Lecture 6: Score-based diffusion models

2025–05–28

1.1 Forward processes

Consider the SDE in $\mathbb{R}^d$

$$\begin{aligned} dX_t =& f(X_t, t)\, dt + \sigma(t) dB_t\,, \quad t \in [0, T]\,, \\ X_0 \sim& p_0\,, \end{aligned} \tag{1.1}$$

where $T > 0$, $f : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ and $\sigma : [0, T] \to \mathbb{R}^+$ are C^1-smooth functions. Note that for simplicity we assume that σ does not depend on the state X_t.

The probability density function of X_t, denoted by $p(x, t)$, satisfies the Fokker-Planck equation

$$\begin{aligned} &\frac{\partial p}{\partial t} = \mathcal{L}_t^\top p\,, \quad t \in [0, T]\,, \\ &p(x, 0) = p_0(x)\,, \end{aligned} \tag{1.2}$$

where $\mathcal{L}_t$ denotes the generator of (1.1) at time t, which is defined as

$$(\mathcal{L}_t g)(x) = f(x, t) \cdot \nabla g(x) + \frac{\sigma^2(t)}{2} \Delta g(x)\,, \quad g : \mathbb{R}^d \to \mathbb{R}\,, \tag{1.3}$$

and $\mathcal{L}_t^\top$ is the adjoint operator of $\mathcal{L}_t$, whose explicit expression can be computed as

$$(\mathcal{L}_t^\top g)(x) = -\mathrm{div}(f(x, t) g(x)) + \frac{\sigma^2(t)}{2} \Delta g(x)\,, \quad g : \mathbb{R}^d \to \mathbb{R}\,. \tag{1.4}$$

Given $x_0 \in \mathbb{R}^d$, we denote by $p(x, t|x_0)$ the probability density of the process (1.1) starting from the fixed state $X_0 = x_0$. Then, $p(x, t|x_0)$ satisfies the Fokker-Planck equation with the initial condition $p(x, 0|x_0) = \delta(x - x_0)$. For the process (1.1) starting from a general initial density p_0, we have

$$p(x, t) = \int_{\mathbb{R}^d} p(x, t|x_0) p_0(x_0) dx_0\,. \tag{1.5}$$

Below we study an example where $p(x, t|x_0)$ has an explicit expression.

Example (Linear SDEs). Consider the process

$$dX_t = -\alpha(t) X_t dt + \sqrt{\beta(t)} dB_t\,, \tag{1.6}$$

where $\alpha(t), \beta(t) : [0, T] \to \mathbb{R}^+$. Apply Ito's formula, we can integrate (1.6) and obtain

$$X_t = \mathrm{e}^{-\int_0^t \alpha(s) ds} X_0 + \int_0^t \mathrm{e}^{-\int_s^t \alpha(r) dr} \sqrt{\beta(s)} dB_s\,, \quad t \geq 0\,. \tag{1.7}$$

Assume that $X_0 = x_0$ is fixed. Then, X_t is a Gaussian random variable at any $t > 0$. The mean and the covariance of X_t can be calculated as

$$\begin{aligned} &\mathbb{E}(X_t) = \mathrm{e}^{-\int_0^t \alpha(s) ds} x_0\,, \\ &\mathbb{E}\Big((X_t - x_0)(X_t - x_0)^\top\Big) = \eta^2(t) \mathbf{1}_d\,, \end{aligned} \tag{1.8}$$

where $\mathbf{1}_d$ denotes the identity matrix of size d, and

$$\eta^2(t) = \int_0^t \mathrm{e}^{-2\int_s^t \alpha(r)dr}\beta(s)ds\,. \tag{1.9}$$

Therefore, we have obtained

$$X_t \sim \mathcal{N}\Big(\mathrm{e}^{-\int_0^t \alpha(s)ds}x_0, \eta^2(t)\mathbf{1}_d\Big)\,, \tag{1.10}$$

and we have

$$p(x,t|x_0) = \big(2\pi\eta^2(t)\big)^{-\frac{d}{2}}\mathrm{e}^{-\frac{1}{2\eta^2(t)}|x-\mathrm{e}^{-\int_0^t \alpha(s)ds}x_0|^2}\,, \quad x\in\mathbb{R}^d\,. \tag{1.11}$$

1.2 Time reversal

Let us define the time-reversal of p:

$$q(x,t) = p(x,T-t), \quad \forall\, x\in\mathbb{R}^d, \quad t\in[0,T]\,. \tag{1.12}$$

The following theorem states that q is the probability density of a (backward) diffusion process.

Theorem 1. The probability density function q in (1.12) is the probability density of the process Y_t that is governed by the following SDE

$$\begin{aligned} dY_t =& f^-(Y_t,t)\,dt + \sigma^-(t)dB_t\,, \quad t\in[0,T]\,,\\ Y_0 \sim& p(\cdot,T)\,, \end{aligned} \tag{1.13}$$

where the coefficients are given by

$$\begin{aligned} f^-(x,t) =& -f(x,T-t) + \sigma^2(T-t)\nabla\ln p(x,T-t)\\ \sigma^-(t) =& \sigma(T-t)\,, \end{aligned} \tag{1.14}$$

for $t\in[0,T]$ and $x\in\mathbb{R}^d$.

Proof. To show that q is the probability density of Y_t, it is sufficient to verify that it satisfies the Fokker-Planck equation associated to (1.13):

$$\begin{aligned} &\frac{\partial q}{\partial t} = (\bar{\mathcal{L}}_t)^\top q\,, \quad t\in[0,T]\,,\\ &q(x,0) = p(x,T)\,. \end{aligned} \tag{1.15}$$

where $\bar{\mathcal{L}}_t$ denotes the generator of (1.13) at time t. Similarly as in (1.4), we have

$$(\bar{\mathcal{L}}_t)^\top g(x) = -\mathrm{div}\Big(f^-(x,t)g(x)\Big) + \frac{(\sigma^-)^2(t)}{2}\Delta g(x)\,. \tag{1.16}$$

Clearly, (1.12) implies that the initial condition in (1.15) is satisfied. Using

(1.12) and the fact that p satisfies (1.2), we can derive

$$\begin{aligned}
\frac{\partial q}{\partial t}(x,t) =& -\frac{\partial p}{\partial t}(x,T-t)\\
=& -\mathcal{L}_{T-t}^{\top} p(x,T-t)\\
=& -\Big[-\operatorname{div}\Big(f(x,T-t)p(x,T-t)\Big)+\frac{\sigma^2(T-t)}{2}\Delta p(x,T-t)\Big]\\
=&\operatorname{div}\Big(f(x,T-t)p(x,T-t)\Big)-\frac{\sigma^2(T-t)}{2}\Delta p(x,T-t)\\
=&\operatorname{div}\Big(f(x,T-t)p(x,T-t)-\sigma^2(T-t)\nabla p(x,T-t)\Big)\\
&+\frac{\sigma^2(T-t)}{2}\Delta p(x,T-t)\\
=&-\operatorname{div}\Big[\Big(-f(x,T-t)+\sigma^2(T-t)\nabla \ln p(x,T-t)\Big)p(x,T-t)\Big]\\
&+\frac{\sigma^2(T-t)}{2}\Delta p(x,T-t)\\
=&-\operatorname{div}\Big(f^-(x,t)q(x,t)\Big)+\frac{(\sigma^-)^2(t)}{2}\Delta q(x,t)\\
=&(\bar{\mathcal{L}}_t)^{\top} q(x,t)\,,
\end{aligned}$$

where we have used (1.14). Therefore, q satisfies the Fokker-Planck equation (1.15). □

1.3 Loss function

Assume that a dataset is given where the data is sampled from a target density p_{target}. Our goal is to learn a model that allows to generate new samples from p_{target}.

Choose the initial distribution $p_0 = p_{\text{target}}$ in (1.1). Then, the forward process (1.1) transforms the target distribution to the density $p(x,T)$. Theorem 1 tells us that the probability densities of the backward process Y_t at time $t=0$ and $t=T$ are

$$q(x,0) = p(x,T), \quad \text{and } q(x,T) = p(x,0) = p_{\text{target}}, \tag{1.17}$$

respectively. Therefore, we can sample the target density p_{target} by simulating the backward process (1.13), provided that we can easily generate samples $Y_0 \sim p(x,T)$ and compute $\nabla \ln p$, which is called the score function and is involved in the drift f^- of the SDE (1.13).

In theory, we can compute the score by solving the optimization problem

$$\begin{aligned}
&\min_{u:\mathbb{R}^d\times[0,T]\to\mathbb{R}^d} \mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x\sim p(\cdot,t)}\Big[\frac{1}{2}\big|u(x,t)-\nabla\ln p(x,t)\big|^2 w(t)\Big]\\
=&\min_{u:\mathbb{R}^d\times[0,T]\to\mathbb{R}^d}\Big[\frac{1}{T}\int_0^T\Big(\int_{\mathbb{R}^d}\frac{1}{2}\big|u(x,t)-\nabla\ln p(x,t)\big|^2 p(x,t)dx\Big)w(t)dt\Big],
\end{aligned} \tag{1.18}$$

where $w(t):[0,T]\to\mathbb{R}^+$ is a weight function, because the score is obviously the minimizer of (1.18). However, (1.18) is not useful in practice because the density $p(x,t)$ is unknown.

Notice that the density $p(x,t)$ can be written as the integral involving $p(x,t|x_0)$ in (1.5) and $p(x,t|x_0)$ has an explicit expression when the forward process is linear (see (1.6) and (1.11)). Therefore, the idea is to rewrite (1.18) using (1.5) and derive an objective involving $p(x,t|x_0)$ instead of $p(x,t)$.

First, we write the objective in (1.18) as

$$\begin{aligned}
&\frac{1}{2}\int_0^T \Big[\int_{\mathbb{R}^d} |u(x,t)-\nabla \ln p(x,t)|^2 p(x,t)dx\Big] w(t)dt \\
=&\frac{1}{2}\int_0^T \Big[\int_{\mathbb{R}^d} \Big|u(x,t)-\frac{\nabla p(x,t)}{p(x,t)}\Big|^2 p(x,t)dx\Big]\, w(t)dt \\
=&\int_0^T \Big[\int_{\mathbb{R}^d} \Big(\frac{1}{2}|u(x,t)|^2 p(x,t) - u(x,t)\cdot\nabla p(x,t) + \frac{1}{2}\frac{|\nabla p(x,t)|^2}{p(x,t)}\Big)dx\Big] w(t)\, dt \\
=&\int_0^T \Big[\int_{\mathbb{R}^d} \Big(\frac{1}{2}|u(x,t)|^2 p(x,t) - u(x,t)\cdot\nabla p(x,t)\Big)dx\Big]\, w(t)dt + C_1\,,
\end{aligned} \tag{1.19}$$

where C_1 is a constant independent of u.

Substituting (1.5) into (1.19), we derive

$$\begin{aligned}
&\int_0^T \Big[\int_{\mathbb{R}^d} \Big(\frac{1}{2}|u(x,t)|^2 p(x,t) - u(x,t)\cdot\nabla p(x,t)\Big)dx\Big]\, w(t)dt \\
=&\int_0^T \Big\{\int_{\mathbb{R}^d} \Big[\frac{1}{2}|u(x,t)|^2 \int_{\mathbb{R}^d} p(x,t|x_0)p_0(x_0)dx_0 \\
&\qquad - u(x,t)\cdot\Big(\int_{\mathbb{R}^d} \nabla p(x,t|x_0)p_0(x_0)dx_0\Big)\Big]dx\Big\}\, w(t)dt \\
=&\int_0^T \Big[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big(\frac{1}{2}|u(x,t)|^2 p(x,t|x_0) - u(x,t)\cdot\nabla p(x,t|x_0)\Big)p_0(x_0)dx_0dx\Big]\, w(t)dt \\
=&\int_0^T \Big[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Big(\frac{1}{2}|u(x,t)|^2 - u(x,t)\cdot\nabla \ln p(x,t|x_0)\Big)p(x,t|x_0)p_0(x_0)dx_0dx\Big]\, w(t)dt \\
=&\int_0^T \Big[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{1}{2}\big|u(x,t)-\nabla \ln p(x,t|x_0)\big|^2 p(x,t|x_0)p_0(x_0)dx_0dx\Big]\, w(t)dt + C_2 \\
=&T\mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim p(\cdot,t|x_0)}\Big[\frac{1}{2}\big|u(x,t)-\nabla \ln p(x,t|x_0)\big|^2 w(t)\Big] + C_2 \\
=&T\mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim p(\cdot,t|x_0)}\Big[\Big(\frac{1}{2}|u(x,t)|^2 - u(x,t)\cdot\nabla \ln p(x,t|x_0)\Big)w(t)\Big] + C_3,
\end{aligned} \tag{1.20}$$

where $C2, C_3$ are constants independent of u.

To summarize, we have obtained the following result.

Proposition 1. For $u : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, we have

$$\begin{aligned}
&\mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x\sim p(\cdot,t)}\Big[\frac{1}{2}\big|u(x,t) - \nabla \ln p(x,t)\big|^2 w(t)\Big] \\
=&\mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim p(\cdot,t|x_0)}\Big[\frac{1}{2}\big|u(x,t) - \nabla \ln p(x,t|x_0)\big|^2 w(t)\Big] + C_2 \\
=&\mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim p(\cdot,t|x_0)}\Big[\Big(\frac{1}{2}|u(x,t)|^2 - u(x,t)\cdot\nabla \ln p(x,t|x_0)\Big) w(t)\Big] + C_3,
\end{aligned} \tag{1.21}$$

where C_2, C_3 are constants independent of u.

Notice that $\nabla \ln p(x,t|x_0)$ diverges as $t \to 0^+$. Therefore, when designing loss function in practice, we often modify $p(x,t|x_0)$ in Proposition 1 and choose $w(t)$ properly in order to avoid numerical issue at $t = 0$. In the following example, we illustrate this point with a concrete forward process.

Example (VESDE). Consider the linear SDE (1.6), where we choose

$$\begin{aligned}
\alpha(t) =& 0\,, \\
\beta(t) =& \frac{2\eta_{\min}^2}{T}\Big(\frac{\eta_{\max}}{\eta_{\min}}\Big)^{2t/T} \ln\Big(\frac{\eta_{\max}}{\eta_{\min}}\Big), \quad t \in [0,T]\,,
\end{aligned} \tag{1.22}$$

and $0 < \eta_{\min} \le \eta_{\max}$. The SDE becomes

$$dX_t = \sqrt{\beta(t)}dB_t\,, \tag{1.23}$$

which can be solved directly as

$$X_t = X_0 + \int_0^t \sqrt{\beta(s)}dB_s\,, \tag{1.24}$$

and from (1.9) we have

$$\eta(t) = \Big(\int_0^t \beta(s)ds\Big)^{\frac{1}{2}} = \eta_{\min}\sqrt{\Big(\frac{\eta_{\max}}{\eta_{\min}}\Big)^{2t/T} - 1}, \quad t \in [0,T]\,. \tag{1.25}$$

Assume that $X_0 \sim \mathcal{N}(x_0, \eta_{\min}^2 \mathbf{1}_d)$. Equivalently,

$$X_0 = x_0 + \eta_{\min} z, \quad \text{where } z \in \mathcal{N}(0, 1_d)\,. \tag{1.26}$$

Then

$$X_t = x_0 + \eta_{\min} z + \int_0^t \sqrt{\beta(s)}dB_s \sim \mathcal{N}(x_0, \tilde{\eta}^2(t)\mathbf{1}_d) \tag{1.27}$$

where

$$\tilde{\eta}(t) = \sqrt{\eta^2(t) + \eta_{\min}^2} = \eta_{\min}\Big(\frac{\eta_{\max}}{\eta_{\min}}\Big)^{t/T}. \tag{1.28}$$

The density of X_t is

$$\tilde{p}(x,t|x_0) = \big(2\pi\tilde{\eta}^2(t)\big)^{-\frac{d}{2}} \mathrm{e}^{-\frac{1}{2}|x-x_0|^2/\tilde{\eta}^2(t)}\,, \quad x \in \mathbb{R}^d\,. \tag{1.29}$$

Replacing $p(x,t|x_0)$ by $\tilde{p}(x,t|x_0)$ and choosing $w(t) = \tilde{\eta}^2(t)$ in Proposition 1, we obtain

$$\begin{aligned}
\mathrm{Loss}(u) =& \mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim\tilde{p}(\cdot,t|x_0)}\Big[\Big(\frac{1}{2}|u(x,t)|^2 - u(x,t)\cdot\nabla\ln\tilde{p}(x,t|x_0)\Big)w(t)\Big] \\
=& \mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{x\sim\tilde{p}(\cdot,t|x_0)}\left[\Big(\frac{1}{2}|u(x,t)|^2 + u(x,t)\cdot\frac{x-x_0}{\tilde{\eta}^2(t)}\Big)\tilde{\eta}^2(t)\right] \\
=& \mathbb{E}_{t\sim U([0,T])}\mathbb{E}_{x_0\sim p_0}\mathbb{E}_{z\sim\mathcal{N}(0,\mathrm{I}_d)}\left[\Big(\frac{1}{2}|u(x,t)|^2 + \frac{u(x,t)\cdot z}{\tilde{\eta}(t)}\Big)\tilde{\eta}^2(t)\right],
\end{aligned} \tag{1.30}$$

where the last equality follows from the fact that samples $x \sim \tilde{p}(\cdot,t|x_0)$ can be directly obtained as $x = x_0 + \tilde{\eta}(t)z$, where $z \sim \mathcal{N}(0,\mathrm{I}_d)$.