Tutorial on Generative Model

Guanyu Chen

January 4, 2025

Abstract

Abstract

Contents

Reference: [6][1][5] [3][4][7][2]

1 Preliminaries

1.1 SDEs

1.1.1 SODEs

Problem 1 *Assume we have a Stochastic Differential Equation like:*

$$dX_t = f(X_t, t)dt + G(X_t, t)dW_t \tag{1}$$

where $X_t \in \mathbf{R}^d, f \in \mathcal{L}(\mathbf{R}^{d+1}, \mathbf{R}^d)$*, and* W_t *is m-dim Brownian Motion with diffusion matrix* Q*,* $G(X_t, t) \in \mathcal{L}(\mathbf{R}^{m+1}, \mathbf{R}^d)$*, with initial condition* $X_0 \sim p(X_0)$*.*

1.1.2 Fokker-Planck-Kolmogorov Equation

Definition 1 (Generator) *The infinitesimal generator of a stochastic process* $X(t)$ *for function* $\phi(x)$*, i.e.* $\phi(X_t)$ *can be defined as*

$$\mathcal{A}\phi(X_t) = \lim_{s\to 0^+} \frac{E[\phi(X(t+s)] - \phi(X(t))}{s} \tag{2}$$

Where ϕ *is a suitable regular function.*

This leads to Dynkin's Formula very naturally.

Theorem 1 (Dynkin's Formula)

$$E[f(X_t)] = f(X_0) + E\left[\int_0^t \mathcal{A}(f(X_s))ds\right] \tag{3}$$

Theorem 2 *If* $X(t)$ *s.t. 1, then the generator is given:*

$$\mathcal{A}(\cdot) = \sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t, t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right)\left[G(X_t, t)QG^\top(X_t, t)\right]_{ij} \tag{4}$$

Proof 1 *See P119 of SDE by Oksendal.*

Example 1 *If* $dX_t = dW_t$*, then* $\mathcal{A} = \frac{1}{2}\Delta$*, where* Δ *is the Laplace operator.*

Definition 2 (Generalized Generator) *For* $\phi(x, t)$*, i.e.* $\phi(X_t, t)$*, the generator can be defined as:*

$$A_t\phi(x, t) = \lim_{s\to 0^+} \frac{E[\phi(X(t+s), t+s)] - \phi(X(t), t)}{s} \tag{5}$$

Theorem 3 *Similarly if* $X(t)$ *s.t. 1, then the generalized generator is given:*

$$\mathcal{A}_t(\cdot) = \frac{\partial(\cdot)}{\partial t} + \sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t, t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right)\left[G(X_t, t)QG^\top(X_t, t)\right]_{ij} \tag{6}$$

We want to consider the density distribution of $X_t, P(x, t)$

Theorem 4 (Fokken-Planck-Kolmogorov equation) *The density function* $P(x, t)$ *of* X_t *s.t. 1 solves the PDE:*

$$\frac{\partial P(x, t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[f_i(x, t)p(x, t)\right] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^\top\right)_{ij} P(x, t)\right] \tag{7}$$

The PDE is called FPK equation / forward Kolmogorov equation.

Proof 2 *Consider the function* $\phi(x)$*, let* $x = X_t$ *and apply Ito's Formula:*

$$\begin{aligned} d\phi &= \sum_i \frac{\partial \phi}{\partial x_i} dx_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right) dx_i dx_j \\ &= \sum_i \frac{\partial \phi}{\partial x_i}\left(f_i\left(X_t, t\right)dt + \left(G\left(X_t, t\right)dW_t\right)\right) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right)\left[G(X_t, t)QG^\top(X_t, t)\right]_{ij} dt. \end{aligned} \tag{8}$$

Take expectation of both sides:

$$\frac{dE[\phi]}{dt} = \sum_i E\left[\frac{\partial \phi}{\partial x_i} f_i(X_t, t)\right] + \frac{1}{2}\sum_{ij} E\left[\frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij}\right] \tag{9}$$

So

$$\begin{cases} \dfrac{dE[\phi]}{dt} = \dfrac{d}{dt}\left[\displaystyle\int \phi(x)P(X_t = x, t)dx\right] = \displaystyle\int \phi(x)\frac{\partial P(x,t)}{\partial t}dx \\ \displaystyle\sum_i E\left[\frac{\partial \phi}{\partial x_i} f_i\right] = \sum_i \int \frac{\partial \phi}{\partial x_i} f_i(X_t = x, t) P dx = -\sum_i \int \phi \cdot \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right] dx. \\ \displaystyle\frac{1}{2}\sum_{ij} E\left[\frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij}\right] = \frac{1}{2}\sum_{ij}\int \frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij} P dx = \frac{1}{2}\sum_{ij}\int \phi(x)\frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij} P\right) dx. \end{cases} \tag{10}$$

then

$$\int \phi \frac{\partial P}{\partial t} dX = -\sum_i \int \phi \frac{\partial}{\partial x_i}\left(f_i P\right) dX + \frac{1}{2}\sum_{ij}\int \phi \frac{\partial^2}{\partial x_i x_j}\left(\left[GQG^\top\right]_{ij} P\right) dx$$

Hence

$$\int \phi \cdot \left[\frac{\partial P}{\partial t} + \sum_i \frac{\partial}{\partial x_i}\left(f_i P\right) - \frac{1}{2}\sum_{ij}\frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij} P\right)\right] dX = 0$$

Therefore P s.t.

$$\frac{\partial P}{\partial t} + \sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)P(x,t)\right) - \frac{1}{2}\sum_{i=1}\frac{\partial^2}{\partial X_i \partial X_j}\left(\left[GQG^\top\right]_{ij} P(x,t)\right) = 0 \tag{11}$$

Which gives the FPK Equation.

Remark 1 *When SDE is time independent:*

$$dX_t = f(X_t)dt + G(X_t)dW_t \tag{12}$$

then the solution of FPK often converges to a stationary solution s.t. $\frac{\partial P}{\partial t} = 0$.

Here is an another way to show FPK equation: Since we have inner product $\langle \phi, \psi \rangle = \int \phi(x)\psi(x)dx$. Then $E[\phi(x)] = \langle \phi, P \rangle$.

As the equation 9 can be written as

$$\frac{d}{dt}\langle \phi, P \rangle = \langle \mathcal{A}\phi, P \rangle \tag{13}$$

Where $\mathcal{A}$ has been mentioned above. If we note the adjoint operator of $\mathcal{A}$ as $\mathcal{A}^*$, then we have

$$\langle \phi, \frac{dP}{dt} - \mathcal{A}^*(P) \rangle = 0, \forall \phi(x) \tag{14}$$

Hence we have

Theorem 5 (FPK Equation)

$$\frac{dP}{dt} = \mathcal{A}^*(P), \text{where } \mathcal{A}^*(\cdot) = -\sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)(\cdot)\right) + \frac{1}{2}\sum_{i=1}\frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij}(\cdot)\right) \tag{15}$$

It can be rewritten as:

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\nabla \cdot \left[f(x,t)p(x,t)\right] + \frac{1}{2}\nabla^2 \cdot \left[\left(GQG^\top\right) p(x,t)\right] \\ &= -\nabla \cdot \left[f(x,t)p(x,t) - \frac{1}{2}\nabla \cdot \left[\left(GQG^\top\right) p(x,t)\right]\right] \end{aligned} \tag{16}$$

Theorem 6 (Transition Density(Forward Komogorov Equation)) *The transition density $P_{t|s}(x_t|x_s), t \geq s$, which means the propability of transition from $X(s) = x_s$ to $X(t) = x_t$, satisfies the FPK equation with initial condition $P_{s|s}(x|x_s) = \delta(x - x_s)$ i.e. for $P_{t|s}(x|y)$, it solves*

$$\frac{\partial P_{t|s}(x|y)}{\partial t} = \mathcal{A}^*(P_{t|s}(x|y)), \text{with } P_{s|s}(x|y) = \delta(x - y) \tag{17}$$

Theorem 7 (Backward Komogorov Equation) *$P_{s|t}(y|x)$ for $t \geq s$ solves:*

$$\frac{\partial P_{s|t}(y|x)}{\partial s} + \mathcal{A}(P_{s|t}(y|x)) = 0, \text{with } P_{s|t}(y|x) = \delta(x - y) \tag{18}$$

1.1.3 Mean and Covariance

After we derived the FPK equation, which is the complete probabilistic description of SDE, we can derive the mean and covariance of SDE. By takeing $\phi(x,t)$, then

$$\frac{dE[\phi]}{dt} = E\left[\frac{\partial\phi}{\partial t}\right] + \sum_i E\left[\frac{\partial\phi}{\partial x_i} f_i(X_t,t)\right] + \frac{1}{2}\sum_{ij} E\left[\frac{\partial^2\phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij}\right] \tag{19}$$

By taking $\phi(X,t) = x_i$ and $\phi(X,t) = x_i x_j - m(t)_i m(t)_j$, we have the mean function $m(t) = E[X_t]$ and covariance function $c(t) = E\left[(X_t - m(t))(X_t - m(t))^T\right]$ respectively, s.t.

$$\begin{cases} \dfrac{dm}{dt} = E\left[f(X_t,t)\right] \\ \dfrac{dc}{dt} = E\left[f(X,t)(X - m(t)^T)\right] + E\left[(X - m(t) f^T(X,t))\right] + E\left[G(X_t,t)QG^T(X_t,t)\right] \end{cases} \tag{20}$$

So we can estimate the mean and covariance of solution to SDE. However, these equations cannot be used as such, because only in the Gaussian case do the expectation and covariance actually characterize the distribution.

1.1.4 Linear SDEs

The linear SDE has explicit solution. Assume the linear SDe has the form

$$dX_t = (K(t)X_t + B(t))\,dt + G(t)dW_t \tag{21}$$

where $K(t) \in \mathbf{R}^{d\times d}, B(t) \in \mathbf{R}^d, G(t) \in \mathbf{R}^{d\times m}$ are given functions. $X_t \in \mathbf{R}^d$ is the state vector, $W_t \in \mathbf{R}^m$ is the Brownian Motion with diffusion matrix Q.

Theorem 8 *The explicit solution to the linear SDE is given by:*

$$X_t = \Psi(t,t_0)X_0 + \int_{t_0}^t \Psi(t,s)B(s)ds + \int_{t_0}^t \Psi(t,s)G(s)dW_s \tag{22}$$

where $\Psi(t,t_0)$ is the transition matrix of the linear SDE, which satisfies the following matrix ODE:

$$\frac{d\Psi}{dt} = K(t)\Psi(t,t_0), \Psi(t_0,t_0) = I \tag{23}$$

Hence, X_t is a Gaussian process(A linear transformation of Brownian Motion which is a Gaussian process).

Proof 3 *Multiply both sides of the SDE by Integrating factor $\Psi(t_0,t)$ and apply Ito's formula to $\Psi(t_0,t)X_t$. See Sarkka P49.*

As discussed above, we can compute the mean and covariance function of solution to linear SDE.

Theorem 9 *The mean and covariance function of solution to linear SDE are given by:*

$$\begin{cases} \dfrac{dm}{dt} = K(t)m(t) + B(t) \\ \dfrac{dc}{dt} = K(t)c(t) + c(t)K^T(t) + G(t)QG^T(t) \end{cases} \tag{24}$$

with initial condition $m_0 = m(t_0) = E[X_0], c_0 = c(t_0) = Cov(X_0)$. Then the solution is given by solving the above ODEs:

$$\begin{cases} m(t) = \Psi(t,t_0)m_0 + \displaystyle\int_{t_0}^t \Psi(t,s)B(s)ds \\ c(t) = \Psi(t,t_0)c_0\Psi^T(t,t_0) + \displaystyle\int_{t_0}^t \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{25}$$

Proof 4 *Apply $F(X,t) = K(t)X + B(t), G(X,t) = G(t)$ to 20.*

Hence the solution to linear SDE is a Gaussian process with mean and covariance function given by the above ODEs.

Theorem 10 *The solution to LSDE is Gaussian:*

$$p(X,t) = \mathcal{N}(X(t)|m(t), c(t)) \tag{26}$$

Specially when $X_0 = x_0$ is fixed, then

$$p(X,t|X_0 = x_0) = \mathcal{N}(X(t)|m(t|x_0), c(t|x_0)) \tag{27}$$

That is, $m_0 = x_0, c_0 = 0$. Then we have:

$$\begin{cases} m(t|x_0) = \Psi(t,t_0)x_0 + \displaystyle\int_{t_0}^{t} \Psi(t,s)B(s)ds \\ c(t|x_0) = \displaystyle\int_{t_0}^{t} \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{28}$$

Proof 5 *The proof is straight foward either by applying $m_0 = x_0, c_0 = 0$ to 25 or by eq 22.*

So, to sum up, linear SDE has great properties! The distribution is completedly decided by the inital condition. Also, if we generate X_0 to X_{t_k}, which means that we begin SDE at t_i with X_{t_i}, we have the equivalent discretization of SDE:

Theorem 11 *Original SDE is weakly, in distribution, equivalent to the following discrete-time SDE:*

$$X_{t_{i+1}} = A_i X_{t_i} + B_i + G_i \tag{29}$$

where

$$\begin{cases} A_i = \Psi(t_{i+1}, t_i) \\ B_i = \displaystyle\int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s)B(s)ds \\ G_i = \displaystyle\int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s)G(s)QG^T(s)\Psi^T(t_{i+1}, s)ds \end{cases} \tag{30}$$

Proof 6 *The proof is straight forward.*

Theorem 12 *The covariance of X_t and $X_s(s < t)$ is given by:*

$$Cov(X_t, X_s) = \Psi(t,s)c(s) \tag{31}$$

Proof 7 *See Sarkka P88-89.*

1.1.5 Numerical Methods

Euler-Maruyama and Milstein methods have been talked before.

1.1.6 Langevin SDE

The Langevin SDE has the following form:

$$X_{t+s} = X_t + \nabla \log p_t(x_t)s + \sqrt{2s}\xi \tag{32}$$

where $X_t \in \mathcal{R}^d, p_t(x_t) = p(X_t = x_t)$, $\xi \sim N(0, I)$, I is identical matrix of $m \times m$. Our goal is to sample from specific $p(x,t)$.

Theorem 13 *The density of Langevin Diffusion Model converges to $p(x)$ over time. In other words, if $X_t \sim p(x)$, then $X_{t+s} \sim p(x)$ for $\forall s > 0$.*

Proof 8 *Let $\mu_t(f) = E\left[f\left(X_t\right)\right]$. Consider $\mu_{t+\tau}(f) = E\left[f\left(X_{t+\tau}\right)\right]$, as $\tau \to 0$. Then*

$$\begin{aligned} \mu_{t+\tau} =& E\left[f\left(X_t + \nabla \log p_t\left(x_t\right) \cdot \tau + \sqrt{2\tau}\xi\right)\right] \\ =& E\Big[f\left(x_t\right) + \nabla^\top f\left(x_t\right)\left(\tau \nabla \log p_t\left(x_t\right) + \sqrt{2\tau}\xi\right) \\ &+ \frac{1}{2}\left(\nabla^\top \log p_t(x_t)\tau + \sqrt{2\tau}\xi\right)\nabla^2 f(x_t) \nabla \log p_t(x_t)\tau + \sqrt{2\tau}\xi\Big] \\ =& E\left[f\left(x_t\right)\right] + E\left[\tau \nabla^\top f\left(x_t\right) \nabla \log p_t\left(x_t\right)\right] \\ &+ \frac{\tau^2}{2} E\left[\nabla^\top \log p\left(x_t\right) \cdot \nabla^2 f\left(x_t\right) \cdot \nabla \log p\left(x_t\right)\right] + E\left[\tau \xi^\top \nabla^2 f\left(x_t\right)\xi\right] \end{aligned} \tag{33}$$

The second term:

$$\begin{aligned}
&\tau E\left[\nabla^{\top} f \nabla \log p_t\right] \\
=&\tau \int \nabla f \cdot \nabla \log p_t p_t dx = \tau \int \nabla f \cdot \nabla p_t dx \\
=&-\tau \int \operatorname{tr}\left(\nabla^2 f\right) \cdot p_t dx = -\tau E\left[\operatorname{tr}\left(\nabla^2 f\right)\right] \\
=&-\tau E\left[\xi^{\top} \nabla^2 f \xi\right]
\end{aligned} \tag{34}$$

Then

$$\mu_{t+\tau} = E\left[\frac{1}{2} \nabla^{\top} \log p_t \nabla^2 f \nabla \log p_t\right] \cdot \tau^2 = O\left(\tau^2\right) \tag{35}$$

Hence we have $\frac{d}{dt}(\mu_t) = 0$, *i.e.* $E[\mu_t] = E[\mu_{t+s}]$ *for* $\forall s > 0$.

Remark 2 *We define the density of normal distribution* $N(x;\mu,\Sigma)$, *and its log-density, gradient of density and score as follows:*

$$\begin{cases}
N(x;\mu,\Sigma) = \dfrac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)} \\
\log N(x;\mu,\Sigma) = -\dfrac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu) - \log\left(\sqrt{(2\pi)^d|\Sigma|}\right). \\
\nabla_x N(x;\mu,\Sigma) = N(x;\mu,\Sigma)\Sigma^{-1}(x-\mu) \\
\nabla_x \log N(x;\mu,\Sigma) = -\Sigma^{-1}(x-\mu).
\end{cases} \tag{36}$$

Actually, Langevin SDE is not necessary be as above i.e. the diffusion term is not necessary to be $\sqrt{2}$. The reason is to guarantee the stationary distribution of $p_t(x)$. i.e. the term $\frac{\partial p(x,t)}{\partial t} = 0$ in FPK equation. If the diffusion term is $g(t)$, then by FPK equation, we have

$$\nabla_x \cdot (fp - \frac{1}{2} g^2(t) \nabla p) = 0$$

then $f(x,t) = \frac{1}{2} g^2(t) \frac{\nabla_x p(x,t)}{p(x,t)} = \frac{1}{2} g^2(t) \nabla_x \log p(x,t)$.

1.2 Conservation Laws

1.2.1 Flow Map

Theorem 14 *Two important theorems in calculus:*

*1.**Divergence Theorem:***

$$\int_{\Omega} \nabla \cdot \mathbf{F} dx = \int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n} dS \tag{37}$$

*2.**Reynolds Transport Theorem:***

$$\frac{d}{dt} \int_{\Omega(t)} f(t,x) dx = \int_{\Omega(t)} \frac{\partial f}{\partial t} dx + \int_{\partial\Omega(t)} f(t,x) \mathbf{v} \cdot \mathbf{n} dS \tag{38}$$

where u *is the velocity at* $\partial\Omega(t)$.

Here the $\Omega(t)$ is the domain of the flow, and the $\partial\Omega(t)$ is the boundary of the flow, which is described by the flow map ϕ_s^t. Here is the definition.

Definition 3 (Flow Map) *Assume a description of some characteristic of particle* **P**, *like the position or the boundary, as* $\mathbf{x} \in \mathcal{R}^m$, *then we have a flow map* $\phi_s^t(\mathbf{x}) \in \mathcal{R}^m$, *which means that the flow transimits the characteristic(position)* $\mathbf{x}$ *from* $\mathbf{x}$ *at* s *to* $\phi_s^t(\mathbf{x})$ *at* t, *controlled by the vector field(velocity field)* $\mathbf{F} : \mathcal{R}^m \times \mathcal{R} \to \mathcal{R}^m$:

$$\begin{cases}
\dfrac{d\phi_s^t(\mathbf{x})}{dt} = \mathbf{F}(\phi_s^t(\mathbf{x}), t) \\
\phi_s^s(\mathbf{x}) = \mathbf{x}
\end{cases} \tag{39}$$

1.2.2 Conservation Laws

If we assume $\Omega(t)$ is composed of particles, i.e. $\Omega(t) = \phi_{t_0}^t(\Omega)$(when $t = t_0$, $\Omega(t_0) = \Omega$), then we by **conservation of mass**, we have the following theorem:

Theorem 15 (Continuity Equation) *By conservation of mass, i.e. $\int_{\Omega(t)} \rho(t, \mathbf{x})d\mathbf{x} = C$, we have:*

$$\begin{aligned} \frac{d}{dt}\int_{\Omega(t)} \rho(t,\mathbf{x})d\mathbf{x} &= \int_{\Omega(t)} \frac{\partial \rho}{\partial t}d\mathbf{x} + \int_{\partial\Omega(t)} \rho(t,\mathbf{x})\mathbf{u}\cdot\mathbf{n}dS \\ &= \int_{\Omega(t)} \left(\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\right)d\mathbf{x} = 0 \end{aligned} \tag{40}$$

Therefore:

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) = 0 \tag{41}$$

*which is also called **continuity equation**.*

Theorem 16 (Conservation of Momentum) *By conservation of momentum, i.e.*

$$\frac{d}{dt}\int_{\Omega(t)} \rho(t,\mathbf{x})\mathbf{v}(t,\mathbf{x})d\mathbf{x} = -\int_{\partial\Omega(t)} p\cdot\mathbf{n}dS \tag{42}$$

we have:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v} + p) = 0 \tag{43}$$

where p is the pressure.

Theorem 17 (Conservation of Energy)

$$\frac{\partial E}{\partial t} + \nabla\cdot(\mathbf{v}(E+p)) = 0 \tag{44}$$

Then we have can get the Euler's equation:

Theorem 18 (Euler's Equation) *The Euler's equation is given by:*

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho \\ \rho\mathbf{v} \\ E\end{bmatrix} + \nabla\cdot\begin{bmatrix}\rho\mathbf{v} \\ \rho\mathbf{v}\otimes\mathbf{v} + p \\ \mathbf{v}(E+p)\end{bmatrix} = 0 \tag{45}$$

So the general form of conservation laws is given by: suppose $U \in \mathcal{R}^d$ is the conserved quantity, F is $\mathcal{R}^d \to \mathcal{R}^d$ is the flux, then we have:

$$\frac{\partial U}{\partial t} + \nabla\cdot(F(U)) = 0 \tag{46}$$

2 What is Diffusion After All?

2.1 From SDEs

At the beginning, the diffusion phenomenon is observed through the motion of particles(Brownian motion). Normally, the SDE can be written as:

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \tag{47}$$

Here, we skip the drift term $f(X_t,t)$ and only consider the diffusion term $G(X_t,t)dW_t$, i.e.

$$dX_t = G(X_t,t)dW_t \tag{48}$$

Then by FPK equation, we can derive

Theorem 19 *The probability density function $p(x,t)$ satisfies:*

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left[\left(GQG^\top\right)_{ij}p(x,t)\right] = \frac{1}{2}\nabla\cdot\left(\nabla\cdot\left(GQG^T p(x,t)\right)\right) \tag{49}$$

Specially, when $G(X_t,t) = G(t)$ and $Q = I$, we have:

$$\frac{\partial p}{\partial t} = \nabla\cdot\left(\frac{GG^T}{2}\nabla p\right) \tag{50}$$

So, when $X_0 \sim p_0$, we can then compute the diffusion density $p(x,t)$ by solving the FPK equation.

2.2 From Flow Map

Since we have the definition of **Flow Map** $\phi_s^t(\mathbf{x})$, which is controlled by vector field $V(\phi_s^t(\mathbf{x}), t)$, then just think the $\phi_0^t(\mathbf{x})$ as the trajectory of the particle beginning at x over time, noted as $\phi_t(x)$. Then the vector field is actually the velocity field of the particle, so we have:

$$\begin{cases} \dfrac{\partial \phi_t(\mathbf{x})}{\partial t} = V(\phi_t(\mathbf{x}), t) \\ \phi_0(\mathbf{x}) = \mathbf{x} \end{cases} \tag{51}$$

The motion of particles described by ϕ_t determines how the density $p_t(x)$ evolves over time.

Theorem 20 *When the initial density $p_0(x)$ is known, the density field can be expressed as:*

$$p(\phi_t(x), t) = \frac{p_0(x)}{|\det J_{\phi_t}(x)|} \tag{52}$$

It should be noted that $\phi_t(x)$ is actually the same as X_t in SDE, then similarly, the density is:

$$\phi_t(x) \sim p_t(x) \tag{53}$$

So, the flow map is an ODE, which is a special case of SDE without diffusion term. Then we have:

Theorem 21 (Continuity Equation) *The probability density function $p(x,t)$ of X_t satisfies:*

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (V(x,t)p(x,t)) \tag{54}$$

which is called ***Continuity Equation****.*

Remark 3 *The continuity equation can also be derived from the Conservation of Mass.*

Theorem 22 *When the incompressible condition is satisfied, that is $\nabla \cdot V = 0$, then the flow $\phi_t(x)$ is* ***measure preserving****, that is:*

$$|\det J_{\phi_t}(x)| = 1, i.e. p(\phi_t(x), t) = p_0(x) \tag{55}$$

Definition 4 (Flux) *We find that $V(x,t)p(x,t)$ is actually the flux $\mathcal{F}(x,t)$ of the particle.*

Then the continuity equation can be rewritten as:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (\mathcal{F}(x,t)) \tag{56}$$

Then we find that if the flux s.t. $F = -\frac{1}{2}\nabla \cdot \left(GQG^T p(x,t)\right)$, then $p(x,t)$ describes the diffusion process. This is the famous Fick's Law.

Theorem 23 (Fick's Law) *Fick's Law describes the relationship between the flux $\mathcal{F}(x,t)$ of the particle and the concentration/density $p(x,t)$.:*

$$\mathcal{F}(x,t) = -\frac{1}{2}\nabla \cdot \left(GQG^T p(x,t)\right) \tag{57}$$

Specifically, when $G(X_t, t) = G(t)$ and $Q = I$, we have:

$$\mathcal{F}(x,t) = -\frac{GG^T}{2}\nabla p(x,t) \tag{58}$$

Then

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot \left(\frac{GG^T}{2}\nabla p(x,t)\right) \tag{59}$$

2.3 Solution

Note $-\frac{GG^T}{2}$ is actually the diffusion coefficient $\mathcal{D}$. Then we have the diffusion equation:

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot (\mathcal{D}\nabla p(x,t)) \tag{60}$$

with initial condition $p(x,0) = p_0(x)$. We can use the Fourier Transform to solve this equation.

Theorem 24 *The solution to the diffusion equation is:*

$$\begin{aligned} p(x,t) &= \mathscr{F}^{-1}\left[\tilde{p}_0(\lambda)\exp\left(-\lambda^T\mathcal{D}\lambda t\right)\right] = \left(p_0 \star \mathcal{G}_{2t\mathcal{D}}\right)(x) \\ &= \frac{1}{\sqrt{(4\pi t)^d \det(\mathcal{D})}}\int_{\mathcal{R}^d}\left(p_0(\xi)\exp\left(-\frac{1}{4t}(x-\xi)^T\mathcal{D}^{-1}(x-\xi)\right)\right)d\xi \end{aligned} \tag{61}$$

where $\tilde{p}_0(\lambda) = \mathscr{F}(p_0(x))$ is the Fourier Transform of $p_0(x)$. $\mathcal{G}_{2t\mathcal{D}}$ is the Gaussian Kernel with variance $2t\mathcal{D}$.

Proof 9 *First, assume the Fourier Transform of $p(x,t)$ is $\tilde{p}(x,t)$:*

$$\begin{cases} \tilde{p}(x,t) = \mathscr{F}\left[p(x,t)\right] = \int_{\mathcal{R}^d} p(x,t)e^{-i\lambda\cdot x}dx \\ p(x,t) = \mathscr{F}^{-1}\left[\tilde{p}(x,t)\right] = \frac{1}{(2\pi)^d}\int_{\mathcal{R}^d}\tilde{p}(x,t)e^{i\lambda\cdot x}dx \end{cases} \tag{62}$$

Then, we have:

$$\begin{cases} \mathscr{F}\left[\nabla\cdot\mathbf{v}\right] = i\lambda\cdot\mathscr{F}\left[\mathbf{v}\right] \\ \mathscr{F}\left[\mathcal{D}\nabla p\right] = i\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{cases} \tag{63}$$

Then,

$$\begin{aligned} \mathscr{F}\left[\frac{\partial p}{\partial t}\right] &= \frac{d}{dt}\mathscr{F}\left[p\right] = \mathscr{F}\left[\nabla\cdot(\mathcal{D}\nabla p)\right] \\ =& i\lambda\cdot\mathscr{F}\left[\mathcal{D}\nabla p\right] = -\lambda^T\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{aligned} \tag{64}$$

where $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_d)^T$.

Therefore, $\mathscr{F}\left[p\right] = \tilde{p}_0\exp\left(-\lambda^T\mathcal{D}\lambda t\right)$. Since $\mathscr{F}\left[N(x|0,2t\mathcal{D})\right] = \exp\left(-\lambda^T\mathcal{D}\lambda t\right)$, which gives the theorem.

Remark 4 *Specially, 1. When the initial density $p_0(x)$ is $\delta(x-x_0)$, the solution is:*

$$p(x,t) = \frac{1}{\sqrt{(4\pi t)^d \det\mathcal{D}}}\exp\left(-\frac{(x-x_0)^T\mathcal{D}^{-1}(x-x_0)}{4t}\right) \sim N(x_0, 2t\mathcal{D}) \tag{65}$$

2. When the initial density $p_0(x)$ is a Gaussian distribution $N(\mu,\Sigma)$, the solution is:

$$p(x,t) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma+2t\mathcal{D})}}\exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma+2t\mathcal{D})^{-1}(x-\mu)\right) \sim N(\mu,\Sigma+2t\mathcal{D}) \tag{66}$$

(The Fourier transform of (μ,Σ) is $\exp\left(-i\lambda^T\mu + \frac{1}{2}\lambda^T\Sigma\lambda\right)$.)

Till here, we can see the insight of diffusion. It is actually a process of smoothing the initial density by the Gaussian Kernel.

3 Variational Auto-Encoder

3.1 Structure

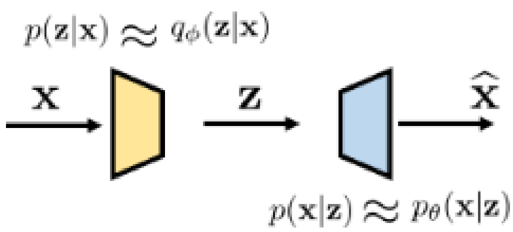

Figure 1: VAE block

where $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the proxy for $p(\mathbf{z} \mid \mathbf{x})$, which is also the distribution associated with the encoder. And $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ is the proxy for $p(\mathbf{x} \mid \mathbf{z})$, which is also the distribution associated with the decoder. Like the encoder, the decoder can be parameterized by a deep neural network.

If we treat ϕ and θ as optimization variables, then we need an objective function (or the loss function) so that we can optimize ϕ and θ through training samples.

3.2 Evidence Lower Bound

Definition 5 (Evidence Lower Bound) *The Evidence Lower Bound (ELBO) is defined as:*

$$\mathrm{ELBO}(x) = \mathbf{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] \tag{67}$$

Remark 5 *The ELBO is a lower bound of the log-likelihood of the data. It is used to estimate* $\log p(x)$*.*

$$\begin{aligned}\log p(x) &= \log \int p(x,z)dz \\ &= \log \int \frac{p(x,z)}{q_\phi(z|x)} \cdot q_\phi(z|x)dz \\ &\geq \mathbf{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] = \mathrm{ELBO}(x)\end{aligned} \tag{68}$$

Theorem 25 (Decomposition of Log-likelihood) *We have*

$$\log p(x) = \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z|x)||p(z|x)) \tag{69}$$

then we can minimize the gap between $\log p(x)$ *and ELBO, and the equality hold if and only if* $q_\phi(z|x) = p(z|x)$*. Since* $p(z|x)$ *is a delta function,*

Proof 10

$$\begin{aligned}\log p(x) &= \log p(x) \int q_\phi(z|x)dz \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log\left(\frac{p(x,z)}{p(z|x)}\frac{q_\phi(z|x)}{q_\phi(z|x)}\right)\right] \\ &= \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z|x)||p(z|x))\end{aligned} \tag{70}$$

Theorem 26 *Also, we can rewrite the ELBO as:*

$$\begin{aligned}\mathrm{ELBO}(x) &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x|z) + \log p(z) - \log q_\phi(z|x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z)) \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z))\end{aligned} \tag{71}$$

where the first term determines how good the decoder is, maxmizing the likelihood of observing the image, and the letter describes how good the encoder is, minimizing the distance between two distributions.

Definition 6 (The objective of VAE) *The optimiztion objective of VAE is to maxmize the ELBO:*

$$(\phi,\theta) = \mathrm{argmax}_{\phi,\theta} \sum_{x\in X} \mathrm{ELBO(x)} \tag{72}$$

where X *is the training set.*

4 Diffusion Model

4.1 DDPM

DDPM is like splitting the encoder and decoder of VAE into controllable parts. For each training data point $x_0 \sim p_{data}$, then a discrete Markov chain $\{x_1, \cdots, x_N\}$ is constructed by transition function:

$$p(x_i|x_{i-1}) = \mathcal{N}(x_i|\sqrt{1-\beta_i}x_{i-1}, \beta_i I) \tag{73}$$

Then we can get

$$p_{\alpha_i}(x_i|x_0) = \mathcal{N}(x_i|\sqrt{\alpha_i}x_0, (1-\alpha_i)I), \alpha_i = \Pi_{j=1}^i(1-\beta_j) \tag{74}$$

Hence we need to train the ELBO:

$$\mathrm{ELBO}(x) = \sum_{i=1}^N (1-\alpha_i)\mathbf{E}_{p_{data}(x)}\left[\mathbf{E}_{p_{\alpha_i}(\hat{x}|x)}\left[\|s_\theta(\hat{x},i) - \nabla_{\hat{x}}\log p_{\alpha_i}(\hat{x}|x)\|\right]\right] \tag{75}$$

Then do the reverse Markov chain.

This is clearly a discrete version. Then we consider the continuous version. We consider linear SDE having the form:

$$dX_t = (a(t)X_t + b(t))dt + g(t)dW_t \tag{76}$$

where $X_t \in \mathcal{R}^d, W_t \in \mathcal{R}^m$ with diffusion factor $Q \in \mathcal{R}^{m\times m}$, then $a(t) \in \mathcal{R}^{d\times d}, b(t) \in \mathcal{R}^d, g(t) \in \mathcal{R}^{d\times m}$. By Euler Maruyama method, it can be approximated By

$$\begin{aligned} X_{t+s} &= X_t + (a(t)X_t + b(t))s + g(t)\sqrt{sQ}\xi \\ &= (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{sQ}\xi \end{aligned} \tag{77}$$

where $\xi \sim N(0, I_m)$. Usually we need to consider the expectation, variance and distribution of X_t. But the stochastic value of X_t is dependent of x_0. Then first we consider

$$\begin{aligned} E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right] &\approx \left(a(t)E\left[X_t|X_t\right] + b(t)\right)s + g(t)\sqrt{sQ}E[\xi] \\ &= \left(a(t)E\left[X_t|X_0\right] + b(t)\right)s. \end{aligned} \tag{78}$$

Note $e(t) = E\left[X_t|X_0\right]$, then

$$e'(t) = \lim_{s\to 0} \frac{E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right]}{s} = a(t)\cdot e(t) + b(t). \quad e(0) = X_0. \tag{79}$$

which is an ODE system, having solution

$$e(t) = e^{\int_0^t a(s)ds}\cdot\left(X_0 + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right) \tag{80}$$

Therefore

$$\begin{aligned} E\left[X_t\right] &= E\left[E\left[X_t|X_0\right]\right] = E[e(t)] \\ &= e^{\int_0^t a(s)ds}\cdot\left(E\left[X_0\right] + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right) \end{aligned} \tag{81}$$

Similarly, Note $\operatorname{Var}\left(X_0|X_0\right) = v(t)$: then $\operatorname{Var}\left(X_{t+s}|X_0\right) = (1 + sa(t))^2\operatorname{Var}\left(X_t|X_0\right) + sgQg^\top$. Then

$$\begin{aligned} V'(t) &= \lim_{s\to 0}\frac{\operatorname{Var}\left(X_{t+s}|X_0\right) - \operatorname{Var}\left(X_t|X_0\right)}{s} \\ &= \left[\left(a^2(t)s + 2a(t)\right)v(t) + g^2(t)\right]|_{s\to 0} \\ &= 2\alpha(t)V(t) + g(t)Qg^\top(t), \qquad V(0) = 0 \end{aligned} \tag{82}$$

Solution is:

$$v(t) = e^{\int_0^t 2a(s)ds}\cdot\left(\int_0^t e^{-\int_0^s 2a(r)dr} g(s)Qg^\top(s)ds\right) \tag{83}$$

By law of total variance:

$$\begin{aligned} \operatorname{Var}\left(X_t\right) &= E\left[X_t^2\right] - E^2\left[X_t\right] = E\left[E\left[X_t^2|X_0\right]\right] - E^2\left[X_t\right] \\ &= E\left[\operatorname{Var}\left(X_t|X_0\right) + E^2\left[X_t|X_0\right]\right] - E^2\left[X_t\right] \\ &= E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + E\left[E^2\left[X_t|X_0\right]\right] - E^2\left[E\left[X_t|X_0\right]\right] \\ &= E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + \operatorname{Var}\left(E\left[X_t|X_0\right]\right) \end{aligned} \tag{84}$$

then

$$\begin{aligned} \operatorname{Var}(X_t) &= E[V(t)] + \operatorname{Var}(e(t)) \\ &= e^{\int_0^t 2a(s)ds}\cdot\left(\int_0^t e^{-\int_0^s 2a(r)dr} g(s)Qg^\top(s)ds\right) + e^{\int_0^t 2a(s)ds}\cdot\operatorname{Var}\left(X_0\right). \end{aligned} \tag{85}$$

We have the following theorem which is crucial for diffusion models. Usually, we assume $Q = I_m$.

Theorem 27 *If $X_{t+s} = (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{s}\xi$ then $X_t|X_0 \sim N\left(E\left[X_t|X_0\right], \operatorname{Var}\left(X_t|X_0\right)\right)$, where $E\left[X_t|X_0\right] = e(t), \operatorname{Var}\left(X_t|X_0\right) = V(t)$.*

It should be noted that $e(t)$ is related to X_0 and t, while $V(t)$ only depends on t!

Next, we will see how the above formula can be applied to diffusion modtels. There are three frameworks to build SDEs for diffusion models, VP, VE and sub-VP.

Definition 7 *Noise function $\beta(t)$. s.t. $\beta(0) = 0; \beta'(t) \geqslant 0; \beta(t) \to \infty$ as $t \to \infty$*

Variance Preserving (VP) SDE

So if we have diffusion model like:

$$\begin{aligned} X_{t_{i+1}} &= \sqrt{1-(\beta(t_{i+1})-\beta(t_i))}X_{t_i} + \sqrt{(\beta(t_{i+1})-\beta(t_i))}\xi \\ &= \sqrt{1-\Delta\beta(t_i)}X_{t_i} + \sqrt{\Delta\beta(t_i)}\xi \end{aligned} \tag{86}$$

Then the conditional distribution is given by:

$$q\left(X_{t_{i+1}}|X_{t_i}\right) = N(x_{t_{i+1}}; \sqrt{1-\Delta\beta(t_i)}X_{t_i}, \Delta\beta(t_i)) \tag{87}$$

Then we need to estimate θ drift term f and diffusion term g:

$$\begin{aligned} f(x,t) &= \lim_{h\to 0}\frac{E\left[X_{t+h}-X_t|X_t=x\right]}{h} \\ &= \lim_{h\to 0}\frac{x\sqrt{1-\Delta\beta(t)}-x}{h} = -\frac{x}{2}\beta'(t). \\ g(t) &= \sqrt{\lim_{h\to 0}\frac{N\left[X_{t+h}|X_t=x\right]}{h}} = \sqrt{\lim_{h\to 0}\frac{\beta(t+h)-\beta(t)}{h}} = \sqrt{\beta'(t)} \end{aligned} \tag{88}$$

Then the model can be written as $dx = -\frac{x}{2}\beta'(t)dt + \sqrt{\beta'(t)}dW_t$

Then by Theorem 27 we have

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{\int_0^t -\frac{1}{2}\beta'(s)ds} = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right]e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s \beta'(r)dr}\beta'(s)ds\cdot e^{-\beta(t)} = 1-e^{-\beta(t)} \\ V\left(X_t\right) = 1-e^{-\beta(t)} + V\left(X_0\right)e^{-\beta(t)} = 1+\left(V\left(X_0\right)-1\right)e^{-\beta(t)}. \end{cases} \tag{89}$$

So as $t\to\infty, \beta(t)\to\infty$, then $E\to 0, V\to 1$, i.e. $X_t|X_0 \sim N\left(E\left[X_t|X_0\right], \mathrm{Var}\left|X_t\right|X_0\right) \to N(0,1)$ as $t\to\infty$.

Variance-Exploding SDE

Here is the model: $X_{t+h} = X_t + \sqrt{\Delta\beta(t)}\xi$

Similarly we can compute the $f(x,t)\equiv 0$ and $g(t)=\sqrt{\beta(t)}$. Hence

$$\begin{cases} E\left[X_0|X_0\right] = X_0 \\ E\left[X_t\right] = E\left[X_0\right] \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s 0dr}\beta'(s)ds = \beta(t) \\ V\left(X_t\right) = V\left[X_0\right] + \beta(t) \end{cases} \tag{90}$$

So the expectation value is constant and the variance is increasing monotonical.
If we rescale X_t as $Y_t = \frac{X_t}{\sqrt{\beta(t)}}$, then $Y_t \to N(0,1), t\to\infty$.

Sub-VP SPE

Here, we set the dift and diffusion term as

$$\begin{aligned} f(x,t) &= -\frac{1}{2}\beta'(t) \\ g(t) &= \sqrt{\beta'(t)\left(1-e^{-2\beta(t)}\right)} \end{aligned} \tag{91}$$

As the same, we can compute that.

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right]e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \left(1-e^{-\beta(t)}\right)^2 \\ V\left(X_t\right) = \left(1-e^{-\beta(t)}\right)^2 + V\left(X_t\right)e^{-\beta(t)}. \end{cases} \tag{92}$$

We can find out that the variance is always smaller that of VP SDE.

Remark 6 *To sum up, finally we hope that X_t converges to a normal distribution by choosing different drift and diffusion functions. For generative model, the goal is to sample from a Data distribution p_{data}. We have known that if we set the initial distribution $p_0(x_0) = p(X_0 = x_0) \sim p_{data}$, then after $t = T$, the distribution of X_t is tend to be $N(0, 1)$ under certain conditions.*

So the idea is backward: if we sample from $X_T \sim N(0, 1)$, and then run SDE backwards, could we get the initial distribution?

Assume we have forward SDE: from $X_0 \sim p_0, X_T \sim p_T$,

$$dX_t = f(X_t, t)dt + G(t)dW_t \tag{93}$$

Then we define the reverse SDE as: from $X_T \sim p_T$,

$$d\bar{X}_t = \bar{f}(\bar{X}_t, t)dt + \bar{G}(t)d\bar{W}_t \tag{94}$$

where $\bar{W}_t$ is Brownian Motion runns backward in time, i.e. $\bar{W}_{t-s} - \bar{W}_t$ is independent of $\bar{W}_t$. We can approximate by EM:

$$\bar{X}_{t-s} - \bar{X}_t = -s\bar{f}(\bar{X}_t, t) + \sqrt{s}\bar{G}(t)\xi \tag{95}$$

So the problem is: If given f, G, are there $\bar{f}, \bar{G}$ s.t. the reverse time diffusion process $\bar{X}_t$ has the same distribution as the forward process X_t? Yes!

Theorem 28 *The reverse SDE with $\bar{f}, \bar{G}$ having the following form has the same distribution as the forward SDE 93:*

$$\begin{cases} \bar{f}(x, t) = f(x, t) - GG^T \nabla_x \log p_t(x) \\ \bar{G} = G(t) \end{cases} \tag{96}$$

i.e.

$$d\bar{X}_t = \left[f(\bar{X}_t, t) - GG^T \nabla_x \log p_t(x_t)\right] dt + G(t)d\bar{W}_t \tag{97}$$

Proof 11 *The proof is skipped.*

This theroem allows us to learn how to generate samples from p_{data}.

Algorithm 1 *:*
Step1. Select $f(x, t)$ and $g(t)$ with affine drift coefficients s.t. $X_T \sim N(0, 1)$
Step2. Train a network $s_\theta(x, t) = \frac{\partial}{\partial x} \log p_t(x)$ where $p_t(x) = p(X_t = x)$ is the forward distribution.
Step3. Sample X_T from $N(0, 1)$, then run reverse SDE from T to 0:

$$\bar{X}_{t-s} = \bar{X}_t + s\left[g^2(t)s_\theta(\bar{X}_t, t) - f(\bar{X}_t, t)\right] + \sqrt{s}g(t)\xi \tag{98}$$

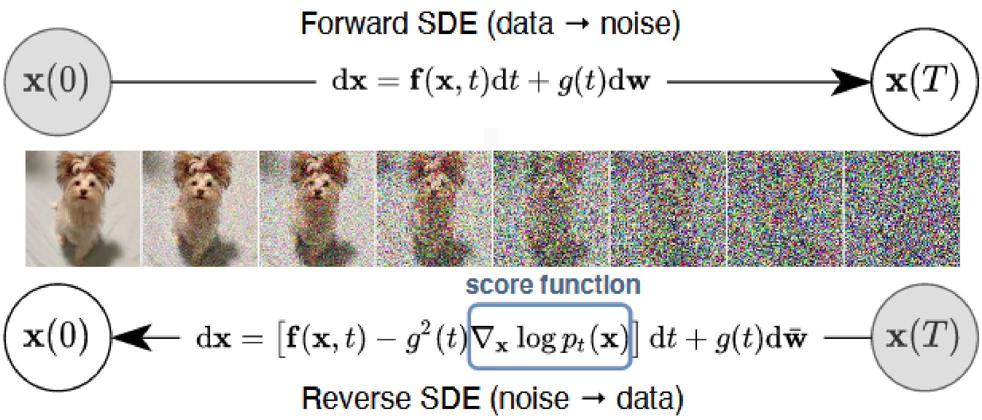

Figure 2: score-based generative model

4.2 Score Matching

The most difficult question on how to obtian $\nabla_x \log p(x)$ because it solves FPK equation.

4.2.1 Explicit Score Matching

Suppose we have a set of samples $x_1, x_2, \cdots, x_n$ from the data distribution $p_{data}(x)$. A classical way is to consider the kernel density estimation $q(x)$ of $p(x)$:

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{99}$$

where $K(x)$ is the kernel function. Since $q(x)$ is an approximation to p_{data}. We can define a loss function to train a network:

$$\begin{aligned}\mathcal{L}_\theta =&\mathbf{E}_{x\sim p(x)}\left[\|s_\theta(x) - \nabla_x \log p(x)\|^2\right]\\ \approx&\mathbf{E}_{x\sim q(x)}\left[\|s_\theta(x) - \nabla_x \log q(x)\|^2\right]\\ =&\int \|s_\theta(x) - \nabla_x \log q(x)\|^2 q(x)dx\\ \approx&\frac{1}{n}\sum_{i=1}^{n}\int \|s_\theta(x) - \nabla_x \log q(x)\|^2 K(x - x_i)dx\end{aligned} \tag{100}$$

However, when the number of samples is limited, the estimation $\nabla_x \log q(x)$ is not accurate.

4.2.2 Implicit Score Matching

4.2.3 Denoising Score Matching

Normally we can define the loss function as follows:

$$\begin{aligned}L_\theta &= \frac{1}{T}\int_0^T \lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right]dt\\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right]\right]\end{aligned} \tag{101}$$

It should be clearified that $p_{t|0}(x_t|x_0) = p(X_t = x_t|X_0 = x_0)$. So

$$p_t(x_t) = \int p_{t|0}(x_t|x_0)p_0(x_0)dx_0 = E_{x_0\sim p_{data}}\left[p_{t|0}(x_t|x_0)\right]$$

where $p_t(x) = p\left(X_t = x\right)$, $p_{t|0}(x|y) = p\left(X_t = x|X_0 = y\right)$. Then

$$\begin{aligned}\nabla \log p_t(x) &= \frac{1}{p_t(x)}\nabla p_t(x).\\ &= \frac{1}{p_t(x)}\nabla\int p_{t|0}(x|y)p_0(y)dy\\ &= \frac{1}{p_t(x)}\int \nabla p_{t|0}(x|y)p_0(y)dy\\ &= \frac{1}{p_t(x)}\int \frac{\nabla p_{t|0}(x|y)}{p_{t|0}(x|y)}p_0(y)\cdot p_{t|0}(x|y)dy\\ &= \int \nabla_x \log\left(p_{t|0}(x|y)\right)\cdot p_{0|t}(y|x)dy\\ &= \underset{y\sim p_{0|t}(y|x)}{E}\left[\nabla_x \log\left(p_{t|0}(x|y)\right)\right]\end{aligned} \tag{102}$$

Where we have used the following lemma:

Lemma 1

$$\underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[\underset{x_0'\sim p_{0|t}(\cdot|x_t)}{E}[f(x_t,x_0')]\right]\right] = \underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}[f(x_t,x_0)]\right] \tag{103}$$

Proof 12 *Easy to prove.*

Then we can rewrite the loss function as:

$$\begin{aligned} L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log p_t\left(x_t\right)\right\|^2\right]\right]\right] \\ &\leqslant \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\underset{y\sim p_{data}}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\left(p_{t|0}(x_t|y)\right)\right\|^2\right]\right]\right]\right] \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\left(p_{t|0}\left(x_t|x_0\right)\right\|^2\right]\right]\right] \end{aligned} \tag{104}$$

Since $p_{t|0}\left(x_t|x_0\right)=p\left(X_t=x_t|X_0=x_0\right)$ has been discussed:

$$p_{t|0}\left(x_t|x_0\right)\sim N\left(x_t;E\left[X_t=x_t|X_0=x_0\right],\operatorname{Var}\left(X_t=x_t|X_0=x_0\right)\right).$$

Then by theorem 27, x can be written as $x=e(t,X_0)+\sqrt{V(t)}\xi$, where $\xi\sim N(0,1)$, then the score function is:

$$\frac{\partial}{\partial x}\log p_{t|0}\left(x|x_0\right)=-\frac{x-E_{t|0}\left[x|x_0\right]}{\operatorname{Var}_{t|0}\left(x|x_0\right)}=-\frac{x-e(t,X_0)}{V(t)}\sim -N\left(0,\frac{1}{V(t)}\right) \tag{105}$$

So

$$\begin{aligned} L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{\xi\sim N(0,1)}{E}\left[\left\|s_\theta\left(\sqrt{V(t)}\xi+e(t,X_0),t\right)+\frac{\xi}{\sqrt{V(t)}}\right\|^2\right]\right]\right] \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\frac{1}{V(t)}\underset{\xi\sim N(0,1)}{E}\left[\left\|\xi_\theta\left(\sqrt{V(t)}\xi+e(t,X_0),t\right)-\xi\right\|^2\right]\right]\right] \end{aligned} \tag{106}$$

where $\xi_\theta=-\sqrt{V(t)}s_\theta$ is called denoising network.

4.3 Denoising Diffusion

4.3.1 With Classifier Guidance

Though we can produce pictures by sampling from normal distribution, we still cannot control what we will generate. What we want to do is something like: "Give me the pictures of number 6", then the model can sample from the normal distribution and do the denoising to generate pics of 6.

Usually, we can do something like: train a model for every class label. This do make the model smaller, but increases number of models. Think about it, when the label is TEXT, it is impossiable to train a model for each sentences.

So, the initial distribution is $p_0(x|y)$ given the label y. Similarly, we will convert the data distribution $p_{data}(x|y)$ to final distribution, normal distribution expected. Then we SDE becomes: $X_t\sim p_t(x|y)$

$$\begin{aligned} &p_t(x\mid y)=p\left(X_t=x\mid y\right)=\frac{p\left(y\mid X_t=x\right)p\left(X_t=x\right)}{p(y)} \\ \Rightarrow&\log\left(p_t(x\mid y)\right)=\log\left(p\left(y\mid X_t=x\right)\right)+\log\left(p\left(X_t=x\right)\right)-\log(p(y)) \\ \Rightarrow&\nabla_x\log\left(p_t(x\mid y)\right)=\nabla_x\log\left(p\left(y\mid X_t=x\right)\right)+\nabla_x\log\left(p\left(X_t=x\right)\right) \end{aligned} \tag{107}$$

We have finished training $\nabla_x\log\left(p\left(X_t=x\right)\right)$ in sampling. Then we need to estimate $\nabla_x\log\left(p\left(y\mid X_t=x\right)\right)$. This is the conditional protability, we end up with a sharp factor s: $p'\left(y\mid X_t=x\right)$, then:

$$\nabla_x\log\left(p_t(x\mid y)\right)=S\nabla_x\log\left(p\left(y\mid x_t=x\right)\right)+\nabla_x\log\left(p\left(x_t=x\right)\right) \tag{108}$$

Note $\omega_\theta(y\mid x,t)$ to learn $s\nabla_x\log\left(p\left(y\mid X_t=x\right)\right)$

4.3.2 Classifier Guidance Free

$$\begin{aligned} &\gamma\nabla_x\log\left(p\left(y\mid X_t=x\right)\right) \\ =&\gamma\left(\nabla_x\log\left(p(X_t=x|y)\right)-\nabla_x\log\left(p_t(x)\right)\right) \end{aligned} \tag{109}$$

Then

$$\begin{aligned} &\nabla_x\log_\gamma\left(p_t(x\mid y)\right) \\ =&(1-\gamma)\nabla_x\log\left(p_t(x)\right)+\gamma\nabla_x\log\left(p(X_t=x|y)\right) \end{aligned} \tag{110}$$

Hence we only need one conditional denoising network, and using null condition to represent the unconditional model.

5 Flow Matching

5.1 FPK Equation

We have discussed the FPK Equation in 'learnsde'.

Theorem 29 (Fokken-Planck-Kolmogorov equation) *The density function $p(x,t)$ of X_t s.t.*

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \tag{111}$$

solves the PDE:

$$\frac{\partial p(x,t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^\top\right)_{ij} p(x,t)\right] \tag{112}$$

The PDE is called FPK equation / forward Kolmogorov equation.

It can be rewritten as:

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\nabla \cdot [f(x,t)p(x,t)] + \frac{1}{2}\nabla^2 \cdot \left[\left(GQG^\top\right)p(x,t)\right] \\ &= -\nabla \cdot \left[f(x,t)p(x,t) - \frac{1}{2}\nabla \cdot \left[\left(GQG^\top\right)p(x,t)\right]\right]\end{aligned} \tag{113}$$

Here, if we only consider $G(X_t,t) = g(t)$, then we notice that $M = GQG^T$ is independent of X_t, so we can write:

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\nabla \cdot \left(fp - \frac{1}{2}\nabla \cdot (Mp)\right) \\ &= -\nabla \cdot \left(fp - \frac{1}{2}M\nabla p\right) \\ &= -\nabla \cdot \left[\left(f - \frac{1}{2}M\frac{\nabla p}{p}\right)p\right] \\ &= -\nabla \cdot \left[\left(f - \frac{1}{2}M\nabla \log p\right)p\right]\end{aligned} \tag{114}$$

So we find out that if we have an ODE s.t. $dZ_t = F(Z_t,t)dt$ with $Z_0 \sim p_0$, instead of a SDE, then by FPK equation, the density $p(z,t)$ satisfies:

$$\frac{\partial p(z,t)}{\partial t} = -\nabla \cdot (F(z,t)p(z,t)) \tag{115}$$

So if we set $F(z,t) = f(z,t) - \frac{1}{2}M(t)\nabla \log p(z,t)$, then $p(z,t)$ is exactly like the density $p(x,t)$ of X_t in SDE. So theoretically, we can do the diffusion like reverse ode!

This is the topic discussed in 'Probability Flow'.

5.2 Probability Flow

Define a flow $\phi : \mathcal{R}^d \times [0,1] \to \mathcal{R}^d$ is a flow generated by a vector field $v : \mathcal{R}^d \times [0,1] \to \mathcal{R}^d$ i.e.

$$\begin{cases}\dfrac{\partial \phi(x,t)}{\partial t} = v(\phi(x,t),t) \\ \phi(0,x) = x\end{cases} \tag{116}$$

The flow means that under the vector field v, if the initial point is x, then the flow push the point after time t to $\phi(x,t)$. That is ϕ gives the evolution trajectory of x under the vector field v. So normally, we can consider the flow $\phi(x,t)$ as X_t in SDE:

$$dX_t = v(X_t,t)dt + 0dW_t \tag{117}$$

with $X_0 = x$. It turns out that it is actually an ODE, a SDE without diffusion term. Similar to SDE, if $X_0 = x \sim p_0(x)$, we have the probability density $p(x,t)$ satisfies FPK equation:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (v(x,t)p(x,t)) \tag{118}$$

which is a special case of FPK equation, called **Continuity Equation**. So, typically, we can solution to an ODE is a flow.

So the objective of Flow Matching Model can be described as: Let $p_t(x)$ be the density with initial $p_0(x)$, which is designed to be a simple distribution, like normal distribution. So let $p_1(x)$ be the approximation equal in distribution top_{data}. Then we need to design a flow to match the flow s.t. p_1 can properly approximate p_{data}.

References

[1] Stanley H. Chan. Tutorial on diffusion models for imaging and vision, 2024.

[2] Peter Holderrieth. The fokker-planck equation and diffusion models,, 2023. https://www.peterholderrieth.com/blog/2023/The-Fokker-Planck-Equation-and-Diffusion-Models/.

[3] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024.

[4] Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, Cambridge, 2014.

[5] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition, January 2014.

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.

[7] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2019.