Mean field limit in inverse problem

Yutian He, Joseph Hunter, Vasily Ilin, Ian McPherson, Kouakou Innocent NDRI, Yantao Wu, Jaeyoung Yoon

SLMath: Particle interactive systems

28th June 2024

Problem Setup

Problem Setup: Mean-field version

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

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

1. We are given the "activation function" f and data $D(x) = u_{\rho_*}(t=1,x)$.
2. Want to approximate the distribution of "weights" $\rho_*(\theta)$.

Solution: Minimizing loss functional

$$E[\rho] = \frac{1}{2}\int_{\Omega_x} |u_\rho(1,x) - D(x)|^2 \, d\mu(x)$$

Problem Setup: Particle version

Function $u_{\rho_N} : [0,1] \times \mathbb{R} \to \mathbb{R}$ has the following dynamics

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} u_{\rho_N} = \frac{1}{N} \sum_{i=1}^{N} f(u_{\rho_N}, \theta_i), \quad t \in [0,1] \\ u_{\rho_N}(t=0, x) = x \end{cases} \tag{2}$$

Use the particle approximation

$$\rho_{N,s} = \frac{1}{N} \sum_{i} \delta_{\theta_i(s)}$$

$$\frac{d}{ds}\theta_i(s) = -\partial_{\theta_i} E[\rho_{N,s}] = W_{\rho_{N,s}}(\theta_i)$$

$$= -\int_0^1 \int v_{\rho_{N,s}}(t,x) \partial_{\theta_i} f(u_{\rho_{N,s}}, \theta) dxdt, \quad v = \text{adjoint term}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} v_{\rho_{N,s}} = -v_{\rho_{N,s}} \int_{\mathbb{R}} \frac{\partial f}{\partial u}(u_{\rho_{N,s}}, \theta) \rho_{N,s}(\theta)\, d\theta, \quad v_{\rho_{N,s}}(t=1) = D - u_{\rho_{N,s}}(t=1).$$

Interpretation via Neural ODE

Dynamics (2) describes a Neural Network:

1. Layer $t \in [0, 1]$ and input x.
2. $u_{\rho_N}(t, x)$ = output from t-th layer. First layer is identity $u(t = 0, x) = x$.
3. Minimize sum of squares $E[\rho_N] = \int |D(x) - u_{\rho_N}(t = 1, x)|^2 d\mu(x)$.
4. Residual network architecture, finite width N, infinitely deep, layer-wise homogeneous neural network.

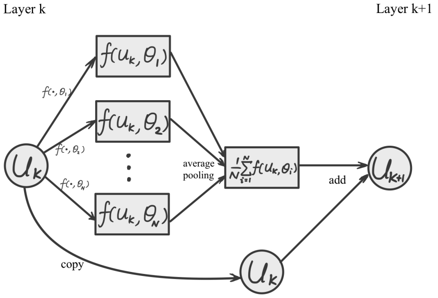

Figure: One layer of NN

Gradient Flows

Wasserstein Gradient Flow

$$\begin{cases} \partial_s \rho_s = -\nabla_{\mathcal{W}_2} E[\rho_s] = \partial_\theta \left(\rho_s \partial_\theta \left. \frac{\delta E}{\delta \rho} \right|_{\rho_s} \right), \quad s \in [0, T] \\ \rho_{s=0} = \rho_0 \end{cases} \tag{3}$$

Euclidean Gradient Flow

$$\begin{cases} \partial_s \Theta_s = -\nabla_\Theta E[\rho_{N,s}] \quad s \in [0, T] \\ \Theta_{s=0} = \Theta_0 \end{cases} \tag{4}$$

Main Goal: Compare how empirical measure

$$\rho_{N,s} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_{i,s}}$$

approximates ρ_s as $N \to \infty$.
Remark: convergence as $s \to \infty$ is a different story.

Theoretical Analysis

Main Theorem

Theorem (Mean-field limit)

Assume the following finiteness conditions on f and ρ_s:

$$C_f^0 = \|\partial_\theta f(y,\theta)\|_{L^\infty(y,\theta)} + \|\partial_y f(y,\theta)\|_{L^\infty(y,\theta)} < \infty$$

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

$$\mathbb{E}\left(E[\rho_s] + E[\rho_{N,s}] + E[\bar{\rho}_{N,s}]\right) < \infty, \quad E[\rho] = \int \left|D(x) - u_{\rho_s}(t=1,x)\right|^2 d\mu(x)$$

$$\int |\theta|^5 d\rho_s(\theta) < \infty.$$ *Then, for all $s \in [0,T]$*

$$\mathbb{E}\mathcal{W}_2(\rho_s, \rho_{N,s}) \le \frac{C}{N^{1/4}} + \frac{C''}{N^{1/2}} \exp\left(C's\right) = \mathcal{O}(N^{-1/4}),$$

$$\mathbb{E}\mathcal{W}_1(\rho_s, \rho_{N,s}) = \mathcal{O}(N^{-1/2})$$

$$C = \text{universal}, \quad C' = 16\left(C_f^1\right)^2 e^{5C_f^0}\mathbb{E}\left(E[\rho_s]^{\frac{1}{2}} + E[\rho_{N,s}]^{\frac{1}{2}} + E[\bar{\rho}_{N,s}]^{\frac{1}{2}}\right) + 1$$

$$C'' = C'\left\|f(u_\rho,\theta)\right\|_{L^4(\mu(x)\otimes\rho(x))}$$

Proof sketch

$$\frac{d}{ds}|\theta_i - \bar{\theta}_i| = |W_{\hat{\rho}_N}(\theta_i) - W_{\rho_s}(\bar{\theta}_i)|$$
$$\leq \left|W_{\hat{\rho}_N}(\theta_i) - W_{\hat{\rho}_N}(\bar{\theta}_i)\right| + \left|W_{\hat{\rho}_N}(\bar{\theta}_i) - W_{\bar{\rho}_N}(\bar{\theta}_i)\right| + \left|W_{\bar{\rho}_N}(\bar{\theta}_i) - W_{\rho_s}(\bar{\theta}_i)\right|$$
$$\leq L_A|\theta_i - \bar{\theta}_i| + L_B\left(\frac{1}{N}\sum_{j=1}^{N}|\theta_j - \bar{\theta}_j|\right) + L_C\frac{1}{\sqrt{N}}$$

Sum over i and combine Lipshitz constants L_A, L_B, L_C:

$$\frac{d}{ds}\left(\frac{1}{N}\sum_{i=1}^{N}|\theta_{i,s} - \bar{\theta}_{i,s}|^2\right) \leq \frac{(C'')^2}{N} + C'\left(\frac{1}{N}\sum_{i=1}^{N}|\theta_{i,s} - \bar{\theta}_{i,s}|^2\right)$$

By Grönwall's inequality,

$$\mathcal{W}_2^2(\bar{\rho}_{N,s}, \rho_{N,s}) \leq \frac{1}{N}\sum_{i=1}^{N}|\theta_{i,s} - \bar{\theta}_{i,s}|^2 \leq \frac{(C'')^2}{N}\exp(C's)$$

$$\mathcal{W}_2(\rho_s, \rho_{N,s}) \leq \mathcal{W}_2(\rho_s, \bar{\rho}_{N,s}) + \mathcal{W}_2(\bar{\rho}_{N,s}, \rho_{N,s}) \leq \frac{C}{N^{1/4}} + \frac{C''}{N^{1/2}}\exp(C's)$$

Lipschitz lemmas

Lemma (Flow Velocity $W_\rho(\theta)$ is Lipschitz in θ)

$$|W_{\rho_{N,s}}(\theta_1) - W_{\rho_{N,s}}(\theta_2)| \le 2C_f^1 e^{C_f^0} E[\rho_{N,s}]^{\frac{1}{2}} |\theta_1 - \theta_2|.$$

Lemma (Flow Velocity $W_\rho(\theta)$ is Lipschitz-Like in $\hat{\rho}_N, \bar{\rho}_N$)

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

Lemma (Flow Velocity $W_\rho(\theta)$ is Lipschitz-Like in $\rho, \bar{\rho}_N$)

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

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

Numeric analysis

Minimization problem

In numeric, we can not have a full information for a distribution ρ. Instead, we approximate $\rho(\theta)$ with $\rho_N = \frac{1}{N}\sum_i \delta_{\theta_i}$ and solve the following minimization problem:

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

where the dynamic $u_{\rho_N}(t, x)$ follows

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

Lagrange Multiplier

Since it is hard to compute $\nabla_{\Theta_N} E(\Theta_N)$ directly, we define a Lagrangian functional $\mathcal{L}$ by

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

where $\eta(t,x)$ and $\tilde{\eta}(x)$ are Lagrange multipliers. Note that the first-order optimality condition is

$$\frac{\delta\mathcal{L}}{\delta\eta} = \frac{\delta\mathcal{L}}{\delta\tilde{\eta}} = \frac{\delta\mathcal{L}}{\delta u(t,x)} = \frac{\delta\mathcal{L}}{\delta u(0,x)} = \frac{\delta\mathcal{L}}{\delta u(1,x)} = \frac{\partial\mathcal{L}}{\partial\Theta_N} = 0.$$

Strategy

While keeping

$$\frac{\delta \mathcal{L}}{\delta \eta} = \frac{\delta \mathcal{L}}{\delta \tilde{\eta}} = \frac{\delta \mathcal{L}}{\delta u(t,x)} = \frac{\delta \mathcal{L}}{\delta u(0,x)} = \frac{\delta \mathcal{L}}{\delta u(1,x)} = 0, \tag{5}$$

make a gradient flow for $\Theta_N(s)$ as

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

The conditions (5) gives the forward system (original constraint) and adjoint system:

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

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

Since it is a final value problem, we convert it into an initial value problem,

$$
\text{(A')} \quad \begin{cases} \partial_t \tilde{\eta}(t,x) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \partial_u f(u(1-t,x),\theta_i), \\ \tilde{\eta}(0,x) = u(1,x) - D(x), \quad \forall\,(t,x) \in (0,1) \times \Omega_x, \end{cases}
$$

where we used the change of variable

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

Algorithm at a glance

1. Flow for Θ in pseudo time s:

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

2. Forward system in physical time t:

$$u(t_{l+1}, x) = u(t_l, x) + \frac{h_t}{N} \sum_{i=1}^{N} f(u(t_l, x), \theta_i), \quad u(0, x) = x,$$

3. Adjoint system in physical time t:

$$\tilde{\eta}(t_{l+1}, x) = \tilde{\eta}(t_l, x) + \frac{h_t}{N} \sum_{i=1}^{N} \partial_u f\big(u\big((N_t - l)h_t, x\big), \theta_i\big),$$
$$\tilde{\eta}(0, x) = u(1, x) - D(x).$$

- Uniform time sequences $\{t_l\}_{l=0}^{N_t}$ and $\{s_k\}_{k=0}^{N_s}$ satisfy

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

How can we observe the convergence as $N \to \infty$?

From now on, we use the measure notation ρ_N defined by

$$\rho_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i(s)}.$$

In the best way, one can consider, for some $\rho \in \mathcal{P}(\Omega_\theta)$ and some metric d,

$$d(\rho_N, \rho) \lesssim N^{-\alpha}, \quad \text{for some } \alpha > 0.$$

In numeric, it produces **high-cost** in computation. Instead, we would like to see the convergence indirectly using cost function E. Roughly speaking, we want to observe,

$$|E(\rho) - E(\rho_N)| \lesssim N^{-\beta}, \quad \text{for some } \beta > 0.$$

Which β?

Lemma

If f is bounded and Lipschitz, then we have

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

Proof. We use

$$\|\rho - \bar{\rho}_N\|_{W_1} \lesssim \frac{1}{\sqrt{N}} \quad \text{and} \quad \|\bar{\rho}_N - \rho_N\|_{W_1} \lesssim \frac{1}{\sqrt{N}}.$$

and Grönwall's inequality to derive

$$|u_\rho(t=1,x) - u_{\rho_N}(t=1,x)| \lesssim \frac{1}{\sqrt{N}}.$$

Thanks to finite spatial measure ($|\Omega_x| < \infty$), we get the desired result. □

Numerical Results

What do we observe in simulation?

Goal: Observe $|E(\rho) - E(\rho_N)| \lesssim \frac{1}{\sqrt{N}}$.

We consider a scalar

$$\alpha_N(s) := \frac{\sqrt{2N}|E(\rho_N) - E(\rho_{2N})|}{\sqrt{2} - 1} \tag{6}$$

and we claim that $\alpha < C$ for any N.
If it is true, then

$$|E(\rho_N) - E(\rho)| < \frac{C}{\sqrt{N}}.$$

We want to show in simulation that $\alpha_N(s) < C$ for all N. For example,

$$\alpha_{25}, \alpha_{50}, \alpha_{100}, \alpha_{200} < C.$$

Test Cases and Setup

- Goal: train $\{\theta_i\}_{i=1}^{N}$ to recover the true distribution $\rho_*(\theta) = \mathcal{N}(0,1)$.
- We approximate ρ_* with $\{\theta_i^*\}_{i=1}^{N^*}$, where $N^* = 3000$ and $\theta_i^* \sim \rho_*$.
- We generate the target data $\{(x_j, u_j)\}_{j=1}^{M^*}$ with $M^* = 1000$, where $u_j = u_*(t=1, x_j)$.
- We then randomly select $M = 500$ samples from the target data set $\{(x_j, u_j)\}_{j=1}^{M^*}$ to train on.
- We initialize $\{\theta_i\}_{i=1}^{N}$ such that $\theta_i \sim \mathcal{U}[0,1]$.
- For $f(u,\theta)$, we choose radial basis functions (RBF). That is $f(u,\theta) = \phi(|u-\theta|)$. Specifically we test $f(u,\theta) = \exp(-|u-\theta|^2)$ and $f(u,\theta) = \frac{1}{1+|u-\theta|^2}$.
- We use a forward Euler timestep of $h_t = 0.1$. We advance Θ_N by its flow map using a step size of $h_s = 100$ (large, but results show it is not unstable). We iterate the system $10,000$ times to allow Θ_N to stabilize for all $N = 25, 50, 100, 200, 400$.
- For each N we run thirty trials and average the results to obtain the final parameter set Θ_N.

Test Case: $f(u,\theta) = \exp(-(u-\theta)^2)$

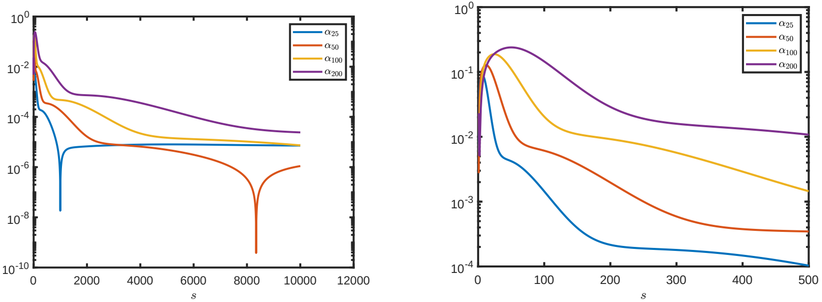

Figure: α (y-axis) versus pseudo-time s (x-axis) for $N = 25, 50, 100, 200$. The right plot zooms in on the region $s \in [0, 500]$.

Test Case: $f(u,\theta) = \exp(-(u-\theta)^2)$

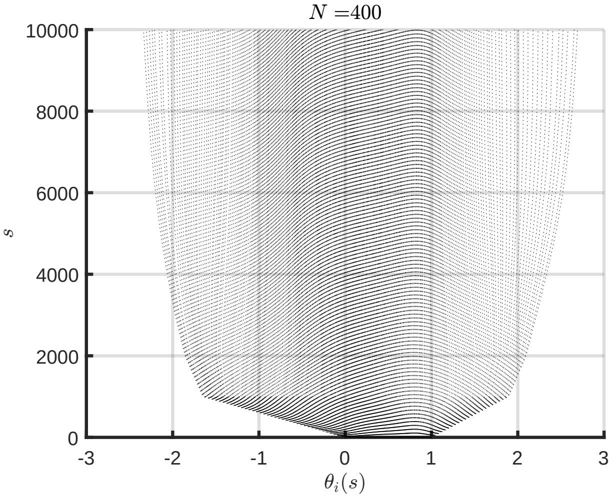

Figure: Trajectories of θ_i (x-axis) as a function of pseudo-time s (y-axis), $N = 400$, $\theta(s=0) \sim \mathcal{U}[0,1]$, $\rho_* = \mathcal{N}(0,1)$.

Test Case: $f(u, \theta) = \exp(-(u-\theta)^2)$

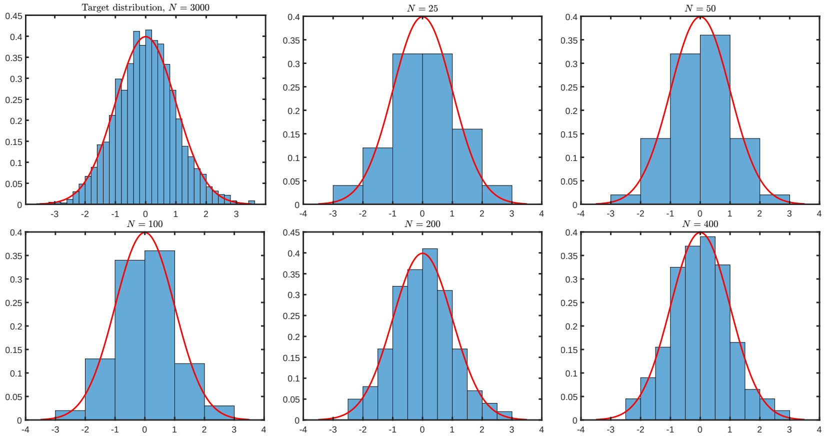

Figure: Convergence $\rho_N \to \rho_* = \mathcal{N}(0,1)$, $N = 25, 50, 100, 200, 400$.

Test Case: $f(u,\theta) = \frac{1}{1+(u-\theta)^2}$

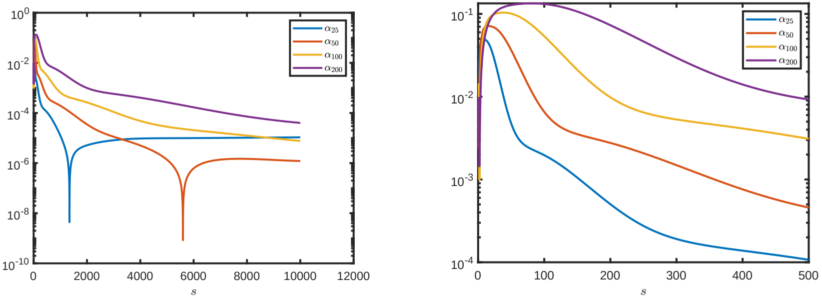

Figure: α (y-axis) versus pseudo-time s (x-axis) for $N = 25, 50, 100, 200$. The right plot zooms in on the region $s \in [0, 500]$.

Test Case: $f(u,\theta) = \frac{1}{1+(u-\theta)^2}$

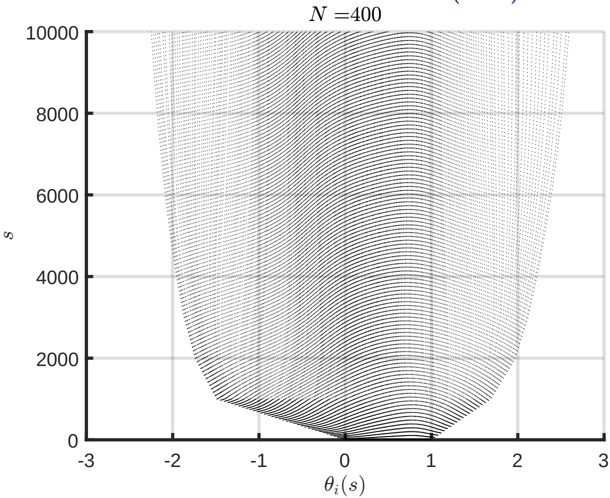

Figure: Trajectories of θ_i (x-axis) as a function of pseudo-time s (y-axis), $N = 400$, $\theta(s=0) \sim \mathcal{U}[0,1]$, $\rho_* = \mathcal{N}(0,1)$.

Test Case: $f(u,\theta) = \frac{1}{1+(u-\theta)^2}$

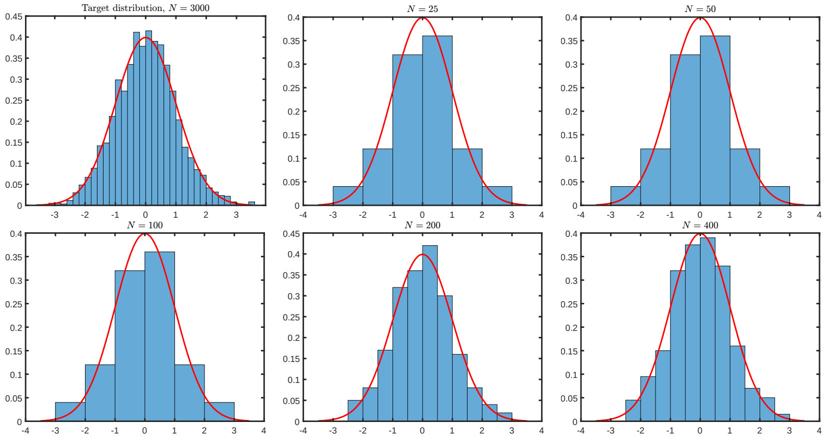

Figure: Convergence $\rho_N \to \rho_* = \mathcal{N}(0,1)$, $N = 25, 50, 100, 200, 400$.

Thank you !