Copyright by Hongyun Wang, UCSC

Diffusion Models

1 The forward diffusion

1 SDE of the forward diffusion

Let $X(0)$ be a random sample from an underlying distribution that, in applications, is represented by a data set. Mathematically, we view $X(0)$ as a random variable (RV) with a continuous distribution. We evolve $X(t)$ stochastically in time by i) making it gradually forget the current position and ii) adding gradually independent noise to it. This process is called the forward diffusion and is achieved by the stochastic differential equation (SDE) below.

$$dX(t) = \underbrace{-\beta X(t)dt}_{\text{drift to 0}} + \underbrace{\sigma dW(t)}_{\text{diffusion}}, \tag{1}$$

$$dX(t) \equiv X(t+dt) - X(t), \quad dW(t) \sim N(0, dt)$$

2 Solution of the forward diffusion

(1) is an Ornstein-Uhlenbeck process. Previously obtained its solution

$$X(t) = X(0)e^{-\beta t} + \sqrt{\sigma^2 \frac{(1-e^{-2\beta t})}{2\beta}}\, Z, \qquad Z \sim N(0,1) \tag{2}$$

As $t \to \infty$, the original information of $X(0)$ is lost and $X(t)$ converges to a pure noise.

$$X(\infty) \equiv \lim_{t\to\infty} X(t) \sim N(0, \sigma_\infty^2), \qquad \sigma_\infty^2 \equiv \frac{\sigma^2}{2\beta} \tag{3}$$

3 Fokker-Planck equation of the forward diffusion

We consider a general SDE. Let $p(x,t)$ be the probability density of $X(t)$ at time t.

$$dX(t) = \underbrace{f(X(t))dt}_{\text{drift}} + \underbrace{\sigma dW(t)}_{\text{diffusion}}, \qquad dW(t) \sim N(0, dt) \tag{4}$$

For SDE (4), the probability density $p(x,t)$ is governed by the Fokker-Planck equation.

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= \underbrace{\nabla\cdot\Big(-f(x)p(x,t)\Big)}_{\text{drift}} + \underbrace{\frac{1}{2}\sigma^2\nabla^2 p(x,t)}_{\text{diffusion}} \\ &= \frac{\partial}{\partial x}\Big(-f(x)p(x,t)\Big) + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2}p(x,t) \end{aligned} \tag{5}$$

In applications, $x \equiv \mathbf{x} \in \mathbb{R}^n$ is in a very high dimensional space (i.e., a 1024×1024 image has $2^{20} \approx 10^6$ pixels). In (5), we used the notations for the general case of vector $\mathbf{x}$.

$$\begin{aligned}&\nabla\cdot\Big(-f(x)p(x,t)\Big) \equiv \nabla\cdot\Big(-f(\mathbf{x})p(\mathbf{x},t)\Big), \\ &\nabla^2 p(x,t) \equiv \nabla\cdot\Big(\nabla p(\mathbf{x},t)\Big)\end{aligned}$$

The first line of (5) is for the general case of vector $\mathbf{x}$; the second line gives the equation for the case of one-dimensional x. Notice in particular the relation between the drift term in the SDE and the that in the Fokker-Planck equation.

$$dX(t) = f(X(t))dt + \sigma dW(t) \quad \Longleftrightarrow \quad \frac{\partial p(x,t)}{\partial t} = \nabla\cdot\Big(-f(x)p(x,t)\Big) + \frac{\sigma^2}{2}\nabla^2 p(x,t)$$

The relation allows us i) to write out the Fokker-Planck equation corresponding to a given SDE or ii) to write out the SDE corresponding to a given Fokker-Planck equation. We will use ii).

In SDE (1), $f(x) = -\beta x$. The Fokker-Planck equation and equilibrium $p_\infty(x)$ are

$$\frac{\partial p(x,t)}{\partial t} = \underbrace{\nabla\cdot\Big((\beta x)p(x,t)\Big)}_{\text{drift}} + \underbrace{\frac{1}{2}\sigma^2\nabla^2 p(x,t)}_{\text{diffusion}} \tag{6}$$

$$\begin{aligned}0 &= \nabla\cdot\Big((\beta x)p_\infty(x) + \frac{1}{2}\sigma^2\nabla p_\infty(x)\Big) \\ &\xrightarrow{\text{eq: zero flux everywhere}} \quad (\frac{2\beta}{\sigma^2}x)p_\infty(x) + \nabla p_\infty(x) = 0 \\ &\xrightarrow{\text{integrating factor}} \quad \nabla\Big(\exp(\frac{\beta}{\sigma^2}\|x\|^2)p_\infty(x)\Big) = 0 \\ &\longrightarrow \quad p_\infty(x) \propto \exp(\frac{-\|x\|^2}{2\sigma_\infty^2}), \qquad \sigma_\infty^2 \equiv \frac{\sigma^2}{2\beta}\end{aligned} \tag{7}$$

which simply confirms the equilibrium (3) we found in SDE (1).

2 The reverse diffusion

The objective of diffusion models is to sample the underlying distribution of $X(0)$. The basic idea of diffusion models is the following.

1. At a large time T, start with a random sample of pure noise: $X(T) \sim N(0, \sigma_\infty^2)$.

2. Evolve $X(t)$ backward in time to reverse the process of the forward diffusion. To make the time reversal well-posed, we the score function, which is learned from data.

3. When we reach time 0, we get a sample of the distribution of $X(0)$.

1 Time reversal of an SDE

We consider SDE (4): $dX(t) = f(X(t))dt + \sigma dW(t)$. Based on the SDE, for small dt, the forward time evolution from $X(t)$ to $X(t+dt)$ is approximately a Gaussian.

$$\Big(X(t+dt)\Big|X(t)=x_0\Big) \sim N\big(x_0 + f(x_0)dt,\ \sigma^2 dt\big) + o(dt)$$

Let $p(x,t) \equiv \rho_{X(t)}(x_0)$ be the probability density of X at time t. The transition density of the reverse time evolution from $X(t+dt) = x_1$ to $X(t)$ is given by Bayes theorem.

$$\begin{aligned}
\rho_{(X(t)\big|X(t+dt)=x_1)}(x_0) &\propto \rho_{(X(t+dt)\big|X(t)=x_0)}(x_1)\rho_{X(t)}(x_0) \\
&\propto \exp\Big(\frac{-\big(x_1 - x_0 - f(x_0)dt\big)^2}{2\sigma^2 dt}\Big)p(x_0,t) \\
&\propto \exp\Big(\frac{-(x_0-x_1)^2}{2\sigma^2 dt} - \frac{2(x_0-x_1)f(x_0) + f^2(x_0)dt}{2\sigma^2} + \ln p(x_0,t)\Big) \qquad (8)
\end{aligned}$$

For small dt, the RHS of (8) is dominated by the factor $\exp\big(\frac{-(x_0-x_1)^2}{2\sigma^2 dt}\big)$, which is significant only in the region of $(x_0 - x_1) = O(\sqrt{dt})$. Note that density $\rho_{(X(t)\big|X(t+dt)=x_1)}(x_0)$ is a function of x_0 with x_1 as a parameter. In (8), we expand $f(x_0)$ and $\ln p(x_0,t)$ about x_1.

$$\begin{aligned}
f(x_0) &= f(x_1) + f'(x_1)(x_0 - x_1) + \cdots \\
\ln p(x_0,t) &= \ln p(x_1,t) + \frac{\partial \ln p(x,t)}{\partial x}\Big|_{x_1}(x_0-x_1) + \frac{1}{2}\frac{\partial^2 \ln p(x,t)}{\partial x^2}\Big|_{x_1}(x_0-x_1)^2 + \cdots
\end{aligned}$$

Using these expansions, we write the exponent in (8) as

$$\begin{aligned}
\text{Exponent in (8)} &= \Big(\frac{-(x_0-x_1)^2}{2\sigma^2 dt} - \frac{2(x_0-x_1)f(x_0) + f^2(x_0)dt}{2\sigma^2} + \ln p(x_0,t)\Big) \\
&= \frac{-1}{2\sigma^2 dt}\Big((x_0-x_1)^2 + 2f(x_1)dt(x_0-x_1) + 2f'(x_1)dt(x_0-x_1)^2 \\
&\quad - 2\sigma^2 dt\frac{\partial \ln p(x,t)}{\partial x}\Big|_{x_1}(x_0-x_1) - \sigma^2 dt\frac{\partial^2 \ln p(x,t)}{\partial x^2}\Big|_{x_1}(x_0-x_1)^2 + \cdots\Big) + const \\
&= \frac{-1}{2\sigma^2 dt}\Big((x_0-x_1)^2 + 2c_1 dt(x_0-x_1) + c_2 dt(x_0-x_1)^2 + \cdots\Big) + const \qquad (9)
\end{aligned}$$

where coefficients c_1 and c_2 are

$$c_1 \equiv f(x_1) - \sigma^2\frac{\partial \ln p(x,t)}{\partial x}\Big|_{x_1}, \qquad c_2 \equiv 2f'(x_1) - \sigma^2\frac{\partial^2 \ln p(x,t)}{\partial x^2}\Big|_{x_1} \qquad (10)$$

Note that there is some "inconsistency" between x and t in coefficients c_1 and c_2: $X(t+dt) = x_1$ is the position of X at time $(t+dt)$. We will address this issue shortly.

In the above, we neglected terms $(dt)^r(x_0 - x_1)^k$ for $r \geqslant 2$ or $k \geqslant 3$ with $r = 1$. We complete the square to write the quadratic form in (9) as (homework)

$$\text{Exponent in (8)} = \frac{-(1+c_2 dt)}{2\sigma^2 dt}\Big(x_0 - x_1 + \frac{c_1 dt}{1+c_2 dt}\Big)^2 + \cdots \tag{11}$$

(11) implies that the reverse time step $\big(X(t)\big|X(t+dt) = x_1\big)$ is also approximately a Gaussian.

$$\Big(X(t)\Big|X(t+dt) = x_1\Big) \sim N\Big(\frac{-c_1 dt}{1+c_2 dt},\ \frac{\sigma^2 dt}{1+c_2 dt}\Big)$$

Using (10) and neglecting $o(dt)$, we write the drift and diffusion terms as

$$\frac{-c_1 dt}{1+c_2 dt} = -c_1 dt + o(dt) = \big(-f(x) + \sigma^2 \nabla \ln p(x, t+dt)\big)\Big|_{x_1} dt + o(dt)$$

$$\frac{\sigma^2 dt}{1+c_2 dt} = \sigma^2 dt + o(dt)$$

Note that in the expressions of drift and diffusion terms above, we have replaced $\ln p(x,t)$ with $\ln p(x, t+dt)$ to make the position and the time both at time $(t+dt)$. The starting point of the time reversal is $X(t+dt) = x_1$. The reverse time evolution from $X(t+dt)$ to $X(t)$ is

$$X(t) = X(t+dt) + \big(-f(x) + \sigma^2 \nabla \ln p(x, t+dt)\big)\Big|_{X(t+dt)} dt + \sigma\sqrt{dt}\, N(0,1)$$

We shift the time and write out the reverse time evolution from $X(t)$ to $X(t-dt)$.

$$X(t-dt) = X(t) + \big(-f(x) + \sigma^2 \nabla \ln p(x, t)\big)\Big|_{X(t)} dt + \sigma\sqrt{dt}\, N(0,1)$$

In SDE (1), $f(x) = -\beta x$. The reverse time evolution of SDE (1) is

$$X(t-dt) = X(t) + \big((\beta x) + \sigma^2 \nabla \ln p(x, t)\big)\Big|_{X(t)} dt + \sigma\sqrt{dt}\, N(0,1) \tag{12}$$

Here $\nabla \ln p(x,t)$ is called the score function. Let $\tau \equiv (T-t)$, $t = (T-\tau)$, and

$$Y(\tau) \equiv X(t) = X(T-\tau), \qquad q(x,\tau) \equiv p(x,t) = p(x, T-\tau)$$

The evolution from $\underbrace{Y(\tau) \text{ to } Y(\tau + d\tau)}_{\text{forward in } \tau}$ corresponds to that from $\underbrace{X(T-\tau) \text{ to } X(T-\tau-d\tau)}_{\text{backward in } t}$.

$$\boxed{dY(\tau) = \big((\beta x) + \sigma^2 \nabla \ln q(x,\tau)\big)\Big|_{Y(\tau)} d\tau + \sigma dW(\tau), \quad Y(T) \sim N(0, \sigma_\infty^2)} \tag{13}$$

2 Fokker-Planck equation for reversing the density

SDE (13) describes the reverse time evolution of stochastic process $X(t)$, which means paths $\{Y(\tau)\colon 0 \leqslant \tau \leqslant T\}$ of SDE (13) are statistically indistinguishable from the flipped forward diffusion paths $\{X(T-t)\colon 0 \leqslant t \leqslant T\}$ of SDE (1).

Recall the objective of diffusion models: drawing samples of the underlying density of $X(0)$. This objective can be achieved by SDE (13), which reverses the pure noise density of $X(T)$ at large T to the density of $X(0)$. However, for the goal of mapping pure noise to a desired density, it does not require the microscopic reverse time evolution of stochastic process $X(t)$. For example, to map $N(0,1)$ at $t=0$ to $N(1,e^{-1})$ at $t=1$, we can do it in many ways.

$$dX = -\beta(X-b)dt + \sqrt{a^2}dW, \quad b = \frac{1}{1-e^{-\beta}}, \quad a^2 = 2\beta\frac{(e^{-1}-e^{-2\beta})}{1-e^{-2\beta}}, \quad \text{for any } \beta > 1.$$

$$dX = \frac{-1}{2}\big(X - \frac{1}{1-e^{-1/2}}\big)dt, \quad \text{a deterministic ODE.}$$

$$dX = \frac{-1}{2}\big(X - t - 2\big)dt, \quad \text{a deterministic ODE.}$$

Fokker-Planck equation (6) evolves the density of $X(0)$ at $t=0$ forward in time to pure noise at large T. We explore reversing (6) in time to evolve the pure noise density at large T to the target density of $X(0)$ at $t=0$. The straightforward time reversal of a diffusion equation is ill-posed. We need to find a well-posed way to reverse (6) in time.

Let $\tau \equiv (T-t)$, $t = (T-\tau)$, and $q(x,\tau) \equiv p(x,t) = p(x,T-\tau)$. Note that evolving backward in t from T to 0 corresponds to evolving forward in τ from 0 to T. For $q(x,\tau)$, (6) becomes

$$\frac{\partial q(x,\tau)}{\partial \tau} = \nabla\cdot\Big(-(\beta x)q(x,\tau)\Big) - \frac{1}{2}\sigma^2\nabla^2 q(x,\tau) \tag{14}$$

This PDE is ill-posed for evolving $q(x,\tau)$ forward in τ because the coefficient of the diffusion term is negative. To make the PDE well-posed, we write the negative diffusion coefficient as the sum of a positive part and a larger negative part.

$$\begin{aligned}-\frac{1}{2}\sigma^2\nabla^2 q(x,\tau) &= \frac{\gamma^2}{2}\sigma^2\nabla^2 q(x,\tau) - \frac{(1+\gamma^2)}{2}\sigma^2\nabla\cdot\Big(\nabla q(x,\tau)\Big)\\ &= \frac{\gamma^2}{2}\sigma^2\nabla^2 q(x,\tau) - \frac{(1+\gamma^2)}{2}\sigma^2\nabla\cdot\Big((\nabla \ln q(x,\tau))q(x,\tau)\Big)\end{aligned}$$

We keep the positive part as the diffusion and move the negative part into the drift term with the help of a key component. We rewrite Fokker-Planck equation (14) as

$$\begin{aligned}\frac{\partial q(x,\tau)}{\partial \tau} &= \nabla\cdot\Big(-(\beta x)q(x,\tau) - \frac{(1+\gamma^2)}{2}\sigma^2(\nabla \ln q(x,\tau))q(x,\tau)\Big) + \frac{\gamma^2}{2}\sigma^2\nabla^2 q(x,\tau)\\ &= \nabla\cdot\Big(\big(-(\beta x) - \frac{(1+\gamma^2)}{2}\sigma^2\nabla \ln q(x,\tau)\big)q(x,\tau)\Big) + \frac{\gamma^2}{2}\sigma^2\nabla^2 q(x,\tau)\end{aligned} \tag{15}$$

(15) is well-posed for evolving $q(x,\tau)$ forward in τ (i.e., evolving $p(x,t)$ backward in t). The key component needed is the score function $\nabla \ln q(x,\tau)$. (15) describes a collection of Fokker-Planck equations, one for each value of $\gamma \geqslant 0$.

3 SDE for reversing the density

For each $\gamma \geqslant 0$, the SDE corresponding to Fokker-Planck equation (15) is

$$\boxed{dY(\tau) = \Big((\beta x) + \frac{(1+\gamma^2)}{2}\sigma^2 \nabla \ln q(x,\tau)\Big)\Big|_{Y(\tau)} d\tau + \gamma\sigma dW(\tau)} \tag{16}$$

For each $\gamma \geqslant 0$, SDE (16) evolves the pure noise density at large T to the target density of $X(0)$ at $t = 0$. We make several comments on SDE (16).

- When $\gamma = 1$, SDE (16) becomes SDE (13), which describes the reverse time evolution of SDE (1). That is, paths $\{Y(\tau)\colon 0 \leqslant \tau \leqslant T\}$ of SDE (13) are statistically indistinguishable from the flipped forward diffusion paths $\{X(T-t)\colon 0 \leqslant t \leqslant T\}$ of SDE (1).
- For $\gamma \neq 1$, paths $\{Y(\tau)\colon 0 \leqslant \tau \leqslant T\}$ of SDE (16) are statistically different from the flipped forward diffusion paths $\{X(T-t)\colon 0 \leqslant t \leqslant T\}$ of SDE (1)
- If at $\tau = 0$ we start $Y(0)$ with an ensemble of samples of pure noise, then at $\tau = T$ the evolved ensemble by SDE (16) contains samples of the target distribution.
- The capability of SDE (16) to sample the target distribution is valid for every $\gamma \geqslant 0$, including $\gamma = 0$, which gives a deterministic ODE instead of an SDE.
- The capability of SDE (16) to sample the target distribution depends on that we know the score function $\nabla \ln q(x,\tau) \equiv \nabla \ln p(x, T-\tau)$.

3 Learning the score function $\nabla \ln p(x,t)$ from data

The objective of sampling from the target distribution of $X(0)$ is achieved by using SDE (16) to evolve the pure noise density to the target density. In SDE (16), we need the score function $\nabla \ln q(x,\tau) = \nabla \ln p(x, T-\tau)$, which is not immediately available even if the distribution $X(0)$ is given. In applications, the underlying distribution of $X(0)$ is represented in a set of data points of $X(0)$. We need to construct (learn) the score function from the given data set.

1 Conceptual framework for learning $\nabla \ln p(x,t)$ at time t

Recall that SDE (12) is the microscopic time reversal of SDE (1). Sample paths obtained $\{X(t)\}$ from the forward diffusion in SDE (1) are also sample paths of the reverse diffusion in SDE (12).

In principle, the score function $\nabla \ln p(x,t)$ can be estimated from these sample paths.

$$\begin{aligned}
\text{SDE (12):} \quad & X(t-dt) = X(t) + \Big((\beta x) + \sigma^2 \nabla \ln p(x,t)\Big)\Big|_{X(t)} dt + \sigma\sqrt{dt}\, N(0,1) \\
\Longrightarrow \quad & E\Big(X(t-dt)\Big|X(t)=x\Big) = x + x\beta dt + \sigma^2 \nabla \ln p(x,t) dt + o(dt) \\
\Longrightarrow \quad & \nabla \ln p(x,t) = E\Big(\frac{X(t-dt) - X(t)(1+\beta dt)}{\sigma^2 dt}\Big|X(t)=x\Big) + o(1)
\end{aligned} \tag{17}$$

Although this will work in principle, there are two practical issues.

1) To estimate $\nabla \ln p(x,t)$ at x, we need a large number of sample paths $\{X(t)\}$ arriving at x at time t. These sample paths are needed in $E\big(\,\bullet\,\big|X(t)=x\big)$. While it is conceptually possible to obtain these sample paths, practically it is very difficult.

2) $\nabla \ln p(x,t)$ is a function of (x,t). How do we represent it in a computational form? especially when we work with a real data set with a finite number of points.

3) SDE (12) is for the limit of infinitesimal dt. At any finite dt, the finite difference version has a discretization error of $o(dt)$, leading to an error of $o(1)$ in the estimated $\nabla \ln p(x,t)$.

We first address issues 1) and 2) above. We rewrite (17) in a simpler form. Let $s(x)$ denote the unknown function we wish to determine. Let $(X(\omega), Y(\omega))$ be a pair of random variables satisfying i) the support of $X(\omega)$ is $(-\infty, +\infty)$, covering the support of the unknown function $s(x)$, and ii) the condition average of $Y(\omega)$ given $X(\omega) = x$ is $s(x)$.

$$s(x) = E\Big(Y\Big|X{=}x\Big) \tag{18}$$

Note that for any random variable U, its mean μ_U satisfies

$$\begin{aligned}
& E\Big((U-\lambda)^2\Big) = E\Big((U-\mu_U)^2\Big) + (\lambda - \mu_U)^2 \geqslant E\Big((U-\mu_U)^2\Big) \\
& \mu_U = \arg\min_{\lambda} E\Big((U-\lambda)^2\Big)
\end{aligned}$$

At any x, applying this result to the conditional average in (18) yields

$$E\Big((Y-\lambda(X))^2\Big|X{=}x\Big) \geqslant E\Big((Y-s(X))^2\Big|X{=}x\Big) \quad \text{for any function } \lambda(x)$$

Using the law of total expectation, we obtain

$$\begin{aligned}
E\Big((Y-\lambda(X))^2\Big) &= E\Big(E\big((Y-\lambda(X))^2\Big|X\big)\Big) \\
&\geqslant E\Big(E\big((Y-s(X))^2\Big|X\big)\Big) = E\Big((Y-s(X))^2\Big) \quad \text{for any function } \lambda(x)
\end{aligned}$$

It follows that

$$\boxed{s(x) = \arg\min_{\{\lambda(x)\}} E\Big((Y-\lambda(X))^2\Big)} \tag{19}$$

(19) is the key component in extracting function $s(x)$ from data of (X, Y) satisfying (18). It tells us that a function can be determined in a functional least squares problem.

In applications, random vector (X, Y) is represented by a data set of finite size:

$$\text{Data set of } (X,Y): \quad D = \left\{(X^{(j)}, Y^{(j)})\right\} \tag{20}$$

Accordingly, function $\lambda(x)$ in (19) is represented by a neural network

$$\text{Neural network representation: } \lambda(x) = \lambda(x;\, \theta)$$

Functional least squares problem (19) becomes

$$\boxed{\begin{cases} \theta^{(opt)} = \arg\min\limits_{\theta} \sum\limits_{j} \left(Y^{(j)} - \lambda(X^{(j)};\, \theta)\right)^2 \\ s(x) = \lambda(x;\, \theta^{(opt)}) \end{cases}} \tag{21}$$

2 Precise formulation for learning $\nabla \ln p(x,t)$ at time t

We now address issue 3): the discretization error for finite dt. Recall that SDE (1), as an Ornstein Uhlenbeck process, has an analytical solution. As a result, sample paths $\{X(t)\}$ in the forward diffusion can be generated exactly with no discretization error. With starting point $X(0)$, the forward diffusion is given in (2). We rewrite is as

$$X(t) = X(0)e^{-\beta t} + \sigma_t\, Z, \quad Z \sim N(0,1), \quad \sigma_t^2 \equiv \sigma^2 \frac{(1 - e^{-2\beta t})}{2\beta} \tag{22}$$

We use this analytical solution to derive an exact formulation for extracting the score function, which has no discretization error. In the forward diffusion, $X(0)$ is from the underlying distribution and noise Z is independent of $X(0)$. Together $X(0)$ and Z determine $X(t)$. To facilitate the discussion, We introduce two conditional probability densities. Let

- $p(x_t|x_0) \equiv p\big((x_t,t)|(x_0,0)\big)$ be the conditional density of $X(t)$ given $X(0) = x_0$;
- $p(x_0|x_t) \equiv p\big((x_0,0)|(x_t,t)\big)$ be the conditional density of $X(0)$ given $X(t) = x_t$.

The forward diffusion solution (22) gives

$$\begin{aligned} &\big(X(t)|X(0) = x_0\big) \sim N(x_0 e^{-\beta t}, \sigma_t^2), \qquad p(x_t|x_0) \propto \exp\Big(\frac{-(x_t - x_0 e^{-\beta t})^2}{2\sigma_t^2}\Big) \\ &\nabla_{x_t} \ln p(x_t|x_0) = \frac{\nabla_{x_t} p(x_t|x_0)}{p(x_t|x_0)} = \frac{-1}{\sigma_t^2}(x_t - x_0 e^{-\beta t}) \end{aligned} \tag{23}$$

The essence of reverse diffusion is denoising. Conditional on $X(t) = x_t$, the added Z that produces $X(t)$ in (22) is no longer Gaussian. We derive the conditional average of Z given $X(t) = x_t$.

$$E\Big(\sigma_t Z \Big| X_t = x_t\Big) = E\Big((X(t) - X(0)e^{-\beta t}) \Big| X(t) = x_t\Big) = \int (x_t - x_0 e^{-\beta t}) p(x_0|x_t) dx_0 \tag{24}$$

We use the expression of $\nabla_{x_t} p(x_t|x_0)$ in (23) to rewrite it as

$$E\Big(\sigma_t Z\Big|X_t = x_t\Big) = -\sigma_t^2 \int \nabla_{x_t} p(x_t|x_0) \frac{p(x_0|x_t)}{p(x_t|x_0)} dx_0 \tag{25}$$

We apply Bayes' theorem.

$$\underbrace{\frac{p(x_0|x_t)}{p(x_t|x_0)} = \frac{p(x_0)}{p(x_t)}}_{\text{Bayes' theorem}}, \qquad E\Big(\sigma_t Z\Big|X_t = x_t\Big) = -\sigma_t^2 \int \nabla_{x_t} p(x_t|x_0) \frac{p(x_0)}{p(x_t)} dx_0 \tag{26}$$

where $p(x_0)$ and $p(x_t)$ are short notations for $p(x_0) \equiv p(x_0, 0)$ and $p(x_t) \equiv p(x_t, t)$. Note that in (26), the integration variable is x_0. We move $p(x_t)$ and ∇_{x_t} out of the integral.

$$E\Big(\sigma_t Z\Big|X_t = x_t\Big) = -\sigma_t^2 \frac{1}{p(x_t)} \nabla_{x_t} \Big(\underbrace{\int p(x_t|x_0) p(x_0) dx_0}_{=p(x_t)} \Big)$$

$$= -\sigma_t^2 \frac{\nabla_{x_t} p(x_t)}{p(x_t)} = -\sigma_t^2 \nabla_{x_t} \ln p(x_t) = \boxed{-\sigma_t^2 \nabla_{x_t} \ln p(x_t, t)} \tag{27}$$

We use $X(t) = X(0)e^{-\beta t} + \sigma_t Z$, and we denote x_t simply as x.

$$\boxed{\begin{array}{c} \nabla \ln p(x,t) = E\Big(\dfrac{-Z}{\sigma_t}\Big|X(0)e^{-\beta t} + \sigma_t Z = x\Big), \\ \\ \sigma_t^2 \equiv \sigma^2 \dfrac{(1 - e^{-2\beta t})}{2\beta}, \qquad Z \sim N(0,1) \end{array}} \tag{28}$$

(28) is an improvement over (17) in two aspects: i) (28) is exact with no discretization error, and ii) (28) is valid exactly for any $t > 0$.

We follow the approach outlined in (19) and (21) to formulate the task of extracting $\nabla \ln p(x,t)$ as that of training a neural network. To apply the result of (19), we identify

$$X = X(0)e^{-\beta t} + \sigma_t Z, \qquad Y = \frac{-Z}{\sigma_t}$$

Based on (19), the functional least squares problem for $\nabla \ln p(x,t)$ is

$$\boxed{\nabla \ln p(x,t) = \underset{\{\lambda(x)\}}{\arg\min}\, E\Big(\Big[\frac{Z}{\sigma_t} + \lambda\big(X(0)e^{-\beta t} + \sigma_t Z\big)\Big]^2\Big), \quad Z \sim N(0,1)} \tag{29}$$

Random vector $\big(X(0), Z\big)$ is represented by the data set below.

$$\text{Data set:} \quad D = \big\{(X_0^{(i)}, Z^{(i,j)})\big\}, \quad Z^{(i,j)} \overset{\text{i.i.d.}}{\sim} N(0,1) \tag{30}$$

In the data set, for each sample $X_0^{(i)}$ of $X(0)$, there are many independent realizations $Z^{(i,j)}$ of Z, corresponding to independent realizations of $X(t)$: $X_t^{(i,j)} = X_0^{(i)} e^{-\beta t} + \sigma_t Z^{(i,j)}$.

$\nabla \ln p(x,t)$ at time t is represented by a trained neural network.

$$\boxed{\begin{cases} \theta^{(opt)} = \arg\min\limits_{\theta} \sum\limits_{i,j} \Big[\dfrac{Z^{(i,j)}}{\sigma_t} + \lambda\big(X_0^{(i)} e^{-\beta t} + \sigma_t Z^{(i,j)};\, \theta\big) \Big]^2, \\ \qquad\qquad Z^{(i,j)} \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \text{neural network: } \lambda(x;\,\theta) \\ \nabla \ln p(x,t) = \lambda(x;\, \theta^{(opt)}) \end{cases}} \tag{31}$$

3 Formulation for learning $\nabla \ln p(x,t)$ as a function of (x,t)

Let $\{X_0^{(i)}\}$ be a set of data points from the underlying distribution of $X(0)$. In applications, the underlying distribution is represented by data set $\{X_0^{(i)}\}$. The objective of diffusion models is to draw new samples from the underlying distribution.

In the time direction, we use a sequence of time instances $\{t_k\}$ to cover $[0,T]$. Random vector $\big(X(0), Z\big)$ used in $X(t) = X(0)e^{-\beta t} + \sigma_t Z,\ t \in [0,T]$ is represented by the data set below.

$$\text{Data set:} \quad D = \big\{(X_0^{(i)}, Z^{(i,j,t_k)})\big\}, \quad Z^{(i,j,t_k)} \overset{\text{i.i.d.}}{\sim} N(0,1) \tag{32}$$

In the data set, for each sample $X_0^{(i)}$ of $X(0)$, there are many independent realizations $Z^{(i,j,t_k)}$ of Z, corresponding to many independent realizations of $X(t_k)$ at many t_k.

$$X_{t_k}^{(i,j)} = X_0^{(i)} e^{-\beta t_k} + \sigma_{t_k} Z^{(i,j,t_k)}$$

To model functions of (x,t), we adopt a neural network of the form $\lambda(x,t;\,\theta)$. The score function $\nabla \ln p(x,t)$ as a function of (x,t) is represented by a trained neural network.

$$\boxed{\begin{cases} \theta^{(opt)} = \arg\min\limits_{\theta} \sum\limits_{i,j,t_k} \Big[\dfrac{Z^{(i,j,t_k)}}{\sigma_{t_k}} + \lambda\big(X_0^{(i)} e^{-\beta t_k} + \sigma_{t_k} Z^{(i,j,t_k)}, t_k;\, \theta\big) \Big]^2, \\ \qquad\qquad Z^{(i,j,t_k)} \overset{\text{i.i.d.}}{\sim} N(0,1), \quad \text{neural network: } \lambda(x,t;\,\theta) \\ \nabla \ln p(x,t) = \lambda(x,t;\, \theta^{(opt)}) \end{cases}} \tag{33}$$