PDE analysis for sampling dynamics and generative models

Jianfeng Lu (鲁剑锋)

Duke University
jianfeng@math.duke.edu

IPAM, UCLA, April 2020

Joint with
Yu Cao (Duke → Courant Institute)
Yulong Lu (Duke → UMass Amherst)
Lihan Wang (Duke)

Duke
UNIVERSITY

Sampling high dimensional probability distributions is a ubiquitous challenge in many fields:

- machine learning;
- Bayesian statistics;
- computational statistical mechanics;
- quantum many-body problems;
- ...

Popular approaches for sampling:

- Markov chain Monte Carlo, e.g., diffusion based sampling;
- Interacting samplers, e.g., particle filters;
- Generative models, e.g., normalizing flows, generative adversarial networks.

In this talk, we will discuss two recent sampling-related works:

Convergence to equilibrium of Langevin dynamics
(joint with Yu Cao and Lihan Wang)

Generative models for high dimensional probability distributions
(joint with Yulong Lu)

Convergence to equilibrium of Langevin dynamics
(joint with Yu Cao and Lihan Wang)

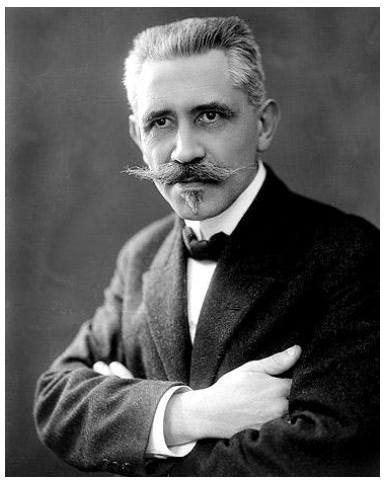
Generative models for high dimensional probability distributions
(joint with Yulong Lu)

Langevin dynamics

Consider the (underdamped) Langevin dynamics:

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \\ m\,\mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma k_B T}\,\mathrm{d}W_t \end{aligned}$$

- x: position; v: velocity;
- U: potential energy;
- γ: friction coefficient;
- m: particle mass;
- T: temperature;
- k_B Boltzmann constant;
- W_t: d-dimensional standard Brownian motion.



Paul Langevin (1872–1946)

For simplicity, we will set $m = 1$ and $k_B T = 1$ (after possibly a rescaling).

Langevin dynamics (with only friction parameter γ)

$$\mathrm{d}x_t = v_t \,\mathrm{d}t$$
$$\mathrm{d}v_t = -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t$$

Langevin dynamics (with only friction parameter γ)

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t \end{aligned}$$

As $\gamma \to 0$, we get the Hamiltonian dynamics for $H(v,x) = \frac{1}{2}|v|^2 + U(x)$:

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t = \frac{\partial H(x_t, v_t)}{\partial v}\,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t = -\frac{\partial H(x_t, v_t)}{\partial x}\,\mathrm{d}t \end{aligned}$$

Langevin dynamics (with only friction parameter γ)

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t \end{aligned}$$

As $\gamma \to 0$, we get the Hamiltonian dynamics for $H(v,x) = \frac{1}{2}|v|^2 + U(x)$:

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t = \frac{\partial H(x_t, v_t)}{\partial v}\,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t = -\frac{\partial H(x_t, v_t)}{\partial x}\,\mathrm{d}t \end{aligned}$$

As $\gamma \to \infty$, on the $O(1)$ time scale, it is dominated by the Ornstein-Uhlenbeck process in the velocity variable with "equilibrium"

$$"v_t \,\mathrm{d}t = -\gamma^{-1}\nabla U(x_t)\,\mathrm{d}t + \sqrt{2\gamma^{-1}}\,\mathrm{d}W_t,"$$

after rescaling $\gamma t \mapsto t$, we obtain the overdamped Langevin dynamics

$$\mathrm{d}x_t = -\nabla U(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

Langevin dynamics

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t \end{aligned}$$

The corresponding backward Kolmogorov equation (known as the kinetic Fokker-Planck equation) is given by

$$\begin{aligned} \partial_t f &= \mathcal{L} f \\ f(0, x, v) &= f_0(x, v) \end{aligned}$$

with the generator given by $\mathcal{L} = \mathcal{L}_{\mathsf{ham}} + \gamma \mathcal{L}_{\mathsf{FD}}$ with

$$\mathcal{L}_{\mathsf{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\mathsf{FD}} = \Delta_v - v \cdot \nabla_v$$

Langevin dynamics

$$\begin{aligned} \mathrm{d}x_t &= v_t \,\mathrm{d}t \\ \mathrm{d}v_t &= -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t \,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t \end{aligned}$$

The corresponding backward Kolmogorov equation (known as the kinetic Fokker-Planck equation) is given by

$$\begin{aligned} \partial_t f &= \mathcal{L} f \\ f(0, x, v) &= f_0(x, v) \end{aligned}$$

with the generator given by $\mathcal{L} = \mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}$ with

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v$$

The invariant measure of the Langevin dynamics is given by

$$\rho_\infty(\,\mathrm{d}x,\,\mathrm{d}v) = \frac{1}{Z} e^{-H(x,v)}\,\mathrm{d}x\,\mathrm{d}v = \frac{1}{Z} e^{-U(x) - \frac{1}{2}|v|^2}\,\mathrm{d}x\,\mathrm{d}v,$$

where Z is the normalizing constant (recall that $\beta = k_B T = 1$).

The invariant measure of Langevin dynamics is

$$\rho_\infty(\,\mathrm{d}x,\,\mathrm{d}v) \propto e^{-U(x)-\frac{1}{2}|v|^2}\,\mathrm{d}x\,\mathrm{d}v,$$

its marginal on x is the Boltzmann-Gibbs distribution $\mu_U \propto e^{-U(x)}$.

The invariant measure of Langevin dynamics is

$$\rho_\infty(\mathrm{d}x, \mathrm{d}v) \propto e^{-U(x)-\frac{1}{2}|v|^2}\,\mathrm{d}x\,\mathrm{d}v,$$

its marginal on x is the Boltzmann-Gibbs distribution $\mu_U \propto e^{-U(x)}$.

Thus, the Langevin dynamics and its overdamped limit can be used to sample the Boltzmann-Gibbs distribution.

- Smart Monte Carlo [Rossky, Doll, Friedman 1978] (which in fact discussed both underdamped and overdamped Langevin);
- The overdamped version is known in the statistics literature as the Metropolis-adjusted Langevin algorithm (MALA) [Besag 1994; Roberts, Tweedie 1996];
- The underdamped version becomes popular in machine learning in recent years, see e.g., [Cheng, Chatterji, Bartlett, Jordan 2018];
- The underdamped version is also closely related to the Hamiltonian Monte Carlo method [Duane, Kennedy, Pendleton, Roweth 1987];
- Often combined with Metropolis algorithm, though can be used without accept/reject, see e.g., [Neal 1993].

Question: How fast Langevin dynamics converges to equilibrium?

Recall the kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v$$

Since $\mathcal{L}^* \rho_\infty = 0$, $\int f(t, \cdot)\, \mathrm{d}\rho_\infty$ is invariant in time. As $t \to \infty$, f will converge to the constant $\int f(0, \cdot)\, \mathrm{d}\rho_\infty$. How fast?

Question: How fast Langevin dynamics converges to equilibrium?

Recall the kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v$$

Since $\mathcal{L}^* \rho_\infty = 0$, $\int f(t, \cdot)\, \mathrm{d}\rho_\infty$ is invariant in time. As $t \to \infty$, f will converge to the constant $\int f(0, \cdot)\, \mathrm{d}\rho_\infty$. How fast?

c.f. the Fokker-Planck equation for overdamped dynamics

$$\partial_t h = -\nabla_x U \cdot \nabla_x h + \Delta_x h, \qquad h(0, x) = h_0(x).$$

The convergence of Fokker-Planck is well understood, as the generator is self-adjiont and coercive with respect to $L^2_{\mu_U}$.

Assumption (Poincaré inequality for μ_U)

$$\int \left(h - \int h \, \mathrm{d}\mu_U\right)^2 \mathrm{d}\mu_U \leq \frac{1}{m} \int |\nabla_x h|^2 \, \mathrm{d}\mu_U$$

This implies that the overdamped dynamics has convergence rate m.
e.g., when U is m-strongly convex.

Question: Any improvement by the underdamped Langevin dynamics?

Assumption (Poincaré inequality for μ_U)

$$\int \left(h - \int h \, \mathrm{d}\mu_U\right)^2 \mathrm{d}\mu_U \leq \frac{1}{m} \int |\nabla_x h|^2 \, \mathrm{d}\mu_U$$

This implies that the overdamped dynamics has convergence rate m.
e.g., when U is m-strongly convex.

Question: Any improvement by the underdamped Langevin dynamics?

gradient descent $\longrightarrow$ accelerated gradient descent

overdamped dynamics $\longrightarrow$ Langevin dynamics

Langevin dynamics can be thought of as a stochastic version of accelerated gradient dynamics.

Thus, if we take the analogy of optimization vs sampling, we would hope that (underdamped) Langevin dynamics gives convergence rate $\mathcal{O}(\sqrt{m})$ for strongly convex U.

Assumption (Poincaré inequality for μ_U)

$$\int \left(h - \int h \, \mathrm{d}\mu_U\right)^2 \mathrm{d}\mu_U \le \frac{1}{m} \int |\nabla_x h|^2 \, \mathrm{d}\mu_U$$

This implies that the overdamped dynamics has convergence rate m.

Question: Any improvement by the underdamped Langevin dynamics?

Theorem (Cao-L.-Wang 2019)

For convex U satisfying $|\mathrm{Hess}\, U| \lesssim 1 + |\nabla U|$ and superlinear as $|x| \to \infty$,

$$\|f(t,\cdot) - \int f(t,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)} \le C_0 \exp(-\lambda t) \|f(0,\cdot) - \int f(0,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)}$$

with explicit estimate of λ as

$$\lambda = \sqrt{m} \log\Big(1 + \frac{\gamma\sqrt{m}}{c_0(\sqrt{m}+\gamma)^2}\Big)$$
$$= \mathcal{O}(\sqrt{m}) \qquad \text{if we take } \gamma = \mathcal{O}(\sqrt{m}).$$

Results available for more general case; we will not discuss those here.

Exponential convergence with rate $\mathcal{O}(\sqrt{m})$ (setting $\int f \, \mathrm{d}\rho_\infty = 0$)

$$\|f(t,\cdot)\|_{L^2(\rho_\infty)} \leq C_0 \exp(-c\sqrt{m}t)\|f(0,\cdot)\|_{L^2(\rho_\infty)}$$

- The $\mathcal{O}(\sqrt{m})$ convergence rate is optimal, as can be seen when U is a Gaussian (so explicit calculation can be done);
- First result in literature for sharp $\sqrt{m}$ convergence rate (acceleration compared with overdamped dynamics with rate m);
- Convergence in L^2 implies convergence of density in χ^2-divergence, and thus in relatively entropy and total variation distance with $\mathcal{O}(\sqrt{m})$ rate;
- The constant $C_0 > 1$ is unavoidable, since the operator $\mathcal{L}$ is not coercive (more on next slide).

Kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v.$$

The operator is hypo-elliptic [Hörmander 1967] (as the diffusion is degenerate in the x direction).

In particular, we cannot hope for the exponential convergence to follow from a Poincaré (coercivity) estimate for $\mathcal{L}$. Hypo-coercivity needs to be established.

Kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v.$$

The analysis of kinetic Fokker-Planck eqns dates back to [Kolmogorov 1934]. Related previous results on convergence:

- Convergence in $H^1_{\rho_\infty}$ norm [Villani 2009];
- Convergence in a modified $L^2_{\rho_\infty}$ norm [Dolbeault, Mouhot, Schmeiser 2009; 2015] (also earlier idea from [Herau 2006]).
  This was applied to kinetic Fokker-Planck equation by [Roussel, Stoltz 2018], which gives explicit constants, though not sharp.
- Convergence in Wasserstein distance:
  Bakry-Émery framework [Boudoin 2016];
  Coupling approaches [Eberle, Guillin, Zimmer 2019; Cheng, Chatterji, Bartlett, Jordan 2018; Dalalyan, Riou-Durand 2018].

Kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\mathsf{ham}} + \gamma \mathcal{L}_{\mathsf{FD}}\big) f; \qquad f(0,x,v) = f_0(x,v),$$

where

$$\mathcal{L}_{\mathsf{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\mathsf{FD}} = \Delta_v - v \cdot \nabla_v.$$

Our analysis method was inspired by a recent variational framework [Armstrong, Mourrat 2019] for kinetic FP equation on compact domain without potential, which implicitly used the bracket condition dating back to [Hörmander 1967].

As $\mathcal{L}$ is not coercive, the idea is to resort to augmenting the state space by a time interval $I = (0,T)$ equipped with Lebesgue measure λ. Since in time, the diffusion in v direction will propagate to the x direction.

Kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\text{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\text{FD}} = \Delta_v - v \cdot \nabla_v.$$

Let κ be the Gaussian measure in velocity ($\rho_\infty(\,\mathrm{d}x\,\mathrm{d}v) = \mu_U(\,\mathrm{d}x)\kappa(\,\mathrm{d}v)$). The exp. conv. follows from separating into intervals with $T = 1/\sqrt{m}$.

Theorem (Poincaré inequality in time augmented state space)

$$\|f - (f)_{\lambda\times\mu_U}\|_{L^2(\lambda\times\mu_U;L^2_\kappa)} \lesssim \big(1 + \frac{1}{T\sqrt{m}}\big)\|\nabla_v f\|_{L^2(\lambda\times\mu_U;L^2_\kappa)} + \big(\frac{1}{\sqrt{m}} + T\big)\|\partial_t f - \mathcal{L}_{ham} f\|_{L^2(\lambda\times\mu_U;H^{-1}_\kappa)},$$

where $(f)_{\lambda\times\mu_U} := \frac{1}{T}\int f(t, x, v)\,\mathrm{d}t\,\mathrm{d}\rho_\infty$.

Convergence to equilibrium of Langevin dynamics
(joint with Yu Cao and Lihan Wang)

Generative models for high dimensional probability distributions
(joint with Yulong Lu)

Generative models: Transform simple probability distributions (such as Gaussian) often by neural networks to the desired probability distribution:

$$\pi \approx g_{\#} p_z.$$

Question? Given source and target distributions, can one construct a DNN to approximate to push-forward?

This is a distribution version of the question of universal approximation.

cf. [Lee, Ge, Ma, Risteski, Arora 2017]

Generative models: Transform simple probability distributions (such as Gaussian) often by neural networks to the desired probability distribution:

$$\pi \approx g_{\#} p_z.$$

Question? Given source and target distributions, can one construct a DNN to approximate to push-forward?

This is a distribution version of the question of universal approximation.

cf. [Lee, Ge, Ma, Risteski, Arora 2017]

To make the question more quantitative, we need to specify the metric we use to measure the difference of two distributions.
Integral probability metric (IPM):

$$d_{\mathcal{F}_D}(p, \pi) := \sup_{f \in \mathcal{F}_D} \left| \mathbb{E}_p f(X) - \mathbb{E}_\pi f(Y) \right|,$$

where f can be understood as a discriminator.

Integral probability metric (IPM):

$$d_{\mathcal{F}_D}(p,\pi) := \sup_{f\in\mathcal{F}_D} \left|\mathbb{E}_p f(X) - \mathbb{E}_\pi f(Y)\right|.$$

1-Wasserstein distance: $\mathcal{F}_D$: 1-Lipschitz functions

$$\mathcal{W}_1(p,\pi) = \inf_{\gamma\in\Gamma(p,\pi)} \int |x-y|\gamma(\mathrm{d}x\,\mathrm{d}y).$$

Maximum mean discrepancy (MMD): $\mathcal{F}_D$: unit ball of a reproducing kernel Hilbert space (RKHS):

$$\mathrm{MMD}(p,\pi) = \sup_{\|f\|_{\mathcal{H}_k}\leq 1} \left|\mathbb{E}_p f(X) - \mathbb{E}_\pi f(Y)\right|$$

Kernelized Stein discrepancy (KSD): $\mathcal{F}_D = \left\{\mathcal{T}_\pi f \mid f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \leq 1\right\}$

$$\mathrm{KSD}(p,\pi) = \mathbb{E}_p\left(\mathcal{T}_\pi f(X)\right)^2$$

where $\mathcal{T}_\pi$ is the Stein operator (for $f : \mathbb{R}^d \to \mathbb{R}^d$)

$$\mathcal{T}_\pi f := \nabla \log \pi \cdot f + \nabla \cdot f.$$

Theorem (L.-Lu 2020)

For source distribution p_z absolutely continuous wrt Lebesgue measure, there exists a fully connected and feed-forward deep neural network u with depth $L = \lceil \log_2 n \rceil$ and width $N = 2^L$ such that

$$d_{\mathcal{F}_D}\big((\nabla u)_{\#} p_z, \pi\big) \leq \varepsilon,$$

where the complexity n depends on the choice of metric.

- *If $d_{\mathcal{F}_D} = \mathcal{W}_1$ and π has finite 3rd moment,*

$$n \lesssim \begin{cases} \varepsilon^{-2}, & d = 1; \\ \log^2(\varepsilon)\varepsilon^{-2}, & d = 2; \\ \varepsilon^{-d}, & d \geq 3. \end{cases}$$

Theorem (L.-Lu 2020)

For source distribution p_z absolutely continuous wrt Lebesgue measure, there exists a fully connected and feed-forward deep neural network u with depth $L = \lceil \log_2 n \rceil$ and width $N = 2^L$ such that

$$d_{\mathcal{F}_D}\big((\nabla u)_{\#} p_z, \pi\big) \leq \varepsilon,$$

where the complexity n depends on the choice of metric.

- *If $d_{\mathcal{F}_D} = \mathrm{MMD}$ with kernel k such that $\sup_x |k(x,x)| \lesssim 1$,*

$$n \lesssim \varepsilon^{-2}.$$

- *If $d_{\mathcal{F}_D} = \mathrm{KSD}$ with kernel k plus some technical assumptions, π is sub-Gaussian and $\nabla \log \pi(x)$ is Lipschitz,*

$$n \lesssim d\varepsilon^{-1}.$$

Remark: It is better to parameterize u using NN and use ∇u to push-forward than directly parameterize the map due to regularity.

Proof sketch: The overall strategy is similar to standard approximation theory: Take a sieve for the whole space and construct approximation for elements in the sieve.

- Approximate π by empirical measures $X_i \sim \pi$:

$$\pi \approx \pi_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

  Wasserstein: [Fourier, Guillin 2015; Weed, Bach 2019; Lei 2020]; MMD: [Sriperumbudur 2016]. KSD: Our work.

- Push-forward the source dist. p_z to the empirical distribution P_n;

  this is based on the theory of semi-discrete optimal transport.

Monge formulation of optimal transport with quadratic loss

$$\inf_{T\colon \mathbb{R}^d \to \mathbb{R}^d} \int \frac{1}{2}|x - T(x)|^2 \mu(\mathrm{d}x) \qquad \text{s.t. } T_{\#}\mu = \nu.$$

Kantorovich formulation of optimal transport

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \frac{1}{2}|x - y|^2 \gamma(\mathrm{d}x\, \mathrm{d}y).$$

Monge formulation of optimal transport with quadratic loss

$$\inf_{T\colon \mathbb{R}^d \to \mathbb{R}^d} \int \frac{1}{2}|x - T(x)|^2 \mu(\mathrm{d}x) \qquad \text{s.t. } T_{\#}\mu = \nu.$$

Kantorovich formulation of optimal transport

$$\mathcal{W}_2^2(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \frac{1}{2}|x-y|^2 \gamma(\mathrm{d}x\,\mathrm{d}y).$$

The Kantorovich dual formulation

$$\begin{aligned}\mathcal{W}_2^2(\mu,\nu) &= \sup_{\substack{\varphi \in L^1(\mu),\, \psi \in L^1(\nu) \\ \varphi(x)+\psi(y) \le \frac{1}{2}|x-y|^2}} \int \varphi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu \\ &= \sup_{\psi \in L^1(\nu)} \int \psi^c \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu,\end{aligned}$$

where ψ^c is given by the c-transform:

$$\psi^c(x) = \inf_{y \in \mathbb{R}^d} \frac{1}{2}|x-y|^2 - \psi(y).$$

Monge formulation of optimal transport with quadratic loss

$$\inf_{T\colon \mathbb{R}^d \to \mathbb{R}^d} \int \frac{1}{2}|x - T(x)|^2 \mu(\mathrm{d}x) \qquad \text{s.t. } T_{\#}\mu = \nu.$$

The Kantorovich dual formulation

$$\mathcal{W}_2^2(\mu, \nu) = \sup_{\psi \in L^1(\nu)} \int \psi^c \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu,$$

where ψ^c is given by the c-transform:

$$\psi^c(x) = \inf_{y \in \mathbb{R}^d} \frac{1}{2}|x - y|^2 - \psi(y).$$

Theorem (Brenier 1987)

Assume that μ is absolutely continuous with respect to the Lebesgue measure. Then there exists a convex function $\bar{\varphi} : \mathbb{R}^d \to \mathbb{R}$

$$T(x) = \nabla \bar{\varphi}(x), \qquad \mu\text{-a.e. } x.$$

In fact, $\bar{\varphi}(x)$ can be chosen as $\frac{1}{2}|x|^2 - \varphi(x)$.

Semi-discrete optimal transport:

- Source measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous: $\mathrm{d}\mu = \rho\,\mathrm{d}x$;
- Target measure $\nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$.

In the discrete case, the Kantorovich dual becomes the following functional depending on $(\psi_1, \ldots, \psi_n)$:

$$\mathcal{F}(\psi_1, \cdots, \psi_n) = \int \psi^c(x)\rho(x)\,\mathrm{d}x + \sum_{j=1}^{n} \psi_j \nu_j$$

$$= \int \inf_j \left(\frac{1}{2}|x - y_j|^2 - \psi_j\right)\rho(x)\,\mathrm{d}x + \sum_{j=1}^{n} \psi_j \nu_j.$$

Semi-discrete optimal transport:

- Source measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ absolutely continuous: $\mathrm{d}\mu = \rho\,\mathrm{d}x$;
- Target measure $\nu = \sum_{j=1}^n \nu_j \delta_{y_j}$.

In the discrete case, the Kantorovich dual becomes the following functional depending on $(\psi_1, \ldots, \psi_n)$:

$$\mathcal{F}(\psi_1, \cdots, \psi_n) = \int \psi^c(x)\rho(x)\,\mathrm{d}x + \sum_{j=1}^n \psi_j \nu_j$$

$$= \int \inf_j \big(\frac{1}{2}|x - y_j|^2 - \psi_j\big)\rho(x)\,\mathrm{d}x + \sum_{j=1}^n \psi_j \nu_j.$$

Theorem (Gu-Luo-Sun-Yau 2016 for compact domain; L.-Lu 2020)

The optimal transport plan T is given by for some $m_j \in \mathbb{R}$.

$$T(x) = \nabla\bar{\varphi}(x) := \nabla \max_j\{x \cdot y_j + m_j\}.$$

In particular, $\bar{\varphi}$ is piecewise linear and can be easily approx. by ReLU NN.

- Underdamped Langevin dynamics has faster convergence to equilibrium compared with overdamped dynamics for convex pot.
- Universal approximation without curse-of-dimensionality for generative models for distributions;

Thank you! Any questions?

Email: jianfeng@math.duke.edu

URL: http://www.math.duke.edu/~jianfeng/

References:

- with Yu Cao and Lihan Wang, *On explicit L^2-convergence rate estimate for underdamped Langevin dynamics*, arXiv:1908.04746
- with Yulong Lu, *A universal approximation theorem of deep neural networks for expressing distributions*, arXiv:2004.08867