[AI2613] Convergence of Langevin Diffusion, DDPM

June 4, 2024

1 Convergence of Langevin diffusion

In this section, we analyse the convergence rate of Langevin diffusion.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-strongly convex function [1]. Recall that when analysing the gradient flow, we calculate $\frac{\mathrm{d}\|X_t - Y_t\|^2}{\mathrm{d}t}$, where X_t and Y_t come from the ODE $\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t$ and $\mathrm{d}Y_t = -\nabla f(Y_t)\,\mathrm{d}t$ and start from some $X_0, Y_0 \in \mathbb{R}^d$ respectively.[2]

[1] A differentiable function f is μ-strongly convex if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

[2] Unless otherwise stated, $\|\cdot\|$ refers to the Euclidean norm in this note.

For the Langevin diffusion, let $\{X_t\}$ and $\{Y_t\}$ be two processes generated by

$$\begin{cases} \mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \mathrm{d}B_t \\ \mathrm{d}Y_t = -\nabla f(Y_t)\,\mathrm{d}t + \mathrm{d}B_t \end{cases}. \tag{1}$$

Assume Y_0 is drawn from the stationary distribution $\pi \propto e^{-f}$ and the distribution of X_0 can be arbitrary. We use the Wasserstein metric to measure the distance between the distribution of X_t and Y_t.

For two distributions μ and ν, let ζ be the set of all couplings between μ and ν. The Wasserstein distance $W_2(\mu, \nu) = \inf_{\omega \in \zeta} \mathbf{E}_{(X,Y)\sim\omega}\left[\|X - Y\|^2\right]^{\frac{1}{2}}$.

We couple X_t and Y_t with the same Brownian motion. From Equation (1), we have

$$\frac{\mathrm{d}\,(X_t - Y_t)}{\mathrm{d}t} = \nabla f(Y_t) - \nabla f(X_t).$$

Consequently,

$$\begin{aligned} \frac{\mathrm{d}\mathbf{E}\left[\|X_t - Y_t\|^2\right]}{\mathrm{d}t} &= 2\mathbf{E}\left[\langle \frac{\mathrm{d}\,(X_t - Y_t)}{\mathrm{d}t}, X_t - Y_t \rangle\right] \\ &= 2\mathbf{E}\left[\langle \nabla f(Y_t) - \nabla f(X_t), X_t - Y_t \rangle\right] \\ &\overset{(\heartsuit)}{\leq} -2\mu\mathbf{E}\left[\|X_t - Y_t\|^2\right], \end{aligned} \tag{2}$$

where $(\heartsuit)$ follows from

$$\begin{cases} f(y) - f(x) \geq +\langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2 \\ f(x) - f(y) \geq +\langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|_2^2 \end{cases}$$

which is due to the strong convexity of f. Equation (2) indicates the convergence of Langevin diffusion

$$\mathbf{E}\left[\|X_t - Y_t\|\right] \leq e^{-2\mu t}\mathbf{E}\left[\|X_0 - Y_0\|^2\right].$$

For the discretized Langevin algorithm, we left its analysis in homework.

2 DDPM

Note that the above analysis relies heavily on the strong log-concavity of π (or equivalently, the strong convexity of f), which does not hold for most distributions in practice. One recent popular method to sample a general distribution π is the so-called *denoising diffusion probabilistic modeling (DDPM)*.

The idea of DDPM is to add Gaussian noise to an initial sample $X_0 \sim \pi$ until the distribution is close enough to a Gaussian. Then it starts with a sample from the Gaussian distribution and executes a reverse process to generate a sample from the target distribution.

To be specific, in the forward procedure, we run an Ornstein-Uhlenbeck process (OU process) starting from $X_0 \sim \pi$ and get $\left\{\overline{X}_t\right\}$. Recall that the OU process follows the stochastic differential equation $\mathrm{d}\overline{X}_t = -\overline{X}_t \,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$. Therefore we have

$$\overline{X}_t = e^{-t}X_0 + \sqrt{1-e^{-2t}}Z_t$$

where $Z_t \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable. Running this for a time of T, we get a series of random variables $\left\{\overline{X}_t\right\}$ and $\overline{X}_T$ is close to a Gaussian random variable.

Let q_t be the law of $\overline{X}_t$. In the reversed process, we first sample $\overleftarrow{X}_T$ from a Gaussian distribution and then calculate $\left\{\overleftarrow{X}_t\right\}$ according to the reversed equation of OU process:

$$\mathrm{d}\overleftarrow{X}_t = \left(\overleftarrow{X}_t + 2\boldsymbol{\nabla}\log q_{T-t}\left(\overleftarrow{X}_t\right)\right)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t. \tag{3}$$

If law of $\overleftarrow{X}_0$ is $\mathcal{N}(0,1)$, then the law of $\overleftarrow{X}_0$ will converge to π. In the following, we will prove that Equation (3) is indeed the reverse of the OU process.

2.1 The Reverse of OU Process

Recall that for a diffusion $\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}B_t$, the generator $\mathcal{L}$ satisfies

$$\mathcal{L}f(x) = \mu(x)\cdot\partial_x f + \frac{1}{2}\sigma^2(x)\cdot\partial_x^2 f \tag{4}$$

for any function f. Also,

$$\mathcal{L}^* f(x) = -\partial_x\left(\mu\cdot f\right) + \frac{1}{2}\partial_x^2\left(\sigma^2\cdot f\right). \tag{5}$$

Recall the Kolmogorov Forward equation $\frac{\partial P_t}{\partial t} = \mathcal{L}^* P_t$. This is equivalent to say

$$\forall s < t,\ \partial_t p[X_t = x | X_s = y] = \mathcal{L}^* p[X_t = x | X_s = y]. \tag{6}$$

With fixed y, for any $s > \tau$, let $f(x) = \delta(y-x)$ and $f_\tau(x) = Q_\tau f(x) = p[X_s = y | X_{s-\tau} = x]$. Then the Kolmogorov backward equation indicates that

$$\partial_\tau p[X_s = x | X_{s-\tau} = y] = \mathcal{L}p[X_s = x | X_{s-\tau} = y],$$

or equivalently,

$$\forall t < s, -\partial_t p[X_s = x | X_t = y] = \mathcal{L}p[X_s = x | X_t = y]. \tag{7}$$

Here we slightly abuse the notation and let $p[X_s = y | X_{s-\tau} = x]$ represent the density of X_s at y given $X_{s-\tau} = x$. Similarly define $p[X_t = x]$ and $p[X_s = y, X_t = x]$.

In the following part, we abbreviate $p[X_s = x_s, X_t = x_t]$ as $p(x_s, x_t)$ and $p[X_s = x_s | X_t = x_t]$ as $p(x_s|x_t)$ (here $s > t$). We have

$$\begin{aligned} -\partial_t p(x_s, x_t) &= -\partial_t \left(p(x_s|x_t) \cdot p(x_t)\right) \\ &= -p(x_t) \cdot \partial_t p(x_s|x_t) - p(x_s|x_t) \cdot \partial_t p(x_t). \end{aligned} \tag{8}$$

From Equations (5) and (6),

$$\begin{aligned} \partial_t p(x_t) &= \mathcal{L}^* p(x_t) \\ &= -\partial_{x_t} \left(\mu(x_t) \cdot p(x_t)\right) + \frac{1}{2} \cdot \partial^2_{x_t} \left(\sigma^2(x_t) \cdot p(x_t)\right). \end{aligned} \tag{9}$$

From Equations (4) and (7),

$$\begin{aligned} \partial_t \left(p(x_s|x_t)\right) &= -\mathcal{L} p(x_s|x_t) \\ &= -\mu(x_t) \cdot \partial_{x_t} p(x_s|x_t) - \frac{1}{2}\sigma^2(x_t)\partial^2_{x_t} p(x_s|x_t). \end{aligned} \tag{10}$$

Plugging Equations (9) and (10) into Equation (8), we have

$$\begin{aligned} -\partial_t p(x_s, x_t) = p(x_t) \cdot &\left(\mu(x_t) \cdot \partial_{x_t} p(x_s|x_t) + \frac{1}{2}\sigma^2(x_t)\partial^2_{x_t} p(x_s|x_t)\right) \\ &+ p(x_s|x_t) \cdot \left(\partial_{x_t} \left(\mu(x_t) \cdot p(x_t)\right) - \frac{1}{2} \cdot \partial^2_{x_t} \left(\sigma^2(x_t) \cdot p(x_t)\right)\right). \end{aligned} \tag{11}$$

Note that

$$\partial_{x_t} p(x_s|x_t) = \partial_{x_t} \left(\frac{p(x_s, x_t)}{p(x_t)}\right) = \frac{\partial_{x_t} p(x_s, x_t)}{p(x_t)} - \frac{p(x_s, x_t) \cdot \partial_{x_t} p(x_t)}{p^2(x_t)}$$

and

$$\partial_{x_t} \left(\mu(x_t) \cdot p(x_t)\right) = p(x_t) \cdot \partial_{x_t} \left(\mu(x_t)\right) + \mu(x_t) \cdot \partial_{x_t} \left(p(x_t)\right).$$

Then by direct calculation, we can further write Equation (11) as

$$\begin{aligned} -\partial_t p(x_s, x_t) = \partial_{x_t} \left(p(x_s, x_t) \cdot \mu(x_t)\right) &+ \frac{1}{2} p(x_t)\sigma^2(x_t) \cdot \partial^2_{x_t} p(x_s|x_t) \\ &- \frac{1}{2} p(x_s|x_t) \cdot \partial^2_{x_t} \left(\sigma^2(x_t) \cdot p(x_t)\right) \end{aligned} \tag{12}$$

We have

$$\begin{aligned} \frac{1}{2}\partial^2_{x_t} \left(\sigma^2(x_t) \cdot p(x_s, x_t)\right) &= \frac{1}{2}\partial^2_{x_t} \left(\sigma^2(x_t) \cdot p(x_s|x_t) p(x_t)\right) \\ &= \frac{1}{2}\sigma^2(x_t) p(x_t) \cdot \partial^2_{x_t} p(x_s|x_t) + \frac{1}{2} p(x_s|x_t) \cdot \partial^2_{x_t} \left(\sigma^2(x_t) p(x_t)\right) \\ &\quad + \partial_{x_t} p(x_s|x_t) \partial_{x_t} \left(\sigma^2(x_t) p(x_t)\right). \end{aligned}$$

Plugging this into Equation (12),

$$\begin{aligned}
-\partial_t p(x_s, x_t) &= \partial_{x_t}\left(p(x_s, x_t)\cdot\mu(x_t)\right) + \frac{1}{2}\partial^2_{x_t}\left(\sigma^2(x_t)\cdot p(x_s, x_t)\right) \\
&\quad - p(x_s|x_t)\cdot\partial^2_{x_t}\left(\sigma^2(x_t)p(x_t)\right) - \partial_{x_t}p(x_s|x_t)\partial_{x_t}\left(\sigma^2(x_t)p(x_t)\right) \\
&= \partial_{x_t}\left(p(x_s, x_t)\cdot\mu(x_t)\right) + \frac{1}{2}\partial^2_{x_t}\left(\sigma^2(x_t)\cdot p(x_s, x_t)\right) \\
&\quad - \partial_{x_t}\left[p(x_s|x_t)\partial_{x_t}\left(\sigma^2(x_t)p(x_t)\right)\right] \\
&= \frac{1}{2}\partial^2_{x_t}\left(\sigma^2(x_t)\cdot p(x_s, x_t)\right) + \partial_{x_t}\left[p(x_s, x_t)\cdot\left(\mu(x_t) - \frac{1}{p(x_t)}\partial_{x_t}\left(\sigma^2(x_t)p(x_t)\right)\right)\right].
\end{aligned}$$

Since $p(x_s, x_t) = p(x_t|x_s)p(x_s)$, we have

$$-\partial_t p(x_t|x_s) = \frac{1}{2}\partial^2_{x_t}\left(\sigma^2(x_t)\cdot p(x_t|x_s)\right)+\partial_{x_t}\left[p(x_t|x_s)\cdot\left(\mu(x_t) - \frac{1}{p(x_t)}\partial_{x_t}\left(\sigma^2(x_t)p(x_t)\right)\right)\right].$$

In the OU process, we have $\mu(x_t) = -x_t$ and $\sigma(x_t) = \sqrt{2}$. Therefore,

$$\begin{aligned}
-\partial_t p(x_t|x_s) &= \partial^2_{x_t}p(x_t|x_s) + \partial_{x_t}\left[p(x_t|x_s)\cdot\left(-x_t - \frac{2}{p(x_t)}\partial_{x_t}p(x_t)\right)\right] \\
&= \partial^2_{x_t}p(x_t|x_s) + \partial_{x_t}\left[p(x_t|x_s)\cdot\left(-x_t - 2\partial_{x_t}\log p(x_t)\right)\right].
\end{aligned}$$

Let $\tau = T - t$. Note that $p(x_t|x_T)$ is exactly the density of $\overleftarrow{X}_\tau$, which is generated by the reversed OU process. From the Kolmogorov forward equation, the reversed process $\left\{\overleftarrow{X}_t\right\}$ satisfies

$$\mathrm{d}\overleftarrow{X}_\tau = \left(\overleftarrow{X}_\tau + 2\nabla\log q_{T-\tau}\left(\overleftarrow{X}_\tau\right)\right)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_\tau,$$

where $q_{T-\tau}$ is the law of $\overline{X}_{T-\tau}$.

2.2 *Score Matching*

To execute the reversed process, one of the most challenging problem is to estimate the score function $\nabla\log q_t(x)$. Let $\mathcal{F}$ be a family of candidate functions, for example, the functions can be represented by neural networks. Our goal is to find

$$\arg\min_{S_t\in\mathcal{F}}\mathbf{E}_{x\sim q_t}\left[\|S_t(x) - \nabla\log q_t(x)\|^2\right].$$

In this subsection, we introduce the idea of score matching.

Since $\nabla\log q_t(x)$ is independent with S_t,

$$\arg\min_{S_t\in\mathcal{F}}\mathbf{E}_{x\sim q_t}\left[\|S_t(x) - \nabla\log q_t(x)\|^2\right] = \arg\min_{S_t\in\mathcal{F}}\mathbf{E}_{x\sim q_t}\left[\|S_t(x)\|^2 - 2\langle S_t(x), \nabla\log q_t(x)\rangle\right].$$

Let $\gamma(\cdot)$ be the density of the standard Gaussian distribution. We then show that estimating $\nabla\log q_t(x)$ is equivalent to find an S_t minimizing $\mathbf{E}_{x\sim q_t}\left[\left\|S_t(x) - \frac{Z_t}{\sqrt{1-e^{-2t}}}\right\|^2\right]$. We give a proof for the one dimensional case. The proof is similar in high-dimension space.

Lemma 1.

$$\arg\min_{S_t \in \mathcal{F}} \mathbf{E}_{x\sim q_t}\left[\|S_t(x) - \nabla \log q_t\ (x)\ \|^2\right] = \arg\min_{S_t \in \mathcal{F}} \mathbf{E}_{x\sim q_t}\left[\left\|S_t(x) - \frac{Z_t}{\sqrt{1-e^{-2t}}}\right\|^2\right].$$

Proof. From direct calculation,

$$\begin{aligned}
\mathbf{E}_{x\sim q_t}\left[\langle S_t(x), \nabla \log q_t\ (x)\rangle\right] &= \int_{\mathbb{R}} S_t(x)\ (\log q_t\ (x))'\ q_t(x)\,\mathrm{d}x \\
&\overset{(\spadesuit)}{=} -\int_{\mathbb{R}} q_t(x) S_t'(x) \mathrm{d}t \\
&= -\int_{\mathbb{R}}\int_{\mathbb{R}} S_t'\left(e^{-t}x_0 + \sqrt{1-e^{-2t}} z_t\right) q_0(x_0)\gamma(z_t)\,\mathrm{d}x_0\,\mathrm{d}z_t \\
&\overset{(\clubsuit)}{=} -\frac{1}{\sqrt{1-e^{-2t}}}\int_{\mathbb{R}} q_0(x_0)\int_{\mathbb{R}} S_t\left(e^{-t}x_0 + \sqrt{1-e^{-2t}} z_t\right)\gamma(z_t)\cdot z_t\,\mathrm{d}x_0\,\mathrm{d}z_t \\
&= \mathbf{E}_{x\sim q_t}\left[\left\langle S_t(x), \frac{Z_t}{\sqrt{1-e^{-2t}}}\right\rangle\right],
\end{aligned}$$

where ($\spadesuit$) and ($\clubsuit$) is derived through integrating by parts. Therefore,

$$\begin{aligned}
\arg\min_{S_t \in \mathcal{F}} \mathbf{E}_{x\sim q_t}\left[\|S_t(x) - \nabla \log q_t\ (x)\ \|^2\right] &= \arg\min_{S_t \in \mathcal{F}} \mathbf{E}_{x\sim q_t}\left[\|S_t(x)\|^2 - 2\left\langle S_t(x), \frac{Z_t}{\sqrt{1-e^{-2t}}}\right\rangle\right] \\
&= \arg\min_{S_t \in \mathcal{F}} \mathbf{E}_{x\sim q_t}\left[\left\|S_t(x) - \frac{Z_t}{\sqrt{1-e^{-2t}}}\right\|^2\right].
\end{aligned}$$

□

References