An Introduction to Optimization over the Space of Probability Measures: From Sampling to Wasserstein Gradient Flow

Handstein Wang
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

January 27, 2026

Outline

Introduction

- **Optimization over the space of probability measures** $\mathcal{P}(\mathbb{R}^d)$:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{F}(\mu).$$

- **Applications:**
  - **Sampling with target distribution** π [Wibisono, 2018]:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{F}(\mu) := \mathrm{KL}(\mu \| \pi).$$

  - **Variational inference (VI)** [Jordan et al., 1999]:

$$\min_{\mu \in \mathcal{A} \subset \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu \| \pi).$$

  - **Other examples:** distributed robust optimization [Xu and Zhu, 2025], deep learning [Chizat, 2022]; single-cell analysis in mathematical biology [Lavenant et al., 2023], etc.

Outline

Sampling

Goal: Sample from a distribution π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.
Two main settings of sampling:

- **Setting 1: π is given in explicit form up to a normalization constant:**

$$\pi(x) = \frac{e^{-f(x)}}{\int_{\mathbb{R}^d} e^{-f(x)} dx}, \quad x \in \mathbb{R}^d,$$

  Applications: Bayesian inference, numerical integration in high dimensions, differential privacy, simulation in physics, finance, and machine learning.

- **Setting 2: π is given with a collection if i.i.d. samples:**

$$\text{only have } \{x_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \pi$$

  Applications: generative models (diffusion models, GANs, etc).

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC): generate a Markov chain $\{X_n\}$ such that

$$\mu_n = \mathrm{Law}(X_n) \xrightarrow{\text{in some sense}} \pi$$

- Algorithms: **Unadjusted Langevin Algorithm (ULA)** (also called Langevin Monte Carlo (LMC)), **Metropolis-Adjusted Langevin Algorithm (MALA)**, etc.

Unadjusted Langevin Algorithm

The Unadjusted Langevin Algorithm for sampling from target distribution $\pi \propto e^{-f(x)}$ is given by

$$X_{k+1,h} = X_{k,h} - h\nabla f(X_{k,h}) + \sqrt{2h}\,\xi_{k+1}, \tag{ULA}$$

where $h > 0$ is the step size, $\xi_1, \cdots, \xi_k, \cdots$ are i.i.d. $N(0, I_d)$ random variables.

probabilistic perspective $\Longrightarrow$ geometric perspective

Unadjusted Langevin Algorithm

Unadjusted Langevin Algorithm: $X_{k+1,h} = X_{k,h} - h\nabla f(X_{k,h}) + \sqrt{2h}\,\xi_{k+1}$

- **Probabilistic perspective:** ULA is the Euler-Maruyama discretization of the Langevin dynamics
$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dB_t, \tag{LD}$$
By Fokker-Planck equation
$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot \left(\mu_t \nabla \log \frac{\mu_t}{\pi}\right), \quad \mu_t \text{ denotes the density of } X_t \text{ for all } t \geq 0,$$
the target distribution π is the unique invariant measure of LD.

- **Geometric perspective:** $(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathrm{W}_2)$ can be viewed as a Riemannian manifold and the Langevin dynamics can be viewed as the Wasserstein gradient flow of KL divergence functional $\mathrm{KL}(\cdot\|\pi)$ [Otto, 2001].

Outline

Wasserstein Space

- The **Wasserstein space** $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ is a metric space, where

$$\mathcal{P}_2(\mathbb{R}^d) := \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int_{\mathbb{R}^d} \|x\|_2^2 \, \mu(dx) < +\infty\},$$

$$\mathrm{W}_2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/2}, \quad \text{for all } \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d),$$

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) | \gamma(A \times \mathbb{R}^d) = \mu(A), \ \gamma(\mathbb{R}^d \times A) = \nu(A), \ \text{ for all } A \in \mathcal{B}(\mathbb{R}^d)\}$$

- **(Isometric embedding property)** The mapping $x \mapsto \delta_x$ is an isometric embedding from $(\mathbb{R}^d, \|\cdot\|)$ to $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$:

$$\mathrm{W}_2(\delta_x, \delta_y) = \|x - y\|, \ \text{for all } x, y \in \mathbb{R}^d.$$

- We will work on

$$\mathcal{P}_{2,ac}(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \mu \ll m \text{ and } \int_{\mathbb{R}^d} \|x\|_2^2 \, \mu(dx) < +\infty\}$$

Tangent Space of $\mathcal{P}_{2,ac}(\mathbb{R}^d)$

For evolution $t \mapsto \mu_t$ in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$, there are two main perspectives:

Lagrangian viewpoint

Let X_t be random variable representing the particle trajectory, evolving according to

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{dt}} X_t = v_t(X_t), \quad a.s. \\ X_0 \sim \mu_0, \end{cases}$$

and let $\mu_t = \mathrm{Law}(X_t)$.

Eulerian viewpoint

Let μ_t be probability density representing the mass density and $v_t : \mathbb{R}^d \to \mathbb{R}^d$ be the velocity field and they satisfies the **continuity equation**

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0.$$

Under some regularity conditions, the two view points above are equivalent.

Tangent Space of $\mathcal{P}_{2,ac}(\mathbb{R}^d)$: Continuity Equation

From continuity equation

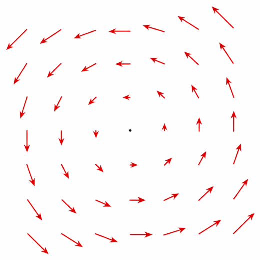
$$\begin{cases} \frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0 \\ \int_{\mathbb{R}^d} \mu_t(x)\,\mathrm{d}x = 1, \quad \forall t \geq 0 \end{cases}$$

- if $(v_t)_{t\geq 0}$ is given, then under some regularity conditions, $(\mu_t)_{t\geq 0}$ is unique;
- but if the curve $t \mapsto \mu_t$ is given, the velocity field $(v_t)_{t\geq 0}$ is often **not unique**.

- example: let $\mu_t \equiv N(0, I_d)$, and $v_t^{(1)} \equiv 0$ and $v_t^{(2)}$ be the rotation vector field, then they both satisfy the continuity equation.
- if vector field $(w_t)_{t\geq 0}$ satisfies

$$\nabla \cdot (\mu_t w_t) \equiv 0,$$

then the vector field $(v_t + w_t)_{t\geq 0}$ also satisfies the continuity equation.



Rotation vector field

Tangent Space of $\mathcal{P}_{2,ac}(\mathbb{R}^d)$

Idea: minimize the kinetic energy

$$\min_{v_t:\mathbb{R}^d\to\mathbb{R}^d} \quad \frac{1}{2}\|v_t\|_{\mu_t}^2 := \frac{1}{2}\int_{\mathbb{R}^d} \|v_t\|^2 \,\mathrm{d}\,\mu_t$$
$$\text{s. t.} \quad \frac{\partial \mu_t}{\partial t} + \nabla\cdot(\mu_t v_t) = 0$$

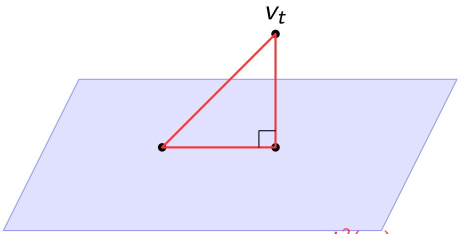
Weak form of continuity equation: for all $\varphi \in C_c^\infty(\mathbb{R}^d, \mathbb{R})$,

$$\frac{\partial}{\partial t}\int_{\mathbb{R}^d} \varphi(x)\mu_t(\mathrm{d}x) = \int_{\mathbb{R}^d} \langle \nabla\varphi(x), v_t(x)\rangle\, \mu_t(\mathrm{d}x).$$

Theorem 1 ([Chewi et al., 2024])

For "nice enough" curve $t \mapsto \mu_t$, there is a unique vector field $t \mapsto v_t$ minimizing the kinetic energy such that the continuity equation holds. Furthermore, v_t is the minimizer if and only if it belongs to the set

$$\overline{\{\nabla\varphi \mid \varphi \in C_c^\infty(\mathbb{R}^d,\mathbb{R})\}}^{L^2(\mu_t)}.$$



$$\overline{\{\nabla\varphi \mid \varphi \in C_c^\infty(\mathbb{R}^d,\mathbb{R})\}}^{L^2(\mu_t)}$$

A Riemannian Manifold View of $(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathrm{W}_2)$

- **Tangent space**: $\mathrm{T}_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) := \overline{\{\nabla\varphi \mid \varphi \in C_c^\infty(\mathbb{R}^d, \mathbb{R})\}}^{L^2(\mu)}$;
- We say vector field (v_t) **is tangent to** (μ_t) if and only if $v_t \in \mathrm{T}_{\mu_t} \mathcal{P}_{2,ac}(\mathbb{R}^d)$ for all $t \geq 0$ and the continuity holds;
- **Riemannian distance**:

$$d(\mu_0, \mu_1) := \inf \left\{ \int_0^1 \|v_t\|_{\mu_t} \,\mathrm{d}t \;\middle|\; (v_t)_{t\in[0,1]} \text{ is tangent to } (\mu_t)_{t\in[0,1]} \right\} = \mathrm{W}_2(\mu_0, \mu_1);$$

- **Geodesics:** let $X_0 \sim \mu_0$, $X_1 \sim \mu_1$ be optimally coupled, and

$$X_t = (1-t)X_0 + tX_1, \;\; t \in [0,1],$$

then $\mu_t = \mathrm{Law}(X_t), \;\; t \in [0,1]$ gives the geodesic from μ_0 to μ_1.

Convexity of KL Divergence Functional

Definition 2 (Strongly convex of a functional)

We say a functional $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ is $m-$strongly convex along Wasserstein geodesics if for geodesic $(\mu_t)_{t\in[0,1]}$ and all $t \in [0,1]$,

$$\mathcal{F}(\mu_t) \leq (1-t)\,\mathcal{F}(\mu_0) + t\,\mathcal{F}(\mu_1) - \frac{mt(1-t)}{2} \mathrm{W}_2(\mu_0, \mu_1)^2$$

Theorem 3 (Convexity of the KL divergence functional)

If $\pi \propto e^{-f(x)}$, where f is $m-$strongly convex, then $\mathrm{KL}(\cdot\|\pi)$ is also $m-$strongly convex along Wasserstein geodesics.

Wasserstein Gradient

Definition 4 (Wasserstein gradient)

Let $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be a functional. The Wasserstein gradient of $\mathcal{F}$ at μ, denoted by $\nabla_{\mathrm{W}_2} \mathcal{F}(\mu)$, is the unique element of $\mathrm{T}_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$ such that for all curves $t \mapsto \mu_t$ with $\mu_0 = \mu$ with tangent vector v_0,

$$\frac{\mathrm{d}}{\mathrm{dt}}\Big|_{t=0} \mathcal{F}(\mu_t) = \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu), v_0 \rangle_\mu$$

Definition 5 (First variational derivative of a functional)

Let $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be a functional. The first variational derivative of $\mathcal{F}$ at μ is a function, denoted by $\frac{\delta \mathcal{F}}{\delta \mu}[\mu]$, such that for any perturbation in measure ν such that $\mu + \varepsilon \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$,

$$\frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}\Big|_{\varepsilon=0} \mathcal{F}(\mu + \varepsilon \nu) = \int_{\mathbb{R}^d} \frac{\delta \mathcal{F}}{\delta \mu}[\mu](x) \,\mathrm{d}\, \nu(x).$$

Wasserstein Gradient

Theorem 6

Let $\mathcal{F} : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ be a functional. Then the Wasserstein gradient of $\mathcal{F}$ at μ is just the gradient of first variational derivative of $\mathcal{F}$ at μ, namely

$$\nabla_{\mathrm{W}_2} \mathcal{F}(\mu) = \nabla \frac{\delta \mathcal{F}}{\delta \mu}[\mu].$$

Example 7

Let $\mathcal{F}(\mu) := \mathrm{KL}(\mu \| \pi)$, then

$$\nabla_{\mathrm{W}_2} \mathcal{F}(\mu) = \nabla \frac{\delta \mathcal{F}}{\delta \mu}[\mu] = \nabla \log \frac{\mu}{\pi}.$$

Wasserstein Gradient Flow

Definition 8 (Wasserstein gradient flow)

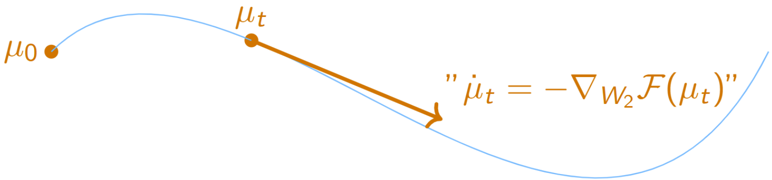
The Wasserstein gradient flow of $\mathcal{F}$ is the curve $t \mapsto \mu_t$ with tangent vector $-\nabla_{W_2} \mathcal{F}(\mu_t)$ at time t, which means

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla_{W_2} \mathcal{F}(\mu_t))$$

Theorem 9

The functional value decreases along the Wasserstein gradient flow trajectory

$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{F}(\mu_t) = -\|\nabla_{W_2} \mathcal{F}(\mu_t)\|^2_{\mu_t} \leq 0.$$



Geometric Perspective of ULA: Wasserstein Gradient Flow

Recall the Langevin dynamics

$$dX_t = -\nabla f(X_t)\,\mathrm{dt} + \sqrt{2}\,\mathrm{d}\,B_t$$

with our target distribution $\pi \propto e^{-f(x)}$. Now we consider the functional $\mathcal{F}$ as KL divergence

$$\mathcal{F}(\mu) = \mathrm{KL}(\mu\|\pi), \quad \nabla_{\mathrm{W}_2}\mathcal{F}(\mu) = \nabla\frac{\delta\,\mathcal{F}}{\delta\mu}[\mu] = \nabla\log\frac{\mu}{\pi}.$$

Wasserstein gradient flow

$$\frac{\partial\mu_t}{\partial t} = \nabla\cdot(\mu_t\nabla\log\frac{\mu_t}{\pi})$$

Fokker-Planck equation

$$\frac{\partial\mu_t}{\partial t} = \nabla\cdot(\mu_t\nabla\log\frac{\mu_t}{\pi})$$

Theorem 10 ([Jordan et al., 1998])

The Langevin dynamics can be viewed as the Wasserstein gradient flow of $\mathrm{KL}(\cdot\|\pi)$.

Wasserstein Gradient Flow Framework for Algorithm Design

Techniques from gradient flow

Wasserstein gradient flow
Functional $\mathcal{F}$

Continuity equation
$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla_{W_2} \mathcal{F}(\mu_t))$

Fokker-Planck equation
$\frac{\partial \mu_t}{\partial t} = -\nabla \cdot (a \mu_t) + \frac{1}{2} \nabla \cdot (b_t b_t^\top \nabla \mu_t)$

Stochastic differential equation (SDE)
$dX_t = a(t, X_t)\, dt + b_t \,\mathrm{d} B_t$

Numerical scheme for SDE
e.g. Euler-Maruyama Method

Outline

Metropolis-Hastings Adjusted Method

- Proposal kernel $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$, with $Q(x, dy) = q(x,y)\,\mathrm{dy}$;
- Accept each move with probability

$$\alpha(x,y) = \min\left[1, \frac{\pi(y)\ q(y,x)}{\pi(x)\ q(x,y)}\right],$$

  otherwise don't move;
- New kernel

$$P(x, \mathrm{dy}) = \alpha(x,y)Q(x,\mathrm{dy}) + \left(1 - \int_{\mathbb{R}^d} \alpha(x,z)Q(x,\mathrm{dz})\right)\delta_x(dy);$$

- Ensure the Markov chain is reversible with respect to probability measure π

$$\pi(\mathrm{dx})P(x,\mathrm{dy}) = \pi(\mathrm{dy})P(y,\mathrm{dx}), \quad \text{for all } x, y \in \mathbb{R}^d,$$

  which implies it has the desired invariant measure π.

Metropolis-Adjusted Langevin Algorithm (MALA)

Input: Initial point x_0 from a starting distribution μ_0, step size h, number of steps n.

Output: Sequence of samples $x_1, x_2, \ldots, x_n$

for $k = 0, 1, \ldots, n-1$ **do**

  Draw from the proposal distribution $y_k \sim \mathcal{N}(x_k - h\nabla f(x_k), 2h\mathbf{I}_d)$;

  Compute the acceptance rate

$$\alpha_k \leftarrow \min\left\{ \frac{\exp\left(-f(y_k) - \|x_k - y_k + h\nabla f(y_k)\|_2^2 / 4h\right)}{\exp\left(-f(x_k) - \|y_k - x_k + h\nabla f(x_k)\|_2^2 / 4h\right)}, 1 \right\};$$

  Draw $u \sim \mathrm{Unif}[0,1]$;

  if $u < \alpha_k$ **then**

    Accept the proposal: $x_{k+1} \leftarrow y_k$;

  else

    Reject the proposal: $x_{k+1} \leftarrow x_k$;

  end

end

Metropolis-Hastings Algorithms Are Fast

Definition 11 ($\varepsilon-$mixing time)

Suppose a MCMC algorithm with target distribution π starts from initial distribution μ_0 and the distribution at step k denoted by μ_k, then the $\varepsilon-$mixing time is defined by

$$t_{\mathrm{mix}}(\varepsilon, \mu_0) := \min\{k \in \mathbb{N} \mid \mathrm{TV}(\mu_k, \pi) \leq \varepsilon\}.$$

Algorithm	Strongly log-concave	Weakly log-concave
ULA [Dalalyan, 2017]	$\mathcal{O}\left(\frac{d\kappa^2 \log^2(\beta/\varepsilon)}{\varepsilon^2}\right)$	$\widetilde{\mathcal{O}}\left(\frac{d^3 L^2}{\varepsilon^4}\right)$
MALA [Dwivedi et al., 2019]	$\mathcal{O}\left(\max\left\{d\kappa,\, d^{0.5}\kappa^{1.5}\right\}\log\left(\frac{\beta}{\varepsilon}\right)\right)$	$\widetilde{\mathcal{O}}\left(\frac{d^2 L^{1.5}}{\varepsilon^{1.5}}\right)$

Table: Scalings of upper bounds on ε-mixing time for ULA and MALA in $\mathbb{R}^d$ with target $\pi \propto e^{-f}$.

Geometric Interpretation of the MH Adjustment

- Proposal kernel Q and $P(Q)$ be the Metropolis-Hastings kernel obtained from Q;
- $\mathcal{R}(\pi)$ be the space of kernels K which are reversible with respect to π and such that for each $x \in \mathbb{R}^d$, $K(x, dy) = k(x, y)\,\mathrm{dy}$, for $y \neq x$;
- Distance on the space of kernels $K(x, dy) = k(x, y)\,\mathrm{dy}$, for $y \neq x$,

$$d(K, K') := \int_{(\mathbb{R}^d \times \mathbb{R}^d)\setminus\Delta} \left|k(x, y) - k'(x, y)\right| \pi(\mathrm{dx})\, \mathrm{dy}$$

where $\Delta := \{(x, x) \mid x \in \mathbb{R}^d\}$ is the diagonal in $\mathbb{R}^d \times \mathbb{R}^d$;

Theorem 12 ([Billera and Diaconis, 2001])

The mapping $Q \mapsto P(Q)$ is a projection of the proposal kernel Q onto the space of reversible Markov chains with stationary distribution π with respect to distance d,

$$P(Q) \in \operatorname*{arg\,min}_{K \in \mathcal{R}(\pi)} d(Q, K).$$

MH Adjustment Based Sampling Algorithms

Optimization algorithms

e.g. Newton method

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

↓

Add some noise at each iteration

$$x_{k+1} = x_k - \alpha_k \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k) + \beta_k \xi_k,$$

where $\xi_k \sim N(0, I_d)$

↓

Do Metropolis-Hastings adjustment at each iteration

Goal: Sampling from target distribution

$$\pi(x) \propto e^{-f(x)}$$

Need to analysis:

- What is the optimal noise level β_k?
- What is the optimal step size α_k?
- Bound the mixing time.

Outline

Generative Modeling: Density Estimation

- **Setting 2:** We only have a collection of i.i.d. samples $x_1, x_2, \cdots, x_N$;
- **Density estimation:** Setting 2 $\Longrightarrow$ Setting 1;
- But density estimation can NOT avoid the curse of dimensionality.

Theorem 13 ([Tsybakov, 2009])

Given N i.i.d. samples $x_1, \ldots, x_N$ from a bounded pdf $f \in C^{m,\alpha}(\mathbb{R}^d)$ and let $\beta = m + \alpha$, then the minimax risk of an estimation $\hat{f}$ of f under the quadratic loss function $\ell(\hat{f}, f) := \|\hat{f} - f\|_2^2 = \int_{[0,1]^n} (f(x) - \hat{f}(x))^2 \, dx$ satisfies

$$\inf_{\hat{f}} \sup_{f \in P_\beta} \|\hat{f} - f\|_2^2 \gtrsim N^{-\frac{2\beta}{d+\beta}}.$$

The infimum is taken over all estimators $\hat{f}$ built on the data $x_1, \ldots, x_N$.

Likelihood-based Models

- Parameterization: $\pi(x) \approx p_\theta(x)$;
- MLE: learn θ by maximizing the log-likelihood of the data

$$\max_\theta \sum_{i=1}^{N} \log p_\theta(x_i).$$

- Examples: autoregressive models, normalizing flow models, energy-based models, variational auto-encoders (VAEs), etc.
- Disadvantage: The normalization constant becomes difficult to compute when the dimension is high.

MLE as Optimization in the Space of Probability Measures

Consider the space of probability measures $\{P_\theta \; ; \; \theta \in \Theta, P_\theta \ll m, \; p_\theta = \frac{\mathrm{d}\, P_\theta}{\mathrm{d}\, m}\}$, then

$$\mathrm{KL}(\pi \| P_\theta) = \int_{\mathbb{R}^d} \log \frac{\mathrm{d}\,\pi}{\mathrm{d}\,P_\theta} \,\mathrm{d}\,\pi = \underbrace{\int_{\mathbb{R}^d} \log \pi(x) \,\mathrm{d}\,\pi}_{\text{independent of } \theta} - \underbrace{\int_{\mathbb{R}^d} \log p_\theta(x) \,\mathrm{d}\,\pi}_{= \mathbb{E}_{X\sim\pi}[\log p_\theta(X)]} .$$

Hence,

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathrm{KL}(\pi \| P_\theta) \iff \theta^* \in \arg\max_{\theta\in\Theta} \mathbb{E}_{X\sim\pi}[\log p_\theta(X)].$$

By the Law of Large Numbers (LLN),

$$\mathbb{E}_{X\sim\pi}[\log p_\theta(X)] \approx \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x_i) \longleftarrow \textbf{log-likelihood}.$$

$$\boxed{\textbf{MLE} \iff \textbf{minimizing } \mathrm{KL}(\pi \| \cdot) + \textbf{LLN}}$$

Score-based Generative Models

- ULA with target $\pi(x) \propto e^{-f(x)}$ needs the information of $-\nabla f(x) = \nabla \log \pi(x)$, which is the definition of **score of** π;

- **Score matching:**

$$\min_{\theta \in \Theta} \ \mathbb{E}_{X \sim \pi} \left\| \nabla \log \pi(X) - s_\theta(X) \right\|_2^2 \iff \min_{Q \in \mathcal{M}} \mathcal{J}(\pi \| Q), \tag{SM}$$

where $\mathcal{J}(\pi \| Q)$ is the Fisher divergence of π w.r.t. Q defined by

$$\mathcal{J}(\pi \| Q) := \mathbb{E}_{X \sim \pi}[\|\nabla \log \pi(X) - \nabla \log q(X)\|^2]$$

and $\mathcal{M} = \{Q \in \mathcal{P}(\mathbb{R}^d) \ ; \ \text{the score of } Q \text{ belongs to } \{s_\theta \ ; \ \theta \in \Theta\}\}$.

Denoising Score Matching

- Denoising score matching problem

$$\min_{\theta \in \Theta} \; DSM(s_\theta) := \mathbb{E} \left\| \nabla \log \mu_Y(Y) - s_\theta(Y) \right\|_2^2, \tag{DSM}$$

  where $Y = aX + \sigma Z$, $a \in \mathbb{R}, \sigma > 0$,$X \sim \pi$, $Z \sim N(0, I_d)$, Z is independent of X and μ_Y is the density of Y w.r.t. Lebesgue measure.
- Equivalent optimization problem:

$$\arg\min_{Q \in \mathcal{M}} \mathcal{J}(\mu_Y \| Q) = \arg\min_{\theta} DSM(s_\theta) = \arg\min_{\theta} \mathbb{E} \left[\left\| s_\theta(Y) + \frac{1}{\sigma} Z \right\|_2^2 \right],$$

- By LLN, the objective can be replaced with

$$\min_{\theta} \; \frac{1}{N} \sum_{i=1}^{N} \left\| s_\theta(ax_i + \sigma z_i) + \frac{1}{\sigma} z_i \right\|_2^2,$$

  where $\{x_i\}_{i=1}^N$ are i.i.d. given samples drawing from π and $\{z_i\}_{i=1}^N \overset{i.i.d.}{\sim} N(0, I_d)$.

Score-based Diffusion Model: DDPM

Forward process:

$$dX_t = -X_t dt + \sqrt{2}\, dB_t, \quad X_0 \sim q_0 := \pi,\ X_t \sim q_t,\ t \in [0, T],$$

by Ito's lemma, whose solution can be written as

$$X_t = e^{-t} X_0 + \left(1 - e^{-2t}\right)^{1/2} Z_t, \quad Z_t \sim N(0, I_d).$$

Reverse process:

$$dX_t^{\leftarrow} = \left[X_t^{\leftarrow} + 2\nabla \log q_{T-t}(X_t^{\leftarrow})\right] dt + \sqrt{2}\, dB_t, \quad X_0^{\leftarrow} \sim q_T \approx \gamma^d := \text{law of } N(0, I_d),$$

which is designed by Fokker-Planck equation to ensure $\text{Law}(X_t^{\leftarrow}) = \text{Law}(X_{T-t})$.

DDPM: Score Matching

For all given t, the score matching problem for X_t

$$\min_{\theta_t \in \Theta_t} \mathbb{E} \left\| \nabla \log q_t(X_t) - s_t^{(\theta_t)}(X_t) \right\|_2^2 \iff \min_{\theta_t \in \Theta_t} \mathbb{E} \left[\left\| s_t^{(\theta_t)}(X_t) + \frac{1}{\sqrt{1 - e^{-2t}}} Z_t \right\|_2^2 \right],$$

By LLN,

$$\min_{\theta_t \in \Theta_t} \frac{1}{N} \sum_{i=1}^{N} \left\| s_t^{(\theta_t)}\left(x_t^{(i)}\right) + \frac{1}{\sqrt{1 - \exp(-2t)}} z_t^{(i)} \right\|^2, \qquad \text{(DDPM-SM)}$$

where $(z_t^{(i)})_{i \in [N]}$ are i.i.d. standard Gaussian samples independent of the data $(x_0^{(i)})_{i \in [N]}$.

DDPM: Error Analysis

Forward process:

$$dX_t = -X_t dt + \sqrt{2}\, dB_t, \quad X_0 \sim q_0 := \pi,\ X_t \sim q_t,\ t \in [0, T].$$

Reverse process:

$$dX_t^{\leftarrow} = \left[X_t^{\leftarrow} + 2\nabla \log q_{T-t}(X_t^{\leftarrow})\right] dt + \sqrt{2}\, dB_t, \quad X_0^{\leftarrow} \sim q_T \approx \gamma^d := \text{law of } N(0, I_d),$$

Partition : Partition the interval $[0, T]$ to $[kh, (k+1)h]$, $k = 0, 1, \cdots, K-1$ with $h > 0,\ K = T/h$. Integrate the reverse process from $[kh, t],\ t \in [kh, (k+1)h]$,

$$X_t^{\leftarrow} = X_{kh}^{\leftarrow} + \int_{kh}^{t} X_s^{\leftarrow} ds + \int_{kh}^{t} 2 \underbrace{\nabla \log q_{T-s}(X_s^{\leftarrow})}_{\approx s_{T-s}^{(\theta_{T-s})}(X_s^{\leftarrow}) \approx s_{T-kh}^{(\theta_{T-kh})}(X_{kh}^{\leftarrow})} ds + \sqrt{2}(B_t - B_{kh}).$$

Score matching and integral approximation:

$$dX_t^{\leftarrow} = \left\{ X_t^{\leftarrow} + 2\, s_{T-kh}^{(\theta_{T-kh})}(X_{kh}^{\leftarrow}) \right\} dt + \sqrt{2}\, dB_t, \qquad t \in [kh, (k+1)h],$$

which is a linear SDE and can be integrated in closed form.

Diffusion Model: Convergence Analysis

Let $p_t := \mathrm{Law}(X_t^{\leftarrow})$, DDPM has mainly three types of errors.

1. The error made at initialization of reverse process, γ^d used instead of q_T.
2. The score matching error, which depends on sample size N and neural network.
3. The discretization of the reverse process, which depends on the step-size h.

Assumption 1 (Lipschitz score)

For any $t \geq 0$, the score $\nabla \log q_t$ is L-Lipschitz.

Assumption 2 (Second moment bound)

Assume that $M_2^2 := \mathbb{E}_{X \sim q_0} \|X\|_2^2 < \infty$.

Assumption 3 (Score estimation error bound)

For $k = 1, \ldots, K$,

$$\mathbb{E}\left[\left\| s_{kh}^{(\theta_{kh})}(X_{kh}) - \nabla \log q_{kh}(X_{kh}) \right\|_2^2\right] \leq \varepsilon_{\text{score}}^2.$$

Diffusion Model: Convergence Analysis

Theorem 14 ([Chen et al., 2023])

Under the three previous assumptions. Let p_T be the output of the DDPM algorithm at time $T > 0$, with $h = T/K$ and K the number of steps, suppose $h \lesssim 1/L$, then

$$\mathrm{TV}(p_T, q_0) \lesssim \underbrace{\sqrt{\mathrm{KL}(q_0 \,\|\, \gamma^d)}\, e^{-T}}_{\text{convergence of forward process}} + \underbrace{(L\sqrt{d\,h} + LM_2 h)\sqrt{T}}_{\text{discretization error}} + \underbrace{\epsilon_{\text{score}}\sqrt{T}}_{\text{score estimation error}} .$$

DDPM can efficiently sample from multi-modal target measures as long as the score estimation is good.

Outline

Summary

- Many applications can be viewed as optimization problems over the space of probability measures;
- From the idea of sampling: **probability perspective** $\Longleftrightarrow$ **geometric perspective**;

  The merit of the right gradient flow formulation of a dissipative evolution equation is that it separates energetics and kinetics: The energetics endow the state space with a functional, the kinetics endow the state space with a (Riemannian) geometry via the metric tensor.

  [Otto, 2001].
- **Functional:** exclusive KL divergence $\mathrm{KL}(\pi\|\cdot)$, inclusive KL divergence $\mathrm{KL}(\mu\|\cdot)$, Fisher divergence $\mathcal{J}(\pi\|\cdot)$, etc.
- **Geometry:** Wasserstein geometry, maximum mean discrepancy (MMD) geometry, Fisher-Rao geometry, etc.

Future Works

- What role does the **Metropolis–Hastings adjustment** play in Wasserstein gradient flows, and how can this approach be applied to other applications?

- From the perspective of optimization over spaces of probability measures, what are many **learning-based methods in generative modeling** actually doing? How can we transfer these methods to other applications?

- Many optimization problems over spaces of probability measures come with **constraints** (e.g., variational inference, sampling for generative models with hard constraints, etc.). Can we systematically summarize and further develop the algorithms and theory for **constrained optimization problems over spaces of probability measures**?

Thanks!

References I

Billera, L. J. and Diaconis, P. (2001).
A geometric interpretation of the metropolis-hastings algorithm.
Statistical Science, 16(4):335–339.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2023).
Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions.

Chewi, S., Niles-Weed, J., and Rigollet, P. (2024).
Statistical optimal transport.
arXiv preprint arXiv:2407.18163, 3.

Chizat, L. (2022).
Mean-field langevin dynamics: Exponential convergence and annealing.

Dalalyan, A. S. (2017).
Theoretical guarantees for approximate sampling from smooth and log-concave densities.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(3):651–676.

References II

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2019).
Log-concave sampling: Metropolis-hastings algorithms are fast.
Journal of Machine Learning Research, 20(183):1–42.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999).
An introduction to variational methods for graphical models.
Machine learning, 37(2):183–233.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

Lavenant, H., Zhang, S., Kim, Y.-H., and Schiebinger, G. (2023).
Towards a mathematical theory of trajectory inference.

Otto, F. (2001).
The geometry of dissipative evolution equations: the porous medium equation.

References III

Tsybakov, A. (2009).

Introduction to nonparametric estimation. springer series in statistics. springer, new york.

Wibisono, A. (2018).

Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem.

In *Conference on learning theory*, pages 2093–3027. PMLR.

Xu, Z. and Zhu, J.-J. (2025).

Gradient flow sampler-based distributionally robust optimization.