The Scaling Limit of Stein Variational Gradient Descent

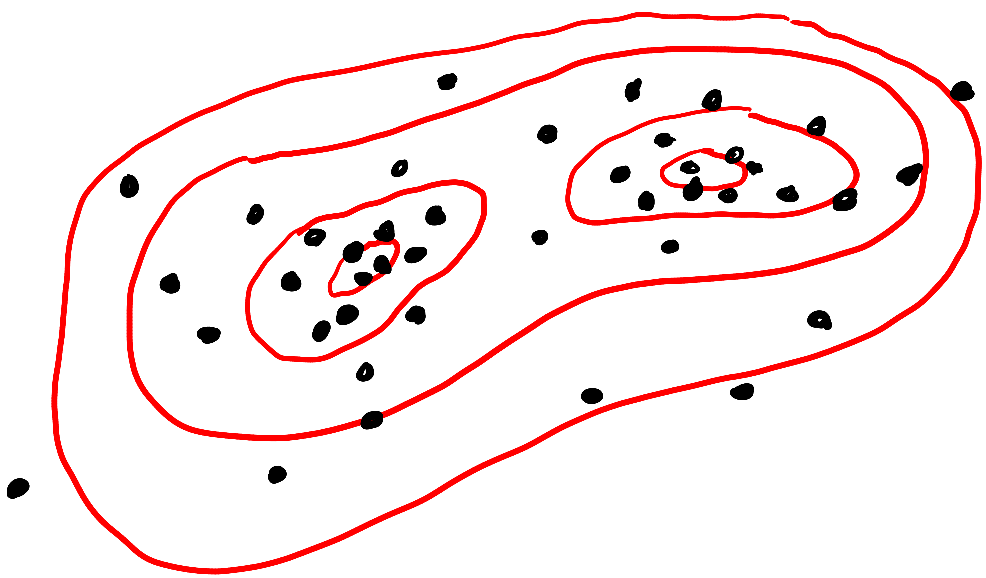
Jim Nolen
Mathematics Department
Duke University

In collaboration with Jianfeng Lu and Yulong Lu

Suppose $\bar{\rho}(x) = \frac{1}{Z} e^{-V(x)}$ is a probability density on $\mathbb{R}^d$, but Z is unknown.

How can we distribute points $x_1, \ldots x_N \in \mathbb{R}^d$ so that

$$\mu^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i} \approx \bar{\rho}(x)\, dx \quad ?$$



Stein Variational Gradient Descent was proposed in context of machine learning and Bayesian posterior approximation, as a deterministic algorithm for distributing the points $x_1, \ldots, x_N$:

$$\frac{d}{dt} x_i(t) = -\frac{1}{N} \sum_{\ell=1}^{N} \nabla K(x_i - x_\ell) \; - \frac{1}{N} \sum_{\ell=1}^{N} K(x_i - x_\ell) \nabla V(x_\ell), \quad i = 1, \ldots, N$$

$\bar{\rho}(x) = \frac{1}{Z} e^{-V(x)}$.

$K(x) : \mathbb{R}^d \to \mathbb{R}$ is a smooth, positive-definite kernel (e.g. a Gaussian).

Q. Liu and D. Wang, NIPS 2016, Q. Liu, NIPS 2017.

Let $\{x_i(t)\}_{i=1}^N \subset \mathbb{R}^d$ solve

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^N \nabla K(x_i - x_\ell) \ - \frac{1}{N}\sum_{\ell=1}^N K(x_i - x_\ell)\nabla V(x_\ell), \quad i = 1, \dots, N$$

$\bar{\rho}(x) = \frac{1}{Z}e^{-V(x)}$.

$K(x) : \mathbb{R}^d \to \mathbb{R}$ is a smooth, positive-definite kernel (e.g. a Gaussian).

Let $\{x_i(t)\}_{i=1}^N \subset \mathbb{R}^d$ solve

$$\frac{d}{dt}x_i(t) = -\mathbf{\frac{1}{N}\sum_{\ell=1}^{N} \nabla K(x_i - x_\ell)} \; - \frac{1}{N}\sum_{\ell=1}^{N} K(x_i - x_\ell)\nabla V(x_\ell), \quad i = 1, \ldots, N$$

$\bar{\rho}(x) = \frac{1}{Z}e^{-V(x)}$.

$K(x) : \mathbb{R}^d \to \mathbb{R}$ is a smooth, positive-definite kernel (e.g. a Gaussian).

The first term in the ODE is $-\nabla_i E(\mathbf{x})$, where E is the interaction energy

$$E(\mathbf{x}) = \frac{1}{N}\sum_{i<j} K(x_i - x_j).$$

Let $\{x_i(t)\}_{i=1}^N \subset \mathbb{R}^d$ solve

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^{N} \nabla K(x_i - x_\ell) \mathbf{-\frac{1}{N}\sum_{\ell=1}^{N} K(x_i - x_\ell)\nabla V(x_\ell)}, \quad i = 1, \dots, N$$

$\bar{\rho}(x) = \frac{1}{Z}e^{-V(x)}$.

$K(x) : \mathbb{R}^d \to \mathbb{R}$ is a smooth, positive-definite kernel (e.g. Gaussian).

The first term in the ODE is $-\nabla_i E(\mathbf{x})$, where E is the interaction energy

$$E(\mathbf{x}) = \frac{1}{N}\sum_{i<j} K(x_i - x_j).$$

The second term in the ODE is an average of $-\nabla V$

$$-\int K(x_i - y)\nabla V(y) d\mu^N(y)$$

Compare this to overdamped Langevin dynamics:

$$dX_i(t) = \sqrt{2}dB_i(t) - \nabla V(X_i)dt$$

for which $\bar{\rho}(x) \sim e^{-V}$ is an invariant distribution. Fokker-Planck equation for the density of a particle:

$$\partial_t q = \Delta q + \nabla \cdot (q \nabla V) \qquad (*)$$

(*) corresponds to the gradient flow for relative entropy (KL-divergence)

$$q \mapsto \mathrm{Ent}(q \mid \bar{\rho}) = \int_{\mathbb{R}^d} q(x) \ln \left(\frac{q(x)}{\bar{\rho}(x)} \right) dx \geq 0.$$

with respect to Wasserstein-2 metric.

SVGD also has formal structure of gradient flow, but with respect to a different metric, involving a RKHS with kernel K.

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^{N}\nabla K(x_i - x_\ell) \; - \frac{1}{N}\sum_{\ell=1}^{N} K(x_i - x_\ell)\nabla V(x_\ell), \quad i = 1, \dots, N$$

Formal derivation of SVGD:

Suppose $\partial_t q + \nabla \cdot (qb) = 0$ for some $b(t,x) \in \mathcal{V}$. Then

$$\frac{d}{dt}\mathrm{Ent}(q \mid \bar{\rho}) = -\int q(\nabla \cdot b - b \cdot \nabla V)\, dx$$

So, choose b to optimize

$$\sup_{b \in \mathcal{V}} \int q(\nabla \cdot b - b \cdot \nabla V)\, dx$$

If $\mathcal{V}$ is an RKHS with kernel K, then the optimal b is

$$b = \nabla K * q + K * (q \nabla V)$$

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^{N} \nabla K(x_i - x_\ell) \; - \frac{1}{N}\sum_{\ell=1}^{N} K(x_i - x_\ell)\nabla V(x_\ell),$$

Questions about SVGD: Behavior as $N \to \infty$? Behavior of system as $t \to \infty$?

The scaling limit as $N \to \infty$ involves a nonlocal, nonlinear pde

$$\partial_t q = \nabla \cdot \left(q \left(\nabla K * q + K * (\nabla V q)\right)\right)$$

First result: Assuming suitable control V, DV, and D^2V as $|x| \to \infty$ (e.g. polynomial growth) and regularity of $q_0(x)$, this PDE has a unique, global classical solution with $q(0,x) = q_0(x)$.

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^{N}\nabla K(x_i - x_\ell) \; - \frac{1}{N}\sum_{\ell=1}^{N}K(x_i - x_\ell)\nabla V(x_\ell),$$

$$\mu_t^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i(t)}$$

Second result: Convergence of μ_t^N to the PDE solution as $N \to \infty$. Let $q(t,x)$ satisfy

$$\partial_t q = \nabla \cdot \left(q\left(\nabla K * q + K * (\nabla V q)\right)\right), \qquad q(0,x) = q_0(x)$$

Then

$$\sup_{t\in[0,T]} \mathcal{W}_p(\mu_t^N \,,\; q(t,\cdot)) \le C\mathcal{W}_p(\mu_0^N \,,\; q_0(\cdot))$$

Third result: Convergence of the PDE solution as $t \to \infty$.

$$\begin{aligned}\partial_t q &= \nabla \cdot \left(q \left(\nabla K * q + K * (\nabla V q)\right)\right) \\ &= \nabla \cdot \left(qK * \left(q \nabla \log \left(\frac{q}{\bar{\rho}}\right)\right)\right)\end{aligned}$$

Assume the kernel K is Gaussian and that $\mathrm{Ent}(q_0 \mid \bar{\rho}) < \infty$. Then

$$q(t,x) \to \bar{\rho} = \frac{1}{Z} e^{-V(x)}$$

weakly as $t \to \infty$.

$\mathrm{Ent}(q \mid \bar{\rho})$ is a Lyapunov function, but we lack a Poincaré or log-Sobolev type inequality to get a rate of convergence.

$$\frac{d}{dt}\mathrm{Ent}(q \mid \bar{\rho}) = -\iint \left(q \nabla \log \frac{q}{\bar{\rho}}\right)(x)\; K(x-y)\; \left(q \nabla \log \frac{q}{\bar{\rho}}\right)(y)\, dx\, dy$$

Unresolved Issues

1. Large time behavior of the particle system. The finite particle system doesn't have gradient structure, and there may be multiple stationary solutions.

$$\frac{d}{dt}x_i(t) = -\frac{1}{N}\sum_{\ell=1}^{N}\nabla K(x_i - x_\ell) \; - \frac{1}{N}\sum_{\ell=1}^{N} K(x_i - x_\ell)\nabla V(x_\ell)$$

2. Rates of convergence for the non-local, nonlinear PDE as $t \to \infty$. Formally, when $K = \delta_0$, equation takes the form

$$\partial_t q = \nabla \cdot (q \nabla q) + \nabla \cdot \left(q^2 \nabla V\right)$$

A related work: The "Blob Method" for the Fokker-Planck equation

$$\partial_t q = \Delta q + \nabla \cdot (q \nabla V)$$

is based on the regulariztion

$$\mathrm{Ent}_\epsilon(q \mid \bar{\rho}) = \int_{\mathbb{R}^d} q(x) \ln \left(\frac{\eta_\epsilon * q(x)}{\bar{\rho}(x)} \right) \, dx$$

For this functional, Wasserstein-2 gradient flow perserves atomic measures, but $\bar{\rho}$ is not invariant. Evolution is described by

$$\partial_t q = \nabla \cdot \left(q \left(\nabla \eta_\epsilon * \left(\frac{q}{\eta_\epsilon * q} \right) + \frac{\nabla \eta_\epsilon * q}{\eta_\epsilon q} \right) \right) + \nabla \cdot (q \nabla V)$$

J. Carrillo, K. Craig, S. Patacchini Francesco (2017).

This is the end!

References:

- J. Lu, Y. Lu, J. Nolen, arxiv:1805.04035, 2018
- Q. Liu and D. Wang, *Stein variational gradient descent: A general purpose bayesian inference algorithm*, NIPS 2016.
- Q. Liu, *Stein variational gradient descent as gradient flow* NIPS 2017.
- J. Carrillo, K. Craig, S. Patacchini Francesco, *A blob method for diffusion* arXiv:1709.09195, 2017.