Solving Fredholm Integral Equations of the First Kind via Wasserstein Gradient Flow

Francesca Crucinio, Arnaud Doucet and Adam M. Johansen

1 Introduction

2 Background

2.1 Fredholm Integral Equations of the First Kind

We want to solve the integral equation

$$\mu(y) = \int_{\mathbb{X}} \rho(x)\, K(x,y) dx, \qquad \forall y \in \mathbb{Y}$$

where ρ and μ are probability densities on $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{Y} \subseteq \mathbb{R}^m$ and K is a Markov transition density, i.e. $\mu = \rho K$ in operator notation.

We can steal some bits from the introduction to the other paper

- applications
- connection with MLE/EM

$$L(\rho) = \mathrm{KL}(\mu, \rho K) + \int \rho \tag{1}$$

The solution to this problem is not unique and we propose to regularize the problem using an entropic penalty; i.e. for a given $\alpha > 0$ we propose to minimize w.r.t. ρ

$$E(\rho) = \mathrm{KL}(\mu, \rho K) - \alpha\, \mathrm{ent}(\rho) \tag{2}$$

where $\mathrm{KL}(\mu, \rho K)$ is the Kullback-Leibler divergence between μ and ρK and $\mathrm{ent}(\rho) = -\int \rho \log \rho$ is the entropy of ρ. This requires the solution of a minimization problem in the space of probability measures. We are going to follow a Wasserstein gradient flow approach.

2.2 Wasserstein Gradient Flows

Gradient flows are...

$$x'(t) = -\nabla F(x(t)) \tag{3}$$

Wasserstein gradient flows are the

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

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

and we define the 2-Wasserstein distance on this set

$$W_2(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \, d\pi(x,y)\right)^{1/2} \tag{4}$$

where $\Pi(\mu,\nu)$ is the set of all possible couplings between μ and ν. A minimiser of (4) always exists but might not be unique, each minimiser is called an optimal transport plan (Ambrosio et al., 2008, Theorem 6.2.4). We denote by $\mathcal{P}_2^{ac}(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ the subset of these measures which is absolutely continuous w.r.t. the appropriate Lebesgue measure. For every pair $\mu,\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ there is a unique π attaining the minimimum of (4), $\pi = (Id, t_\mu^\nu)$, where t_μ^ν denotes the unique transport map between μ and ν (see, for example, Ambrosio et al. (2008, page 150)). It is easy to check that for all $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ we have $\mathrm{ent}(\nu) < +\infty$.

To define gradient flows on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ we also need a notion of curve, among the several definitions proposed (Ambrosio et al., 2008, Chapter 9) we consider constant speed geodesics: for $\mu,\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$, the constant speed geodesic originating from μ and with endpoint ν is

$$\mu_s : s \in [0,1] \mapsto ((1-s)Id + st_\mu^\nu)_{\#}\mu \tag{5}$$

where t_μ^ν is the unique transport map between μ and ν and $T_{\#}\mu$ denotes the push-forward measure $T_{\#}\mu(A) = \mu(T^{-1}(A))$.

We then consider functionals $F : \mathcal{P}_2^{ac}(\mathbb{R}^d) \to \mathbb{R}$ defined on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ and build a gradient flow solving the minimisation problem

$$\min_{\rho\in\mathcal{P}_2^{ac}(\mathbb{R}^d)} F(\rho). \tag{6}$$

The construction of a gradient flow equation for F allows us to transform the minimisation problem (6) into a PDE whose solution is a density ρ solving (6). We will restrict our attention to functionals which are proper (i.e. $F(\rho) < +\infty$ for some $\rho \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$), continuous w.r.t. the W_2 metric (which metrizes weak convergence, Santambrogio (2017, Theorem 4.4)) and coercive. Following Ambrosio et al. (2008, Definition 2.1b), a functional F defined on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ is coercive if there exist $\tau > 0$ and $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ such that

$$\inf_{\rho\in\mathcal{P}_2^{ac}(\mathbb{R}^d)} \frac{1}{2\tau} W_2^2(\nu,\rho) + F(\rho) > -\infty.$$

Should we try to connect this with the normal definition of coercivity? AMJ

Since we are dealing with an optimization problem, it is natural to assume that the functional F is convex in an appropriate sense; in $\mathcal{P}_2^{ac}(\mathbb{R}^d)$ the standard notion of convexity in $\mathbb{R}^d$ corresponds to convexity along geodesics: for all $\nu,\mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ take a geodesic connecting μ to ν (5), then the functional F is λ-geodesically convex if

$$F\left(((1-s)Id + st_\mu^\nu)_{\#}\mu\right) \le (1-s)F(\mu) + sF(\nu) - \frac{\lambda}{2}s(1-s)W_2^2(\nu,\mu).$$

It is easy to see that when $\lambda = 0$ the above corresponds to the standard notion of convexity, in this case the functional is called displacement convex, if $\lambda > 0$ the above is stronger than convexity, and for $\lambda < 0$ is weaker (Ambrosio et al., 2008, page 202).

The last ingredient in the definition of a gradient flow for functional F in $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$ is a notion of gradient for F. As standard definitions of derivatives do not apply outside vector spaces, it is standard practice to replace the gradient in (3)with the sub-gradient $\partial F(\mu)$ of F.

For any proper, lower semicontinuous, λ-geodesically convex functional F defined on $\mathcal{P}_2^{ac}(\mathbb{R}^d)$, ξ belongs to the sub-differential $\partial F(\mu)$ if

$$F(\nu) - F(\mu) \geq \int \langle \xi(x), t_\mu^\nu(x) - x \rangle d\mu(x) + \frac{\lambda}{2} W_2^2(\nu, \mu) \tag{7}$$

for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2008, page 231). Equipped with this notion of gradient, we can now define the gradient flow equation for a functional F on $\left(\mathcal{P}_2^{ac}(\mathbb{R}^d), W_2\right)$: ρ_t is a solution of the gradient flow equation for F if

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t) \tag{8}$$

with $v_t \in -\partial F(\rho_t)$ Ambrosio et al. (2008, Section 11.1.2). Existence of solutions of (8) given an initial condition $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ is ensured for all functionals F which are proper, continuous and coercive if the sub-differential (7) is single-valued (Ambrosio et al., 2008, Theorem 11.1.6 and Corollary 11.1.8). The solution of (8) with initial condition ρ_0 is unique for $\lambda-$convex functionals (Ambrosio et al., 2008, Theorem 11.1.4). In particular we have the estimate

$$W_2(\rho_t^1, \rho_t^2) \leq e^{-\lambda t} W_2(\rho_0^1, \rho_0^2)$$

for all $t > 0$,for ρ_t^i solution of the gradient flow equation with initial condition ρ_0^i, $i = 1, 2$, which shows that the gradient flow is contractive w.r.t. W_2 when $\lambda > 0$ and non-expansive when $\lambda = 0$.

3 Gradient Flow Approach

In order to apply the gradient flow construction described above to (2) we make the following assumptions:

(A0) $\mathbb{X}$ and $\mathbb{Y}$ are bounded and open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, μ, ρ are probability densities with finite second moment and K is a the density of a Markov kernel for each x.

(A1) the kernel $K(x, y)$ is bounded above and below

$$\exists m_K > 0 \text{ such that} \qquad 0 < \frac{1}{m_K} \leq K(x, y) \leq m_K < \infty \qquad \forall (x, y) \in \mathbb{X} \times \mathbb{Y},$$

is $\lambda-$concave in x, uniformly in y, i.e. $x \mapsto K(x, y) + \frac{\lambda}{2}\|x\|_2^2$ is concave for some $\lambda > 0$ for all $y \in \mathbb{Y}$, and is Lipschitz continuous in x, uniformly in y, with Lipschitz constant L

$$\|K(x, y) - K(x', y)\|_2 \leq L\|x - x'\|_2 \qquad \forall (x, y), (x', y) \in \mathbb{X} \times \mathbb{Y}.$$

(A2) the kernel $K(x, y)$ is differentiable with gradient $\nabla K(x, y)$ bounded, $\|\nabla K(x, y)\|_2 \leq B$, and Lipschitz continuous in x, uniformly in y, with Lipschitz constant L'

$$\|\nabla K(x, y) - \nabla K(x', y)\|_2 \leq L'\|x - x'\|_2 \qquad \forall (x, y), (x', y) \in \mathbb{X} \times \mathbb{Y}.$$

Assumption, (A1), on the kernel K, is used to demonstrate existence and uniqueness of the solution of the gradient flow and of the corresponding PDE. On a bounded set, the $\lambda-$concavity is satisfied for all twice continuously differentiable functions with some suitable $\lambda > 0$ (Santambrogio, 2017, page 91). As this is a very active area of research, there are directions to relax the $\lambda-$concavity assumption to weaker moduli of convexity (Craig, 2017).

The Lipschitz assumptions on K and ∇K are necessary to deal with the SDE corresponding to the gradient flow PDE. In principle it would suffice to assume that the gradient $\nabla K(x, y)$ is locally Lipschitz with polynomial growth and satisfies a monotonic growth condition (Dos Reis et al., 2019), but this would lead to more complicated Euler time-discretisation schemes (Reis et al., 2018).

3.1 Properties of the Functional E

Under assumptions (A0), (A1), we can show that the functional E in (2) defined on $\mathcal{P}_2^{ac}(\mathbb{X})$ is proper, coercive, continuous and geodesically convex; as a consequence we can build a gradient flow targetting the minimum of (2). To see that E is a proper functional, observe that the image of E is $(-\infty, +\infty]$, since the KL divergence is always positive and the $\operatorname{ent}(\rho) < +\infty$ for all $\rho \in \mathcal{P}_2^{ac}(\mathbb{X})$, and that there exists at least one $\rho \in \mathcal{P}_2^{ac}(\mathbb{X})$ such that $E(\rho) < +\infty$, e.g. take ρ to be uniform on $\mathbb{X} \subset \mathbb{R}^n$, the entropy is finite and the boundedness of K ensures that the KL divergence is finite too

$$\mathrm{KL}(\mu, \rho K) = \int \mu(dy) \log \rho K(y) + \operatorname{ent}(\mu) \leq \log m_K \int \mu(dy) \log \rho(\mathbb{X}) + \operatorname{ent}(\mu) < +\infty$$

since $\rho(\mathbb{X}) \equiv 1$ and $\operatorname{ent}(\mu) < +\infty$. Continuity of E follows from the properties of the KL divergence, of the entropy function ent and from the assumptions on K in (A1):

Proposition 1. *The functional E is continuous in $(\mathcal{P}_2^{ac}(\mathbb{X}), W_2)$.*

Proof. As W_2 metrizes weak convergence, take $\rho_n \rightharpoonup \rho$. Then

$$|E(\rho_n) - E(\rho)| \leq \left| \int \mu(dy) \left[\log \rho K(y) - \log \rho_n K(y) \right] \right| + \alpha \left| \int \left[\rho_n(dx) \log \rho_n(x) - \rho(dx) \log \rho(x) \right] \right|,$$

since $x \mapsto K(x, y)$ is a continuous function for all $y \in \mathbb{Y}$, weak convergence, $\rho_n \rightharpoonup \rho$, implies $\rho_n K(y) \to \rho K(y)$ and the continuity of the logarithm gives $\log \rho K(y) \to \log \rho_n K(y)$. The dominated convergence theorem gives

$$\left| \int \mu(dy) \left[\log \rho K(y) - \log \rho_n K(y) \right] \right| \to 0.$$

Similarly, continuity of the second term is given by the continuity of the entropy function and the dominated convergence theorem. □

Proposition 2. *The functional E is coercive in $(\mathcal{P}_2^{ac}(\mathbb{X}), W_2)$.*

Proof. Under (A1) we have that

$$\begin{aligned}
\frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) &= \frac{1}{2\tau} W_2^2(\nu, \rho) - \int \mu(dy) \log \rho K(y) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu) \\
&\geq \frac{1}{2\tau} W_2^2(\nu, \rho) - m_K \int \mu(dy) \log \rho(\mathbb{X}) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu) \\
&= \frac{1}{2\tau} W_2^2(\nu, \rho) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu).
\end{aligned}$$

Since $\mu \in \mathcal{P}_2^{ac}(\mathbb{X})$, $\operatorname{ent}(\mu) \leq C < +\infty$, the above is bounded below by

$$\begin{aligned}
\inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{X})} \frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) &\geq \inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{X})} \frac{1}{2\tau} W_2^2(\nu, \rho) - \alpha \operatorname{ent}(\rho) - \operatorname{ent}(\mu) \\
&\geq \inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{X})} \frac{1}{2\tau} W_2^2(\nu, \rho) - \alpha \operatorname{ent}(\rho) - C.
\end{aligned}$$

If we take $\nu = \rho$ in the above, we get, for all $\tau > 0$,

$$\inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{X})} \frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) \geq -C - \alpha \inf_{\rho \in \mathcal{P}_2^{ac}(\mathbb{X})} \operatorname{ent}(\rho) > -\infty$$

since $\rho \in \mathcal{P}_2^{ac}(\mathbb{X})$. □

Finally, we show that E is displacement convex, this result is crucial for the definition of the gradient flow for E, as it ensures that, given an initial condition $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{X})$, the gradient flow equation has a unique solution.

Proposition 3. *The functional E is displacement convex in $(\mathcal{P}_2^{ac}(\mathbb{X}), W_2)$.*

Proof. We have

$$\begin{aligned} E(\rho) &= \mathrm{KL}(\mu, \rho K) - \alpha\, \mathrm{ent}(\rho) \\ &= -\int \mu(dy) \log \rho K(y) - \alpha\, \mathrm{ent}(\rho) - \mathrm{ent}(\mu). \end{aligned}$$

The entropy $\mathrm{ent}(\mu)$ is constant w.r.t. ρ and $\mathrm{ent}(\rho)$ is displacement convex in ρ (Santambrogio, 2017, page 130).
By (A1), K is $\lambda-$concave in x and the functional $F : \rho \mapsto \rho K(y) = \int K(x,y)\ d\rho(x)$ is $\lambda-$geodesically concave for all $y \in \mathbb{Y}$; in particular, we have that for all $s \in [0,1]$ and $\lambda > 0$

$$F\left(((1-s)Id + st_\rho^\nu)_{\#}\rho\right) \geq (1-s)F(\rho) + sF(\nu) + \frac{\lambda}{2} s(1-s) W_2^2(\nu, \rho) \geq (1-s)F(\rho) + sF(\nu)$$

where t_ρ^ν is the unique transport map between ν and ρ (Santambrogio, 2017, page 128).
Since $-\log x$ is a convex decreasing function

$$\begin{aligned} -\log\left(F\left(((1-s)Id + st_\mu^\nu)_{\#}\rho\right)\right) &\leq -\log\left((1-s)F(\rho) + sF(\nu)\right) \\ &\leq (1-s)\left(-\log F(\rho)\right) + s\left(-\log F(\nu)\right), \end{aligned}$$

the same inquality holds after integration w.r.t. μ giving the result. □

For displacement convex functionals, the coercivity in Proposition 2 is equivalent to the existence of a finite infimum on $\mathcal{P}_2(\mathbb{X})$ Ambrosio et al. (2008, page 295).

3.2 Gradient Flow

The last step towards the definition of the gradient flow equation (8) for the functional E consists of finding the quantity v_t which belongs to its (negative) sub-differential. A good candidate is the first variation (or functional derivative) of E, $\frac{\delta E}{\delta \rho}(x)$, the unique (up to additive constant) function such that

$$\lim_{\epsilon \to 0} \epsilon^{-1} \left(E(\rho + \epsilon\chi) - E(\rho)\right) = \int \chi(dx) \frac{\delta E}{\delta \rho}(x)$$

for every signed measure χ and every $\epsilon > 0$ small enough such that $\rho + \epsilon\chi \in \mathcal{P}_2(\mathbb{X})$ (Santambrogio, 2017). If $\frac{\delta E}{\delta \rho}(x)$ exists and satisfies (7), then the gradient flow equation for the functional E is

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta E}{\delta \rho_t}\right).$$

Consider the functional E

$$\begin{aligned} E(\rho) &= \mathrm{KL}(\mu, \rho K) - \alpha\, \mathrm{ent}(\rho) \\ &= -\int \mu(dy) \log \rho K(y) - \alpha\, \mathrm{ent}(\rho) - \mathrm{ent}(\mu), \end{aligned}$$

the last term, $\mathrm{ent}(\mu)$, does not depend on ρ and does not constribute to $\frac{\delta E}{\delta \rho}$, additionally

$$\frac{-\delta \,\mathrm{ent}}{\delta \rho}(\rho) = 1 + \log \rho \tag{9}$$

is a sub-differential for $-\mathrm{ent}(\rho)$ (Carrillo et al., 2006, Lemma 8).
We are then left with the computation of the first variation of the KL divergence as a function of ρ, $F(\rho) := \mathrm{KL}(\mu, \rho K)$, then

$$\begin{aligned}
\lim_{\epsilon \to 0} \epsilon^{-1} \left(F\left(\rho + \epsilon\chi\right) - F(\rho) \right) &= \lim_{\epsilon \to 0} \epsilon^{-1} \left(- \int \mu(dy) \log\left(\left(\rho + \epsilon\chi\right) K(y) \right) + \int \mu(dy) \log \rho K(y) \right) \\
&= \lim_{\epsilon \to 0} \epsilon^{-1} \left(- \int \mu(dy) \log \left(1 + \epsilon \frac{\chi K(y)}{\rho K(y)} \right) \right) \\
&= \lim_{\epsilon \to 0} \epsilon^{-1} \left(- \int \mu(dy) \left(\epsilon \frac{\chi K(y)}{\rho K(y)} + o\left(\epsilon \frac{\chi K(y)}{\rho K(y)} \right) \right) \right) \\
&= - \int \mu(dy) \frac{\chi K(y)}{\rho K(y)}
\end{aligned}$$

where the third equality follows from the Taylor expansion of the logarithm as $\epsilon \to 0$. Rearrenging the above we obtain

$$\lim_{\epsilon \to 0} \epsilon^{-1} \left(F\left(\rho + \epsilon\chi\right) - F(\rho) \right) = - \int \chi(dx) \int \mu(dy) \frac{K(x,y)}{\rho K(y)}$$

showing that

$$\frac{\delta F}{\delta \rho}(x) = - \int \mu\left(dy\right) \frac{K(x,y)}{\rho K(y)}. \tag{10}$$

Remark 4. The EM algorithm for Fredholm integral equations of Kondor (1983) aims at minimising (1) by finding the zeros of the first variations of L: given the first variation

$$\frac{\delta L}{\delta \rho}(x) = - \int \mu(dy) \frac{K(x,y)}{\rho K(y)} + 1,$$

multiplying by $\rho(x) \geq 0$ and setting the result to 0 leads to the EM iteration

$$\rho(x) = \rho(x) \int \mu(dy) \frac{K(x,y)}{\rho K(y)}.$$

To show that $-\int \mu\left(dy\right) \frac{K(x,y)}{\rho K(y)}$ is a sub-differential for F, take $\rho, \nu \in \mathcal{P}_2^{ac}(\mathbb{X})$, the unique transport map between ν and ρ, t_ν^ρ, and use the definition of sub-differential in (7):

$$\begin{aligned}
& F(\nu) - F(\rho) - \int \left\langle \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle d\rho(x) \\
&= \int \mu(dy) \left[-\log \nu K(y) + \log \rho K(y) \right] - \int \left\langle \int \mu\left(dy\right) \frac{K(x,y)}{\rho K(y)}, t_\rho^\nu(x) - x \right\rangle d\rho(x) \\
&= \int \mu(dy) \left[-\log \nu K(y) + \log \rho K(y) \right] - \int \frac{\mu\left(dy\right)}{\rho K(y)} \int \left\langle K(x,y), t_\rho^\nu(x) - x \right\rangle d\rho(x) \\
&= \int \mu(dy) \left[-\log \frac{\nu K(y)}{\rho K(y)} + \frac{1}{\rho K(y)} \int \left\langle -K(x,y), t_\rho^\nu(x) - x \right\rangle d\rho(x) \right].
\end{aligned}$$

Carrillo et al. (2006, Lemma 9) show that if M is a $\lambda-$convex function, then for all $\nu \in \mathcal{P}_2^{ac}(\mathbb{X})$

$$\nu M(y) - \rho M(y) = \int \langle M(x,y), t_\rho^\nu(x) - x \rangle d\rho(x);$$

since K is $\lambda-$concave, $-K$ is $\lambda-$convex, thus

$$-\nu K(y) + \rho K(y) = \int \langle -K(x,y), t_\rho^\nu(x) - x \rangle d\rho(x)$$

and

$$\begin{aligned} F(\nu) - F(\rho) - \int \left\langle \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle d\rho(x) &= \int \mu(dy) \left[-\log \frac{\nu K(y)}{\rho K(y)} + \frac{-\nu K(y) + \rho K(y)}{\rho K(y)} \right] \\ &= \int \mu(dy) \left[-\log \frac{\nu K(y)}{\rho K(y)} - \frac{\nu K(y)}{\rho K(y)} + 1 \right]. \end{aligned}$$

The first order Taylor expansion with Lagrange remainder $-\log x = -\log 1 - (x-1) + (x-a)^2/(2a^2)$ with a between x and 1 gives

$$F(\nu) - F(\rho) - \int \left\langle \frac{\delta F}{\delta \rho}(x), t_\rho^\nu(x) - x \right\rangle d\rho(x) = \int \mu(dy) \frac{1}{2a(y)^2} \left(\frac{\nu K(y)}{\rho K(y)} - a(y) \right)^2 \geq 0$$

with $a(y)$ a value in between $\nu K(y)$ and $\rho K(y)$, showing that $\frac{\delta F}{\delta \rho}(x)$ is a subdifferential for F.since F is $\lambda-$geodesically convex with $\lambda = 0$.

Putting (9) and (10) together we obtain the subdifferential for E

$$\frac{\delta E}{\delta \rho}(x) = -\int \mu(dy) \frac{K(x,y)}{\rho K(y)} + \alpha(1 + \log \rho(x)) \tag{11}$$

and the gradient flow equation

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta E}{\delta \rho_t} \right). \tag{12}$$

The infinitesimal reduction of E along ρ_t is

$$\frac{dE(\rho_t)}{dt} = -\int \left\| \nabla \frac{\delta E}{\delta \rho_t}(x) \right\|^2 d\rho_t(x). \tag{13}$$

by construction (Arbel et al., 2019, page 14).

Since E is proper, continuous, coercive and displacement convex (see Section (3.1)) and the first variation (11) is single-valued, the gradient flow equation (12) has a unique solution for each initial condition $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{X})$ (Ambrosio et al., 2008, Chapter 11).

4 Nonlinear SDE Approach

The main focus of this section is to connect the Fokker-Plank/Kolmogorov forward equation in (12) with its corresponding SDE and to investigate the properties of the latter in terms of existence of solutions, ergodicity and numerical implementation. Starting from the first variation of E in (11), we observe that

$$\nabla \frac{\delta E}{\delta \rho}(x) = \nabla \left[- \int \mu(dy) \frac{K(x,y)}{\rho K(y)} + \alpha(1 + \log \rho(x)) \right]$$
$$= - \int \mu(dy) \frac{\nabla K(x,y)}{\rho K(y)} + \alpha \nabla \log \rho(x).$$

where the equality follows from (A0)-(A2) and Leibniz's integral rule.
Observing that

$$\nabla \cdot (\rho_t \nabla \log \rho_t) = \nabla \cdot \nabla \rho_t = \triangle \rho_t,$$

where $\triangle f = \sum_i \partial_i^2 f_i$ is the Laplacian, the PDE in (12) becomes

$$\partial_t \rho_t = - \nabla \cdot \left(\rho_t \int \mu(dy) \frac{\nabla K(x,y)}{\rho_t K(y)} \right) + \alpha \triangle \rho_t \tag{14}$$

with correspoding SDE

$$dX_t = \int \mu(dy) \frac{\nabla K(X_t, y)}{\rho_t K(y)} dt + \sqrt{2\alpha} dW_t, \quad X_0 \sim \rho_0, \tag{15}$$

where W_t is a standard $n-$dimensional Brownian motion. It is well known that the distribution of X_t satisfies (14) with initial distribution ρ_0 (e.g. Jordan et al. (1998)).

The SDE (15) is a McKean-Vlasov SDE(McKean, 1966), since the drift coefficient involves the distribution ρ_t itself; in the following, we show that under assumptions ((A0))-((A2)) the McKean-Vlasov SDE in (15) admits a unique solution. We then focus on the numerical implementation of (15), considering both discretisation in time (as for standard SDEs) and in space (required by the presence of ρ_t in the drift coefficient).

Remark 5. We can use the following equilibrium condition to obtain a characterisation of the solution ρ_t of (12): if $\rho^\star$ is a solution, then (13) must be 0

$$\int \left\| \nabla \frac{\delta E}{\delta \rho^\star}(x) \right\|^2 d\rho^\star(x) = 0.$$

Since $\rho^\star \in \mathcal{P}_2^{ac}(\mathbb{X})$ is a positive density it follows that I think I'm missing some subtlety here... AMJ

$$\nabla \frac{\delta E}{\delta \rho^\star}(x) = - \int \mu(dy) \frac{\nabla K(x,y)}{\rho^\star K(y)} + \alpha \nabla \log \rho^\star(x) = 0$$

for all $\alpha > 0$.

4.1 Existence and uniqueness

Existence and uniqueness of a strong solution to (15) follows from standard result for non-linear SDEs (Jourdain et al., 2007) provided that the drift coefficient and the diffusion coefficient are Lipschitz continuous in (x, ρ), since the diffusion coefficient is constant, we just have to show that the drift coefficient

$$\int \mu(dy) \frac{\nabla K(X_t, y)}{\rho_t K(y)}$$

is Lipschitz continuous in (x, ρ).

Take $x, x' \in \mathbb{R}^n$ and $\rho, \rho' \in \mathcal{P}_2^{ac}(\mathbb{X})$ and consider

$$\begin{aligned}
\| \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho K(y)} - \int \mu\,(dy)\, \frac{\nabla K(x',y)}{\rho' K(y)} \|_2 &\leq \| \int \mu\,(dy) \left[\frac{\nabla K(x,y)}{\rho K(y)} - \frac{\nabla K(x',y)}{\rho' K(y)} \right] \|_2 \\
&\leq \| \int \mu\,(dy) \left[\frac{\nabla K(x,y) - \nabla K(x',y)}{\rho K(y)} \right] \|_2 \\
&+ \| \int \mu\,(dy)\, \frac{\nabla K(x',y)}{\rho' K(y) \rho K(y)} \left[\rho' K(y) - \rho K(y) \right] \|_2
\end{aligned}$$

The first term is bounded under (A2)

$$\begin{aligned}
\| \int \mu\,(dy) \left[\frac{\nabla K(x,y) - \nabla K(x',y)}{\rho K(y)} \right] \|_2 &\leq L' \|x' - x\|_2 \left| \int \frac{\mu\,(dy)}{\rho K(y)} \right| \\
&\leq L' \|x' - x\|_2 m_K
\end{aligned}$$

and the last inequality follows from (A1). For the second term, take an optimal transport plan π between ρ, ρ', as defined in (4), and consider

$$\begin{aligned}
|\rho' K(y) - \rho K(y)| &= | \int \rho'(dx') K(x',y) - \int \rho(dx) K(x,y) | \\
&= | \int \pi(dx, dx') \left[K(x',y) - K(x,y) \right] | \\
&\leq \int \pi(dx, dx') \|K(x',y) - K(x,y)\|_2 \\
&\leq L \int \pi(dx, dx') \|x' - x\|_2 \\
&\leq L \left(\int \pi(dx, dx') \|x' - x\|_2^2 \right)^{1/2} \\
&= L W_2(\rho, \rho')
\end{aligned}$$

where the second inequality follows from (A1) and the second-to-last inequality is a consequence of Jensen's inequality. Then,

$$\begin{aligned}
\| \int \mu\,(dy)\, \frac{\nabla K(x',y)}{\rho' K(y) \rho K(y)} \left[\rho' K(y) - \rho K(y) \right] \|_2 &\leq m_K^2 \| \int \mu\,(dy)\, \nabla K(x',y) \left[\rho' K(y) - \rho K(y) \right] \|_2 \\
&\leq m_K^2 \int \mu(dy) \| \nabla K(x',y) \left[\rho' K(y) - \rho K(y) \right] \|_2 \\
&\leq m_K^2 m_K^2 \int \mu(dy) |\rho' K(y) - \rho K(y)| \|\nabla K(x',y)\|_2 \\
&\leq m_K^2 B L W_2(\rho, \rho') \int \mu(dy) \\
&\leq m_K^2 B L W_2(\rho, \rho').
\end{aligned}$$

Thus, the SDE (15) admits a unique strong solution (strong because the proof in Jourdain et al. (2007) is obtained through a contraction argument).

Existence and uniquess of the solution of (15) can also be obtained under slightly weaker assumptions on the drift, in particular Lipschitz continuity in ρ, local Lipschitz continuity and a growth condition w.r.t. x Dos Reis et al. (2019, Theorem 3.3). However, the lack of global Lipschitz continuity w.r.t. x influences the stability of time discretisation schemes, and the standard Euler scheme cannot be applied, alternative tamed schemes are described in Reis et al. (2018).

4.2 Ergodicity

Needs extra work

The generator of (15) is

$$\mathcal{L}u(x) = (\nabla u(x)) \cdot \int \mu\,(dy)\,\frac{\nabla K(x,y)}{\rho K(y)} + \frac{2\alpha}{2}\Delta u(x) \tag{16}$$

defined for all continuously twice differentiable u.
with adjoint

$$\mathcal{L}^\star\phi(x) = \nabla \cdot \left(\phi(x)\int \mu\,(dy)\,\frac{\nabla K(x,y)}{\rho K(y)}\right) - \frac{2\alpha}{2}\Delta\phi(x). \tag{17}$$

The first condition that we check is non-explosion. To do so we use Theorem 2.1 in (Meyn and Tweedie, 1993b) and we make the following additional assumption:

(A3) the gradient $\nabla K(x,y)$ and μ satisfy

$$\int \mu(dy)\nabla K(x,y)\cdot x \le a\|x\|_2^2 + b$$

for $a, b < \infty$.

This condition is satisfied by the 1D gaussian example as we obtain

$$\int \mu(dy)\nabla K(x,y) = \frac{(\mu-1)\sigma_K^2 + (x-1)\sigma_\nu^2}{\sigma_K^2(\sigma_K^2+\sigma_\nu^2)^2}\,\mathcal{N}(x;\mu,\sigma_K^2+\sigma_\nu^2)$$

and

$$\mathcal{N}(x;\mu,\sigma_K^2+\sigma_\nu^2)\cdot x^2 \le x^2 \text{ for all } x\in\mathbb{R} \qquad \text{and} \qquad \mathcal{N}(x;\mu,\sigma_K^2+\sigma_\nu^2)\cdot x \le x^2 \text{for all } |x| \ge \sqrt{W(2)}$$

where $W(z)$ is the Lambert W function.
Condition (CD0) in (Meyn and Tweedie, 1993b) holds with $V(x) = \|x\|_2^2/2$:

$$\begin{aligned}
\mathcal{L}V(x) &= \sum_{i=1}^n x_i \int \frac{\partial K(x,y)}{\partial x_i}\frac{\mu\,(dy)}{\rho K(y)} + \alpha\sum_{i=1}^n 1 \\
&= \int \mu(dy)\frac{\nabla K(x,y)\cdot x}{\rho K(y)} + \alpha n \\
&\le m_K \int \mu(dy)\frac{\nabla K(x,y)\cdot x}{\rho(\mathbb{R}^n)} + \alpha n \\
&\le m_K \int \mu(dy)\nabla K(x,y)\cdot x + \alpha n \\
&\le m_K(a\|x\|_2^2 + b) + \alpha n \\
&\le 2m_K a V(x) + m_K b + \alpha n.
\end{aligned}$$

Next: we want to show that 15 is Feller and irreducible.
Without additional assumptions, the drift is not bounded, but locally bounded as by (A2) is Lipschitz. We know that boundedness implies that X_t is strong Feller (Bhattacharya (1978, Theorem 2.1) and Stroock and Varadhan (Section 7, 1967)). But maybe local boundedness is

enough? This is what Roberts and Tweedie (1996) claim but I still don't know why. This should also imply that X_t is irreducible w.r.t. Lebsgue, but again I do not know why.

As a consequence of the Feller property and the irreducibility, by Tweedie (1994, Theorem 7.1) X_t is an irreducible T-model (the support of the Lebesgue measure is $\mathbb{R}^n$ which clearly has open interior). Hence, all compact sets are petite Tweedie (1994, Theorem 5.1). Since X_t is irreducible, so are all the skeleton chains (easy to see from the definition of occupation time) and thus, X_t is aperiodic by Meyn and Tweedie (1993a, Theorem 5.2).

Existence of an invariant $\pi$is given by Meyn and Tweedie (1993b, Theorem 4.4) if we can find $V \geq 0, c, d \geq 0, f \geq 1, C$ compact such that V is bounded on C and

$$\mathcal{L}V(x) \leq -df(x) + c\mathbf{1}_C(x).$$

If we can find $V \geq 1, c, d \geq 0, C$compact such that V is bounded on $C$and

$$\mathcal{L}V(x) \leq -dV(x) + c\mathbf{1}_C(x)$$

then we have V-uniform erdocity by Down et al. (1995, Theorem 5.2).

We know that any invariant distribution $\pi(x)$ must satisfy

$$\mathcal{L}^\star\pi(x) = \nabla \cdot \left(\pi(x) \int \mu\left(dy\right) \frac{\nabla K(x,y)}{\pi K(y)}\right) - \alpha\Delta\pi(x) = 0$$
$$\mathcal{L}^\star\pi(x) = \int \frac{\mu(dy)}{\pi K(y)} \left[\nabla\pi(x) \cdot \nabla K(x,y) + \pi(x)\Delta K(x,y)\right] - \alpha\Delta\pi(x) = 0.$$

Guess: $V(x) = \pi(x)^{-d}$ with $d \in (0,1)$. Then

$$\nabla V(x) = -d\pi(x)^{-(d+1)}\nabla\pi(x) = -d\pi(x)^{-d}\left(\pi(x)^{-1}\nabla\pi(x)\right)$$
$$\Delta V(x) = -d\pi(x)^{-d}\left(-(d+1)\pi(x)^{-2}\|\nabla\pi(x)\|_2^2 + \pi(x)^{-1}\Delta\pi(x)\right)$$

and

$$\begin{aligned}
\mathcal{L}V(x) &= -d\pi(x)^{-d}\pi(x)^{-1}\int \frac{\mu(dy)}{\pi K(y)}\nabla K(x,y)\cdot\nabla\pi(x) - \alpha d\pi(x)^{-d}\left(-(d+1)\pi(x)^{-2}\|\nabla\pi(x)\|_2^2 + \pi(x)^{-1}\Delta\pi(x)\right) \\
&= -d\pi(x)^{-d}\left[\pi(x)^{-1}\int \frac{\mu(dy)}{\pi K(y)}\nabla K(x,y)\cdot\nabla\pi(x) - \alpha(d+1)\pi(x)^{-2}\|\nabla\pi(x)\|_2^2 + \alpha\pi(x)^{-1}\Delta\pi(x)\right] \\
&= -d\pi(x)^{-d}\left[\pi(x)^{-1}\left(\int \frac{\mu(dy)}{\pi K(y)}\nabla K(x,y)\cdot\nabla\pi(x) + \alpha\Delta\pi(x)\right) - \alpha(d+1)\pi(x)^{-2}\|\nabla\pi(x)\|_2^2\right].
\end{aligned}$$

5 Numerical Implementation

The first step towards solving (15) is the introduction of a space discretisation, this is necessary since McKean-Vlasov SDEs present a dependence on the distribution ρ_t of X_t (McKean, 1966). We follow the approach of Bossy and Talay (1997) and consider N particles $(X_t^1, ..., X_t^N)$ such that, at initialization, we sample i.i.d. particles $X_0^i \sim \rho_0$ and then they evolve according to the non-linear SDE

$$dX_t^{i,N} = \int \mu\left(dy\right) \frac{\nabla K(X_t^{i,N}, y)}{\rho_t^N K(y)} dt + \sqrt{2\alpha} dW_t^i, \quad \rho_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{i,N}}. \tag{18}$$

where W_t^i for $i = 1, \ldots, N$ are N independent n-dimensional standard Brownian motions and ρ_t^N is the empirical measure given by the N particles. Existence and uniqueness of a solution of (18) is given in Protter (2005, Theorem 7, page 253), and Jourdain et al. (2007, Theorem 3) show that over a finite time interval $[0, T]$ the error introduced by the interacting particle system in (18) goes to 0 as the number of particles N increases

$$\lim_{N\to\infty} \sup_{i\leq N} \mathbb{E}\left[\sup_{t\leq T} |X_t^{i,N} - X_t^i|^2\right] = 0, \tag{19}$$

this result is commonly known as propagation of chaos (Sznitman, 1991).

In practice, the integral w.r.t. μ in the drift coefficient of (18) cannot be computed analytically, but can be approximated through the sample average

$$\int \mu^M(dy) \frac{\nabla K(X_t^{i,N}, y)}{\rho_t^N K(y)}, \qquad \mu^M = \frac{1}{M}\sum_{j=1}^M \delta_{Y^{j,M}}$$

with $Y^{j,M}$ for $j = 1, \ldots, M$ i.i.d. samples from μ. The propagation of chaos result does not apply if we consider

$$dX_t^{i,N,M} = \int \mu^M(dy) \frac{\nabla K(X_t^{i,N,M}, y)}{\rho_t^{N,M} K(y)} dt + \sqrt{2\alpha} dW_t^i, \quad \rho_t^{N,M} = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^{i,N,M}} \tag{20}$$

instead of $X_t^{i,N}$ in (18), we extend this result in Section (5.1).

As for all SDEs, we then introduce a time discretisation; the simple Euler scheme with discretisation step Δt

$$X_{n+1}^{i,N} = X_n^{i,N} + \int \mu(dy) \frac{\nabla K(X_n^{i,N}, y)}{\rho_n^N K(y)} \Delta t + \sqrt{2\alpha}\Delta W^i, \quad \rho_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_n^{i,N}}$$

with $\Delta W^i \sim \mathcal{N}(0, \Delta t)$ is well behaved if the drift and diffusion coefficient are smooth with bounded derivatives (Antonelli et al., 2002). Clearly this is true for the constant diffusion coefficient $\sqrt{2\alpha}$, which is also bounded below and therefore satisfies the Hörmander condition (H1) (Antonelli et al., 2002, page 428). Under the assumption that K and ∇K are smooth with bounded derivatives Antonelli et al. (2002, Theorem 3.1) gives the following estimates of the error in terms of time discretisation step Δt and number of particles N:

$$\int_{\mathbb{X}} \mathbb{E}\left[\left|\rho_t(x) - \frac{1}{N}\sum_{i=1}^N \phi_{\Delta t}(X_t^{i,N} - x)\right|\right] dx \leq C\left(\Delta t + \frac{1}{\sqrt{N}} + \frac{1}{\Delta t^{1/4}\sqrt{N}}\right)$$

where $\phi_{\Delta t}$ is a Gaussian kernel with variance Δt (i.e. a kernel density estimator with Gaussian kernel). If we choose $N = O(1/\Delta t)^k$ for some $k > 0$ we can get uniform bounds:

$$\sup_{x\in\mathbb{X}} \mathbb{E}\left[\left|\rho_t(x) - \frac{1}{N}\sum_{i=1}^N \phi_{\Delta t}(X_t^{i,N} - x)\right|\right] \leq C_p\left(h + \frac{1}{\sqrt{N}} + \frac{1}{\Delta t^{1-1/2p}\sqrt{N}}\right)$$

for all $p > 1$. If the drift coeffient presents super-linear growth, tamed Euler schemes can be employed (Reis et al., 2018).

The Milstein scheme in Bao et al. (2020) coincides with the Euler scheme above since the diffusion coefficient is constant. This scheme can be shown to have strong order of convergence 1 in time under some additional assumptions involving Lions derivatives of the drift and the diffusion coefficient.

5.1 Propagation of chaos

The propagation of chaos result (19) in Jourdain et al. (2007) can be easily extended to (20) provided that both N, M tend to infinity:

$$\lim_{N,M\to\infty} \sup_{i\leq N} \mathbb{E}\left[\sup_{t\leq T} |X_t^{i,N,M} - X_t^i|^2\right] = 0.$$

The proofs is similar in structure to that of Jourdain et al. (2007, Theorem 3)

Proof. The following decomposition allows us to treat the influence of N and M separately

$$\mathbb{E}\left[\sup_{t\leq T} \|X_t^{i,N,M} - X_t^i\|_2^2\right] \leq \mathbb{E}\left[\sup_{t\leq T} \|X_t^{i,N,M} - X_t^{i,N}\|_2^2\right] + \mathbb{E}\left[\sup_{t\leq T} \|X_t^{i,N} - X_t^i\|_2^2\right].$$

The result for the last expectation is given in (19). Lemma 1 of Jourdain et al. (2007) gives the following bound

$$\begin{aligned}
\mathbb{E}\left[\sup_{s\leq t} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] &\leq C\int_0^t \mathbb{E}\left[\| \int \mu^M(dy) \frac{\nabla K(X_s^{i,N,M}, y)}{\rho_s^{N,M} K(y)} - \int \mu(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} \|_2^2\right] ds \\
&\leq C\int_0^t \mathbb{E}\left[\| \int \mu^M(dy) \frac{\nabla K(X_s^{i,N,M}, y)}{\rho_s^{N,M} K(y)} - \int \mu^M(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} \|_2^2\right] ds \\
&+ C\int_0^t \mathbb{E}\left[\| \int \mu^M(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} - \int \mu(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} \|_2^2\right] ds.
\end{aligned}$$

The first integral is bounded using the Lipschitz continuity of the drift as in the existence and uniqueness argument in Section (4.1), for $D_1, D_2 < \infty$

$$\begin{aligned}
&C\int_0^t \mathbb{E}\left[\| \int \mu^M(dy) \frac{\nabla K(X_s^{i,N,M}, y)}{\rho_s^{N,M} K(y)} - \int \mu^M(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} \|_2^2\right] ds \\
&\leq C\int_0^t \mathbb{E}\left[D_1 \|X_s^{i,N,M} - X_s^{i,N}\|_2^2 + D_2 W_2(\rho_s^{N,M}, \rho_s^N)^2\right] ds \\
&\leq C\int_0^t \mathbb{E}\left[D_1 \|X_s^{i,N,M} - X_s^{i,N}\|_2^2 + D_2 \frac{1}{N} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds \\
&\leq C\int_0^t \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds
\end{aligned}$$

where the second-to-last inequality follows from Jourdain et al. (2007, page 5).

For the second term, we exploit the properties of sample means of i.i.d. samples

$$\begin{aligned}
&\mathbb{E}\left[\| \int \mu^M(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} - \int \mu(dy) \frac{\nabla K(X_s^{i,N}, y)}{\rho_s^N K(y)} \|_2^2\right] \\
&= \mathbb{E}\left[\frac{1}{M} \operatorname{Var}\left(\frac{\nabla K(X_s^{i,N}, \cdot)}{\rho_s^N K(\cdot)} \mid \sigma(X_s^{i,N}, i = 1:N)\right)\right] \leq \frac{1}{M} B^2 m_K^2.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\sup_{s\leq t} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] \leq C\int_0^t \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T$$

by Gronwall's Lemma applied to

$$u(t) = \int_0^t \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T$$

we obtain

$$\int_0^t \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T \leq u(0)\exp(Ct) = 0$$

for all $t < T$ since $X_0^{i,N,M} = X_0^{i,N} = X_0^i$. For $t = T$ we can write

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\leq T} \|X_t^{i,N,M} - X_t^{i,N}\|_2^2\right] \\
&\leq \int_0^T \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T \\
&= \int_0^{T-\varepsilon} \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \int_{T-\varepsilon}^T \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T \\
&= \int_{T-\varepsilon}^T \mathbb{E}\left[\sup_{r\leq s} \|X_s^{i,N,M} - X_s^{i,N}\|_2^2\right] ds + \frac{C}{M} B^2 m_K^2 T
\end{aligned}$$

as $\varepsilon \to 0$ we have

$$\mathbb{E}\left[\sup_{t\leq T} \|X_t^{i,N,M} - X_t^{i,N}\|_2^2\right] \leq \frac{C}{M} B^2 m_K^2 T$$

which tends to 0 as $M \to \infty$. The above and (19) give the result.I haven't looked at this in detail yet, but it's a nice result. It is perhaps interesting that the size of the error is O(1/M) whereas Monte Carlo error will be O(1/N), presumably, which gives some guidance on balancing costs where both M and N can be specified by the user. AMJ □

6 Examples

We now consider some examples. The first one is a 1D toy example which we use to get some insight on the choice of number of particles N, discretisation step Δt, initial distribution ρ_0 and regularisation parameter α.

6.1 Toy Example

We consider the toy Fredholm integral equation

$$\mathcal{N}(y; m, \sigma_\mu^2 := \sigma_K^2 + \sigma_\rho^2) = \int_{\mathbb{X}} \mathcal{N}(x; m, \sigma_\rho^2)\mathcal{N}(y; x, \sigma_K^2)\, dx, \qquad y \in \mathbb{Y}$$

with $\mathbb{X} = \mathbb{Y} = [0, 1]$. For this toy example we find the minimiser of (2), under the assumption that such a minimiser is $\rho_\alpha = \mathcal{N}(m, \sigma_\alpha^2)$, a Normal distribution with mean m (equal to those of the data distribution μ) and variance depending on α

$$\sigma_\alpha^2 = \frac{-(\sigma_K^2 - \sigma_\mu^2 - 2\alpha\sigma_K^2) + \sqrt{\sigma_K^4 + \sigma_\mu^4 - 2\sigma_K^2\sigma_\mu^2(1-2\alpha)}}{2(1-\alpha)}.$$

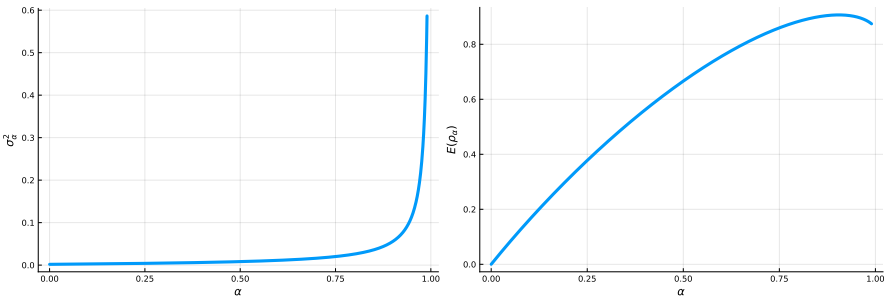

Figure 1: Variance and E as functions of $\alpha \in [0, 0.99]$ for the toy example.

It is clear that when $\alpha = 0$ (no entropy constraint), $\sigma_\alpha^2 = \sigma_\rho^2$, and that $\alpha > 1$ give negative variance. In particular we have

$$E(\rho_\alpha) = \frac{1}{2}\log\frac{\sigma_\alpha^2 + \sigma_K^2}{\sigma_\mu^2} + \frac{\sigma_\mu^2}{2(\sigma_\alpha^2 + \sigma_K^2)} - \frac{1}{2} - \alpha\left(\frac{1}{2} + \frac{1}{2}\log(2\pi\sigma_\alpha^2)\right).$$

The functional dependence of σ_α^2 and $E(\rho_\alpha)$ on α is shown in Figure 6.1.

Since the gradient flow PDE (12) admits a unique solution for each starting condition ρ_0, we analyse the influence of the starting distribution for this simple toy example with $m = 0.5$, $\sigma_\rho^2 = 0.043^2$, $\sigma_K^2 = 0.045^2$. We consider 6 initial distributions: three Dirac δs centred at 0, 0.5 and 1 respectively, a Uniform on $[0, 1]$, the solution $\mathcal{N}(x; m, \sigma_\rho^2)$ and a more dispersed Gaussian $\mathcal{N}(x; m, \sigma_\rho^2 + \varepsilon)$. Figure 2 shows that better results are obtained when ρ_0 is a point mass or a Gaussian distribution, this is coherent with the observations of Bossy and Talay (1997); Antonelli et al. (2002).

The choice $\rho_0 \sim U([0, 1])$ results in slower convergence, indeed for 1000 iterations the values of MISE and E are still far from those of the solution (Figure 3).

Now that we have a good intuition for the choice of ρ_0 we investigate the influence of α. We set $\Delta t = 10^{-3}$ and we compare reconstructions through variance σ_α^2, MISE, $E(\rho)$ and the entropy $\mathrm{ent}(\rho)$ after 100 iterations. We use $N = 500, 1000, 5000$ and 1000 repetitions for each α. The initial distribution ρ_0 is a point mass concentrated at a random point in $[0, 1]$. Figure 4 shows that the dependence of σ_α^2, MISE and $E(\rho)$ on α is mostly linear for all Ns considered. To further investigate we zoom in to $\alpha \in [0, 0.1]$ (Figure 5). Values smaller than 0.25 give the smallest values of E.

Figure 6.1 shows the reconstruction of $\mathcal{N}(x; m, \sigma_\rho^2)$ with $m = 0.5$, $\sigma_\rho^2 = 0.043^2$, $\sigma_K^2 = 0.045^2$ with $N = 1000$, $\Delta t = 10^{-3}, 100$ iterations, $\rho_0 = \delta_{0.5}$. On the left hand side we compare the reconstruction for $\alpha = 0.01$ with the corresponding exact minimiser of (2) given by $\mathcal{N}(x; m, \sigma_\alpha^2)$. On the right hand side we compare different values of α.

6.2 Comparison with SMC

We now compare the WGF approach with the SMC one. Parameter setting:

- SMC: Niter $= 100$, $\varepsilon = 10^{-3}$
- WGF: $\Delta t = 10^{-3}, \alpha = 0.01$, Niter $= 100$

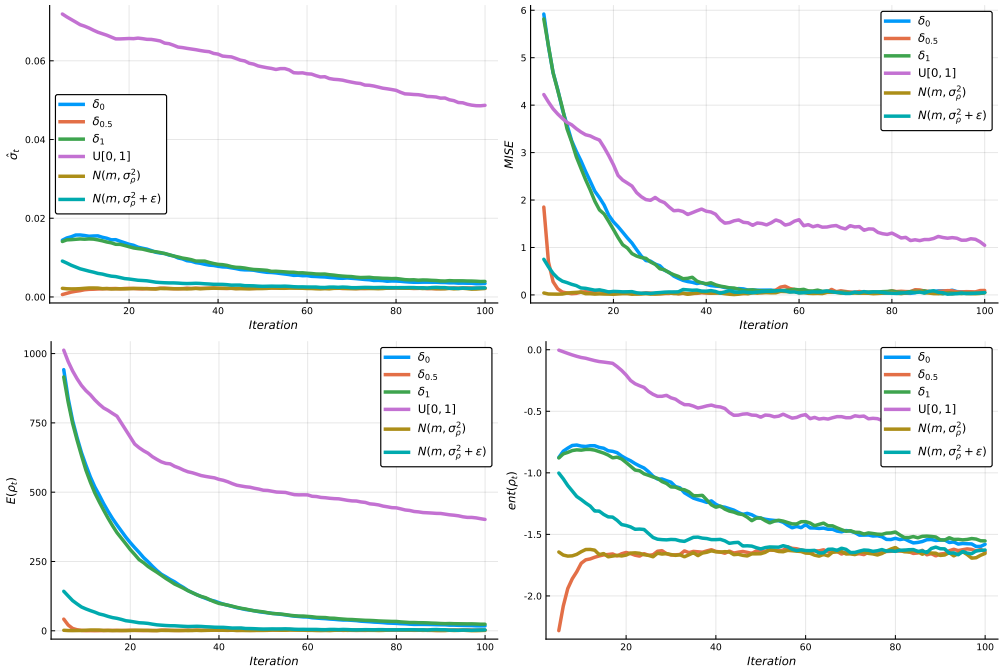

Figure 2: Effect of initial distribution for the toy example. $N = 1000, dt = 10^{-3}, \alpha = 0.025, 100$ iterations.

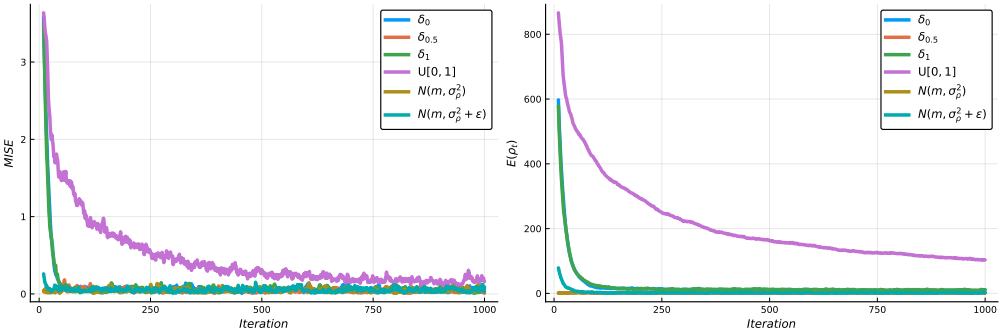

Figure 3: Effect of initial distribution for the toy example. $N = 1000, dt = 10^{-3}, \alpha = 0.025, 1000$ iterations.

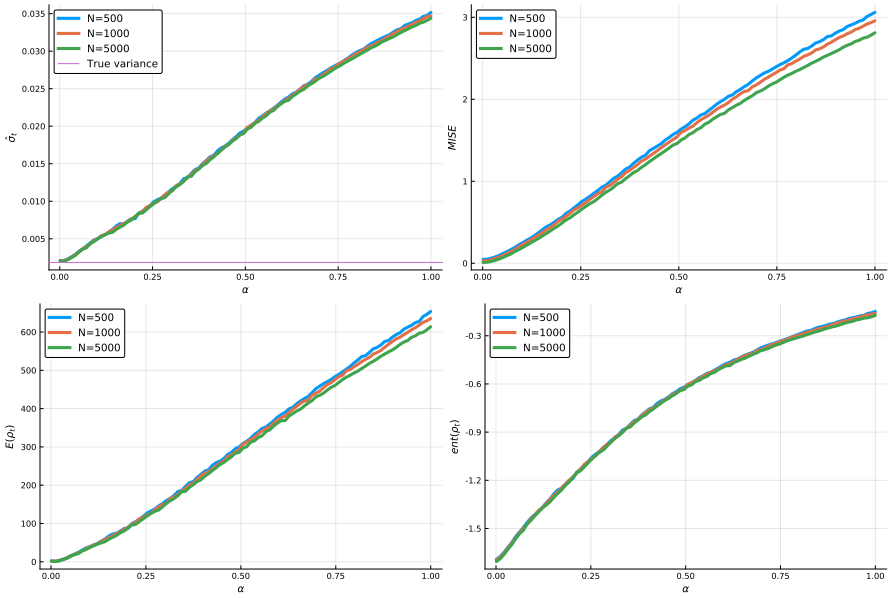

Figure 4: $\Delta t = 10^{-3}, \alpha \in [0, 1]$, 100 iterations, 1000 repetitions for each α

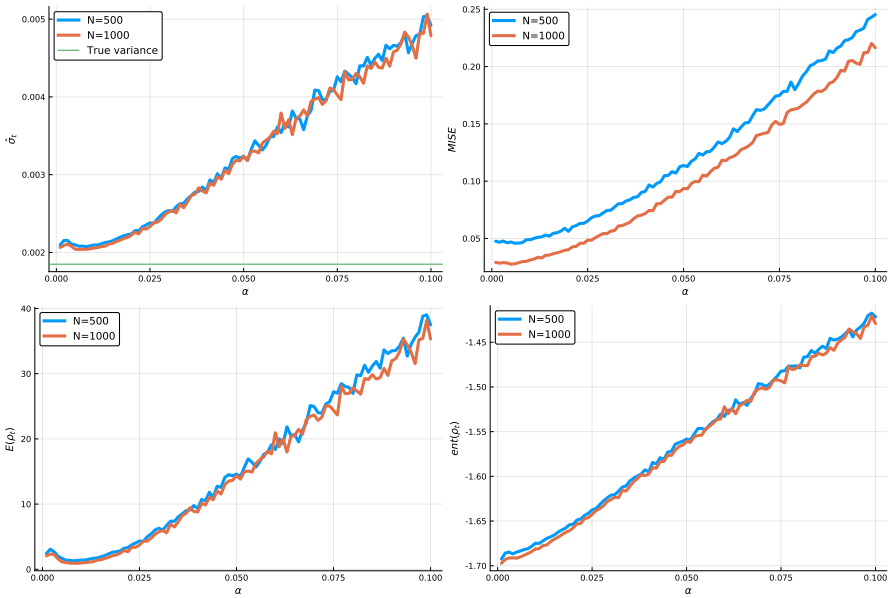

Figure 5: $\Delta t = 10^{-3}, \alpha \in [0, 0.1]$, 100 iterations, 1000 repetitions for each α

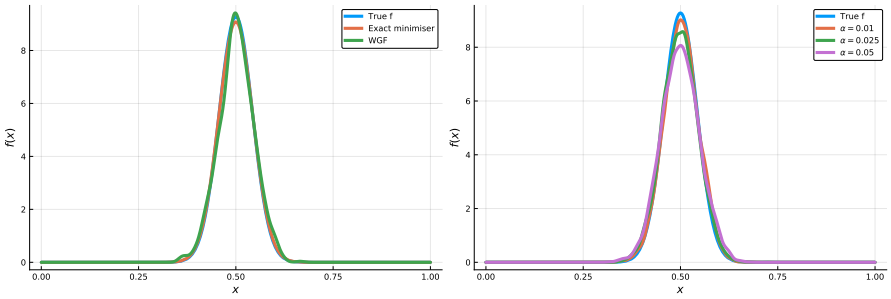

Figure 6: $N = 1000, \Delta t = 10^{-3}$, 100 iterations.

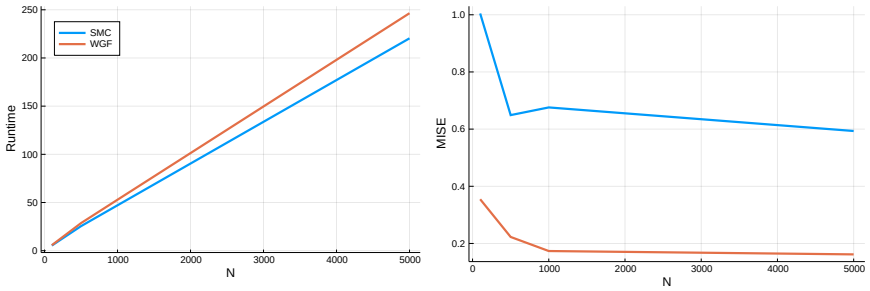

Figure 7: Comparison of SMC and WGF

7 Gaussian Mixture 1D

This is the indirect density estimation example with

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

Figure 7 shows the reconstruction of $\mathcal{N}(x; m, \sigma_\rho^2)$ with $m = 0.5$, $\sigma_\rho^2 = 0.043^2$, $\sigma_K^2 = 0.045^2$. We set $N = 5000$, $\alpha = 0.005$, $\Delta t = 10^{-3}$. The initial distribution ρ_0 is a point mass concentrated at 0.5. We also compare different values of α.

8 Multivariate Gaussian

Take

$$\mathcal{N}(y; \mathbf{m}, \Sigma_\mu^2 := \Sigma_K^2 + \Sigma_\rho^2) = \int \mathcal{N}(x; \mathbf{m}, \Sigma_\rho^2)\mathcal{N}(y; x, \Sigma_K^2)\, dx, \qquad y \in \mathbb{R}^2$$

with $\mathbf{m} = (0,0)^T, \Sigma_\rho^2 = 0.2 \cdot Id, \Sigma_K^2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

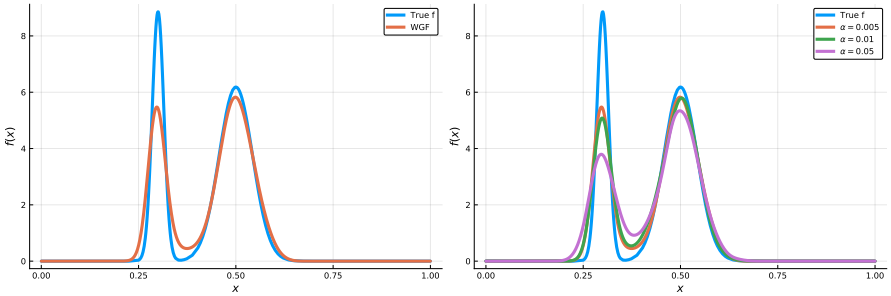

Figure 8: $N = 5000, dt = 10^{-3}, \alpha = 0.005$

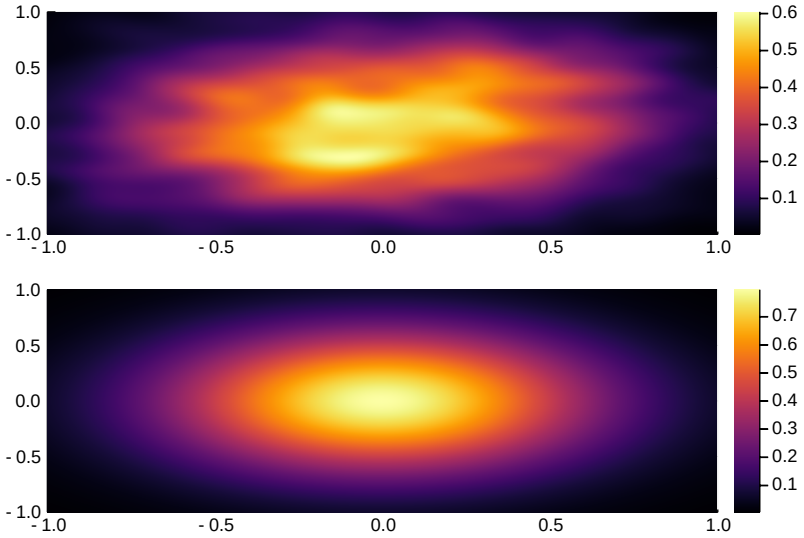

Figure 9: $N = 5000, dt = 10^{-2}, T = 1, \alpha = 0.005$, 1500 iterations

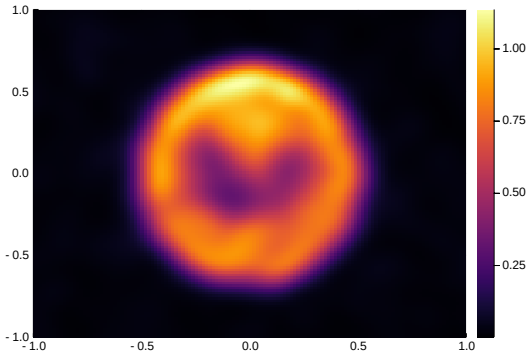

Figure 10: $N = 5000, dt = 10^{-3}, \alpha = 0.1, 1000$ iterations.

9 PET

In this case we have the sinogram data as input data $h(\phi, \xi)$ and the kernel $K(\phi, \xi \mid x, y) = \mathcal{N}(x \cos\phi + y \sin\phi - \xi; 0, \sigma^2)$ for σ^2 small. ξ gives the displacement from the centre and ϕ the angle.

References

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008. 2.2, 2.2, 2.2, 2.2, 3.1, 3.2

F. Antonelli, A. Kohatsu-Higa, et al. Rate of convergence of a particle method to the solution of the McKean–Vlasov equation. *The Annals of Applied Probability*, 12(2):423–476, 2002. 5, 6.1

M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*, pages 6481–6491, 2019. 3.2

J. Bao, C. Reisinger, P. Ren, and W. Stockinger. First order convergence of Milstein schemes for mckean equations and interacting particle systems. *arXiv preprint arXiv:2004.03325*, 2020. 5

R. Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability*, 6(4):541–553, 1978. 4.2

M. Bossy and D. Talay. A stochastic particle method for the McKean-Vlasov and the Burgers equation. *Mathematics of Computation*, 66(217):157–192, 1997. 5, 6.1

J. A. Carrillo, R. J. McCann, and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Archive for Rational Mechanics and Analysis*, 179(2): 217–263, 2006. 3.2, 3.2

K. Craig. Nonconvex gradient flow in the Wasserstein metric and applications to constrained nonlocal interactions. *Proceedings of the London Mathematical Society*, 114(1):60–102, 2017. 3

G. Dos Reis, W. Salkeld, J. Tugaut, et al. Freidlin–Wentzell LDP in path space for McKean–Vlasov equations and the functional iterated logarithm law. *The Annals of Applied Probability*, 29(3):1487–1540, 2019. 3, 4.1

D. Down, S. P. Meyn, and R. L. Tweedie. Exponential and uniform ergodicity of Markov processes. *The Annals of Probability*, pages 1671–1691, 1995. 4.2

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. 4

B. Jourdain, S. Méléard, and W. Woyczynski. Nonlinear SDEs driven by L\'evy processes and related PDEs. *arXiv preprint arXiv:0707.2723*, 2007. 4.1, 5, 5.1

A. Kondor. Method of convergent weights—An iterative procedure for solving Fredholm's integral equations of the first kind. *Nuclear Instruments and Methods in Physics Research*, 216(1-2): 177–181, 1983. 4

H. McKean. A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences of the United States of America*, 56(6):1907–1911, 1966. 4, 5

S. P. Meyn and R. L. Tweedie. Stability of Markovian processes II: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993a. 4.2

S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548, 1993b. 4.2, 4.2

P. Protter. Stochastic Integration and Differential Equations. *Stochastic Modelling and Applied Probability*, 21, 2005. 5

G. d. Reis, S. Engelhardt, and G. Smith. Simulation of McKean Vlasov SDEs with super linear growth. *arXiv preprint arXiv:1808.05530*, 2018. 3, 4.1, 5

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. 4.2

F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017. 2.2, 3, 3.1, 3.2

D. W. Stroock and S. S. Varadhan. Diffusion processes with continuous coefficients, II. *Communications on Pure and Applied Mathematics*, 22(4):479–530, 1967. 4.2

A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991. 5

R. L. Tweedie. Topological conditions enabling use of Harris methods in discrete and continuous time. *Acta Applicandae Mathematica*, 34(1-2):175–188, 1994. 4.2