Solving Fredholm integral equations of the first kind via Wasserstein gradient flow

Francesca Crucinio, Arnaud Doucet and Adam M. Johansen

1 Fredholm integral equation of first kind

We want to solve the integral equation

$$\mu(y) = \int \rho(x)\, K(x,y) dx,$$

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

$$E(\rho) = \mathrm{KL}(\mu, \rho K) - \alpha\, \mathrm{Ent}(\rho) \tag{1}$$

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.

1.1 Assumptions

(A0) 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 We've assumed that the spaces are $\mathbb{R}^n$and $\mathbb{R}^m$, respectively, but perhaps we ought to make them subsets of these to allow consistency with this assumption? AMJ

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

is $\lambda-$concave in x, uniformly in y, for some $\lambda \in \mathbb{R}$ and is Lipschitz continuous in x, uniformly in y, with Lipschitz constant L. Should we give explicit expressions for these? AMJ

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

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 C^2 functions Somewhere we should define C^2 and λ-concavity. AMJ 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).

2 Gradient flow approach

2.1 Notation

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

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

where $\Pi(\mu, \nu)$ is the set of all possible couplings between μ and ν. 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 exists a unique optimal transport map t_μ^ν (see, for example, Ambrosio et al., 2008, page 150). It is easy to check that for all $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ the entropy $\mathrm{Ent}(\nu)$ is finite Is it? It's not $+\infty$ because its bounded above by the variance of a normal of the same variance, but does anything stop it reaching $-\infty$? AMJ.

We now give some useful definitions.

A functional F defined on $\mathcal{P}_2^{ac}(\mathbb{R}^d)$ is λ-geodesically convex if Presumably this uses the fact that there exists a unique optimal transport map in the space and so the general definition simplifies, but perhaps this needs to be made a bit more explicit? AMJ

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

for all $\nu, \mu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ (Ambrosio et al., 2008, page 202). If the above holds true for $\lambda = 0$, the functional is called displacement convex.

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

$$F(\nu) - F(\mu) \geq \int \langle \xi(x), t_\mu^\nu(x) - x \rangle d\mu(x) + o\left(W_2(\nu, \mu)\right)$$

for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2008, Definition 10.1.1). This is perhaps pedantic, but we should probably be explicit about the small-o notation; particularly that it's being used in a limit as $||\nu - \mu||$ vanishes. AMJ

2.2 Properties of the functional E

We consider the functional E in (1) defined on $\mathcal{P}_2^{ac}(\mathbb{R}^d)$. The image of this functional is $(-\infty, +\infty]$, since the KL divergence is always positive and the entropy of both μ and ρ is finiteDoes the entropy of \mu have any significance here? Also, is the entropy of \rho really

necessarily finite rather than just $\neq +\infty$, which suffices to obtain the claim here, anyway? . This functional is proper, i.e. there exists at least one ρ such that $E(\rho) < +\infty$ (e.g. take ρ to be uniform on some subset of $\mathbb{R}^n$, the entropy is finite and the boundedness of K ensures that the KL divergence is finite too) I don't follow this example; doesn't μ enter into which ρ will lead to finite KL divergences? AMJ.

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

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

$$\begin{aligned} |E(\rho_n) - E(\rho)| &= \left| -\int \mu \log \rho_n K + \alpha \int \rho_n \log \rho_n + \int \mu \log \rho K - \alpha \int \rho \log \rho \right| \\ &\leq \left| \int \mu \left[\log \rho K - \log \rho_n K \right] \right| + \alpha \left| \int \left[\rho_n \log \rho_n - \rho \log \rho \right] \right| \end{aligned}$$

if K is continuousIs continuity enough or do we need it to be Feller? AMJ, since $\rho_n \rightharpoonup \rho$, we also have that $\rho_n K \to \rho K$ and the continuity of the logarithm gives $\log \rho K \to \log \rho_n K$. The dominated convergence theorem gives

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

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 $(\mathcal{P}_2^{ac}(\mathbb{R}^n), W_2)$. Should we try to connect this with the normal definition of coercivity? AMJ*

Proof. Following Definition 2.1b in Ambrosio et al. (2008, page 43) we want to show that there exist $\tau > 0$ and $\nu \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ such that

$$\inf_{\rho \in \mathcal{Q}} \frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) > -\infty.$$

Under (A1) we have that

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

Since the entropy of μ is finite, $\int \mu \log \mu > C > -\infty$, we have that

$$\begin{aligned} \inf_{\rho \in \mathcal{Q}} \frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) &\geq \inf_{\rho \in \mathcal{Q}} \frac{1}{2\tau} W_2^2(\nu, \rho) + \alpha \int \rho \log \rho + \int \mu \log \mu \\ &\geq \inf_{\rho \in \mathcal{Q}} \frac{1}{2\tau} W_2^2(\nu, \rho) + \alpha \int \rho \log \rho + C. \end{aligned}$$

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

$$\inf_{\rho \in \mathcal{Q}} \frac{1}{2\tau} W_2^2(\nu, \rho) + E(\rho) \geq C + \alpha \inf_{\rho \in \mathcal{Q}} \int \rho \log \rho > -\infty$$

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

Proposition 3. *The functional E is displacement convex.*

Proof. We have

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

The integral $\int \mu \log \mu$ is constant w.r.t. ρ and the negative entropy $\alpha \int \rho \log \rho$ is displacement convex in ρ (Santambrogio, 2017, page 130).
Regarding the integral $-\int \mu \log \rho K$, we have that the functional $F : \rho \mapsto \rho K(y) = \int K(x, y)\, d\rho(x)$ is $\lambda-$geodesically convex if K is $\lambda-$geodesically convex in x (Santambrogio, 2017, page 128). In particular, we have that for all $s \in [0, 1]$

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

If $\lambda > 0$, this is stronger than convexity, thus F is also displacement convex (i.e. $\lambda-$geodesically convex with $\lambda = 0$)

$$F\left(((1-s)Id + st_\rho^\nu)_\# \rho\right) \leq (1-s)F(\rho) + sF(\nu),$$

for $\lambda < 0$ it is weaker than convexity (Santambrogio, 2017, page 91). I rather lost the thread of the argument at this point; perhaps a bit more signposting would be helpful. Is it clear that K has to be either convex or concave? AMJ It is easy to see that if K is $\lambda-$geodesically concave, then

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

Thus, if $\lambda > 0$, $F\left(((1-s)Id + st_\rho^\nu)_\# \rho\right) \geq (1-s)F(\rho) + sF(\nu)$ and

$$\begin{aligned} &-\log\left(F\left(((1-s)Id + st_\mu^\nu)_\# \rho\right)\right) \\ &\leq -\log\left((1-s)F(\rho) + sF(\nu) + \frac{\lambda}{2} s(1-s) W_2^2(\nu, \rho)\right) \\ &\leq -\log\left((1-s)F(\rho) + sF(\nu)\right) \\ &\leq (1-s)\left(-\log F(\rho)\right) + s\left(-\log F(\nu)\right) \end{aligned}$$

since $-\log x$ is a convex decreasing function. The functional E is displacement convex as integrating w.r.t μ does not change the inequality. □

2.3 Gradient flow

Following Ambrosio et al. (2008, Section 11.1.2), ρ_t is a solution of the gradient flow equation for E with the W_2 metric if

$$\partial_t \rho_t = -\nabla \cdot (\rho_t v_t)$$

with

$$v_t \in -\partial E(\rho_t)$$

where $\|v_t\|_{\mathbb{L}_2(\rho_t)}$ is locally integrable. The second line states that v_t is a sub-differential for E at Isn't it the negative of a sub-differential? AMJρ_t.

2.3.1 First variation

To arrive to an expression for v_t we compute the first variation of E It wouldn't hurt to define a bit more of the notation here; it's not immediately apparent what is defined in terms of what... AMJ

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

where χ is any signed measure such that $\rho + \epsilon\chi$ is a probability measure, and then we need to show that this is a sub-differential for E.

We have

$$\begin{aligned} E(\rho) &= \mathrm{KL}(\mu, \rho K) - \alpha \operatorname{Ent}(\rho) \\ &= -\int \mu \log \rho K + \alpha \int \rho \log \rho + \int \mu \log \mu. \end{aligned}$$

We know that

$$\frac{-\delta \operatorname{Ent}}{\delta \rho}(\rho) = 1 + \log \rho$$

is a sub-differential for $-\operatorname{Ent}(\rho)$ (Carrillo et al., 2006, Lemma 8).
The first variation of the KL divergence as a function of ρ is given by

$$\begin{aligned} \mathrm{KL}(\mu, (\rho + \epsilon\chi) K) - \mathrm{KL}(\mu, \rho K) &= -\int \mu \log \left((\rho + \epsilon\chi) K \right) + \int \mu \log (\rho K) \\ &= -\int \mu \log \left(1 + \frac{\epsilon \chi K}{\rho K} \right) \\ &= -\int \mu \left(\frac{\epsilon \chi K}{\rho K} + o\left(\frac{\epsilon \chi K}{\rho K} \right) \right) \\ &= -\epsilon \int \mu \frac{\chi K}{\rho K} + o\left(\epsilon \int \mu \frac{\chi K}{\rho K} \right). \end{aligned}$$

We have Mixing notation for integration in a single expression seems unnecessarily cruel to the reader... is there a reason not to write $\chi(dx)$ rather than $d\chi(x)$ here? This type of thing happens a few times, it wasn't clear to me whether something was being emphasized or we might perhaps be able to slightly simplify the presentation. AMJ

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

so The abuse of notation in what follows isn't ideal; can we do anything to avoid it? AMJ

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

Next, we show that $-\int \mu(dy) \frac{K(x,y)}{\rho K(y)}$ is a sub-differential for $\mathrm{KL}(\mu, \rho K)$. Take $\rho, \nu \in \mathcal{P}_2^{ac}(\mathbb{R}^n)$

and consider

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

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

$$\nu M(y)-\rho M(y)=\int\langle M(x,y),t_\rho^\nu(x)-x\rangle d\rho(x)+o\left(W_2(\nu,\rho)\right).$$

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

$$-\nu K(y)+\rho K(y)=\int\langle -K(x,y),t_\rho^\nu(x)-x\rangle d\rho(x)+o\left(W_2(\nu,\rho)\right).$$

This yields

$$\begin{aligned}
&\mathrm{KL}(\mu,\nu K)-\mathrm{KL}(\mu,\rho K)-\int\langle\frac{\delta\,\mathrm{KL}}{\delta\rho}(x),t_\rho^\nu(x)-x\rangle d\rho(x)\\
&=\int\mu(dy)\left[-\log\frac{\nu K(y)}{\rho K(y)}+\frac{-\nu K(y)+\rho K(y)+o\left(W_2(\nu,\rho)\right)}{\rho K(y)}\right]\\
&=\int\mu(dy)\left[-\log\frac{\nu K(y)}{\rho K(y)}-\frac{\nu K(y)}{\rho K(y)}+1+o\left(W_2(\nu,\rho)\right)\right].
\end{aligned}$$

Now let $o\left(W_2(\nu,\rho)\right)\to 0$. It follows that $\nu K(y)\to\rho K(y)$ for all $y\in\mathbb{R}^m$,by continuity of K. The Taylor expansion of $\log x$ at $x=1$ then gives $-\log\frac{\nu K(y)}{\rho K(y)}-\frac{\nu K(y)}{\rho K(y)}+1=o\left(W_2(\nu,\rho)\right)$ and therefore

$$\begin{aligned}
\mathrm{KL}(\mu,\nu K)-\mathrm{KL}(\mu,\rho K)-\int\langle\frac{\delta\,\mathrm{KL}}{\delta\rho}(x),t_\rho^\nu(x)-x\rangle d\rho(x)&=\int\mu(dy)o\left(W_2(\nu,\rho)\right)\\
&=o\left(W_2(\nu,\rho)\right).
\end{aligned}$$

Hence, it follows that It might help the reader if you explain how what remains follows... you compute the first variation and establish that it's a sub-differential, but there seems to be a bit of a gap between that and the conclusion (which I think you've mentioned before is filled by standard arguments, but it would be good to make that explicit). AMJ

$$\frac{\delta E}{\delta\rho}(x)=-\int\mu(dy)\frac{K(x,y)}{\rho K(y)}+\alpha\left(1+\log\rho(x)\right)\tag{2}$$

is a sub-differential for E and the gradient flow ρ_t is a solution of

$$\partial_t\rho_t=\nabla\cdot\left(\rho_t\nabla\frac{\delta E}{\delta\rho_t}\right),\tag{3}$$

2.3.2 Existence and uniqueness of the gradient flow

Theorem 11.1.6 and Corollary 11.1.8 in Ambrosio et al. (2008) give existence of a gradient flow solution of (3) since E is continuous and coercive and the first variation (2) is single-valued. Thus, for every $\rho_0 \in \mathcal{P}_2^{ac}(\mathbb{R}^d)$ there exist a solution to the gradient flow equation (3). Uniqueness follows straightforwardly from Theorem 11.1.4 in Ambrosio et al. (2008) since the functional E is displacement convex in ρ. In particular we have the estimate This looks odd at first glance, but on reflection it's just the statement that the flow is non-expansive in W_2, isn't it? Perhaps we could tell people that a bit more directly. Is it possible to obtain strict contraction if $\rho_0^1 \rho_0^2$?AMJ

$$W_2(\rho_t^1, \rho_t^2) \leq W_2(\rho_0^1, \rho_0^2)$$

for all $t > 0$,for ρ_t^i solution of the gradient flow equation with initial condition ρ_0^i, $i = 1, 2$.

2.4 PDE

(3) is a Fokker-Plank/Kolmogorov forward equation with corresponding nonlinear ODE (Jordan et al., 1998)

$$dX_t = -\nabla \frac{\delta E}{\delta \rho_t}(X_t)\, dt, \quad X_0 \sim \rho_0 \tag{4}$$

such that $\text{Law}(X_t) = \rho_t$. The terminology nonlinear ODE is here used to indicate that the drift depends not only on X_t but on its distribution too.

Then, by construction (see Arbel et al. (2019, page 14)), one has the the infinitesimal reduction of E along ρ_t is

$$\frac{dE(\rho_t)}{dt} = -\int \left\| \nabla \frac{\delta E}{\delta \rho_t}(x) \right\|^2 d\rho_t(x). \tag{5}$$

Practically, what we would like to do is to simulate N particles $(X_t^1, ..., X_t^N)$ such that, at initialization, we sample iid particles $X_0^i \sim \rho_0$ and then implement numerically the N nonlinear ODEs Somehow this isn't really N nonlinear ODEs so much as a coherent system of approximate solutions for different initial conditions... AMJ

$$dX_t^i = \int \mu(dy) \frac{\nabla K(X_t^i, y)}{\rho_t^N K(y)} dt - \alpha \nabla \log\left(\rho_t^N * H_\epsilon\left(X_t^i\right)\right) dt, \quad \rho_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{X_t^i}.$$

Approximating the first term on the r.h.s. is fine as practically $\mu(dy)$ is a discrete measure What exactly is μ here? AMJbut approximating $\nabla \log \rho_t(x)$ from the empirical measure ρ_t^N is difficult and would require say convolution by some kernel H_ϵ. This is ugly and would be most likely highly inefficient. In the next section, we show how to address this issue.

Also, standard existence/uniqueness theorems do not apply because the drift $-\nabla \frac{\delta E}{\delta \rho_t}(X_t)$ is not Lipschitz continuous.

3 Nonlinear SDE approach and numerical implementation

We can now compute the gradient of this functional derivative equation w.r.t. x

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

Swapping integral and gradient is not a problem because ∇ is over x and the integral is over y. There's still a need to be slightly careful. If $f : (x, y) \mapsto 1$ then $\nabla \int f(x, y)dy$ is not well defined whereas $\int \nabla f(x, y)dy = 0$. AMJ

Hence (3) is equivalent to

$$\begin{aligned}\partial_t \rho_t =& \nabla \cdot \left(\rho_t \nabla \frac{\delta E}{\delta \rho_t} \right) \\ =& - \nabla \cdot \left(\rho_t \int \mu \left(dy \right) \frac{\nabla K(x, y)}{\rho_t K(y)} \right) + \alpha \nabla \cdot \left(\rho_t \nabla \log \rho_t \right).\end{aligned}$$

However, we have

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

where $\triangle f = \sum_i \partial_i^2 f_i$ is the Laplacian, so that the PDE above becomes

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

So we can consider the following non-linear SDE (McKean-Vlasov)

$$dX_t = \int \mu \left(dy \right) \frac{\nabla K(X_t, y)}{\rho_t K(y)} dt + \sqrt{2\alpha} dW_t, \quad X_0 \sim \rho_0, \tag{6}$$

where W_t is a standard $n-$dimensional Brownian motion. The SDE (6) has the same marginal distributions as the nonlinear ODE (4) (Itô's Lemma) This seems to need a bit more explanation... does an ODE have a marginal distribution? Under random initial conditions, perhaps... and it would be good to give a reference for this claim unless we show it explicitly. AMJ. So to solve the minimization problem of interest, we will simulate in practice N particles $(X_t^1, ..., X_t^N)$ such that, at initialization, we sample iid particles $X_0^i \sim \rho_0$ and then they evolve according to the non-linear (McKean-Vlasov) SDE Should we note that the W_t^i are independent standard Wiener processes? AMJ

$$dX_t^{i,N} = \int \mu \left(dy \right) \frac{\nabla K(X_t^{i,N}, y)}{\rho_t^N K(y)} dt + \sqrt{2\alpha} dW_t^i, \quad \rho_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{X_t^{i,N}}. \tag{7}$$

To permit practical implementation, it will also be necessary to discretize in time these SDEs.

Remark 4. We can use the following equilibrium condition to obtain a characterisation of the solution ρ_t of (3): if $\rho^\star$ is a solution, then (5) must be 0

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

Since $\rho^\star$ is positive it follows that I think I'm missing some subtlety here... AMJ

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

for all $\alpha > 0$.

3.1 Existence and uniqueness

To show that the nonlinear SDE (6) admits a unique solution we need to show that the drift is Lipschitz continuous in (x, ρ) (Jourdain et al., 2007).

Take $x, x' \in \mathbb{R}^n$ and $\rho, \rho' \in \mathcal{Q}$ 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 Do you mean a maximal coupling here? AMJcoupling π between ρ, ρ' and consider

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

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 (6) admits a unique strong solution (strong because the proof in Jourdain et al. (2007) is obtained through a contraction argument). Existence and uniqueness of a solution of (7) is given in Protter (2005, Theorem 7, page 253), Jourdain et al. (2007) give error estimates.

The above can be slightly relaxed using Dos Reis et al. (2019, Theorem 3.3) which only requires Lipschitz continuity in ρ and some other conditions w.r.t x.

- Lipschitz in ρ:

$$\begin{aligned}\| \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho K(y)} - \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho' K(y)} \|_2 &= \| \int \mu\,(dy)\, \nabla K(x,y) \left[\frac{1}{\rho K(y)} - \frac{1}{\rho' K(y)} \right] \|_2 \\ &= \| \int \mu\,(dy)\, \nabla K(x,y) \left[\frac{\rho' K(y) - \rho K(y)}{\rho K(y) \rho' K(y)} \right] \|_2 \\ &\leq \int \frac{\mu\,(dy)}{\rho K(y) \rho' K(y)} |\rho' K(y) - \rho K(y)| \|\nabla K(x,y)\|_2 \\ &\leq L W_2(\rho, \rho') \int \frac{\mu\,(dy)}{\rho K(y) \rho' K(y)} \|\nabla K(x,y)\|_2\end{aligned}$$

  if the integral is bounded we have the Lispchitz continuity.

- Locally Lipschitz with polynomial growth: for $q > 1$

$$\begin{aligned}\| \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho K(y)} - \int \mu\,(dy)\, \frac{\nabla K(x',y)}{\rho K(y)} \|_2 &= \| \int \frac{\mu\,(dy)}{\rho K(y)} \left[\nabla K(x,y) - \nabla K(x',y) \right] \|_2 \\ &\leq C(1 + \|x\|_2^q + \|x'\|_2^q) \|x - x'\|_2\end{aligned}$$

- Monotone growth condition

$$\begin{aligned}&\langle x - x', \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho K(y)} - \int \mu\,(dy)\, \frac{\nabla K(x',y)}{\rho K(y)} \rangle = \int \frac{\mu\,(dy)}{\rho K(y)} \langle x - x', \nabla K(x,y) - \nabla K(x',y) \rangle \\ &\leq C \|x - x'\|_2^2\end{aligned}$$

This way, we can relax the Lipschitz continuity of $\nabla K(x,y)$ (but we need boundedness). The diffusion coefficient is constant, thus Lipschitz continuous.

3.2 Propagation of chaos

The propogation of chaos results in Jourdain et al. (2007) only apply to (7) when the integral w.r.t. μ is computed analytically. In practice we would approximate that by M i.i.d. samples from μWe perhaps what to give a bit more context earlier on to make this idea explicit, I think a reader who wasn't familiar with this might have got a bit lost by this point. AMJ, thus

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

We can show that

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

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 Jourdain et al. (2007, Theorem 3). Using similar techniques we can show the same result for the first term:

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

The first term is bounded using the Lipschitz continuity of the drift (see existence and uniqueness proof), for $D_1, D_2 < \infty$

$$\begin{aligned}&C\int_0^t \mathbb{E}\left[\|\int \mu^M(dy)\,\frac{\nabla K(X_s^{i,N,M},y)}{\rho_s^{N,M}K(y)}-\int \mu^M(dy)\,\frac{\nabla K(X_s^{i,N},y)}{\rho_s^{N}K(y)}\|_2^2\right]ds\\ &\leq C\int_0^t \mathbb{E}\left[D_1\|X_s^{i,N,M}-X_s^{i,N}\|_2^2+D_2W_2(\rho_s^{N,M},\rho_s^N)^2\right]ds\\ &\leq C\int_0^t \mathbb{E}\left[D_1\|X_s^{i,N,M}-X_s^{i,N}\|_2^2+D_2\frac{1}{N}\|X_s^{i,N,M}-X_s^{i,N}\|_2^2\right]ds\\ &\leq C\int_0^t \mathbb{E}\left[\sup_{r\leq s}\|X_s^{i,N,M}-X_s^{i,N}\|_2^2\right]ds.\end{aligned}$$

For the second term

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

Therefore,

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

by Gronwall's Lemma applied to

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

we obtain

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

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 $M \to \infty$.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 □

3.3 Ergodicity

The generator of (6) is

$$\mathcal{L}u(x) = (\nabla u(x)) \cdot \int \mu\,(dy)\,\frac{\nabla K(x,y)}{\rho K(y)} + \frac{2\alpha}{2}\Delta u(x) \tag{8}$$

defined for all continuously twice differentiable u.
with adjoint

$$\mathcal{L}^\star\phi(x) = \nabla \cdot \left(\phi(x)\int \mu\,(dy)\,\frac{\nabla K(x,y)}{\rho K(y)}\right) - \frac{2\alpha}{2}\Delta\phi(x). \tag{9}$$

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

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

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

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

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

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

Next: we want to show that 6 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

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

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

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

and

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

3.4 Time Discretisation

We now consider time discretisations of the Mc-Kean Vlasov SDE 7. We start by considering a simple Euler scheme with discretisation step Δt

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

Antonelli et al. (2002) show that this scheme is well behaved if the drift and diffusion coefficient are smooth with bounded derivatives. 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:

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

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:

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

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) coincide 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.

4 Connections with other methods

Minimisation of a functional involving the KL divergence is a common method to solve Fredholm integral equations of the first kind. For example, if we consider the functional

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

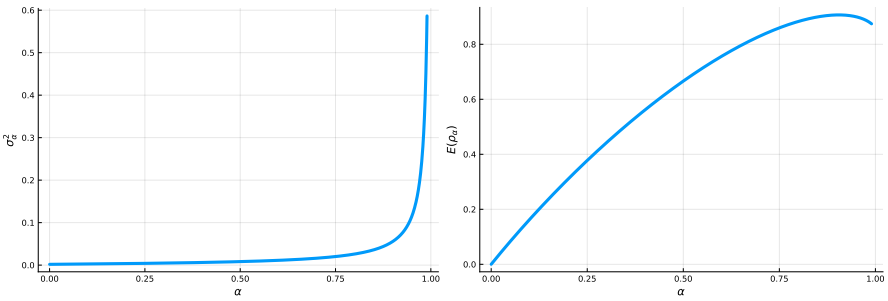

Figure 1: Variance and E as functions of $\alpha \in [0, 0.99]$ for the toy example.

the corresponding first variation is

$$\frac{\delta L}{\delta \rho} = -\int \mu \frac{K}{\rho K} + 1.$$

Setting the first variation to 0 and multiplying by ρ leads to the EM iteration

$$\rho \int \mu \frac{K}{\rho K} - \rho = 0$$

5 Toy Example

We consider the toy Fredholm integral equation

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

For this toy example we find the minimiser of (1), 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

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

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 5.

Since the gradient flow PDE (3) 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).

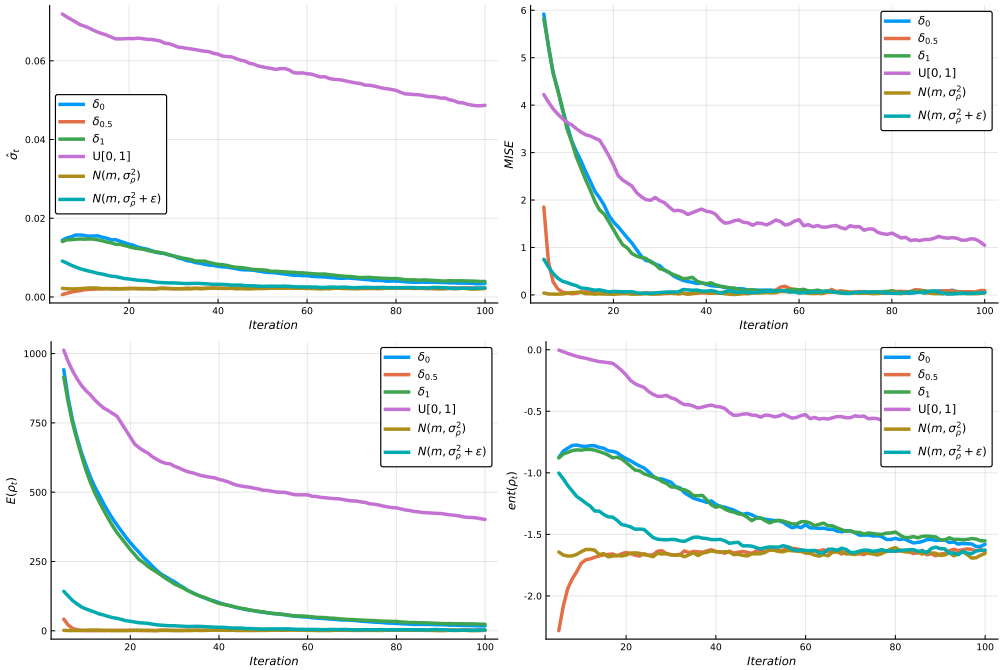

Figure 2: Effect of initial distribution for the toy example. $N = 1000, dt = 10^{-3}, \alpha = 0.025, 100$ iterations.

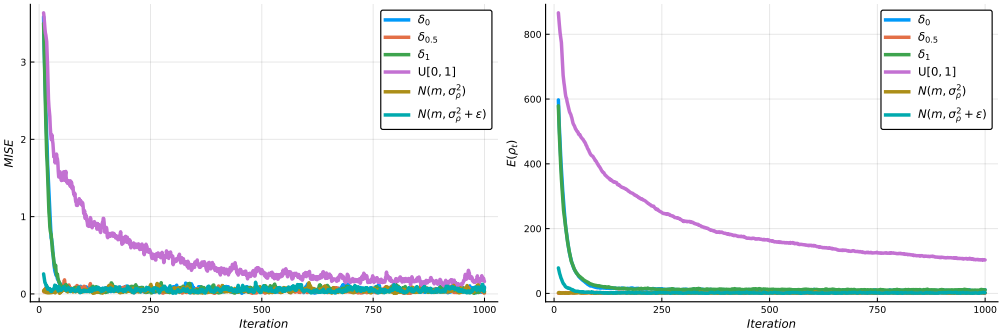

Figure 3: Effect of initial distribution for the toy example. $N = 1000, dt = 10^{-3}, \alpha = 0.025, 1000$ iterations.

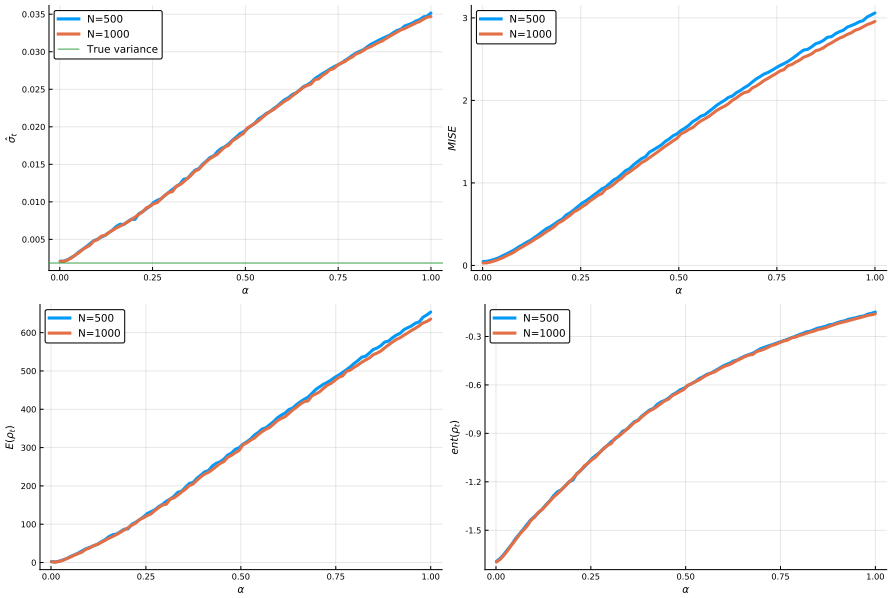

Figure 4: $dt = 10^{-3}, \alpha \in [0, 1]$, 100 iterations, 1000 repetitions for each α

The choice $\rho_0 \sim Uniform([0, 1])$ 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 $dt = 10^{-3}$ and we compare reconstructions of mean μ, variance σ_α^2, 95th percentile of MSE (to check smoothness), MISE, $E(\rho)$ and the entropy $\text{Ent}(\rho)$ 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 5 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 $N = 1000$, $\alpha = 0.025$, $dt = 10^{-3}, 100$ iterations, and the corresponding minimiser of (1) given by $\mathcal{N}(x; m, \sigma_\alpha^2)$ (left hand side). The initial distribution ρ_0 is a point mass concentrated at 0.5. On the right hand side we compare different values of α.

5.1 Comparison with SMC

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

- SMC: Niter $= 100$, $\varepsilon = 10^{-3}$
- WGF: $dt = 10^{-3}, \alpha = 0.025$, Niter $= 100$

6 Gaussian Mixture 1D

This is the indirect density estimation example with

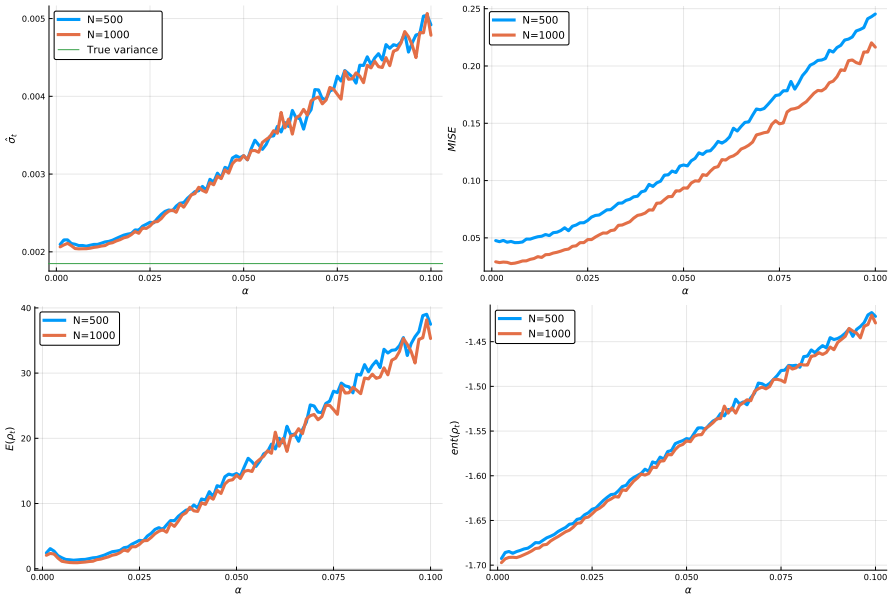

Figure 5: $dt = 10^{-3}, \alpha \in [0, 0.1]$, 100 iterations, 1000 repetitions for each α

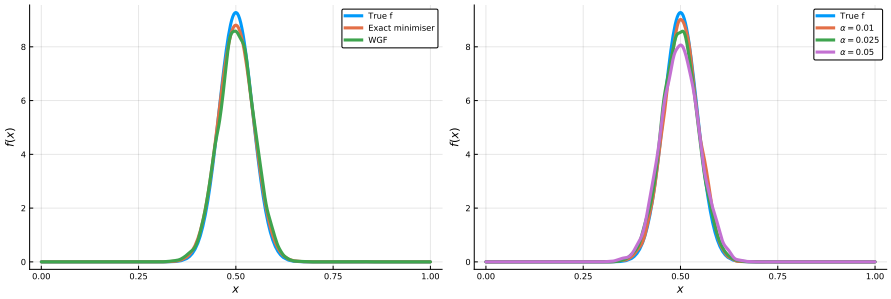

Figure 6: $N = 1000, dt = 10^{-3}, \alpha = 0.025$, 100 iterations.

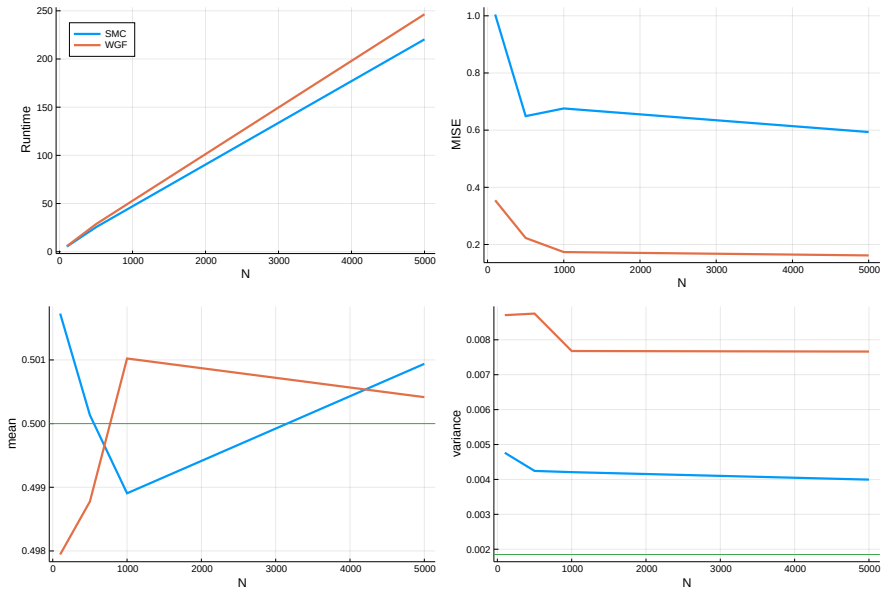

Figure 7: Comparison of SMC and WGF

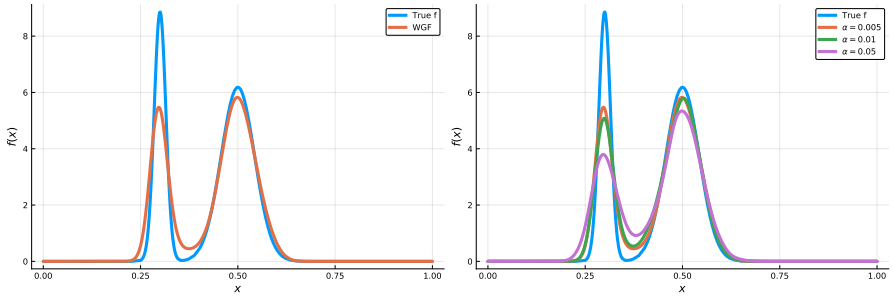

Figure 8: $N = 5000, dt = 10^{-3}, \alpha = 0.005$

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

Figure 6 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 $N = 5000$, $\alpha = 0.005$, $dt = 10^{-3}$. The initial distribution ρ_0 is a point mass concentrated at 0.5. We also compare different values of α.

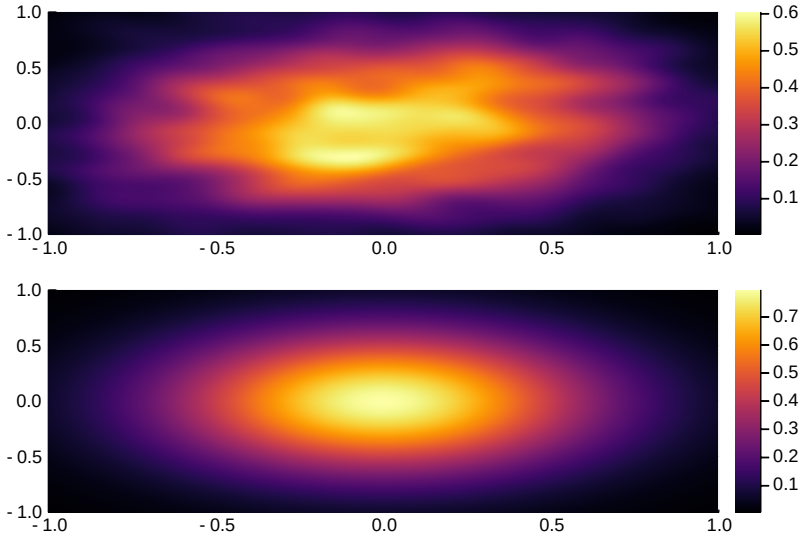

Figure 9: $N = 5000, dt = 10^{-2}, T = 1, \alpha = 0.005$, 1500 iterations

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

8 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.

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.

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.

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.

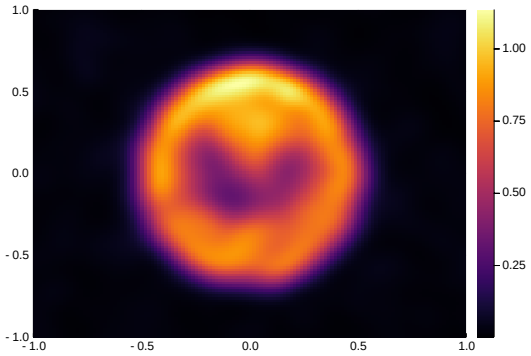

Figure 10: $N = 5000, dt = 10^{-3}, \alpha = 0.1, 1000$ iterations.

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

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.

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.

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.

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.

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

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

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

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.

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.

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

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

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

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

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

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