Notebook

Guanyu Chen

June 27, 2025

Zhejiang University

Abstract

Here is my notebook.

Keywords: SDE · SPDE · Generative Model · Neural SDE · Neural SPDE

Contents

List of Tables

List of Figures

Chapter 1

SDE

1.1 Introduction to SDEs

1.1.1 SODEs

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

$$dX_t = f(X_t, t)dt + G(X_t, t)dW_t \tag{1.1}$$

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 The It'so and Stratonovich Stochastic Integrals

1.1.3 Ito's Formula

Theorem 1 (Ito's Formula) *Let X_t be an Ito process defined by 1.1, and let $\phi(X_t)$ be a twice continuously differentiable function of x and t. Then the process $f(X_t, t)$ is also an Ito process, and its dynamics are given by:*

$$d\varphi(X_t) = \nabla\varphi(X_t) \cdot dX_t + \frac{1}{2}\operatorname{Tr}\left(GQG^T\nabla^2\varphi(X_t)dt\right) \tag{1.2}$$

1.1.4 Mean and Covariance

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

$$\frac{dE[\phi]}{dt} = E\left[\frac{\partial\phi}{\partial t}\right] + \sum_i E\left[\frac{\partial\phi}{\partial x_i}f_i(X_t, t)\right] + \frac{1}{2}\sum_{ij} E\left[\frac{\partial^2\phi}{\partial x_i\partial x_j}\left[GQG^\top\right]_{ij}\right] \tag{1.3}$$

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.

$$\begin{cases} \dfrac{dm}{dt} = E\left[f(X_t, t)\right] \\ \dfrac{dc}{dt} = E\left[f(X, t)(X - m(t)^T)\right] + E\left[(X - m(t)f^T(X, t))\right] + E\left[G(X_t, t)QG^T(X_t, t)\right] \end{cases} \tag{1.4}$$

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

$$dX_t = (K(t)X_t + B(t))\,dt + G(t)dW_t \tag{1.5}$$

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

$$X_t = \Psi(t,t_0)X_0 + \int_{t_0}^{t} \Psi(t,s)B(s)ds + \int_{t_0}^{t} \Psi(t,s)G(s)dW_s \tag{1.6}$$

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

$$\frac{d\Psi}{dt} = K(t)\Psi(t,t_0), \Psi(t_0,t_0) = I \tag{1.7}$$

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

$$\begin{cases} \dfrac{dm}{dt} = K(t)m(t) + B(t) \\ \dfrac{dc}{dt} = K(t)c(t) + c(t)K^T(t) + G(t)QG^T(t) \end{cases} \tag{1.8}$$

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:

$$\begin{cases} m(t) = \Psi(t,t_0)m_0 + \displaystyle\int_{t_0}^{t} \Psi(t,s)B(s)ds \\ c(t) = \Psi(t,t_0)c_0\Psi^T(t,t_0) + \displaystyle\int_{t_0}^{t} \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{1.9}$$

Proof 2 *Apply $F(X,t) = K(t)X + B(t), G(X,t) = G(t)$ to 1.4.*

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

Theorem 4 *The solution to LSDE is Gaussian:*

$$p(X,t) = \mathcal{N}(X(t)|m(t),c(t)) \tag{1.10}$$

Specially when $X_0 = x_0$ is fixed, then

$$p(X,t|X_0 = x_0) = \mathcal{N}(X(t)|m(t|x_0),c(t|x_0)) \tag{1.11}$$

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

$$\begin{cases} m(t|x_0) = \Psi(t,t_0)x_0 + \displaystyle\int_{t_0}^{t} \Psi(t,s)B(s)ds \\ c(t|x_0) = \displaystyle\int_{t_0}^{t} \Psi(t,s)G(s)QG^T(s)\Psi^T(t,s)ds \end{cases} \tag{1.12}$$

Proof 3 *The proof is straight foward either by applying $m_0 = x_0, c_0 = 0$ to 1.9 or by eq 1.6.*

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

$$X_{t_{i+1}} = A_i X_{t_i} + B_i + G_i \tag{1.13}$$

where

$$\begin{cases} A_i = \Psi(t_{i+1}, t_i) \\ B_i = \int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s) B(s) ds \\ G_i = \int_{t_i}^{t_{i+1}} \Psi(t_{i+1}, s) G(s) Q G^T(s) \Psi^T(t_{i+1}, s) ds \end{cases} \tag{1.14}$$

Proof 4 *The proof is straight forward.*

Theorem 6 *The covariance of X_t and $X_s (s < t)$ is given by:*

$$Cov(X_t, X_s) = \Psi(t, s) c(s) \tag{1.15}$$

Proof 5 *See Sarkka P88-89.*

1.2 Fokker-Planck-Kolmogorov Equation

1.2.1 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*

$$\mathcal{A}\phi(X_t) = \lim_{s \to 0^+} \frac{E[\phi(X(t+s)] - \phi(X(t))}{s} \tag{1.16}$$

Where ϕ is a suitable regular function.

This leads to Dynkin's Formula very naturally.

Theorem 7 (Dynkin's Formula)

$$E[f(X_t)] = f(X_0) + E\left[\int_0^t \mathcal{A}(f(X_s)) ds\right] \tag{1.17}$$

Theorem 8 *If $X(t)$ s.t. 1.1, then the generator is given:*

$$\mathcal{A}(\cdot) = \sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t, t) + \frac{1}{2} \sum_{i,j} \left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j} \right) \left[G(X_t, t) Q G^\top(X_t, t) \right]_{ij} \tag{1.18}$$

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:*

$$A_t \phi(x, t) = \lim_{s \to 0^+} \frac{E[\phi(X(t+s), t+s)] - \phi(X(t), t)}{s} \tag{1.19}$$

Theorem 9 *Similarly if $X(t)$ s.t. 1.1, then the generalized generator is given:*

$$\mathcal{A}_t(\cdot) = \frac{\partial(\cdot)}{\partial t} + \sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(X_t,t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right)\left[G(X_t,t)QG^\top(X_t,t)\right]_{ij} \tag{1.20}$$

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

Theorem 10 (Fokken-Planck-Kolmogorov equation) *The density function $P(x,t)$ of X_t s.t. 1.1 solves the PDE:*

$$\frac{\partial P(x,t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right] + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[\left(GQG^\top\right)_{ij} P(x,t)\right] \tag{1.21}$$

The PDE is called FPK equation / forward Kolmogorov equation.

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

$$\begin{aligned} d\phi &= \sum_i \frac{\partial \phi}{\partial x_i} dx_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right) dx_i dx_j \\ &= \sum_i \frac{\partial \phi}{\partial x_i}\left(f_i\left(X_t,t\right)dt + \left(G\left(X_t,t\right)dW_t\right)\right) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right)\left[G(X_t,t)QG^\top(X_t,t)\right]_{ij} dt. \end{aligned} \tag{1.22}$$

Take expectation of both sides:

$$\frac{dE[\phi]}{dt} = \sum_i E\left[\frac{\partial \phi}{\partial x_i} f_i(X_t,t)\right] + \frac{1}{2}\sum_{ij} E\left[\frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij}\right] \tag{1.23}$$

So

$$\begin{cases} \dfrac{dE[\phi]}{dt} = \dfrac{d}{dt}\left[\displaystyle\int \phi(x)P(X_t = x,t)dx\right] = \displaystyle\int \phi(x)\frac{\partial P(x,t)}{\partial t}dx \\ \displaystyle\sum_i E\left[\frac{\partial \phi}{\partial x_i} f_i\right] = \sum_i \int \frac{\partial \phi}{\partial x_i} f_i(X_t = x,t)Pdx = -\sum_i \int \phi \cdot \frac{\partial}{\partial x_i}\left[f_i(x,t)p(x,t)\right]dx. \\ \displaystyle\frac{1}{2}\sum_{ij} E\left[\frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij}\right] = \frac{1}{2}\sum_{ij}\int \frac{\partial^2 \phi}{\partial x_i \partial x_j}\left[GQG^\top\right]_{ij} Pdx = \frac{1}{2}\sum_{ij}\int \phi(x)\frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij} P\right)dx. \end{cases} \tag{1.24}$$

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.

$$\frac{\partial P}{\partial t} + \sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)P(x,t)\right) - \frac{1}{2}\sum_{i=1}\frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij} P\left(x,t\right)\right) = 0 \tag{1.25}$$

Which gives the FPK Equation.

Remark 1 *When SDE is time independent:*

$$dX_t = f(X_t)dt + G(X_t)dW_t \tag{1.26}$$

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 1.23 can be written as

$$\frac{d}{dt}\langle \phi, P \rangle = \langle \mathcal{A}\phi, P \rangle \tag{1.27}$$

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

$$\langle \phi, \frac{dP}{dt} - \mathcal{A}^*(P) \rangle = 0, \forall \phi(x) \tag{1.28}$$

Hence we have

Theorem 11 (FPK Equation)

$$\frac{dP}{dt} = \mathcal{A}^*(P), \text{where}\, \mathcal{A}^*(\cdot) = -\sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)(\cdot)\right) + \frac{1}{2}\sum_{i=1} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij}(\cdot)\right) \tag{1.29}$$

It can be rewritten as:

$$\begin{aligned} \frac{\partial P}{\partial t} &= -\nabla \cdot [f(x,t)P(x,t)] + \frac{1}{2}\nabla^2 \cdot \left[\left(GQG^\top\right)P(x,t)\right] \\ &= -\nabla \cdot \left[f(x,t)P(x,t) - \frac{1}{2}\nabla \cdot \left[\left(GQG^\top\right)P(x,t)\right]\right] \end{aligned} \tag{1.30}$$

We define the probability flux to be:

$$J(x,t) = f(x,t)p(x,t) - \frac{1}{2}\nabla \cdot [M(x)p(x,t)], M(x) = G(x,t)Q(x,t)G(x,t)^T \tag{1.31}$$

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

$$\frac{d}{dt}\int_{R^d} p(x,t)dx = \int_{R^d} \nabla \cdot J(x,t)dx = 0 \tag{1.32}$$

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

$$\nabla \cdot J(x,t) = 0 \tag{1.33}$$

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

1.2.2 Forward and backward Komogorov Equation

Theorem 12 *Fix $t > s$, let $u(x,s) := E\left[g(X_t)|X_s = x\right] = \int g(y)P(y,t|x,s)dy$, then $u(x,s)$ satisfies the following equation:*

$$\frac{\partial u}{\partial s} + f(x,s) \cdot \nabla u + \frac{1}{2}\nabla \cdot (M \nabla u) = 0, \qquad u(x,s) = g(x) \tag{1.34}$$

Theorem 13 (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*

$$\frac{\partial P_{t|s}(x|y)}{\partial t} = \mathcal{A}^*(P_{t|s}(x|y)), \text{with}\, P_{s|s}(x|y) = \delta(x-y) \tag{1.35}$$

The Feynman-Kac Formula bridges PDE and certain stochastic value of SDE solutions. Consider $u(x,t)$ satisfied the following PDE:

$$\frac{\partial u}{\partial t} + f(x)\frac{\partial u}{\partial x} + \frac{1}{2}L^2(x)\frac{\partial^2 u}{\partial x^2} = 0. \quad u(x,T) = \psi(x). \tag{1.36}$$

Then we define a stochastic process $X(t)$ on $[t', T]$ as

$$dX = f(X)dt + L(X)dW_t \quad X\left(t'\right) = x' \tag{1.37}$$

By Ito formula:

$$\begin{aligned} du &= \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial x}dx + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}dx^2 \\ &= \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial x}\left(f(x)dt + L(x)dW_t\right) + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}L^2(x)dt \\ &= \left(\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}f(x) + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}L^2(x)\right)dt + \frac{\partial u}{\partial x}L(x)dW_t. \\ &= \frac{\partial u}{\partial x}L(x)dW_t. \end{aligned} \tag{1.38}$$

Integrating both sises from t' to T:

$$\begin{aligned} \int_{t'}^{T} \frac{\partial u}{\partial x}L(x)dW_t &= u(X(T),T) - u(X(t'),t') \\ &= \psi(X(T)) - u(x',t') \end{aligned} \tag{1.39}$$

Take expectation of both sides:

$$u\left(x',t'\right) = E[\psi(X(T))] \tag{1.40}$$

Theorem 14 (Feynman-Kac Formula) *This can be generalized to PDE like:*

$$\frac{\partial u}{\partial t} + f(x)\frac{\partial u}{\partial x} + \frac{1}{2}L^2(x)\frac{\partial^2 u}{\partial x^2} - V(x,t)u = 0. \quad u(x,T) = \psi(x). \tag{1.41}$$

By consider the Ito formula of $e^{-\int_0^t V(x,s)ds}u(x,t)$, we can similarly compute the resulting Feynman-Kac equation as

$$u(x',t') = e^{-\int_0^t V(x,s)ds}E\left[\psi(X(t))\right] \tag{1.42}$$

This means we can get the value of PDE at (x',t') by simulating SDE paths beginning at (x',t'), and compute corresponding $E\left[\psi(X(T))\right]$.

Reversely, if we consider the PDE the inital condition $u(x,0)$:

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \mathcal{L}u(t,x) + V(t,x)u(t,x), \quad t > 0,\ x \in \mathbb{R}^d \\ u(0,x) = \psi(x) \end{cases} \tag{1.43}$$

where $\mathcal{L}u(t,x) = \sum_{i=1}^d f_i(t,x)\frac{\partial u}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^d L_{ij}^2(t,x)\frac{\partial^2 u}{\partial x_i \partial x_j}$, then we can get the solution of PDE by Feynman-Kac formula:

$$\begin{aligned} &u(x,t) = E\left[\psi(X(t))e^{\int_0^t V(X_s,s)ds}|X(0) = x\right] \\ &dX_t = f(X_t,t)dt + L(X_t,t)dW_t \,\text{with}\, X(0) = x \end{aligned} \tag{1.44}$$

We can get more generalized conclusion:

Algorithm 1 (Solve Backward PDE) *To compute the backward PDE: $(\mathcal{A}_t - r)(u) = 0$, i.e.*

$$\frac{\partial u}{\partial t} + \sum_i \frac{\partial u(x,t)}{\partial x_i} f_i(x,t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 u(x,t)}{\partial x_i \partial x_j}\right)\left[G(x,t)QG^\top(x,t)\right]_{ij} - ru(x,t) = 0 \tag{1.45}$$

with boundary condition $u(x,T) = \psi(x)$. Then for any fixed points (x', t') where $t' \leq T, x' \in D$, $u(x', t')$ can be computed as:
Step1. Simulate N sample paths of SDE from t' to T:

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \,\text{with}\, X(t') = x' \tag{1.46}$$

Step2. Estimate $u(x', t') = e^{-r(T-t')}E\left[\psi(X(T))\right]$

Algorithm 2 (Solve Forward PDE) *Consider the solution $u(x,t)$ of forward PDE: $\frac{\partial u}{\partial t} = (\mathcal{A} - r)(u)$, i.e.*

$$\frac{\partial u}{\partial t} = \sum_i \frac{\partial u(x,t)}{\partial x_i} f_i(x,t) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 u(x,t)}{\partial x_i \partial x_j}\right)\left[G(x,t)QG^\top(x,t)\right]_{ij} - ru(x,t) \tag{1.47}$$

with initial condition $u(x,0) = \psi(x)$. Then for any fixed points (x', t') where $t' \leq T, x' \in D$, $u(x', t')$ can be computed as:
Step1. Simulate N sample paths of SDE from 0 to t':

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \,\text{with}\, X(0) = x' \tag{1.48}$$

Step2. Estimate $u(x', t') = e^{-rt'}E\left[\psi(X(t'))\right]$

Algorithm 3 (Solve Boundary Value Problem) *For solution $u(x)$ to the following elliptic PDE defined on some domain D:*

$$\sum_i \frac{\partial u(x)}{\partial x_i} f_i(x) + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 u(x)}{\partial x_i \partial x_j}\right)\left[G(x)QG^\top(x)\right]_{ij} - ru(x) = 0 \tag{1.49}$$

with boundary condition $u(x) = \psi(x)$ on ∂D. Then for any fixed points in D can be computed as:
Step1. Simulate N sample paths of SDE from t' to the first exit time T_e:

$$dX_t = f(X_t)dt + G(X_t)dW_t \,\text{with}\, X(t') = x' \tag{1.50}$$

Step2. Estimate $u(x') = e^{-r(T_e-t')}E\left[\psi(X(T_e))\right]$

1.2.3 Ornstein-Uhlenbeck Process

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

$$dX_t = -\alpha X_t dt + \sqrt{2D}dW_t \tag{1.51}$$

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

By FPK equation, we have:

$$\begin{cases} \dfrac{\partial p}{\partial t} = \alpha \dfrac{\partial}{\partial x}(xp) + D\dfrac{\partial^2 p}{\partial x^2} \\ p_0(x|x_0) = \delta(x - x_0) \end{cases} \tag{1.52}$$

When (1.52) 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 (1.52) can be computed:

$$X_t \sim N(x_0 e^{-\alpha t}, \frac{D}{\alpha}(1-e^{-2\alpha t})) \tag{1.53}$$

The generator of OU process is:

$$\mathcal{L} = -\alpha x \cdot \nabla + D\Delta \tag{1.54}$$

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

$$\mathcal{L}^*(h\rho) = \rho \mathcal{L} h \tag{1.55}$$

The IVP for FPK equation:

$$\frac{\partial p}{\partial t} = \mathcal{L}^* p, \qquad p(x,0) = p_0(x) \tag{1.56}$$

becomes:

$$\frac{\partial h}{\partial t} = \mathcal{L} h, \qquad h(x,0) = \rho^{-1} p_0(x) \tag{1.57}$$

Theorem 15 *Consider the eigenpairs problem for the generator operator $\mathcal{L}$ of OU process:*

$$\begin{cases} \lambda_n = \alpha n \\ \phi_n(x) = \frac{1}{n!} H_n(\sqrt{\alpha\beta} x) \end{cases} \qquad n = 0, \cdots, \infty \tag{1.58}$$

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

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n}(e^{-x^2/2}) \tag{1.59}$$

1.2.4 Langevin SDE

The Langevin SDE has the following form:

$$X_{t+s} = X_t + \nabla \log p_t(x_t) s + \sqrt{2s}\xi \tag{1.60}$$

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 16 *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*

$$\begin{aligned} \mu_{t+\tau} =& E\left[f\left(X_t + \nabla \log p_t(x_t) \cdot \tau + \sqrt{2\tau}\xi\right)\right] \\ =& E\Big[f(x_t) + \nabla^\top f(x_t)\left(\tau \nabla \log p_t(x_t) + \sqrt{2\tau}\xi\right) \\ &+ \frac{1}{2}\left(\nabla^\top \log p_t(x_t)\tau + \sqrt{2\tau}\xi\right) \nabla^2 f(x_t) \nabla \log p_t(x_t)\tau + \sqrt{2\tau}\xi\Big] \\ =& E[f(x_t)] + E\left[\tau \nabla^\top f(x_t) \nabla \log p_t(x_t)\right] \\ &+ \frac{\tau^2}{2} E\left[\nabla^\top \log p(x_t) \cdot \nabla^2 f(x_t) \cdot \nabla \log p(x_t)\right] + E\left[\tau \xi^\top \nabla^2 f(x_t)\xi\right] \end{aligned} \tag{1.61}$$

The second term:

$$\begin{aligned}&\tau E\left[\nabla^\top f \nabla \log p_t\right]\\ =&\tau\int \nabla f\cdot\nabla\log p_t p_t dx = \tau\int\nabla f\cdot\nabla p_t dx\\ =&-\tau\int \operatorname{tr}\left(\nabla^2 f\right)\cdot p_t dx = -\tau E\left[\operatorname{tr}\left(\nabla^2 f\right)\right]\\ =&-\tau E\left[\xi^\top\nabla^2 f\xi\right]\end{aligned} \tag{1.62}$$

Then

$$\mu_{t+\tau} = E\left[\frac{1}{2}\nabla^\top \log p_t \nabla^2 f \nabla \log p_t\right]\cdot\tau^2 = O\left(\tau^2\right) \tag{1.63}$$

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:*

$$\begin{cases} N(x;\mu,\Sigma) = \dfrac{1}{\sqrt{(2\pi)^d|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)}\\ \log N(x;\mu,\Sigma) = -\dfrac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) - \log\left(\sqrt{(2\pi)^d|\Sigma|}\right).\\ \nabla_x N(x;\mu,\Sigma) = N(x;\mu,\Sigma)\Sigma^{-1}(x-\mu)\\ \nabla_x \log N(x;\mu,\Sigma) = -\Sigma^{-1}(x-\mu).\end{cases} \tag{1.64}$$

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

1.3 What is Diffusion After All?

1.3.1 From SDEs

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

$$dX_t = f(X_t,t)dt + G(X_t,t)dW_t \tag{1.65}$$

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

$$dX_t = G(X_t,t)dW_t \tag{1.66}$$

Then by FPK equation, we can derive

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

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}\left[\left(GQG^\top\right)_{ij}p(x,t)\right] = \frac{1}{2}\nabla\cdot\left(\nabla\cdot(GQG^T p(x,t))\right) \tag{1.67}$$

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

$$\frac{\partial p}{\partial t} = \nabla\cdot\left(\frac{GG^T}{2}\nabla p\right) \tag{1.68}$$

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

1.3.2 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:

$$\begin{cases} \dfrac{\partial \phi_t(\mathbf{x})}{\partial t} = V(\phi_t(\mathbf{x}), t) \\ \phi_0(\mathbf{x}) = \mathbf{x} \end{cases} \tag{1.69}$$

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

Theorem 18 *When the initial density $p_0(x)$ is known, the density field can be expressed as:*

$$p(\phi_t(x), t) = \frac{p_0(x)}{|\det J_{\phi_t}(x)|} \tag{1.70}$$

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

$$\phi_t(x) \sim p_t(x) \tag{1.71}$$

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

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

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (V(x,t)p(x,t)) \tag{1.72}$$

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

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

Theorem 20 *When the incompressible condition is satisfied, that is $\nabla \cdot V = 0$, then the flow $\phi_t(x)$ is* ***measure preserving****, that is:*

$$|\det J_{\phi_t}(x)| = 1, i.e. p(\phi_t(x), t) = p_0(x) \tag{1.73}$$

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:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (\mathcal{F}(x,t)) \tag{1.74}$$

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

Theorem 21 (Fick's Law) *Fick's Law describes the relationship between the flux $\mathcal{F}(x,t)$ of the particle and the concentration/density $p(x,t)$.:*

$$\mathcal{F}(x,t) = -\frac{1}{2}\nabla \cdot \Big(GQG^T p(x,t)\Big) \tag{1.75}$$

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

$$\mathcal{F}(x,t) = -\frac{GG^T}{2}\nabla p(x,t) \tag{1.76}$$

Then

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot \left(\frac{GG^T}{2}\nabla p(x,t)\right) \tag{1.77}$$

1.3.3 Solution

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

$$\frac{\partial p(x,t)}{\partial t} = \nabla \cdot (\mathcal{D}\nabla p(x,t)) \tag{1.78}$$

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

Theorem 22 *The solution to the diffusion equation is:*

$$\begin{aligned} p(x,t) &= \mathscr{F}^{-1}\left[\tilde{p}_0(\lambda)\exp\left(-\lambda^T\mathcal{D}\lambda t\right)\right] = \left(p_0 \star \mathcal{G}_{2t\mathcal{D}}\right)(x) \\ &= \frac{1}{\sqrt{(4\pi t)^d \det(\mathcal{D})}} \int_{\mathcal{R}^d} \left(p_0(\xi)\exp\left(-\frac{1}{4t}(x-\xi)^T\mathcal{D}^{-1}(x-\xi)\right)\right) d\xi \end{aligned} \tag{1.79}$$

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

$$\begin{cases} \tilde{p}(x,t) = \mathscr{F}\left[p(x,t)\right] = \int_{\mathcal{R}^d} p(x,t)e^{-i\lambda\cdot x}dx \\ p(x,t) = \mathscr{F}^{-1}\left[\tilde{p}(x,t)\right] = \frac{1}{(2\pi)^d}\int_{\mathcal{R}^d} \tilde{p}(x,t)e^{i\lambda\cdot x}dx \end{cases} \tag{1.80}$$

Then, we have:

$$\begin{cases} \mathscr{F}\left[\nabla\cdot\mathbf{v}\right] = i\lambda\cdot\mathscr{F}\left[\mathbf{v}\right] \\ \mathscr{F}\left[\mathcal{D}\nabla p\right] = i\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{cases} \tag{1.81}$$

Then,

$$\begin{aligned} \mathscr{F}\left[\frac{\partial p}{\partial t}\right] &= \frac{d}{dt}\mathscr{F}\left[p\right] = \mathscr{F}\left[\nabla\cdot(\mathcal{D}\nabla p)\right] \\ =& i\lambda\cdot\mathscr{F}\left[\mathcal{D}\nabla p\right] = -\lambda^T\mathcal{D}\lambda\mathscr{F}\left[p\right] \end{aligned} \tag{1.82}$$

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:*

$$p(x,t) = \frac{1}{\sqrt{(4\pi t)^d \det\mathcal{D}}}\exp\left(-\frac{(x-x_0)^T\mathcal{D}^{-1}(x-x_0)}{4t}\right) \sim N(x_0, 2t\mathcal{D}) \tag{1.83}$$

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

$$p(x,t) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma+2t\mathcal{D})}}\exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma+2t\mathcal{D})^{-1}(x-\mu)\right) \sim N(\mu,\Sigma+2t\mathcal{D}) \tag{1.84}$$

(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.

1.4 Reversible Diffusions

1.4.1 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 23 *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 (1.1), the generator operator $\mathcal{L}$ and its self adjoint operator $\mathcal{L}^*$ are given by:

$$\begin{cases} \mathcal{L}(\cdot) = \sum_i \frac{\partial(\cdot)}{\partial x_i} f_i(x,t) + \frac{1}{2}\sum_{i,j} \left(\frac{\partial^2(\cdot)}{\partial x_i \partial x_j}\right) \left[GQG^\top\right]_{ij} \\ \qquad = f \cdot \nabla(\cdot) + \frac{1}{2}\left(M : \nabla \cdot \nabla\right)(\cdot) \\ \mathcal{L}^*(\cdot) = -\sum_i \frac{\partial}{\partial x_i}\left(f_i(x,t)(\cdot)\right) + \frac{1}{2}\sum_{i=1} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left[GQG^\top\right]_{ij}(\cdot)\right) \\ \qquad = \nabla \cdot (-f(\cdot) + \frac{1}{2}\nabla \cdot (M(\cdot))) \end{cases} \tag{1.85}$$

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

$$\mathcal{L}^* \rho_s = 0 \tag{1.86}$$

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

$$\rho_s = e^{-\Phi} \tag{1.87}$$

where Φ is a potential function.

Theorem 24 *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:*

$$f = \frac{1}{2}\rho_s^{-1}\nabla \cdot (M\rho_s) \tag{1.88}$$

Consider now an arbitrary ergodic diffusion process X_t, the solution of (1.1) 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 25 *The generator of an arbitrary diffusion process in $\mathbb{R}^d$ can we written as:*

$$\mathcal{L} = \rho_s^{-1} J_s \cdot \nabla + \frac{1}{2}\rho_s^{-1}\nabla \cdot (M\rho_s\nabla) := \mathcal{S} + \mathcal{A} \tag{1.89}$$

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*

$$f = \rho^{-1}\nabla\rho = \nabla \log \rho \tag{1.90}$$

which is the form of Langevin equation.

1.4.2 Schrödinger Operator

...

Chapter 2

Generative Model

2.1 Flow and Diffusion Model

As discussed before, the SDE and ODE are strong related through the Fokker-Planck equation. In this section, we will discuss how to use the SDE and ODE to generate data. [4]

2.1.1 The training targets

We constructed flow and diffusion models where we obtain trajectories $(X_t)_{0\leq t\leq 1}$ by simulating the ODE/SDE

$$X_0 \sim p_{\text{init}}, \quad \mathrm{d}X_t = u_t^\theta(X_t)\,\mathrm{d}t \quad \text{(Flow model)} \tag{10}$$

$$X_0 \sim p_{\text{init}}, \quad \mathrm{d}X_t = u_t^\theta(X_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t \quad \text{(Diffusion model)} \tag{11}$$

where u_t^θ is a neural network and σ_t is a fixed diffusion coefficient. Naturally, if we just randomly initialize the parameters θ of our neural network u_t^θ, simulating the ODE/SDE will just produce nonsense. As always in machine learning, we need to train the neural network. We accomplish this by minimizing a loss function $\mathcal{L}(\theta)$, such as the mean-squared error:

$$\mathcal{L}(\theta) = \left\| u_t^\theta(x) - \underbrace{u_t^{\text{target}}(x)}_{\text{training target}} \right\|^2,$$

where $u_t^{\text{target}}(x)$ is the training target that we would like to approximate. To derive a training algorithm, we proceed in two steps: In this chapter, our goal is to **find an equation for the training target** u_t^{target}. Naturally, like the neural network u_t^θ, the training target should itself be a vector field $u_t^{\text{target}} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. Further, u_t^{target} should do what we want u_t^θ to do: convert noise into data. **Therefore, the goal of this chapter is to derive a formula for the training target $u_t^{\mathbf{ref}}$ such that the corresponding ODE/SDE converts $p_{\mathbf{init}}$ into $p_{\mathbf{data}}$.** Along the way we will encounter two fundamental results from physics and stochastic calculus: the continuity equation and the Fokker-Planck equation, which we have discussed in the previous chapter. Next, we will derive the conditional and marginal probability path, vector fields and score functions.

For a data point $z \in \mathbb{R}^d$, we have the Dirac delta distribution δ_z. Then we can construct a conditional probability path: a set of distributions $p_t(x|z)$ over $x \in \mathbb{R}^d$ for $t \in [0,1]$ s.t.

$$p_0(\cdot|z) = p_{init}, p_1(\cdot|z) = \delta_z, \qquad \forall z \in \mathbb{R}^d \tag{2.1}$$

In other words, a conditional probability path gradually converts a single data point into the distribution p_{init}. You can think of a probability path as a trajectory in the space of distributions. Every conditional probability path $p_t(x|z)$ induces a marginal probability path $p_t(x)$ defined as the distribution that we obtain by first sampling a data point $z \sim p_{\text{data}}$ from the data distribution and then sampling from $p_t(\cdot|z)$:

$$z \sim p_{\text{data}}, \quad x \sim p_t(\cdot|z) \quad \Rightarrow \quad x \sim p_t, \quad p_t(x) = \int p_t(x|z) p_{\text{data}}(z)\,\mathrm{d}z \tag{2.2}$$

Note that we know how to sample from p_t but we don't know the density values $p_t(x)$ as the integral is intractable. Check for yourself that because of the conditions on $p_t(\cdot|z)$, the marginal probability path p_t interpolates between p_{init} and p_{data}:

$$p_0 = p_{\text{init}} \quad \text{and} \quad p_1 = p_{\text{data}}. \tag{2.3}$$

Normally, we choose p_{init} as a Gaussian distribution.

For every data point $z \in \mathbb{R}^d$, let $u_t^{\text{target}}(\cdot|z)$ denote a conditional vector field, defined so that the corresponding ODE yields the conditional probability path $p_t(\cdot|z)$,

$$X_0 \sim p_{\text{init}}, \quad \frac{d}{dt}X_t = u_t^{\text{target}}(X_t|z) \quad \Rightarrow \quad X_t \sim p_t(\cdot|z), \quad (0 \leq t \leq 1). \tag{2.4}$$

Then the marginal vector field $u_t^{\text{target}}(x)$ is defined by

$$u_t^{\text{target}}(x) = \int u_t^{\text{target}}(x|z)\frac{p_t(x|z)\, p_{\text{data}}(z)}{p_t(x)}, dz. \tag{2.5}$$

follows the marginal probability path, i.e.

$$X_0 \sim p_{\text{init}}, \quad \frac{d}{dt}X_t = u_t^{\text{target}}(X_t) \quad \Rightarrow \quad X_t \sim p_t \quad (0 \leq t \leq 1). \tag{2.6}$$

In particular, $X_1 \sim p_{\text{data}}$ for this ODE, so that we might say "u_t^{target} converts noise p_{init} into data p_{data}."

Similarly, we can express the marginal score function via conditional score function:

$$\begin{aligned} \nabla \log p_t(x) &= \frac{\nabla p_t(x)}{p_t(x)} = \frac{\nabla \int p_t(x|z)\, p_{\text{data}}(z)\, \mathrm{d}z}{p_t(x)} \\ &= \frac{\int \nabla p_t(x|z)\, p_{\text{data}}(z)\, \mathrm{d}z}{p_t(x)} \\ &= \int \nabla \log p_t(x|z)\, \frac{p_t(x|z)\, p_{\text{data}}(z)}{p_t(x)}\, \mathrm{d}z \end{aligned} \tag{2.7}$$

Sum up:

Conditional	Notion	Gaussian Exp.
Prob. Path	$p_t(\cdot\|z)$	$N(\alpha_t z, \beta_t^2 I_d)$
Vector Field	$u_t^{\text{target}}(x\|z)$	$(\dot{\alpha}_t - \frac{\dot{\beta}_t}{\beta_t})z + \frac{\dot{\beta}_t}{\beta_t}x$
Score Function	$\nabla \log p_t(x\|z)$	$-\frac{x-\alpha_t z}{\beta_t^2}$

Table 2.1: Conditional

Marginal	Notion	Gaussian Exp.
Prob. Path	$p_t(\cdot)$	$p_t(x) = \int p_t(x\|z) p_{\text{data}}(z)$
Vector Field	$u_t^{\text{target}}(x)$	$\int u_t^{\text{target}}(x\|z)\frac{p_t(x\|z)\, p_{\text{data}}(z)}{p_t(x)}$
Score Function	$\nabla \log p_t(x)$	$\int \nabla \log p_t(x\|z)\, \frac{p_t(x\|z)\, p_{\text{data}}(z)}{p_t(x)}$

Table 2.2: Marginal

In the next section, we will discuss how to train the model. First, we restrict ourselves to ODEs again, in doing so recovering flow matching. Second, we explain how to extend the approach to SDEs via score matching. Finally, we consider the special case of Gaussian probability paths, in doing so recovering denoising diffusion models. With these tools, we will at last have an end-to-end procedure to train and sample from a generative model with ODEs and SDEs.

2.1.2 Flow Matching

As before, let us consider a flow model given by

$$X_0 \sim p_{\text{init}}, \quad \mathrm{d}X_t = u_t^\theta(X_t)\,\mathrm{d}t. \tag{2.8}$$

As we learned, we want the neural network u_t^θ to equal the marginal vector field u_t^{target}. In other words, we would like to find parameters θ so that $u_t^\theta \approx u_t^{\text{target}}$. In the following, we denote by Unif $=$ Unif$[0,1]$ the uniform distribution on the interval $[0,1]$, and by $\mathbb{E}$ the expected value of a random variable. An intuitive way of obtaining $u_t^\theta \approx u_t^{\text{target}}$ is to use a mean-squared error, i.e. to use the flow matching loss defined as

$$\begin{aligned}\mathcal{L}_{\text{FM}}(\theta) &= \mathbb{E}_{t\sim\text{Unif},\, x\sim p_t}\left[\left\|u_t^\theta(x) - u_t^{\text{target}}(x)\right\|^2\right] \\ &= \mathbb{E}_{t\sim\text{Unif},\, z\sim p_{\text{data}},\, x\sim p_t(\cdot|z)}\left[\left\|u_t^\theta(x) - u_t^{\text{target}}(x)\right\|^2\right]\end{aligned} \tag{2.9}$$

where $p_t(x) = \int p_t(x|z)\,p_{\text{data}}(z)\,\mathrm{d}z$ is the marginal probability path and in the second line we used the sampling procedure given by equation (13). Intuitively, this loss says: First, draw a random time $t \in [0,1]$. Second, draw a random point z from our data set, sample from $p_t(\cdot|z)$ (e.g., by adding some noise), and compute $u_t^\theta(x)$. Finally, compute the mean-squared error between the output of our neural network and the marginal vector field $u_t^{\text{target}}(x)$. Unfortunately, we are not done here. While we do know the formula for u_t^{target} by:

$$u_t^{\text{target}}(x) = \int u_t^{\text{target}}(x|z)\frac{p_t(x|z)\,p_{\text{data}}(z)}{p_t(x)}\,\mathrm{d}z, \tag{2.10}$$

we cannot compute it efficiently because the above integral is intractable. Instead, we will exploit the fact that the conditional velocity field $u_t^{\text{target}}(x|z)$ is tractable. To do so, let us define the conditional flow matching loss

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{z\sim p_{\text{data}},\, x\sim p_t(\cdot|z)}\left[\left\|u_t^\theta(x) - u_t^{\text{target}}(x|z)\right\|^2\right]. \tag{2.11}$$

Note the difference to equation above: we use the conditional vector field $u_t^{\text{target}}(x|z)$ instead of the marginal vector $u_t^{\text{target}}(x)$. As we have an analytical formula for $u_t^{\text{target}}(x|z)$, we can minimize the above loss easily. But wait, what sense does it make to regress against the conditional vector field if it's the marginal vector field we care about? As it turns out, by explicitly regressing against the tractable conditional vector field, we are implicitly regressing against the intractable, marginal vector field. The next result makes this intuition precise.

Theorem 26 *The marginal flow matching loss equals the conditional flow matching loss up to a constant. That is,*

$$\mathcal{L}_{FM}(\theta) = \mathcal{L}_{CFM}(\theta) + C, \tag{2.12}$$

where C is independent of θ. Therefore, their gradients coincide:

$$\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta). \tag{2.13}$$

Hence, minimizing $\mathcal{L}_{CFM}(\theta)$ with e.g., stochastic gradient descent (SGD) is equivalent to minimizing $\mathcal{L}_{FM}(\theta)$ in the same fashion. In particular, for the minimizer θ^* of $\mathcal{L}_{CFM}(\theta)$, it will hold that $u_t^{\theta^*} = u_t^{target}$ (assuming an infinitely expressive parameterization).

Proof 10 *The proof works by expanding the mean-squared error into three components and removing constants:*

$$\begin{aligned}\mathcal{L}_{FM}(\theta) &= \mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[\left\|u_t^\theta(x) - u_t^{target}(x)\right\|^2\right]\\ &= \mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x)\|^2 - 2u_t^\theta(x)^T u_t^{target}(x) + \|u_t^{target}(x)\|^2\right]\\ &= \mathbb{E}_{t\sim\,Unif, x\sim p_t}\left[\|u_t^\theta(x)\|^2\right] - 2\mathbb{E}_{t\sim\,Unif, x\sim p_t}\left[u_t^\theta(x)^T u_t^{target}(x)\right] + \mathbb{E}_{t\sim\,Unif[0,1], x\sim p_t}\left[\|u_t^{target}(x)\|^2\right]\\ &= \mathbb{E}_{t\sim\,Unif, x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x)\|^2\right] - 2\mathbb{E}_{t\sim\,Unif, x\sim p_t}\left[u_t^\theta(x)^T u_t^{target}(x)\right] + C_1\end{aligned} \tag{2.14}$$

Let us reexpress the second summand:

$$\begin{aligned}&\mathbb{E}_{t\sim\,Unif, x\sim p_t}\left[u_t^\theta(x)^T u_t^{target}(x)\right]\\ &= \int_0^1 \int p_t(x) u_t^\theta(x)^T u_t^{target}(x)\,dx\,dt\\ &= \int_0^1 \int u_t^\theta(x)^T \left[\int u_t^{target}(x|z)\frac{p_t(x|z)p_{data}(z)}{p_t(x)}dz\right] p_t(x)dx\,dt\\ &= \int_0^1 \int\int u_t^\theta(x)^T u_t^{target}(x|z)p_t(x|z)p_{data}(z)dzdx\,dt\\ &= \mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[u_t^\theta(x)^T u_t^{target}(x|z)\right]\end{aligned} \tag{2.15}$$

Plug this into the expression for $\mathcal{L}_{FM}$:

$$\begin{aligned}\mathcal{L}_{FM}(\theta) =&\mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x)\|^2 - 2u_t^\theta(x)^T u_t^{target}(x|z)\right] + C_1\\ =&\mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x)\|^2 - 2u_t^\theta(x)^T u_t^{target}(x|z) + \|u_t^{target}(x|z)\|^2\right]\\ &+ \mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[-\|u_t^{target}(x|z)\|^2\right] + C_1\\ =&\mathbb{E}_{t\sim\,Unif, z\sim p_{data}, x\sim p_t(\cdot|z)}\left[\|u_t^\theta(x) - u_t^{target}(x|z)\|^2\right] + C_2 + C_1\\ =&\mathcal{L}_{CFM}(\theta) + C\end{aligned} \tag{2.16}$$

This finishes the proof.

2.1.3 Score Matching Diffusion

Let us extend the algorithm we just found from ODEs to SDEs. Remember we can extend the target ODE to an SDE with the same marginal distribution given by

$$\begin{aligned}&\mathrm{d}X_t = \left[u_t^{\text{target}}(X_t) + \frac{\sigma_t^2}{2}\nabla\log p_t(X_t)\right]\mathrm{d}t + \sigma_t \mathrm{d}W_t\\ &X_0 \sim p_{\text{init}} \Rightarrow \quad X_t \sim p_t \quad (0 \le t \le 1)\end{aligned} \tag{2.17}$$

where u_t^{target} is the marginal vector field and $\nabla\log p_t(x)$ is the marginal score function represented via the formula

$$\nabla\log p_t(x) = \int \nabla\log p_t(x|z)\frac{p_t(x|z)\,p_{\text{data}}(z)}{p_t(x)}\,\mathrm{d}z \tag{2.18}$$

To approximate the marginal score $\nabla\log p_t$, we can use a neural network that we call score network $s_t^\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. In the same way as before, we can design a score matching loss and a conditional score matching loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{SM}}(\theta) &= \mathbb{E}_{t\sim\mathrm{Unif},\, z\sim p_{\mathrm{data}},\, x\sim p_t(\cdot|z)}\left[\|s_t^\theta(x) - \nabla\log p_t(x)\|^2\right]\\ \mathcal{L}_{\mathrm{CSM}}(\theta) &= \mathbb{E}_{t\sim\mathrm{Unif},\, z\sim p_{\mathrm{data}},\, x\sim p_t(\cdot|z)}\left[\|s_t^\theta(x) - \nabla\log p_t(x|z)\|^2\right]\end{aligned} \tag{2.19}$$

where again the difference is using the marginal score $\nabla\log p_t(x)$ vs. using the conditional score $\nabla\log p_t(x|z)$. As before, we ideally would want to minimize the score matching loss but can't because we don't know $\nabla\log p_t(x)$. But similarly as before, the conditional score matching loss is a tractable alternative:

Theorem 27 *The score matching loss equals the conditional score matching loss up to a constant:*

$$\mathcal{L}_{SM}(\theta) = \mathcal{L}_{CSM}(\theta) + C, \tag{2.20}$$

where C is independent of parameters θ. Therefore, their gradients coincide:

$$\nabla_\theta\mathcal{L}_{SM}(\theta) = \nabla_\theta\mathcal{L}_{CSM}(\theta). \tag{2.21}$$

In particular, for the minimizer θ^*, it will hold that $s_t^{\theta^*} = \nabla\log p_t$.

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:

$$p(x_i|x_{i-1}) = \mathcal{N}(x_i|\sqrt{1-\beta_i}x_{i-1}, \beta_i I) \tag{2.22}$$

Then we can get

$$p_{\alpha_i}(x_i|x_0) = \mathcal{N}(x_i|\sqrt{\alpha_i}x_0, (1-\alpha_i)I), \alpha_i = \Pi_{j=1}^i(1-\beta_j) \tag{2.23}$$

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:

$$p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}|\mu_\theta(x_i, i), \Sigma_\theta(x_i, i)) \tag{2.24}$$

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:

$$\begin{aligned}&KL(q(x_0, x_1, \cdots, x_N)||p_\theta(x_0, x_1, \cdots, x_N))\\ =&-\mathbf{E}_q\left[\log p_\theta(x_0, x_1, \cdots, x_N)\right] + const\\ =&-\mathbf{E}_q\left[\log p(x_N) + \sum_{i=1}^N \log\frac{p_\theta(x_{i-1}|x_i)}{q(x_i|x_{i-1})}\right]\end{aligned} \tag{2.25}$$

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

Score SDEs We consider linear SDE having the form:

$$dX_t = (a(t)X_t + b(t))dt + g(t)dW_t \tag{2.26}$$

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

$$\begin{aligned} X_{t+s} &= X_t + (a(t)X_t + b(t))s + g(t)\sqrt{sQ}\xi \\ &= (1 + a(t)s)X_t + b(t)s + g(t)\sqrt{sQ}\xi \end{aligned} \tag{2.27}$$

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

$$\begin{aligned} E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right] &\approx \left(a(t)E\left[X_t|X_t\right] + b(t)\right)s + g(t)\sqrt{sQ}E[\xi] \\ &= \left(a(t)E\left[X_t|X_0\right] + b(t)\right)s. \end{aligned} \tag{2.28}$$

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

$$e'(t) = \lim_{s\to 0} \frac{E\left[X_{t+s}|X_0\right] - E\left[X_t|X_0\right]}{s} = a(t)\cdot e(t) + b(t). \quad e(0) = X_0. \tag{2.29}$$

which is an ODE system, having solution

$$e(t) = e^{\int_0^t a(s)ds} \cdot \left(X_0 + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right) \tag{2.30}$$

Therefore

$$\begin{aligned} E\left[X_t\right] &= E\left[E\left[X_t|X_0\right]\right] = E[e(t)] \\ &= e^{\int_0^t a(s)ds} \cdot \left(E\left[X_0\right] + \int_0^t e^{-\int_0^s a(r)dr} b(s)ds\right) \end{aligned} \tag{2.31}$$

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

$$\begin{aligned} V'(t) &= \lim_{s\to 0} \frac{\operatorname{Var}\left(X_{t+s}|X_0\right) - \operatorname{Var}\left(X_t|X_0\right)}{s} \\ &= \left[\left(a^2(t)s + 2a(t)\right)v(t) + g^2(t)\right]|_{s\to 0} \\ &= 2\alpha(t)V(t) + g(t)Qg^\top(t), \qquad V(0) = 0 \end{aligned} \tag{2.32}$$

Solution is:

$$v(t) = e^{\int_0^t 2a(s)ds} \cdot \left(\int_0^t e^{-\int_0^s 2a(r)dr} g(s)Qg^\top(s)ds\right) \tag{2.33}$$

By law of total variance:

$$\begin{aligned} \operatorname{Var}\left(X_t\right) =& E\left[X_t^2\right] - E^2\left[X_t\right] = E\left[E\left[X_t^2|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right) + E^2\left[X_t|X_0\right]\right] - E^2\left[X_t\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + E\left[E^2\left[X_t|X_0\right]\right] - E^2\left[E\left[X_t|X_0\right]\right] \\ =& E\left[\operatorname{Var}\left(X_t|X_0\right)\right] + \operatorname{Var}\left(E\left[X_t|X_0\right]\right) \end{aligned} \tag{2.34}$$

then

$$\begin{aligned} \operatorname{Var}(X_t) =& E[V(t)] + \operatorname{Var}(e(t)) \\ =& e^{\int_0^t 2a(s)ds} \cdot \left(\int_0^t e^{-\int_0^s 2a(r)dr} g(s)Qg^\top(s)ds\right) + e^{\int_0^t 2a(s)ds} \cdot \operatorname{Var}\left(X_0\right). \end{aligned} \tag{2.35}$$

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

Theorem 28 *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:

$$\begin{aligned} X_{t_{i+1}} &= \sqrt{1-(\beta(t_{i+1})-\beta(t_i))}X_{t_i} + \sqrt{(\beta(t_{i+1})-\beta(t_i))}\xi \\ &= \sqrt{1-\Delta\beta(t_i)}X_{t_i} + \sqrt{\Delta\beta(t_i)}\xi \end{aligned} \tag{2.36}$$

Then the conditional distribution is given by:

$$q\left(X_{t_{i+1}}|X_{t_i}\right) = N(x_{t_{i+1}}; \sqrt{1-\Delta\beta(t_i)}X_{t_i}, \Delta\beta(t_i)) \tag{2.37}$$

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

$$\begin{aligned} f(x,t) &= \lim_{h\to 0}\frac{E\left[X_{t+h}-X_t|X_t=x\right]}{h} \\ &= \lim_{h\to 0}\frac{x\sqrt{1-\Delta\beta(t)}-x}{h} = -\frac{x}{2}\beta'(t). \\ g(t) &= \sqrt{\lim_{h\to 0}\frac{N\left[X_{t+h}|X_t=x\right]}{h}} = \sqrt{\lim_{h\to 0}\frac{\beta(t+h)-\beta(t)}{h}} = \sqrt{\beta'(t)} \end{aligned} \tag{2.38}$$

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

Then we have

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{\int_0^t -\frac{1}{2}\beta'(s)ds} = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right]e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s \beta'(r)dr}\beta'(s)ds \cdot e^{-\beta(t)} = 1-e^{-\beta(t)} \\ V\left(X_t\right) = 1-e^{-\beta(t)} + V\left(X_0\right)e^{-\beta(t)} = 1+\left(V\left(X_0\right)-1\right)e^{-\beta(t)}. \end{cases} \tag{2.39}$$

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

$$\begin{cases} E\left[X_0|X_0\right] = X_0 \\ E\left[X_t\right] = E\left[X_0\right] \\ V\left(X_t|X_0\right) = \int_0^t e^{\int_0^s 0dr}\beta'(s)ds = \beta(t) \\ V\left(X_t\right) = V\left[X_0\right] + \beta(t) \end{cases} \tag{2.40}$$

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

$$\begin{aligned} f(x,t) &= -\frac{1}{2}\beta'(t) \\ g(t) &= \sqrt{\beta'(t)\left(1-e^{-2\beta(t)}\right)} \end{aligned} \tag{2.41}$$

As the same, we can compute that.

$$\begin{cases} E\left[X_t|X_0\right] = X_0 e^{-\frac{1}{2}\beta(t)} \\ E\left[X_t\right] = E\left[X_0\right] e^{-\frac{1}{2}\beta(t)} \\ V\left(X_t|X_0\right) = \left(1-e^{-\beta(t)}\right)^2 \\ V\left(X_t\right) = \left(1-e^{-\beta(t)}\right)^2 + V\left(X_t\right) e^{-\beta(t)}. \end{cases} \tag{2.42}$$

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 $X_0 \sim p_0, X_T \sim p_T$,

$$dX_t = f(X_t, t)dt + G(t)dW_t \tag{2.43}$$

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

$$d\bar{X}_t = \bar{f}(\bar{X}_t, t)dt + \bar{G}(t)d\bar{W}_t \tag{2.44}$$

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:

$$\bar{X}_{t-s} - \bar{X}_t = -s\bar{f}(\bar{X}_t, t) + \sqrt{s}\bar{G}(t)\xi \tag{2.45}$$

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 29 *The reverse SDE with $\bar{f}, \bar{G}$ having the following form has the same distribution as the forward SDE 2.43:*

$$\begin{cases} \bar{f}(x,t) = f(x,t) - GG^T \nabla_x \log p_t(x) \\ \bar{G} = G(t) \end{cases} \tag{2.46}$$

i.e.

$$d\bar{X}_t = \left[f(\bar{X}_t, t) - GG^T \nabla_x \log p_t(x_t)\right] dt + G(t)d\bar{W}_t \tag{2.47}$$

Proof 11 *The proof is skipped.*

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

Algorithm 4 *:*
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:

$$\bar{X}_{t-s} = \bar{X}_t + s\left[g^2(t)s_\theta(\bar{X}_t, t) - f(\bar{X}_t, t)\right] + \sqrt{s}g(t)\xi \tag{2.48}$$

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 $p(x)$:

$$q(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i) \tag{2.49}$$

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:

$$\begin{aligned}
\mathcal{L}_\theta =& \mathbf{E}_{x\sim p(x)}\left[\|s_\theta(x) - \nabla_x \log p(x)\|^2\right] \\
\approx& \mathbf{E}_{x\sim q(x)}\left[\|s_\theta(x) - \nabla_x \log q(x)\|^2\right] \\
=& \int \|s_\theta(x) - \nabla_x \log q(x)\|^2 q(x)dx \\
\approx& \frac{1}{n}\sum_{i=1}^{n}\int \|s_\theta(x) - \nabla_x \log q(x)\|^2 K(x-x_i)dx
\end{aligned} \tag{2.50}$$

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:

$$\begin{aligned}
L_\theta &= \frac{1}{T}\int_0^T \lambda(t) \underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right] dt \\
&= \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\left[\|s_\theta(x_t,t) - \nabla_{x_t}\log p_t(x_t)\|^2\right]\right]\right]
\end{aligned} \tag{2.51}$$

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

$$\begin{aligned}
\nabla \log p_t(x) &= \frac{1}{p_t(x)}\nabla p_t(x). \\
&= \frac{1}{p_t(x)}\nabla \int p_{t|0}(x|y)p_0(y)dy \\
&= \frac{1}{p_t(x)}\int \nabla p_{t|0}(x|y)p_0(y)dy \\
&= \frac{1}{p_t(x)}\int \frac{\nabla p_{t|0}(x|y)}{p_{t|0}(x|y)}p_0(y)\cdot p_{t|0}(x|y)dy \\
&= \int \nabla_x \log\left(p_{t|0}(x|y)\right)\cdot p_{0|t}(y|x)dy \\
&= \underset{y\sim p_{0|t}(y|x)}{E}\left[\nabla_x \log\left(p_{t|0}(x|y)\right)\right]
\end{aligned} \tag{2.52}$$

Where we have used the following lemma:

Lemma 1

$$\underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[\underset{x_0'\sim p_{0|t}(\cdot|x_t)}{E}\left[f(x_t,x_0')\right]\right]\right] = \underset{x_0\sim p_0}{E}\left[\underset{x_t\sim p_{t|0}(\cdot|x_0)}{E}\left[f(x_t,x_0)\right]\right] \tag{2.53}$$

Proof 12 *Easy to prove.*

Then we can rewrite the loss function as:

$$\begin{aligned} L_\theta = & \underset{t\sim U(0,T)}{E}\Bigg[\lambda(t)\underset{x_0\sim p_{data}}{E}\Bigg[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\Big[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log p_t\left(x_t\right)\right\|^2\Big] \\ \leqslant & \underset{t\sim U(0,T)}{E}\Bigg[\lambda(t)\underset{x_0\sim p_{data}}{E}\Bigg[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\Big[\underset{y\sim p_{data}}{E}\Big[\Big\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\Big(p_{t|0}(x_t|y)\Big)\Big\|^2\Big]\Big]\Bigg]\Bigg] \\ = & \underset{t\sim U(0,T)}{E}\Bigg[\lambda(t)\underset{x_0\sim p_{data}}{E}\Bigg[\underset{x_t\sim p_{t|0}(x_t|x_0)}{E}\Big[\left\|S_\theta\left(x_t,t\right)-\nabla_{x_t}\log\Big(p_{t|0}\left(x_t|x_0\right)\right\|^2\Big]\Big]\Bigg] \end{aligned} \tag{2.54}$$

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 28, 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:

$$\frac{\partial}{\partial x}\log p_{t|0}\left(x|x_0\right)=-\frac{x-E_{t|0}\left[x|x_0\right]}{\operatorname{Var}_{t|0}\left(x|x_0\right)}=-\frac{x-e(t,X_0)}{V(t)}\sim -N\left(0,\frac{1}{V(t)}\right) \tag{2.55}$$

So

$$\begin{aligned} L_\theta = & \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\underset{\xi\sim N(0,1)}{E}\left[\left\|s_\theta\left(\sqrt{V(t)}\xi+e(t,X_0),t\right)+\frac{\xi}{\sqrt{V(t)}}\right\|^2\right]\right]\right] \\ = & \underset{t\sim U(0,T)}{E}\left[\lambda(t)\underset{x_0\sim p_{data}}{E}\left[\frac{1}{V(t)}\underset{\xi\sim N(0,1)}{E}\left[\left\|\xi_\theta\left(\sqrt{V(t)}\xi+e(t,X_0),t\right)-\xi\right\|^2\right]\right]\right] \end{aligned} \tag{2.56}$$

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

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

$$\begin{aligned} & p_t(x\mid y)=p\left(X_t=x\mid y\right)=\frac{p\left(y\mid X_t=x\right)p\left(X_t=x\right)}{p(y)} \\ \Rightarrow & \log\left(p_t(x\mid y)\right)=\log\left(p\left(y\mid X_t=x\right)\right)+\log\left(p\left(X_t=x\right)\right)-\log(p(y)) \\ \Rightarrow & \nabla_x\log\left(p_t(x\mid y)\right)=\nabla_x\log\left(p\left(y\mid X_t=x\right)\right)+\nabla_x\log\left(p\left(X_t=x\right)\right) \end{aligned} \tag{2.57}$$

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: $p^{\prime}\left(y \mid X_t=x\right)$, then:

$$\nabla_x \log\left(p_t(x \mid y)\right)=S \nabla_x \log\left(p\left(y \mid x_t=x\right)\right)+\nabla_x \log\left(p\left(x_t=x\right)\right) \tag{2.58}$$

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

$$\begin{aligned} & \gamma \nabla_x \log\left(p\left(y \mid X_t=x\right)\right) \\ = & \gamma\left(\nabla_x \log\left(p(X_t=x \mid y)\right)-\nabla_x \log\left(p_t(x)\right)\right) \end{aligned} \tag{2.59}$$

Then

$$\begin{aligned} & \nabla_x \log_\gamma\left(p_t(x \mid y)\right) \\ = & (1-\gamma) \nabla_x \log\left(p_t(x)\right)+\gamma \nabla_x \log\left(p(X_t=x \mid y)\right) \end{aligned} \tag{2.60}$$

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

2.2 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.

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:*

$$\operatorname{ELBO}(x)=\mathbf{E}_{q_\phi(z \mid x)}\left[\log \frac{p(x, z)}{q_\phi(z \mid x)}\right] \tag{2.61}$$

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

$$\begin{aligned} \log p(x) & =\log \int p(x, z) d z \\ & =\log \int \frac{p(x, z)}{q_\phi(z \mid x)} \cdot q_\phi(z \mid x) d z \\ & \geq \mathbf{E}_{q_\phi(z \mid x)}\left[\log \frac{p(x, z)}{q_\phi(z \mid x)}\right]=\operatorname{ELBO}(x) \end{aligned} \tag{2.62}$$

Theorem 30 (Decomposition of Log-likelihood) *We have*

$$\log p(x)=\operatorname{ELBO}(x)+\operatorname{KL}(q_\phi(z \mid x) || p(z \mid x)) \tag{2.63}$$

then we can minimize the gap between $\log p(x)$ *and ELBO, and the equality hold if and only if* $q_\phi(z \mid x)=p(z \mid x)$.

Since $p(z \mid x)$ is a delta function,

Proof 13

$$\begin{aligned}\log p(x) &= \log p(x) \int q_\phi(z|x) dz \\ &= \mathbf{E}_{q_\phi(z|x)} \left[\log p(x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)} \left[\log \left(\frac{p(x,z)}{p(z|x)} \frac{q_\phi(z|x)}{q_\phi(z|x)}\right)\right] \\ &= \mathrm{ELBO}(x) + \mathrm{KL}(q_\phi(z|x)||p(z|x))\end{aligned} \tag{2.64}$$

Theorem 31 *Also, we can rewrite the ELBO as:*

$$\begin{aligned}\mathrm{ELBO}(x) &= \mathbf{E}_{q_\phi(z|x)} \left[\log p(x|z) + \log p(z) - \log q_\phi(z|x)\right] \\ &= \mathbf{E}_{q_\phi(z|x)} \left[\log p(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z)) \\ &= \mathbf{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \mathrm{KL}(q_\phi(z|x)||p(z))\end{aligned} \tag{2.65}$$

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:*

$$(\phi, \theta) = \mathrm{argmax}_{\phi,\theta} \sum_{x \in X} \mathrm{ELBO(x)} \tag{2.66}$$

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.

Chapter 3

Random Field

Here we discussed about the classical numerical methods and SPDE-based approaches for simulating the random field.

3.1 Random Field

3.1.1 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:*

$$\begin{cases} \mu(x) = \mathbf{E}[u(x)] \\ C(x,y) = Cov(u(x), u(y)) = \mathbf{E}[(u(x) - m(x))(u(y) - m(y))] \end{cases} \tag{3.1}$$

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

$$u = u[u(x_1), u(x_2), \cdots, u(x_n)]^T \sim \mathcal{N}(\mu(x), C(x,y)),\ \forall x_i \in D \tag{3.2}$$

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

$$\begin{aligned} \langle \mathscr{C}\phi, \psi \rangle &= Cov\left(\langle u, \phi \rangle_{L^2(D)}, \langle u, \psi \rangle_{L^2(D)}\right) \\ &= E\left[\left(\int_D \phi(x)(u(x) - \mu(x))dx\right)\left(\int_D \psi(y)(u(y) - \mu(y))dy\right)\right] \\ &= \int_D \int_D \phi(x)\psi(y) E[(u(x) - mu(x))(u(y) - mu(y))]dxdy \\ &= \int_D \int_D \phi(x)\psi(y) Cov(u(x), u(y))dxdy \end{aligned} \tag{3.3}$$

So that

$$(\mathscr{C}\phi)(x) = \int_D Cov(u(x), u(y))\phi(y)dy \tag{3.4}$$

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 32 (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*

$$c(x) = \int_{\mathbb{R}^d} e^{iv \cdot x} dF(v) = (2\pi)^{\frac{d}{2}} \hat{f}(x) \tag{3.5}$$

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.*

$$Cov(x) = c(\|x\|_2) = c^0(r) \tag{3.6}$$

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

3.1.2 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:*

$$x_k = x_{i,j} = (i\Delta x_1, j\Delta x_2)^T,\ i = 0, 1, \cdots, n_1 - 1,\ j = 0, 1, \cdots, n_2 - 1, k = i + jn_1 \tag{3.7}$$

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:

$$C_{kl} = Cov(u_k, u_l) = c(x_{i+jn_1} - x_{r+sn_1}) \tag{3.8}$$

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

Theorem 33 *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 5 *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:*

$$S = \begin{pmatrix} \tilde{C}_0 & \tilde{C}_1^T & \cdots & \tilde{C}_{n_2-1}^T \\ \tilde{C}_1 & \tilde{C}_2 & \cdots & \tilde{C}_{n_2-2}^T \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{C}_{n_2-1} & \tilde{C}_{n_2-2} & \cdots & \tilde{C}_0 \end{pmatrix}, \quad \tilde{C}_i = \begin{pmatrix} C_i & B_i \\ B_i & C_i \end{pmatrix} \tag{3.9}$$

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.

3.1.3 KL expansion of R.F.

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

$$(\mathscr{C}\phi)(x) = \int_D Cov(u(x), u(y))\phi(y)dy = \int_D c(x-y)\phi(y)dy \tag{3.10}$$

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 34 (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:*

$$u(x) = \mu(x) + \sum_{j=0}^{\infty} \sqrt{v_j}\phi_j(x)\xi_j \tag{3.11}$$

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

$$\xi_j = \frac{1}{\sqrt{v_j}} \int_D (u(x) - \mu(x))\phi_j(x)dx \tag{3.12}$$

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.

3.2 For Stationary RF on $\mathbb{R}^d$

First we define stationary random field on $\mathbb{R}^d$ as:

Definition 13 (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. $\mu(x) = \mu$, $C(x, y) = C(x - y)$.*

Then we can define the covariance operator $\mathcal{C}$ as:

$$\mathcal{C}\phi = \int_{\mathbb{R}^d} C(x-y)\phi(y)dy \tag{3.13}$$

We find that it is actually the convolution operator of $C(x)$ with $\phi(x)$.

Stationary random fields have some beautiful properties.

Theorem 35 (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*

$$c(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{ik\cdot x} dF(k) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{ik\cdot x} S(k)dk = \left(\mathcal{F}^{-1}S\right)(x) \tag{3.14}$$

where $F(k)$ is some measure on $\mathbb{R}^d$ called spectral distribution and $\hat{S}(x)$ is the Fourier transform of $S(k)$, called spectral density. Reversely, $S(k) = (\mathcal{F}c)(k) = \hat{c}(k)$. If $S(k)$ is non-negative and integrable, then $c(x)$ is a valid covariance function.

Theorem 36 (Spectral Density of Random Field) *Assume $u(x)$ has zero mean, then*

$$S_u(k) = \frac{1}{(2\pi)^d}\mathbb{E}[|\hat{u}(k)|^2] \tag{3.15}$$

By defining the pseudo-differential operators, the class of SPDEs is defined by:

$$\mathcal{L}_g u = W, \mathcal{L}_g = \mathcal{F}^{-1} g \mathcal{F} \tag{3.16}$$

where $g : \mathbb{R}^d \to \mathbb{C}$ must be a sufficiently regular and Hermitian-symmetric function, that is it must satisfy: $g(k) = \overline{g(-k)}$, $\bar{\cdot}$ denotes the complex conjugate. So if we have $\mathcal{L}_g u = W$, then:

$$u = \mathcal{L}_{\frac{1}{g}} W \tag{3.17}$$

Theorem 37 *The spectral density of $\mathcal{L}_g u$ and of u are related by:*

$$S_{\mathcal{L}_g u}(k) = |g(k)|^2 \, S_u(k) \tag{3.18}$$

Generally, if

$$\mathcal{L}_g u = w \tag{3.19}$$

where w is a GeRF source term, then $S_w(k) = |g(k)|^2 \, S_u(k)$. Therefore, when $w = W$, $S_u = \frac{1}{(2\pi)^d |g(k)|^2}\mathbb{E}[|\hat{W}(k)|^2] = \frac{1}{|g(k)|^2}$. Then,

$$u(x) = \mathcal{L}_{\frac{1}{g}} w(x) = \mathcal{L}_{\sqrt{\frac{S_u}{S_w}}} w(x) \tag{3.20}$$

Then consider the exitence of the function.

Theorem 38 *Let $w(x)$ be a real stationary GeRF over $\mathbb{R}^d$, and let $g : \mathbb{R}^d \to \mathbb{C}$ be a symbol function. Then for (3.19), there exists a unique stationary solution $u(x)$ if and only if: there exists $N \in \mathbb{N}$ s.t.*

$$\int_{\mathbb{R}^d} \frac{dS_w(k)}{|g(k)|^2 \, (1 + \|k\|^2)^N} < \infty \tag{3.21}$$

and

$$S_u(k) = |g(k)|^{-2} \, S_w(k) \tag{3.22}$$

Moreover, $S_u(k)$ is unique if and only if $|g| > 0$.

Hence the key is the symbol function $g(k)$. The following theorem shows that solutions of SPDEs with White Noise source term is the starting point of more general solutions, when the source term can be any stationary GeRF.

Theorem 39 *Let $w(x)$ be a real stationary GeRF over $\mathbb{R}^d$ with covariance distribution $C_w(x)$. Let g be a symbol function over $\mathbb{R}^d$ such that $\frac{1}{g}$ is smooth with polynomially bounded derivatives of all orders. Then, there exists a unique stationary solution to (3.19) and its covariance distribution is given by*

$$C_u(x) = C_u^W * C_w(x) \tag{3.23}$$

where C_u^W is the covariance function of the solution to the SPDE with White Noise source term.

Proof 14 *The proof is straightforward by using Wiener-Khinchin theorem.*

For any precision operator which is a polynomial in the Laplacian, $Q = p(-\Delta)$, such as the Matern operator with $\nu \in \mathbb{N}$, this results in a polynomial $F(Q) = p(\|k\|^2)$.

3.2.1 Matern Field

The important relationship that we will make use of is that a Gaussian field $u(x)$ with the Matern covariance is a solution to the linear fractional stochastic partial differential equation (SPDE):

$$\mathcal{L}^{\alpha/2}u(x) = (\kappa^2 - \Delta)^{\alpha/2}u(x) = W(x), \qquad x \in D \in \mathbb{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0, \tag{3.24}$$

where $\nu = \alpha - d/2, \rho = \frac{\sqrt{2\nu}}{\kappa}$ is the range parameter, Δ is the Laplacian operator, $W(x)$ is a spatial Gaussian white noise with unit variance. We will name any solution to Equ (3.24) a Matern field in the following.

Theorem 40 (Spectral Solution of Matern Field) *The solution of u solved by Equ (3.24) is given by:*

$$u(x) = \mathcal{F}^{-1}\left[\frac{\hat{W}(k)}{(\kappa^2 + \|k\|^2)^{\alpha/2}}\right](x) \tag{3.25}$$

where $\mathcal{F}$ is defined in (A.1). And the covariance function of u is given by:

$$c(x) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa\|x\|)^\nu K_\nu(\kappa\|x\|) \tag{3.26}$$

where $\nu = \alpha - d/2, \rho = \frac{\sqrt{2\nu}}{\kappa}, \sigma^2 = \frac{\Gamma(\nu)}{(4\pi)^{d/2}\kappa^{2\nu}\Gamma(\alpha)}$

Wiener-Khinchin theorem + Spectral Theorem.

3.2.2 Generalized Matern Field

Consider the following SPDE:

$$(\kappa^2 + (-\Delta)^\gamma)^{\alpha/2}u(x) = \mathcal{F}^{-1}\left((\kappa^2 + \|k\|^{2\gamma})^{\alpha/2}\mathcal{F}u\right)(x) = W(x), \quad x \in D \in \mathbb{R}^d \tag{3.27}$$

Hence the solution is:

$$u(x) = \mathcal{F}^{-1}\left[\frac{\hat{W}(k)}{(\kappa^2 + \|k\|^{2\gamma})^{\alpha/2}}\right](x) \tag{3.28}$$

Therefore the spectral density is:

$$S_u(k) = \frac{1}{(\kappa^2 + \|k\|^{2\gamma})^\alpha} \tag{3.29}$$

So when $\gamma = 1$, it becomes the Matern Field. Since the spectral density $S_u(k) \in L^2(\mathbb{R}^d)$ if and only if $\alpha\gamma > \frac{d}{2}$.

Generally, we can define the pseudo-differential operator through symbol function $g(k)$.

3.3 Spatial-Temporal General Random Field on $\mathbb{R}^d \times (0,T)$

3.3.1 Stein Model

Proposed in [8], we define the spatial-temporal white noise $\mathcal{W}(x,t)$ as Gaussian noise that is white in time but correlated in space.

$$\left(b(s^2 - \frac{d}{dt^2})^\beta + a(\kappa^2 - \Delta)^\alpha\right)^{\nu/2} u(x,t) = W, \qquad (x,t) \in D \times (0,T) \tag{3.30}$$

Consider when $D = \mathbb{R}^d$, where the space-time spectral density of the stationary solution is given by:

$$S_u(k_s, k_t) = \frac{1}{(b(s^2 + k_t^2)^\beta + a(\kappa^2 + k_s^2)^\alpha)^\nu} \tag{3.31}$$

We note the spatio-temporal symbol function as:

$$g(k_s, k_t) : (k_s, k_t) \to (b(s^2 + k_t^2)^\beta + a(\kappa^2 + k_s^2)^\alpha)^{\nu/2} \tag{3.32}$$

When α, β, ν are positive and $\frac{d}{\alpha\nu} + \frac{1}{\beta\nu} = 2$, [8] shows that the spectral density is finite and the corresponding random field is mean square continuous.

Theorem 41 *When $\kappa, s, a, b > 0$ and α, β, ν are not null, $g(k_s, k_t)$ satisfies Thm 39, then for any stationary GeRF X, the SPDE:*

$$\left(b(s^2 - \frac{d}{dt^2})^\beta + a(\kappa^2 - \Delta)^\alpha\right)^{\nu/2} U(x,t) = X(x,t) \tag{3.33}$$

has a unique stationary solution $U(x,t)$ with covariance function:

$$C_U(x, y, t, s) = C_U^W * C^X(x - y, t - s) \tag{3.34}$$

3.3.2 Evolution Equations Model

Here we consider the following model:

$$\frac{\partial^\beta u}{\partial t^\beta} + \mathcal{L}_g u = w(x,t) \tag{3.35}$$

where $\mathcal{L}_g$ is a pseudo-differential operator with symbol $g(k)$ and $w(x,t)$ is a stationary spatio-temporal GeRF.

$$g(k_s, k_t) = (ik_t)^\beta + g(k_s) \tag{3.36}$$

3.3.3 Advection-Diffusion SPDE

This is poeposed in [7]. The equation is given by:

$$\left[\frac{\partial}{\partial t} - \nabla \cdot (\Sigma \nabla) + \mu \nabla + C\right] u(x,t) = w(x,t) \tag{3.37}$$

where Σ is the diffusion matrix, μ is the advection velocity, C is the drift coefficient. Here we set the diffusion matrix as:

$$\Sigma = \frac{1}{\rho^2} \begin{pmatrix} \cos\theta & \sin\theta \\ -\gamma\sin\theta & \gamma\cos\theta \end{pmatrix}^T \begin{pmatrix} \cos\theta & \sin\theta \\ -\gamma\sin\theta & \gamma\cos\theta \end{pmatrix} \tag{3.38}$$

where ρ is the correlation length and γ is the anisotropy ratio, $\theta \in [0, \pi/2)$. With $\gamma = 1$, it becomes isotropic. Similarly the spectral density is given by:

$$\begin{aligned} S_u(k_s, k_t) &= \frac{S_w(k_s, k_t)}{|i(k_t + \mu k_s) + (C + k_s^T \Sigma k_s)|^2} \\ &= \frac{S_w(k_s, k_t)}{(k_t + \mu k_s)^2 + (C + k_s^T \Sigma k_s)^2} \\ &= \frac{S_w(k_s, k_t)}{|g_u|^2} \end{aligned} \tag{3.39}$$

By Wiener-Khinchin theorem, the covariance function is given by:

$$C_u(x,t) = \frac{1}{(2\pi)^d} \int S_w \frac{e^{-i\mu k_s t - (k_s^T \Sigma k_s + C)|t|}}{2(k_s^T \Sigma k_s + C)} e^{i k_s x} dk_s \tag{3.40}$$

Specifically, when $\mu = 0, \Sigma = 0$, the covariance function is given by:

$$C_u(x,t) = \frac{e^{-C|t|}}{2C} C_w(x,t) \tag{3.41}$$

However $\mu(x)$ may not be constant.

3.3.4 Generic class of non-stationary models

Similar to ADSPDE, we consider:

$$\frac{\partial u}{\partial t} + \left[-\nabla \cdot (\Sigma(x,t)\nabla) + \mu(x,t) \cdot \nabla + \kappa^2(x,t) \right]^{\alpha/2} u(x,t) = w(x,t) \tag{3.42}$$

where μ, Σ, κ are functions of x, t, and $w(x,t)$ is a GeRF driven by Equ (3.24).

Chapter 4

SDE in Function Space

Reference: [3, 6, 1]

4.1 Introduction

Definition 14 *Random field $\mathcal{M}(x,\omega)$, where $x \in D$ and $\omega \in \Omega$, is defined as:*

$$\begin{aligned}&\mathcal{M}(x,\cdot) \text{ is a random variable defined on the probability space } (\Omega,\mathcal{F},P),\\&\mathcal{M}(\cdot,\omega) \text{ is a deterministic function of } x.\end{aligned} \tag{4.1}$$

In fact, a random field is an infinite-dimensional distribution over functions, and can therefore be understood as a Function Distribution/Function Space. Classical methods to simulate random field are based on polynomial chaos expansion A.2 and Karhunen-Loeve expansion A.3, See [6]. Random field can be regarded as a Hilbert space($L^2(\Omega,H)$)-valued random variable.

Definition 15 ($L^p(\Omega,H)$ 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:*

$$\|X\|_{\mathcal{L}^p(\Omega,H)} := \left(\int_\Omega \|X(\omega)\|^p dP(\omega)\right)^{\frac{1}{p}} = \mathbf{E}[\|X\|^p]^{\frac{1}{p}} \tag{4.2}$$

Then we can define the inner product:

$$\langle X,Y\rangle_{\mathcal{L}^2(\Omega,H)} := \int_\Omega \langle X(\omega),Y(\omega)\rangle_H dP(\omega) \tag{4.3}$$

Definition 16 (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*

$$\langle \mathcal{C}_{XY}\phi,\psi\rangle_H = \mathrm{Cov}\left(\langle X,\phi\rangle_H,\langle Y,\psi\rangle_H\right), \forall\phi,\psi\in H \tag{4.4}$$

The covariance operator plays a special role as it characterizes the properties of the random field, like regularity and smoothness.

Example 6 ($H=\mathbb{R}^d$) *In finite dimensional case $H=\mathbb{R}^d$, the covariance matrix conincides with the covariance operator.*

$$\begin{aligned}&\mathrm{Cov}\left(\langle X,\phi\rangle,\langle Y,\psi\rangle\right) = \mathrm{Cov}\left(\phi^T X,\psi^T Y\right)\\=&\mathbf{E}\left[\phi^T(X-\mu_X)(Y-\mu_Y)^T\psi\right] = \phi^T\mathbf{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right]\psi\\=&\phi^T Cov(X,Y)\psi = \langle C_{XY}\phi,\psi\rangle\end{aligned} \tag{4.5}$$

Example 7 (X=Y) *When $X = Y$ noted as $u(x, \omega) \in H$, the covariance operator: The covariance operator $\mathcal{C}$ can be defined as:*

$$\begin{aligned}
&Cov\left(\langle u, \phi\rangle_H, \langle u, \psi\rangle_H\right) \\
=&\mathbb{E}_m\left[\langle u, \phi\rangle_H \langle u, \psi\rangle_H\right] \\
=&\int_H \langle u, \phi\rangle_H \langle u, \psi\rangle_H m(du) \\
=&\int_D \left(\int_D Cov(u(x), u(y))\phi(x)dx\right)\psi(y)dy \\
=&\langle \mathcal{C}_u \phi, \psi\rangle_H
\end{aligned} \tag{4.6}$$

So that for $\forall x \in D$,

$$(\mathcal{C}_u \phi)(x) = \int_D Cov(u(x), u(y))\phi(y)dy \tag{4.7}$$

That is any $L^2(D)$-valued random variable $u(x)$ can defines a R.F. with $\mu(x)$ and $C(x, y)$ equal to the integral kernel of $\mathcal{C}$. The measure m is called probability measure.

Definition 17 (Trace class) *For random field $u(x, \omega)$ with the covariance operator $\mathcal{C}$, suppose $\mathcal{C}$ is trace class with eigenpairs (λ_i, ϕ_i), then the second moment of $u(x, \omega)$ is given by:*

$$\mathbf{E}[\|u(x, \omega)\|_H^2] = \sum_{i=1}^{\infty} \lambda_i \leq \infty \tag{4.8}$$

Definition 18 (H-valued Gaussian random variable) *Let H be a Hilbert space. An H-valued random variable $u(x, \omega)$ is Gaussian if $\langle u(x, \omega), \phi\rangle_H$ is a real-valued Gaussian random variable for all $\phi \in H$. Here the real-valued Gaussian Random Variable is defined as:*

$$\langle u, \phi\rangle_H \sim N(\langle \mu, \phi\rangle_H, \langle \mathcal{C}_u \phi, \phi\rangle_H) \tag{4.9}$$

This actually defines the Gaussian Measure m on H: $u \sim N(\mu, \mathcal{C}_u) := m$. The covariance operator of u is the symmetric, positive-definite operator $\mathcal{C}_u : H \to H$.

Proposition 1 *If $u(x, \omega)$ is a Gaussian random field, then $\mathcal{C}_u$ is trace class. Reversely, if $\mathcal{C}_u$ is a positive, symmetric, trace class operator in H, then there exists a Gasussian measure $m = N(0, \mathcal{C}_u)$ on H.*

4.1.1 Gaussian Measure

Since we consider infinite dimensional case, unlike finite dimensions, not all translations preserve the measure. We need to cansider those directions in H along which translating a Gaussian measure does not change its essential nature (i.e., keeps it equivalent).

Including the SPDE, also these problems are dealed with in Hilbert Space. First we have some definitions:

Definition 19 *Let m_1, m_2 be two measures on H.*

- *$m_1 \ll m_2$ means measure m_1 is absoutely continuous respect to m_2: if $m_1(A) = 0$ for all A s.t. $m_2(A) = 0$. This means that the support of m_1 is a subset of the support of m_2.*
- *If $m_1 \ll m_2, m_2 \ll m_1$, m_1, m_2 are said to be equivalent $m_1 \sim m_2$.*
- *If there exists a measurable set A s.t. $m_1(A) = 0$ and $m_2(A^c) = 1$, then m_1, m_2 are singular $m_1 \perp m_2$.*

Theorem 42 (Radon-Nikodym Theorem) *Let $S = (H, \mathcal{B}(H))$ be a measurable space. m_1, m_2 are two σ-finite measures on S. If $m_1 \ll m_2$, then there exists a measurable function $f : H \to [0, \infty)$ s.t. $m_1(A) = \int_A f m_2(du), \forall A \in \mathcal{B}(H)$.*

The function f is called the Radon-Nikodym derivative of m_1 with respect to m_2 and is denoted by $f = \frac{dm_1}{dm_2}$.

Example 8 *In finite dimensional case, the Radon-Nikodym derivative is the probability density function. Like m_2 is the Lebesgue measure, $m_1(dx) = p(x)dx$, then the Radon-Nikodym derivative is the probability density function $p(x)$.*

Definition 20 (KL-divergence) *The KL-divergence is actually related to the Radon-Nikodym derivative: if $m_1 \ll m_2$, then:*

$$\mathrm{KL}(m_1||m_2) = \int_H \log\left(\frac{dm_1}{dm_2}(x)\right) dm_1(x) = \int_H \log\left(f(x)\right) f(x) dm_2(x) \tag{4.10}$$

This is the KL divergence between measures. We define the Radon-Nikodym derivative to be infinite if $m_1 \not\ll m_2$. So for $H = \mathbb{R}^d, dm_1 = p_1(x)dx, dm_2 = p_2(x)dx$, the KL-divergence can be written as:

$$\mathrm{KL}(m_1||m_2) = \int_{\mathbb{R}^d} \log\left(\frac{p_1(x)}{p_2(x)}\right) p_1(x)dx \tag{4.11}$$

The Radon-Nikodym derivative captures local density ratios, and the KL divergence is their global average.

Then we need to consider the characteristic functional of Gaussian measure, cause one can show that characteristic functions of two measures are identical, then the measures are identical as well:

Definition 21 (Characteristic Functional) *The characteristic functional of a measure m is defined as:*

$$\hat{m}(\lambda) = \int_H e^{i\langle \lambda, u\rangle_H} m(du) \tag{4.12}$$

For Gaussian measure $m = N(\mu, \mathcal{C})$, we have the characteristic functional:

$$\hat{m}(\lambda) = \int_H e^{i\langle \lambda, u\rangle_H} m(du) = e^{i\langle \lambda, \mu\rangle_H - \frac{1}{2}\langle \lambda, \mathcal{C}\lambda\rangle_H}, \qquad \lambda \in H \tag{4.13}$$

So conversely, let $\mu \in H, \mathcal{C}$ is a trace class operator on H. Then there exists a unique probability measure m on H s.t. $\hat{m}(\lambda) = e^{i\langle \lambda, \mu\rangle_H - \frac{1}{2}\langle \lambda, \mathcal{C}\lambda\rangle_H}, \forall \lambda \in H$. Here $m = N(\mu, \mathcal{C})$. Meanwhile, we have $\int_H \|x\|_H^2 m(dx) = \mathrm{Tr}(\mathcal{C}) + \|\mu\|_H^2$.

Definition 22 (Reproducing Kernel Hilbert Space) *The subspace $\mathcal{C}^{1/2}H$ is called the reproducing kernel of measure $N_{\mathcal{C}}$. If $\mathrm{Ker}(\mathcal{C}) = 0$, then $\mathcal{C}^{1/2}H$ is dense in H.*

Then we have an isomorphism between H and $L^2(H, N_{\mathcal{C}})$:

$$f \in \mathcal{C}^{1/2}H \to W_f \in L^2(H, N_{\mathcal{C}}), \qquad W_f(x) = \langle \mathcal{C}^{-1/2} f, x\rangle_H, \quad x \in H \tag{4.14}$$

for $\int_H W_f(x) W_g(x) N_{\mathcal{C}}(dx) = \langle f, g\rangle_{\mathcal{C}}, \quad \forall f, g \in H$.

4.1.2 Camaron Martin Space

Unlike the finite dimensional case, there is no natural Brownian Motion process in infinite dimensions. Here is why: For example, the space time white noise W_t is defined on $H = L^2([0,1])$, but it doesn't take values in H a.s. Since the white noise has covariance operator I, it is not trace class in H. So for Hilbert space H, we need to define the Brownian Motion process on H by using the Cameron-Martin space U. In other words, if we want to define the Diffusion Process on infinite dimensional space, we need to consider the direction in H along which translating a measure preserves the absolutely continuity.

Definition 23 (Cameron-Martin Space) *The Cameron-Martin space U is defined as:*

$$U := \{h \in H | m_h \ll m\}, m_h(A) = T_h^{\#} m(A) := m(A - h), \forall A \in \mathcal{B}(H) \tag{4.15}$$

where T_h is the translation operator: $T_h(u) = u + h$, $T_h^{\#} m$ is the push-forward measure of m by T_h,

So the Cameron-Martin space U is more regular than the original space H.

Theorem 43 (Cameron-Martin Theorem) *For $m = N(0, \mathcal{C})$, we have:*

$$m_h \sim m \text{ if and only if } h \in \mathcal{C}^{\frac{1}{2}} H := U \tag{4.16}$$

Since $\mathcal{C}^{-1}$ is unbounded, U is a proper subspace of $U = \mathcal{C}^{1/2} H$. The inner product is defined as:

$$\langle \phi, \psi \rangle_U = \langle \mathcal{C}^{-1/2} \phi, \mathcal{C}^{-1/2} \psi \rangle_H \tag{4.17}$$

This can be generalized to general Gaussian measure cases:

Theorem 44 (Feldman-Hajek Theorem[3]) *The following statements hold:*
1. $m_1 = N(\mu_1, \mathcal{C}_1), m_2 = N(\mu_2, \mathcal{C}_2)$ are either singular or equivalent.
2. They are equivalent if and only if

- $\mathcal{C}_1^{1/2}(H) = \mathcal{C}_2^{1/2}(H) = H_0$, $\mu_1 - \mu_2 \in H_0$
- $(\mathcal{C}_1^{-1/2} \mathcal{C}_2^{1/2})(\mathcal{C}_1^{-1/2} \mathcal{C}_2^{1/2})^* - I$ *is a Hilbert-Schmidt operator.*

Specifically, if $\mathcal{C}_1 = \mathcal{C}_2 = \mathcal{C}$, we have the Radon Nikodym derivative:

$$\frac{dm_1}{dm_2}(u) = \exp\left(\langle \mu_1 - \mu_2, \mathcal{C}^{-1}(u - \mu_2) \rangle_H - \frac{1}{2} \| \mathcal{C}^{-1/2}(\mu_1 - \mu_2) \|_H^2 \right) \tag{4.18}$$

Therefore, the KL divergence between m_1 and m_2 is:

$$\mathrm{KL}[m_1 | m_2] = \frac{1}{2} \left(\| \Delta\mu, C^{-1} \Delta\mu \|_H^2 \right) \tag{4.19}$$

First we assume define the Cylindrical Wiener Process $\hat{W}_t$ on H, but it does not take values in H cause I is not a trace class. Here we suppose $\hat{W}_t$ is H_1-valued:

$$\hat{W}_t = \sum_{i=1}^{\infty} \beta_i(t) \phi_i \tag{4.20}$$

where $\beta_i(t)$ are independent standart Brownian motions and ϕ_i are the eigenbasis of H. Normally, we can define the $\mathcal{C}$-Wiener Process on H as:

Definition 24 ($\mathcal{C}$-Wiener Process) *A H-valued process W_t is called $\mathcal{C}$-Wiener process if: 1. $W_0 = 0$ 2. W_t has continuous trajectories: $\mathbb{R}^+ \to H$ 3. W_t has independent Gaussian increments: $m(W_t - W_s) = N(0, (t-s)\mathcal{C})$*

Therefore by applying the covariance operator $\mathcal{C}$ to cylindrical Wiener Process $\hat{W}_t$, we can define the H-valued $\mathcal{C}$-Wiener process W_t as:

$$W_t = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\beta_i(t)\phi_i = \mathcal{C}^{1/2}\hat{W}_t \tag{4.21}$$

where λ_i are the eigenvalues of $\mathcal{C}$. So H is actually the Cameron-Matin space of H_1.

To sum up, in infinite dimensional diffusion model, suppose we have $u \sim m_0, \eta \sim m_1$, where m_0 is the initial measure and m_1 is the target measure, all defined on H. In the diffusion process, we add noise process η to u, getting $v = u + \eta$. Then $v \sim \nu = m_0 * m_1$. Here we need to do some assumptions on measure m_0 and m_1: m_0 needs to be supported on the Cameron-Martin space U of m_1, that is $m_0 \ll m_1$, m_0 should be more regular than m_1. Only in this way, we can define the Radon-Nikodym derivative of ν with respect to m_1, and the perturbed measure ν is equivalent to m_1.

Normally, to guarantee the diffusion progress, we have two choices: 1. The support of m_{data} is contained in Camaron-Martin space of $N(0, \mathcal{C}_{data})$ 2. Choose a large space K, s.t. the support of m_{data} and $N(0, \mathcal{C})$ are supported on K.

Definition 25 *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:*

$$A^\alpha u = \sum_{j=1}^{\infty} \lambda_j^\alpha \langle u, \phi_j \rangle \phi_j \tag{4.22}$$

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

Definition 26 (Problem Setting) *Assume we have H-valued Random Fields $\mathcal{M}(\S, \omega)$ defined on $(H, \mathcal{B}(H), m)$ and $\mathcal{N}(\S, \omega)$ defined on $(H, \mathcal{B}(H), m_0)$, where m_0 is the initial probability measure. We consider the following problem:*

- *How to build a generative model to sample from the unknown measure m_0? Conversly, given the target measure and initial measure, how to build the push forward operator $\mathcal{L}$?*
- *Given the initial measure m_0 and the push forward operator $\mathcal{L}$, what is the measure $m_t, \forall t \in [0,1]$?*

4.2 Stationary SPDEs

4.2.1 Definition

Definition 27 (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.:*

$$\begin{cases} -\nabla \cdot (a(x,w)\nabla u(x,w)) = f(x,w), & x \in D \\ u(x,w) = g(x), & x \in \partial D \end{cases} \tag{4.23}$$

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

Definition 28 (Weak solution on $D \times \Omega$) *A weak solution to Eq(4.23) with $g = 0$ is a function $u \in V = L^2(\Omega, H_0^1(D))$ s.t. for any $v \in V$,*

$$a(u, v) = l(v) \tag{4.24}$$

where

$$\begin{cases} a(u, v) = E\left[\int_D a(x, \cdot)\nabla u(x, \cdot) \cdot \nabla v(x, \cdot)dx\right] \\ \quad l(v) = E\left[\int_D f(x, \cdot)v(x, \cdot)dx\right] \end{cases} \tag{4.25}$$

If $g \neq 0$, a weak solution to Eq(4.23) is a function $u \in W = L^2(\Omega, H_g^1(D))$ s.t. for any $v \in V$,

$$a(u, v) = l(v) \tag{4.26}$$

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

$$\begin{cases} a(u, v) = E\left[\int_D a(x, \cdot)\nabla u(x, \cdot) \cdot \nabla v(x, \cdot)dx\right] \\ \quad l(v) = E\left[\int_D f(x, \cdot)v(x, \cdot)dx\right] \end{cases} \tag{4.27}$$

Theorem 45 (Existence and uniqueness of weak solution) *Note for all $x \in D$*

$$0 < a_{\min} \leq a(x, \cdot) \leq a_{\max} < \infty \tag{4.28}$$

as a basic assumption.

If $f \in L^2(\Omega, L^2(D)), g = 0$, and Assumption (4.28) holds, then SPDE 4.24 has a unique weak solution $u \in V$.

If Assumption (4.28) holds, $f \in L^2(\Omega, L^2(D))$, and $g \in H^{\frac{1}{2}}(\partial D)$, then SPDE 4.26 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. (4.28) holds.

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

$$\begin{cases} a(x, w) = \mu_a(x) + \sum_{i=1}^{N_a} \sqrt{v_i^a}\phi_i^a(x)\xi_i^a(w) \\ f(x, w) = \mu_f(x) + \sum_{i=1}^{N_f} \sqrt{v_i^f}\phi_i^f(x)\xi_i^f(w) \end{cases} \tag{4.29}$$

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:

$$\begin{aligned} a(u, v) &= E\left[\int_D a(x, \cdot)\nabla u(x, \cdot) \cdot \nabla v(x, \cdot)dx\right] \\ &= \int_\Omega \int_D a(x, w)\nabla u(x, w) \cdot \nabla v(x, w)dxdP(w) \end{aligned} \tag{4.30}$$

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(4.23) on $D \times \Gamma$, where $\Gamma = \prod_{i=1}^{N_a} \Gamma_i$.

Definition 29 (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 30 (Weak solution on $D \times \Gamma$) *Let $\tilde{a}(x)$ and $\tilde{f}(x)$ be finite-dimensional noises defined in Eq(4.29), then the solution to Eq (4.23) is also finite-dimensional noise. Define*

$$W := L^2_p(\Gamma, H^1_g(D)) = \left\{ v : D \times \Gamma \to \mathbb{R} : \int_\Gamma \|v(\xi, \cdot)\|^2_{H^1_g(D)} d\xi < \infty \right\} \tag{4.31}$$

A weak solution to Eq(4.23) on $D \times \Gamma$ is a function $u \in W = L^2_p(\Gamma, H^1_g(D))$ s.t. for any $v \in V = L^2_p(\Gamma, H^1_0(D))$,

$$a(u, v) = l(v) \tag{4.32}$$

where

$$\begin{cases} a(u, v) = \displaystyle\int_\Gamma p(\xi) \int_D \tilde{a}(x, \xi) \nabla u(x, \xi) \cdot \nabla v(x, \xi) dx d\xi \\ \quad l(v) = \displaystyle\int_\Gamma p(\xi) \int_D \tilde{f}(x, \xi) v(x, \xi) dx d\xi \end{cases} \tag{4.33}$$

4.2.2 Stochastic Galerkin Method

Therefore, we have the stochastic Galerkin solution: seek $u_{hk} \in W^{hk} \subset L^2(\Gamma, H^1_g(D))$ s.t. for any $v_{hk} \in V^{hk} \subset L^2(\Gamma, H^1_0(D))$.

By define the inner product:

$$\langle v, w \rangle_p = \int_\Gamma v(\xi) w(\xi) P(\xi) d\xi \tag{4.34}$$

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

$$L^2_p(\Gamma) := \{ v : \Gamma \to \mathbb{R} : \|v\|^2_{L^2_p(\Gamma)} = \langle v, v \rangle_p < \infty \} \tag{4.35}$$

Definition 31 *Note S^k be the set of polynomials of degree k or less on Γ:*

$$\begin{aligned} S^k &= \text{span}\{ \prod_{i=1}^M P_i^{\alpha_i}(\xi_i) : \alpha_i \in \mathbb{N}_0, \sum_{i=1}^M \alpha_i \leq k \} \\ &= \text{span}\{ \psi_1, \psi_2, \cdots, \psi_Q \} \end{aligned} \tag{4.36}$$

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

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

$$p(\xi) = \prod_{i=1}^M p_i(\xi_i) \tag{4.37}$$

Recall $V^h = \text{span}\{\phi_i\}_{i=1}^J \subset H^1_0(D)$ is the finite element space, we have tensor product space:

$$V^{hk} := V^h \otimes S^k = \text{span}\{\phi_i \psi_j\}_{i=1,j=1}^{J,Q} \tag{4.38}$$

Then

$$W^{hk} := V^{hk} \oplus \text{span}\{\phi_{J+1}, \cdots, \phi_{J+J_b}\} \tag{4.39}$$

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

Theorem 46 (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:

$$u_{hk}(x,\xi) = \sum_{i=1}^{J}\sum_{j=1}^{Q} u_{ij}\phi_i(x)\psi_j(\xi) + w_g \tag{4.40}$$

Theorem 47 (Mean and covariance) *The Galerkin solution can be rewritten as:*

$$\begin{aligned} u_{hk}(x,\xi) &= \sum_{j=1}^{Q}\left(\sum_{i=1}^{J} u_{ij}\phi_i(x)\right)\psi_j(\xi) + w_g \\ &= \sum_{j=1}^{Q} u_j\psi_j(\xi) + w_g \qquad (4.41)\\ &= (u_1(x) + w_g(x))\psi_1(\xi) + \sum_{j=2}^{Q} u_j(x)\psi_j(\xi) \end{aligned}$$

Then the mean and covariance is

$$\begin{cases} E[u_{hk}] = u_1 + w_g \\ Var(u_{hk}) = \sum_{j=2}^{Q} u_j^2 \end{cases} \tag{4.42}$$

4.2.3 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:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1Q} \\ A_{21} & A_{22} & \cdots & A_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ A_{Q1} & A_{Q2} & \cdots & A_{QQ} \end{pmatrix}, \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_Q \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_Q \end{pmatrix} \tag{4.43}$$

where

$$\mathbf{u}_j = [u_{1j}, u_{2j}, \cdots, u_{Jj}]^T, j = 1,2,\cdots,Q \tag{4.44}$$

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

$$A_{ij} = K_0\langle\psi_i,\psi_j\rangle_p + \sum_{l=1}^{P} K_l\langle\psi_i,\psi_j\xi_l\rangle_p \tag{4.45}$$

where

$$\begin{cases} [K_0]_{rs} = \int_D \mu_a(x)\nabla\phi_r\cdot\nabla\phi_s dx \\ [K_l]_{rs} = \int_D (\sqrt{v_l^a}\phi_l^a)\nabla\phi_r\cdot\nabla\phi_s dx \end{cases}, \qquad r,s = 1,2,\cdots,J \tag{4.46}$$

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

$$\mathbf{b}_s = \langle\psi_1,\psi_s\rangle_p F_0 + \sum_{l=1}^{P} F_l\langle\xi_l,\psi_s\rangle_p - \langle W,\psi_s\rangle_p \tag{4.47}$$

where

$$\begin{cases} [F_0]_i = \int_D \mu_f(x)\phi_i(x)dx \\ [F_l]_i = \int_D (\sqrt{v_l^f}\phi_l^f)\phi_i(x)dx \\ W = K_{0B}^T \mathbf{w}_B + \sum_{l=1}^{P} K_{lB}^T \xi_l \mathbf{w}_B \end{cases} \tag{4.48}$$

4.3 Semilinear Stochastic PDEs

4.3.1 Definition

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

$$du = [\Delta u + f(u)]dt + G(u)dW(t,x) \tag{4.49}$$

Definition 32 (Semilinear SPDE) *Simmilar to normal time-dependent PDE, we treat SPDE like this as semilinear SODEs on a Hilbert space, like*

$$du = [-Au + f(u)]dt + G(u)dW(t) \tag{4.50}$$

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

Example 9 (Phase-field model)

$$du = [\epsilon\Delta u + u - u^3]dt + \sigma dW(t,x) \tag{4.51}$$

Example 10 (Fluid Flow)

$$\begin{aligned} u_t &= \epsilon\Delta u - \nabla p - (u \cdot \nabla)u \\ \nabla \cdot u &= 0 \end{aligned} \tag{4.52}$$

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 33 (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 34 (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 48 (Q-Wiener Process) *Assume we have Q defined in 33. Then, $W(t)$ is a Q-Wiener process if and only if*

$$W(t) = \sum_{j=1}^{\infty} \sqrt{q_j} \mathcal{X}_j \beta_j(t) \tag{4.53}$$

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

Theorem 49 ($H^r_{\text{per}}(0,a)$-valued process) *...*

Theorem 50 ($H^r_0(0,a)$-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 35 (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 51 *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 $L^2(U, U_1)$).*

4.3.2 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 36 (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.*

$$\|B\|_{L^2_0} := \left(\sum_{j=1}^{\infty} \|BQ^{\frac{1}{2}}\mathcal{X}_j\|^2\right)^{\frac{1}{2}} = \|BQ^{\frac{1}{2}}\|_{\text{HS}(U_0,H)} < \infty \tag{4.54}$$

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

Definition 37 *The stochastic integral can be defined by*

$$\int_0^t B(s)dW(s) := \sum_{j=1}^{\infty} \int_0^t B(s)\sqrt{q_j}\mathcal{X}_j d\beta_j(s) \tag{4.55}$$

So, we can have the truncated form:

$$\int_0^t B(s)dW^J(s) = \sum_{j=1}^{J} \int_0^t B(s)\sqrt{q_j}\mathcal{X}_j d\beta_j(s) \tag{4.56}$$

4.3.3 Solution

Consider the semilinear SPDE:

$$du = [-Au + f(u)]dt + G(u)dW(t) \tag{4.57}$$

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 11 *Consider the stochastic heat equation:*

$$du = \Delta u dt + \sigma dW(t,x), u(0,x) = u_0(x) \in L^2(D) \tag{4.58}$$

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,

$$H = U = L^2(D), f(u) = 0, G(u) = \sigma I \tag{4.59}$$

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

$$u(t) = u_0 + \int_0^t [-Au(s) + f(u(s))]ds + \int_0^t G(u(s))dW(s), \quad \forall t \in [0,T] \tag{4.60}$$

Definition 39 (weak solution) *A predictable H -valued process $\{u(t) : t \in [0,T]\}$ is called a weak solution if*

$$\langle u(t), v\rangle = \langle u_0, v\rangle + \int_0^t [-\langle u(s), Av\rangle + \langle f(u(s)), v\rangle]ds + \int_0^t \langle G(u(s))dW(s), v\rangle, \quad \forall t \in [0,T], v \in \mathcal{D}(A) \tag{4.61}$$

where

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

Definition 40 (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 12 (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: $\forall v \in \mathcal{D}(A)$,*

$$\begin{aligned}\langle u(t), v\rangle_{L^2(0,\pi)} &= \langle u_0, v\rangle_{L^2(0,\pi)} + \int_0^t \langle -u(s), Av\rangle_{L^2(0,\pi)} ds \\ &+ \sum_{j=1}^{\infty}\int_0^t \sigma\sqrt{q_j}\,\langle \phi_j, v\rangle_{L^2(0,\pi)}\, d\beta_j(s)\end{aligned} \tag{4.62}$$

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

$$\hat{u}_j(t) = \hat{u}_j(0) + \int_0^t (-\lambda_j)\,\hat{u}_j(s) ds + \int_0^t \sigma\sqrt{q}_j d\beta_j(s). \tag{4.63}$$

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

$$d\hat{u}_j = -\lambda_j \hat{u}_j dt + \sigma\sqrt{q_j} d\beta_j(t) \tag{4.64}$$

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

$$\mathrm{Var}\left(\hat{u}_j(t)\right) = \frac{\sigma^2 q_j}{2\lambda_j}\left(1 - \mathrm{e}^{-2\lambda_j t}\right) \tag{4.65}$$

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

$$\|u(t)\|^2_{L^2(\Omega, L^2(0,\pi))} = \mathbb{E}\left[\sum_{j=1}^{\infty} |\hat{u}_j(t)|^2\right] = \sum_{j=1}^{\infty}\frac{\sigma^2 q_j}{2\lambda_j}\left(1 - \mathrm{e}^{-2\lambda_j t}\right). \tag{4.66}$$

4.4 SPDE as infinite-dimensional SDEs

4.4.1 Kolmogorov's backward equation

Refer to Chapter 7 of [2].

We have studied the semilinear SPDE, which can be rewritten as infinite-dimensional SDEs $X(t, x), x \in H$, here x can be viewed as $u(t)$:

$$\begin{cases} dX(t,x) = [AX(t,x) + F(t, X(t,x))]dt + G(X(t,x))dW(t) \\ X(0,x) = x \in H \end{cases} \tag{4.67}$$

where A is the generator of a semigroup $S(t) = e^{tA}$ and G is a Hilbert-Schmidt operator. We also assume U, H are separable Hilbert spaces, W_t is a Q weiner process taking values in $U \supset U_0$, U_0 is the Cameron-Martin space of U, x is actually u_0, an H-valued random variable. Note $\mathcal{L}_0^2 = \mathcal{L}(U_0, H)$. F, G need to be formulated Lipschitz and linear growth. Here are the details:

Theorem 52 (Hypothesis) *We fix $T > 0$ and impose first the following conditions on coefficients A, F, G of the equation.*

- *A is the generator of a semigroup $S(t) = e^{tA}$ on H*

- *The drift:*
$$F : [0,T] \times \Omega \times H \to H : (t, \omega, x) \mapsto F(t, \omega, x) \tag{4.68}$$
is a Borel measurable function from $(\Omega_T \times H, \mathcal{B}(\Omega_T) \times \mathcal{B}(H))$ *to* $(H, \mathcal{B}(H))$

- *the diffusion:*
$$G : [0,T] \times \Omega \times H \to \mathcal{L}_0^2 : (t, \omega, x) \mapsto G(t, \omega, x) \tag{4.69}$$
is a Borel measurable function from $(\Omega_T \times H, \mathcal{B}(\Omega_T) \times \mathcal{B}(H))$ *to* $(\mathcal{L}_0^2, \mathcal{B}(\mathcal{L}_0^2))$

- *There exists a constant* $C > 0$ *such that for all* $t \in [0,T]$, $\omega \in \Omega$, $x, y \in H$*:*
$$\|F(t,\omega,x) - F(t,\omega,y)\|_H + \|G(t,\omega,x) - G(t,\omega,y)\|_{L_0^2} \leq C(\|x-y\|_H \tag{4.70}$$
and
$$\|F(t,\omega,x)\|_H^2 + \|G(t,\omega,x)\|_{L_0^2}^2 \leq C^2(1 + \|x\|_H^2) \tag{4.71}$$

Consider the drift and diffusion term depend only on x:

$$dX(t) = [AX(t) + F(X(t))]dt + G(X(t))dW(t) \tag{4.72}$$

and the initial value is $X(0) = x$. As discussed in 40, the mild solution of 4.72 is given by:

$$X(t) = S(t)x + \int_0^t S(t-s)F(X(s))ds + \int_0^t S(t-s)G(X(s))dW(s) \tag{4.73}$$

where $S(t) = e^{tA}$ is the semigroup generated by A.

Then assume the initial measure m, test function $\varphi(x), x \in H$, we shall study the transition semigroup P_t defined by

$$v(t,x) = P_t\varphi(x) = \mathbb{E}\left[\varphi(X(t,x))\right], \qquad \varphi \in \mathcal{L}_b(H) \tag{4.74}$$

which is a solution of Kolmogorov's backward equation:

$$\begin{cases} v_t = \langle Ax + F(x), D_x v\rangle + \dfrac{1}{2}\operatorname{Tr}\left[D_{xx}v\left(GQ^{\frac{1}{2}}\right)\left(GQ^{\frac{1}{2}}\right)^*\right], & x \in \mathcal{D}(A), t > 0 \\ v(0,x) = \varphi(x), \qquad x \in H \end{cases} \tag{4.75}$$

where:

- $A : D(A) \subset H \to H$ is a linear operator
- $D\phi(x)$ denotes the Fréchet derivative (gradient) of ϕ,
- $D^2\phi(x)$ denotes the second Fréchet derivative (Hessian operator),
- Tr denotes the trace.

Here we note $v_t = \mathcal{L}v$, the generator operator $\mathcal{L}$ is on test function $v : [0 + \infty) \times H \to \mathbb{R}$. Under curtein condition, the BK equation 4.75 has a unique solution $v \in C_b^{1,2}([0,T]; \mathcal{L}_b(H))$. See [3] Chapter 9.0.

Then if we define the measure m_t on H, we can have the adjoint generator operator $\mathcal{L}^*$ on m_t:

$$\frac{d}{dt}m_t = \mathcal{L}^* m_t \tag{4.76}$$

This is called the infinite-dimensional Fokker-Planck-Kolmogorov equation. Since we have

$$\int_H \mathcal{L}\varphi(x)m_t(dx) = \int_H \varphi(x)\mathcal{L}^* m_t(dx) \tag{4.77}$$

If there exists a density $\rho(t,x)$ with respect to a reference measure ν (e.g., a Gaussian measure), guarantee absolutely continuous, such that:

$$d\mu(t,x) = \rho(t,x)\,d\nu(x), \tag{4.78}$$

then $\rho(t,x)$ satisfies the strong form of the Fokker-Planck-Kolmogorov equation:

$$\partial_t \rho(t,x) = \mathcal{L}^* \rho(t,x), \tag{4.79}$$

To complete the proof, we need to generalize the Ito's formula to infinite dimension case:

Theorem 53 (Ito's Formula for Infinite Dimension) *Consider test function $\varphi(X_t) : H \to \mathbb{R}$ which is twice Frechet differentiable and X_t is the solution to:*

$$dX_t = F(X_t,t)dt + G(X_t,t)dW_t \tag{4.80}$$

where $F : [0,T] \times \Omega \times H \to H$ and $G : [0,T] \times \Omega \times H \to \mathcal{L}_0^2$ are Hilbert-Schmidt operators. Then we have:

$$d\varphi(X_t) = \langle D\varphi(X_t), dX_t\rangle_H + \frac{1}{2}\operatorname{Tr}\left(GQG^* D^2\varphi(X_t)\right)dt \tag{4.81}$$

Then the proof is straightforward.

Assume the measure are all relatively absolutely continuous with respect to the Gaussian measure $N(0,Q)$, then we have measure density $\rho(t,x)$ satisfies:

$$\begin{aligned}\frac{d}{dt}\rho(t,x) &= -\operatorname{div}\left((Ax+F(x))\rho\right) + \frac{1}{2}\operatorname{div}^2\left(GQG^*\rho\right)\\ &= -\operatorname{div}\left((Ax+F(x))\,\rho - \frac{1}{2}D\left(GQG^*\rho\right)\right)\end{aligned} \tag{4.82}$$

If G is independent of x, then we have:

$$\frac{d}{dt}\rho(t,x) = -\operatorname{div}\left(\left((Ax+F(x)) - \frac{1}{2}GQG^*D\log\rho\right)\rho\right) \tag{4.83}$$

Similar to the case of finite dimension, we can define:

$$V(t,x) = (Ax+F(x)) - \frac{1}{2}GQG^*D\log\rho \tag{4.84}$$

Then we can define the Flow ϕ_t decided by field V by:

$$\frac{d\phi_t(x)}{dt} = V(t,\phi_t(x)), \qquad x \in H, \phi_0(x) = x \tag{4.85}$$

4.4.2 Absolutely Continuous Measure

Assume we have $X_t, \tilde{X}_t$ are weak solutions of:

$$\begin{aligned} dX_t &= [AX_t + F(X_t)]dt + G(X_t)dW_t, \qquad & X(0) = x\\ d\tilde{X}_t &= [\tilde{A}\tilde{X}_t + \tilde{F}(\tilde{X}_t)]dt + \tilde{G}(\tilde{X}_t)d\tilde{W}_t, \qquad & \tilde{X}(0) = x\end{aligned} \tag{4.86}$$

where W is a cylindrical Wiener process on Hilbert Space U. We denote by $\{e_k\}$ a complete orthonormal basis of U and by $\{\beta_k\}$ a sequence of independent Brownian motions s.t.

$$\langle W(t), x\rangle = \sum_{k=1}^{\infty} \langle x, e_k\rangle \beta_k(t), \qquad x \in H \tag{4.87}$$

We hope to define derivative and integral on H which is related to Gateaux derivative and Fomin differentiable measure.

Definition 41 (smooth cylinder function) *The smooth cylinder function is the set of functions of the form:*

$$\mathcal{F}\mathcal{C}_b^\infty := \{f(\varphi_1, \varphi_2, \cdots, \varphi_n) : \mathcal{H} \to \mathbb{R} | f \in \mathcal{C}_b^\infty(\mathbb{R}^m), \varphi_i \in \mathcal{H}^*, m \in \mathbb{N}\} \tag{4.88}$$

where $\varphi_i \in \mathcal{H}^*$ is a linear functional: $\mathcal{H} \to \mathbb{R}$.

Definition 42 (Gateaux derivative) *The Gateaux derivative of $f_{\varphi_{1:m}} \in \mathcal{F}\mathcal{C}_b^\infty$ at $u \in \mathcal{H}$ in the direction of $\xi \in \mathcal{H}_0$, which is the Camaron Martin Space of measure $\mathcal{C}$, is defined by:*

$$\langle Df_{\varphi_{1:m}}, \xi\rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \partial_i f_{\varphi_{1:m}}(u) \langle \mathcal{C}\varphi_i, \xi\rangle_{\mathcal{H}_0} \tag{4.89}$$

4.4.3 Girsanov's theorem

Here U_0 is the Camaron-Martin space of U and note the norm of U_0 is $\|\cdot\|_{U_0}$.

Theorem 54 *Assume $\psi(\cdot)$ is a U_0-valued $\mathscr{F}_t$-adapted process s.t.*

$$\mathbb{E}\left[\exp\left(\int_0^T \langle\psi(s), dW(s)\rangle_{U_0} - \frac{1}{2}\int_0^T \|\psi(s)\|_{U_0}^2 ds\right)\right] = 1 \tag{4.90}$$

Then the process

$$\widehat{W}(t) = W(t) - \int_0^t \psi(s) ds, \quad t \in [0, T] \tag{4.91}$$

is Q-Wiener process with respect to the filtration $\{\mathscr{F}_t\}$ on $(\Omega, \mathscr{F}, \widehat{\mathbb{P}})$ where

$$d\widehat{\mathbb{P}} = \exp\left(\int_0^T \langle\psi(s), dW(s)\rangle_{U_0} - \frac{1}{2}\int_0^T \|\psi(s)\|_{U_0}^2 ds\right) d\mathbb{P} \tag{4.92}$$

Chapter **5**

About KPZ Equation, some singluar SPDEs

After we talk about some theory about SPDE, we find that the random noise in SPDE plays a very important role in the dynamics of the system. For example, when the noise has good regularity, the solution of SPDE is smooth. But when the noise is singular, like space-time white noise, the solution is not regular functions but distributions. In this chapter, we will talk about some singular SPDEs, such as KPZ equation and some other SPDEs[5].

5.1 Edwards–Wilkinson (EW) Equation

The equation is given by:

$$\frac{\partial h(t,x)}{\partial t} = \nu \nabla^2 h(t,x) + \eta(t,x)$$

where

- $h(t,x)$: Surface height at time t and position x.
- ν: Diffusion coefficient (surface tension).
- ∇^2: Laplacian operator (models smoothing).
- $\eta(t,x)$: Space-time white noise, typically modeled as Gaussian with

$$\mathbb{E}[\eta(t,x)\eta(t',x')] = 2D\,\delta(t-t')\,\delta^d(x-x')$$

Theorem 55 *The EW equation describes the stochastic evolution of a surface under linear diffusion and random fluctuations. It is analytically solvable and is used to model symmetric surface growth or roughening when $d=1$.*

Proof 15 *Consider the Edwards–Wilkinson (EW) equation with $h(0,x)=0$. We define the* ***mild solution*** *as:*

$$h(t,x) = \int_0^t \int_{\mathbb{R}^d} G(t-s, x-y)\,\eta(s,y)\,dy\,ds$$

where $G(t,x)$ is the heat kernel:

$$G(t,x) = \frac{1}{(4\pi\nu t)^{d/2}} \exp\left(-\frac{|x|^2}{4\nu t}\right)$$

We compute the second moment of $h(t,x)$ to check whether it is square integrable:

$$\mathbb{E}[h(t,x)^2] = \int_0^t \int_{\mathbb{R}^d} G(t-s, x-y)^2\,dy\,ds$$

Substitute the explicit form of G:

$$G(t-s,y)^2 = \left(\frac{1}{(4\pi\nu(t-s))^{d/2}} \exp\left(-\frac{|y|^2}{4\nu(t-s)}\right)\right)^2 = \frac{1}{(4\pi\nu(t-s))^d} \exp\left(-\frac{|y|^2}{2\nu(t-s)}\right)$$

Then,

$$\int_{\mathbb{R}^d} G(t-s,y)^2\,dy = \frac{1}{(4\pi\nu(t-s))^d}\int_{\mathbb{R}^d} \exp\left(-\frac{|y|^2}{2\nu(t-s)}\right)\,dy$$

Using the Gaussian integral:

$$\int_{\mathbb{R}^d} \exp\left(-\frac{|y|^2}{2\nu(t-s)}\right)\,dy = (2\pi\nu(t-s))^{d/2}$$

So the integral becomes:

$$\int_{\mathbb{R}^d} G(t-s,y)^2\,dy = \frac{(2\pi\nu(t-s))^{d/2}}{(4\pi\nu(t-s))^d} = C_d\,(t-s)^{-d/2}$$

Therefore:

$$\mathbb{E}[h(t,x)^2] = \int_0^t C_d\,(t-s)^{-d/2}\,ds$$

This integral converges if and only if:

$$\int_0^t (t-s)^{-d/2}\,ds < \infty \quad\Longleftrightarrow\quad \frac{d}{2} < 1 \quad\Longleftrightarrow\quad d < 2$$

So, for $d < 2$, the second moment is finite, and the mild solution exists pointwise in $L^2(\Omega)$. For $d \geq 2$, the integral diverges, thus $\mathbb{E}[h(t,x)^2] = \infty$, and the mild solution does not exist as an L^2-valued random variable at each point x.

Hence, for $d \geq 2$, the solution can only be interpreted as a generalized function (distribution), not a classical or square-integrable function.

5.2 Kardar-Parisi-Zhang (KPZ) Equation

The equation is given by:

$$\frac{\partial h(t,x)}{\partial t} = \nu\nabla^2 h(t,x) + \frac{\lambda}{2}\left(\nabla h(t,x)\right)^2 + \eta(t,x)$$

where the nonlinear term $\frac{\lambda}{2}(\nabla h)^2$ models slope-dependent growth (faster growth along steeper gradients).It generalizes the EW model by incorporating asymmetric growth behavior.

The KPZ equation captures more realistic surface growth phenomena where the rate of height increase depends on local slope, such as in flame propagation, deposition, or biological growth.

To linearize KPZ in one spatial dimension, define

$$h(t,x) = \frac{2\nu}{\lambda}\log Z(t,x)$$

as Cole-Hopf Transformation.

Then $Z(t,x)$ satisfies:

$$\frac{\partial Z}{\partial t} = \nu\nabla^2 Z + \frac{\lambda}{2\nu}\eta(t,x)Z$$

It is a multiplicative stochastic heat equation, amenable to probabilistic and path integral methods (e.g., Feynman-Kac formula). If the noise term $\eta(t) = 0$, it can be easily solved by Fourier transform:

$$h(t,x) = \frac{2\nu}{\lambda} \log \left\{ \int_{\mathbb{R}^d} \frac{d^d\xi}{(4\pi\nu t)^{d/2}} \exp\left(-\frac{(x-\xi)^2}{4\nu t} + \frac{\lambda}{2\nu} h_0(\xi) \right) \right\} \tag{5.1}$$

By Feynman-Kac formula, we can get the solution of KPZ equation:

$$h(t,x) = \frac{2\nu}{\lambda} \log \left\{ \mathbb{E}\left[Z_0(X_t) \exp\left(\frac{\lambda}{2\nu} \int_0^t \eta(t-s, X_s) ds \right) \right] \right\} \tag{5.2}$$

We mainly describe the KPZ equation by roughness exponent and the dynamic exponent.

Bibliography

[1] Castellet Solanas, M.: Kolmogorov Equations for Stochastic PDEs, Advanced Courses in Mathematics CRM Barcelona, vol. 7. 1 edn. (Jan 2004)

[2] Da Prato, G., Zabczyk, J.: Second Order Partial Differential Equations in Hilbert Spaces. London Mathematical Society Lecture Note Series, Cambridge University Press (2002)

[3] Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2 edn. (2014)

[4] Holderrieth, P., Erives, E.: Introduction to flow matching and diffusion models (2025), https://diffusion.csail.mit.edu/

[5] Kardar, M., Parisi, G., Zhang, Y.C.: Dynamic scaling of growing interfaces. Physical Review Letters **56**(9), 889 (1986)

[6] Lord, G.J., Powell, C.E., Shardlow, T.: An Introduction to Computational Stochastic PDEs. Cambridge Texts in Applied Mathematics, Cambridge University Press (2014)

[7] Sigrist, F., Künsch, H.R., Stahel, W.A.: Stochastic partial differential equation based modelling of large space–time data sets. Journal of the Royal Statistical Society Series B: Statistical Methodology **77**(1), 3–33 (2015)

[8] Stein, M.L.: Space–time covariance functions. Journal of the American Statistical Association **100**(469), 310–321 (2005)

Appendix A

Appendix

A.1 Fourier Transform

Definition 43 *Define the Fourier transform of $u(x)$ as:*

$$\begin{cases} \mathcal{F}u(k) = \int_{\mathbb{R}^d} u(x)e^{-ikx}dx, \\ \mathcal{F}^{-1}u(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} u(k)e^{ikx}dk. \end{cases} \tag{A.1}$$

A.2 Polynomial Chaos Expansion

A spectral expansion in $L_\mu(D)$ is called chasos expansion. By defining the inner product in $L_\mu(D)$ as:

$$\langle f, g\rangle_\mu = \int_D f(x)g(x)d\mu(x), \tag{A.2}$$

the space of $L_\mu(D)$ is a Hilbert space.

Theorem 56 *Let $\{\phi_i(x)\}_{i=1}^\infty$ be an orthonormal basis of $L_\mu(D)$, i.e.*

$$\begin{aligned} &1.\int_D \phi_i(x)\phi_j(x)d\mu(x) = \delta_{ij}, \\ &2.\phi_i(x) \textit{ is dense in } L_\mu(D). \end{aligned} \tag{A.3}$$

Then the chasos expansion of $\mathcal{M}(x,\omega) \in L_\mu(D)$ is given by:

$$\mathcal{M}(x,\omega) = \sum_{i=1}^\infty c_i\phi_i(x), \tag{A.4}$$

where $c_i(x) = \langle \mathcal{M}, \phi_i\rangle_\mu$ are the coefficients of the expansion.

A.3 Karhunen-Loeve Expansion

Let $\mu(x)$ be the mean function and $C(x,y)$ be the covariance function of $\mathcal{M}(x,\omega)$. Assume that D is bounded, $C(x,y)$ is continuous and $\mathcal{M}(x,\cdot)$ has finite variables for all $x \in D$.

Theorem 57 *The Karhunen-Loeve expansion of $\mathcal{M}(x,\omega)$ is given by:*

$$\mathcal{M}(x,\omega) = \mu(x) + \sum_{i=1}^\infty \lambda_i\phi_i(x)\xi_i(\omega), \tag{A.5}$$

where λ_i and $\phi_i(x)$ are the eigenvalues and eigenfunctions of the covariance operator $\mathcal{C}$:

$$(\mathcal{C}\phi_i)(x) = \int_D Cov(\mathcal{M}(x,\omega),\mathcal{M}(y,\omega))\phi_i(y)dy = \int_D C(x,y)\phi_i(y)dy = \lambda_i\phi_i(x). \tag{A.6}$$

where $C(x,y), x, y \in D$ is the covariance function of $\mathcal{M}(x,\omega)$. The KL-random variables $\xi_i(\omega)$ are the result of the projection of $\mathcal{M}(x,\omega)$ onto the eigenfunctions $\phi_i(x)$:

$$\xi_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_D (\mathcal{M}(x,\omega) - \mu(x))\phi_i(x)dx. \tag{A.7}$$

Note that both $\{\phi_i(x)\}_{i=1}^{\infty}$ and $\{\xi_i(\omega)\}_{i=1}^{\infty}$ are orthonormal bases, one capturing the "spatial" variation of $\mathcal{M}(x,\omega)$ over D (in terms of x), the other capturing the stochastic variation of $\mathcal{M}(x,\omega)$ (in terms of ω).

Theorem 58 (Mercer's theorem) *The covariance function $C(x,y)$ of $\mathcal{M}(x,\omega)$ can be expressed as:*

$$C(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y). \tag{A.8}$$

It follows that the average variance of the random field over the domain D is equal to $\sum_{i=1}^{\infty}\lambda_i$.

In most cases, the integral eigenvalue problem in Equ (57) is difficult: analytically solved and high-dimensional.

A.4 Some proofs

Proof 16 (Proof of Thm 17)

$$\begin{aligned}
\mathbb{E}\left[\|u(x,\omega)\|_H^2\right] &= \mathbb{E}\left[\sum_{i=1}^{\infty}\langle u,\phi_i\rangle_H^2\right] \\
&= \sum_{i=1}^{\infty}\mathbb{E}\left[\langle u,\phi_i\rangle_H^2\right] \\
&= \sum_{i=1}^{\infty}\langle \mathcal{C}_u\phi_i,\phi_i\rangle_H = \sum_{i=1}^{\infty}\lambda_i
\end{aligned} \tag{A.9}$$

Proof 17 (Proof of Thm 36)

$$\begin{aligned}
S_u(k) = (\mathcal{F}c)(k) &= \int_{\mathbb{R}^d} e^{-ikh}c(h)dh \\
&= \int_{\mathbb{R}^d} e^{-ik(x+h-x)}\mathbb{E}\left[u(x+h)u(x)\right]dh \\
&= \mathbb{E}\left[\int_{\mathbb{R}^d} e^{-ik(x+h)}u(x+h)e^{ikx}u(x)dh\right] \\
&= \frac{1}{(2\pi)^d}\mathbb{E}\left[|(\mathcal{F}u)(k)|^2\right]
\end{aligned} \tag{A.10}$$

Proof 18 (Proof of Thm 40) *Do Fourier transform on Equ (3.24):*

$$\left\{\mathcal{F}(\kappa^2-\Delta)^{\alpha/2}u\right\}(k) = (\kappa^2+\|k\|^2)^{\alpha/2}(\mathcal{F}u)(k) \tag{A.11}$$

then we have

$$(\mathcal{F}u)(k) = \hat{u}(k) = \frac{\hat{W}(k)}{(\kappa^2+\|k\|^2)^{\alpha/2}} \tag{A.12}$$

Therefore, u can be written as:

$$u(x) = \mathcal{F}^{-1}\left[\frac{\hat{W}(k)}{(\kappa^2+\|k\|^2)^{\alpha/2}}\right] \tag{A.13}$$

Then the stationary covariance function of u is given by:

$$c(x) = Cov(u(x), u(0)) \tag{A.14}$$

By the definition of spectral density Equ (3.15) and Equ (A.12) we have:

$$S_u(k) = \frac{1}{(2\pi)^d} \frac{\mathbb{E}\left[\left|\hat{W}(k)\right|^2\right]}{(\kappa^2 + \|k\|^2)^\alpha} = \frac{1}{(\kappa^2 + \|k\|^2)^\alpha} \tag{A.15}$$

Then we have the variance of u:

$$c(0) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} S_u(k) dk = \frac{\Gamma(\nu)}{(4\pi)^{d/2} \kappa^{2\nu} \Gamma(\alpha)} := \sigma^2 \tag{A.16}$$

By Wiener-Khinchin theorem, we have:

$$\begin{aligned} c(x) &= (\mathcal{F}^{-1} S_u)(x) = \mathcal{F}^{-1}\left[\frac{1}{(\kappa^2 + \|k\|^2)^\alpha}\right] \\ &= \frac{\|x\|^\nu K_\nu(\kappa\|x\|)}{(4\pi)^{d/2} 2^{\nu-1} \kappa^\nu \Gamma(\alpha)} \\ &= \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa\|x\|)^\nu K_\nu(\kappa\|x\|) \end{aligned} \tag{A.17}$$

Remark 8 *To make $c(0) = 1$, we can multiple a constant factor σ_1 to $S_u(k)$:*

$$\sigma_1^2 = \frac{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}{\Gamma(\nu)} \tag{A.18}$$

Then the corresponding function of u is:

$$u(x) = \mathcal{F}^{-1}\left[\frac{\sigma_1 \hat{W}(k)}{(\kappa^2 + \|k\|^2)^{\alpha/2}}\right] \tag{A.19}$$