Improved Finite-Particle Convergence Rates for Stein Variational Gradient Descent

krishna balasubramanian (uc davis)
Sayan Banerjee (UNC Chapel Hill)
Promit Ghosal (U Chicago)

ICLR, 2025

Sample from density $\pi \propto e^{-V}$
where potential $V : \mathbb{R}^d \to \mathbb{R}$.

- Consider N particles $x_1, \ldots, x_N$ and a symmetric, positive definite, bi-variate function $K(\cdot, \cdot)$.

- The SVGD update is given by

$$x_i \leftarrow x_i - \frac{\eta}{N} \left(\sum_{j=1}^{N} K(x_i, x_j) \nabla V(x_j) - \sum_{j=1}^{N} \nabla K(x_i, x_j) \right)$$

  where η is the step size.

- Deterministic interacting particle system (as opposed to randomized MCMC algorithms).

Particle Flow: The small step-size limit ($\eta \to 0$) takes the following form:

$$\frac{dx_i(t)}{dt} = \underbrace{\frac{1}{N}\sum_{j=1}^{N} \nabla_2 K(x_i(t), x_j(t))}_{\textit{Repulsive term}} - \underbrace{\frac{1}{N}\sum_{j=1}^{N} K(x_i(t), x_j(t)) \nabla V(x_j(t))}_{\textit{Weighted average of potential}}$$

Mean-field Limit: As $N \to \infty$, empirical distribution of $x_i(t)$'s converges weakly to μ_t (Lu, Lu, Nolen, 2021) where $\mu_t(\boldsymbol{x}) = \rho_t(\boldsymbol{x})d\boldsymbol{x}$ and,

$$\partial_t \rho_t(\boldsymbol{x}) = \underbrace{\nabla \cdot \left(\rho_t(\boldsymbol{x}) \int K(\boldsymbol{x}, \boldsymbol{y}) \big(\nabla \rho_t(\boldsymbol{y}) + \rho_t(\boldsymbol{y}) \nabla V(\boldsymbol{y}) \big) d\boldsymbol{y} \right)}_{\nabla \cdot (\nabla \rho_t(\boldsymbol{x}) + \rho_t(\boldsymbol{x}) \nabla V(\boldsymbol{x})) = \text{WGF}}$$

The above PDE is the gradient flow of KL divergence w.r.t. the Stein metric in the space of probability distributions.

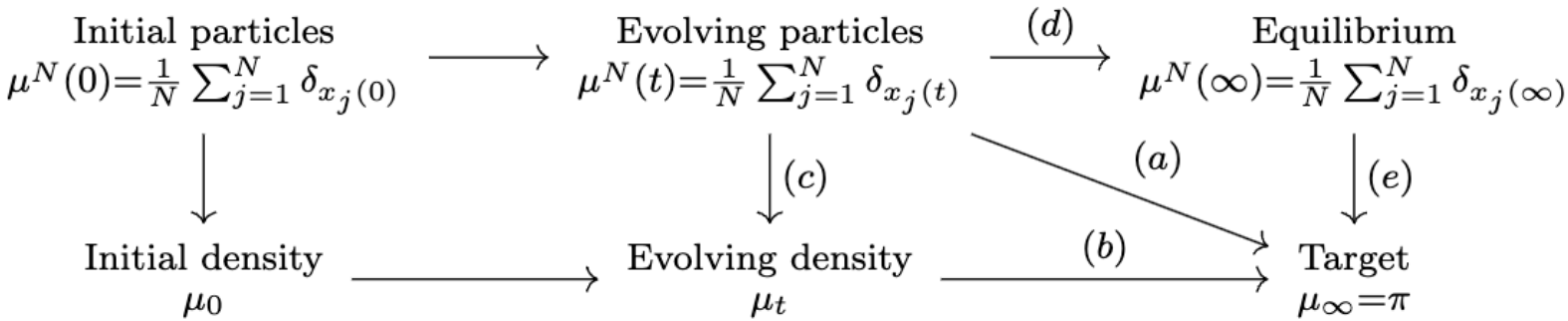
Particle Flow: The small step-size limit ($\eta \to 0$) takes the following form:

$$\frac{dx_i(t)}{dt} = \underbrace{\frac{1}{N}\sum_{j=1}^{N} \nabla_2 K(x_i(t), x_j(t))}_{\textit{Repulsive term}} - \underbrace{\frac{1}{N}\sum_{j=1}^{N} K(x_i(t), x_j(t)) \nabla V(x_j(t))}_{\textit{Weighted average of potential}}$$

Mean-field Limit: As $N \to \infty$, empirical distribution of $x_i(t)$'s converges weakly to μ_t (Lu, Lu, Nolen, 2021) where $\mu_t(\boldsymbol{x}) = \rho_t(\boldsymbol{x})d\boldsymbol{x}$ and,

$$\partial_t \rho_t(\boldsymbol{x}) = \underbrace{\nabla \cdot \left(\rho_t(\boldsymbol{x}) \int K(\boldsymbol{x}, \boldsymbol{y}) \big(\nabla \rho_t(\boldsymbol{y}) + \rho_t(\boldsymbol{y}) \nabla V(\boldsymbol{y}) \big) d\boldsymbol{y} \right)}_{\nabla \cdot (\nabla \rho_t(\boldsymbol{x}) + \rho_t(\boldsymbol{x}) \nabla V(\boldsymbol{x})) = \text{ WGF}}$$

The above PDE is the gradient flow of KL divergence w.r.t. the Stein metric in the space of probability distributions.



(a) Unified convergence of the empirical measure for $N < \infty$ particles to the continuous target as time t and N jointly grow to infinity;
(b) Convergence of mean-field SVGD to the target distribution over time;
(c) Convergence of the empirical measure for finite particles to the mean-field distribution at any finite given time $t \in [0, \infty)$;
(d) Convergence of finite-particle SVGD to equilibrium over time;
(e) Convergence of the empirical measure for finite particles to the continuous target at time $t = \infty$.

- Most of the prior approach towards this was based on combining (b) with (c).

- Liu (2017), Korba et al. (2020), Chewi et al. (2020), Salim et al. (2022), Sun et al. (2023) and Duncan et al. (2023) showed the convergence of mean-field SVGD in (b).

- Lu et al. (2019), Gorham et al. (2020) and Korba et al. (2020) obtained time-dependent mean-field convergence in (c) under various assumptions using techniques from partial differential equations.

- Shi & Mackey (2024) obtained refined results for (c) and combined them with (b) to get the *first unified convergence* (a) in terms of KSD.

- Rate is rather slow rate of order $1/\sqrt{\log \log N}$.
  - For the finite-particle versions there is no gradient structure to the dynamics.
  - This leads to the slow convergence rate.

- For Nonparametric SVGD, we directly characterize (a).
  - Key insight: track evolution of joint density rather than the empirical measure.

- Tractability of the mean-field SVGD: Has a (projected) gradient structure which leads to the following monotonicity property of the KL-divergence:

$$\partial_t \mathrm{KL}(\mu_t||\pi) = -\mathrm{KSD}^2(\mu_t||\pi), \quad t \geq 0.$$

- The non-negativity of KL then leads to bounds on the KSD.

- KSD stands for Kernel Stein Discrepancy (widely used in ML/MCMC communities). Intuitively: kernelized version of Fisher information.

- Key insight: Track the evolution of relative entropy of *joint density* of the N-particles with respect to $\pi^{\otimes N}$.

- Result: Time derivative of this relative entropy has:
  - Negative part: N times the expected KSD^2 of the empirical measure at time t with respect to π.

  - Positive part: can be separately handled and shown to be small in comparison to the negative part

- Novel connection between the joint particle dynamics and the empirical measure evolution mimicking the mean-field system.

Theorem: Define

$$\mu^N(t) := \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i(t)}, t \geq 0 \quad \text{and}$$

$$\mu^N_{\text{av}}(dx) := \frac{1}{N}\int_0^{T=N} \mu^N(t, dx)dt.$$

Then

$$\text{KSD}^2(\mu^{(N)}_{\text{av}} \| \pi) \leq \frac{\sup_L \frac{\text{KL}(p^L(0)||\pi^{\otimes L})}{L} + C^*}{N},$$

$$C^* := \sup_z \left(\nabla_2 K(z,z) \cdot \nabla V(z) + K(z,z)\Delta V(z) - \Delta_2 K(z,z)\right)$$

We make the following regularity assumptions.

- ▷ Kernel K and up to order 2 derivatives are bounded.
- ▷ For some $A > 0$, $\alpha \in [0, 1/2]$, $\|\nabla V(x)\| \leq AV(x)^{\alpha}$ for all $x \in \mathbb{R}^d$. Parameter α interpolates between exponential tails and Gaussian tails as α varies from 0 to $1/2$.
- ▷ Initial entropy bound: $\mathrm{KL}(p(0)||\pi^{\otimes N}) \lesssim Nd$ for some constant $C_{KL} > 0$.

Theorem: Setting

$$\eta \approx d^{-(\frac{1+\alpha}{2(1-\alpha)} \vee 1)} N^{-\frac{1+\alpha}{1-\alpha}} \qquad T \approx N^{\frac{2}{1-\alpha}},$$

we have

$$\frac{1}{T} \sum_{n=0}^{T-1} \mathrm{KSD}^2 \left(\mu^N(n) || \pi \right) = O\left(\frac{d^{(\frac{3-\alpha}{2(1-\alpha)} \vee 1)}}{N} \right).$$

Define average occupancy measure of k particles out of total N particles as $\bar{\mu}_k^N(dx_1, \ldots, dx_k) =$

$$\frac{1}{N} \int_0^N \mathbb{P}(x_1(t) = x_1 + dx_1, \ldots, x_k(t) = x_k + dx_k) dt.$$

Theorem: $W_2^2(\bar{\mu}_k^N, \pi^{\otimes k}) \to 0$ as $N \to 0$.

- ▷ POC: The particle marginal laws at a fixed time become asymptotically independent as $N \to \infty$.
- ▷ Prior works show POC only for "short-times", i.e., $t = t_N = O(\log \log N)$ when the particle marginal laws are not necessarily close to π.
- ▷ Our result POC extends to the time interval $[0, N]$ ("long-time" POC) and the time-averaged particle trajectories essentially produce i.i.d. samples from π.

OPEN QUESTION

- Is there uniform-in-time propagation for chaos for SVGD?

Thank you!
Question? Email: kbala@ucdavis.edu