes uniqueness of the minimizers of a discrete version of (3) and studies the behaviour of the minimizers as $\alpha \to 0$; in the continuous setting, [80] shows existence and uniqueness of a minimizer of (3) under the assumption that (1) admits a maximum entropy solution.

In related works, functionals involving the Kullback–Leibler divergence $\mathrm{KL}(\mu|\pi\mathrm{K})$ are minimized iteratively by discretizing the domain of π, μ [15,20,42,85,95], while maximum entropy estimators are usually obtained approximating π through a set of basis functions [49,52,56,65]. However, when π, μ are probability measures it is natural to consider Monte Carlo strategies, and approximate π using particles [26]. This is particularly convenient when the only information on μ is given by a sample drawn from it, as it is the case in most applications [31,39,40,45,61,63,68,77,95]. Recently, [26] introduced a sequential Monte Carlo method (SMC-EMS) to approximately solve (1), this algorithm provides an adaptive stochastic discretization of the expectation maximization smoothing algorithm first studied in [84], which aims at obtaining solutions of (1) which achieve a good trade-off between solving (2) and having enough smoothness.

The functional (3) can be seen as a probabilistic counterpart of the classic Tikhonov regularization setting (e.g [43]); [38] propose an interacting particle system based on a Wasserstein gradient flow construction to approximately solve the Tikhonov-regularized problem for a wide class of (nonlinear) inverse problems. While the particle system introduced in [38] is similar in spirit to the approach taken here, the former requires an a priori discretization of the support of π, which makes it impractical when we deal with Fredholm integral equations, since the object of interest (*i.e.* π) is infinite dimensional. By focussing on the case of linear integral equations, we can avoid the discretization step required by the particle system of [38], and work directly with the measures π, μ.

In this work, we adopt a probabilistic approach and present a novel particle algorithm to solve Fredholm integral equations of the first kind. More precisely, we derive a particle approximation to the Stochastic Differential Equation (SDE) associated with the Wasserstein gradient flow [2,81] minimizing surrogates of (3). Wasserstein gradient flows have been widely used to solve a variety of optimization problems in the space of probability measures [5,35,53,59], including finding solutions to finite dimensional inverse problems [38], but, to the best of our knowledge, not for infinite dimensional inverse problems like the integral equation (1). The SDE associated with this gradient flow is non-linear in the sense that it depends not only upon the position of the solution but also

on its marginal distribution. Such SDEs are called McKean–Vlasov SDEs (MKVSDE), and have been actively studied during the past decades [18,41,64,66,67,75,86,87]. In this work, we derive a particular MKVSDE whose solution converges towards a minimizer of (3). However, this MKVSDE cannot be solved analytically, and we introduce a numerical approximation using an Euler–Maruyama discretization and a particle system.

We perform an extensive simulation study in which we compare our method with SMC-EMS and problem-specific algorithms and show that it achieves comparable performances. The results obtained with our method and SMC-EMS are particularly similar, since both methods aim at minimizing a regularized Kullback–Leibler divergence. However, while in this work we explicitly have a minimization problem for the regularized functional (3), the SMC-EMS algorithm does not minimize a particular functional but corresponds to a smoothed version of the EM algorithm minimizing the Kullback–Leibler divergence (2). When α and the smoothness parameter controlling the amount of regularization in SMC-EMS are both small, the two algorithms will therefore give very similar reconstructions. Nevertheless, the method proposed in this work comes with a number of additional theoretical guarantees and practical improvements. From the theoretical point of view, we show that (3) admits a unique minimizer for any α and that this minimizer can be seen as the probabilistic counterpart of the classical solution to the Tikhonov regularized least-squares problem. On the other hand, uniqueness of the fixed point of the EMS recursion approximated by SMC-EMS has not been established. From the practical point of view, minimizing (3) allows us to introduce additional information about the solution through π_0, this proves particularly beneficial when the kernel K leads to model misspecification (see Section 5.2). When compared on higher dimensional problems, the proposed method outperforms both standard discretization-based methods and SMC-EMS (Section 5.4), opening the door to solving Fredholm integral equations for high-dimensional problems.

In addition, the value of the regularization parameter α can be chosen through common approaches for selecting the regularization parameter in Tikhonov regularization settings (e.g. cross–validation [1,91]), while in the case of SMC-EMS the link between the particle system and the functional which is minimized is less clear, and cannot be exploited as straightforwardly.

To summarize, our main contributions are as follows:

(a) First, we establish that the functional in (3) is continuous and admits a unique minimizer under mild conditions on μ, K. In addition, we show that a modification of (3) is stable with respect to perturbations of the observation μ and study the limit behaviour of $\pi^\star$ as $\alpha \to 0$.
(b) Second, we derive the Wasserstein gradient flow used to minimize a surrogate of (3) and show that the associated MKVSDE converges towards a unique minimizer using recent results from optimal transport [46].
(c) Third, we present and study a discrete-time particle approximation of this MKVSDE. In particular, we establish geometric ergodicity of the particle system under regularity conditions.
(d) Finally, we illustrate the efficiency of our method to approximately solve Fredholm integral equations of the first kind on a range of different numerical examples and show that it achieves superior or comparable performances to those of problem-specific algorithms and SMC-EMS, outperforming the latter for high dimensional problems. In addition, we show on a task from epidemiology that, when the kernel K does not correctly model the problem at hand, the presence of a reference measure π_0 allows us to outperform even problem-specific estimators.

The rest of the manuscript is organized as follows. In Section 2 we establish that the functional in (3) is well-founded and study its key properties. The link between gradient flows and MKVSDE in the context of Fredholm integral equations of the first kind is studied in Section 3. In Section 4 we describe a numerical scheme which approximates the continuous time dynamics of the particle system and provide error bounds. Finally, in Section 5, we test the algorithm on a number of applications of Fredholm integral equations from statistics, epidemiology and image processing, and explore the scaling properties of our method through a comparison with the SMC-EMS algorithm recently proposed in [26].

Notation

We endow $\mathbb{R}^d$ with the Borel σ-field $\mathcal{B}(\mathbb{R}^d)$ with respect to the Euclidean norm $\|\cdot\|$. For a matrix A we denote $\|A\| = (\sum_{ij} a_{ij}^2)^{1/2}$ the Hilbert–Schmidt norm. We denote by $\mathrm{C}(\mathbb{R}^d)$ the set of continuous functions defined over $\mathbb{R}^d$ and by $\mathrm{C}^n(\mathbb{R}^d)$ the set of n-times differentiable functions defined over $\mathbb{R}^d$ for any $n \in \mathbb{N}^\star$, where $\mathbb{N}^\star$ denotes the non-zero natural numbers. For all $f \in \mathrm{C}^1(\mathbb{R}^d)$, we denote by ∇f its gradient. Furthermore, if $f \in \mathrm{C}^2(\mathbb{R}^d)$ we denote by $\nabla^2 f$ its Hessian and by Δf its Laplacian. For all $f \in \mathrm{C}^1(\mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_m})$ with $m \in \mathbb{N}^\star$, we denote by $\nabla_i f$ its gradient w.r.t. component i. We say that a function $f : \mathsf{X} \to \mathbb{R}$ (where X is a metric space) is coercive if for any $t \in \mathbb{R}$, $f^{-1}((-\infty, t])$ is relatively compact. This definition can be extended to the case where X is only a topological space, see [28, Definition 1.12]. A family of functions $\{f_\alpha : \alpha \in \mathsf{A}\}$ such that for any $\alpha \in \mathsf{A}$, $f_\alpha : \mathsf{X} \to \mathbb{R}$, is said to be equicoercive if there exists $g : \mathsf{X} \to \mathbb{R}$ such that for any $\alpha \in \mathsf{A}$, $f_\alpha \geq g$ and g is coercive. For any $p \in \mathbb{N}$ we denote by $\mathscr{P}_p(\mathbb{R}^d) = \{\pi \in \mathscr{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_p^p \,\mathrm{d}\pi(x) < +\infty\}$ the set of probability measures over $\mathcal{B}(\mathbb{R}^d)$ with finite pth moment. For ease of notation, we define $\mathscr{P}(\mathbb{R}^d) = \mathscr{P}_0(\mathbb{R}^d)$ the set of probability measures over $\mathcal{B}(\mathbb{R}^d)$ and endow this space with the topology of weak convergence. For any $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$ we define the p-Wasserstein distance $\mathbf{W}_p(\mu, \nu)$ between μ and ν by

$$\mathbf{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \mathbf{T}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p \,\mathrm{d}\gamma(x, y) \right)^{1/p}$$

where $\mathbf{T}(\mu, \nu) = \{\gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) : \gamma(\mathsf{A} \times \mathbb{R}^d) = \mu(\mathsf{A}),\ \gamma(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})\ \forall \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\}$ denotes the set of all transport plans between μ and ν. In the following, we metrize $\mathscr{P}_p(\mathbb{R}^d)$ with $\mathbf{W}_p$. Assume that $\mu \ll \nu$ and denote by $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ its Radon–Nikodym derivative. The

Kullback–Leibler divergence, $\mathrm{KL}(\mu|\nu)$, between μ and ν is defined by

$$\mathrm{KL}(\mu|\nu) = \int_{\mathsf{X}} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)\mathrm{d}\mu(x).$$

We say that $\mathrm{K} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^p) \to [0,+\infty)$ is a Markov kernel if for any $x \in \mathbb{R}^d$, $\mathrm{K}(x,\cdot)$ is a probability measure over $\mathcal{B}(\mathbb{R}^p)$ and for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^p)$, $\mathrm{K}(\cdot,\mathsf{A})$ is a Borel-measurable function over $\mathbb{R}^d$.

2. A minimization problem for Fredholm integral equations

In this section, we study the properties of the regularized functional $\mathscr{F}_\alpha : \mathscr{P}(\mathbb{R}^d) \to \mathbb{R} \cup +\infty$ given for any $\alpha \geq 0$ and $\pi \in \mathscr{P}(\mathbb{R}^d)$ by

$$\mathscr{F}_\alpha(\pi) = \mathrm{KL}(\mu|\pi\mathrm{K}) + \alpha\mathrm{KL}\left(\pi|\pi_0\right),$$

where we recall that $\mu \in \mathscr{P}(\mathbb{R}^p)$, $\pi_0 \in \mathscr{P}(\mathbb{R}^d)$ and the Markov kernel $\mathrm{K} : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^p) \to [0,1]$ are given. For ease of notation we denote $\mathscr{F} = \mathscr{F}_0$.

In Section 2.1 we introduce a surrogate functional which enjoys better stability properties than $\mathscr{F}_\alpha$. We prove that under mild assumption on μ and K, a minimizer of the surrogate exists, is unique and continuous w.r.t the parameters of the functional. In Section 2.2, we investigate the links between our approach and other regularization methods for Fredholm integration equations of the first kind.

2.1. A surrogate functional: stability and regularization

First, we introduce a surrogate functional $\mathscr{G}_\alpha$ whose minimizers are the same as $\mathscr{F}_\alpha$ if μ is absolutely continuous w.r.t the Lebesgue measure but which is also easier to study when μ is approximated by an empirical distribution. Assume that there exists $\mathrm{k} : \mathbb{R}^d \times \mathbb{R}^p \to [0,+\infty)$ such that for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^p)$, $\mathrm{K}(x,\mathsf{A}) = \int_{\mathsf{A}} \mathrm{k}(x,y)\mathrm{d}y$. Then, using the Fubini–Tonelli theorem we have that for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^p)$ and $\pi \in \mathscr{P}(\mathbb{R}^d)$,

$$\pi\mathrm{K}(\mathsf{A}) = \int_{\mathbb{R}^d} \mathrm{K}(x,\mathsf{A})\mathrm{d}\pi(x) = \int_{\mathsf{A}}\left(\int_{\mathbb{R}^d} \mathrm{k}(x,y)\mathrm{d}\pi(x)\right)\mathrm{d}y.$$

Therefore $\pi\mathrm{K}$ admits a density w.r.t. the Lebesgue measure given by $y \mapsto \pi[\mathrm{k}(\cdot,y)]$. In addition, assume that μ admits a density w.r.t. the Lebesgue measure $\mathrm{d}\mu(y) = \mu(y)\mathrm{d}y$ such that $\int_{\mathbb{R}^p} |\log(\mu(y))|\mathrm{d}\mu(y) < +\infty$. Then for any $\pi \in \mathscr{P}(\mathbb{R}^d)$

$$\mathrm{KL}(\mu|\pi\mathrm{K}) = -\mathrm{H}(\mu) - \int_{\mathbb{R}^p} \log(\pi[\mathrm{k}(\cdot,y)])\mathrm{d}\mu(y),$$

where $\mathrm{H}(\mu) := -\int_{\mathbb{R}^p} \log(\mu(y))\mathrm{d}\mu(y)$ is the entropy of μ.

Since $\int_{\mathbb{R}^p} |\log(\pi[\mathrm{k}(\cdot,y)])|\,\mathrm{d}\mu(y) < +\infty$ if $\mathrm{KL}(\mu|\pi\mathrm{K}) < +\infty$, minimizing $\mathscr{F}_\alpha$ on the set $\{\pi \in \mathscr{P}(\mathbb{R}^d) : \mathrm{KL}(\mu|\pi\mathrm{K}) < +\infty\}$ for any $\alpha \geq 0$ is equivalent to minimizing $\mathscr{G}_\alpha$ on the same set, where for any $\pi \in \mathscr{P}(\mathbb{R}^d)$

$$\mathscr{G}_\alpha(\pi) = -\int_{\mathbb{R}^p} \log(\pi[\mathrm{k}(\cdot,y)])\mathrm{d}\mu(y) + \alpha\mathrm{KL}\left(\pi|\pi_0\right), \tag{4}$$

with $\mathscr{G} = \mathscr{G}_0$ for ease of notation.

The functional (4) is not defined if $\int_{\mathbb{R}^p} \log(\pi[\mathrm{k}(\cdot,y)])\mathrm{d}\mu(y) = +\infty$ or $\mathrm{KL}\left(\pi|\pi_0\right) = +\infty$. We consider the following assumption on the kernel K so that the functional is well-defined for any $\pi \in \mathscr{P}(\mathbb{R}^d)$ and regular enough, see Proposition 1.

A1. There exists $\mathrm{k} \in \mathrm{C}^\infty(\mathbb{R}^d \times \mathbb{R}^p, [0,+\infty))$ such that for any $x \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^p)$, $\mathrm{K}(x,\mathsf{A}) = \int_{\mathsf{A}} \mathrm{k}(x,y)\mathrm{d}y$ and $-\int_{\mathbb{R}^p} \log(\mathrm{k}(0,y))\mathrm{d}\mu(y) < +\infty$. In addition, there exists $\mathtt{M} \geq 0$ such that for any $(x,y) \in \mathbb{R}^d \times \mathbb{R}^p$ we have $\mathrm{k}(x,y) + \|\nabla\mathrm{k}(x,y)\| + \|\nabla^2\mathrm{k}(x,y)\| \leq \mathtt{M}$. In addition, $\|\partial^3_{1,ijl}\mathrm{k}(x,y)\| \leq \mathtt{M}$, where ∂_i denotes the derivative w.r.t. the ith variable.

In this section, we could restrict ourselves to bounded second derivatives. The condition $\|\partial^3_{1,ijl}\mathrm{k}(x,y)\| \leq \mathtt{M}$ is used only in Proposition 10, in which we establish a convergence result for the time discretization of the MKVSDE. Under A1, the density k of the Markov kernel K is Lipschitz continuous for every $y \in \mathbb{R}^p$.

To address stability issues we also consider the following regularized functional: for any $\alpha, \eta \geq 0$ and $\pi \in \mathscr{P}(\mathbb{R}^d)$

$$\mathscr{G}^\eta_\alpha(\pi) = -\int_{\mathbb{R}^p} \log(\pi[\mathrm{k}(\cdot,y)] + \eta)\mathrm{d}\mu(y) + \alpha\mathrm{KL}\left(\pi|\pi_0\right), \tag{5}$$

where $\eta \geq 0$ is a hyperparameter. Using $\mathscr{G}^\eta_\alpha$ with $\eta > 0$ will be crucial in order to establish the Lipschitz continuity of the drift of the MKVSDE in Section 3, and hence the stability of the proposed procedure. For clarity, we write $\mathscr{G}_\alpha = \mathscr{G}^0_\alpha$, $\mathscr{G}^\eta = \mathscr{G}^\eta_0$ and $\mathscr{G} = \mathscr{G}^0_0$. In the following proposition, we derive regularity properties for the family of functionals $\{\mathscr{G}^\eta : \eta \geq 0\}$.

Proposition 1. *Assume A1, then the following hold:*

(a) For any $\eta \geq 0$, $\mathscr{G}^\eta$ is lower bounded, lower-semi continuous, convex and is not coercive.
(b) For any $\eta > 0$, $\mathscr{G}^\eta$ is proper and $\mathscr{G}^\eta \in \mathrm{C}(\mathscr{P}(\mathbb{R}^d), \mathbb{R})$.

Proof. See Appendix A.2. □

Proposition 1 shows that $\mathscr{G}^\eta$ enjoys better regularity properties for $\eta > 0$. In Proposition 3, we will also show that if $\eta, \alpha > 0$ then the minimizers of $\mathscr{G}^\eta_\alpha$ are stable w.r.t μ. In addition, note that for any $\eta \geq 0$, $\mathscr{G}^\eta$ is *not* coercive. As a result, the uniqueness of the minimizer of $\mathscr{G}^\eta$ is difficult to establish. In what follows we study the function $\mathscr{G}^\eta_\alpha$ for $\alpha, \eta > 0$ and show that $\mathscr{G}^\eta_\alpha$ is coercive in this case. Hence, $\mathscr{G}^\eta_\alpha$ admits a unique minimizer $\pi^\star_{\alpha,\eta}$, and we show that the family of minimizers $\{\pi^\star_{\alpha,\eta} : \alpha, \eta > 0\}$ is smooth with respect to the parameters α and η. In Proposition 4 we will characterize the limit of $\pi^\star_{\alpha,\eta}$ when both $\alpha, \eta \to 0$.

Proposition 2. *Assume A1. Then the following hold:*

(a) For any $\alpha, \eta > 0$, $\mathscr{G}^\eta_\alpha$ is proper, strictly convex, coercive and lower semi-continuous. In particular, $\mathscr{G}^\eta_\alpha$ admits a unique minimizer $\pi^\star_{\alpha,\eta} \in \mathscr{P}(\mathbb{R}^d)$.

(b) $\mathtt{m} : (\alpha, \eta) \mapsto \pi^\star_{\alpha,\eta} \in \mathrm{C}((0,+\infty)^2, \mathscr{P}(\mathbb{R}^d))$ and $\mathtt{d} : (\alpha, \eta) \mapsto \inf_{\mathscr{P}(\mathbb{R}^d)} \mathscr{G}^\eta_\alpha \in \mathrm{C}((0,+\infty)^2, \mathbb{R})$.

Proof. See Appendix A.3. □

In most applications, μ is unknown but we have access to samples $y^{1:M} = \{y^{k,M}\}_{k=1}^M \in (\mathbb{R}^p)^M$ drawn from μ, see [31,45,61,63,68,77,95]. The following stability result guarantees that if we substitute μ with the empirical measure $\mu^M = (1/M)\sum_{k=1}^M \delta_{y^{k,M}}$ in $\mathscr{G}^\eta_\alpha$ we obtain a minimization problem whose solution converges to the minimizer we would obtain if we had full knowledge of μ as the number of samples M increases.

Proposition 3. *Assume A1. Let $\alpha, \eta > 0$ and for any $\nu \in \mathscr{P}(\mathbb{R}^p)$ denote $\pi^{\nu,\star}_{\alpha,\eta}$ the unique minimizer of (5) with $\mu \leftarrow \nu$. Let $(\mu_n)_{n\in\mathbb{N}} \in (\mathscr{P}_1(\mathbb{R}^d))^{\mathbb{N}}$ such that $\lim_{n\to+\infty} \mathbf{W}_1(\mu_n, \mu) = 0$ with $\mu \in \mathscr{P}_1(\mathbb{R}^d)$. Then, we have $\lim_{n\to+\infty} \mathbf{W}_1(\pi^{\mu_n,\star}_{\alpha,\eta}, \pi^{\mu,\star}_{\alpha,\eta}) = 0$.*

Proof. See Appendix A.4. □

We emphasize that in the proof of Proposition 3, the fact that $\eta > 0$ plays a crucial role. Indeed, if $\eta > 0$ and A1 holds we can exploit the fact that $y \mapsto \log(\pi[\mathrm{k}(\cdot, y)] + \eta)$ is Lipschitz continuous.

2.2. Variants and connections with other methods

2.2.1. Maximum entropy methods

The functional $\mathscr{F}_\alpha$ can be seen as the Lagrangian associated with the following primal problem

$$\arg\min\{\mathrm{KL}\left(\pi|\pi_0\right) : \pi \in \mathscr{P}(\mathbb{R}^d),\ \mathrm{KL}\left(\mu|\pi\mathrm{K}\right) = 0\}.$$

The latter problem is a maximum entropy problem in the sense of [27] (where the entropy functional is replaced by the Kullback–Leibler divergence). Closely related to (3) is the functional $\tilde{\mathscr{F}}_\alpha$ given (when it is defined) by

$$\tilde{\mathscr{F}}_\alpha(\pi) = \mathrm{KL}\left(\mu|\pi\mathrm{K}\right) - \alpha\mathrm{H}(\pi). \tag{6}$$

The choice of an entropic penalty does not require specifying a reference measure π_0 and connects (3) with the classical maximum entropy methods [51]. Indeed, (6) is the Lagrangian associated with the following primal problem

$$\arg\max\{\mathrm{H}(\pi) : \pi \in \mathscr{P}_{\mathrm{H}}(\mathbb{R}^d),\ \mathrm{KL}\left(\mu|\pi\mathrm{K}\right) = 0\},$$

where $\mathscr{P}_{\mathrm{H}}(\mathbb{R}^d)$ is the set of probability distributions with finite entropy. However, the functional $\tilde{\mathscr{F}}_\alpha$ is not lower bounded and the corresponding minimization problem is not well-defined and therefore is not considered here. On the contrary, $\mathscr{F}_\alpha$ is always non-negative making the optimization problem well-defined. A number of maximum entropy approximations of solutions of Fredholm integral equations have been proposed in the literature, most of which maximize the entropy subject to moment constraints obtained by integrating μ and $\pi\mathrm{K}$ w.r.t. a (possibly infinite) set of basis functions [49,52,56,65]. These approaches work particularly well in the one-dimensional case, as in this case the maximum entropy solution can be written analytically (see, e.g., [49, Proposition 3.1]). For higher dimensional problems, the maximum entropy solution is usually approximated by discretizing the support of π (see, e.g., [70]).

2.2.2. Generalized Bayesian inference

As mentioned in the introduction, minimizing $\mathscr{F}$ results in the maximum likelihood estimator (2). Given a set of observations $y_1, \ldots, y_M$ from μ, the log-likelihood for the infinite-dimensional parameter π is [19]

$$(1/M)\sum_{j=1}^M \log \pi[\mathrm{k}(\cdot, y_j)] = \int_{\mathbb{R}^p} \log \pi[\mathrm{k}(\cdot, y)]\mathrm{d}\mu^M(y), \tag{7}$$

with $\mu^M = (1/M)\sum_{j=1}^M \delta_{y_j}$. For any $\alpha > 0$, minimizing the functional $\mathscr{G}_\alpha$ in (4) is equivalent to minimizing

$$-\frac{1}{\alpha}\int_{\mathbb{R}^p} \log(\pi[\mathrm{k}(\cdot, y)])\mathrm{d}\mu(y) + \mathrm{KL}\left(\pi|\pi_0\right),$$

which can be interpreted as generalized Bayesian inference in the sense of adjusting the weight (or learning rate) of the likelihood (7) w.r.t. the prior π_0 (see, e.g., [11, Section 3] and [44]). In particular, if we assume that the target measure π is a discrete distribution with a fixed and finite number of components (a straightforward case which is not considered here), *i.e.*$\pi = \sum_{i=1}^N \delta_{x^i} p_i$, with $\sum_{i=1}^N p_i = 1$ and $\{x^i\}_{i=1}^N \in (\mathbb{R}^d)^N$, then (1) is a finite mixture model and minimizing (4) corresponds to posterior inference for the locations $\{x^i\}_{i=1}^N$ and the weights $\{p_i\}_{i=1}^N$ with prior π_0 and learning rate α^{-1}.

Choosing the appropriate α then amounts to the calibration problem in generalized Bayesian inference with standard Bayesian inference obtained when $\alpha = 1$ [11,44]. Contrary to the standard setting, in which values of $0 < \alpha^{-1} < 1$ are preferred since the model is believed to be misspecified, we are in the setting in which we want to reduce the influence of the prior π_0 which is used as a regularizer in (4) and favour values of $\alpha^{-1} > 1$.

2.2.3. Tikhonov regularization

We now move onto considering the link between $\mathscr{G}_\alpha$ and Tikhonov regularization. In fact, minimizing $\mathscr{G}_\alpha$ can be seen as the probabilistic counterpart to the classical Tikhonov regularization setting [43] which we recall briefly.

Let $\mathrm{A} \in \mathbb{R}^{p\times d}$ and $y \in \mathbb{R}^p$ (note that this discussion is still valid in more general Hilbert spaces but we restrict ourselves to the finite-dimensional case for the sake of clarity). Define $g : \mathbb{R}^d \to [0, +\infty)$ for any $x \in \mathbb{R}^d$ by $g(x) = \|\mathrm{A}x - y\|^2$. Denote by $\mathrm{A}^\dagger$ the Moore–Penrose inverse of A, see [72] for a definition. $\mathrm{A}^\dagger y \in \arg\min_{\mathbb{R}^d} g$ and moreover $\mathrm{A}^\dagger y \in \arg\min\{x \mapsto \|x\| : x \in \arg\min_{\mathbb{R}^d} g\}$, *i.e.* $\mathrm{A}^\dagger y$ is the solution of the minimization of g with least norm. For any $\alpha \geq 0$, the Tikhonov regularization with level α of g is given by $g_\alpha : \mathbb{R}^d \to [0, +\infty)$ such that $g_\alpha(x) = g(x) + \alpha \|x\|^2$ for any $x \in \mathbb{R}^d$. There exists a unique $x_\alpha \in \mathbb{R}^d$ such that $x_\alpha \in \arg\min_{\mathbb{R}^d} g_\alpha$. It is known that $\lim_{\alpha\to 0} x_\alpha = \mathrm{A}^\dagger y$, *i.e.* the solution of the Tikhonov regularization converges towards the solution of g with minimal norm [43].

In Proposition 4, we show a similar property in the probabilistic setting. In this case the norm function is replaced by the Kullback–Leibler divergence with respect to some reference measure $\pi \mapsto \mathrm{KL}(\pi|\pi_0)$. Given the functional $\mathscr{G} : \mathscr{P}(\mathbb{R}^d) \to (-\infty, \infty]$ in (5), we obtain for any $\alpha, \eta \geq 0$ a regularized functional $\mathscr{G}_\alpha^\eta(\pi) = \mathscr{G}^\eta(\pi) + \alpha\mathrm{KL}(\pi|\pi_0)$. Similarly to the Tikhonov regularization, there exists a unique minimizer to $\mathscr{G}_\alpha^\eta$, $\pi^\star_{\alpha,\eta}, \in \mathscr{P}(\mathbb{R}^d)$, see Proposition 2. In Proposition 4, we show that any limiting point of the sequences $(\pi^\star_{\alpha_n,\eta_n})_{n\in\mathbb{N}}$ with $\lim_{n\to+\infty} \alpha_n = 0$ and $\lim_{n\to+\infty} \eta_n = 0$, is a minimizer of $\mathscr{G}$ with least Kullback–Leibler divergence with respect to π_0.

To control the behaviour of the family of minima $\{\inf_{\pi\in\mathscr{P}(\mathbb{R}^d)} \mathscr{G}_\alpha^\eta(\pi) : \alpha, \eta > 0\}$ when α is close to 0, we make use of the following assumption which controls the tail behaviour of π_0 and the tail behaviour of the density k.

A2. The following hold:

(a) π_0 admits a density w.r.t. the Lebesgue measure, $\mathrm{d}\pi_0(x) = \pi_0(x)\mathrm{d}x$, with $\pi_0(x) \propto \exp[-U(x)]$, where, $U : \mathbb{R}^d \to \mathbb{R}$ is such that there exist $\tau, C_1 > 0$ satisfying for any $x \in \mathbb{R}^d$

$$-C_1 - \tau\|x\|^2 \leq U(x) \leq C_1 + \tau\|x\|^2.$$

(b) $\mu \in \mathscr{P}_2(\mathbb{R}^p)$ and there exists $C_2 \geq 0$ such that for any $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^p$

$$\mathrm{k}(x, y) \geq C_2^{-1} \exp[-C_2(1 + \|x\|^2 + \|y\|^2)].$$

Under this assumption we have the following result.

Proposition 4. *Assume A1 and A2. Then, for any $\alpha > 0$, $\eta \geq 0$, $\pi^\star_{\alpha,\eta} \in \mathscr{P}_2(\mathbb{R}^d)$ admits a density w.r.t the Lebesgue measure. If there exist $\pi^\star \in \mathscr{P}_2(\mathbb{R}^d)$, $(\alpha_n)_{n\in\mathbb{N}} (0, +\infty)^{\mathbb{N}}$, $(\eta_n)_{n\in\mathbb{N}} \in [0, +\infty)^{\mathbb{N}}$ such that $\lim_{n\to+\infty} \alpha_n = 0$, $\lim_{n\to+\infty} \eta_n = 0$ and $\lim_{n\to+\infty} \mathbf{W}_2(\pi^\star_{\alpha_n,\eta_n}, \pi^\star) = 0$, then*

$$\pi^\star \in \arg\min\{\mathrm{KL}(\pi|\pi_0) : \pi \in \arg\min_{\mathscr{P}_2(\mathbb{R}^d)} \mathscr{G}\}.$$

Proof. See Appendix A.5. □

3. McKean–Vlasov stochastic differential equation and minimization of functionals

In most interesting cases the direct minimization of (5) is not tractable. In this work, we propose to follow a Wasserstein gradient flow approach in order to approximate these minimizers. Our methodology stems from the connection between minimization of functionals in the space of probability measures and partial differential equations (PDEs) pointed out in [53,76]. In particular, we draw links between the minimization of $\mathscr{G}_\alpha^\eta$ and a SDE whose invariant measure is the minimizer of $\mathscr{G}_\alpha^\eta$. An informal presentation of the bridges between Wasserstein gradient flows and a certain class of SDEs, namely McKean–Vlasov SDEs (MKVSDEs), is presented in Section 3.1. We investigate the long-time behaviour of these MKVSDEs in the context of the minimization of $\mathscr{G}_\alpha^\eta$ in Section 3.2. Finally, we show that there exists a system of interacting SDEs which approximates the obtained MKVSDE in Section 3.3.

3.1. Minimization and gradient flows

In this section, we set $\alpha, \eta > 0$ and informally derive the continuous-time dynamics we study in the rest of this paper by drawing a link with gradient flows; formal results are given in Appendix B.1. We assume that all the probability measures we consider admit densities w.r.t the Lebesgue measure and do not distinguish between the distribution and its density. Let $(\pi_t)_{t\geq 0}$ be a family of probability measures satisfying (in a weak sense)

$$\partial_t \pi_t = -\mathrm{div}(\pi_t \mathbf{v}_t), \tag{8}$$

where $\mathbf{v}_t$ belongs to the Wasserstein subdifferential of a functional $\mathscr{F} : \mathscr{P}_2(\mathbb{R}^d) \to [0, +\infty]$, see [2, Definition 10.1.1] for a definition. Then $(\pi_t)_{t\geq 0}$ is called a Wasserstein gradient flow associated with $\mathscr{F}$ [2, Definition 11.1.1] and it has been shown in numerous settings that such gradient flows converge towards the minimizers of $\mathscr{F}$ when they exist [2,81]. In Proposition 19 in Appendix B.1 we show that

$$x \mapsto -\int_{\mathbb{R}^p} (\eta + \pi[\mathrm{k}(x,y)])^{-1} \nabla_1 \mathrm{k}(x,y) \mathrm{d}\mu(y) + \alpha\{\nabla U(x) + \nabla \log \pi(x)\},$$

belongs to the subdifferential of $\mathscr{G}_\alpha^\eta$ at π under mild conditions and assuming that $\pi_0(x) \propto \exp[-U(x)]$. We also derive the corresponding gradient flow equation

$$\partial_t \pi_t = -\mathrm{div}\left(\pi_t \left[\int_{\mathbb{R}^p} (\pi_t[\mathrm{k}(\cdot,y)] + \eta)^{-1} \nabla_1 \mathrm{k}(x,y) \mathrm{d}\mu(y) - \alpha \nabla U(x)\right]\right) + \alpha \Delta \pi_t. \tag{9}$$

For strongly geodesically convex (i.e. convex along geodesics) functionals the gradient flow (8) converges geometrically towards the unique minimizer. However, in our setting $\mathscr{G}_\alpha^\eta$ is not geodesically convex but only convex. Using recent results from [46] we establish the convergence of the Wasserstein gradient flow in Section 3.2 but without quantitative convergence rates.

We use the gradient flow equation (9) to draw a link between the minimization of $\mathscr{G}_\alpha^\eta$ and MKVSDEs [64], a class of SDEs in which the drift and diffusion coefficients depend not only on the current position of the solution but also on its distribution. Since (9) is a Fokker–Planck equation, we informally derive the corresponding SDE

$$\mathrm{d}\mathbf{X}_t^\star = \{\int_{\mathbb{R}^p} \tilde{b}_\eta(\mathbf{X}_t^\star, \lambda_t^\star, y) \mathrm{d}\mu(y) - \alpha \nabla U(\mathbf{X}_t^\star)\} \mathrm{d}t + \sqrt{2\alpha} \mathrm{d}\mathbf{B}_t \qquad \mathbf{X}_0^\star \in \mathbb{R}^d, \tag{10}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion, $\lambda_t^\star$ is the distribution of $(\mathbf{X}_t^\star)_{t\geq 0}$ and for any $\nu \in \mathscr{P}(\mathbb{R}^d)$ and $(x,y) \in \mathbb{R}^d \times \mathbb{R}^p$

$$\tilde{b}_\eta(x, \nu, y) = \nabla_1 \mathrm{k}(x,y)/(\nu[\mathrm{k}(\cdot,y)] + \eta). \tag{11}$$

In particular, note that $\tilde{b}_\eta$ is always defined since $\eta > 0$. In the following section, we establish various theoretical properties of (10): we show that its solution exists, is unique and converges towards the minimizer of $\mathscr{G}_\alpha^\eta$.

3.2. Existence, uniqueness and long-time behaviour

We are now concerned with the theoretical properties of (10). In particular, we show that under A1 the MKVSDE in (10) admits a unique strong solution and is ergodic provided that the gradient of the potential U is Lipschitz continuous. We present here several assumptions on the regularity of U which we use throughout the remainder of this paper. However, the results presented in this section only rely on A3-(a) and A3-(b).

A3. The following hold:

(a) There exists $\mathtt{L} \geq 0$ such that $\|\nabla U(x_1) - \nabla U(x_2)\| \leq \mathtt{L} \|x_1 - x_2\|$, for any $x_1, x_2 \in \mathbb{R}^d$.
(b) There exist $\mathtt{m}, \mathtt{c} > 0$ such that for any $x_1, x_2 \in \mathbb{R}^d$, $\langle \nabla U(x_1) - \nabla U(x_2), x_1 - x_2 \rangle \geq \mathtt{m} \|x_1 - x_2\|^2 - \mathtt{c}$.
(c) There exists $\mathtt{L}_2 \geq 0$ such that for any $x_1, x_2 \in \mathbb{R}^d$, $\left\|\nabla^2 U(x_1) - \nabla^2 U(x_2)\right\| \leq \mathtt{L}_2 \|x_1 - x_2\|$.

Under these conditions, if $\eta > 0$ the drift given in (11) is Lipschitz continuous and we can use standard tools for McKean–Vlasov processes to establish existence and uniqueness.

Proposition 5. *Assume A1 and A3-(a). Then for any $\alpha, \eta > 0$ there exists a unique strong solution to (10) for any initial condition $\mathbf{X}_0^\star \in \mathscr{P}_1(\mathbb{R}^d)$.*

Proof. See Appendix B.2. □

The previous proposition is limited to the case where $\eta > 0$. Indeed, if $\eta = 0$ the drift is *not* Lipschitz continuous and the SDE (10) might be unstable, with solutions existing up to a (possibly small) explosion time. Having shown that (10) admits a unique strong solution when $\eta > 0$, we verify that the distribution $\lambda_t^\star$ converges to the unique minimizer $\pi_{\alpha,\eta}^\star$ of $\mathscr{G}_\alpha^\eta$ when $t \to +\infty$.

Proposition 6. *Assume A1, A3-(a) and A3-(b). Then for any $\alpha, \eta > 0$, we have $\lim_{t\to+\infty} \mathbf{W}_2(\lambda_t^\star, \pi_{\alpha,\eta}^\star) = 0$.*

Proof. See Appendix B.3. □

There exists a rich literature on the exponential ergodicity of MKVSDE, see [12,16,36,62] for instance. However, these results require the interaction term $\tilde{b}_\eta$ to be small when compared to $\alpha\nabla U$. This is not the case in our applications where we study the behaviour of the McKean–Vlasov process for small values of $\alpha > 0$. The original approach of [46] differs from these previous works and relies on the fact the MKVSDE under consideration stems from a Wasserstein gradient flow. However, [46, Theorem 2.11] does not establish quantitative bounds contrary to [12,16,36,62] which derive explicit convergence rates. Under additional assumptions on the kernel k (e.g. strong convexity), the result above can be strengthened to show that, if $\alpha_t \to 0$ at the appropriate rate, then $\mathscr{G}^\eta(\lambda_t^\star)$ converges towards the minimum of the unregularized functional [22, Theorem 4.1].

3.3. Particle system

The MKVSDE (10) has several shortcomings from a methodological point of view. First, since the drift of the SDE depends not only on the current position of the solution but also on its distribution, it is not possible to sample from (10) even in simplistic settings. Second, as is the case for classical SDEs, it is usually difficult to sample from continuous-time processes. We circumvent the first issue by introducing a particle system [13,64] $(\mathbf{X}_t^{1:N})_{t\geq 0} = \{(\mathbf{X}_t^{k,N})_{t\geq 0}\}_{k=1}^N$ for any $N \in \mathbb{N}^\star$ which satisfies a classical SDE. In the limit $N \to +\infty$ we have that $(\mathbf{X}_t^{1,N})_{t\geq 0}$ approximates $(\mathbf{X}_t^\star)_{t\geq 0}$. In Section 4 we will investigate a time discretization scheme for this particle system SDE in order to obtain a particle system Markov chain which can be implemented.

We introduce the particle system $(\mathbf{X}_t^{1:N})_{t\geq 0}$ which satisfies the following SDE: for any $k \in \{1, \dots, N\}$, $\mathbf{X}_0^{k,N} \in \mathbb{R}^d$ and

$$\mathbf{X}_t^{k,N} = \mathbf{X}_0^{k,N} + \int_0^t b(\mathbf{X}_s^{k,N}, \lambda_s^N(\mathbf{X}_0^{1:N}))\mathrm{d}s + \sqrt{2\alpha}\mathbf{B}_t^k, \tag{12}$$

where for any $x \in \mathbb{R}^d$ and $\nu \in \mathscr{P}(\mathbb{R}^d)$ we have

$$b(x, \nu) = \int_{\mathbb{R}^p} \tilde{b}_\eta(x, \nu, y)\mathrm{d}\mu(y) - \alpha\nabla U(x), \tag{13}$$

$\{(\mathbf{B}_t^k)_{t\geq 0} : k \in \mathbb{N}\}$ is a family of independent Brownian motions and $\lambda_t^N(\mathbf{X}_0^{1:N})$ is the empirical measure associated with $\{(\mathbf{X}_t^{k,N})_{t\geq 0}\}_{k=1}^N$ with starting point $\mathbf{X}_0^{1:N}$, *i.e.* we have for any $N \in \mathbb{N}^\star$

$$\lambda_t^N(\mathbf{X}_0^{1:N}) = (1/N)\sum_{k=1}^N \delta_{(\mathbf{X}_t^{k,N})}.$$

In the following proposition we show using classical propagation of chaos tools that there exists a unique strong solution to (12) and that this solution approximates (10) for any finite time horizon.

Proposition 7. *Assume A1 and A3-(a). Then, for any $\alpha, \eta > 0$ and $N \in \mathbb{N}^\star$ there exists a unique strong solution to (12) for any initial condition $\mathbf{X}_0^{1:N}$ such that $\mathcal{L}(\mathbf{X}_0^{1:N}) \in \mathscr{P}_1((\mathbb{R}^d)^N)$ and $\{\mathbf{X}_0^{k,N}\}_{k=1}^N$ is exchangeable. In addition, if $\mathbf{X}_0^{\ell,N} = \mathbf{X}_0^\star$, for any $T \geq 0$ there exists $C_T \geq 0$ such that for any $N \in \mathbb{N}^\star$ and $\ell \in \{1, \dots, N\}$*

$$\mathbb{E}^{1/2}[\sup_{t\in[0,T]} \|\mathbf{X}_t^\star - \mathbf{X}_t^{\ell,N}\|^2] \leq C_T N^{-1/2}.$$

Proof. See Appendix B.4. □

We now turn to the exponential ergodicity of the particle system. Using recent advances in the theory of convergence of Markov chains [30] we derive quantitative bounds. However, these bounds are not uniform w.r.t to the number of particles and therefore we cannot use this result to conclude that the McKean–Vlasov process itself is exponentially ergodic by letting $N \to +\infty$.

Proposition 8. *Assume A1, A3-(a) and A3-(b). Then, for any $\alpha, \eta > 0$ and $N \in \mathbb{N}^\star$ there exist $C_N \geq 0$ and $\rho_N \in [0, 1)$ such that for any $x_1^{1:N}, x_2^{1:N} \in (\mathbb{R}^d)^N$ and $t \geq 0$*

$$\mathbf{W}_1(\lambda_t^N(x_1^{1:N}), \lambda_t^N(x_2^{1:N})) \leq C_N \rho_N^t \|x_1^{1:N} - x_2^{1:N}\|,$$

where for any $x^{1:N} \in (\mathbb{R}^d)^N$, $\lambda_t^N(x^{1:N})$ is the distribution of $(\mathbf{X}_t^{1:N})_{t\geq 0}$ with initial condition $x^{1:N}$. In particular, (12) admits a unique invariant probability measure denoted $\pi^N \in \mathscr{P}_p((\mathbb{R}^d)^N)$ and for any $x^{1:N} \in (\mathbb{R}^d)^N$ and $t \geq 0$

$$\mathbf{W}_1(\lambda_t^N(x^{1:N}), \pi^N) \leq C_N \rho_N^t \left(\|x^{1:N}\| + \int_{\mathbb{R}^d} \|\tilde{x}\|\mathrm{d}\pi^N(\tilde{x})\right).$$

Proof. See Appendix B.5 □

In the two previous propositions, the exponential ergodicity and the approximation between the particle system and the McKean–Vlasov process are only valid for a finite number of particles or a finite time-horizon, respectively. In particular, we cannot directly conclude that the projection of the invariant distribution of the particle system onto its first component on $\mathscr{P}(\mathbb{R}^d)$, denoted by $\pi^{1,N}$, is close to $\pi_{\alpha,\eta}^\star$. This is the topic of the next proposition.

Proposition 9. *Assume A1, A3-(a) and A3-(b). Then, for any $\alpha, \eta > 0$, $\lim_{N\to+\infty} \mathbf{W}_1(\pi^{1,N}, \pi^\star_{\alpha,\eta}) = 0$. In addition, $\pi^\star_{\alpha,\eta}$ is the unique invariant probability measure of (10).*

Proof. See Appendix B.6. □

4. Numerical implementation

In the previous section, we introduced a continuous time Markov process $(\mathbf{X}_t^{1:N})_{t\geq 0}$ such that the distribution $\mathbf{X}_t^{1,N}$ approximates $\pi^\star_{\alpha,\eta}$ for large values of $t \geq 0$ and $N \in \mathbb{N}$. In this section, we derive an algorithm approximating this particle system.

In order to obtain a numerical approximation of (12) we need two consider two factors: first, as in the case of classical SDEs, the continuous time process (12) cannot be simulated directly, and we need to introduce time discretization, secondly, in practice the integral w.r.t. μ in (13) is intractable.

First, in Section 4.1 we consider an Euler–Maruyama discretization of the particle system. For stability issues we consider a tamed version of the classical Euler–Maruyama discretization. To address the second issue, given that in applications we will often be in the situation in which μ is only known through a sample, we approximate the integral w.r.t. μ in (13) with a sample average In Section 4.2, we introduce a second discrete-time particle system, obtained by approximating μ with its empirical average $(1/M)\sum_{k=1}^M \delta_{y^{k,M}}$ where $y^{1:M} = \{y^{k,M}\}_{k=1}^M$ is a sample from $\mu^{\otimes M}$. We show that the two particle systems are close for large values of M. In particular, Proposition 11 ensures the quantitative stability of the particle system. Finally, in Section 4.3 we discuss the choice of parameters for the algorithm.

4.1. Tamed Euler–Maruyama discretization

In order to obtain a Markov chain which can be implemented, we consider a time discretization of (12). More precisely, for any $N \in \mathbb{N}$, we consider the following tamed Euler–Maruyama discretization given by $X_0^{1:N} \in (\mathbb{R}^d)^N$ and for any $n \in \mathbb{N}$ and $k \in \{1, \dots, N\}$

$$X_{n+1}^{k,N} = X_n^{k,N} + \gamma b(X_n^{k,N}, \lambda_n^N)/(1 + \gamma\|b(X_n^{k,N}, \lambda_n^N)\|) + \sqrt{2\alpha}Z_{n+1}^k, \tag{14}$$

where $\{Z_n^k\}_{k,n\in\mathbb{N}}$ is a family of independent Gaussian random variables, $\gamma > 0$ is a stepsize, b is given by (13) and for any $n \in \mathbb{N}$, we have that $\lambda_n^N = (1/N)\sum_{k=1}^N \delta_{X_n^{k,N}}$. The choice of a tamed scheme is motivated by the behaviour of the drift (13). As shown in the Appendix, b is Lipschitz continuous for any $\eta > 0$, however, for small values of η the Lipschitz constant can be large. In practice, we observe that the use of a tamed Euler–Maruyama discretization scheme prevents some numerical stability issues appearing for small values of $\eta > 0$.

The tamed Euler–Maruyama discretization (and its extension to the Milstein scheme) for classical SDEs has been investigated in [14,47,48,92]. More recently, strong approximation results have been established for McKean–Vlasov SDEs in [8]. In the following proposition, we use the results of [8, Theorem 2.5] to derive strong approximation results between $(X_n^{k,N})_{n\in\mathbb{N}}$ and $(\mathbf{X}_t^{k,N})_{t\geq 0}$ for any $k \in \{1, \dots, N\}$; since the diffusion coefficient of (12) is constant, the tamed scheme (14) coincides with a Milstein scheme, allowing us to show strong convergence of order 1.

Proposition 10. *Assume A1 and A3. Then for any $\eta, \alpha > 0$ and any $T \geq 0$ there exists $C_T \geq 0$ such that for any $N \in \mathbb{N}^\star$, $\ell \in \{1, \dots, N\}$ and $\gamma > 0$*

$$\mathbb{E}[\sup_{n\in\{0,\dots,n_T\}} \|\mathbf{X}_{n\gamma}^{\ell,N} - X_n^{\ell,N}\|] \leq C_T\gamma$$

where $n_T = \lfloor T/\gamma \rfloor$.

Proof. See Appendix C.2. □

A similar result holds for a continuous-time interpolation of $(X_n^{\ell,N})_{n\in\mathbb{N}}$, see [8, Theorem 2.5] for more details. For small values of $\gamma > 0$ and large values of $n\gamma$ and $N \in \mathbb{N}$ we get that $(1/N)\sum_{k=1}^N \delta_{X_n^{k,N}}$ is an approximation of $\pi^\star_{\alpha,\eta}$.

4.2. Stability of the particle system w.r.t the observed measure

Proposition 3 shows that if μ is approximated by the empirical measure μ^M obtained from a sample $y^{1:M} = \{y^{k,M}\}_{k=1}^M$ then the minimizer of $\mathscr{G}_\alpha^\eta$ obtained using μ^M instead of μ converges to the minimizer of $\mathscr{G}_\alpha^\eta$ obtained using μ as $M \to +\infty$. We now consider the discrete-time alternative particle system $\{(X_n^{k,N,M})_{n\in\mathbb{N}}\}_{k=1}^N$ which evolves according to (14) with b replaced by its empirical approximation obtained by approximating μ with the empirical measure of the M samples $y_1, \dots, y_M$:

$$X_{n+1}^{k,N,M} = X_n^{k,N,M} + \gamma b^M(X_n^{k,N,M}, \lambda_n^{N,M})/(1 + \gamma\|b^M(X_n^{k,N,M}, \lambda_n^{N,M})\|) + \sqrt{2\alpha}Z_{n+1}^k \tag{15}$$

where $\{Z_n^k\}_{k,n\in\mathbb{N}}$ is a family of independent Gaussian random variables, $\gamma > 0$ is a stepsize, $\lambda_n^{N,M} = (1/N)\sum_{k=1}^N \delta_{X_n^{k,N,M}}$ for any $n \in \mathbb{N}$ and for any $x \in \mathbb{R}^d$ and $\nu \in \mathscr{P}(\mathbb{R}^d)$ we have

$$b^M(x,\nu) = \int_{\mathbb{R}^p} \tilde{b}_\eta(x,\nu,y)\mathrm{d}\mu^M(y) - \alpha\nabla U(x) = (1/M)\sum_{j=1}^M \tilde{b}_\eta(x,\nu,y_j) - \alpha\nabla U(x). \tag{16}$$

In the next proposition, we show that the two particle systems (14) and (15) are close when M is large.

Proposition 11. *Assume A1, A3-(a) and A3-(b). Let $\eta, \alpha > 0$. For any $N, M \in \mathbb{N}^\star$, let $(X_n^{1,N})_{t\geq 0}$ and $(X_n^{1,N,M})_{t\geq 0}$ be obtained with (14) and (15) respectively, driven by the same underlying family of independent Gaussian random variables $\{Z_n^k\}_{k,n\in\mathbb{N}}$, with initial condition $X_0^{1:N}$ such that $\mathcal{L}(X_0^{1:N}) \in \mathscr{P}_1((\mathbb{R}^d)^N)$, $\{X_0^{k,N}\}_{k=1}^N$ is exchangeable and $\{X_0^{k,N}\}_{k=1}^N = \{X_0^{k,N,M}\}_{k=1}^N$. Then, for any $T \geq 0$ there exists $C_T \geq 0$ such that for any $\ell \in \{1, \dots, N\}$ and $\gamma > 0$*

$$\mathbb{E}[\sup_{n\in\{0,\dots,n_T\}} \|X_n^{\ell,N} - X_n^{\ell,N,M}\|] \leq C_T \gamma M^{-1/2}.$$

Proof. See Appendix C.3. □

An equivalent result showing convergence of the continuous process

$$\mathbf{X}_t^{k,N} = \mathbf{X}_0^{k,N} + \int_0^t b^M(\mathbf{X}_s^{k,N}, \lambda_s^N(\mathbf{X}_0^{1:N}))\mathrm{d}s + \sqrt{2\alpha}\mathbf{B}_t^k,$$

with b^M as in (16), to (12) could also be obtained. The argument follows the same line of that used to prove Proposition 7 and exploits the convergence of b^M to b established in the proof of Proposition 11. The resulting rate is $M^{-1/2}$ (see Appendix C.5).

As we discuss in the next section, in some cases it might be beneficial to resample from the empirical measure μ^M at each time step n. In the following proposition we show that this resampling operation does not change the error bound obtained in Proposition 11.

Proposition 12. *Assume A1, A3-(a) and A3-(b). Let $\eta, \alpha > 0$. For any $N, M \in \mathbb{N}^\star$, let $(X_n^{1,N})_{t\geq 0}$ be obtained with (14) and $(X_n^{1,N,m})_{t\geq 0}$ with (15) where at each time n an i.i.d. sample of size m is drawn from μ^M. Assume that $(X_n^{1,N})_{t\geq 0}$ and $(X_n^{1,N,m})_{t\geq 0}$ are driven by the same underlying family of independent Gaussian random variables $\{Z_n^k\}_{k,n\in\mathbb{N}}$, with initial condition $X_0^{1:N}$ such that $\mathcal{L}(X_0^{1:N}) \in \mathscr{P}_1((\mathbb{R}^d)^N)$, $\{X_0^{k,N}\}_{k=1}^N$ is exchangeable and $\{X_0^{k,N}\}_{k=1}^N = \{X_0^{k,N,m}\}_{k=1}^N$. Then, for any $T \geq 0$ there exists $C_T \geq 0$ such that for any $\ell \in \{1, \dots, N\}$ and $\gamma > 0$*

$$\mathbb{E}[\sup_{n\in\{0,\dots,n_T\}} \|X_n^{\ell,N} - X_n^{\ell,N,m}\|] \leq C\gamma m^{-1/2}.$$

Proof. See Appendix C.4. □

The empirical measure $\lambda_n^{N,M}$ provides an approximation of $\pi_{\alpha,\eta}^\star$. In low dimensional settings, for the purpose of visualization we derive a smooth approximation of $\pi_{\alpha,\eta}^\star$ using standard kernel density estimation tools and define $\hat{\pi}_n^{N,M} : \mathbb{R}^d \to \mathbb{R}$ such that for any $x \in \mathbb{R}^d$

$$\hat{\pi}_n^{N,M}(x) = (1/N)\sum_{k=1}^N \det(\mathrm{H})^{-1/2}\varphi(\mathrm{H}^{-1/2}(x - X_n^{k,N,M})), \tag{17}$$

where φ is the density of a d-dimensional Gaussian distribution with zero mean and identity covariance matrix and H is a positive definite bandwidth matrix, see, e.g., [83]. Combining the results in Propositions 7 and 10 with standard arguments from the kernel density estimation literature, the estimator (17) can be shown to converge to $\pi_{\alpha,\eta}^\star$ as $N \to \infty$ and $\gamma \to 0$ (e.g., [4, Theorem 3.1] and [13, Theorem 2.2]). Our final algorithm is summarized in Algorithm 1.

Algorithm 1 Solving Fredholm integral equations with Wasserstein gradient flows (FE-WGF)

Require: $N, M, n_T \in \mathbb{N}$, $\alpha, \eta, \gamma > 0$, $\mathrm{H} \in \mathbb{R}^{d\times d}$, $\mu, \pi_0, \pi_{\text{init}} \in \mathscr{P}(\mathbb{R}^d)$.
Draw $\{X_0^{k,N,M}\}_{k=1}^N$ from $\pi_{\text{init}}^{\otimes N}$
for $n = 0 : n_T$ **do**
  Draw $\{y^{k,M}\}_{k=1}^M$ from $\mu^{\otimes M}$
  Compute $b_M(X_n^{k,N,M}, \lambda_n^{N,M})$ as in (16)
  Update $X_{n+1}^{k,N,M} = X_n^{k,N,M} + \gamma b^M(X_n^{k,N,M}, \lambda_n^{N,M})/(1 + \gamma\|b^M(X_n^{k,N,M}, \lambda_n^{N,M})\|) + \sqrt{2\alpha}Z_{n+1}$ as in (14)
end for
return $\hat{\pi}_n^{N,M}(x)$ as in (17)

4.3. Implementation guidelines

Obtaining the estimator (17) of the minimizer of $\mathscr{G}_\alpha^\eta$ requires specification of a number of parameters: the reference measure π_0 and the regularization parameter α control the properties of the measure π we want to reconstruct (e.g. its smoothness), while the parameter η, the number of particles N, M and the time discretization step γ control accuracy and stability of the numerical implementation.

Choice of π_0. There are of course, many possible choices for the reference measure π_0 to which we want the regularized solution to be close depending on the constraints we want to impose on the reconstructed solution. The choice of an improper reference measure $\pi_0 \propto C$ with $C > 0$ results in a McKean–Vlasov SDE in which there is no dependence on π_0

$$\mathrm{d}\mathbf{X}_t^\star = \{\int_{\mathbb{R}^p} \tilde{b}_\eta(\mathbf{X}_t^\star, \lambda_t^\star, y)\mathrm{d}\mu(y)\}\mathrm{d}t + \sqrt{2\alpha}\mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0^\star \in \mathbb{R}^d,$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $\lambda_t^\star$ is the distribution of $(\mathbf{X}_t^\star)_{t\geq 0}$. This scheme corresponds to what we would obtain by applying the gradient flow procedure to the functional (6). Under additional assumptions on the tail behaviour of the kernel k, it is possible to show that this functional is proper, lower semi-continuous and coercive in an appropriate sense; [25, Chapter 7] shows empirically that the corresponding particle system is stable. Nevertheless, in many scenarios some knowledge of the target will be embedded in the problem at hand. In deconvolution problems, μ is the distribution of noisy samples from the target distribution π; in this case it is sensible to expect that π itself will be close to μ and therefore choose π_0 to be reasonably close to μ. More generally, for problems in which $\mathrm{k}(x, y) = \mathrm{k}(y - x)$ is a deconvolution kernel, the distribution μ will be the convolution of π and k and therefore encapsulates information on π. In these problems we choose π_0 to be a Gaussian distribution with mean and variance given by the empirical mean and variance of the sample from μ. For the epidemiology examples in Section 5.2, we set π_0 to be μ shifted back by a number of days equal to the mean [39] or the mode [93] of the infection-to-confirmed delay distribution k. For image reconstruction, π_0 could be selected to enforce particular characteristics in the reconstructed image, e.g. smoothness [71] or sparsity [60,97]. Alternatively, if a large data set of previously reconstructed images is available one could use score matching priors [54] to obtain a reference measure π_0 encapsulating all the relevant features of the image to reconstruct. We study the influence of the reference measure π_0 on a toy example in Appendix D.1 and find that π_0 influences the shape of the minimizer of $\mathscr{G}_\alpha^\eta$ and its minimum but not the speed at which convergence to the minimum occurs.

Choice of α. As the parameter $\alpha > 0$ controls the amount of regularization introduced by the cross-entropy penalty, its value should be chosen to give a good trade-off between the distance from the data distribution μ, $\mathrm{KL}(\mu|\pi\mathrm{K})$, and that from π_0, $\mathrm{KL}(\pi|\pi_0)$. A common approach for selecting the regularization parameter in Tikhonov regularization is cross-validation [1,91]. 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 sample from μ into L subsets and find the value of α which minimizes

$$\mathrm{CV}(\alpha) = (1/L)\sum_{j=1}^{L} \hat{\mathscr{G}}_\alpha^\eta(\hat{\pi}^j),$$

where $\hat{\pi}^j$ is obtained using Algorithm 1 with the samples from μ which are not in group $j \in \{1, \dots, L\}$ and $\hat{\mathscr{G}}_\alpha^\eta$ is an approximation of $\mathscr{G}_\alpha^\eta$

$$\hat{\mathscr{G}}_\alpha^\eta(\hat{\pi}^j) := -\frac{1}{M}\sum_{\ell=1}^{M} \log\left(\frac{1}{N}\sum_{k=1}^{N} \mathrm{k}(X_n^{k,N,M}, y_j)\right) + \frac{\alpha}{N}\sum_{k=1}^{N} \log\left(\frac{\hat{\pi}^j(X_n^{k,N,M})}{\pi_0(X_n^{k,N,M})}\right). \tag{18}$$

If some prior information on the smoothness of the solution π is known (e.g. its variance), one could chose α so that the smoothness of (17), matches the expected smoothness of π. We show how the value of α influences the reconstructions of π on simple problems in Appendix D.2.

Choice of η. The parameter η has been introduced in (5) to deal with the possible instability of the functional $\mathscr{G}_\alpha^\eta$; we did not find performances to be significantly influenced by this parameter as long as its value is sufficiently small. In practice, in the experiments in Section 5 we set $\eta \equiv 0$. This choice might seem counter-intuitive, since most of our theoretical results are obtained with $\eta > 0$, On one hand, this suggests that for regular enough k, the functional $\mathscr{G}_\alpha$ is well-behaved; on the other hand, we observe that the introduction of the tamed Euler–Maruyama scheme (14) prevents most of the numerical instability issues arising when discretizing (12).

Choice of N, γ and m. The values of the number of particles N, the time discretization step γ and the number of samples m from μ to use at each iteration control the quality of the numerical approximation of (10), their choice is therefore largely application dependent. However, the results in Propositions 7, 10 and 12 give the following global error estimate: for any $T \geq 0$ we have

$$\mathbb{E}[\sup_{n\in\{0,\dots,n_T\}} \|\mathbf{X}_{n\gamma}^\star - X_n^{\ell,N,m}\|] \leq C_T(N^{-1/2} + \gamma + \gamma m^{-1/2}). \tag{19}$$

Choosing N amounts to the classical task of selecting an appropriate sample size for Monte Carlo approximations, while the choice of γ corresponds to the specification of a timescale on which to discretize a continuous time (stochastic) process; hence, one can exploit the vast literature on Monte Carlo methods and discrete time approximations of SDEs to select these values [55]. In practice, we found that the choice of N largely depends on the dimensionality of the problem; for one-dimensional examples values between $N = 200$ and $N = 1000$ achieve high enough accuracy to compete with specialized algorithms (see Sections 5.1, 5.2); as the dimension d increases larger values of N should be considered (see Section 5.3). Similar considerations apply to the choice of the time discretization step γ. In particular, this parameter 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 discretization error. In practice, we observed good results with γ between 10^{-1} and 10^{-3}.

The number of samples m from μ to employ at each iteration is again problem dependent. In many cases, we are given a number M of observations from μ, if M is very large, the resulting algorithm might become computationally too expensive. In this setting,

we propose the following batch strategy to approximate μ: fix N and 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 (12) and the total number of samples available from μ; if N is larger than the total number of observed 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 $\mathcal{O}(N^2)$ per time step.

In cases in which we have access to a sampling mechanism from μ, and we can therefore draw m samples at each iteration, we simply set $m = N$.

Choice of n_T. It is straightforward to choose the number of time steps n_T, adaptively. The number of steps necessary to give convergence of (12) to its stationary distribution is estimated by approximating the value of $\mathscr{G}_\alpha^\eta$ through numerical integration as in (18) and once the value of $\mathscr{G}_\alpha^\eta$ stops decreasing, a minimizer has apparently been reached and the iteration in Algorithm 1 can be stopped.

5. Experiments

We test FE-WGF on a number of examples of Fredholm integral equations (1) and compare our density estimator (17) with the reconstructions given by state-of-the-art methods for each experiment.

First, in Section 5.1 we consider a classical problem in statistics, density deconvolution. In this case, $\mathrm{k}(x, y) = \mathrm{k}(y - x)$ and (1) models the task of reconstructing the density of a random variable X from noisy measurements Y. In Section 5.2 we test FE-WGF on a problem drawn from epidemiology. More precisely, we aim at reconstructing the incidence profile of a disease from the observed number of cases. This is a particular instance of the deconvolution problem in which a smoothness constraint on the solution π of (1) is necessary to make the estimator robust to noise in the observations Y [69]. In Section 5.3, we consider an application to computed tomography (CT), in which a cross-sectional image of the lungs needs to be recovered from the radial measurements provided by the CT scanner. Finally, in Section 5.4, we study how the performances of the proposed method scale with the dimensionality of the support of π on a toy model.

In most examples, μ is known through a sample and at each iteration we resample m times from its empirical distribution. For image restoration problems we consider the observed distorted image as the empirical distribution of a sample from μ and use it to draw m samples at each iteration in Algorithm 1. In order to implement FE-WGF we follow the guidelines provided in Section 4.3. Julia code to reproduce all examples is available online at https://github.com/FrancescaCrucinio/FE_WGF.

5.1. Density deconvolution

The focus of this section is on deconvolution problems, *i.e.* those in which μ and π satisfy (1) with $\mathrm{k}(x, y) = \mathrm{k}(y-x)$. In particular, we consider the case in which we want to recover the density of a random variable X from observations of $Y = X + Z$ where Z is a zero mean error variable independent of X. The deconvolution problem is widely studied, and recently a number of approaches based on normalizing flows [34], deep learning [3], and generative adversarial networks [29] have been explored.

We consider the Gaussian mixture model of [61] (see also [26]) where, with a slight abuse of notation, we denote both a measure and its density w.r.t. the Lebesgue measure with the same symbol:

$$\begin{aligned}
\pi(x) &= (1/3)\mathcal{N}(x; 0.3, 0.015^2) + (2/3)\mathcal{N}(x; 0.5, 0.043^2),\\
\mathrm{k}(x, y) &= \mathcal{N}(y; x, 0.045^2),\\
\mu(y) &= (1/3)\mathcal{N}(y; 0.3, 0.045^2 + 0.015^2) + (2/3)\mathcal{N}(y; 0.5, 0.045^2 + 0.043^2),
\end{aligned}$$

and compare our method with the deconvolution kernel density estimators with plug-in bandwidth (DKDEpi) [32], a class of estimators for deconvolution problems which given a sample from μ produce a kernel density estimator for π, and SMC-EMS, a particle method to solve (1) which achieves state-of-the-art performance.

We check the accuracy of the reconstructions through the integrated square error

$$\mathrm{ISE}(\hat{\pi}) = \int_{\mathbb{R}} \{\pi(x) - \hat{\pi}(x)\}^2 \mathrm{d}x, \tag{20}$$

with $\hat{\pi}$ an estimator of π. Even if the analytic form of μ is known, for SMC-EMS and (17) we draw a sample of size $M = 10^3$ and draw without replacement $m = \min(N, M)$ samples at each iteration; for DKDE we draw a sample of size N for each N.

Both SMC-EMS and the estimator (17) require specification of a number of parameters: we fix the number of time steps for both algorithms to $n_T = 100$ as we observed that convergence occurs in less than 100 iterations, the initial distribution is that of the samples from Y, since we expect the distribution of Y to be close to that of X. The regularization parameter α for FE-WGF and the smoothing parameter for SMC-EMS are chosen through cross validation, using (5) as target functional for the former and the Kullback–Leibler divergence (2) for the latter. To implement FE-WGF we also need to specify the reference measure π_0, a Gaussian with mean and variance given by the empirical mean and variance of Y, and the time step $\gamma = 10^{-3}$.

We consider different particle sizes (from 10^2 to 10^4) and compare reconstruction accuracy and total runtime of SMC-EMS, FE-WGF and DKDE (Fig. 1). In particular, we compare the cost per iteration since we run both algorithms for a fixed number of steps, however, we found that SMC-EMS and FE-WGF converge in a similar number of steps for this example. The computational cost could be reduced by considering stopping criteria using approximations of the functional $\mathscr{G}_\alpha^\eta$. The performances of FE-WGF and SMC-EMS

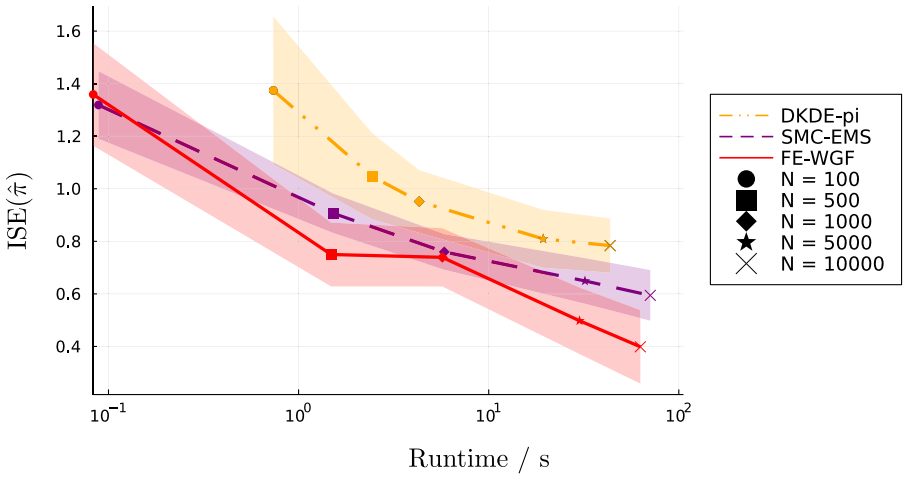

Fig. 1. Average accuracy and runtime for FE-WGF, SMC-EMS and DKDE with number of particles N ranging between 10^2 and 10^4. The shaded regions represent a interval of two standard deviations over 100 repetitions centred at the average ISE.

are similar both in terms of runtime and in terms of accuracy, with FE-WGF generally giving more accurate reconstructions for the same computational cost.

The DKDE-pi estimator has runtime comparable to that of SMC-EMS and FE-WGF, but poorer accuracy. Despite being better than SMC-EMS on average, the accuracy of FE-WGF has a slightly higher variance (about twice that of SMC-EMS), while DKDE-pi has still higher variance (about three times that of SMC-EMS).

As both SMC-EMS and FE-WGF are regularized versions of the inconsistent maximum likelihood estimator for π, it is natural to compare the smoothness of the reconstructions. To characterize smoothness, we take 100 points x_c in the support of π and approximate (with 100 replicates) the mean squared error

$$\mathrm{MSE}(x_c) = \mathbb{E}\left[\left(\pi(x_c) - \hat{\pi}(x_c)\right)^2\right].$$

The MSE measures locally the fit of the reconstructions to the solution π, which is known to be smooth, providing information on the (relative) smoothness of the reconstructions.

The distribution of the MSE over the 100 points (Fig. 2) shows that SMC-EMS and FE-WGF can achieve considerably lower MSE than DKDE-pi, and therefore that the smoothness of the reconstructions provided by these two methods is closer to that of the true density π.

5.2. Epidemiology

In epidemiology, Fredholm integral equations link the unknown infection incidence curve π to the number of reported cases, deaths or hospitalizations. This is a particular instance of the deconvolution problem, in which the kernel $\mathrm{k}(x, y) = \mathrm{k}(y - x)$ describes the delay distribution between time of infection and time of death or hospitalization. Deconvolution techniques have been used to infer π in the case of HIV [9] and influenza [39] and have been recently applied to estimate the incidence curve of COVID-19 [21,63,69,93].

As an example, we take the spread of the pandemic influenza in the city of Philadelphia between the end of September and the beginning of October 1918 [39]. The count of daily deaths and the distribution of delay between infection and death are available through the R package `incidental` [69].

To obtain a parametric form for k we fit a mixture of Gaussians to the delay data using the expectation maximization algorithm (`normalmixEM` function in R; [10])

$$\mathrm{k}(x, y) = 0.595\mathcal{N}(y - x; 8.63, 2.56^2) + 0.405\mathcal{N}(y - x; 15.24, 5.39^2).$$

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 [74]) and also benefits from a bounded derivative and therefore satisfies A1 and A2.

We compare the reconstructions obtained using FE-WGF with the robust incidence deconvolution estimator (RIDE) of [69], the Richardson–Lucy (RL) deconvolution described in [39], an iterative algorithm popular in the image processing literature which aims at minimizing the Kullback–Leibler divergence (2), and SMC-EMS. To benchmark the performances of the four methods we consider synthetic data obtained simulating from the slow decay incidence curve in [69]

$$\pi(x) \propto \begin{cases} \exp(-0.05(8 - x)^2) & 0 \le x \le 8, \\ \exp(-0.001(x - 8)^2) & 8 \le x \le 100, \end{cases} \tag{21}$$

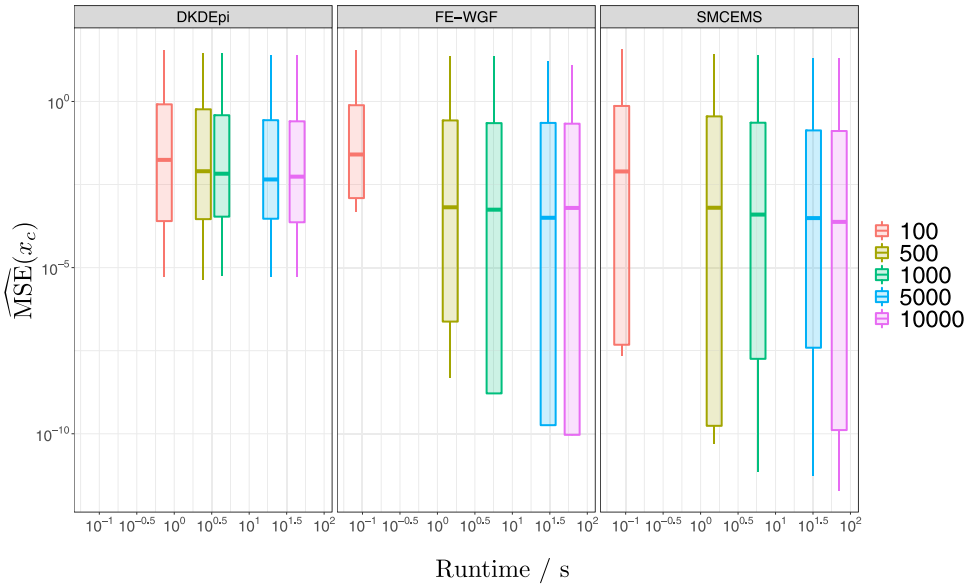

Fig. 2. Distribution of MSE as a function of runtime (in log seconds) for FE-WGF, SMC-EMS and DKDE. The number of particles N ranges between 100 and 10,000.

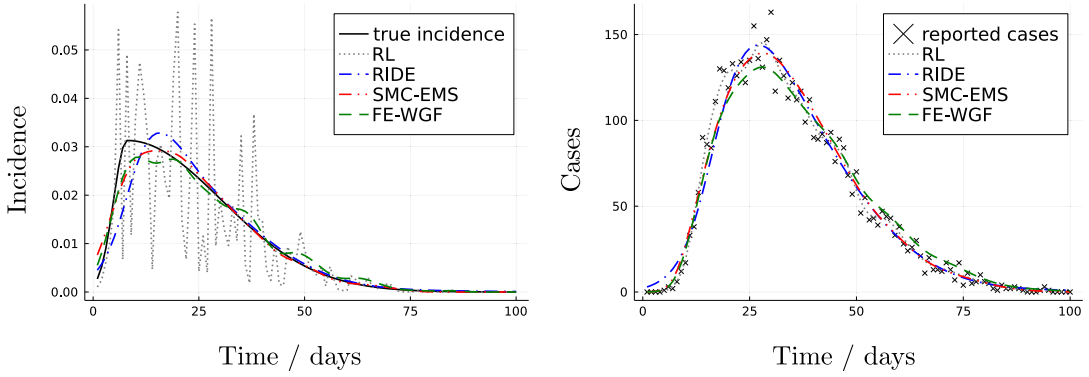

Fig. 3. Example fit of the reconstructions of the synthetic incidence curve (21) and of the corresponding reconstruction of the number of cases.

Table 1
Comparison of reconstructions of the synthetic incidence curve (21), the average reconstruction accuracy and fit to data over 100 repetitions are reported.

Method	Well-specified			Misspecified		
	ISE $(\hat{\pi})$	ISE (μ^{rec})	runtime (s)	ISE $(\hat{\pi})$	ISE (μ^{rec})	runtime (s)
RL	$7.7 \cdot 10^{-3}$	$\mathbf{1.5 \cdot 10^{-4}}$	**<1**	$7.9 \cdot 10^{-3}$	$\mathbf{1.5 \cdot 10^{-4}}$	**<1**
RIDE	$9.0 \cdot 10^{-4}$	$3.4 \cdot 10^{-4}$	58	$1.0 \cdot 10^{-3}$	$3.4 \cdot 10^{-4}$	58
SMC-EMS	$3.3 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	3	$3.7 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	3
FE-WGF	$\mathbf{2.7 \cdot 10^{-4}}$	$2.5 \cdot 10^{-4}$	96	$\mathbf{3.1 \cdot 10^{-4}}$	$2.5 \cdot 10^{-4}$	95

whose shape is similar to that of the incidence of the 1918 pandemic influenza, rescaled to have 5000 infections distributed over 100 time steps.

We consider both the case in which the model is well-specified, and each case on the incidence curve is randomly propagated forward s days according to k, and the case in which data are noisy and the delay model is misspecified: every sixth and seventh day a uniform random number between 0.3 and 0.5 is drawn and that proportion of cases are recorded two days later (e.g. cases from day six move to day eight). This approximates reporting delays for testing and death records [69].

Since the average delay between infection and death is 9 days [39], we set the reference measure π_0 to be a Gaussian distribution with mean equal to the mean of μ (≈34.6 days) shifted backwards by 9 and variance equal to the variance of μ (≈228 days). Estimates of the unknown delay distribution are usually obtained by previous studies or from similar diseases (see, e.g. [9,21] and references

Fig. 4. Reconstruction of a lung CT scan via FE-WGF and FBP. FBP provides reconstructions which preserve sharp edges but present speckle noise, while the reconstructions obtained with FE-WGF are smooth but with blurry edges.

therein), in this case, influenza-like infections. Following similar considerations, we set the initial distribution to be the death curve shifted back by 9 days for both the RL algorithm, SMC-EMS and FE-WGF. For FE-WGF we take $N = 500$, $\gamma = 10^{-1}$ and iterate for 3000 time steps (necessary to achieve convergence).

As discussed in Section 4.3 we set $\eta = 0$ precision, we pick $\alpha = 1 \times 10^{-3}$ and the parameter $\varepsilon = 2 \times 10^{-4}$ for SMC-EMS by cross-validation over $L = 5$ dataset for which the delay distribution k is misspecified and use this value for both the well-specified and the misspecified case. To approximate μ we draw $m = N = 500$ samples without repetition from the $M = 5000$ observed cases at each iteration in Algorithm 1. The RIDE estimator does not require specification of any parameter while for RL and SMC-EMS we set the maximum number of iterations to 100.

The estimators' quality is evaluated using the integrated squared error (20) and the fit to the observed number of cases measured through the ISE between the observed cases and the reconstructed cases given by reconvolution of the estimates of π with k

$$\mu^{rec}(y) = \int_{\mathbb{R}} \mathrm{k}(y - x)\mathrm{d}\hat{\pi}(x).$$

The RL algorithm gives the best fit the observed death counts and has the lowest runtime (<1 second), however, the reconstruction of the incidence curve is not smooth and sensitive to noise [69]. RIDE, SMC-EMS and FE-WGF address the lack of smoothness introducing regularization and provide smoother reconstructions (Fig. 3) which are considerably closer to the incidence curve both in the well-specified and in the misspecified setting (Table 1) with FE-WGF achieving the best fit to π. The runtime of RIDE and FE-WGF are comparable, while SMC-EMS is cheaper as convergence is achieved in only 100 steps, and the reconstructions do not change considerably after that. This is not unexpected, as the reconstructions provided by RIDE, SMC-EMS and FE-WGF are a regularized version of the measure π minimizing the Kullback–Leibler divergence (2), while that given by RL minimizes (2) with no regularization and inherits the well-known inconsistency of the maximum likelihood estimator in the infinite dimensional setting [58].

Comparing SMC-EMS and FE-WGF, we find that the latter always achieves better accuracy, but the increase in accuracy is more pronounced for the misspecified model. In fact, even if SMC-EMS and FE-WGF both aim at minimizing a regularized Kullback–Leibler divergence (2), the latter minimizes (3) which allows us to impose that the reconstructions are close (in the Kullback–Leibler sense) to the reference measure π_0, which for this example is particularly informative, since the corresponding term in the drift (13) pushes the particles towards the mode of the distribution of the number of cases shifted back by 9 days. On the contrary, SMC-EMS does not allow us to exploit this information, and π_0 is only used to initialize the particle system for SMC-EMS.

5.3. Computed tomography

Fredholm integral equations of the first kind find wide application in medical imaging [94]; in this context they model the reconstruction of cross-section images of the organ of interest from the noisy measurements provided by positron emission tomography (PET) and CT scanners. CT scanners provide noisy measurements by mapping each point of the organ's cross-section π onto its radial projection μ defined for angles ϕ between 0 and 2π and depths $\xi \in \mathbb{R}$ with 0 corresponding to the centre of the scanner. The distribution of the radial projections μ is linked to the cross-section image of the organ of interest π through the Radon transform [78], *i.e.* for any $\phi \in [0, 2\pi]$ and $\xi \in \mathbb{R}$

$$\mu(\phi, \xi) = \int_{\mathbb{R}} \pi(\xi \cos\phi - t \sin\phi, \xi \sin\phi + t \cos\phi)\mathrm{d}t, \tag{22}$$

(with a slight abuse of notation we denote both a measure and its density w.r.t. the Lebesgue measure with the same symbol) where the right hand side is the line integral along the line with equation $x \cos\phi + y \sin\phi = \xi$. We model the alignment between the

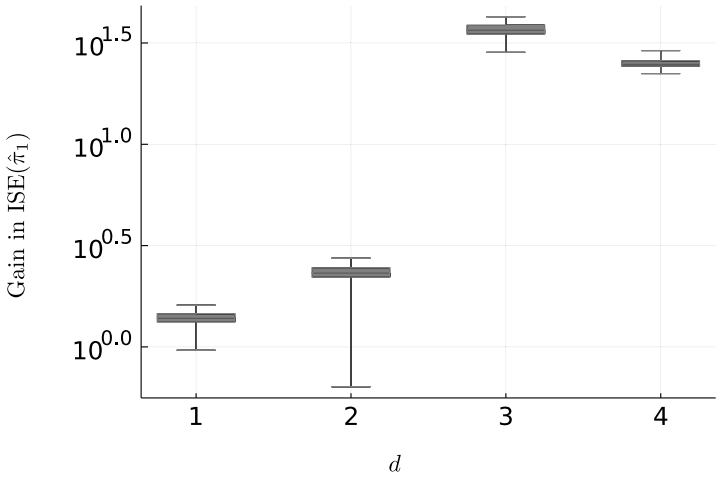

Fig. 5. Distribution of ISE ratios (ISE of OSL-EM divided by ISE of FE-WGF) over 100 repetitions. The runtime of FE-WGF is 1.3 times that of OSL-EM on average, while the average gain ranges from 1.3 for $d = 1$ to ≈ 30 for $d = 3, 4$.

projections onto (ϕ, ξ) and the corresponding location (x, y) in the reference image in (22) using a Gaussian distribution with small variance, renormalized to integrate to 1 for fixed $(x, y) \in \mathbb{R}^2$ and obtain the following Fredholm integral equation (1)

$$\mu(\phi, \xi) = \int_{\mathbb{R}^2} (2\pi\sigma^2)^{-1} \exp[(x\cos\phi + y\sin\phi - \xi)^2/(2\sigma^2)]\pi(x, y)\mathrm{d}x\mathrm{d}y,$$

We test FE-WGF on a 512×512 pixels lung CT scan from the LIDC-IDRI database [6] (in particular, scan LIDC_IDRI_0683_1_048). The data image is obtained using the ASTRA toolbox [90] by projecting the CT scan at 512 equally spaced angles ϕ in $[0, 2\pi]$ and at 729 depths, and then corrupting the obtained projections with Poisson noise. We compare the reconstructions given by FE-WGF with those obtained by filtered back projection (FBP), one of the most common methods for analytic image reconstruction [89], available in the ASTRA toolbox [90].

The reference measure is a Gaussian distribution with mean corresponding to the centre of the image and variance $s^2\,\mathrm{Id}$ with $s > 0$ large enough to approach a uniform distribution over the image (in our case $s^2 = 0.35^2$), this guarantees that A3 is satisfied. The number of particles $N = 10^4$ and the time step $\gamma = 10^{-3}$ are chosen to achieve high enough resolution, the regularization parameter $\alpha = 7 \times 10^{-3}$ is selected by cross validation over $L = 5$ replicates. Convergence is measured by numerically approximating $\mathscr{G}_\alpha^\eta$ and occurs in 200 steps.

The reconstructions provided by FBP are less robust to noise in the data, which reflects in a higher ISE, on the contrary FE-WGF is more robust to noise and achieves lower ISE (see Fig. 4). However, the reconstructions obtained with FE-WGF rarely result in sharp edges due to the inherent features of the estimator (17) (*i.e.* the diffusive behaviour of (12) and the kernel density estimator).

5.4. Scaling with dimension

To explore the scaling with the dimension d of the support of π we take the generalization of the Gaussian mixture model in Section 5.1 described in [26] (with a slight abuse of notation we denote both a measure and its density with the same symbol)

$$\begin{aligned}
\pi(x) &= (1/3)\mathcal{N}(x; 0.3 \cdot \mathbf{1}_d, 0.07^2\,\mathrm{Id}_d) + (2/3)\mathcal{N}(x; 0.7 \cdot \mathbf{1}_d, 0.1^2\,\mathrm{Id}_d), \\
\mathrm{k}(x, y) &= \mathcal{N}(y; x, 0.15^2\,\mathrm{Id}_d), \\
\mu(y) &= (1/3)\mathcal{N}(y; 0.3 \cdot \mathbf{1}_p, (0.07^2 + 0.15^2)\,\mathrm{Id}_p) \\
&\quad + (2/3)\mathcal{N}(y; 0.7 \cdot \mathbf{1}_p, (0.1^2 + 0.15^2)\,\mathrm{Id}_p),
\end{aligned} \tag{23}$$

where $p = d$ and $\mathbf{1}_d, \mathrm{Id}_d$ denote the unit function in $\mathbb{R}^d$ and the $d \times d$ identity matrix, respectively.

We compare the reconstructions given by FE-WGF with those obtained with the one-step-late expectation maximization (OSL-EM) algorithm [42], an iterative algorithm for maximum penalized likelihood estimation, and SMC-EMS [26]. First, we focus on the comparison with OSL-EM since this algorithm can be implemented to minimize $\mathscr{G}_\alpha^\eta$ in (5) as does FE-WGF. In particular, FE-WGF is a stochastic approach to the problem of minimizing (3), while OSL-EM relies on a deterministic discretization of π. Secondly, we compare the proposed approach with SMC-EMS [26].

5.4.1. Comparison with OSL-EM

As briefly discussed in Section 2.2, the first term in the definition of the functional $\mathscr{G}_\alpha$ in (4) (with μ replaced by μ^M) is the likelihood associated with the incomplete-data $\{y^{k,M}\}_{k=1}^M$. Minimizing $\mathscr{G}_\alpha$ is equivalent to maximizing a penalized likelihood with

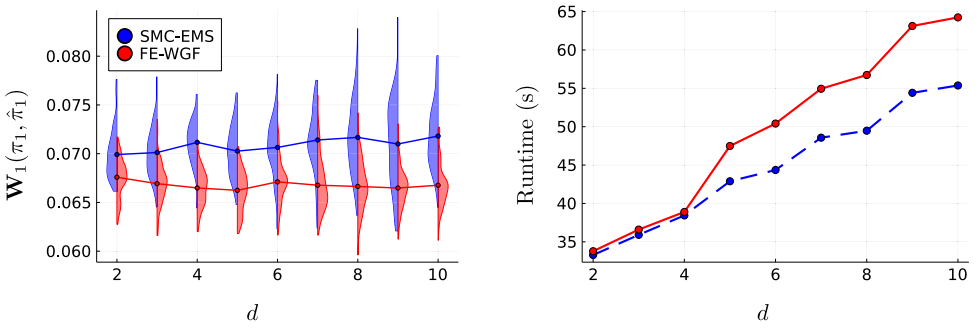

Fig. 6. Reconstruction accuracy over 100 of SMC-EMS and FE-WGF for the Gaussian mixture model in (23). The left panel shows the distribution of the Wasserstein-1 distance between the first marginal of π and its particle approximations for $d = 2, \dots, 10$, the solid lines highlight the average accuracy. The right panel reports average runtime.

penalty $\mathrm{KL}(\pi|\pi_0)$. The one-step-late expectation maximization (OSL-EM) algorithm [42] is a variant of the well-known EM algorithm [33] to maximize any penalized likelihood.

In order to implement OSL-EM we need to discretize the support of π, which essentially requires π to have known compact support. This is not the case for the measure π considered in this example, however, $|\int_{[0,1]^d} \pi(x)\mathrm{d}x - 1| < 10^{-2}$ for $d \leq 10$; therefore we define a discretization grid for OSL-EM over the d-dimensional hypercube $[0,1]^d$ (similar considerations apply to μ). We consider $B \in \mathbb{N}$ equally spaced bins for each dimension and obtain the discretizations of π and μ, $\{\pi_b\}_{b=1}^{B^d}, \{\mu_b\}_{b=1}^{B^d}$, setting $\pi_b = \pi(x_b)$, where x_b is the centre of the bth bin. The same discretization mechanism is used for k to obtain $\{\mathrm{k}_{b,c}\}_{b,c=1}^{B^d}$, where $\mathrm{k}_{b,c} = \mathrm{k}(x_b, x_c)$. The one-dimensional OSL-EM iteration is given for any $n \in \mathbb{N}$ by

$$\pi_b^{(n+1)} = \pi_b^{(n)} / (1 + \alpha(1 + \log \pi_b^{(n)} - \log \pi_{0,b})) \sum_{c=1}^{B} \{\mu_c \mathrm{k}_{b,c} / \sum_{d=1}^{B} \pi_d^{(n)} \mathrm{k}_{d,c}\},$$

where $\pi_{0,b} = \pi_0(x_b)$, for $b \in \{1, \dots, B^d\}$. Contrary to the estimator provided by FE-WGF, the reconstructions provided by OSL-EM are deterministic and might not converge to the unique minimizer of $\mathscr{G}_\alpha^\eta$. For the mixture model in (23), we observe empirical convergence to a unique fixed point for reasonable values of α, e.g. $\alpha < 1$.

To compare OSL-EM and FE-WGF we fix $\alpha = 0.01$ (as both algorithms aim at minimizing $\mathscr{G}_\alpha^\eta$ the value of $\alpha = 0.01$ is chosen to compare them in a regime of practical interest, with the aim of obtaining good approximations of π), set the reference measure π_0 to be a Gaussian distribution with mean $(0.5, \dots, 0.5)$ and covariance matrix $\sigma_0^2 \mathrm{Id}_d$ with $\sigma_0^2 = 0.25^2$ and use π_0 as initial condition for FE-WGF. With this choice of parameters, convergence (assessed empirically by approximating the value of $\mathscr{G}_\alpha^\eta$) occurs within 50 steps for both algorithms and $d \in \{1, \dots, 4\}$. The number of particles is set to $N = 10^4$, and we obtain the number of bins per dimension, B, so that $B^d \approx N$, and the two algorithms have similar runtime. This corresponds to coarser discretizations as d increases which give low resolution reconstructions, for instance, for $d = 4$ each dimension is discretized into $B = 10$ bins while for $d = 1$ we have $B = N = 10^4$ bins; however, such fine discretization schemes are impractical for $d \gg 1$ since the number of points at which π needs to be approximated grows exponentially with d.

To assess the quality of the reconstructions we use the integrated squared error (20) of the first marginal, $\mathrm{ISE}(\hat{\pi}_1)$, where $\hat{\pi}$ is the estimated density. Since OSL-EM is a deterministic algorithm we take its ISE as reference and investigate the gains obtained by using FE-WGF for dimension $d = 1, 2, 3, 4$, see Fig. 5. For $d = 1, 2$ the gain with respect to OSL-EM is moderate, but for $d = 3, 4$ the reconstructions obtained with FE-WGF are at least 10 times more accurate that those obtained by OSL-EM with an average runtime which is only 1.3 times that of OSL-EM.

5.4.2. Comparison with SMC-EMS

We now turn to comparing the reconstructions obtained with FE-WGF with those given by SMC-EMS [26]. Both methods approximate π through a family of interacting particles from which smooth estimates can be obtained via standard kernel density estimation as in (17). Since typical applications of Fredholm integral equations are low-dimensional, mostly one to three dimensional with some notable exceptions [82], we focus on d up to 10.

To compare the two algorithms we consider $N = 10^3$ and we assume that μ is not known but a sample of size $M = 10^6$ is available and resample without replacement $m = N = 10^3$ times from it at each iteration of both SMC-EMS and FE-WGF. SMC-EMS converges quickly to an approximation of π [26], for the model in (23) we observe convergence in less than 50 steps (measured by numerically approximating the Kullback–Leibler divergence (2) using the samples from μ and $\hat{\pi}$, the estimated distribution). For FE-WGF we set $\gamma = 10^{-2}$ and select the reference measure π_0 to be a Gaussian distribution with mean $(0.5, \dots, 0.5)$ and covariance matrix $\sigma_0^2 \mathrm{Id}_d$ with $\sigma_0^2 = 0.25^2$; with this choice of parameters convergence occurs in less than 50 iterations also for FE-WGF. To specify the smoothing parameter for SMC-EMS and α for FE-WGF we use the cross-validation approach in Section 4.3 with $L = 10$.

Since kernel density estimators do not perform well for $d \gg 1$ [83], we do not build the kernel density estimator (17) and compare the reconstructions obtained by SMC-EMS and FE-WGF using the particle population given by each method as an approximation for

π (as justified by Proposition 9), and compute the Wasserstein-1 distance between the first marginal π_1 and its particle approximation (Fig. 6).

SMC-EMS and FE-WGF give similar results for all d, with the average runtime of SMC-EMS being slightly lower than that of FE-WGF for $d = 2, 3, 4$ and about 15% lower for $d \geq 5$. The increase in runtime for $d \geq 5$ is caused by the computation of the norm of the drift for the tamed Euler scheme (15) and is likely due to the increased algebraic complexity of this operation with dimension. In terms of accuracy, FE-WGF performs slightly better than SMC-EMS with $\mathbf{W}_1$ distance 5% smaller on average for FE-WGF.

6. Discussion

We consider a probabilistic approach to the solution of Fredholm integral equations of the first kind (1) by introducing the minimization problem (3) and its surrogate (5). Under mild smoothness assumptions on the kernel K and the reference measure π_0, we show that the surrogate functional admits a unique minimizer which is stable w.r.t. μ. Leveraging a Wasserstein gradient flow construction, we introduce an interacting particle system (12) which approximates the minimizer of the surrogate functional and is ergodic for any finite N. Combining standard propagation of chaos results and strong convergence of tamed Euler–Maruyama schemes we obtain the bound (19) which provides a practical guideline for the selection of the time discretization step γ, the number of particles N and the number of samples m from μ to use at each iteration.

In Section 5 we compare FE-WGF with both problem specific and state-of-the-art methods to solve (1); specifically, we compare our algorithm with SMC-EMS, a particle method to solve (1) which also aims at minimizing a regularized Kullback–Leibler divergence. This particle method has been shown to achieve state-of-the-art performances in a number of settings and to scale with the dimension of the state space d better than standard discretization based algorithms [26]. Our results show that FE-WGF has always comparable or superior performances to problem specific algorithms and outperforms SMC-EMS in high dimensional setting. Moreover, FE-WGF allows the introduction of a reference measure π_0, which can be particularly beneficial when some features of the measure π to be reconstructed are known (as, for instance, in epidemiology applications as that in Section 5.2).

From a theoretical perspective, Proposition 2 and the convergence properties in Section 3 guarantee that the particle system (12) approximates the unique minimizer of (5); an equivalent uniqueness property has not yet been established for SMC-EMS. However, the results presented in Section 5 also provide some additional confidence in the use of SMC-EMS, demonstrating empirically that it recovers comparable solutions to the provably convergent method developed in this paper in a range of settings.

Moreover, while both algorithms minimize a regularized Kullback–Leibler divergence, and for small values of the regularization parameters the reconstruction they provide are similar, the methods through which they approximate the minimizer present some notable differences. FE-WGF is based on a numerical approximation of the MKVSDE (10) and therefore the associated particle system has a diffusive behaviour, while SMC-EMS uses a resampling mechanism to eliminate the particles which are not in the support of π and replicate those close to high-probability regions. The effect of this different algorithmic approach can be observed by considering different initial distributions for $\{\mathbf{X}_0^{k,N}\}_{k=1}^N$: SMC-EMS benefits from overly diffused initial distributions, since the resampling mechanism allows to quickly move the particles towards high probability regions, for concentrated distributions (e.g. Dirac delta) convergence still occurs but in a larger number of iterations [26, Appendix E.1]. The convergence speed of FE-WGF is less influenced by the initial state $\{\mathbf{X}_0^{k,N}\}_{k=1}^N$ as shown in Appendix D. In contrast with SMC-EMS, FE-WGF requires specification of a reference measure π_0. While this might seem a limitation of our method, we found that in practice we often have sufficient knowledge about the problem at hand to build π_0. For instance, in image processing applications usually a large number of images needs to be processed and it is possible to build reference measures using previously processed images (e.g. [54]), while in epidemiology, one can exploit previous studies on similar diseases, since they often present incidence curves with the same characteristics (e.g. [39]). In fact, it is in this latter setting that we find that FE-WGF outperforms SMC-EMS in terms of accuracy (Section 5.2).

FE-WGF describes one possible numerical approximation of the particle system (12), however, we anticipate that more sophisticated time discretization schemes will prove beneficial in some settings. For instance, in our experiments we observed that in the first few iterations the drift component in (12) has larger values, and pushes the particles towards high π-probability regions; after the first iterations, the magnitude of the drift component decreases and the shape of the approximated solution is refined. In terms of the functional $\mathscr{G}_\alpha^\eta$, the first few iterations correspond to a steepest decrease in $\mathscr{G}_\alpha^\eta$, which then stabilizes around the minimum (see e.g., Appendix D). To speed up convergence to the minimizer, one could consider having a larger time step γ for earlier iterations, and then gradually reduce γ to refine the approximation. More sophisticated adaptive strategies for MKVSDE have been recently studied in [79].

In conclusion, FE-WGF offers a general recipe to obtain regularized solutions to Fredholm integral equations of the first kind, which can be refined exploiting specific knowledge of the problem at hand; for example, choosing informative reference measures π_0 or adapting the time discretization step Δt to the magnitude of the drift (13).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

FRC, VDB, AD and AMJ acknowledge support from the EPSRC, UK (grant # EP/R034710/1). AMJ acknowledges further support from the EPSRC, UK (grant # EP/T004134/1) and the Lloyd's Register Foundation, UK Programme on Data-Centric Engineering at the Alan Turing Institute. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission. No new data was created in this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spa.2024.104374.

References

[1] Umberto Amato, W. Hughes, Maximum entropy regularization of Fredholm integral equations of the first kind, Inverse Problems 7 (6) (1991) 793.
[2] Luigi Ambrosio, Nicola Gigli, Giuseppe Savaré, Gradient Flows: in Metric Spaces and in the Space of Probability Measures, Springer Science & Business Media, 2008.
[3] Anders Andreassen, Patrick T Komiske, Eric M Metodiev, Benjamin Nachman, Adi Suresh, Jesse Thaler, Scaffolding simulations with deep learning for high-dimensional deconvolution, in: ICLR 2021 SimDL Workshop, 2021.
[4] Fabio Antonelli, Arturo Kohatsu-Higa, Rate of convergence of a particle method to the solution of the McKean–Vlasov equation, Ann. Appl. Probab. 12 (2) (2002) 423–476.
[5] Michael Arbel, Anna Korba, Adil Salim, Arthur Gretton, Maximum mean discrepancy gradient flow, in: Advances in Neural Information Processing Systems, 2019, pp. 6481–6491.
[6] Samuel G Armato, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931.
[7] Richard C. Aster, Brian Borchers, Clifford H. Thurber, Parameter Estimation and Inverse Problems, Elsevier, 2018.
[8] Jianhai Bao, Christoph Reisinger, Panpan Ren, Wolfang Stockinger, First order convergence of Milstein schemes for McKean–Vlasov equations and interacting particle systems, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. (2020).
[9] Niels G. Becker, Lyndsey F. Watson, John B. Carlin, A method of non-parametric back-projection and its application to AIDS data, Stat. Med. 10 (1991) 1527–1542.
[10] Tatiana Benaglia, Didier Chauveau, David R. Hunter, Derek Young, mixtools: An r package for analyzing finite mixture models, J. Stat. Softw. 32 (6) (2009) 1–29.
[11] Pier Giovanni Bissiri, Chris C. Holmes, Stephen G. Walker, A general framework for updating belief distributions, J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (5) (2016) 1103–1130.
[12] Vladimir I Bogachev, Michael Röckner, Stanislav V Shaposhnikov, Convergence in variation of solutions of nonlinear Fokker–Planck–Kolmogorov equations to stationary measures, J. Funct. Anal. 276 (12) (2019) 3681–3713.
[13] Mireille Bossy, Denis Talay, A stochastic particle method for the McKean–Vlasov and the Burgers equation, Math. Comp. 66 (217) (1997) 157–192.
[14] Nicolas Brosse, Alain Durmus, Éric Moulines, Sotirios Sabanis, The tamed unadjusted Langevin algorithm, Stochastic Process. Appl. 129 (10) (2019) 3638–3663.
[15] Martin Burger, Elena Resmerita, Martin Benning, An entropic Landweber method for linear ill-posed problems, Inverse Problems 36 (1) (2019) 015009.
[16] Oleg A. Butkovsky, On ergodic properties of nonlinear Markov chains and stochastic McKean–Vlasov equations, Theory Probab. Appl. 58 (4) (2014) 661–674.
[17] Charles Byrne, Paul P.B. Eggermont, EM algorithms, in: Handbook of Mathematical Methods in Imaging, Springer, 2015, pp. 305–388.
[18] René Carmona, François Delarue, Probabilistic analysis of mean-field games, SIAM J. Control Optim. 51 (4) (2013) 2705–2734.
[19] Minwoo Chae, Ryan Martin, Stephen G. Walker, Convergence of an iterative algorithm to the nonparametric MLE of a mixing distribution, Statist. Probab. Lett. 140 (2018) 142–146.
[20] Minwoo Chae, Ryan Martin, Stephen G. Walker, On an algorithm for solving Fredholm integrals of the first kind, Statist. Comput. 29 (2018) 645–654.
[21] Pui Hing Chau, Wei Ying Li, Paul S.F. Yip, Construction of the infection curve of local cases of COVID-19 in Hong Kong using back-projection, Int. J. Environ. Res. Public Health 17 (18) (2020) 6909.
[22] Lénaïc Chizat, Mean-field langevin dynamics: Exponential convergence and annealing, Trans. Mach. Learn. Res. (2022).
[23] Christian Clason, Barbara Kaltenbacher, Elena Resmerita, Regularization of ill-posed problems with non-negative solutions, in: Splitting Algorithms, Modern Operator Theory, and Applications, Springer, 2020, pp. 113–135.
[24] David Colton, Rainer Kress, Inverse Acoustic and Electromagnetic Scattering Theory, vol. 93, Springer, 2012.
[25] Francesca R. Crucinio, Some Interacting Particle Methods with Non-Standard Interactions (Ph.D. thesis), University of Warwick, 2021.
[26] Francesca R. Crucinio, Arnaud Doucet, Adam M. Johansen, A particle method for solving Fredholm equations of the first kind, J. Amer. Statist. Assoc. 118 (542) (2023) 937–947.
[27] Imre Csiszár, I-divergence geometry of probability distributions and minimization problems, Ann. Probab. 3 (1975) 146–158.
[28] Gianni Dal Maso, An Introduction to Γ-Convergence, in: Progress in Nonlinear Differential Equations and their Applications, vol. 8, Birkhäuser Boston, Inc., Boston, MA, 1993, p. xiv+340.
[29] Kaustuv Datta, Deepak Kar, Debarati Roy, Unfolding with generative adversarial networks, 2018, arXiv preprint arXiv:1806.00433.
[30] Valentin De Bortoli, Alain Durmus, Convergence of diffusions and their discretizations: from continuous to discrete processes and back, 2019, arXiv preprint arXiv:1904.09808.
[31] Aurore Delaigle, An alternative view of the deconvolution problem, Statist. Sinica (2008) 1025–1045.
[32] Aurore Delaigle, Irène Gijbels, Practical bandwidth selection in deconvolution kernel density estimation, Comput. Statist. Data Anal. 45 (2) (2004) 249–267.
[33] Arthur P. Dempster, Nan M. Laird, Donald B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977) 2–38.
[34] Tim Dockhorn, James A. Ritchie, Yaoliang Yu, Iain Murray, Density deconvolution with normalizing flows, in: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models 2020, 2020.
[35] Alain Durmus, Szymon Majewski, Błażej Miasojedow, Analysis of Langevin Monte Carlo via convex optimization, J. Mach. Learn. Res. 20 (1) (2019) 2666–2711.
[36] Andreas Eberle, Reflection couplings and contraction rates for diffusions, Probab. Theory Relat. Fields 166 (3–4) (2016) 851–886.
[37] Paul P.B. Eggermont, Maximum entropy regularization for Fredholm integral equations of the first kind, SIAM J. Math. Anal. 24 (6) (1993) 1557–1576.

[38] Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, Andrew M Stuart, Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler, SIAM J. Appl. Dyn. Syst. 19 (1) (2020) 412–441.
[39] Edward Goldstein, Jonathan Dushoff, Junling Ma, Joshua B Plotkin, David JD Earn, Marc Lipsitch, Reconstructing influenza incidence by deconvolution of daily mortality time series, Proc. Natl. Acad. Sci. 106 (51) (2009) 21825–21829.
[40] Katelyn M Gostic, Lauren McGough, Edward B Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, Rebecca Kahn, Rene Niehus, James A Hay, Pablo M De Salazar, et al., Practical considerations for measuring the effective reproductive number, Rt, PLoS Comput. Biol. 16 (12) (2020) e1008409.
[41] Alexander David Gottlieb, Markov transitions and the propagation of chaos, 2000, arXiv preprint math/0001076.
[42] Peter J. Green, On use of the EM for penalized likelihood estimation, J. R. Stat. Soc. Ser. B Stat. Methodol. 52 (3) (1990) 443–452.
[43] Charles W. Groetsch, The Theory of Tikhonov Regularization for Fredholm Equations, Pitman Advanced Publishing Program, 1984, p. xiii+104.
[44] Peter Grünwald, The safe Bayesian, in: Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, Thomas Zeugmann (Eds.), Algorithmic Learning Theory, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 169–183.
[45] Peter Hall, Joel L. Horowitz, Nonparametric methods for inference in the presence of instrumental variables, Ann. Statist. 33 (6) (2005) 2904–2929.
[46] Kaitong Hu, Zhenjie Ren, David Šiška, Łukasz Szpruch, Mean-field Langevin dynamics and energy landscape of neural networks, in: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, Vol. 57, Institut Henri Poincaré, 2021, pp. 2043–2065.
[47] Martin Hutzenthaler, Arnulf Jentzen, Numerical Approximations of Stochastic Differential Equations with Non-Globally Lipschitz Continuous Coefficients, American Mathematical Society, 2015.
[48] Martin Hutzenthaler, Arnulf Jentzen, Peter E. Kloeden, Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients, Ann. Appl. Probab. 22 (4) (2012) 1611–1641.
[49] Md Shafiqul Islam, Adam Smith, Approximating solutions of Fredholm integral equations via a general spline maximum entropy method, Int. J. Appl. Comput. Math. 6 (3) (2020).
[50] Alfredo Noel Iusem, B.F. Svaiter, A new smoothing-regularization approach for a maximum-likelihood estimation problem, Appl. Math. Optim. 29 (3) (1994) 225–241.
[51] Edwin T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (4) (1957) 620–630.
[52] Congming Jin, Jiu Ding, Solving Fredholm integral equations via a piecewise linear maximum entropy method, J. Comput. Appl. Math. 304 (2016) 130–137.
[53] Richard Jordan, David Kinderlehrer, Felix Otto, The variational formulation of the Fokker–Planck equation, SIAM J. Math. Anal. 29 (1) (1998) 1–17.
[54] Diederik P. Kingma, Yann LeCun, Regularized estimation of image statistics by score matching, in: Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1, 2010, pp. 1126–1134.
[55] Peter E. Kloeden, Eckhard Platen, Numerical Solution of Stochastic Differential Equations, Springer, 1992.
[56] Sławomir Kopeć, On application of maxent to solving Fredholm integral equations, in: Maximum Entropy and Bayesian Methods, Springer, 1993, pp. 63–66.
[57] Rainer Kress, Linear Integral Equations, in: Applied Mathematical Sciences, vol. 82, Springer, 2014.
[58] Nan Laird, Nonparametric maximum likelihood estimation of a mixing distribution, J. Amer. Statist. Assoc. 73 (364) (1978) 805–811.
[59] Qiang Liu, Stein variational gradient descent as gradient flow, Adv. Neural Inf. Process. Syst. 30 (2017).
[60] Jiaming Liu, Yu Sun, Xiaojian Xu, Ulugbek S. Kamilov, Image restoration using total variation regularized deep image prior, in: International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 7715–7719.
[61] Jun Ma, Indirect density estimation using the iterative Bayes algorithm, Comput. Statist. Data Anal. 55 (3) (2011) 1180–1195.
[62] Florient Malrieu, Logarithmic Sobolev inequalities for some nonlinear PDEs, Stochastic Process. Appl. 95 (1) (2001) 109–132.
[63] I.C. Marschner, Back-projection of COVID-19 diagnosis counts to assess infection incidence and control measures: analysis of Australian data, Epidemiol. Infect. 148 (2020) e97.
[64] Henry P. McKean, A class of Markov processes associated with nonlinear parabolic equations, Proc. Natl. Acad. Sci. USA 56 (6) (1966) 1907–1911.
[65] Lawrence R. Mead, Approximate solution of Fredholm integral equations by the maximum-entropy method, J. Math. Phys. 27 (12) (1986) 2903–2907.
[66] Sylvie Méléard, Asymptotic behaviour of some interacting particle systems; McKean–Vlasov and Boltzmann models, in: Probabilistic Models for Nonlinear Partial Differential Equations, Springer, 1996, pp. 42–95.
[67] Sylvie Méléard, Sylvie Roelly-Coppoletta, A propagation of chaos result for a system of particles with moderate interaction, Stochastic Process. Appl. 26 (1987) 317–332.
[68] Wang Miao, Zhi Geng, Eric J. Tchetgen Tchetgen, Identifying causal effects with proxy variables of an unmeasured confounder, Biometrika 105 (4) (2018) 987–993.
[69] Andrew C Miller, Lauren Hannah, Joseph Futoma, Nicholas J Foti, Emily B Fox, Alexander D'Amour, Mark Sandler, Rif A Saurous, Joseph Lewnard, Statistical Deconvolution for Inference of Infection Time Series, Technical Report 4, Wolters Kluwer Health, 2022, p. 470.
[70] Rafael Molina, Ascension del Olmo, Jaime Perea, Brian D Ripley, Bayesian deconvolution in optical astronomy, Astron. J. 103 (1992) 666–675.
[71] Rafael Molina, Brian D. Ripley, Using spatial models as priors in astronomical image analysis, J. Appl. Stat. 20 (5–6) (1993) 281–298.
[72] Eliakim H. Moore, On the reciprocal of the general algebraic matrix, Bull. Amer. Math. Soc. 26 (1920) 394–395.
[73] Heinrich N. Mülthei, Benno Schorr, Willi Törnig, On an iterative method for a class of integral equations of the first kind, Math. Methods Appl. Sci. 9 (1) (1987) 137–168.
[74] Thomas Obadia, Romana Haneef, Pierre-Yves Boëlle, The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 1–9.
[75] Karl Oelschlager, A martingale approach to the law of large numbers for weakly interacting stochastic processes, Ann. Probab. (1984) 458–479.
[76] Felix Otto, The geometry of dissipative evolution equations: the porous medium equation, Comm. Partial Differential Equations (2001).
[77] Marianna Pensky, et al., Minimax theory of estimation of linear functionals of the deconvolution density with or without sparsity, Ann. Statist. 45 (4) (2017) 1516–1541.
[78] Johann Radon, On the determination of functions from their integral values along certain manifolds, IEEE Trans. Med. Imaging 5 (4) (1986) 170–176.
[79] Christoph Reisinger, Wolfgang Stockinger, An adaptive Euler–Maruyama scheme for McKean–Vlasov SDEs with super-linear growth and application to the mean-field FitzHugh–Nagumo model, J. Comput. Appl. Math. 400 (2022) 113725.
[80] Elena Resmerita, Robert S. Anderssen, Joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems, Math. Methods Appl. Sci. 30 (13) (2007) 1527–1544.
[81] Filippo Santambrogio, {Euclidean, metric, and Wasserstein} gradient flows: an overview, Bull. Math. Sci. 7 (1) (2017) 87–154.
[82] Alberto Signoroni, Mattia Savardi, Annalisa Baronio, Sergio Benini, Deep learning meets hyperspectral image analysis: a multidisciplinary review, J. Imaging 5 (5) (2019) 52.
[83] Bernard W. Silverman, Density Estimation for Statistics and Data Analysis, in: Monographs on Statistics and Applied Probability, vol. 26, Chapman & Hall, 1986.
[84] Bernard W. Silverman, M.C. Jones, J.D. Wilson, D.W. Nychka, A smoothed EM approach to indirect estimation problems, with particular, reference to stereology and emission tomography, J. R. Stat. Soc. Ser. B Stat. Methodol. 52 (2) (1990) 271–324.
[85] Donald L. Snyder, Timothy J. Schulz, Joseph A. O'Sullivan, Deblurring subject to nonnegativity constraints, IEEE Trans. Signal Process. 40 (5) (1992) 1143–1150.

[86] Alain-Sol Sznitman, Topics in propagation of chaos, in: Ecole d'Été de Probabilités de Saint-Flour XIX—1989, in: Lecture Notes in Mathematics, vol. 1464, Springer, Berlin, 1991, pp. 165–251.
[87] Hiroshi Tanaka, Limit theorems for certain diffusion processes with interaction, in: North-Holland Mathematical Library, vol. 32, Elsevier, 1984, pp. 469–488.
[88] Vitalii Pavlovich Tanana, E Yu Vishnyakov, Anna Ivanovna Sidikova, An approximate solution of a Fredholm integral equation of the first kind by the residual method, Numer. Anal. Appl. 9 (1) (2016) 74–81.
[89] Shan Tong, Adam M. Alessio, Paul E. Kinahan, Image reconstruction for PET/CT scanners: past achievements and future challenges, Imaging Med. 2 (5) (2010) 529.
[90] Wim Van Aarle, Willem Jan Palenstijn, Jan De Beenhouwer, Thomas Altantzis, Sara Bals, K Joost Batenburg, Jan Sijbers, The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography, Ultramicroscopy 157 (2015) 35–47.
[91] Grace Wahba, Practical approximate solutions to linear operator equations when the data are noisy, SIAM J. Numer. Anal. 14 (4) (1977) 651–667.
[92] Xiaojie Wang, Siqing Gan, The tamed Milstein method for commutative stochastic differential equations with non-globally Lipschitz continuous coefficients, J. Difference Equ. Appl. 19 (3) (2013) 466–490.
[93] Shuo Wang, Xian Yang, Ling Li, Philip Nadler, Rossella Arcucci, Yuan Huang, Zhongzhao Teng, Yike Guo, A Bayesian Updating Scheme for Pandemics: Estimating the Infection Dynamics of COVID-19, IEEE Comput. Intell. Mag. 15 (4) (2020) 23–33.
[94] Andrew G. Webb, Introduction to Biomedical Imaging, John Wiley & Sons, 2017.
[95] Ran Yang, Daniel W. Apley, Jeremy Staum, David Ruppert, Density deconvolution with additive measurement errors using quadratic programming, J. Comput. Graph. Statist. 29 (3) (2020) 580–591.
[96] Chen Zhang, Simon Arridge, Bangti Jin, Expectation propagation for Poisson data, Inverse Problems 35 (8) (2019) 085006.
[97] Haichao Zhang, David Wipf, Yanning Zhang, Multi-observation blind deconvolution with an adaptive sparse prior, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1628–1643.

Further reading

[1] Patrick Billingsley, Probability and Measure, John Wiley & Sons, 1995.
[2] Rainer Buckdahn, Juan Li, Shige Peng, Catherine Rainer, Mean-field stochastic differential equations and associated PDEs, Ann. Probab. 45 (2) (2017) 824–878.
[3] Pierre Cardaliaguet, Notes on Mean Field Games, Technical Report, Collége de France, 2010.
[4] René Carmona, Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications, SIAM, 2016.
[5] René Carmona, François Delarue, Probabilistic Theory of Mean Field Games with Applications I-II, Springer, 2018.
[6] Paul Dupuis, Richard S. Ellis, A Weak Convergence Approach to the Theory of Large Deviations, in: Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., New York, 1997, p. xviii+479, A Wiley-Interscience Publication.
[7] William RP Hammersley, David Šiška, Łukasz Szpruch, McKean–Vlasov SDEs under measure dependent Lyapunov conditions, in: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, Vol. 57, Institut Henri Poincaré, 2021, pp. 1032–1057.
[8] Ioannis Karatzas, Steven E. Shreve, Brownian Motion and Stochastic Calculus, second ed., in: Graduate Texts in Mathematics, vol. 113, Springer-Verlag, New York, 1991, p. xxiv+470.
[9] Daniel Revuz, Marc Yor, Continuous Martingales and Brownian Motion, third ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 293, Springer-Verlag, Berlin, 1999, p. xiv+602.
[10] Cédric Villani, Optimal Transport: Old and New, vol. 338, Springer Science & Business Media, 2008.