SLMath IPS Project3: Mean-field in inverse problems

Yantao Wu, Vasily Ilin, Ian McPherson,
Yutian He, Jaeyoung Yoon, Joseph Hunter,
Kouakou Innocent NDRI,
under the guidance of Prof. Qin Li

June 2024

Contents

1 Introduction of inverse learning problem

1.1 Mean-field/Continuum version

Consider function $u_*(t,x) : [0,1] \times \mathbb{R} \to \mathbb{R}$ given by the dynamics

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} u_* = \displaystyle\int_{\mathbb{R}} f(u_*,\theta)\rho_*(\theta)\,d\theta, \quad t \in [0,1] \\ u_*(t=0,x) = x \end{cases} \tag{1.1}$$

where $\rho_*(\theta)$ is a fixed unknown probability density function over $\mathbb{R}$ and formula of $f(u,\theta)$ is given. We are interested in the inverse learning problem, that is, find an estimator for ρ_* given terminal-time data $D(x) = u_*(t=1,x)$.

We denote

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

to be the space of probability distribution over $\mathbb{R}$ with finite second moment. For any $\rho \in \mathcal{P}_2(\mathbb{R})$, we define $u_\rho(t,x) \in [0,1] \times \mathbb{R}$ given by the same dynamics as above:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} u_\rho = \displaystyle\int_{\mathbb{R}} f(u_\rho, \theta)\rho(\theta)\, d\theta, \quad t \in [0,1] \\ u_\rho(t=0,x) = x \end{cases} \tag{1.2}$$

Obviously, if $\rho = \rho_*$ then uniqueness of ODE guarantees that $u_{\rho_*}(t=1,x) \equiv u_*(t=1,x) = D(x)$. It is natural to consider the following error functional

$$E[\rho] = \frac{1}{2} \int_{\mathbb{R}} |u_\rho(t=1,x) - D(x)|^2 \mu(x)\, dx, \tag{1.3}$$

where $\mu(x)$ indicates that when we are working with a finite measure over $\mathbb{R}$ when considering the x coordinate. We can use $E[\rho]$ to measure how ρ deviates from the underlying truth ρ_*. Clearly, ρ_* is one global minimizer of E. We thus transform the inverse learning problem into an optimization problem.

We will use Weisserstain Gradient Flow to do the optimization: We start with arbitrary $\rho_0 \in \mathcal{P}_2(\mathbb{R})$ then introduce a one-parameter family $\{\rho_s\}_{s\in[0,T]}$ given by the following dynamics:

$$\begin{cases} \partial_s \rho_s = -\nabla_{\mathcal{W}_2} E[\rho_s] = \partial_\theta \left(\rho_s \partial_\theta \dfrac{\delta E}{\delta \rho}\Big|_{\rho_s} \right), \quad s \in [0,T] \\ \rho_{s=0} = \rho_0 \end{cases} \tag{1.4}$$

Here $s \in [0,T]$ is pesudo-time indicating the iteration during optimization, $\nabla_{\mathcal{W}_2}$ means the Weisserstain Gradient, details are shown in the next section. Ideally, we expect that as $s \to \infty$, $\rho_s \to \rho_*$. But this would require that functional $E[\cdot]$ to be convex over $\mathcal{P}_2(\mathbb{R})$.

1.2 Particle-model/Discrete version

In analogy, we consider a similar problem setup in discrete space. Function $U : [0,1] \times \mathbb{R} \to \mathbb{R}$ has the following dynamics

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} U = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} f(U, \theta_i^*), \quad t \in [0,1] \\ U(t=0,x) = x \end{cases} \tag{1.5}$$

Here $\Theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_N^*) \in \mathbb{R}^N$ and each θ_i^* are independent identical distributed (not necessarily according to underlying truth $\rho_* \in \mathcal{P}_2(\mathbb{R})$).

For any $\Theta \in \mathbb{R}^N$, we define $U_\Theta(t,x) \in [0,1] \times \mathbb{R}$ given by the above discrete dynamics:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} U_\Theta = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} f(U_\Theta, \theta_i), \quad t \in [0,1] \\ U_\Theta(t=0,x) = x \end{cases} \tag{1.6}$$

Clearly, if one view Θ as empirical measure $\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i}$, then using notation from previous subsection, we have $U_\Theta(t,x) \equiv u_{\hat{\rho}}(t,x)$.

Heuristically speaking, when $N \to \infty$ we have $\hat{\rho} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i} \to \rho_*$, $\frac{1}{N}\sum_{i=1}^{N} f(\cdot, \theta_i) \to \int_{\mathbb{R}} f(\cdot, \theta)\, d\theta$ and hence $U_\Theta(t, x) \to u_{\rho_*}(t, x) = u_*(t, x)$.

We still use expression $E[\hat{\rho}]$ from (1.3) to measure error. Because now $\Theta \in \mathbb{R}^N$, the error $E(\Theta)$ is a function over $\mathbb{R}^N$. The reader should distinguish it from Energy functional $E[\rho]$ from last subsection. We still use $D(x) = u_*(t = 1, x)$ in the error function.

$$E(\Theta) = E[\hat{\rho}] = \frac{1}{2}\int_{\mathbb{R}} |U_\Theta(t = 1, x) - D(x)|^2 \mu(x)\, dx \tag{1.7}$$

Similar to the continuum version, we also use the Gradient Flow to minimize the error functional. Here $\Theta \in \mathbb{R}^N$ is in a finite dimensional space so the Gradient is the standard gradient in multi-variable calculus $(\partial_{\theta_1}, \ldots, \partial_{\theta_N})$. We start with arbitrary $\Theta_0 = (\theta_{1,0}, \ldots, \theta_{N,0}) \in \mathbb{R}^N$ then introduce a one-parameter family $\{\Theta_s = (\theta_{1,s}, \ldots, \theta_{N,s})\}_{s\in[0,T]}$ given by the following dynamics:

$$\begin{cases} \partial_s \Theta = -\nabla_\Theta E(\Theta_s), \quad s \in [0, T] \\ \Theta_{s=0} = \Theta_0 \end{cases} \tag{1.8}$$

Here $s \in [0, T]$ is pseudo-time indicating the iteration during optimization.

1.3 Interpretation via Neural ODE

Interpretation: above dynamics (1.6) can be interpreted as Neural ODE: we can view U as a finite width N, infinitely deep, layer-wise homogeneous neural network where $t \in [0, 1]$ indicating the layer and x in the input. Then for any fixed $t \in [0, 1], U_\Theta(t, x)$ is the output from the t-th layer of NN. We setup $U_\Theta(t = 0, \cdot)$, the first layer of NN, to be identity function on $x \in \mathbb{R}$. The dynamics of (1.5) can be understood as ResNet-type architecture where $\Theta \in \mathbb{R}^N$ indicates the finitely many parameters.

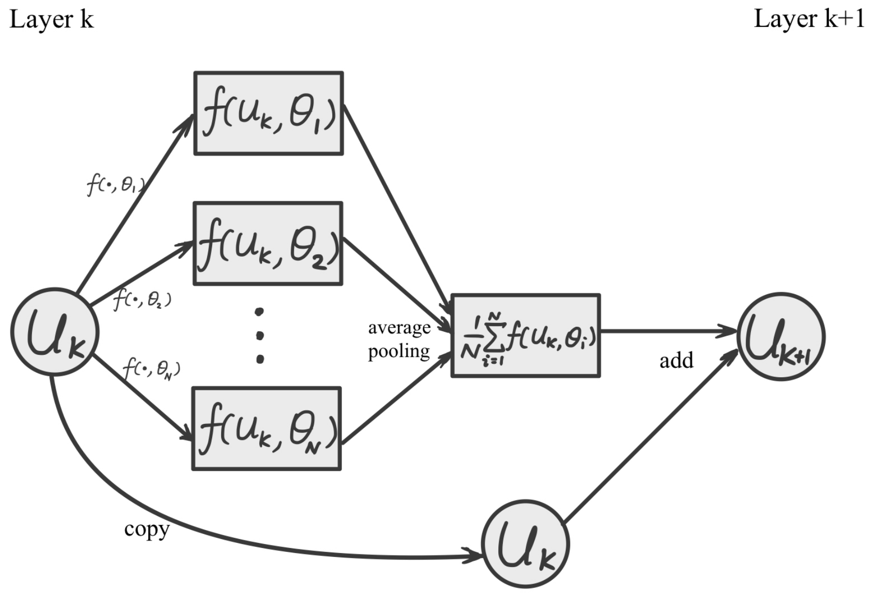

Figure 1: One layer of NN

We are given from data $D(x)$ which describes the true function that we are approximating, and we compare it with our output $U_\Theta(t = 1, \cdot)$ from the final layer of NN. This gives us the loss functional (1.7) and we minimizes it to train the NN to approximate Θ^* which is the true parameters of NN.

The dynamics of (1.1) is the Neural ODE of infinitely wide, infinitely deep, and layer-wide homogeneous NN with ResNet architecture. This gives us the loss functional (1.3) and we minimizes it to train the NN to approximate $\rho_*(\theta)$ which is the true parameters of NN.

The natural question is, how well the particle-model approximates the mean-field model? Heuristically speaking, we might expect them to converge as $N \to \infty$.

We define empirical measure

$$\hat{\rho}_{N,s} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_{i,s}}$$

and we want to show that it will approximate the underlying truth ρ_* by passing $s \to \infty$ and $N \to \infty$. Such an approximation can be illustrated by the following diagram:

$$\begin{tikzcd}
\hat{\rho}_{N,s} \arrow[r, "N\to\infty"] \arrow[d, "s=T"] \arrow[dr] & \rho_s \arrow[d, "s\to\infty"] \\
\hat{\rho}_{N,T} \arrow[r, dashed, "\approx"] & \rho_*
\end{tikzcd}$$

The goal of this project is to justify the horizontal arrow $\hat{\rho}_{N,s} \xrightarrow{N\to\infty} \rho_s$. That says, for each $s \in [0, T]$ (this is iteration time not physical time), particle-based modelling (1.8) gives us an empirical distribution $\hat{\rho}_s = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_{i,s}}$ which should approximate the estimator ρ_s coming from Wasserstein Gradient Flow (1.4). Our final estimator $\hat{\rho}_T$ should approximate ρ_T which should approximate ρ_*.

2 Theoretical Estimates

2.1 Wasserstein Gradient Flows, Fréchet derivatives

We introduce the adjoint dynamics $v_\rho(t, x)$ which evolves backwards in time $t \in [0, 1]$:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} v_\rho = -v_\rho \displaystyle\int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_\rho, \theta)\rho(\theta)\, d\theta, \quad t \in [0, 1] \\ v_\rho(t = 1, x) = D(x) - u_\rho(t = 1, x), \end{cases} \tag{2.1}$$

which runs the dynamics of (1.2) or (1.6), with the terminal condition being the mismatch in the data from the derived u_ρ. With this notation set, we have an explicit form for the Fréchet Derivative for E.

Lemma 2.1. *(Fréchet Derivative for E) Consider functionals $E : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ as defined in* (1.3) *and* (1.7)*. Then, we have that*

$$\frac{\delta E}{\delta \rho}\Big|_\rho (\theta) = \int_0^1 \int_{\mathbb{R}} v_\rho(t, x) f(u_\rho(t, x), \theta)\, \mu(x) dx dt, \tag{2.2}$$

where $v_\rho(t, x)$ is given by (2.1)*.*

As a consequence of Lemma 2.1, we can obtain an explicit formula of $W_\rho(\theta) := -\partial_\theta \frac{\delta E}{\delta \rho}\Big|_\rho(\theta)$, which is the velocity of the gradient flow on the functional E.

$$W_\rho(\theta) := -\partial_\theta \frac{\delta E}{\delta \rho}\Big|_\rho(\theta) = -\int_0^1 \int_{\mathbb{R}} v_\rho(t,x)\partial_\theta f(u_\rho(t,x),\theta)\, d\mu(x)dt \tag{2.3}$$

This such expression, we can rewrite the dynamics of θ_i is the following way:

$$\partial_s \theta_{i,s} = -\partial_{\theta_i} E[\hat{\rho}_{N,s}] = W_{\hat{\rho}_{N,s}}(\theta_{i,s})$$

2.2 Error bounds on difference between Mean-field and Discrete estimators

We introduce an auxiliary Gradient Flow dynamics as follows: for $i = 1, \dots, N$, let $\bar{\theta}_{i,0} = \theta_{i,0}$ independent identical distributed (not according to the underlying truth $\rho_* \in \mathcal{P}_2(\mathbb{R})$). For each i, we introduce a one-parameter family $\{\bar{\theta}_{i,s}\}_{s\in[0,T]}$ given by the following dynamics

$$\begin{cases} \partial_s \bar{\theta}_{i,s} = W_{\rho_s}(\bar{\theta}_{i,s}), \quad s \in [0,T] \\ \bar{\theta}_{i,s=0} = \bar{\theta}_{i,0} \end{cases} \tag{2.4}$$

It is worthy mentioning that above dynamics is similar to (1.8) except that it follows the Gradient Flow velocity W_{ρ_s} which is given by continuum ρ_s instead of empirical measure $\hat{\rho}_{N,s}$. As a consequence, there is no coupling between $\bar{\theta}_{i,s}$ and $\bar{\theta}_{j,s}$. That says $\bar{\theta}_{i,s}$ are independent among i's. As a consequence, for any i, each $\theta_{i,s}$ follows that law of ρ_s.

We denote $\bar{\rho}_{N,s} = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{\theta}_{i,s}}$ to be empirical distribution which evolves during Auxiliary Discrete Gradient Flow (2.4). The subscript s emphasize the Gradient Flow evolution over $s \in [0,T]$.

Theorem 2.2 (Mean-field limit). *Let $s \in [0,T]$ be pseudo-time. Let $\bar{\rho}_{N,s}$ be the empirical measures built from N iid particles $\bar{\theta}_{i,s}$ sampled from ρ_s, which follow dynamics* (2.4)*, and let $\hat{\rho}_{N,s}$ be the empirical measure built from N particles $\theta_{i,s}$ which follow dynamics* (1.8)*. We take $\bar{\rho}_0 = \hat{\rho}_0$.*

Assume that f is such that

$$\begin{aligned} C_f^0 &:= \|\partial_\theta f(y,\theta)\|_{L^\infty(y,\theta)} + \|\partial_y f(y,\theta)\|_{L^\infty(y,\theta)} < \infty, \\ C_f^1 &:= \left\|\partial_\theta^2 f(y,\theta)\right\|_{L^\infty(y,\theta)} + \left\|\partial_{\theta,y}^2 f(y,\theta)\right\|_{L^\infty(y,\theta)} + \left\|\partial_y^2 f(y,\theta)\right\|_{L^\infty(y,\theta)} + 1 < \infty, \end{aligned}$$

and the error functional E is finite in expectation on $\rho_s, \hat{\rho}_{N,s}$ and $\bar{\rho}_{N,s}$, i.e.

$$\begin{aligned} &\mathbb{E}E[\rho_s] = \int |D(x) - u_{\rho_s}(t=1,x)|^2 d\mu(x) < \infty, \\ &\mathbb{E}E[\hat{\rho}_{N,s}] < \infty, \quad \mathbb{E}E[\bar{\rho}_{N,s}] < \infty \end{aligned}$$

Assume that for all s, ρ_s has five bounded moments

$$\int |\theta|^5 d\rho_s(\theta) < \infty.$$

Then, we have for all $s \in [0,T]$,

$$\begin{aligned} \mathbb{E}\mathcal{W}_2(\rho_s, \hat{\rho}_{N,s}) &\le \frac{C}{N^{1/4}} + \frac{C''}{N}\exp\big((C'+1)s\big) \\ &= \mathcal{O}(N^{-1/4}), \end{aligned}$$

where

$$\begin{aligned} C &= \text{universal constant} \\ C' &= 32\left(C_f^1\right)^2 e^{5C_f^0}\mathbb{E}\left(E[\rho_s]^{\frac{1}{2}} + E[\hat{\rho}_{N,s}]^{\frac{1}{2}} + E[\bar{\rho}_{N,s}]^{\frac{1}{2}}\right) \\ C'' &= \left(C'\left\|f(u_\rho,\theta)\right\|_{L^4(\mu(x)\otimes\rho(x))}\right)^2 \end{aligned}$$

Proof. First, we get $\frac{d}{ds}|\theta_i - \bar{\theta}_i|^2$ into a form so we may apply *generalized Grönwall's Inequality*:

$$\begin{aligned} &\partial_s|\theta_{i,s} - \bar{\theta}_{i,s}|^2 = 2|\theta_{i,s} - \bar{\theta}_{i,s}|\ \partial_s(\theta_{i,s} - \bar{\theta}_{i,s}) \\ &\leq 2(\theta_i - \bar{\theta}_i)\left(\left|W_{\hat{\rho}_{N,s}}(\theta_{i,s}) - W_{\hat{\rho}_{N,s}}(\bar{\theta}_{i,s})\right| + \left|W_{\hat{\rho}_{N,s}}(\bar{\theta}_{i,s}) - W_{\bar{\rho}_{N,s}}(\bar{\theta}_{i,s})\right| + \left|W_{\bar{\rho}_{N,s}}(\bar{\theta}_{i,s}) - W_{\rho_s}(\bar{\theta}_{i,s})\right|\right) \end{aligned}$$

The first term is controlled by the Lipshitzness of $W_\rho(\theta)$ w.r.t. θ (proposition 2.4 Part 1). The second term is controlled by Lipschitzness with respect to ρ (proposition 2.4 Part 2). The third term is bounded by a constant times $\frac{1}{\sqrt{N}}$ in expectation (proposition 2.4 Part 3). So, in expectation over the sample $\bar{\theta}_i$, we get

$$\begin{aligned} \partial_s|\theta_{i,s} - \bar{\theta}_{i,s}|^2 &\leq |\theta_{i,s} - \bar{\theta}_{i,s}|\left(L_A|\theta_{i,s} - \bar{\theta}_{i,s}| + L_B\left(\frac{1}{N}\sum_{j=1}^N|\theta_{j,s} - \bar{\theta}_{j,s}|\right) + L_C\frac{1}{\sqrt{N}}\right) \\ &\leq \left(L_A + \frac{L_B}{2} + \frac{1}{2}\right)|\theta_{i,s} - \bar{\theta}_{i,s}|^2 + \frac{L_B}{2}\frac{1}{N}\sum_{j=1}^N|\theta_{j,s} - \bar{\theta}_{j,s}|^2 + \frac{L_C^2}{2N} \end{aligned}$$

Summing over i, we obtain

$$\partial_s\left(\frac{1}{N}\sum_{i=1}^N|\theta_{i,s} - \bar{\theta}_{i,s}|^2\right) \leq \frac{L_C^2}{2N} + \left(L_A + L_B + \frac{1}{2}\right)\left(\frac{1}{N}\sum_{i=1}^N|\theta_{i,s} - \bar{\theta}_{i,s}|^2\right)$$

Recall that $\bar{\theta}_{i,s} = \theta_{i,s}$ when $s = 0$, then by *generalized Grönwall's Inequality*,

$$\mathcal{W}_2(\bar{\rho}_{N,s}, \hat{\rho}_{N,s}) \leq \frac{1}{N}\sum_{i=1}^N|\theta_{i,s} - \bar{\theta}_{i,s}|^2 \leq \frac{sL_C^2}{2N}\exp\left(\left(L_A + L_B + \frac{1}{2}\right)s\right)$$

in expectation over Ω. Finally, we combine the above estimate with theorem 2.3 to obtain the desired bound. □

Theorem 2.3 (Theorem 1 from [FG13]). *Let $\rho \in \mathcal{P}(\mathbb{R}^d)$ and let $p > 0$. Assume that*

$$\int|\theta|^q\rho(d\theta) < \infty$$

for some $q > p$ and define

$$\bar{\rho} := \frac{1}{N}\sum_i\delta_{\theta_i}, \quad \theta_i \sim \rho \ \textit{\textbf{iid}}.$$

There exists a constant C depending only on p, d, q such that, for all $N \geq 1$,

$$\mathbb{E}\left(\mathcal{W}_p^p(\bar{\rho}_N, \rho)\right) \leq CM_q^{p/q}(\rho)\begin{cases} N^{-1/2} + N^{-(q-p)/q} & \textit{if } p > d/2 \textit{ and } q \neq 2p, \\ N^{-1/2}\log(1+N) + N^{-(q-p)/q} & \textit{if } p = d/2 \textit{ and } q \neq 2p, \\ N^{-p/d} + N^{-(q-p)/q} & \textit{if } p \in (0, d/2) \textit{ and } q \neq d/(d-p). \end{cases}$$

2.3 Lipshitz properties

Proposition 2.4 (Lipschitz Properties of Gradient Flow Dynamics)**.** *Assume that the first and second derivatives of f are bounded, i.e. $C_f^0, C_f^1 < \infty$ and $E[\hat{\rho}_{N,s}], E[\bar{\rho}_{N,s}] < \infty$. Then, we have the following three Lipschitz Properties:*

1.
$$|W_{\hat{\rho}_{N,s}}(\theta_{i,s}) - W_{\hat{\rho}_{N,s}}(\bar{\theta}_{i,s})| \leq L_A |\theta_{i,s} - \bar{\theta}_{i,s}|,$$
where
$$L_A := 2C_f^1 e^{C_f^0} E[\rho]^{\frac{1}{2}}.$$

2.
$$|W_{\hat{\rho}_{N,s}}(\bar{\theta}_{i,s}) - W_{\bar{\rho}_{N,s}}(\bar{\theta}_{i,s})| \leq L_B \left(\frac{1}{N} \sum_{j=1}^{N} |\bar{\theta}_{j,s} - \theta_{j,s}| \right)$$
where
$$L_B := 8 \left(C_f^1\right)^2 e^{5C_f^0} \left(E[\hat{\rho}_{N,s}]^{\frac{1}{2}} + E[\bar{\rho}_{N,s}]^{\frac{1}{2}} \right)$$

3.
$$\mathbb{E}_\Omega |W_\rho(\theta) - W_{\bar{\rho}_N}(\theta)| \leq L_C \frac{1}{\sqrt{N}}$$
where
$$L_C := 8C_f^1 e^{5C_f^0} \left(E[\rho]^{\frac{1}{2}} + E[\bar{\rho}_N]^{\frac{1}{2}} \right) \|f(u_\rho, \theta)\|_{L^4(\mu(x)\otimes\rho(x))}$$

Theorem 2.5 (Flow Velocity W is Lipschitz in θ)**.** *Consider any probability measure ρ. Then, we have that*
$$|W_\rho(\theta_1) - W_\rho(\theta_2)| \leq 2C_f^1 e^{C_f^0} E[\rho]^{\frac{1}{2}} |\theta_1 - \theta_2|.$$

Proof. We directly compute that
$$\begin{aligned}
&|W_\rho(\theta_1) - W_\rho(\theta_2)| \leq |\theta_1 - \theta_2| \sup_\theta |\partial_\theta W_\rho(\theta)| \\
=&|\theta_1 - \theta_2| \sup_{\theta\in\mathbb{R}} \left| \int_0^1 \int_{\mathbb{R}} v_\rho(t,x) \partial_\theta^2 f(u_\rho(t,x), \theta) \, d\mu(x) dt \right| \\
\leq& |\theta_1 - \theta_2| \sup_{\theta\in\mathbb{R}} \int_0^1 \|v_\rho(t,x)\|_{L^2(\mu(x)} \left\|\partial_\theta^2 f(u_\rho(t,x), \theta\right\|_{L^2(\mu(x)))} dt \\
\leq& |\theta_1 - \theta_2| \left\|\partial_\theta^2 f\right\|_{L^\infty(y,\theta)} \sup_\theta \int_0^1 \|v_\rho(t,x)\|_{L^2(\mu(x))} dt \\
\leq& |\theta_1 - \theta_2| \, 2E[\rho]^{\frac{1}{2}} \exp(C_f^0) C_f^1
\end{aligned}$$

where we consecutively apply mean-value inequality, Cauchy-Schwartz inequality, Jensen Inequality, use a L^∞ bound on $\partial_\theta^2 f$, and then Lemma 2.10. □

We define the linear operator I acting on $f(y,\cdot)$ and $\partial_y f(y,\cdot)$

$$(I_{\rho_1,\rho_2}f)(y) := \int_{\mathbb{R}} f(y,\theta)(\rho_1(\theta)-\rho_2(\theta))d\theta, \quad (I_{\rho_1,\rho_2}\partial_y f)(y) := \int_{\mathbb{R}} \partial_y f(y,\theta)(\rho_1(\theta)-\rho_2(\theta))d\theta,$$

where we henceforth assume $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R})$ such that $|u_{\rho_1}(t=0,x) - u_{\rho_2}(t=0,x)| = 0$ for all $x \in \mathbb{R}$.

Theorem 2.6 (Flow Velocity W is Lipschitz-Like in ρ)**.** *For any two probability measures ρ_1 and ρ_2 and for any $t \in [0,1]$,*

$$|W_{\rho_1}(\theta) - W_{\rho_2}(\theta)| \le 4C_f^1 e^{4C_f^0}\left(E[\rho_1]^{\frac{1}{2}} + E[\rho_2]^{\frac{1}{2}}\right)\left(\|If(u_{\rho_1},\cdot)\|_{L^2(\mu(x))} + \left\|I\partial_{u_{\rho_1}} f\right\|_{L^2(\mu(x))}\right)$$

Corollary 2.7. *When $\rho_1 = \hat{\rho}_{N,s}$ and $\rho_2 = \bar{\rho}_{N,s}$, by lemma 2.13, we obtain*

$$|W_{\rho_1}(\theta) - W_{\rho_2}(\theta)| \le 8\left(C_f^1\right)^2 e^{5C_f^0}\left(E[\hat{\rho}_{N,s}]^{\frac{1}{2}} + E[\bar{\rho}_{N,s}]^{\frac{1}{2}}\right)\left(\frac{1}{N}\sum_{i=1}^N |\theta_i - \bar{\theta}_i|\right)$$

Corollary 2.8. *When $\rho_1 = \rho_s$ and $\rho_2 = \bar{\rho}_{N,s}$, and*

$$\|f(u_\rho,\theta)\|_{L^4(\mu(x)\otimes\rho(x))} < \infty,$$

by lemma 2.12, we obtain

$$\mathbb{E}_\Omega|W_\rho(\theta) - W_{\bar{\rho}_N}(\theta)| \le \frac{8}{\sqrt{N}}C_f^1 e^{5C_f^0}\left(E[\rho]^{\frac{1}{2}} + E[\bar{\rho}_N]^{\frac{1}{2}}\right)\|f(u_\rho,\theta)\|_{L^4(\mu(x)\otimes\rho(x))}$$

Proof of theorem 2.6. Recall that $W_\rho(\theta) = -\int_0^1\int_{\mathbb{R}} v_\rho(t,x)\partial_\theta f(u_\rho(t,x),\theta)d\mu(x)dt$, then we consecutively use triangle inequality, Cauchy-Schwartz inequality, and Lemma 2.10, Lemma 2.9, and Lemma 2.11, to compute that

$$\begin{aligned}
|W_{\rho_1}(\theta) - W_{\rho_2}(\theta)| \le &\int_0^1\int_{\mathbb{R}} |v_{\rho_2}(t,x)||\partial_\theta f(u_{\rho_1}(t,x),\theta) - \partial_\theta f(u_{\rho_2}(t,x),\theta)|d\mu(x)dt \\
&+ \int_0^1\int_{\mathbb{R}} |v_{\rho_1}(t,x) - v_{\rho_2}(t,x)||\partial_\theta f(u_{\rho_1}(t,x),\theta)|d\mu(x)dt \\
=&C_f^1\,\||v_{\rho_1}|\cdot|u_{\rho_1} - u_{\rho_2}|\|_{L^1(x)} + \||v_{\rho_1} - v_{\rho_2}|\cdot|\partial_\theta f(u_{\rho_2},\theta)|\|_{L^1(\mu(x))} \\
\le&C_f^1\,\|v_{\rho_1}\|_{L^2(\mu(x))}\,\|u_{\rho_1} - u_{\rho_2}\|_{L^2(\mu(x))} + C_f^0\,\|v_{\rho_1} - v_{\rho_2}\|_{L^1(\mu(x))} \\
\le&2C_f^1 E[\rho_1]^{\frac{1}{2}} e^{C_f^0}\,\|If(u_{\rho_1}(t,x),\theta)\|_{L^2(\mu(x)),L^\infty(t)} \\
&+ C_f^0 2e^{C_f^0}\left(E[\rho_2]^{\frac{1}{2}} e^{C_f^0}\cdot\left\|I\partial_{u_{\rho_1}} f\right\|_{L^2(\mu(x))} + \left(E[\rho_2]^{\frac{1}{2}} e^{C_f^0} C_f^1 + 1\right) e^{C_f^0}\left\|If(u_{\rho_1(t,x)})\right\|_{L^2(\mu(x))}\right) \\
\le&4C_f^1 e^{4C_f^0}\left(E[\rho_1]^{\frac{1}{2}} + E[\rho_2]^{\frac{1}{2}}\right)\left(\|If(u_{\rho_1},\cdot)\|_{L^2(\mu(x))} + \left\|I\partial_{u_{\rho_1}} f\right\|_{L^2(\mu(x))}\right).
\end{aligned}$$

The last equality is the result of an elementary algebraic manipulation. □

Below are technical lemmas to control the deviation coming from difference of measure.

Lemma 2.9. *Let $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R})$ as before. Then, for arbitrary $x \in \mathbb{R}$ we have*

$$|u_{\rho_1}(t,x) - u_{\rho_2}(t,x)| \leq e^{C_f^0} \left\|I_{\rho_1,\rho_2} f(u_{\rho_1}(t,x)\right\|_{L^\infty(t)}$$

This can be used to get $\|u_{\rho_N} - u_\rho\|_{L^p(x)}$ bounds for $p = 1, 2, \infty$. Explicitly, for $p = 2$, we have that

$$\|u_{\rho_1}(t,x) - u_{\rho_2}(t,x)\|_{L^2(\mu(x))} \leq e^{C_f^0} \left\|I_{\rho_1,\rho_2} f(u_{\rho_1(t,x)})\right\|_{L^2(\mu(x)),L^\infty(t)}.$$

Proof. To apply *Grönwall's Inequality*, we first compute that

$$\begin{aligned}
&\frac{d}{dt}|u_{\rho_1}(t,x) - u_{\rho_2}(t,x)| \leq \left|\frac{du_{\rho_1}}{dt}(t,x) - \frac{du_{\rho_2}}{dt}(t,x)\right| = \left|\int_{\mathbb{R}} f(u_{\rho_1}(t,x),\theta)\rho_1(\theta) - f(u_{\rho_2}(t,x),\theta)\rho_2(\theta)d\theta\right| \\
\leq &\left|\int_{\mathbb{R}} f(u_{\rho_1}(t,x),\theta)[\rho_1(\theta) - \rho_2(\theta)]d\theta\right| + \left|\int_{\mathbb{R}} f(u_{\rho_1}(t,x),\theta)\rho_2(\theta) - f(u_{\rho_2}(t,x),\theta)\rho_2(\theta)d\theta\right| \\
\leq &|I_{\rho_1,\rho_2} f(u_{\rho_1}(t,x))| + \int_{\mathbb{R}} C_f^0 |u_{\rho_1}(t,x) - u_{\rho_2}(t,x)|\rho_2(\theta)d\theta \\
= &|I_{\rho_1,\rho_2} f(u_{\rho_1}(t,x))| + C_f^0 |u_{\rho_1}(t,x) - u_{\rho_2}(t,x)|,
\end{aligned}$$

where the last inequality we use the C_f^0 Lipschitz bound on the first argument of f. Then, by *Generalized Grönwall Inequality*, and assumptions on ρ_1, ρ_2, we have that for all $t \in [0,1]$, $x \in \mathbb{R}$,

$$|u_{\rho_1}(t,x) - u_{\rho_2}(t,x)| \leq e^{C_f^0 t} \int_0^t |I_{\rho_1,\rho_2} f(u_{\rho_1}(\tau,x))| d\tau \leq e^{C_f^0} \|I_{\rho_1,\rho_2} f(u_{\rho_1}(t,x))\|_{L^\infty(t)}.$$

Lastly, since this is a point-wise bound in x, it immediately follows that

$$\|u_{\rho_1}(t,x) - u_{\rho_2}(t,x)\|_{L^2(\mu(x))} \leq e^{C_f^0} \|I_{\rho_1,\rho_2} f(u_{\rho_1}(t,x))\|_{L^2(\mu(x)),L^\infty(t)}.$$

□

Lemma 2.10. *Fix arbitrary $t \in [0,1]$. Then, we have that*

$$\|v_\rho(t,x)\|_{L^2(\mu(x))} \leq 2E[\rho]^{\frac{1}{2}} e^{C_f^0}.$$

Proof. Recall that the solution v_ρ of dynamics (2.1) can be written explicitly as follows:

$$v_\rho(t,x) = (D(x) - u_\rho(t=1,x)) \exp\left(\int_0^{1-t} \int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_\rho(1-t',x),\theta)\rho(\theta) d\theta dt'\right)$$

As a consequence, for any $x \in \mathbb{R}$, we have L^∞-norm:

$$\|v_\rho(\cdot,x)\|_{L^\infty(t\in[0,1])} \leq |D(x) - u_\rho(t=1,x)| \exp(C_f^0)$$

where in the inequality we use that ρ is a probability measure. Then, we have that

$$\|v_\rho(t,x)\|_{L^2(\mu(x))} \leq \exp(C_f^0)\left(\int_{\mathbb{R}} |D(x) - u_\rho(t=1,x)|^2 d\mu(x)\right)^{\frac{1}{2}} = 2E[\rho]^{\frac{1}{2}} \exp(C_f^0).$$

□

Lemma 2.11. *We take I to mean I_{ρ_1,ρ_2} and take a* $\sup$ *over $t \in [0,1]$ below.*

$$|v_{\rho_1} - v_{\rho_2}| \le \exp\left(\|\partial_y f\|_{L^\infty(y,\theta)}\right)\left(|v_{\rho_2}| \cdot \left|I\partial_{u_{\rho_1}} f\right| + (|v_{\rho_2}|C_f^1 + 1)|u_{\rho_1} - u_{\rho_2}|\right)$$

In particular, taking the $L^1(\mu(x))$ norm and using lemmas 2.10 and 2.9, we obtain

$$\|v_{\rho_1} - v_{\rho_2}\|_{L^1(\mu(x))} \le 2e^{C_f^0}\left(E[\rho_2]^{\frac{1}{2}}e^{C_f^0} \cdot \left\|I\partial_{u_{\rho_1}} f\right\|_{L^2(\mu(x))} + \left(E[\rho_2]^{\frac{1}{2}}e^{C_f^0}C_f^1 + 1\right)e^{C_f^0}\left\|If(u_{\rho_1}(t,x))\right\|_{L^2(\mu(x))}\right)$$

Proof. Recall that $\frac{\mathrm{d}}{\mathrm{d}t}v_\rho = -v_\rho \int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_\rho, \theta)\rho(\theta)\,d\theta$. Therefore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(v_{\rho_1} - v_{\rho_2}) \le{}& |v_{\rho_1} - v_{\rho_2}| \int_{\mathbb{R}} \left|\frac{\partial f}{\partial u}(u_{\rho_1}, \theta)\right| \rho_1(\theta)\,d\theta \\ &+ |v_{\rho_2}| \left|\int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_{\rho_1}, \theta)(\rho_1(\theta) - \rho_2(\theta)\,d\theta\right| + |v_{\rho_2}| \int_{\mathbb{R}} \left|\frac{\partial f}{\partial u}(u_{\rho_1}, \theta) - \frac{\partial f}{\partial u}(u_{\rho_2}, \theta)\right| \rho_2(\theta)\,d\theta \\ \le{}& |v_{\rho_1} - v_{\rho_2}|\, C_f^0 + |v_{\rho_2}||I\frac{\partial f}{\partial u}(u_{\rho_1}, \cdot)| + |v_{\rho_2}|C_f^1|u_{\rho_1} - u_{\rho_2}|\end{aligned}$$

Thus we apply Gronwall's inequality to obtain, for any $t \in [0,1]$ and any $x \in \mathbb{R}$,

$$\begin{aligned}&|v_{\rho_1}(t,x) - v_{\rho_2}(t,x)| \\ \le{}& e^{C_f^0}\left(|u_{\rho_1}(t=1,x) - u_{\rho_2}(t=1,x)| + \int_0^t |v_{\rho_2}||I\frac{\partial f}{\partial u}(u_{\rho_1}, x)| + |v_{\rho_2}|C_f^1|u_{\rho_1} - u_{\rho_2}|\right) d\tau \\ \le{}& e^{C_f^0}\left((\|v_{\rho_2}\|_{L^\infty(t\in[0,1])}\, C_f^0 + 1)\, \|u_{\rho_1} - u_{\rho_2}\|_{L^\infty(t\in[0,1])} + \|v_{\rho_2}\|_{L^\infty(t\in[0,1])} \left\|I\partial_{u_{\rho_1}} f\right\|_{L^\infty(t\in[0,1])}\right)\end{aligned}$$

In particular, taking the $L^1(\mu(x))$ norm and using lemmas 2.10 and 2.9, we obtain

$$\|v_{\rho_1} - v_{\rho_2}\|_{L^1(\mu(x))} \le 2e^{C_f^0}\left(E[\rho_2]^{\frac{1}{2}}e^{C_f^0} \cdot \left\|I\partial_{u_{\rho_1}} f\right\|_{L^2(\mu(x))} + \left(E[\rho_2]^{\frac{1}{2}}e^{C_f^0}C_f^1 + 1\right)e^{C_f^0}\left\|If(u_{\rho_1}(t,x))\right\|_{L^2(\mu(x))}\right)$$

□

We use the following lemma with $h(\theta) = f(u(x,t),\theta)$ and $h(\theta) = \partial_u f(u(x,t),\theta)$, in the bound on $|W_\rho - W_{\bar{\rho}_N}|$. We consider two types of deviation of measure: ρ v.s. $\bar{\rho}_N$ and $\hat{\rho}_N$ v.s. $\bar{\rho}_N$.

Lemma 2.12 (Quantitative law of large numbers)**.** *If $\|h\|_{L^2(\rho(\theta)} < \infty$, then*

$$\mathbb{E}_\Omega\, |I_{\rho,\bar{\rho}_N} h| \le \left(\mathbb{E}_\Omega\, |I_{\rho,\bar{\rho}_N} h|^2\right)^{1/2} \le \frac{1}{\sqrt{N}}\,\mathrm{Var}_\rho(h)^{1/2} \le \frac{1}{\sqrt{N}}\,\|h\|_{L^2(\rho(\theta))}\,.$$

In particular, for any $u(x)$, if

$$\|f(u,\theta)\|_{L^4(\mu(x)\otimes\rho(x))} < \infty,$$

then

$$\mathbb{E}_\Omega\, \|I_{\rho,\bar{\rho}_N} f(u,\theta)\|_{L^2(\mu(x))} \le \frac{1}{\sqrt{N}}\,\|f(u,\theta)\|_{L^4(\mu(x)\otimes\rho(x))}\,.$$

Proof. Recall that empirical measure $\hat{\rho}_{N,s} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{\theta}_{i,s}}$ where $\bar{\theta}_{i,s}$ are iid with distribution ρ_s. Let $Z_i = h(\bar{\theta}_{i,s})$, then $I_{\rho_s, \bar{\rho}_{N,s}} = \frac{1}{N}\sum_{i=1}^{N} h(\bar{\theta}_{i,s}) - \mathbb{E}_\Omega[Z_i]$. Thus, we use Jensen inequality to obtain

$$(\mathbb{E}_\Omega \left| I_{\rho,\bar{\rho}_N} h \right|)^2 \leq \mathbb{E}_\Omega |I_{\rho_s,\bar{\rho}_{N,s}}|^2 = \frac{1}{N^2}\sum_{i=1}^{N} \left| h(\bar{\theta}_{i,s}) - \mathbb{E}_\Omega[Z_i] \right|^2 = \frac{1}{N}\operatorname{Var}_\rho(h) \leq \frac{1}{N}\, \|h\|^2_{L^2(\rho(\theta))}$$

□

We use the following lemma with $h(\theta) = f(u(x,t),\theta)$ and $h(\theta) = \partial_u f(u(x,t),\theta)$, in the bound on $|W_{\hat{\rho}_N} - W_{\bar{\rho}_N}|$.

Lemma 2.13. *If* $h(\theta)$ *is* C^1*, then*

$$|I_{\hat{\rho}_N,\bar{\rho}_N} h| \leq \|\partial_\theta h\|_{L^\infty(\theta)} \frac{1}{N}\sum_i |\theta_i - \bar{\theta}_i|.$$

Proof. Recall that empirical measures $\hat{\rho}_{N,s} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_{i,s}}$ and $\bar{\rho}_{N,s} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\bar{\theta}_{i,s}}$, so

$$|I_{\hat{\rho}_{N,s},\bar{\rho}_{N,s}} h| = \left| \frac{1}{N}\sum_{i=1}^{N} h(\theta_{i,s}) - h(\bar{\theta}_{i,s}) \right| \leq \|\partial_\theta h\|_{L^\infty(\theta)} \frac{1}{N}\sum_{i=1}^{N} |\theta_{i,s} - \bar{\theta}_{i,s}|$$

□

Lemma 2.14 (Generalized Gronwall's inequality)**.** *Suppose* $b \geq 0$ *and* $A'(t) \leq CA(t) + b(t)$*. Then*

$$y(t) \leq e^{Ct}\left(y(0) + \int_0^t b(\tau)d\tau \right)$$

3 Numerical Simulations

In this section, we introduce the numeric algorithm for (1.6) and (1.8) based on the forward Euler scheme. The problem that we want to solve is minimization problem:

$$\min_{\Theta_N \in \mathbb{R}^N} E(\Theta_N) = \min_{\Theta_N \in \mathbb{R}^N} \frac{1}{2} \int_{\Omega_x} |u_{\Theta_N}(1,x) - D(x)|^2 dx, \tag{3.1}$$

where the dynamics $u_{\Theta_N}(t,x)$ follows

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} u_{\Theta_N}(t,x) = \frac{1}{N} \sum_{i=1}^{N} f(u_{\Theta_N}(t,x), \theta_i), \quad \forall\, t \in (0,1), \\ u_{\Theta_N}(0,x) = x, \quad \forall\, x \in \Omega_x. \end{cases}$$

To solve the constrained minimization problem (3.1), we use the method of Lagrange multiplier with a Lagrangian functional defined by

$$\begin{aligned} \mathcal{L}(u, \Theta_N, \eta, \tilde{\eta}) := &\frac{1}{2} \int_{\Omega_x} |u(1,x) - D(x)|^2 dx - \int_{\Omega_x} \Big(u(0,x) - x \Big) \tilde{\eta}(x) dx \\ &- \int_0^1 \int_{\Omega_x} \left(\frac{d}{dt} u(t,x) - \frac{1}{N} \sum_{i=1}^{N} f(u(t,x), \theta_i) \right) \eta(t,x) dx dt, \end{aligned}$$

where $\eta(t,x)$ and $\tilde{\eta}(x)$ are Lagrange multipliers. For the first-order optimality condition, we set

$$\frac{\delta \mathcal{L}}{\delta \eta} = \frac{\delta \mathcal{L}}{\delta \tilde{\eta}} = \frac{\delta \mathcal{L}}{\delta u} = 0,$$

and make a flow of $\Theta_N(s)$, where s is a pseudo-time, as follows:

$$\partial_s \Theta_N(s) = -\frac{\partial \mathcal{L}}{\partial \Theta_N} = -\frac{1}{N} \sum_{i=1}^{N} \int_0^1 \int_{\Omega_x} \nabla_\Theta f(u(t,x), \theta_i(s)) \eta(t,x) dx dt. \tag{3.2}$$

Obviously, the first condition $\delta \mathcal{L} / \delta \eta$ gives a forward system

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t,x) = \frac{1}{N} \sum_{i=1}^{N} f(u(t,x), \theta_i), \quad \forall\, (t,x) \in (0,1) \times \Omega_x.$$

The second condition $\frac{\delta \mathcal{L}}{\delta \tilde{\eta}}$ provides the initial condition for the forward system

$$u(0,x) - x = 0.$$

The last one $\frac{\delta \mathcal{L}}{\delta u}$ deduces an adjoint system

$$\frac{\mathrm{d}}{\mathrm{d}t} \eta(t,x) = -\frac{1}{N} \sum_{i=1}^{N} \partial_u f(u(t,x), \theta_i), \quad \forall\, (t,x) \in (0,1) \times \Omega_x.$$

Finally, we derive the boundary condition for the adjoint system from

$$\frac{\delta \mathcal{L}}{\delta u(0,x)} = \frac{\delta \mathcal{L}}{\delta u(1,x)} = 0.$$

They suggest that

$$-\tilde{\eta}(x) + \eta(0, x) = 0,$$

and

$$u(1, x) - D(x) - \eta(1, x) = 0,$$

respectively. In short, we have the following forward and adjoint system with initial data:

$$\begin{cases} \partial_t u(t, x) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} f(u(t, x), \theta_i), \\ u(0, x) = x, \quad \forall \ (t, x) \in (0, 1) \times \Omega_x, \end{cases} \tag{3.3}$$

and

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \eta(t, x) = -\dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \partial_u f(u(t, x), \theta_i), \\ \eta(1, x) = u(1, x) - D(x), \quad \forall \ (t, x) \in (0, 1) \times \Omega_x. \end{cases}$$

With the change of variable

$$\tilde{\eta}(t, x) = \eta(1 - t, x),$$

the adjoint system can be rewritten as

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} \tilde{\eta}(t, x) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \partial_u f(u(1 - t, x), \theta_i), \\ \tilde{\eta}(0) = u(1, x) - D(x), \quad \forall \ (t, x) \in (0, 1) \times \Omega_x, \end{cases} \tag{3.4}$$

which now turns into an initial value problem so that we can use the forward Euler method.

Here, we breifly mention the iterations by the forward Euler scheme for (3.2), (3.3) and (3.4).

$$\begin{aligned} \Theta_N(s_{k+1}) &= \Theta_N(s_k) - \frac{h_s}{N} \sum_{i=1}^{N} \int_0^1 \int_{\Omega_x} \nabla_\Theta f(u(t, x), \theta_i) \eta(t, x) dx dt, \quad \forall \ k \geq 0, \\ u(t_{l+1}, x) &= u(t_l, x) + \frac{h_t}{N} \sum_{i=1}^{N} f(u(t_l, x), \theta_i), \quad \forall \ 0 \leq l < N_t, \\ \tilde{\eta}(t_{l+1}, x) &= \tilde{\eta}(t_l, x) + \frac{h_t}{N} \sum_{i=1}^{N} \partial_u f(u(1 - t_l, x), \theta_i), \quad \forall \ 0 \leq l < N_t, \end{aligned} \tag{3.5}$$

where the physical time t (for u and η) and the pseudo time s (for Θ_N) are approximated by uniform time sequences satisfying

$$\begin{aligned} &0 = t_0 < t_1 < \cdots < t_{N_t} = 1 \quad \text{with} \quad |t_{l+1} - t_l| = h_t, \\ &0 = s_0 < s_1 < s_2 < \cdots, \quad \text{with} \quad |s_{k+1} - s_k| = h_s, \end{aligned}$$

respectively. We denote the final step in time ($t = 1$) as N_t so that $h_t N_t = 1$. With this notation, the recurrence iteration for Θ_N and $\tilde{\eta}$ in (3.5) can be rewritten as

$$\Theta_N(s_{k+1}) = \Theta_N(s_k) - \frac{h_s h_t}{N} \sum_{i=1}^{N} \sum_{l=1}^{N_t} \int_{\Omega_x} \nabla_\Theta f(u(t_l, x), \theta_i) \eta(t_l, x) dx, \tag{3.6}$$

and

$$\tilde{\eta}(t_{l+1}, x) = \tilde{\eta}(t_l, x) + \frac{h_t}{N} \sum_{i=1}^{N} \partial_u f\big(u((N_t - l)h_t, x), \theta_i\big),$$

respectively.

For discretization of spatial domain Ω_x, recall the meaning of the spatial integral. It is proposed to measure all error for given all data, so for given data set $\{(x_j, u_j)\}_{j=1}^{N}$, it is natural to convert (3.6) into

$$\Theta_N(s_{k+1}) = \Theta_N(s_k) - \frac{h_s h_t}{NM} \sum_{i=1}^{N} \sum_{l=1}^{N_t} \sum_{j=1}^{M} \nabla_\Theta f(u(t_l, x_j), \theta_i) \eta(t_l, x_j).$$

In detail, in the component-wise description,

$$\theta_N^i(s_{k+1}) = \theta_N^i(s_k) - \frac{h_s h_t}{NM} \sum_{l=1}^{N_t} \sum_{j=1}^{M} \partial_\theta f(u(t_l, x_j), \theta_i) \eta(t_l, x_j).$$

Now, we discuss the initial setting for numeric simulations. To train $\{\theta_i\}_{i=1}^N$ using the gradient flow above, we need data $\{(x_j, u_j)\}_{j=1}^{M^*}$ satisfying

$$u_j = u_*(t = 1, x_j),$$

where u_* is the solution to (1.1) with the true distribution ρ_* of θ. To approximate the distribution ρ_*, we use a sufficiently larger number of N^* compared to N to generate a particle distribution $\{\theta_i^*\}_{i=1}^{N^*}$. Then, we choose the number of given data as $M \leq M^*$ and number of θ as $N \leq N^*$, and extract data $\{x_j, u_j\}_{j=1}^{M}$ from the obtained data $\{(x_j, u_j)\}_{j=1}^{M^*}$.

In the numeric simulation, we compare the results for various N:

$$N = 25,\ 50,\ 100,\ 200,$$

and choose parameters

$$h_s = 100, \quad h_t = 0.2, \quad M = 500, \quad N^* = 3000, \quad M^* = 1000.$$

The solution distribution ρ_* is the standard normal distribution and the initial data in given data $\{x_j\}_{j=1}^{M^*}$ follow i.i.d. $\mathcal{U}[0, 1]$. We stop the iteration at $1,000^{\text{th}}$ iteration. For the kernel functions, empirically, we observed that the test with radial basis kernel function, i.e.,

$$f(u, \theta) = \phi(\|u - \theta\|),$$

exhibits good results. Here, we list radial basis kernel functions that we used in numeric simulations:

$$\begin{aligned} (i)\ & f(u, \theta) = \exp(-\|u - \theta\|)' \\ (ii)\ & f(u, \theta) = \frac{1}{1 + \|u - \theta\|^2}. \end{aligned}$$

For the comparison, we The kernel function f is chosen to be

$$f(u,\theta) = \log\left(1+e^{u-\theta}\right),$$

which is the approximation of ReLU function. Or

$$f(u,\theta) = \exp\left(-(u-\theta)^2\right).$$

Proposition 3.1. *If f is bounded and Lipschitz, we have*

$$|E(\rho) - E(\mu_N)| \lesssim \frac{1}{\sqrt{N}}.$$

Proof. By the definition, we have

$$E(\rho) = \frac{1}{2}\int_{\Omega_x} |u_\rho(t=1,x) - D(x)|^2 dx,$$

and

$$E(\Theta_N) = \frac{1}{2}\int_{\Omega_x} |u_{\Theta_N}(t=1,x) - D(x)|^2 dx,$$

which are followed by

$$\begin{aligned} E(\rho) - E(\mu_N) = \int_{\Omega_x} & \frac{1}{2}\Big(u_\rho(t=1,x)^2 - u_{\mu_N}(t=1,x)^2\Big) \\ & - D(x)\Big(u_\rho(t=1,x) - u_{\mu_N}(t=1,x)\Big)dx. \end{aligned} \tag{3.7}$$

Now, we compute the difference between $\partial_t u_\rho(t,x)$ and $\partial_t u_{\mu_N}(t,x)$.

$$\begin{aligned} &\frac{d}{dt}\Big|u_\rho(t,x) - u_{\mu_N}(t,x)\Big| \\ &\quad= \left|\int_{\Omega_\theta} f(u_\rho(t,x),\theta)\rho(\theta)d\theta - \frac{1}{N}\sum_{i=1}^{N} f(u_{\mu_N}(t,x),\theta_i)\right| \\ &\quad= \left|\int_{\Omega_\theta} f(u_\rho(t,x),\theta)\rho(\theta)d\theta - \int_{\Omega_\theta} f(u_{\mu_N}(t,x),\theta)\rho(\theta)d\theta\right| \\ &\qquad+ \left|\int_{\Omega_\theta} f(u_{\mu_N}(t,x),\theta)d\rho(\theta) - \int_{\Omega_\theta} f(u_{\mu_N}(t,x),\theta)d\tilde{\mu}_N(\theta)\right| \\ &\qquad+ \left|\int_{\Omega_\theta} f(u_{\mu_N}(t,x),\theta)d\tilde{\mu}_N(\theta) - \int_{\Omega_\theta} f(u_{\mu_N}(t,x),\theta)d\mu_N(\theta)\right| \\ &\quad\le L_f|u_\rho(t,x) - u_{\mu_N}(t,x)| + \|f\|_\infty\big(\|\rho-\tilde{\mu}_N\|_{W_1} + \|\tilde{\mu}_N - \mu_N\|_{W_1}\big), \end{aligned}$$

where L_f is the Lipschitz coefficient of f. From Theorem 2.2, We have

$$\|\rho - \tilde{\mu}_N\|_{W_1} \lesssim \frac{1}{\sqrt{N}} \quad \text{and} \quad \|\tilde{\mu}_N - \mu_N\|_{W_1} \lesssim N^{-\frac{1}{2}}.$$

Hence, by Grönwall's inequality, one can obtain

$$|u_\rho(t=1,x) - u_{\mu_N}(t=1,x)| \leq \|f\|_\infty \big(\|\rho - \tilde{\mu}_N\|_{W_1} + \|\tilde{\mu}_N - \mu_N\|_{W_1}\big) e^{L_f},$$

where we used that the initial data satisfies

$$u_\rho(t=0,x) = u_{\mu_N}(t=0,x).$$

Back to (3.7), since D and $|u_\rho + u_{\mu_N}|$ are bounded on finite spatial measure ($|\Omega_x| < \infty$), we have

$$|E(\rho) - E(\Theta_N)| \lesssim \frac{1}{\sqrt{N}},$$

which is the desired result. □

The goal of this section is as follows:

1. Compare Θ_N^* and $\Theta_N(T)$ in terms of N and T. [Joe: I'm finding comparing Θ_N^* and $\Theta_N(T)$ directly is not very useful. But comparing $U_{\Theta_N^*}$ and $U_{\Theta_N(T)}$ shows good results. I suppose this is what we want?] [Jaeyoung: As Joe's comment, we compare $U_{\Theta_N^*}$ and $U_{\Theta_N(T)}$ according to N. Still, we don't know when the distribution is recovered although the structure seems to be covered. We need to more think about f.]

2. Can we make an numeric algorithm for (1.2) and (1.4)? → No!

3. If then, compare the two numeric results.

4. Consider what would be a good choice for f and ρ_* so that we can explain how f and ρ_* effect to dynamics and also can approach to good approximation $\rho(s) \to \rho_*$.

4 Appendix

Lemma 4.1. *(Fréchet Derivative for E) Consider functionals $E : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ as defined in (1.3) and (1.7). Then, we have that*

$$\frac{\delta E}{\delta \rho}\Big|_{\rho}(\theta) = \int_0^1 \int_{\mathbb{R}} v_\rho(t,x) f(u_\rho(t,x),\theta)\, \mu(x) dx dt, \tag{4.1}$$

where $v_\rho(t,x)$ is given by (2.1).

Proof. Step 1: To study the Fréchet derivative $\frac{\delta E[\rho]}{\delta \rho}\big|_{\rho}$, we first we study how $E[\rho]$ changes under arbitrary small perturbation. Explicitly, fixing arbitrary $\tilde{\rho} \in L^2(\mathbb{R})$, we have that

$$\begin{aligned}
E[\rho + \epsilon\tilde{\rho}] &= \frac{1}{2}\int_{\mathbb{R}} (u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - D(x))^2 \mu(x) dx \\
=&\frac{1}{2}\int_{\mathbb{R}} (u_\rho(t=1,x) - D(x))^2 \mu(x) dx + \frac{1}{2}\int [u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)]^2 \mu(x) dx \\
&+ \int (u_\rho(t=1,x) - D(x))[u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)]\mu(x) dx \\
=&E[\rho] + \int_{\mathbb{R}} [u_\rho(t=1,x) - D(x)][u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)]\mu(x) dx + O(\epsilon^2),
\end{aligned}$$

which implies after a rearrangement

$$\lim_{\epsilon\to 0^+} \frac{E[\rho+\epsilon\tilde{\rho}] - E[\rho]}{\epsilon} = \int_{\mathbb{R}} (u_\rho(t=1,x) - D(x)) \lim_{\epsilon\to 0^+} \frac{u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)}{\epsilon} \mu(x) dx$$

which can be simplified to be

$$\frac{\delta E}{\delta \rho}\Big|_{\rho}(\theta) = \int_{\mathbb{R}} (u_\rho(t=1,x) - D(x)) \frac{\delta u_\rho(t=1,x)}{\delta \rho}\Big|_{\rho}(\theta) \mu(x) dx \tag{4.2}$$

where the first variation $\frac{\delta u_\rho(t=1,x)}{\delta \rho}\big|_{\rho}$, is defined in the same way as above. That is, for arbitrary $\tilde{\rho}(\theta) \in L^2(\mathbb{R})$, we have the pairing

$$\left\langle \frac{\delta u_\rho(t=1,x)}{\delta \rho}\Big|_{\rho}(\theta), \tilde{\rho}(\theta) \right\rangle_{L^2(d\theta)} = \lim_{\epsilon\to 0^+} \frac{u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)}{\epsilon}$$

and hence $\frac{\delta u_\rho(t=1,x)}{\delta \rho}\Big|_{\rho}(\theta) \in L^2(d\theta)$ is a function of θ on $\mathbb{R}$. In this expression, $t = 1$ is fixed, x is viewed as a parameter, and both $\tilde{\rho}$ and $\frac{\delta u_\rho(t=1,\tilde{x})}{\delta \rho}\Big|_{\rho}$ are function of $\theta \in \mathbb{R}$.

Above equation (4.2) shows that we now need to find an explicit formula for first variation $\frac{\delta u_\rho(t=1,x)}{\delta \rho}\big|_{\rho}(\theta)$. Our strategy is to estimate the difference $u_{\rho+\epsilon\tilde{\rho}}(t=1,x) - u_\rho(t=1,x)$ by the physical dynamics (1.2).

Step 2: We can use dynamics (1.2) to derive an explicit expression for how difference $\tilde{u}(t,\tilde{x}) := u_{\rho+\epsilon\tilde{\rho}}(t,x) - u_\rho(t,x)$ evolve over time. Notice that for each $t \in [0,1]$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}(u_{\rho+\epsilon\tilde{\rho}}(t,x) - u_\rho(t,x)) \\
=&\int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_\rho,\theta)(u_{\rho+\epsilon\tilde{\rho}}(t,x) - u_\rho(t,x))\rho(\theta)\, d\theta + \int_{\mathbb{R}} f(u_\rho,\theta)\epsilon\tilde{\rho}(\theta)\, d\theta + \mathcal{O}(\epsilon^2)
\end{aligned}$$

Therefore, by dropping higher order term, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{u} = \tilde{u}\int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_\rho, \theta)\rho(\theta)\, d\theta + \int_{\mathbb{R}} f(u_\rho, \theta)\epsilon\tilde{\rho}(\theta)\, d\theta$$

As a consequence, the above equation, together with (2.1), shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}(v_\rho \tilde{u}) = \frac{\mathrm{d}}{\mathrm{d}t}v_\rho \tilde{u} + v_\rho \frac{\mathrm{d}}{\mathrm{d}t}\tilde{u} = v_\rho \int_{\mathbb{R}} f(u_\rho, \theta)\epsilon\tilde{\rho}\, d\theta$$

We integrate above over time $t \in [0, 1]$ and notice that initial condition $\tilde{u}(t = 0, x) = 0$ because we have same initial condition for both $u_\rho(t = 0, x)$ and $u_{\rho+\epsilon\tilde{\rho}}(t = 0, x)$. Hence, we know that

$$\begin{aligned} &(D(x) - u_\rho(t = 1, x))u_{\rho+\epsilon\tilde{\rho}}(t = 1, x) - u_\rho(t = 1, x) \\ =&v_\rho(t = 1, x)\tilde{u}(t = 1, x) \\ =&\int_{\mathbb{R}} \left(\int_0^1 v_\rho(t, x) f(u_\rho, \theta) dt \right) \epsilon\tilde{\rho}(\theta)\, d\theta \end{aligned}$$

It follows that definition of first variation that we have

$$(D(x) - u_\rho(t = 1, x))\frac{\delta u_\rho(t = 1, x)}{\delta \rho}\Big|_\rho = \int_0^1 v_\rho(t, x) f(u_\rho, \theta) dt$$

This expression depends only on x and θ. We substitute above in (4.2) to obtain (2.2). □

References

[FG13] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013.