Analogy between Sampling and Optimization

Handstein Wang

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

September 26, 2025

Outline

Outline

Introduction to sampling: what is sampling?

The goal of sampling: We want to sample from a distribution π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which admits a density w.r.t. Lebesgue measure m, also denoted by $\pi(x)$ having the form

$$\pi(x) = \frac{e^{-f(x)}}{\int_{\mathbb{R}^d} e^{-f(x)} dx}, \quad x \in \mathbb{R}^d, \tag{1}$$

where, throughout this talk, we assume $f \in C^2(\mathbb{R}^d)$ strongly convex with

$$0 \preceq m I_d \preceq \nabla^2 f(x) \preceq M I_d, \ \forall x \in \mathbb{R}^d \quad \text{and} \quad \int_{\mathbb{R}^d} e^{-f(x)} dx < +\infty.$$

Remark: Here the target distribution π is supported on $\mathbb{R}^d$, we refer to the problem as **(unconstrained) sampling** problem. If the target distribution is supported on a subset K of $\mathbb{R}^d$, namely the corresponding density having the form

$$\pi(x) = \frac{e^{-f(x)} 1_{\{x \in K\}}}{\int_K e^{-f(x)} dx}, \quad x \in \mathbb{R}^d, , \tag{2}$$

we refer to the problem as **constraint sampling** problem.

Introduction to sampling: why we need sampling?

Why we need sampling:

- Many target distributions $\pi(x)$ are **intractable**:
  - Normalizing constant is unknown
  - Dimension d is very high
  - Only unnormalized density is available

- Sampling provides a way to compute expectations:

$$\mathbb{E}_\pi[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i), \quad x_i \sim \pi$$

- Applications:
  - Numerical integration in high dimensions
  - Bayesian inference and posterior analysis
  - Simulation in physics, finance, and machine learning

Introduction to sampling: how can we do sampling?

How can we solve sampling problems: Generate a sequence of random variables X_n such that the law of X_n converge to the target distribution π in some sense, namely

$$\mu_n = \mathrm{Law}(X_n) \xrightarrow{(?)} \pi$$

Methods: Direct Sampling, Markov Chain Monte Carlo (MCMC), Langevin Monte Carlo (LMC), etc. In this talk, we mainly talk about Langevin Monte Carlo.

How can we quantify the convergence: Kullback–Leibler (KL) divergence, Total Variation (TV) distance, Wasserstein distance.

Quantify the convergence: KL divergence

Definition (KL divergence)

Let μ and ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then the KL divergence of μ with respect to ν is defined by

$$\mathrm{KL}(\mu\|\nu) = \begin{cases} \int_{\mathbb{R}^d} \log \dfrac{d\mu}{d\nu}(x)\ \mu(dx) & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}.$$

Lemma

For unconstrained sampling problem $\pi \propto e^{-f(x)}$ which is supported on the whole $\mathbb{R}^d$, we have

$$\mathrm{KL}(\mu\|\pi) = \begin{cases} \int_{\mathbb{R}^d} \dfrac{d\mu}{dx}(x) \log \left[\dfrac{d\mu}{dx}(x) / \pi(x) \right] dx & \text{if } \mu \ll m \\ +\infty & \text{otherwise} \end{cases}.$$

Quantify the convergence: TV distance

Definition (TV distance)

Let μ and ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, then the Tv distance of μ and ν is defined by

$$\mathrm{TV}(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|.$$

Example

Let $x, y \in \mathbb{R}^d$, then the TV distance of Dirac measure at x and y is

$$\mathrm{TV}(\delta_x, \delta_y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}.$$

Quantify the convergence: Wasserstein distance

Definition (Wasserstein distance)

Let μ and ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with finite second moment, $\Gamma(\mu,\nu)$ denotes the set of all couplings of μ and ν, namely

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

Then the 2-Wasserstein distance between μ and ν is defined as

$$W_2(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \gamma(dx,dy) \right)^{1/2}$$

Remark: By the Prokhorov theorem, it is easy to show that the set $\Gamma(\mu,\nu)$ is compact and the infimum in the definition of Wasserstein distance is attained and we call a probability measure γ^* that achieves this infimum an **optimal coupling**.

Quantify the convergence: Wasserstein distance

Example

Let $x, y \in \mathbb{R}^d$, then the 2-Wasserstein distance of Dirac measure at x and y is

$$W_2(\delta_x, \delta_y) = \|x - y\|_2.$$

Theorem

Let $\{\mu_n\}_{n\geq 1}$ and μ be probability measures in $\mathcal{P}_2(\mathbb{R}^d)$ (i.e. each has finite second moment). Then the following are equivalent:

1. *$W_2(\mu_n, \mu) \to 0$ as $n \to \infty$;*
2. *$\mu_n \rightharpoonup \mu$ (weak convergence) and*

$$\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu_n(x) \;\to\; \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(x).$$

Outline

Langevin Dynamics

We want to sample from a distribution π on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ which admits a density w.r.t. Lebesgue measure m, also denoted by $\pi(x)$ having the form

$$\pi(x) = \frac{e^{-f(x)}}{\int_{\mathbb{R}^d} e^{-f(x)} dx}, \quad x \in \mathbb{R}^d, \tag{1}$$

where we assume $f \in C^2(\mathbb{R}^d)$ strongly convex with

$$0 \preceq m I_d \preceq \nabla^2 f(x) \preceq M I_d, \ \forall x \in \mathbb{R}^d \quad \text{and} \quad \int_{\mathbb{R}^d} e^{-f(x)} dx < +\infty.$$

Definition (Langevin Dynamics)

The Langevin dynamics (also called Langevin diffusion) is defined by the following SDE:

$$dX_t = -\nabla f(X_t) dt + \sqrt{2} dW_t, \tag{LD}$$

where W_t is a standard d-dimensional Brownian motion (also called Wiener process).

How does Langevin dynamics work for sampling?

Let X_t be the unique solution of (LD) and $\mu_t = \text{Law}(X_t)$. Then by **Malliavin calculus**, for ant initial condition and all $t > 0$, μ_t always has a density denoted by ρ_t which is also $C^\infty-$smooth. Hence

$$\text{KL}(\mu_t \| \pi) = \int_{\mathbb{R}^d} \rho_t \log \frac{\rho_t}{\pi} dx.$$

Theorem (Fokker-Planck Equation)

The density ρ_t satisfies the following **Fokker-Planck** *equation,*

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi} \right),$$

where $\nabla\cdot$ denotes the divergence operator, that is, for a vector field $v : \mathbb{R}^d \to \mathbb{R}^d$,

$$\nabla \cdot v(x) = \sum_{i=1}^{d} \frac{\partial v_i(x)}{\partial x_i}.$$

How does Langevin dynamics work for sampling?

Definition (Invariant measure)

We say a probability measure μ is the invariant measure of Langevin dynamics if $X_0 \sim \mu$, then $X_t \sim \mu$ for all $t \geq 0$.

Theorem

The target distribution π is the unique invariant measure of Langevin dynamics.

Analogous to ordinary differential equation dynamical systems, where the existence of a unique fixed point guarantees that trajectories starting from any initial condition converge to this point, Langevin dynamics exhibits a similar property: as $t \to \infty$, the law of the process "converges" to its invariant measure, regardless of the initial condition.

Langevin Monte Carlo

Recall the Langevin dynamics

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t. \tag{LD}$$

Definition (Langevin Monte Carlo algorithm)

The Langevin Monte Carlo (LMC) algorithm is given by

$$X^{(k+1,h)} = X^{(k,h)} - h\nabla f(X^{(k,h)}) + \sqrt{2h}\,\xi^{k+1}, \tag{LMC}$$

where $h > 0$ is the step size, $\xi^{(1)}, \cdots, \xi^{(k)}, \cdots$ are i.i.d. $N(0, I_d)$ random variables.

Remark: Langevin Monte Carlo (LMC) algorithm can be seen as the Euler-Maruyama discretization of Langevin dynamics (LD). Based on the results for numerical solutions of stochastic differential equations with additive noise, the (LMC) algorithm achieves a mean-square convergence rate of $O(h)$ for the (LD).

Outline

Analysis of Langevin Monte Carlo

As the Euler-Maruyama discretization of Langevin dynamics,

$$X^{(k+1,h)} = X^{(k,h)} - h\nabla f(X^{(k,h)}) + \sqrt{2h}\,\xi^{k+1}, \tag{LMC}$$

where $h > 0$ is the step size, $\xi^{(1)}, \cdots, \xi^{(k)}, \cdots$ are i.i.d. $N(0, I_d)$ random variables, the law of (LMC) iterates may converge to the target distribution.

There are various approaches to establish the convergence of (LMC):

Approach	Reference
Proof via Wasserstein coupling	[Dalalyan, 2017]
Proof via interpolation argument	[Vempala and Wibisono, 2019]
Proof via convex optimization	[Wibisono, 2018], [Durmus et al., 2019]
Proof via Girsanov's theorem	[Dalalyan and Tsybakov, 2012]

Table: Various approaches to establish the convergence of LMC. See [Chewi, 2024] for more details.

Analysis of Langevin Monte Carlo via Convex Optimization

With loss of generality, we can assume the target distribution $\pi(x) = e^{-f(x)}$, we now view sampling as the composite optimization problem of minimizing the objective

$$\mathrm{KL}(\mu\|\pi) = \underbrace{\int_{\mathbb{R}^d} f \; d\mu}_{:=\mathcal{E}(\mu)} + \underbrace{\int_{\mathbb{R}^d} \mu \log \mu \; dx}_{:\mathcal{H}(\mu)},$$

where the two terms are the **energy** and the **negative entropy**. Accordingly, we break up the iterates of LMC into the steps:

$$X^+_{k,h} = X_{k,h} - h\nabla f(X_{k,h}), \qquad \text{(foward-step)}$$

$$X_{k+1,h} = X^+_{k,h} + \sqrt{2h}\, \xi^{(k+1)}. \qquad \text{(flow-step)}$$

The main idea of the proof is to show that the forward step of LMC dissipates the energy while not increasing the entropy too much, and that the flow step of LMC dissipates the entropy while not increasing the energy too much.

Analysis of Langevin Monte Carlo via Convex Optimization

Theorem ([Durmus et al., 2019])

Suppose that $\pi(x) = \exp(-f(x))$ *is the target distribution and that* f *is convex with* $\|\nabla^2 f(x)\|_2 \leq M$. *Let* $(\mu_{nh})_{n\in\mathbb{N}}$ *denote the laws of LMC iterates. For any* $\varepsilon \in [0, \sqrt{d}]$, *if we take step size* $h \asymp \frac{\varepsilon^2}{Md}$, *then for the mixture distribution* $\bar{\mu}_{Nh} := \frac{1}{N}\sum_{n=1}^{N} \mu_{nh}$, *it holds that* $\sqrt{\mathrm{KL}(\bar{\mu}_{Nh} \,\|\, \pi)} \leq \varepsilon$ *after*

$$N = O\left(\frac{Md\ \mathrm{W}_2^2(\mu_0, \pi)}{\varepsilon^4}\right)$$

iterations.

Proof Sketch

Step 1: The forward step dissipates the energy $\mathcal{E}$. Since $X^+_{k,h}$ is obtained from $X_{k,h}$ via a gradient step on f, one can show that

$$\mathcal{E}(\mu^+_{kh}) - \mathcal{E}(\pi) \leq \frac{1}{2h}[W_2^2(\mu_{kh}, \pi) - W_2^2(\mu^+_{kh}, \pi)]$$

Step 2: The flow step does not substantially increase the energy $\mathcal{E}$. Using the $M-$smoothness of f, one can show that

$$\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\mu^+_{kh}) \leq Mdh.$$

Step 3: The flow step dissipates the entropy $\mathcal{H}$. Using some properties of Wasserstein gradient flow, one can show that

$$\mathcal{H}(\mu_{(k+1)h}) - \mathcal{H}(\pi) \leq \frac{1}{2h}[W_2^2(\mu^+_{kh}, \pi) - W_2^2(\mu_{(k+1)h}, \pi)].$$

Proof Sketch

Adding the above three inequalities yields

$$\mathrm{KL}(\mu_{(k+1)h}\|\pi) \le \frac{1}{2h}[W_2^2(\mu_{kh},\pi) - W_2^2(\mu_{(k+1)h},\pi)] + Mdh.$$

By summing the above inequality and using the convexity of the KL divergence, we have

$$\mathrm{KL}(\bar{\mu}_{Nh}\|\pi) \le \frac{1}{N}\,\mathrm{KL}(\mu_{kh}\|\pi) \le \frac{\mathrm{W}_2^2(\mu_0,\pi)}{2Nh} + Mdh.$$

The result follows from our choice of h and N.

Outline

Sampling can be used to solve optimization problems

Consider the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x),$$

where $f \in C^2(\mathbb{R}^d)$ is strongly convex with

$$0 \preceq mI_d \preceq \nabla^2 f(x) \preceq MI_d, \ \forall x \in \mathbb{R}^d \quad \text{and} \quad \int_{\mathbb{R}^d} e^{-f(x)} dx < +\infty.$$

Hence for any $\tau > 0$, the unique minimizer of the above optimization satisfies

$$x^* = \arg\min_{x\in\mathbb{R}^d} f(x) = \arg\min_{x\in\mathbb{R}^d} f(x)/\tau = \arg\max_{x\in\mathbb{R}^d} e^{-f(x)/\tau}$$

If we consider the sampling problem of $\pi_\tau \propto e^{-f(x)}/\tau$, then

$$x^* = \arg\max_{x\in\mathbb{R}^d} \pi_\tau(x)$$

Sampling can be used to solve optimization problems

Theorem (Laplace Theorem)

Under the previous assumption of f, we have

$$\overline{x_\tau} := \int_{\mathbb{R}^d} x\pi_\tau(x)dx \to x^*, \quad \text{as } \tau \to 0^+,$$

and the distribution $\pi_\tau \rightharpoonup \delta_{x^*}$ as $\tau \to 0^+$.

This theorem tells us that **if τ is chosen sufficiently small, the samples obtained by sampling from π_τ will, with high probability, be close to x^*.**

Numerical Experiments

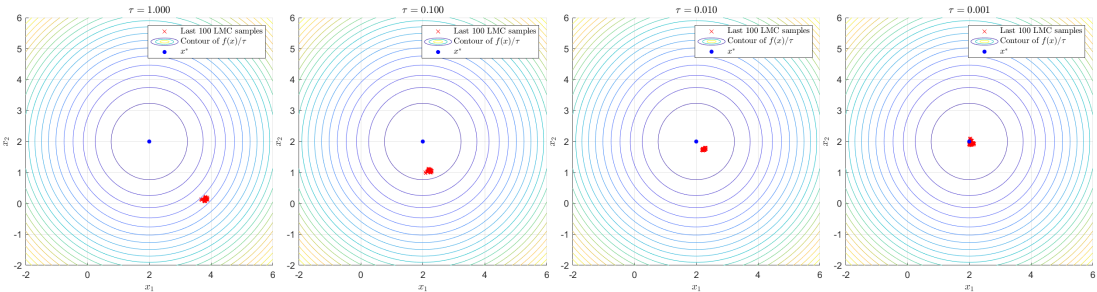

Figure: LMC sampling for the function $f(x_1, x_2)/\tau = [0.05(x_1 - 2)^2 + 0.05(x_2 - 2)^2]/\tau$. The initial distribution is chosen as the Dirac measure at $(x_1, x_2) = (1, 1)$, with step size $h = 0.001$ and total iterations $N = 10000$ with the last 200 samples chosen. The four subplots correspond to $\tau = 1, 0.1, 0.01, 0.001$.

Numerical Experiments

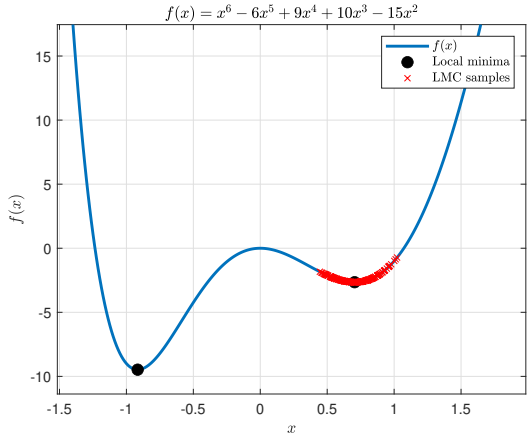

Figure: LMC sampling for the function $f(x) = x^6 - 6x^5 + 9x^4 + 10x^3 - 15x^2$. The initial distribution is chosen as a Dirac measure at $x_0 = 1$, with step size $h = 0.001$ and total number of samples $N = 10000$. The last 200 samples are highlighted.

Sampling can be used to solve optimization problems

Theorem ([Seyoum and You, 2025])

Under the previous assumption of f and $d = 1$, assume we have i.i.d. a set of samples $\mathcal{X} = \{X_1, X_2, \cdots\}$ drawn from the distribution $\pi \propto e^{-f(x)}$. Then, the minimum number of samples N required to guarantee that the probability of at least one sample lying within an ε-ball of x^* is greater than δ is:

$$\frac{\ln(1-\delta)}{\ln\left(1 - \frac{Z_{\rho_2}}{Z_\pi}\left(\Phi(\sqrt{M}\varepsilon) - \Phi(-\sqrt{M}\varepsilon)\right)\right)} \leq N \leq \frac{\ln(1-\delta)}{\ln\left(1 - \frac{Z_{\rho_1}}{Z_\pi}\left(\Phi(\sqrt{m}\varepsilon) - \Phi(-\sqrt{m}\varepsilon)\right)\right)},$$

where $Z_{\rho_1} = \int_{\mathbb{R}} e^{-g_1(x)} dx$ with $g_1(x) = f(x^*) + \frac{m}{2}\|x - x^*\|_2^2$, $Z_{\rho_2} = \int_{\mathbb{R}} e^{-g_2(x)} dx$ with $g_2(x) = f(x^*) + \frac{M}{2}\|x - x^*\|_2^2$ and $\Phi(x)$ is the c.d.f. of standard Gaussian distribution.

Relationship between LMC and Gradient Descent

Theorem ([Dalalyan, 2017])

Under the previous assumption of f, let ν_k be the distribution of the LMC algorithm, that is $X^{(k,h)} \sim \nu_k$. If $h \in (0, \frac{2}{m+M})$, then

$$\mathrm{W}_2(\nu_k, \pi) \leq (1 - mh)^k \, \mathrm{W}_2(\nu_0, \pi) + 2(M/m)(hd)^{1/2}.$$

Since $f_\tau(x) := f(x)/\tau$ satisfies the previous assumption with constant $m_\tau = m/\tau$ and $M_\tau = M/\tau$, we can apply π_τ and $\nu_0 = \delta_{x^{(0)}}$ to this theorem, which tell us that if τ is sufficiently small and we choose $h = 1/M_\tau = \tau/M$, then

$$\mathrm{W}_2(\nu_k, \pi_\tau) \leq \left(1 - \frac{m}{M}\right)^k \mathrm{W}_2(\delta_{x^{(0)}}, \pi_\tau) + 2\left(\frac{M}{m}\right)\left(\frac{\tau d}{M}\right)^{1/2}.$$

Relationship between LMC and Gradient Descent

On the other hand, the LMC algorithm with the step size $h = \tau/M$ applies to f_τ reads as

$$X^{(k+1,h)} = X^{(k,h)} - \frac{1}{M}\nabla f(X^{(k,h)}) + \sqrt{\frac{2\tau}{M}}\,\xi^{k+1}, \quad k = 0, 1, 2\cdots$$

When the parameter τ goes to 0, the above LMC sequence tends to the gradient descent sequence

$$x^{(k+1)} = x^{(k)} - \frac{1}{M}\nabla f(x^{(k)}), \quad k = 0, 1, 2\cdots$$

The limiting case in the previous slide corresponding to $\tau \to 0$ writes as

$$\|x^{(k)} - x^*\|_2 \leq \left(1 - \frac{m}{M}\right)^k \|x^{(0)} - x^*\|_2,$$

which is the well-known result in optimization.

Outline

Sampling as optimization in the space of measures

According to the goal of sampling, the sampling problem can be seen as the following optimization in the space of measures:

$$\min_{\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)} \mathcal{F}(\mu) := \mathrm{KL}(\mu \| \pi),$$

where

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

Theorem

$(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathrm{W}_2)$ *is a Riemann manifold.*

Outline

Gradient flow in Euclidean space

If we want to $\min_{x\in\mathbb{R}^d} f(x)$, where f satisfies the previous assumptions, we can consider the following gradient flow,

$$\begin{cases} \dfrac{dx(t)}{dt} = -\nabla f(x(t)) \\ x(0) = x_0 \end{cases}$$

Theorem

(a) *The trajectory of $f(x(t))$ is decreasing;*

(b) *Let $f^* = \min_{x\in\mathbb{R}^d} f(x)$, then*

$$f(x(t)) - f^* \leq e^{-2mt}[f(x(0)) - f^*];$$

(c) $\lim_{t\to\infty} f(x(t)) = f^*$.

Gradient flow in Wasserstein space

Let $t \mapsto v_t$ be a family of vector fields such that the random variables $t \mapsto X_t$ evolve according to

$$\begin{cases} \dfrac{dX_t}{dt} = v_t(X_t) \\ X_0 \sim \mu_0. \end{cases}$$

Here, we let $\mu_t = \mathrm{Law}(X_t)$ and we want to minimize the KL-divergence functional over a Riemann manifold,

$$\min_{\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)} \mathcal{F}(\mu) := \mathrm{KL}(\mu \| \pi).$$

Theorem

The Wasserstein gradient of a functional $\mathcal{F}$ in Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathrm{W}_2)$ is just the gradient of the first variation of $\mathcal{F}$ in the Euclidean space, that is

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

Gradient flow in Wasserstein space

Definition (first variation of a functional)

The first variation $\delta \mathcal{F}$ of a functional $\mathcal{F}$ is defined by

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathcal{F}(\mu + \varepsilon\varphi) = \int_{\mathbb{R}^d} \varphi(x)\delta \mathcal{F}(x)dx, \qquad \text{for all } \varphi \in C_c^\infty(\mathbb{R}^d).$$

Remark: In $(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathrm{W}_2)$, all probability measure has a unique density function in the sense of L^2. Hence, with a slight abuse of notation, we may regard a functional as a function defined on the space of probability density functions.

Example

For KL-divergence $\mathcal{F}(\mu) := \mathrm{KL}(\mu\|\pi)$, we have

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

Gradient flow in Wasserstein space

Now, we need to consider what is the relationship between μ_t and v_t in the following flow system

$$\begin{cases} \dfrac{dX_t}{dt} = v_t(X_t), \quad X_t \sim \mu_t \\ X_0 \sim \mu_0. \end{cases}$$

Theorem (Continuity Equation)

Let $t \mapsto v_t$ be a family of vector fields such that the random variables $t \mapsto X_t$ evolve according to $\dfrac{dX_t}{dt} = v_t(X_t)$. Then the law of X_t evolves according to the **continuity equation**

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

where $\nabla\cdot$ also denotes the divergence operator.

Sampling as a Wasserstein Gradient Flow

In the previous slides, we have shown that the Wasserstein gradient of KL-divergence functional $\mathcal{F}(\mu) = \mathrm{KL}(\mu\|\pi)$ is

$$\nabla_{\mathrm{W}_2} \mathcal{F}(\mu) = \nabla \log \frac{\mu}{\pi},$$

here in the right hand side, with a slight abuse of notation, μ and π denote the density of probability of μ and π, respectively.
Hence, if we choose the vector fields

$$v_t(\mu_t) = -\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_t) = -\nabla \log \frac{\mu_t}{\pi}$$

and by the continuity equation, we have

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

which exactly matches the Fokker–Planck equation of (LD).

Sampling as a Wasserstein Gradient Flow

We have already known that the Langevin dynamics

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{LD}$$

the density of μ_t of X_t satisfies the Fokker-Planck equation

$$\begin{cases} \dfrac{\partial \mu_t}{\partial t} = \nabla \cdot \left(\mu_t \nabla \log \dfrac{\mu_t}{\pi} \right), \\ \int_{\mathbb{R}^d} \mu_t(x)dx = 1, \ \mu_0 = \mathrm{Law}(X_0). \end{cases}$$

This is also precisely the PDE corresponding to the Wasserstein gradient flow of the KL divergence functional in Wasserstein space. Therefore, by the uniqueness of the solution to this PDE, **the Langevin dynamics is exactly the Wasserstein gradient flow of the KL divergence functional in Wasserstein space**. This amazing connection between the Langevin dynamics and gradient flow was first drawn in [Jordan et al., 1998].

Outline

Constraint Sampling

If the target distribution is supported on a subset K of $\mathbb{R}^d$, namely the corresponding density having the form

$$\pi(x) = \frac{e^{-f(x)} 1_{\{x \in K\}}}{\int_K e^{-f(x)} dx}, \quad x \in \mathbb{R}^d, , \tag{2}$$

we refer to the problem as **constraint sampling** problem.

There are many applications of constraint sampling:

- **Bayesian Inference:** Ensures posterior samples respect prior knowledge or physical constraints on parameters.
- **Constrained Optimization:** Sampling can approximate solutions by concentrating on feasible regions near the optimum.
- **Finance:** Generates feasible portfolios under budget or risk constraints for risk analysis.

Projected Langevin Monte Carlo

Projected Langevin Monte Carlo Algorithm

Denote P_K as the Euclidean projection onto K, then the projected LMC algorithm is defined as

$$X_{k+1} = P_K(X_k - \frac{\eta}{2}\nabla f(X_k) + \sqrt{\eta}\,\xi^{k+1}), \qquad \text{(PLMC)}$$

where $\eta > 0$ is the step size, $\xi^{(1)}, \cdots, \xi^{(k)}, \cdots$ are i.i.d. $N(0, I_d)$ random variables.

Theorem ([Bubeck et al., 2018])

Let $K \subset \mathbb{R}^d$ be a convex body such that $0 \in K$, K contains a Euclidean ball of radius $r = 1$, and K is contained in a Euclidean ball of radius R. Assume f is $L-$Lipschitz and $\varepsilon > 0$. Then one has $\mathrm{TV}(X_N, \pi) \leq \varepsilon$, provided that $\eta = \tilde{\Theta}(R^2/N)$ and that N satisfies the following:

$$N = \tilde{\Theta}\left(\frac{R^6 \max(d, RL, RM)^{12}}{\varepsilon^{12}}\right).$$

Outline

Analogy between Sampling and Optimization

There is a strong analogy between sampling and optimization algorithms.

Optimization Algorithm	**Sampling Algorithm**
Gradient Descent (GD)	Langevin Monte Carlo (LMC)
Projected Gradient Descent (PGD)	Projected LMC [Bubeck et al., 2018]
Penalty Method	Penalized Langevin [Karagulyan and Dalalyan, 2020]
Stochastic Gradient Descent (SGD)	Noisy LMC [Dalalyan, 2017]
Mirror Descent	Mirror-Langevin Dynamics [Hsieh et al., 2020]
Newton Method	Newton-Langevin [Wang and Li, 2020]
Quasi-Newton Method	Quasi-Newton LMC [Simsekli et al., 2016]
Trust Region Method	?
...	?

Table: Comparison between Optimization and Sampling Algorithms

References I

Bubeck, S., Eldan, R., and Lehec, J. (2018).

Sampling from a log-concave distribution with projected langevin monte carlo.

Discrete & Computational Geometry, 59(4):757–783.

Chewi, S. (2024).

Log-concave Sampling.

Unfinished Draft.

Dalalyan, A. (2017).

Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent.

In Kale, S. and Shamir, O., editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689. PMLR.

Dalalyan, A. S. and Tsybakov, A. B. (2012).

Sparse regression learning by aggregation and langevin monte-carlo.

Journal of Computer and System Sciences, 78(5):1423–1443.

References II

Durmus, A., Majewski, S., and Miasojedow, B. (2019).
Analysis of langevin monte carlo via convex optimization.
Journal of Machine Learning Research, 20(73):1–46.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. (2020).
Mirrored langevin dynamics.

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.

Karagulyan, A. and Dalalyan, A. S. (2020).
Penalized langevin dynamics with vanishing penalty for smooth and log-concave targets.

Seyoum, N. and You, H. (2025).
Beyond smoothness and convexity: Optimization via sampling.

References III


Simsekli, U., Badeau, R., Cemgil, T., and Richard, G. (2016).
Stochastic quasi-newton langevin monte carlo.
In *International Conference on Machine Learning*, pages 642–651. PMLR.


Vempala, S. and Wibisono, A. (2019).
Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices.
Advances in neural information processing systems, 32.


Wang, Y. and Li, W. (2020).
Information newton's flow: second-order optimization method in probability space.
arXiv preprint arXiv:2001.04341.


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.

Thanks!