Unit 12: Denoising Diffusion Models

12.1 Introduction

Recall that the Langevin diffusion is given by

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where π is the stationary distribution. Hence, if we simulate the process for a sufficiently long period T, then X_T can be thought of as a sample from π. However, determining a sufficiently large value of T can be quite difficult, and it highly depends on the property of π.

Denoising diffusion models have dynamics similar to the Langevin diffusion. But one specifies the value of T first and then modifies the drift term according to how much time is left so that X_T is guaranteed to be distributed according to π (at least approximately). The following theorem gives an example, which is a special case of Theorem 12.2.

Theorem 12.1. *Let π be a probability distribution on $\mathbb{R}^d$ satisfying certain regularity conditions. Denote the density function of $N(0, \sigma^2 I_d)$ by $\phi_\sigma(x)$, and denote the convolution of two probability distributions μ, ν by $\mu * \nu$. Let $(X_t)_{0 \le t \le 1}$ be a diffusion process given by*

$$\begin{aligned} &X_0 \sim \pi * \phi_\sigma, \\ &\mathrm{d}X_t = \sigma^2 \nabla_x \log h(x,t)\mathrm{d}t + \sigma \mathrm{d}B_t, \ \textit{where } h(x,t) = \int_{\mathbb{R}^d} \pi(y)\phi_{\sigma\sqrt{1-t}}(x-y)\mathrm{d}y. \end{aligned} \tag{1}$$

Then, $X_1 \sim \pi$.

Example 12.1. Consider part (ii) with $d = 1$ and π being the standard normal distribution $N(0,1)$. Then, $X_0 \sim N(0, 1+\sigma^2)$ and

$$\mathrm{d}X_t = A_t X_t \mathrm{d}t + \sigma \mathrm{d}B_t, \text{ where } A_t = -\frac{\sigma^2}{\sigma^2(1-t)+1}.$$

Due to the linearity, the solution can be explicitly expressed by

$$X_t = \Phi_t X_0 + \sigma \Phi_t \int_0^t \Phi_s^{-1} \mathrm{d}B_s, \text{ where } \Phi_t = e^{\int_0^t A_s \mathrm{d}s} = \frac{\sigma^2(1-t)+1}{\sigma^2+1}.$$

By a result known as Itô isometry,

$$\mathrm{Var}\left(\int_0^t \Phi_s^{-1} \mathrm{d}B_s \right) = \int_0^t \Phi_s^{-2} \mathrm{d}s.$$

This can be used to compute the distribution of X_t for every $t \in [0,1]$; the answer is given in the exercise below. We will see that a similar result holds for any general choice of π.

Exercise 12.1. Show that in Example 12.1, $X_t \sim N(0, \sigma^2(1-t)+1)$ for every $t \in [0,1]$.

In this unit and the next, we will address the following questions.

(i) How to understand Theorem 12.1, at least intuitively. See Section 12.2.

(ii) How to utilize the SDE given in (1) to devise sampling algorithms. The function $h(x,t)$ is an integral with respect to the distribution π. Approximating this integral can be as challenging as sampling from π. See Section 12.3.

(iii) How to relax the assumption $X_0 \sim \pi * N(0, \sigma^2 I_d)$. Exact sampling from $\pi * N(0, \sigma^2 I_d)$ does not seem easier than sampling from π. See the next unit.

Before proceeding, an important remark on the background is needed. So far in this course, we have considered sampling problems where π is typically known up to a normalizing constant or at least has an explicit expression. Diffusion models like (1) are widely used in *generative modeling*, where the problem setup is quite different: π is completely unknown, but we have samples drawn from π. In theory, one can use these samples to first estimate π and then apply the sampling algorithms we have discussed. However, as we will see, a better approach is to directly estimate $\nabla_x \log h(x,t)$ using the samples without learning π and then simulate the SDE (1).

12.2 Reverse-time SDE

Given a diffusion process Y_t, if we observe the process backward in time, can its dynamics still be described by a SDE? The following result of [1] provides an answer.

Theorem 12.2. *Let $(Y_t)_{0 \le t \le T}$ be a diffusion over the time interval $[0,T]$, evolving by*

$$\mathrm{d}Y_t = b(Y_t, t)\mathrm{d}t + \sigma(Y_t, t)\mathrm{d}B_t,$$

where b, t are continuously differentiable. Let $a = \sigma\sigma^\top$ and denote the Lebesgue density of the distribution of Y_t by $p_t(y)$ (assumed to exist).[1] *Under certain regularity conditions, the time-reversed process, $(Y_{T-t})_{0 \le t \le T}$ has the same distribution as the diffusion process $(X_t)_{0 \le t \le T}$ such that*

$$\begin{aligned} &X_0 \sim p_T, \\ &\mathrm{d}X_t = -b^*(X_t,\, T-t)\mathrm{d}t + \sigma(X_t,\, T-t)\mathrm{d}B_t, \end{aligned} \tag{2}$$

where

$$\begin{aligned} b^*(x,t) &= b(x,t) - \nabla_x \cdot a(x,t) - a(x,t)\nabla_x \log p_t(x), \\ \text{i.e., } b_i^*(x,t) &= b_i(x,t) - \frac{1}{p_t(x)} \sum_{1 \le j,k \le d} \frac{\partial}{\partial x_j} \left\{ p_t(x)\sigma_{ik}(x,t)\sigma_{jk}(x,t) \right\}. \end{aligned}$$

[1] By a slight abuse of notation, we will also use p_t to denote the distribution of Y_t.

Example 12.2. Suppose $\sigma(x,t) = \sigma_t I$ for some $\sigma_t > 0$. Then the SDE (2) is simplified to

$$\mathrm{d}X_t = \left[-b(X_t, T-t) + \sigma_t^2 \nabla_x \log p_{T-t}(X_t)\right] \mathrm{d}t + \sigma_t \mathrm{d}B_t, \tag{3}$$

which has been widely used in the literature on denoising diffusion probabilistic models [2, 4].

If we further assume $T = 1$, $b \equiv 0$ and $\sigma_t \equiv \sigma$, we have

$$Y_t = Y_0 + \sigma B_t,$$

and we obtain Theorem 12.1. Note that p_0 is also the distribution of X_1 (i.e., the distribution π in Theorem 12.1), and the distribution of X_t is p_{1-t}, the convolution of p_0 and $N(0, \sigma^2(1-t)I)$, which has density $h(x,t)$ as given in (1). This generalizes the claim in Exercise 12.1.

Example 12.3. Suppose that Y_t is a Langevin diffusion with dynamics $\mathrm{d}Y_t = \nabla \log \pi(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$. Let $Y_0 \sim \pi$, which implies $Y_t \sim \pi$ for every t. One can check that in this case, the time-reversed process X_t follows exactly the same SDE, which is expected since Langevin diffusions are reversible.

Remark 12.1. The proof of Theorem 12.2 requires Kolmogorov forward/backward equations, and we refer readers to [1] for details. Here we provide a heuristic argument to offer some insights into Theorem 12.2. We consider the special case $d = 1$, $b(x,t) = b(x)$ and $\sigma(x,t) = \sigma > 0$. Assume that $b(x), \partial b(x)/\partial x, \partial p_t(x)/\partial t$ and $\partial^2 p_t(x)/\partial x^2$ all exist and are bounded over all (x,t).

As we have argued in Remark 10.2, for sufficiently regular function f,

$$\lim_{h \downarrow 0} \frac{\mathbb{E}\left[f(Y_{t+h}) \,|\, Y_t = y\right] - f(y)}{h} = b(y) f'(y) + \frac{1}{2}\sigma^2 f''(y). \tag{4}$$

Hence, if $X_t = Y_{T-t}$ has the dynamics given in (3), it should satisfy that

$$\lim_{h \downarrow 0} \frac{\mathbb{E}\left[f(Y_{t-h}) \,|\, Y_t = x\right] - f(x)}{h} = \left\{-b(x) + \sigma^2 \nabla_x \log p_t(x)\right\} f'(x) + \frac{1}{2}\sigma^2 f''(x). \tag{5}$$

We now show that (4) indeed implies (5). As in Remark 10.2, we use the second-order Taylor expansion of f, and our main task is to determine

$$\lim_{h \downarrow 0} \frac{\mathbb{E}\left[Y_{t-h} - Y_t \,|\, Y_t = x\right]}{h}, \text{ and } \lim_{h \downarrow 0} \frac{\mathbb{E}\left[(Y_{t-h} - Y_t)^2 \,|\, Y_t = x\right]}{h}.$$

The key distinction from the forward-time analysis is that $Y_{t-h} - Y_t$ and Y_t are dependent. Indeed, if we approximate the conditional distribution of Y_t given Y_{t-h} using Euler-Maruyama discretization, i.e.,

$$Y_t \,|\, Y_{t-h} = y \sim N\left(y + hb(y),\, h\sigma^2\right)$$

then the conditional density of $Y_{t-h} = y$ given $Y_t = x$ is

$$p_{t-h|t}(y \,|\, x) = \frac{p_{t-h}(y)}{p_t(x)} \frac{1}{\sqrt{2\pi h \sigma^2}} \exp\left\{-\frac{(x - y - hb(y))^2}{2h\sigma^2}\right\}. \tag{6}$$

By Taylor expansion and our assumption on the derivatives of $p_t(x)$,

$$\begin{aligned}\log p_{t-h}(y) - \log p_t(x) &= \log p_{t-h}(y) - \log p_t(y) + \log p_t(y) - \log p_t(x) \\ &= (y-x)\nabla_x \log p_t(x) + O(h) + O((y-x)^2).\end{aligned}$$

Similarly, we have $b(y) = b(x) + O(|y-x|)$. Plugging these approximations into (6), we get

$$p_{t-h|t}(y \,|\, x) \propto \exp\left[-\frac{1+O(h)}{2h\sigma^2}\left\{y - x + hb(x) - h\sigma^2\nabla_x \log p_t(x)\right\}^2 + O(h)\right].$$

It then follows that

$$\begin{aligned}\lim_{h\downarrow 0} \frac{\mathbb{E}\left[Y_{t-h} - Y_t \,|\, Y_t = x\right]}{h} &= -b(x) + \sigma^2 \nabla_x \log p_t(x), \\ \lim_{h\downarrow 0} \frac{\mathbb{E}\left[(Y_{t-h} - Y_t)^2 \,|\, Y_t = x\right]}{h} &= \sigma^2.\end{aligned}$$

12.3 Generative Modeling and Score Matching

Consider a general modeling problem where we have access to samples from an unknown distribution π. In this section, we explain how to use Theorem 12.1 to build algorithms that can output new samples from π. As we have discussed in Section 12.1, the initial condition in Theorem 12.1 is difficult to simulate exactly. So we simply assume that σ is sufficiently large so that X_0 approximately follows the normal distribution $N(0, \sigma^2 I)$. To simulate the SDE given in (1),

$$\mathrm{d}X_t = \sigma^2 \nabla_x \log h(x,t)\mathrm{d}t + \sigma \mathrm{d}B_t, \quad t \in [0,1],$$

we use the so-called "score matching" technique [3, 5], which we present in Theorem 12.3 below. It allows us to directly estimate $\nabla_x \log h(x,t)$ without learning π.

Let $s(x,t)$ be an estimator for the score $\nabla_x \log h(x,t)$ parameterized by θ (e.g., s can be a neural network model with parameter vector θ). We can train this estimator (i.e., learn the value of θ) by minimizing some loss function. The question is how to define the loss function. Let's first fix an arbitrary $t \in [0,1]$ and consider measuring the loss at time t. As explained in Example 12.2, the joint distribution of (X_t, X_1) can be described by

$$X_1 \sim \pi, \quad X_t \,|\, X_1 \sim N(X_1, \, \sigma^2(1-t)I), \tag{7}$$

and the marginal distribution of X_t has density

$$h(x,t) = \int_{\mathbb{R}^d} \pi(y)\phi_{\sigma\sqrt{1-t}}(x-y)\mathrm{d}y.$$

So we measure the loss (at time t) of the estimator $s(x,t)$ by

$$J_t(s) = \mathbb{E}\|s(X_t,t) - \nabla_x \log h(X_t,t)\|_2^2 = \int \|s(x,t) - \nabla_x \log h(x,t)\|_2^2 \, h(x,t)\mathrm{d}x.$$

By Theorem 12.3, instead of matching the marginal score $\nabla_x \log h(X_t, t)$, we can also match the score $\nabla_{x_t} \log q(X_t \,|\, X_1)$ by conditioning on X_1, where $q(X_t \,|\, X_1)$ denotes the conditional density given in (7). Explicitly,

$$J_t(s) = \mathbb{E}\|s(X_t, t) + \sigma^{-2}(1-t)^{-1}(X_t - X_1)\|_2^2 + C$$

where C is a constant that does not depend on s, and the expectation is taken over the joint distribution given in (7). Estimating this loss is straightforward since we have samples from π; denote them by $X_{1,1}, X_{1,2}, \dots, X_{1,n}$. By generating i.i.d. $Z_1, \dots, Z_n$ from $N(0, I_d)$, we get the empirical loss

$$L_t(s) = \frac{1}{n}\sum_{i=1}^{n} \|s(X_{1,i} + \eta_t Z_i, t) + \eta_t^{-1} Z_i\|_2^2, \text{ where } \eta_t = \sqrt{\sigma^2(1-t)},$$

which measures how well $s(x,t)$ approximates $\nabla_x \log h(x,t)$ at time t. To measure the performance of $s(x,t)$ across the time interval $[0,1]$, we can sample $t_1, \dots, t_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0,1)$ and define the overall empirical loss by

$$L(s) = \frac{1}{n}\sum_{i=1}^{n} w(t_i)\|s(X_{1,i} + \eta_i Z_i, t_i) + \eta_i^{-1} Z_i\|_2^2, \text{ where } \eta_i = \sqrt{\sigma^2(1-t_i)},$$

and $w\colon [0,1] \to (0,\infty)$ is a weighting function chosen by the user.

Theorem 12.3. *Let (X,Y) be random vectors with joint Lebesgue density function $q_{X,Y}(x,y)$. Denote the marginal densities by $q_X(x), q_Y(y)$, and denote the conditional density functions by $q_{X|Y}(x \,|\, y), q_{Y|X}(y \,|\, x)$. Assume that $q_{X,Y}$ is sufficiently regular so that for any y,*[2]

$$\int q_{X|Y}(x \,|\, y) \nabla_y \log q_{X|Y}(x \,|\, y) \mathrm{d}x = 0 \tag{8}$$

Then, for any function $s(y)$,

$$\int \|s(y) - \nabla \log q_Y(y)\|_2^2 \, q_Y(y) \mathrm{d}y = \int \|s(y) - \nabla_y \log q_{Y|X}(y \,|\, x)\|_2^2 \, q_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y + C$$

where C is a constant independent of s.

Proof. By Fisher's identity which is given in Exercise (12.2) below,

$$\begin{aligned}\int \left\{s(y)^\top \nabla \log q_Y(y)\right\} q_Y(y)\mathrm{d}y &= \int \left\{ s(y)^\top \int q_{X|Y}(x \,|\, y) \nabla_y \log q_{Y|X}(y \,|\, x) \mathrm{d}x \right\} q_Y(y) \mathrm{d}y \\ &= \int \left\{s(y)^\top \nabla_y \log q_{Y|X}(y \,|\, x)\right\} q_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y.\end{aligned}$$

[2] All we need is that differentiation and integration can be interchanged in (8).

At the same time,

$$\int \|s(y)\|_2^2 \, q_Y(y) \mathrm{d}y = \int \|s(y)\|_2^2 \, q_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Hence,

$$\mathbb{E}\|s(Y) - \nabla \log q_Y(Y)\|_2^2 = \mathbb{E}\left[\|s(Y)\|_2^2 - 2s(Y)^\top \nabla_y \log q_{Y|X}(Y \mid X)\right] + C',$$

where C' is some constant independent of s. A simple calculation completes the proof. $\square$

Exercise 12.2. Prove that (8) implies

$$\int q_{X|Y}(x \mid y) \nabla_y \log q_{Y|X}(y \mid x) \mathrm{d}x = \nabla \log q_Y(y).$$

References

[1] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[3] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695–709, 2005.

[4] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

[5] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.