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

Arnaud -¿ Adam, Francesca

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 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 entropy constraint; i.e. for a given $\lambda > 0$ we propose to minimize w.r.t. ρ

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

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 solving 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

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

is convex in x uniformly in y and is Lipschitz continuous in x uniformly in y with Lipschitz constant L.

(A2) the gradient $\nabla K(x,y)$ is Lipschitz continuous in x uniformly in y with Lipschitz constant L'.

(A1) implies that μ is bounded too.

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

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

where $\Pi(\mu,\nu)$ is the set of all possible couplings between μ and ν. We denote by $\mathcal{P}_2^r(\mathbb{R}^d) \subset \mathcal{P}_2(\mathbb{R}^d)$ the subset of absolutely continuous measures w.r.t. Lebesgue. For every pair $\nu, \mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ there exists a unique optimal transport map t_μ^ν.

We now give some useful definitions.

A functional F defined on $\mathcal{P}_2^r(\mathbb{R}^d)$ is displacement convex (or geodesically convex) if the map

$$s \in [0,1] \mapsto F\left((Id + s(t_\mu^\nu - Id))_{\#}\mu \right)$$

is convex for all $\nu, \mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ (Ambrosio et al., 2008, page 202).

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

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

for all ν.

2.2 Properties of the functional E

We consider the functional E in (1) defined on $\mathcal{P}_2^r(\mathbb{R}^n)$.

It is easy to see that this functional is geodesically convex in ρ.

Proof. We have

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

The integral $\int \mu \log \mu$ is constant w.r.t. ρ and the negative entropy $\lambda \int \rho \log \rho$ is geodesically convex in ρ (Santambrogio, 2017, page 130). Regarding the integral $-\int \mu \log \rho K$, we have that

- $\rho K = \int K(x,y) \; d\rho(x)$ is geodesically convex if K is convex in x (Santambrogio, 2017, page 128)
- $-\log s$ is a convex function

hence $\int \mu \left(-\log \rho K\right)$ is convex in ρ (integrating w.r.t. μ does not change convexity). □

The functional E is also continuous in $(\mathcal{P}_2^r(\mathbb{R}^n), W_2)$.

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

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

if K is continuous, 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. □

2.3 Gradient flow

Following Ambrosio et al. (2008, Definition 11.1.1), we say that ρ_t is a solution to the gradient flow equation for E with the W_2 metric if

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

where $\partial E(\rho_t)$ is the sub-differential of E evaluated at ρ_t, and v_t is a "gradient" for E at ρ_t such that

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

holds.

To arrive to an expression for v_t we compute the first variation of E

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

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. I don't think we can immediately say that the first variation is a sub-differential because E is not a free energy (Carrillo et al., 2006, Lemmata 8-10)

2.3.1 First variation

We have

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

It follows directly that

$$\frac{\delta \, \mathrm{Ent}}{\delta \rho}(\rho) = -(1 + \log \rho).$$

and

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

We have

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

so

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

Hence, it follows that

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

We know $1 + \log \rho(x)$ is a sub-differential for $-\operatorname{Ent}(\rho)$ (Carrillo et al., 2006, Lemma 8) and $\int \mu \log \mu$ is constant w.r.t ρ, so we just need to show that $-\int \mu(dy) \frac{K(x,y)}{\rho K(y)}$ is a sub-differential for $\int \mu \log(\rho K)$.

2.3.2 Existence and uniqueness of the gradient flow

Corollary 11.1.8 in Ambrosio et al. (2008) give existence of a gradient flow solution of (3) since the first variation (2) is single-valued if we can prove that it's a sub-differential.

Uniqueness is given by Theorem 11.1.4 in Ambrosio et al. (2008) since the functional E is geodesically convex in ρ.

2.4 PDE

From Ambrosio et al. (2008, Definition 11.1.1) we know that the continuity-equation must hold

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

where $\nabla \cdot f = \sum_i \partial_i f_i$ is the divergence operator.

This is a Fokker-Plank/Kolmogorov forward equation type 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 $Law(X_t) = \rho_t$. 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

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

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

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

Approximating the first term on the r.h.s. is fine as practically $\mu(dy)$ is a discrete measure but 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

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

Swapping integral and gradient is not a problem because ∇ is over x and the integral is over y.

Hence (3) is equivalent to

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

However, we have

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

where $\triangle f = \sum_i \partial_i^2 f_i$ is the Laplacian. So we can consider the following non-linear SDE (McKean-Vlasov)

$$dX_t = \int\mu(dy)\frac{\nabla K(X_t,y)}{\rho_t K(y)}dt + \sqrt{2\lambda}dW_t, \quad X_0\sim\rho_0, \tag{5}$$

where W_t is a standard $n-$dimensional Brownian motion. The SDE (5) has the same marginal distributions as the nonlinear ODE (4) (Itô's Lemma). 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

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

We will also need to further discretize in time these SDEs obviously.

3.1 Existence and uniqueness

To show that the nonlinear SDE (5) admits a unique solution we need to show that the drift is Lipschitz continuous in (x,ρ) (Jourdain et al., 2007).

We can use the following equivalence: in $\mathbb{R}^n$ the W_2 distance is equal to

$$W_2(\mu,\nu) = \sup\left|\int\varphi\ d(\mu-\nu)\right|$$

where φ is 1-Lipschitz continuous (Santambrogio, 2017, page 112).

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

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

and the last inequality follows from (A1). Using (A1), the second term is bounded by

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

3.2 Ergodicity

The generator of (5) is

$$Lu(x) = \nabla u(x) \int \mu\,(dy)\, \frac{\nabla K(x,y)}{\rho K(y)} + \frac{2\lambda}{2} \Delta u(x) \tag{6}$$

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

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

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

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

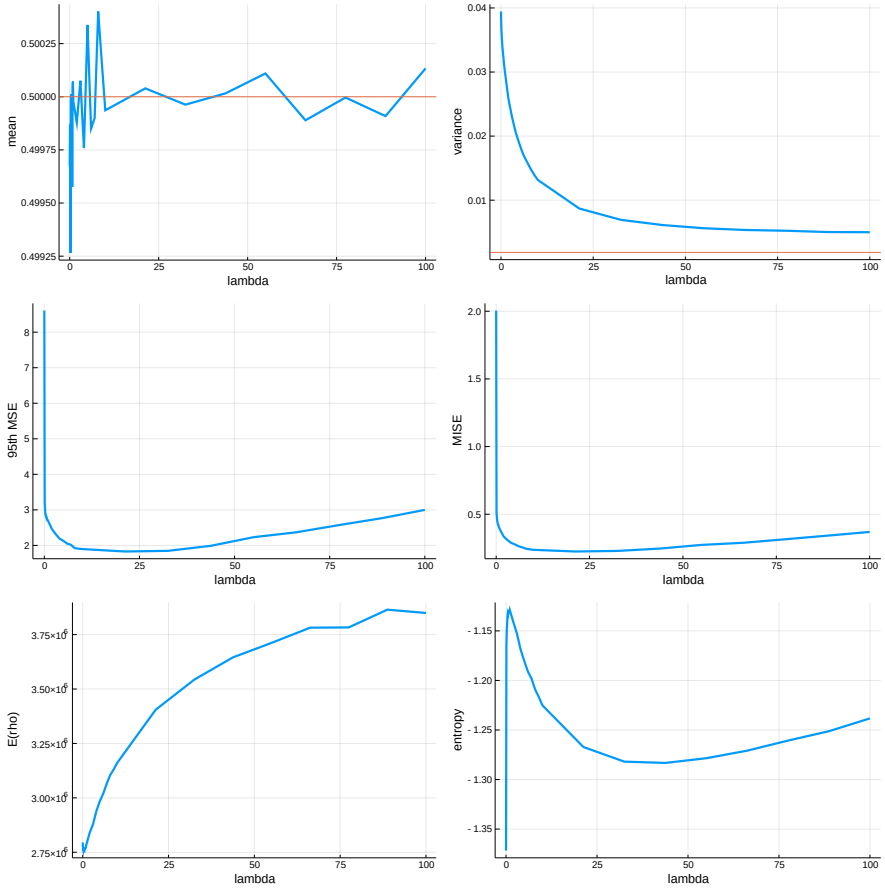

Figure 1: $N = 500, dt = 10^{-3}, T = 1, \lambda \in [0, 100]$, 1000 repetitions for each λ

5 Toy Example

We consider the toy Fredholm integral equation

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

We can use this toy example to get an idea of which values of λ are "good". Thus, we set $dt = 10^{-3}, T = 1$ and we compare reconstructions of mean μ, variance σ_ρ^2, 95th percentile of MSE (to check smoothness), MISE, $E(\rho)$ and the entropy $\mathrm{Ent}(\rho)$. We use $N = 500, 1000$ and 1000 repetitions for each λ.

Figure 5 shows the reconstruction of $\mathcal{N}(x; m, \sigma_\rho^2)$ with $m = 0.5$, $\sigma_\rho^2 = 0.043^2$, $\sigma_K^2 = 0.045^2$. We set $N = 10^4$, $\lambda = 10$, $dt = 10^{-3}$. The initial distribution ρ_0 is uniform on $[0, 1]$.

5.1 Comparison with SMC

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

- SMC: Niter $= 1000$, $\varepsilon = 10^{-3}$

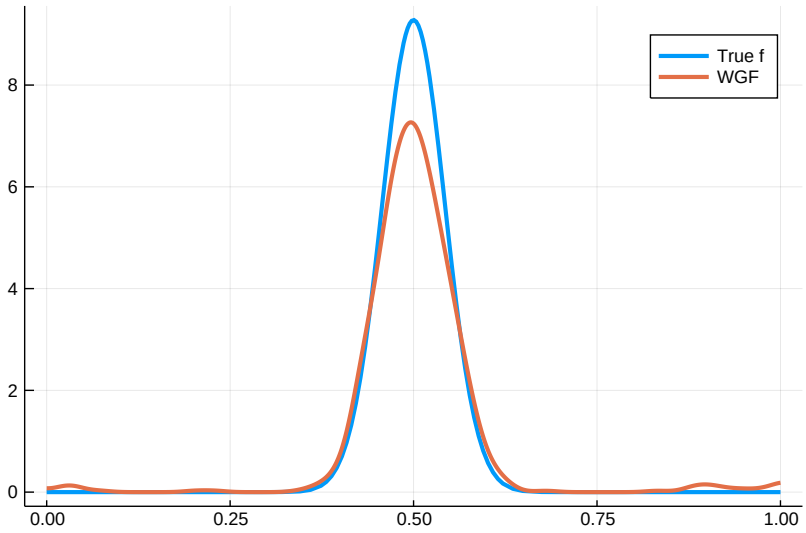

Figure 2: $N = 1000, dt = 10^{-3}, T = 1, \lambda = 25$

- WGF: $dt = 10^{-3}, T = 1, \lambda = 25$

The final number of iterations is the same for both BUT we know that SMC only needs 100 iterations to converge. WGF does not give comparable results if $dt = 10^{-2}$ but gives better MISEwhen $dt = 10^{-3}$.

6 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 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 $N = 10^4$, $\lambda = 10$, $dt = 10^{-3}$. The initial distribution ρ_0 is uniform on $[0, 1]$.

7 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.

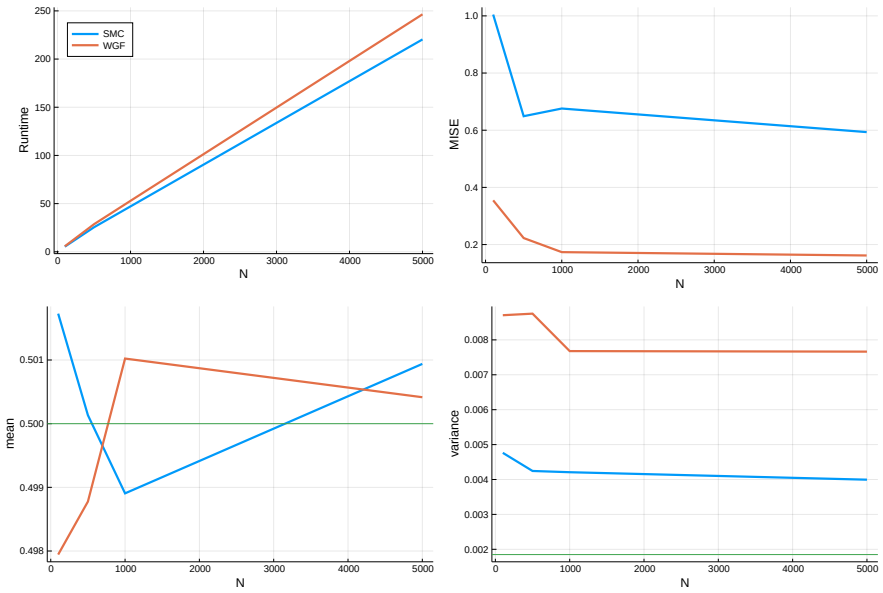

Figure 3: Comparison of SMC and WGF

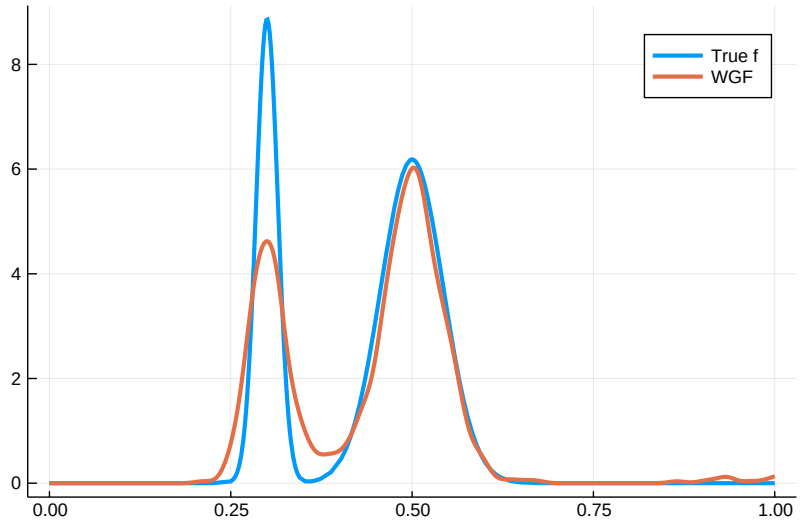

Figure 4: $N = 1000, dt = 10^{-3}, T = 1, \lambda = 25$

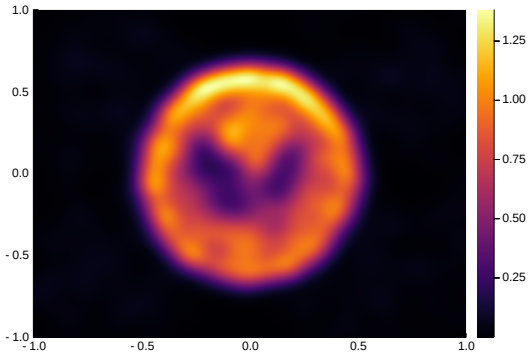

Figure 5: $N = 5000, dt = 10^{-3}, T = 1, \lambda = 25$

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. 2.1, 2.3, 2.3.2, 2.4

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. 2.4

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. 2.3, 2.3.1

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. 2.4

B. Jourdain, S. Méléard, and W. Woyczynski. Nonlinear SDEs driven by L\'evy processes and related PDEs. *arXiv preprint arXiv:0707.2723*, 2007. 3.1

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