Stochastic Differential Equations

Guanyu Chen

February 23, 2025

Abstract

This paper is about the basic theory and applications of SDEs. Also, this is a sum up of last term's SDE course. Reference: [1][2][3] [4][5][6][7] [8] [9]

Contents

1 Introduction to SDEs

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* $X_0 \sim p(X_0)$*.*

1.2 The It'so and Stratonovich Stochastic Integrals

1.3 Ito's Formula

1.4 Mean and Covariance

We can derive the mean and covariance of SDE. By applying Ito's formula to $\phi(x,t)$, then

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

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.

$$\begin{cases} \dfrac{dm}{dt} = E\left[f(X_t,t)\right] \\ \dfrac{dc}{dt} = E\left[f(X,t)(X - m(t)^T)\right] + E\left[(X - m(t)f^T(X,t))\right] + E\left[G(X_t,t)QG^T(X_t,t)\right] \end{cases} \tag{3}$$

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.

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

$$dX_t = \left(K(t)X_t + B(t)\right)dt + G(t)dW_t \tag{4}$$

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

$$X_t = \Psi(t,t_0)X_0 + \int_{t_0}^t \Psi(t,s)B(s)ds + \int_{t_0}^t \Psi(t,s)G(s)dW_s \tag{5}$$

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

$$\frac{d\Psi}{dt} = K(t)\Psi(t,t_0), \Psi(t_0,t_0) = I \tag{6}$$

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

Proof 1 *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 2 *The mean and covariance function of solution to linear SDE are given by:*

$$\begin{cases} \dfrac{dm}{dt} = K(t)m(t) + B(t) \\ \dfrac{dc}{dt} = K(t)c(t) + c(t)K^T(t) + G(t)QG^T(t) \end{cases} \tag{7}$$

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:*

$$\begin{cases} m(t) = \Psi(t,t_0)m_0 + \displaystyle\int_{t_0}^t \Psi(t,s)B(s)ds \\ c(t) = \Psi(t,t_0)c_0\Psi^T(t,t_0) + \displaystyle\int_{t_0}^t \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{8}$$

Proof 2 *Apply* $F(X,t) = K(t)X + B(t), G(X,t) = G(t)$ *to* 3.

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

Theorem 3 *The solution to LSDE is Gaussian:*

$$p(X,t) = \mathcal{N}(X(t)|m(t), c(t)) \tag{9}$$

Specially when $X_0 = x_0$ *is fixed, then*

$$p(X,t|X_0 = x_0) = \mathcal{N}(X(t)|m(t|x_0), c(t|x_0)) \tag{10}$$

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

$$\begin{cases} m(t|x_0) = \Psi(t,t_0)x_0 + \int_{t_0}^{t} \Psi(t,s)B(s)ds \\ c(t|x_0) = \int_{t_0}^{t} \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{11}$$

Proof 3 *The proof is straight foward either by applying* $m_0 = x_0, c_0 = 0$ *to* 8 *or by eq* 5.

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

$$X_{t_{i+1}} = A_i X_{t_i} + B_i + G_i \tag{12}$$

where

$$\begin{cases} A_i = \Psi(t_{i+1}, t_i) \\ B_i = \int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s)B(s)ds \\ G_i = \int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s)G(s)QG^T(s)\Psi^T(t_{i+1}, s)ds \end{cases} \tag{13}$$

Proof 4 *The proof is straight forward.*

Theorem 5 *The covariance of* X_t *and* $X_s (s < t)$ *is given by:*

$$Cov(X_t, X_s) = \Psi(t,s)c(s) \tag{14}$$

Proof 5 *See Sarkka P88-89.*

2 Fokker-Planck-Kolmogorov Equation

2.1 FPK 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*

$$\mathcal{A}\phi(X_t) = \lim_{s \to 0^+} \frac{E[\phi(X(t+s)] - \phi(X(t))}{s} \tag{15}$$

Where ϕ *is a suitable regular function.*

This leads to Dynkin's Formula very naturally.

Theorem 6 (Dynkin's Formula)

$$E[f(X_t)] = f(X_0) + E\left[\int_0^t \mathcal{A}(f(X_s))ds\right] \tag{16}$$

Theorem 7 *If $X(t)$ s.t. 1, then the generator is given:*

$$\mathcal{A}(\cdot)=\sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t,t)+\frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right)\left[G(X_t,t)QG^\top(X_t,t)\right]_{ij} \tag{17}$$

Proof 6 *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:*

$$A_t\phi(x,t)=\lim_{s\to 0^+}\frac{E[\phi(X(t+s),t+s)]-\phi(X(t),t)}{s} \tag{18}$$

Theorem 8 *Similarly if $X(t)$ s.t. 1, then the generalized generator is given:*

$$\mathcal{A}_t(\cdot)=\frac{\partial(\cdot)}{\partial t}+\sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t,t)+\frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right)\left[G(X_t,t)QG^\top(X_t,t)\right]_{ij} \tag{19}$$

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

Theorem 9 (Fokken-Planck-Kolmogorov equation) *The density function $P(x,t)$ of X_t s.t. 1 solves the PDE:*

$$\frac{\partial P(x,t)}{\partial t}=-\sum_i \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right]+\frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left[\left(GQG^\top\right)_{ij}P(x,t)\right] \tag{20}$$

The PDE is called FPK equation / forwand Kolmogorov equation.

Proof 7 *Consider the function $\phi(x)$, let $x=X_t$ and apply Ito's Formula:*

$$\begin{aligned} d\phi &= \sum_i \frac{\partial \phi}{\partial x_i} dx_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2\phi}{\partial x_i\partial x_j}\right) dx_i dx_j \\ &= \sum_i \frac{\partial \phi}{\partial x_i}\left(f_i\left(X_t,t\right)dt+\left(G\left(X_t,t\right)dW_t\right)\right)+\frac{1}{2}\sum_{i,j}\left(\frac{\partial^2\phi}{\partial x_i\partial x_j}\right)\left[G(X_t,t)QG^\top(X_t,t)\right]_{ij}dt. \end{aligned} \tag{21}$$

Take expectation of both sides:

$$\frac{dE[\phi]}{dt}=\sum_i E\left[\frac{\partial\phi}{\partial x_i}f_i(X_t,t)\right]+\frac{1}{2}\sum_{ij}E\left[\frac{\partial^2\phi}{\partial x_i\partial x_j}\left[GQG^\top\right]_{ij}\right] \tag{22}$$

So

$$\begin{cases} \frac{dE[\phi]}{dt}=\frac{d}{dt}\left[\int \phi(x)P(X_t=x,t)dx\right]=\int\phi(x)\frac{\partial P(x,t)}{\partial t}dx \\ \sum_i E\left[\frac{\partial\phi}{\partial x_i}f_i\right]=\sum_i\int\frac{\partial\phi}{\partial x_i}f_i(X_t=x,t)Pdx=-\sum_i\int\phi\cdot\frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right]dx. \\ \frac{1}{2}\sum_{ij}E\left[\frac{\partial^2\phi}{\partial x_i\partial x_j}\left[GQG^\top\right]_{ij}\right]=\frac{1}{2}\sum_{ij}\int\frac{\partial^2\phi}{\partial x_i\partial x_j}\left[GQG^\top\right]_{ij}Pdx=\frac{1}{2}\sum_{ij}\int\phi(x)\frac{\partial^2}{\partial x_i\partial x_j}\left(\left[GQG^\top\right]_{ij}P\right)dx. \end{cases} \tag{23}$$

then

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

Hence

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

Therefore P s.t.

$$\frac{\partial P}{\partial t}+\sum_i\frac{\partial}{\partial x_i}\left(f_i(x,t)P(x,t)\right)-\frac{1}{2}\sum_{i=1}\frac{\partial^2}{\partial X_i\partial X_j}\left(\left[GQG^\top\right]_{ij}P\left(x,t\right)\right)=0 \tag{24}$$

Which gives the FPK Equation.

Remark 1 *When SDE is time independent:*

$$dX_t = f(X_t)dt + G(X_t)dW_t \tag{25}$$

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 22 can be written as

$$\frac{d}{dt}\langle \phi, P \rangle = \langle \mathcal{A}\phi, P \rangle \tag{26}$$

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

$$\langle \phi, \frac{dP}{dt} - \mathcal{A}^*(P) \rangle = 0, \forall \phi(x) \tag{27}$$

Hence we have

Theorem 10 (FPK Equation)

$$\frac{dP}{dt} = \mathcal{A}^*(P), \text{where}\, \mathcal{A}^*(\cdot) = -\sum_i \frac{\partial}{\partial x_i}(f_i(x,t)(\cdot)) + \frac{1}{2}\sum_{i=1} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij}(\cdot)\right) \tag{28}$$

It can be rewritten as:

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\nabla \cdot [f(x,t)P(x,t)] + \frac{1}{2}\nabla^2 \cdot \left[\left(GQG^\top\right) P(x,t)\right] \\ &= -\nabla \cdot \left[f(x,t)P(x,t) - \frac{1}{2}\nabla \cdot \left[\left(GQG^\top\right) P(x,t)\right]\right] \end{aligned} \tag{29}$$

We define the probability flux to be:

$$J(x,t) = f(x,t)p(x,t) - \frac{1}{2}\nabla \cdot [M(x)p(x,t)], M(x) = G(x,t)Q(x,t)G(x,t)^T \tag{30}$$

Integrating the Fokker-Planck equation over $\mathbb{R}^d$ *and using the divergence theorem on the right hand side of the equation, we have:*

$$\frac{d}{dt}\int_{R^d} p(x,t)dx = \int_{R^d} \nabla \cdot J(x,t)dx = 0 \tag{31}$$

The stationary Fokker-Planck equation, whose solutions give us the invariant distributions of the diffusion process X_t*, can be written in the form*

$$\nabla \cdot J(x,t) = 0 \tag{32}$$

Consequently, the equilibrium probability flux is a divergence-free vector field.

2.2 Forward and backward Komogorov Equation

Theorem 11 *Fix* $t > s$*, let* $u(x,s) := E[g(X_t)|X_s = x] = \int g(y)P(y,t|x,s)dy$*, then* $u(x,s)$ *satisfies the following equation:*

$$\frac{\partial u}{\partial s} + f(x,s) \cdot \nabla u + \frac{1}{2}\nabla \cdot (M\nabla u) = 0, \qquad u(x,s) = g(x) \tag{33}$$

Theorem 12 (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*

$$\frac{\partial P_{t|s}(x|y)}{\partial t} = \mathcal{A}^*(P_{t|s}(x|y)), \text{with}\, P_{s|s}(x|y) = \delta(x-y) \tag{34}$$

2.3 Ornstein-Uhlenbeck Process

Definition 3 (Ornstein-Uhlenbeck Process) *The Ornstein-Uhlenbeck Process is defined as:*

$$dX_t = -\alpha X_t dt + \sqrt{2D}dW_t \tag{35}$$

where $\alpha > 0, D > 0$*, normally* $D = \frac{1}{\beta}$.

By FPK equation, we have:

$$\begin{cases} \frac{\partial p}{\partial t} = \alpha \frac{\partial}{\partial x}(xp) + D\frac{\partial^2 p}{\partial x^2} \\ p_0(x|x_0) = \delta(x - x_0) \end{cases} \tag{36}$$

When (36) is used to model the velocity or position of a particle, the noisy term on the right hand side of the equation is related to thermal fluctuations. The solution of (36) can be computed:

$$X_t \sim N(x_0 e^{-\alpha t}, \frac{D}{\alpha}(1 - e^{-2\alpha t})) \tag{37}$$

The generator of OU process is:

$$\mathcal{L} = -\alpha x \cdot \nabla + D\Delta \tag{38}$$

We need to study the properties of the generator $\mathcal{L}$. When the unique invariant density of OU is ρ, do transformation:

$$\mathcal{L}^*(h\rho) = \rho \mathcal{L} h \tag{39}$$

The IVP for FPK equation:

$$\frac{\partial p}{\partial t} = \mathcal{L}^* p, \qquad p(x, 0) = p_0(x) \tag{40}$$

becomes:

$$\frac{\partial h}{\partial t} = \mathcal{L} h, \qquad h(x, 0) = \rho^{-1} p_0(x) \tag{41}$$

Theorem 13 *Consider the eigenpairs problem for the generator operator $\mathcal{L}$ of OU process:*

$$\begin{cases} \lambda_n = \alpha n \\ \phi_n(x) = \frac{1}{n!} H_n(\sqrt{\alpha\beta} x) \end{cases} \qquad n = 0, \cdots, \infty \tag{42}$$

where $H_n(x)$ is the n-th Hermite polynomial:

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n}(e^{-x^2/2}) \tag{43}$$

2.4 Langevin SDE

The Langevin SDE has the following form:

$$X_{t+s} = X_t + \nabla \log p_t(x_t) s + \sqrt{2s}\xi \tag{44}$$

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

$$\begin{aligned} \mu_{t+\tau} =& E\left[f\left(X_t + \nabla \log p_t(x_t) \cdot \tau + \sqrt{2\tau}\xi\right)\right] \\ =& E\Big[f(x_t) + \nabla^\top f(x_t)\left(\tau \nabla \log p_t(x_t) + \sqrt{2\tau}\xi\right) \\ &+ \frac{1}{2}\left(\nabla^\top \log p_t(x_t)\tau + \sqrt{2\tau}\xi\right)\nabla^2 f(x_t) \nabla \log p_t(x_t)\tau + \sqrt{2\tau}\xi\Big] \\ =& E[f(x_t)] + E\left[\tau \nabla^\top f(x_t) \nabla \log p_t(x_t)\right] \\ &+ \frac{\tau^2}{2} E\left[\nabla^\top \log p(x_t) \cdot \nabla^2 f(x_t) \cdot \nabla \log p(x_t)\right] + E\left[\tau \xi^\top \nabla^2 f(x_t)\xi\right] \end{aligned} \tag{45}$$

The second term:

$$\begin{aligned} &\tau E\left[\nabla^\top f \nabla \log p_t\right] \\ =& \tau \int \nabla f \cdot \nabla \log p_t p_t dx = \tau \int \nabla f \cdot \nabla p_t dx \\ =& -\tau \int \operatorname{tr}\left(\nabla^2 f\right) \cdot p_t dx = -\tau E\left[\operatorname{tr}\left(\nabla^2 f\right)\right] \\ =& -\tau E\left[\xi^\top \nabla^2 f \xi\right] \end{aligned} \tag{46}$$

Then

$$\mu_{t+\tau} = E\left[\frac{1}{2}\nabla^\top \log p_t \nabla^2 f \nabla \log p_t\right] \cdot \tau^2 = O\left(\tau^2\right) \tag{47}$$

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:*

$$\begin{cases} N(x;\mu,\Sigma) = \dfrac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)} \\ \log N(x;\mu,\Sigma) = -\dfrac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) - \log\left(\sqrt{(2\pi)^d|\Sigma|}\right). \\ \nabla_x N(x;\mu,\Sigma) = N(x;\mu,\Sigma)\Sigma^{-1}(x-\mu) \\ \nabla_x \log N(x;\mu,\Sigma) = -\Sigma^{-1}(x-\mu). \end{cases} \tag{48}$$

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

3 What is Diffusion After All?

3.1 From SDEs

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

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \tag{49}$$

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

$$dX_t = G(X_t,t)dW_t \tag{50}$$

Then by FPK equation, we can derive

Theorem 15 *The probability density function* $p(x,t)$ *satisfies:*

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^\top\right)_{ij} p(x,t)\right] = \frac{1}{2}\nabla\cdot\left(\nabla\cdot\left(GQG^T p(x,t)\right)\right) \tag{51}$$

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

$$\frac{\partial p}{\partial t} = \nabla\cdot\left(\frac{GG^T}{2}\nabla p\right) \tag{52}$$

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

3.2 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:

$$\begin{cases} \dfrac{\partial \phi_t(\mathbf{x})}{\partial t} = V(\phi_t(\mathbf{x}),t) \\ \phi_0(\mathbf{x}) = \mathbf{x} \end{cases} \tag{53}$$

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

Theorem 16 *When the initial density $p_0(x)$ is known, the density field can be expressed as:*

$$p(\phi_t(x), t) = \frac{p_0(x)}{|\det J_{\phi_t}(x)|} \tag{54}$$

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

$$\phi_t(x) \sim p_t(x) \tag{55}$$

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

Theorem 17 (Continuity Equation) *The probability density function $p(x,t)$ of X_t satisfies:*

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

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

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

Theorem 18 *When the incompressible condition is satisfied, that is $\nabla \cdot V = 0$, then the flow $\phi_t(x)$ is* ***measure preserving****, that is:*

$$|\det J_{\phi_t}(x)| = 1, i.e. p(\phi_t(x), t) = p_0(x) \tag{57}$$

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:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (\mathcal{F}(x,t)) \tag{58}$$

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

$$\mathcal{F}(x,t) = -\frac{1}{2}\nabla \cdot \left(GQG^T p(x,t)\right) \tag{59}$$

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

$$\mathcal{F}(x,t) = -\frac{GG^T}{2}\nabla p(x,t) \tag{60}$$

Then

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot \left(\frac{GG^T}{2}\nabla p(x,t)\right) \tag{61}$$

3.3 Solution

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

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot (\mathcal{D}\nabla p(x,t)) \tag{62}$$

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

Theorem 20 *The solution to the diffusion equation is:*

$$\begin{aligned} p(x,t) &= \mathscr{F}^{-1}\left[\tilde{p}_0(\lambda)\exp\left(-\lambda^T \mathcal{D}\lambda t\right)\right] = (p_0 \star \mathcal{G}_{2t\mathcal{D}})(x) \\ &= \frac{1}{\sqrt{(4\pi t)^d \det(\mathcal{D})}} \int_{\mathcal{R}^d} \left(p_0(\xi)\exp\left(-\frac{1}{4t}(x-\xi)^T \mathcal{D}^{-1}(x-\xi)\right)\right) d\xi \end{aligned} \tag{63}$$

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 $\tilde{p}(x,t)$:*

$$\begin{cases} \tilde{p}(x,t) = \mathscr{F}\left[p(x,t)\right] = \int_{\mathcal{R}^d} p(x,t)e^{-i\lambda\cdot x}dx \\ p(x,t) = \mathscr{F}^{-1}\left[\tilde{p}(x,t)\right] = \frac{1}{(2\pi)^d}\int_{\mathcal{R}^d} \tilde{p}(x,t)e^{i\lambda\cdot x}dx \end{cases} \tag{64}$$

Then, we have:

$$\begin{cases} \mathscr{F}\left[\nabla\cdot\mathbf{v}\right] = i\lambda\cdot\mathscr{F}\left[\mathbf{v}\right] \\ \mathscr{F}\left[\mathcal{D}\nabla p\right] = i\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{cases} \tag{65}$$

Then,

$$\begin{aligned} \mathscr{F}\left[\frac{\partial p}{\partial t}\right] &= \frac{d}{dt}\mathscr{F}\left[p\right] = \mathscr{F}\left[\nabla\cdot(\mathcal{D}\nabla p)\right] \\ =& i\lambda\cdot\mathscr{F}\left[\mathcal{D}\nabla p\right] = -\lambda^T\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{aligned} \tag{66}$$

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

Therefore, $\mathscr{F}[p] = \tilde{p_0}\exp\left(-\lambda^T\mathcal{D}\lambda t\right)$. Since $\mathscr{F}[N(x|0, 2t\mathcal{D})] = \exp\left(-\lambda^T\mathcal{D}\lambda t\right)$, which gives the theorem.

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

$$p(x,t) = \frac{1}{\sqrt{(4\pi t)^d\det\mathcal{D}}}\exp\left(-\frac{(x-x_0)^T\mathcal{D}^{-1}(x-x_0)}{4t}\right) \sim N(x_0, 2t\mathcal{D}) \tag{67}$$

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

$$p(x,t) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma+2t\mathcal{D})}}\exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma+2t\mathcal{D})^{-1}(x-\mu)\right) \sim N(\mu, \Sigma+2t\mathcal{D}) \tag{68}$$

(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.

4 Reversible Diffusions

4.1 Definition

Definition 5 (Time-reversible) *A stationary stochastic process X_t is called time-reversible if for every $T \in (0,+\infty)$, the process X_{T-t} has the same distribution as X_t.*

Theorem 21 *A stationary Markov process X_t in $\mathbb{R}^d$ with generator $\mathcal{L}$ and invariant measure μ is time-reversible if and only if $\mathcal{L}$ is self-adjoint in $L^2(\mathbb{R}^d;\mu)$.*

Since for general SDE (1), the generator operator $\mathcal{L}$ and its self adjoint operator $\mathcal{L}^*$ are given by:

$$\begin{cases} \mathcal{L}(\cdot) = \sum_i \frac{\partial(\cdot)}{\partial x_i}f_i(x,t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i\partial x_j}\right)\left[GQG^\top\right]_{ij} \\ \qquad = f\cdot\nabla(\cdot) + \frac{1}{2}\left(M:\nabla\cdot\nabla\right)(\cdot) \\ \mathcal{L}^*(\cdot) = -\sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)(\cdot)\right) + \frac{1}{2}\sum_{i=1}\frac{\partial^2}{\partial x_i\partial x_j}\left(\left[GQG^\top\right]_{ij}(\cdot)\right) \\ \qquad = \nabla\cdot\left(-f(\cdot) + \frac{1}{2}\nabla\cdot\left(M(\cdot)\right)\right) \end{cases} \tag{69}$$

We assume that the diffusion process has a unique invariant distribution which is the solution of the stationary Fokker-Planck equation:

$$\mathcal{L}^*\rho_s = 0 \tag{70}$$

Notice that we can write the invariant distribution ρ_s in the form:

$$\rho_s = e^{-\Phi} \tag{71}$$

where Φ is a potential function.

Theorem 22 *For stationary process X_t with invariant distribution ρ_s. To guarantee the operator $\mathcal{L}$ is symmetric if and only if $J(\rho_s) = 0$. This is the detailed balance condition. So, expand the stationary probability flux:*

$$f = \frac{1}{2}\rho_s^{-1}\nabla \cdot (M\rho_s) \tag{72}$$

Consider now an arbitrary ergodic diffusion process X_t, the solution of (1) with invariant distribution ρ_s. We can decompose this process into a reversible and a nonreversible part in the sense that the generator can be decomposed into a symmetric and antisymmetric part.

Theorem 23 *The generator of an arbitrary diffusion process in $\mathbb{R}^d$ can we written as:*

$$\mathcal{L} = \rho_s^{-1} J_s \cdot \nabla + \frac{1}{2}\rho_s^{-1}\nabla \cdot (M\rho_s\nabla) := \mathcal{S} + \mathcal{A} \tag{73}$$

where $\mathcal{S}$ is the symmetric part of $\mathcal{L}$ and $\mathcal{A}$ is the antisymmetric part of $\mathcal{L}$.

Example 2 *When $M = 2I$, then*

$$f = \rho^{-1}\nabla\rho = \nabla \log \rho \tag{74}$$

which is the form of Langevin equation.

4.2 Schrödinger Operator

...

5 Generative Model

5.1 Diffusion Model

5.1.1 Diffusion

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:

$$p(x_i|x_{i-1}) = \mathcal{N}(x_i|\sqrt{1-\beta_i}x_{i-1}, \beta_i I) \tag{75}$$

Then we can get

$$p_{\alpha_i}(x_i|x_0) = \mathcal{N}(x_i|\sqrt{\alpha_i}x_0, (1-\alpha_i)I), \alpha_i = \Pi_{j=1}^{i}(1-\beta_j) \tag{76}$$

So, when $\alpha_i \to 0$, $p_{\alpha_i}(x_i|x_0)$ is close to $N(0, I)$. For generating new data samples, DDPMs start by first generating an unstructured noise vector from the prior distribution (which is typically trivial to obtain), then gradually remove noise therein by running a learnable Markov chain in the reverse time direction. The reverse is to learn transition kernel $p_\theta(x_{i-1}|x_i)$ having form:

$$p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}|\mu_\theta(x_i, i), \Sigma_\theta(x_i, i)) \tag{77}$$

where θ denotes model parameters. Key to the success of this sampling process is training the reverse Markov chain to match the actual time reversal of the forward Markov chain. That is, $p_\theta(x_0, x_1, \cdots, x_N)p(x_N)\Pi_{i=1}^{N}p_\theta(x_{i-1}|x_i)$ should be close to $q(x_0, x_1, \cdots, x_N) = p(x_0)\Pi_{i=1}^{N}p_{\alpha_i}(x_i|x_{i-1})$. This is achieved by minimizing the KL divergence between these two:

$$\begin{aligned}
&KL(q(x_0, x_1, \cdots, x_N)||p_\theta(x_0, x_1, \cdots, x_N)) \\
=& -\mathbf{E}_q\left[\log p_\theta(x_0, x_1, \cdots, x_N)\right] + const \\
=& -\mathbf{E}_q\left[\log p(x_N) + \sum_{i=1}^{N}\log\frac{p_\theta(x_{i-1}|x_i)}{q(x_i|x_{i-1})}\right]
\end{aligned} \tag{78}$$

This is clearly a discrete version. Then we consider the continuous version.

Score SDEs We consider linear SDE having the form:

$$dX_t = (a(t)X_t + b(t))dt + g(t)dW_t \tag{79}$$

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

$$\begin{aligned} X_{t+s} &= X_t + (a(t)X_t + b(t))s + g(t)\sqrt{sQ}\xi \\ &= (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{sQ}\xi \end{aligned} \tag{80}$$

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

$$\begin{aligned} E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right] \approx \left(a(t)E\left[X_t|X_t\right] + b(t)\right)s + g(t)\sqrt{sQ}E[\xi] \\ = \left(a(t)E\left[X_t|X_0\right] + b(t)\right)s. \end{aligned} \tag{81}$$

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

$$e'(t) = \lim_{s\to 0}\frac{E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right]}{s} = a(t)\cdot e(t) + b(t). \quad e(0) = X_0. \tag{82}$$

which is an ODE system, having solution

$$e(t) = e^{\int_0^t a(s)ds}\cdot\left(X_0 + \int_0^t e^{-\int_0^s a(r)dr}b(s)ds\right) \tag{83}$$

Therefore

$$\begin{aligned} E\left[X_t\right] &= E\left[E\left[X_t|X_0\right]\right] = E[e(t)] \\ &= e^{\int_0^t a(s)ds}\cdot\left(E\left[X_0\right] + \int_0^t e^{-\int_0^s a(r)dr}b(s)ds\right) \end{aligned} \tag{84}$$

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

$$\begin{aligned} V'(t) &= \lim_{s\to 0}\frac{\operatorname{Var}\left(X_{t+s}|X_0\right) - \operatorname{Var}\left(X_t|X_0\right)}{s} \\ &= \left[\left(a^2(t)s + 2a(t)\right)v(t) + g^2(t)\right]|_{s\to 0} \\ &= 2\alpha(t)V(t) + g(t)Qg^\top(t), \qquad V(0) = 0 \end{aligned} \tag{85}$$

Solution is:

$$v(t) = e^{\int_0^t 2a(s)ds}\cdot\left(\int_0^t e^{-\int_0^s 2a(r)dr}g(s)Qg^\top(s)ds\right) \tag{86}$$

By law of total variance:

$$\begin{aligned} \operatorname{Var}\left(X_t\right) =& E\left[X_t^2\right] - E^2\left[X_t\right] = E\left[E\left[X_t^2|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right) + E^2\left[X_t|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + E\left[E^2\left[X_t|X_0\right]\right] - E^2\left[E\left[X_t|X_0\right]\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + \operatorname{Var}\left(E\left[X_t|X_0\right]\right) \end{aligned} \tag{87}$$

then

$$\begin{aligned} \operatorname{Var}(X_t) =& E[V(t)] + \operatorname{Var}(e(t)) \\ =& e^{\int_0^t 2a(s)ds}\cdot\left(\int_0^t e^{-\int_0^s 2a(r)dr}g(s)Qg^\top(s)ds\right) + e^{\int_0^t 2a(s)ds}\cdot\operatorname{Var}\left(X_0\right). \end{aligned} \tag{88}$$

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

Theorem 24 *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 6 *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:

$$\begin{aligned} X_{t_{i+1}} &= \sqrt{1-(\beta(t_{i+1})-\beta(t_i))}X_{t_i} + \sqrt{(\beta(t_{i+1})-\beta(t_i))}\xi \\ &= \sqrt{1-\Delta\beta(t_i)}X_{t_i} + \sqrt{\Delta\beta(t_i)}\xi \end{aligned} \tag{89}$$

Then the conditional distribution is given by:

$$q\left(X_{t_{i+1}}|X_{t_i}\right) = N(x_{t_{i+1}}; \sqrt{1-\Delta\beta(t_i)}X_{t_i}, \Delta\beta(t_i)) \tag{90}$$

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

$$\begin{aligned} f(x,t) &= \lim_{h\to 0}\frac{E\left[X_{t+h}-X_t|X_t=x\right]}{h} \\ &= \lim_{h\to 0}\frac{x\sqrt{1-\Delta\beta(t)}-x}{h} = -\frac{x}{2}\beta'(t). \\ g(t) &= \sqrt{\lim_{h\to 0}\frac{N\left[X_{t+h}|X_t=x\right]}{h}} = \sqrt{\lim_{h\to 0}\frac{\beta(t+h)-\beta(t)}{h}} = \sqrt{\beta'(t)} \end{aligned} \tag{91}$$

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

Then we have

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{\int_0^t -\frac{1}{2}\beta'(s)ds} = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right]e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s \beta'(r)dr}\beta'(s)ds\cdot e^{-\beta(t)} = 1-e^{-\beta(t)} \\ V\left(X_t\right) = 1-e^{-\beta(t)} + V\left(X_0\right)e^{-\beta(t)} = 1+\left(V\left(X_0\right)-1\right)e^{-\beta(t)}. \end{cases} \tag{92}$$

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

$$\begin{cases} E\left[X_0|X_0\right] = X_0 \\ E\left[X_t\right] = E\left[X_0\right] \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s 0dr}\beta'(s)ds = \beta(t) \\ V\left(X_t\right) = V\left[X_0\right] + \beta(t) \end{cases} \tag{93}$$

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 SDE Here, we set the dift and diffusion term as

$$\begin{aligned} f(x,t) &= -\frac{1}{2}\beta'(t) \\ g(t) &= \sqrt{\beta'(t)\left(1-e^{-2\beta(t)}\right)} \end{aligned} \tag{94}$$

As the same, we can compute that.

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right]e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \left(1-e^{-\beta(t)}\right)^2 \\ V\left(X_t\right) = \left(1-e^{-\beta(t)}\right)^2 + V\left(X_t\right)e^{-\beta(t)}. \end{cases} \tag{95}$$

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

Remark 5 *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 $X_0 \sim p_0, X_T \sim p_T$,

$$dX_t = f(X_t, t)dt + G(t)dW_t \tag{96}$$

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

$$d\bar{X}_t = \bar{f}(\bar{X}_t, t)dt + \bar{G}(t)d\bar{W}_t \tag{97}$$

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:

$$\bar{X}_{t-s} - \bar{X}_t = -s\bar{f}(\bar{X}_t, t) + \sqrt{s}\bar{G}(t)\xi \tag{98}$$

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 25 *The reverse SDE with $\bar{f}, \bar{G}$ having the following form has the same distribution as the forward SDE 96:*

$$\begin{cases} \bar{f}(x,t) = f(x,t) - GG^T \nabla_x \log p_t(x) \\ \bar{G} = G(t) \end{cases} \tag{99}$$

i.e.

$$d\bar{X}_t = \left[f(\bar{X}_t, t) - GG^T \nabla_x \log p_t(x_t)\right] dt + G(t)d\bar{W}_t \tag{100}$$

Proof 10 *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:

$$\bar{X}_{t-s} = \bar{X}_t + s\left[g^2(t)s_\theta(\bar{X}_t, t) - f(\bar{X}_t, t)\right] + \sqrt{s}g(t)\xi \tag{101}$$

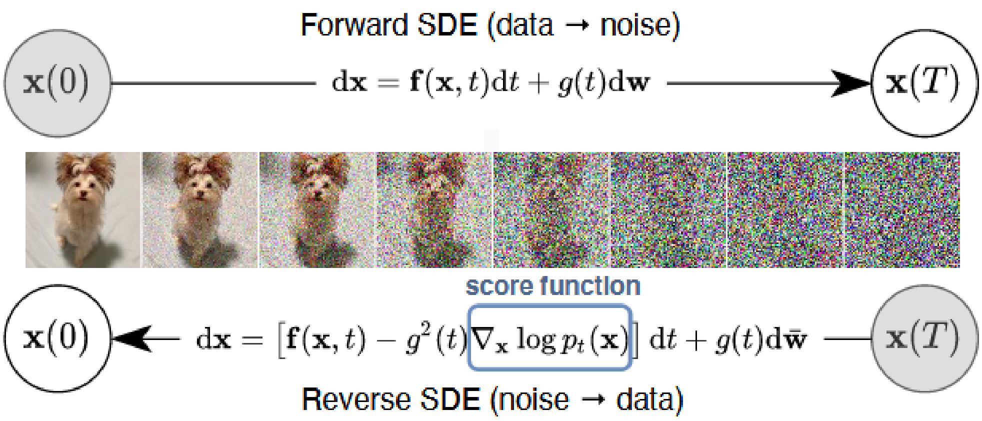

Figure 1: score-based generative model

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

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 $p(x)$:

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{102}$$

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:

$$\begin{aligned}\mathcal{L}_\theta =&\mathbf{E}_{x\sim p(x)}\left[\|s_\theta(x)-\nabla_x \log p(x)\|^2\right]\\ \approx&\mathbf{E}_{x\sim q(x)}\left[\|s_\theta(x)-\nabla_x \log q(x)\|^2\right]\\ =&\int \|s_\theta(x)-\nabla_x \log q(x)\|^2 q(x)dx\\ \approx&\frac{1}{n}\sum_{i=1}^{n}\int \|s_\theta(x)-\nabla_x \log q(x)\|^2 K(x-x_i)dx\end{aligned} \tag{103}$$

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

Implicit Score Matching

Denoising Score Matching Normally we can define the loss function as follows:

$$\begin{aligned}L_\theta &= \frac{1}{T}\int_0^T \lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t)-\nabla_{x_t}\log p_t(x_t)\|^2\right]\right]dt\\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t)-\nabla_{x_t}\log p_t(x_t)\|^2\right]\right]\right]\end{aligned} \tag{104}$$

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

$$\begin{aligned}\nabla \log p_t(x) &= \frac{1}{p_t(x)}\nabla p_t(x).\\ &= \frac{1}{p_t(x)}\nabla\int p_{t|0}(x|y)p_0(y)dy\\ &= \frac{1}{p_t(x)}\int \nabla p_{t|0}(x|y)p_0(y)dy\\ &= \frac{1}{p_t(x)}\int \frac{\nabla p_{t|0}(x|y)}{p_{t|0}(x|y)}p_0(y)\cdot p_{t|0}(x|y)dy\\ &= \int \nabla_x \log\left(p_{t|0}(x|y)\right)\cdot p_{0|t}(y|x)dy\\ &= \underset{y\sim p_{0|t}(y|x)}{E}\left[\nabla_x\log\left(p_{t|0}(x|y)\right)\right]\end{aligned} \tag{105}$$

Where we have used the following lemma:

Lemma 1

$$\underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[\underset{x_0'\sim p_{0|t}(\cdot|x_t)}{E}[f(x_t,x_0')]\right]\right] = \underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}[f(x_t,x_0)]\right] \tag{106}$$

Proof 11 *Easy to prove.*

Then we can rewrite the loss function as:

$$\begin{aligned}L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log p_t\left(x_t\right)\right\|^2\right]\right.\right.\\ &\leqslant \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\underset{y\sim p_{data}}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\left(p_{t|0}(x_t|y)\right)\right\|^2\right]\right]\right]\right]\\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\left(p_{t|0}\left(x_t|x_0\right)\right\|^2\right]\right]\right]\end{aligned} \tag{107}$$

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 24, 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:

$$\frac{\partial}{\partial x} \log p_{t|0}(x|x_0) = -\frac{x - E_{t|0}[x|x_0]}{\mathrm{Var}_{t|0}(x|x_0)} = -\frac{x - e(t, X_0)}{V(t)} \sim -N\left(0, \frac{1}{V(t)}\right) \tag{108}$$

So

$$\begin{aligned} L_\theta &= \underset{t\sim U(0,T)}{E}\left[\lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{\xi\sim N(0,1)}{E}\left[\left\| s_\theta\left(\sqrt{V(t)}\xi + e(t, X_0), t\right) + \frac{\xi}{\sqrt{V(t)}}\right\|^2\right]\right]\right] \\ &= \underset{t\sim U(0,T)}{E}\left[\lambda(t) \underset{x_0\sim p_{data}}{E}\left[\frac{1}{V(t)} \underset{\xi\sim N(0,1)}{E}\left[\left\| \xi_\theta\left(\sqrt{V(t)}\xi + e(t, X_0), t\right) - \xi\right\|^2\right]\right]\right] \end{aligned} \tag{109}$$

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

5.1.2 With labels

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: $X_t \sim p_t(x|y)$

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

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:

$$\nabla_x \log\left(p_t(x \mid y)\right) = S\nabla_x \log\left(p\left(y \mid x_t = x\right)\right) + \nabla_x \log\left(p\left(x_t = x\right)\right) \tag{111}$$

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

Classifier Guidance Free

$$\begin{aligned} &\gamma \nabla_x \log\left(p\left(y \mid X_t = x\right)\right) \\ =&\gamma\left(\nabla_x \log\left(p(X_t = x|y)\right) - \nabla_x \log\left(p_t(x)\right)\right) \end{aligned} \tag{112}$$

Then

$$\begin{aligned} &\nabla_x \log_\gamma\left(p_t(x \mid y)\right) \\ =&(1-\gamma)\nabla_x \log\left(p_t(x)\right) + \gamma\nabla_x \log\left(p(X_t = x|y)\right) \end{aligned} \tag{113}$$

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

5.2 Flow Matching

We have discussed the FPK Equation in 'learnsde'.

Theorem 26 (Fokken-Planck-Kolmogorov equation) *The density function $p(x,t)$ of X_t s.t.*

$$dX_t = f(X_t, t)dt + G(X_t, t)dW_t \tag{114}$$

solves the PDE:

$$\frac{\partial p(x,t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right] + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^\top\right)_{ij} p(x,t)\right] \tag{115}$$

The PDE is called FPK equation / forward Kolmogorov equation.

It can be rewritten as:

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\nabla \cdot [f(x,t)p(x,t)] + \frac{1}{2}\nabla^2 \cdot \left[\left(GQG^\top\right)p(x,t)\right] \\ &= -\nabla \cdot \left[f(x,t)p(x,t) - \frac{1}{2}\nabla \cdot \left[\left(GQG^\top\right)p(x,t)\right]\right]\end{aligned} \tag{116}$$

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:

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} &= -\nabla \cdot \left(fp - \frac{1}{2}\nabla \cdot (Mp)\right) \\ &= -\nabla \cdot \left(fp - \frac{1}{2}M\nabla p\right) \\ &= -\nabla \cdot \left[\left(f - \frac{1}{2}M\frac{\nabla p}{p}\right)p\right] \\ &= -\nabla \cdot \left[\left(f - \frac{1}{2}M\nabla \log p\right)p\right]\end{aligned} \tag{117}$$

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:

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

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'.

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.

$$\begin{cases}\dfrac{\partial \phi(x,t)}{\partial t} = v(\phi(x,t),t) \\ \phi(0,x) = x\end{cases} \tag{119}$$

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:

$$dX_t = v(X_t,t)dt + 0dW_t \tag{120}$$

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:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (v(x,t)p(x,t)) \tag{121}$$

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}.

5.3 Variational Auto-Encoder

Variational Autoencoders aim to learn both an encoder and a decoder to map input data to values in a continuous latent space. In these models, the embedding can be interpreted as a latent variable in a probabilistic generative model, and a probabilistic decoder can be formulated by a parameterized likelihood function. In addition, the data x is assumed to be generated by some unobserved latent variable z.

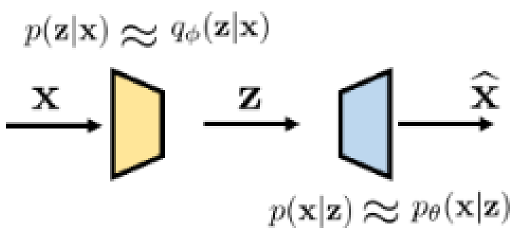

Figure 2: VAE block

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.

Definition 7 (Evidence Lower Bound) *The Evidence Lower Bound (ELBO) is defined as:*

$$\mathrm{ELBO}(x) = \mathbf{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] \tag{122}$$

Remark 6 *The ELBO is a lower bound of the log-likelihood of the data. It is used to estimate* $\log p(x)$*.*

$$\begin{aligned}\log p(x) &= \log \int p(x,z)dz \\ &= \log \int \frac{p(x,z)}{q_\phi(z|x)} \cdot q_\phi(z|x)dz \\ &\geq \mathbf{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] = \mathrm{ELBO}(x)\end{aligned} \tag{123}$$

Theorem 27 (Decomposition of Log-likelihood) *We have*

$$\log p(x) = \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z|x)||p(z|x)) \tag{124}$$

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 12

$$\begin{aligned}\log p(x) &= \log p(x) \int q_\phi(z|x)dz \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log\left(\frac{p(x,z)}{p(z|x)}\frac{q_\phi(z|x)}{q_\phi(z|x)}\right)\right] \\ &= \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z|x)||p(z|x))\end{aligned} \tag{125}$$

Theorem 28 *Also, we can rewrite the ELBO as:*

$$\begin{aligned}\mathrm{ELBO}(x) &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x|z) + \log p(z) - \log q_\phi(z|x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z)) \\ &= \mathbf{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z))\end{aligned} \tag{126}$$

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

$$(\phi, \theta) = \mathrm{argmax}_{\phi,\theta} \sum_{x \in X} \mathrm{ELBO(x)} \tag{127}$$

where X *is the training set.*

The DDPM can be conceptualized as a hierarchical Markovian VAE with a fixed encoder. Specifically, DDPM's forward process functions as the encoder. The DDPM's reverse process, on the other hand, corresponds to the decoder, which is shared across multiple decoding steps. The latent variables within the decoder are all the same size as the sample data.

6 Random Field

6.1 Definitions

Definition 9 (Random Field) *For a set* $D \subset \mathbb{R}^d$*, a (real-valued) random field* $u(x) : x \in D$ *is a set of real-valued random variables on a probability space* $(\Omega, \mathcal{F}, P)$*. We usually speak of realizations of random field, instead of sample paths.*

Definition 10 (second-order random field) *A random field is called second-order random field if $u(x) \in L^2(\Omega)$ for $\forall x \in D$. With its mean and covariance function:*

$$\begin{cases} \mu(x) = \mathbf{E}[u(x)] \\ C(x,y) = Cov(u(x), u(y)) = \mathbf{E}[(u(x) - m(x))(u(y) - m(y))] \end{cases} \tag{128}$$

Definition 11 (Gaussian Random Field) *A second-order random field $u(x) : x \in D$ is called Gaussian random field if*

$$u = u[u(x_1), u(x_2), \cdots, u(x_n)]^T \sim \mathcal{N}(\mu(x), C(x,y)),\ \forall x_i \in D \tag{129}$$

Example 3 ($L^2(D)$-valued random variable) *For $D \subset \mathbb{R}^d$, consider $L^2(D)$-valued R.V. u with $\mu \in L^2(D)$ and $\mathscr{C}$. Then $u(x)$ is a real-valued random field for each $x \in D$, and mean and covariance are well defined.*

Meanwhile, for $\phi, \psi \in L^2(D)$, we have

$$\begin{aligned} \langle \mathscr{C}\phi, \psi \rangle &= Cov\left(\langle u, \phi \rangle_{L^2(D)}, \langle u, \psi \rangle_{L^2(D)}\right) \\ &= E\left[\left(\int_D \phi(x)(u(x) - \mu(x))dx\right)\left(\int_D \psi(y)(u(y) - \mu(y))dy\right)\right] \\ &= \int_D \int_D \phi(x)\psi(y)E[(u(x) - mu(x))(u(y) - mu(y))]dxdy \\ &= \int_D \int_D \phi(x)\psi(y)Cov(u(x), u(y))dxdy \end{aligned} \tag{130}$$

So that

$$(\mathscr{C}\phi)(x) = \int_D Cov(u(x), u(y))\phi(y)dy \tag{131}$$

which is the covariance function of the random field $u(x)$. So, any $L^2(D)$-valued random variable defines a second-order random field, with mean $\mu(x)$ and covariance $C(x,y) = Cov(u(x), u(y))$ which is the kernel of the covariance operator $\mathscr{C}$.

Example 4 (Stationary Random Field) *A second-order random field $u(x) : x \in D$ is called stationary if the mean is constant and covariance function depends only on the difference $x-y$, i.e. $\mu(x) = \mu$, $C(x,y) = C(x-y)$.*

Theorem 29 (Wiener-Khinchin Theorem) *There exists a stationary random field $u(x) : x \in D$ with mean μ and covariance function $c(x)$ that is mean square continuous if and only if the function $c(x) : \mathbb{R}^d \to \mathbb{R}$ is such that*

$$c(x) = \int_{\mathbb{R}^d} e^{iv\cdot x} dF(v) = (2\pi)^{\frac{d}{2}} \hat{f}(x) \tag{132}$$

where $F(v)$ is some measure on $\mathbb{R}^d$ and $\hat{f}(x)$ is the Fourier transform of $f(x)$, f is the density function of F.

Reversely, $f(v) = (2\pi)^{\frac{d}{2}} \hat{c}(v)$. If f is non-negative and integrable, then $c(x)$ is a valid covariance function.

Example 5 (Isotropic Random Field) *A stationary random field is called isotropic if its covariance function depends only on the distance between points, i.e.*

$$Cov(x) = c(\|x\|_2) = c^0(r) \tag{133}$$

where c^0 is known as the isotropic covariance function.

6.2 Algorithms

In 2D cases, the covariance matrices of samples of stationary random fields $u(x)$ at uniformly spaced points $x \in D$ are symmetric BTTB matrices.

Definition 12 (Uniformly spaced points) *Let $D = [0, a_1] \times [0, a_2]$, the uniformly spaced points are given by:*

$$x_k = x_{i,j} = (i\Delta x_1, j\Delta x_2)^T,\ i = 0, 1, \cdots, n_1 - 1,\ j = 0, 1, \cdots, n_2 - 1, k = i + jn_1 \tag{134}$$

where $\Delta x_1 = \frac{a_1}{n_1 - 1}$ and $\Delta x_2 = \frac{a_2}{n_2 - 1}$.

With $N = n_1 n_2$, $u = [u_0, u_1, \cdots, u_{N-1}]^T \sim \mathcal{N}(0, C)$ is the vector of samples of $u(x)$ at the uniformly spaced points. Since $u(x)$ is stationary, C is a $N \times N$ symmetric BTTB matrix with elements:

$$C_{kl} = Cov(u_k, u_l) = c(x_{i+jn_1} - x_{r+sn_1}) \tag{135}$$

where $c(x_k - x_l)$ is the covariance function of $u(x)$.

Theorem 30 *The covariance matrix C is always a symmetric BTTB matrix.*

Since we have the Fourier representation of BCCB matrix and BTTB matrix can by extended to BCCB by even extension, we can use the following algorithm to generate the samples of $u(x)$. So, when the even BCCB extension $\tilde{C} \in \mathbb{R}^{4N\times 4N}$ is non-negative definite, then $N(0,\tilde{C})$ is a valid Gaussian distribution.

Algorithm 2 *Suppose the even BCCB extension $\tilde{C} \in \mathbb{R}^{4N\times 4N}$ is non-negative definite, and the leading principle submatrix $S \in \mathbb{R}^{2N\times 2N}$ is:*

$$S = \begin{pmatrix} \tilde{C}_0 & \tilde{C}_1^T & \cdots & \tilde{C}_{n_2-1}^T \\ \tilde{C}_1 & \tilde{C}_2 & \cdots & \tilde{C}_{n_2-2}^T \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{C}_{n_2-1} & \tilde{C}_{n_2-2} & \cdots & \tilde{C}_0 \end{pmatrix}, \quad \tilde{C}_i = \begin{pmatrix} C_i & B_i \\ B_i & C_i \end{pmatrix} \tag{136}$$

where $C_i, B_i \in \mathbb{R}^{n_1\times n_1}$, $i = 0, 1, \cdots, n_2 - 1$.

Now given $\tilde{u} \sim N(0,\tilde{C})$, let v be the first $2n_1n_2$ elements of $\tilde{u}$, then $v \sim N(0,S)$. Take the first n_1 elements of v per $2n_1$ elements to get $\tilde{v} \sim N(0,C)$.

However, when the even BCCB extension $\tilde{C} \in \mathbb{R}^{4N\times 4N}$ is indefinite, we can avoid this by padding. But sometimes, padding leads to the size of matrix explosion. Approximate circulant embedding may be the only option.

6.3 KL expansion of R.F.

As mentioned before, we have the underlying covariance operator defined by:

$$(\mathscr{C}\phi)(x) = \int_D Cov(u(x), u(y))\phi(y)dy = \int_D c(x-y)\phi(y)dy \tag{137}$$

Hence, for the covariance operator $\mathscr{C}$, we have the eigenfunctions with corresponding eigenvalues $\{v_j, \phi_j\}_{j=1}^{\infty}, v_j \geq v_{j-1}$.

Theorem 31 **(L^2 convergence of KL expansion)** *Let $D \subset \mathbb{R}^d$, consider a random field $u(x) : x \in D$ and $u \in L^2(\Omega, L^2(D))$, then:*

$$u(x) = \mu(x) + \sum_{j=0}^{\infty} \sqrt{v_j}\phi_j(x)\xi_j \tag{138}$$

where the sum converges in $L^2(\Omega, L^2(D))$,

$$\xi_j = \frac{1}{\sqrt{v_j}} \int_D (u(x) - \mu(x))\phi_j(x)dx \tag{139}$$

The random variables ξ_j have mean zero, unit variance and are pairwise uncorrelated. If u is Gaussian, then ξ_j are i.i.d. Gaussian random variables with zero mean and unit variance.

7 Stationary SPDEs

7.1 Definition

Definition 13 (Stationary SPDE) *Assume given $a, f \in L^2(\Omega, L^2(D))$ are random fields, try to seek $u : \bar{D} \times \Omega \to \mathbb{R}$ in weak sense s.t. $\mathbb{P}$-a.s.:*

$$\begin{cases} -\nabla \cdot (a(x,w)\nabla u(x,w)) = f(x,w), & x \in D \\ u(x,w) = g(x), & x \in \partial D \end{cases} \tag{140}$$

To ensure the existence of solution, we need to impose some conditions on g.

Definition 14 (Weak solution on $D \times \Omega$) *A weak solution to Eq(140) with $g = 0$ is a function $u \in V = L^2(\Omega, H_0^1(D))$ s.t. for any $v \in V$,*

$$a(u,v) = l(v) \tag{141}$$

where

$$\begin{cases} a(u,v) = E\left[\int_D a(x,\cdot)\nabla u(x,\cdot)\cdot\nabla v(x,\cdot)dx\right] \\ l(v) = E\left[\int_D f(x,\cdot)v(x,\cdot)dx\right] \end{cases} \tag{142}$$

If $g \neq 0$, a weak solution to Eq(140) is a function $u \in W = L^2(\Omega, H_g^1(D))$ s.t. for any $v \in V$,

$$a(u,v) = l(v) \tag{143}$$

where $a(\cdot,\cdot): W \times V \to \mathbb{R}$ and $l: V \to \mathbb{R}$:

$$\begin{cases} a(u,v) = E\left[\int_D a(x,\cdot)\nabla u(x,\cdot)\cdot\nabla v(x,\cdot)dx\right] \\ l(v) = E\left[\int_D f(x,\cdot)v(x,\cdot)dx\right] \end{cases} \tag{144}$$

Theorem 32 (Existence and uniqueness of weak solution) *Note for all $x \in D$*

$$0 < a_{\min} \leq a(x,\cdot) \leq a_{\max} < \infty \tag{145}$$

as a basic assumption.

If $f \in L^2(\Omega, L^2(D)), g = 0$, and Assumption (145) holds, then SPDE 141 has a unique weak solution $u \in V$.

If Assumption (145) holds, $f \in L^2(\Omega, L^2(D))$, and $g \in H^{\frac{1}{2}}(\partial D)$, then SPDE 143 has a unique weak solution $u \in W$.

Assume we have the approximate random fields $\tilde{a}, \tilde{f}: D \times \Omega \to \mathbb{R}$ s.t. (145) holds.

Then as mentioned before, we can expand a, f in terms of (truncated) Karhunen-Loeve expansion as:

$$\begin{cases} a(x,w) = \mu_a(x) + \sum_{i=1}^{N_a} \sqrt{v_i^a}\phi_i^a(x)\xi_i^a(w) \\ f(x,w) = \mu_f(x) + \sum_{i=1}^{N_f} \sqrt{v_i^f}\phi_i^f(x)\xi_i^f(w) \end{cases} \tag{146}$$

where $(v_i^a, \phi_i^a), (v_i^f, \phi_i^f)$ are the eigenpairs of the covariance operators of a, f respectively, and ξ_i^a, ξ_i^f are i.i.d. random variables.

The next question is how to compute:

$$\begin{aligned} a(u,v) &= E\left[\int_D a(x,\cdot)\nabla u(x,\cdot)\cdot\nabla v(x,\cdot)dx\right] \\ &= \int_\Omega\int_D a(x,w)\nabla u(x,w)\cdot\nabla v(x,w)dxdP(w) \end{aligned} \tag{147}$$

Since the truncated KL expansion of $a(x,w)$ depends on a finite number N_a of random variables $\xi_i^a: \Omega \to \Gamma_i$(same as $f(x,w)$), we consider weak form of Eq(140) on $D \times \Gamma$, where $\Gamma = \prod_{i=1}^{N_a}\Gamma_i$.

Definition 15 (finite-dimensional noise) *A function $v \in L^2(\Omega, L^2(D))$ of the form $v(x, \xi(w))$ for $\forall x \in D, w \in \Omega$, where $\xi = [\xi_1, \cdots, \xi_N]^T: \Omega \to \Gamma$, is called a finite-dimensional noise.*

Definition 16 (Weak solution on $D \times \Gamma$) *Let $\tilde{a}(x)$ and $\tilde{f}(x)$ be finite-dimensional noises defined in Eq(146), then the solution to Eq (140) is also finite-dimensional noise. Define*

$$W := L_p^2(\Gamma, H_g^1(D)) = \left\{v: D \times \Gamma \to \mathbb{R}: \int_\Gamma \|v(\xi,\cdot)\|_{H_g^1(D)}^2 d\xi < \infty\right\} \tag{148}$$

A weak solution to Eq(140) on $D \times \Gamma$ is a function $u \in W = L_p^2(\Gamma, H_g^1(D))$ s.t. for any $v \in V = L_p^2(\Gamma, H_0^1(D))$,

$$a(u,v) = l(v) \tag{149}$$

where

$$\begin{cases} a(u,v) = \int_\Gamma p(\xi)\int_D \tilde{a}(x,\xi)\nabla u(x,\xi)\cdot\nabla v(x,\xi)dxd\xi \\ l(v) = \int_\Gamma p(\xi)\int_D \tilde{f}(x,\xi)v(x,\xi)dxd\xi \end{cases} \tag{150}$$

7.2 Stochastic Galerkin Method

Therefore, we have the stochastic Galerkin solution: seek $u_{hk} \in W^{hk} \subset L^2(\Gamma, H_g^1(D))$ s.t. for any $v_{hk} \in V^{hk} \subset L^2(\Gamma, H_0^1(D))$.

By define the inner product:

$$\langle v, w\rangle_p = \int_\Gamma v(\xi)w(\xi)P(\xi)d\xi \tag{151}$$

We can construct a sequence of polynomials $P_i(\xi)$ on Γ. Hence:

$$L_p^2(\Gamma) := \{v : \Gamma \to \mathbb{R} : \|v\|^2_{L_p^2(\Gamma)} = \langle v, v\rangle_p < \infty\} \tag{152}$$

Definition 17 *Note S^k be the set of polynomials of degree k or less on Γ:*

$$\begin{aligned} S^k &= \operatorname{span}\{\prod_{i=1}^M P_i^{\alpha_i}(\xi_i) : \alpha_i \in \mathbb{N}_0, \sum_{i=1}^M \alpha_i \leq k\} \\ &= \operatorname{span}\{\psi_1, \psi_2, \cdots, \psi_Q\} \end{aligned} \tag{153}$$

where $P_i(\xi_i)$ is some polynomial. And $Q = \dim S^k = \binom{M+k}{k}$.

We need $S^k \subset L_p^2(\Gamma)$ where $\Gamma \subset \mathbb{R}^M$. If $\{\xi_i\}$ are independent, then the joint density p is:

$$p(\xi) = \prod_{i=1}^M p_i(\xi_i) \tag{154}$$

Recall $V^h = \operatorname{span}\{\phi_i\}_{i=1}^J \subset H_0^1(D)$ is the finite element space, we have tensor product space:

$$V^{hk} := V^h \otimes S^k = \operatorname{span}\{\phi_i\psi_j\}_{i=1,j=1}^{J,Q} \tag{155}$$

Then

$$W^{hk} := V^{hk} \oplus \operatorname{span}\{\phi_{J+1}, \cdots, \phi_{J+J_b}\} \tag{156}$$

where J_b is finite element functions associated with Dirichlet boundary vertices.

Theorem 33 (Stochastic basis functions) *If $\{\xi_i\}$ are independent, suppose that $\{P_i^{\alpha_i}(\xi_i)\}_{\alpha_i=1}^M$ are orthonormal with $\langle\cdot,\cdot\rangle_{p_i}$ on Γ_i. Then the complete orthonormal polynomials $\{\psi_j\}_{j=1}^Q$ are orthonormal with $\langle\cdot,\cdot\rangle_p$ on Γ.*

Then u_{hk} can be written as:

$$u_{hk}(x,\xi) = \sum_{i=1}^J\sum_{j=1}^Q u_{ij}\phi_i(x)\psi_j(\xi) + w_g \tag{157}$$

Theorem 34 (Mean and covariance) *The Galerkin solution can be rewritten as:*

$$\begin{aligned} u_{hk}(x,\xi) &= \sum_{j=1}^Q\left(\sum_{i=1}^J u_{ij}\phi_i(x)\right)\psi_j(\xi) + w_g \\ &= \sum_{j=1}^Q u_j\psi_j(\xi) + w_g \\ &= (u_1(x) + w_g(x))\psi_1(\xi) + \sum_{j=2}^Q u_j(x)\psi_j(\xi) \end{aligned} \tag{158}$$

Then the mean and covariance is

$$\begin{cases} E[u_{hk}] = u_1 + w_g \\ Var(u_{hk}) = \sum_{j=2}^Q u_j^2 \end{cases} \tag{159}$$

7.3 Algorithm

Expand u_{hk} in terms of basis functions $v = \phi_r\psi_s$ for $r = 1, 2, \cdots, J, s = 1, 2, \cdots, Q$, we have the linear system:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1Q} \\ A_{21} & A_{22} & \cdots & A_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ A_{Q1} & A_{Q2} & \cdots & A_{QQ} \end{pmatrix}, \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_Q \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_Q \end{pmatrix} \tag{160}$$

where

$$\mathbf{u}_j = [u_{1j}, u_{2j}, \cdots, u_{Jj}]^T, j = 1, 2, \cdots, Q \tag{161}$$

and each submatrix A_{ij} is a $J \times J$ matrix, $i, j = 1, 2, \cdots, Q$:

$$A_{ij} = K_0\langle\psi_i, \psi_j\rangle_p + \sum_{l=1}^{P} K_l\langle\psi_i, \psi_j\xi_l\rangle_p \tag{162}$$

where

$$\begin{cases} [K_0]_{rs} = \int_D \mu_a(x)\nabla\phi_r \cdot \nabla\phi_s dx \\ [K_l]_{rs} = \int_D (\sqrt{v_l^a}\phi_l^a)\nabla\phi_r \cdot \nabla\phi_s dx \end{cases}, \qquad r, s = 1, 2, \cdots, J \tag{163}$$

And $\mathbf{b}_s$ is a $J \times 1$ vector:

$$\mathbf{b}_s = \langle\psi_1, \psi_s\rangle_p F_0 + \sum_{l=1}^{P} F_l\langle\xi_l, \psi_s\rangle_p - \langle W, \psi_s\rangle_p \tag{164}$$

where

$$\begin{cases} [F_0]_i = \int_D \mu_f(x)\phi_i(x)dx \\ [F_l]_i = \int_D (\sqrt{v_l^f}\phi_l^f)\phi_i(x)dx \\ W = K_{0B}^T\mathbf{w}_B + \sum_{l=1}^{P} K_{lB}^T\xi_l\mathbf{w}_B \end{cases} \tag{165}$$

8 Semilinear Stochastic PDEs

8.1 Definition

Then we come to the time-dependent SPDE. We study the stochastic semilinear evolution equation:

$$du = [\Delta u + f(u)]dt + G(u)dW(t, x) \tag{166}$$

Definition 18 (Semilinear SPDE) *Simmilar to normal time-dependent PDE, we treat SPDE like this as semilinear SODEs on a Hilbert space, like*

$$du = [-Au + f(u)]dt + G(u)dW(t) \tag{167}$$

where $-A$ is a linear operator that generates a semigroup $S(t) = e^{-tA}$.

Example 6 (Phase-field model)

$$du = [\epsilon\Delta u + u - u^3]dt + \sigma dW(t, x) \tag{168}$$

Example 7 (Fluid Flow)

$$\begin{aligned} u_t &= \epsilon\Delta u - \nabla p - (u \cdot \nabla)u \\ \nabla \cdot u &= 0 \end{aligned} \tag{169}$$

So like we deal with integration of stochastic process like Itos or stratonovich, we need to generalize the Brownian Motion by introducing spatial variable to W(t). Here we define Q-Wiener Process.

First, we assume U is a Hilbert space. And $(\Omega, \mathbf{F}, \mathbf{F}_t, \mathbb{R})$ is a filtered probability space.

Definition 19 (Q) *$Q \in \mathcal{L}(U)$ is non-negative definite and symmetric. Further, Q has an orthonormal basis $\{\mathcal{X}_j : j \in \mathcal{N}\}$ of eigenfunctions with corresponding eigenvalues $q_j \geq 0$ such that $\sum_{j\in\mathcal{N}} q_j < \infty$ (i.e., Q is of trace class).*

Definition 20 (Q-Wiener Process) *A U-valued stochastic process $\{W(t) : t \geq 0\}$ is Q-Wiener process if*

- *W(0) = 0 a.s.*
- *W(t) is a continuous function $\mathbb{R}^+ \to U$, for each $\omega \in \Omega$.*
- *$W(t)$ is $\mathcal{F}_t$-adapted and $W(t) - W(s)$ is independent of $\mathcal{F}_s$ for $s \leq t$*
- *$W(t) - W(s) \sim N(0, (t-s)Q)$ for all $0 \leq s \leq t$*

Theorem 35 (Q-Wiener Process) *Assume we have Q defined in 19. Then, $W(t)$ is a Q-Wiener process if and only if*

$$W(t) = \sum_{j=1}^{\infty} \sqrt{q_j} \mathcal{X}_j \beta_j(t) \tag{170}$$

which is converges in $L^2(\Omega, C([0,T], U))$ and $\beta_j(t)$ are iid $\mathcal{F}_t$-Brownian motions and the series converges in $L^2(\Omega, U)$.

Theorem 36 ($H^r_{\text{per}}(0,a)$-valued process) *...*

Theorem 37 ($H^r_0(0,a)$-valued process) *...*

So, in place of $L^2(D)$, we develop the theory on a separable Hilbert space U with norm $\|\cdot\|_U$ and inner product $\langle\cdot,\cdot\rangle_U$ and define the Q-Wiener process $W(t) : t \geq 0$ as a U-valued process.

We mention the important case of Q = I, which is not trace class on an infinite-dimensional space U (as $q_j = 1$ for all j) so that the series does not converge in $L^2(\Omega, U)$. To extend the definition of a Q-Wiener process, we introduce the cylindrical Wiener process.

The key point is to introduce a second space U1 such that $U \subset U_1$ and Q = I is a trace class operator when extended to U_1.

Then we can define cylindrical Wiener process:

Definition 21 (Cylindrical Wiener Process) *Let U be a separable Hilbert space. The cylindrical Wiener process (also called space-time white noise) is the U-valued stochastic process W(t) defined by*

$$W(t) = \sum_{j=1}^{\infty} \mathcal{X}_j \beta_j(t)$$

where $\{\mathcal{X}_j\}$ is any orthonormal basis of U and $\beta_j(t)$ are iid $\mathcal{F}_t$-Brownian motions.

Theorem 38 *If for the second Hilbert space U_1, and the inclusion map $\mathcal{I} : U \to U_1$ is Hilbert-Schmidt. Then, the cylindrical Wiener process is a Q-Wiener process well-defined on U_1(Converges in $L^2(U, U_1)$).*

8.2 Ito integral solution

Here we consider the Ito integral $\int_0^t B(s)dW(s)$ for a Q-Wiener process $W(s)$. Since dW_t takes value in Hilbert space U, and we treat SPDE in Hilbert space H, the integral will also take value in Hilbert space H.

Hence, $B(s)$ should be $\mathcal{L}^2_0(U_0, H)$-valued process, where $U_0 \subset U$ known as Cameron-Martin space. So, $B(s)$ is an operator from U_0 to H. Then, we consider the set of operator B.

Definition 22 (L^2_0 space) *Let $U_0 := \{Q^{\frac{1}{2}}u : u \in U\}$, the set of linear operators $B : U_0 \to H$ is noted as L^2_0 s.t.*

$$\|B\|_{L^2_0} := \left(\sum_{j=1}^{\infty} \|BQ^{\frac{1}{2}}\mathcal{X}_j\|^2\right)^{\frac{1}{2}} = \|BQ^{\frac{1}{2}}\|_{\text{HS}(U_0,H)} < \infty \tag{171}$$

Remark 7 *If G is invertible, L^2_0 is the space of Hilbert-Schmidt operators $\text{HS}(U_0, H)$.*

Definition 23 *The stochastic integral can be defined by*

$$\int_0^t B(s)dW(s) := \sum_{j=1}^{\infty} \int_0^t B(s)\sqrt{q_j}\mathcal{X}_j d\beta_j(s) \tag{172}$$

So, we can have the truncated form:

$$\int_0^t B(s)dW^J(s) = \sum_{j=1}^{J} \int_0^t B(s)\sqrt{q_j}\mathcal{X}_j d\beta_j(s) \tag{173}$$

8.3 Solution

Consider the semilinear SPDE:

$$du = [-Au + f(u)]dt + G(u)dW(t) \tag{174}$$

given the initial condition $u_0 \in H$ and $A : \mathcal{D} \subset H \to H$ is a linear operator, $f : H \to H$ and $G : H \to L_0^2$.

Example 8 *Consider the stochastic heat equation:*

$$du = \Delta u dt + \sigma dW(t,x), u(0,x) = u_0(x) \in L^2(D) \tag{175}$$

where D is a bounded domain in $\mathbb{R}^d$ and σ is a constant. Also, homogeneous Dirichlet boundary condition is imposed on D. Hence,

$$H = U = L^2(D), f(u) = 0, G(u) = \sigma I \tag{176}$$

We see that $A = -\Delta$ with domain $\mathcal{D}(A) = H^2(D) \cap H_0^1(D)$.

In the deterministic setting of PDEs, there are a number of different concepts of solution. Here is the same for SPDEs. We can also define strong solution, weak solution and mild solution.

Definition 24 (strong solution) *A predictable H -valued process $\{u(t) : t \in [0,T]\}$ is called a strong solution if*

$$u(t) = u_0 + \int_0^t [-Au(s) + f(u(s))]ds + \int_0^t G(u(s))dW(s), \quad \forall t \in [0,T] \tag{177}$$

Definition 25 (weak solution) *A predictable H -valued process $\{u(t) : t \in [0,T]\}$ is called a weak solution if*

$$\langle u(t), v\rangle = \langle u_0, v\rangle + \int_0^t [-\langle u(s), Av\rangle + \langle f(u(s)), v\rangle]ds + \int_0^t \langle G(u(s))dW(s), v\rangle, \quad \forall t \in [0,T], v \in \mathcal{D}(A) \tag{178}$$

where

$$\int_0^t \langle G(u(s))dW(s), v\rangle := \sum_{j=1}^{\infty} \int_0^t \left\langle G(u(s))\sqrt{q_j}\mathcal{X}_j, v\right\rangle d\beta_j(s).$$

Definition 26 (mild solution) *A predictable H -valued process $\{u(t) : t \in [0,T]\}$ is called a mild solution if for $t \in [0,T]$*

$$u(t) = \mathrm{e}^{-tA}u_0 + \int_0^t \mathrm{e}^{-(t-s)A} f(u(s))ds + \int_0^t \mathrm{e}^{-(t-s)A}G(u(s))dW(s),$$

where e^{-tA} is the semigroup generated by $-A$. The right hand side is also called stochastic convolution.

Example 9 (stochastic heat equation in one dimension) *Consider the weak solution of 1D heat SPDE with $D = (0,\pi)$, so that $-A$ has eigenfunctions $\phi_j(x) = \sqrt{2/\pi}\sin(jx)$ and eigenvalues $\lambda_j = j^2$ for $j \in \mathbb{N}$. Suppose that $W(t)$ is a Q -Wiener process and the eigenfunctions $\mathcal{X}_j$ of Q are the same as the eigenfunctions ϕ_j of A. A weak solution satisfies: $\forall v \in \mathcal{D}(A)$,*

$$\begin{aligned}\langle u(t), v\rangle_{L^2(0,\pi)} &= \langle u_0, v\rangle_{L^2(0,\pi)} + \int_0^t \langle -u(s), Av\rangle_{L^2(0,\pi)} ds \\ &+ \sum_{j=1}^{\infty} \int_0^t \sigma\sqrt{q_j}\,\langle \phi_j, v\rangle_{L^2(0,\pi)}\, d\beta_j(s)\end{aligned} \tag{179}$$

Assume $u(t) = \sum_{j=1}^{\infty} \hat{u}_j(t)\phi_j$ for $\hat{u}_j(t) := \langle u(t), \phi_j\rangle_{L^2(0,\pi)}$. Take $v = \phi_j$, we have

$$\hat{u}_j(t) = \hat{u}_j(0) + \int_0^t (-\lambda_j)\,\hat{u}_j(s)ds + \int_0^t \sigma\sqrt{q}_j d\beta_j(s). \tag{180}$$

Hence, $\hat{u}_j(t)$ satisfies the SODE

$$d\hat{u}_j = -\lambda_j\hat{u}_j dt + \sigma\sqrt{q_j}d\beta_j(t) \tag{181}$$

Therefore, each coefficient $\hat{u}_j(t)$ is an Ornstein-Uhlenbeck (OU) process (see Examples 8.1 and 8.21), which is a Gaussian process with variance

$$\operatorname{Var}(\hat{u}_j(t)) = \frac{\sigma^2 q_j}{2\lambda_j}\left(1 - \mathrm{e}^{-2\lambda_j t}\right) \tag{182}$$

For initial data $u_0 = 0$, we obtain, by the Parseval identity (1.43),

$$\|u(t)\|^2_{L^2(\Omega, L^2(0,\pi))} = \mathbb{E}\left[\sum_{j=1}^{\infty} |\hat{u}_j(t)|^2\right] = \sum_{j=1}^{\infty} \frac{\sigma^2 q_j}{2\lambda_j}\left(1 - \mathrm{e}^{-2\lambda_j t}\right). \tag{183}$$

References

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

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

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

[4] 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.

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

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

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

[8] Rabi N. Bhattacharya and Edward C. Waymire. *Stochastic Processes with Applications*. Society for Industrial and Applied Mathematics, 2009.

[9] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2024.

A Probability Theory

We have known the definition of semigroup and generator of Markov process.

We consider a Markov process X_t in $\mathbb{R}^d$ with generator $\mathcal{L}$ and Markov semigroup P_t.

Definition 27 (Ergodic Markov Chain) *We say X_t is ergodic provided that the generator operator $\mathcal{L}$ has only constant solutions. Or equivalently, the Markov semigroup P_t has a unique invariant distribution π. $P_t g = g$ has only constant solutions.*

The invariant measure (distribution) governs the long-time dynamics of the Markov process. In particular, when $X_0 \sim \mu_0$ initially, we have:

$$\lim_{t\to\infty} P_t^* \mu_0 = \mu \tag{184}$$

Furthermore,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(X_s)ds = \int_{\mathbb{R}^d} f(x)\mu(dx) \tag{185}$$

B Conservation Laws

Theorem 39 *Two important theorems in calculus:*

*1.**Divergence Theorem**:*

$$\int_\Omega \nabla\cdot \mathbf{F} dx = \int_{\partial\Omega} \mathbf{F}\cdot\mathbf{n} dS \tag{186}$$

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

$$\frac{d}{dt}\int_{\Omega(t)} f(t,x)dx = \int_{\Omega(t)} \frac{\partial f}{\partial t}dx + \int_{\partial\Omega(t)} f(t,x)\mathbf{v}\cdot\mathbf{n}dS \tag{187}$$

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 28 (Flow Map) *Assume a description of some characteristic of particle $\mathbf{P}$, 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) $\mathbf{F}:\mathcal{R}^m\times\mathcal{R}\to\mathcal{R}^m$:*

$$\begin{cases} \dfrac{d\phi_s^t(\mathbf{x})}{dt} = \mathbf{F}(\phi_s^t(\mathbf{x}),t) \\ \phi_s^s(\mathbf{x}) = \mathbf{x} \end{cases} \tag{188}$$

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

$$\begin{aligned} \frac{d}{dt}\int_{\Omega(t)} \rho(t,\mathbf{x})d\mathbf{x} &= \int_{\Omega(t)} \frac{\partial \rho}{\partial t}d\mathbf{x} + \int_{\partial\Omega(t)} \rho(t,\mathbf{x})\mathbf{u}\cdot\mathbf{n}dS \\ &= \int_{\Omega(t)} \left(\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\right)d\mathbf{x} = 0 \end{aligned} \tag{189}$$

Therefore:

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}) = 0 \tag{190}$$

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

Theorem 41 (Conservation of Momentum) *By conservation of momentum, i.e.*

$$\frac{d}{dt}\int_{\Omega(t)} \rho(t,\mathbf{x})\mathbf{v}(t,\mathbf{x})d\mathbf{x} = -\int_{\partial\Omega(t)} p\cdot\mathbf{n}dS \tag{191}$$

we have:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\otimes\mathbf{v}+p) = 0 \tag{192}$$

where p is the pressure.

Theorem 42 (Conservation of Energy)

$$\frac{\partial E}{\partial t} + \nabla \cdot (\mathbf{v}(E+p)) = 0 \tag{193}$$

Then we have can get the Euler's equation:

Theorem 43 (Euler's Equation) *The Euler's equation is given by:*

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho \\ \rho\mathbf{v} \\ E\end{bmatrix} + \nabla \cdot \begin{bmatrix}\rho\mathbf{v} \\ \rho\mathbf{v}\otimes\mathbf{v}+p \\ \mathbf{v}(E+p)\end{bmatrix} = 0 \tag{194}$$

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:

$$\frac{\partial U}{\partial t} + \nabla \cdot (F(U)) = 0 \tag{195}$$

C Linear Algebra

Definition 29 (Matrix Kronecker Product) *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ be two matrices. The Kronecker product of A and B is defined as:*

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix} \tag{196}$$

Definition 30 (Fourier Matrix) *Let n be a positive integer. The Fourier matrix W_n is defined as:*

$$W_n = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & w & \cdots & w^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & w^{n-1} & \cdots & 1 \end{bmatrix} \tag{197}$$

Definition 31 (Two dimensional Fourier Matrix) *Let W_1, W_2 be $n_1 \times n_1$ and $n_2 \times n_2$ Fourier matrices respectively. Then the two dimensional Fourier matrix is defined as:*

$$W = W_2 \otimes W_1 \tag{198}$$

which is the $n_1n_2 \times n_1n_2$ matrix.

Theorem 44 (Two-dimensional DFT) *The two-dimentional DFT of $V \in \mathbb{C}^{n_1\times n_2}$ is $\hat{V} \in \mathbb{C}^{n_1\times n_2}$ whose elements are given by:*

$$\hat{V}_{ij} = \sum_{k=0}^{n_1-1}\left(\sum_{l=0}^{n_2-1} V_{kl}w_1^{ik}\right)w_2^{jl}, w_1 = e^{-\frac{2\pi i}{n_1}}, w_2 = e^{-\frac{2\pi i}{n_2}} \tag{199}$$

where $i = 0, \cdots, n_1 - 1, j = 0, \cdots, n_2 - 1$.

$$\hat{V} = \sqrt{n_1n_2}\,\mathrm{array}\left((W_2 \otimes W_1)\tilde{v}\right),\ \tilde{v} = vec(V) \tag{200}$$

Do DFT on column and row respectively.

Theorem 45 (Two-dimensional IDFT) *The two-dimentional IDFT of $\hat{V} \in \mathbb{C}^{n_1\times n_2}$ is $V \in \mathbb{C}^{n_1\times n_2}$ whose elements are given by:*

$$V_{kl} = \frac{1}{n_1n_2}\sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1}\hat{V}_{ij}w_1^{-ik}w_2^{-jl}, w_1 = e^{\frac{2\pi i}{n_1}}, w_2 = e^{\frac{2\pi i}{n_2}} \tag{201}$$

where $k = 0, \cdots, n_1 - 1, l = 0, \cdots, n_2 - 1$.

$$V = \frac{1}{\sqrt{n_1n_2}}\mathrm{array}\left((W_2 \otimes W_1)^*\hat{v}\right),\ \hat{v} = vec(\hat{V}) \tag{202}$$

Do IDFT on column and row respectively.

Theorem 46 (Fourier representation of BCCB matrix) *Let $C \in \mathbb{R}^{n_1\times n_2}$ be a BCCB matrix, then $C = WDW^*$, where W is the two-dimensional Fourier matrix, and D is a diagonal matrix with diagonal $d = vec(\Lambda)$, where*

$$\Lambda = \sqrt{n_1n_2}\,\mathrm{array}\left((W)^*c_{red}\right),\ c_{red} = \mathrm{vec}(C_{red}) \tag{203}$$

D Priori

D.1 Hilbert space-valued random variable

Definition 32 ($L^p(\Omega, H)$ space) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and H is a Hilbert space with norm $\|\cdot\|$. Then $\mathcal{L}^p(\Omega, H)$ with $1 \le p < \infty$ is the space of H-valued $\mathcal{F}$-measurable random vaiables $X : \Omega \to H$ with $\mathbf{E}[\|X\|^p] < \infty$ and a Banach space with norm:*

$$\|X\|_{\mathcal{L}^p(\Omega,H)} := \left(\int_\Omega \|X(\omega)\|^p dP(\omega)\right)^{\frac{1}{p}} = \mathbf{E}[\|X\|^p]^{\frac{1}{p}} \tag{204}$$

Then we can define the inner product:

$$\langle X, Y\rangle_{\mathcal{L}^2(\Omega,H)} := \int_\Omega \langle X(\omega), Y(\omega)\rangle dP(\omega) \tag{205}$$

Definition 33 (uncorrelated, covariance operator) *Let H be a Hilbert space. A linear operator $\mathcal{C} : H \to H$ is the covariance of H-valued random variables X and Y if*

$$\langle \mathcal{C}\phi, \psi\rangle = \mathrm{Cov}\left(\langle X, \phi\rangle, \langle Y, \psi\rangle\right), \forall \phi, \psi \in H \tag{206}$$

specially, we show that in finite dimensional case, the covariance matrix conincides with the covariance operator. when $H = \mathbb{R}^d$,

$$\begin{aligned} &\mathrm{Cov}\left(\langle X, \phi\rangle, \langle Y, \psi\rangle\right) = \mathrm{Cov}\left(\phi^T X, \psi^T Y\right) \\ =&\mathbf{E}\left[\phi^T(X-\mu_X)(Y-\mu)^T\psi\right] = \phi^T\mathbf{E}\left[(X-\mu_X)(Y-\mu)^T\right]\psi \\ =&\langle C\phi, \psi\rangle \end{aligned} \tag{207}$$

Definition 34 (H-valued Gaussian random variable) *Let H be a Hilbert space. An H-valued random variable X is Gaussian if $\langle X, \phi\rangle$ is a real-valued Gaussian random variable for all $\phi \in H$.*

D.2 Hilbert-Schmidt operator

Definition 35 (Hilbert-Schmidt operator) *Let U, H be two separable Hilbert spaces with norms $\|\cdot\|, \|\cdot\|_U$ respectively. For an orthonormal basis $\{\phi_j\}$ of U, define the Hilbert-Schmidt norm:*

$$\|L\|_{\mathrm{HS}(U,H)} := \left(\sum_{j=1}^{\infty} \|L\phi_j\|_H^2\right)^{\frac{1}{2}} \tag{208}$$

where $\mathrm{HS}(U, H) := \{L \in \mathcal{L}(U, H) : \|L\|_{\mathrm{HS}(U,H)} < \infty\}$ is a Banach space with Hilbert-Schmidt norm. And $L \in \mathrm{HS}(U, H)$ is called Hilbert-Schmidt operator.

Definition 36 (Integral operator with kernel G) *For a domain D and a kernel $G \in L^2(D \times D)$, define the integral operator L by*

$$(Lu)(x) = \int_D G(x, y)u(y)dy, x \in D, u \in L^2(D) \tag{209}$$

Furthermore, L is a Hilbert-Schmidt operator.

D.3 Operator theory

Theorem 47 (Sobolev embedding theorem) *1.Let $W^{r,p}(\mathbf{R}^n)$. Here k is a non-negative integer and $1 \le p < \infty$. If $k > \ell, p < n$ and $1 \le p < q < \infty$ are two real numbers such that $\frac{1}{p} - \frac{r}{n} = \frac{1}{q} - \frac{\ell}{n}$, then*

$$W^{r,p}(\mathbf{R}^n) \subseteq W^{\ell,q}(\mathbf{R}^n) \tag{210}$$

Specially, if $\ell = 0$, then $\frac{1}{p} - \frac{r}{n} = \frac{1}{q}$, then $W^{r,p}(\mathbf{R}^n) \subseteq L^q(\mathbf{R}^n)$.

2.If $n < pr$ and $\frac{1}{p} - \frac{r}{n} = -\frac{s+\alpha}{n}$, then $W^{r,p}(\mathbf{R}^n) \subseteq C^{s,\alpha}(\mathbf{R}^n)$.

Definition 37 (domain of operator) *For a linear operator $A : \mathcal{D}(A) \subset H \to H$, the domain of A is defined as $\mathcal{D}(A)$*

Theorem 48 (Dirichlet Boundary Condition) *Consider the Dirichlet problem for Possion equation: for $f \in L^2(0,1)$, find $u \in H^2(0,1)$ s.t.*

$$\begin{aligned} u_{xx} &= f, \quad x \in (0,1) \\ u(0) &= u(1) = 0 \end{aligned} \tag{211}$$

We also assume $u \in H_0^1(0,1)$. By Sobolev embedding theorem, $u \in H_0^1(0,1) \subset C([0,1])$. Then, Laplacian with Dirichlet conditions can be defined as:

$$Au := -u_{xx}, u \in \mathcal{D}(A) = H^2(0,1) \cap H_0^1(0,1) \tag{212}$$

Definition 38 (Periodic Boundary Condition) *...*

Definition 39 *If A is a linear operator from $\mathcal{D}(A) \subset H$ to Hilbert space H, with an orthonormal basis of eigenfunctions $\{\phi_j\}$ and corresponding increasing eigenvalues $\{\lambda_j\}$, then A^α is defined as:*

$$A^\alpha u = \sum_{j=1}^{\infty} \lambda_j^\alpha \langle u, \phi_j \rangle \phi_j \tag{213}$$

and the domain $\mathcal{D}(A^\alpha)$ is the set of all $u \in H$ such that $A^\alpha u \in H$.