Stochastic Differential Equations and Continuous Diffusion Models

Liu Yang

In this lecture, we will cover an introduction to the stochastic differential equations. Some concepts could be hard to understand rigorously, but an intuitive understanding should be enough for this class.

1 Brownian Motion

1.1 Stochastic Process and Brownian Motion

Definition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A stochastic process is a measurable function $X(t, \omega)$ defined on the product space $[0, \infty) \times \Omega$. In particular,*

(a) for each t, $X(t, \cdot)$ is a random variable,

(b) for each ω, $X(\cdot, \omega)$ is a measurable function (called a sample path).

For convenience, the random variable $X(t, \cdot)$ will be written as $X(t)$ or X_t. Thus, a stochastic process $X(t, \omega)$ can also be expressed as $X(t)(\omega)$ or simply as $X(t)$ or X_t.

In most cases, we are tackling X with continuous sample paths, so $X(\cdot, \omega)$ is measurable.

Definition 2. *A stochastic process $B(t, \omega)$ is called a Brownian motion if it satisfies the following conditions:*

1. *$P\{\omega; B(0, \omega) = 0\} = 1$, i.e. $B(0, \omega) = 0$ a.e.*
2. *For any $0 \le s < t$, the random variable $B(t) - B(s)$ is normally distributed with mean 0 and variance $t - s$, i.e., for any $a < b$,*

$$P\{a \le B(t) - B(s) \le b\} = \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b e^{-\frac{x^2}{2(t-s)}}\, dx.$$

3. *$B(t, \omega)$ has independent increments, i.e., for any $0 \le t_1 < t_2 < \cdots < t_n$, the random variables*

$$B(t_1), B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1}),$$

are independent.

4. *Almost all sample paths of $B(t, \omega)$ are continuous functions, i.e.,*

$$P\{\omega; B(\cdot, \omega) \textit{ is continuous}\} = 1.$$

From the definition we have:

1. $B(t)$ is normally distributed with mean 0 and variance t.

2. (Translation invariance) For fixed $t_0 \geq 0$, the stochastic process $B'(t) = B(t + t_0) - B(t_0)$ is also a Brownian motion.

3. (Scaling invariance) For any real number $\lambda > 0$, the stochastic process $B'(t) = B(\lambda t)/\sqrt{\lambda}$ is also a Brownian motion.

We can also define $\mathbb{R}^n$-valued Brownian motion, which means that each dimension is a Brownian motion, and they are independent of each other.

1.2 Numerical Approximation of Brownian Motion

The Brownian motion at time t_{n+1} can be approximated by

$$B(t_{n+1}) = \sum_{i=0}^{n} \Delta W_i = \sum_{i=0}^{n} \sqrt{\Delta t_i} \xi_i$$

where

$$\Delta W_i = B(t_{i+1}) - B(t_i), \quad \text{and} \quad \Delta t_i = t_{i+1} - t_i,$$

and ξ_i are independent standard normal random variables.

One popular approximation of Brownian motion in continuous time is piecewise linear approximation, i.e. linear interpolation of $B(t_i)$.

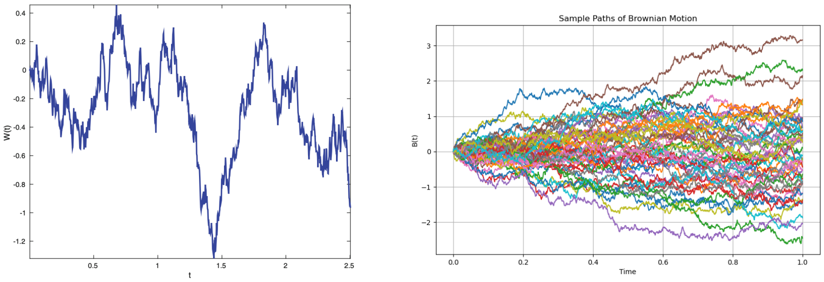

Figure 1: An illustration of sample paths of Brownian motion using cumulative summation of increments.

2 Itô Calculus

2.1 Wiener Integral

Recall the Riemann–Stieltjes integral:

Let g be a monotonically increasing function on a finite closed interval $[a, b]$. A bounded function f defined on $[a, b]$ is said to be Riemann-Stieltjes integrable with respect to g if the following limit exists:

$$\int_a^b f(t)\, dg(t) = \lim_{\|\Delta_n\| \to 0} \sum_{i=1}^{n} f(\tau_i)[g(t_i) - g(t_{i-1})],$$

where $\Delta_n = \{t_0, t_1, \dots, t_{n-1}, t_n\}$ is a partition of $[a, b]$ with the convention $a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b$, $\|\Delta_n\| = \max_{1 \le i \le n}(t_i - t_{i-1})$, and τ_i is an evaluation point in the interval $[t_{i-1}, t_i]$. It is a well-known fact that continuous functions on $[a, b]$ are Riemann-Stieltjes integrable with respect to any monotonically increasing function g on $[a, b]$. In particular, if we set $g(t) = t$, it is reduced to the Riemann integral.

Now let's consider the following integral:

$$\int_a^b f(t) dB(t, \omega)$$

where $f \in L^2([a, b])$, i.e. f is square-integrable on $[a, b]$. We define the Wiener integral in two steps:

Step 1. Suppose f is a step function given by $f = \sum_{i=1}^{n} a_i 1_{[t_{i-1}, t_i)}$, where $t_0 = a$ and $t_n = b$. In this case, define

$$I(f) = \sum_{i=1}^{n} a_i (B(t_{i-1}) - B(t_i))$$

Step 2. Let $f \in L^2[a, b]$. Choose a sequence $\{f_n\}_{n=1}^{\infty}$ of step functions such that $f_n \to f$ in $L^2[a, b]$. We can show that $\{I(f_n)\}_{n=1}^{\infty}$ converges in $L^2(\Omega)$, and we define

$$I(f) = \lim_{n \to \infty} I(f_n), \text{ in } L^2(\Omega).$$

We can show that $I(f)$ is well-defined (see details in [1]) We call $I(f)$ the Wiener integral of f. Note that $I(f)$ is a random variable, i.e., we should write it as $I(f)(\omega)$. It will be denoted by $\int_a^b f(t) dB(t)$ or $\int_a^b f(t) dB(t, \omega)$.

Remark: The convergence in $L^2(\Omega)$ means

$$\lim_{n \to \infty} \mathbb{E}[(I(f_n) - I(f))^2] = 0.$$

2.2 Itô Integral

We have shown how to define $\int_a^b f(t) dB(t, \omega)$ where $f \in L^2([a, b])$, now we can extend it to stochastic f adapted to the Brownian motion $B(t)$. "f is adapted to

$B(t)$ " basically means that $f(t)$ depends only on the information of B available up to time t and not on future information. This condition is satisfied if $f(t)$ is a function of $B(t)$ and t only. A counter-example is $f(t) = B(T)$ for $t < T$, which is not adapted to $B(t)$.

We use the notation $L^2_{ad}([a,b]\times\Omega)$ to denote the set of stochastic processes $f(t,\omega)$ that are adapted to the Brownian motion $B(t)$ and square-integrable on $[a,b]\times\Omega$, i.e. $\int_a^b \mathbb{E}[f^2(t)]\,dt < \infty$.

In an informal way, we can define the Itô integral of $f \in L^2_{ad}([a,b]\times\Omega)$ as follows:

$$\int_a^b f(t,\omega)dB(t,\omega) = \lim_{\|\Delta_n\|\to 0}\sum_{i=1}^n f(t_{i-1},\omega)(B(t_i)-B(t_{i-1})),\ \text{in}\ L^2(\Omega).$$

Here the limit is taken in $L^2(\Omega)$ means that

$$\lim_{\|\Delta_n\|\to 0}\mathbb{E}[\left(\int_a^b f(t,\omega)dB(t,\omega) - \sum_{i=1}^n f(t_{i-1},\omega)(B(t_i)-B(t_{i-1}))\right)^2] = 0.$$

Note that the finite sum in Itô integral is defined at the left-hand points in each subinterval of the partition. This is very important. Indeed, if we define the finite sum at other points, we could get different limits. This is very different from the Riemann integral where the choice of evaluation points does not affect the limit.

For example, if we use the midpoint as the evaluation, We have Stratonovich calculus:

$$\int_a^b f(t,\omega)dB(t,\omega) = \lim_{\|\Delta_n\|\to 0}\sum_{i=1}^n f(\frac{t_{i-1}+t_i}{2},\omega)(B(t_i)-B(t_{i-1})),\ \text{in}\ L^2(\Omega).$$

Example:

$$\int_0^t B(s)dB(s) = \frac{1}{2}B(t)^2 - \frac{1}{2}t.$$

$$\int_0^t B(s)\circ dB(s) = \frac{1}{2}B(t)^2.$$

Let's check it:

$$\begin{aligned}
&\mathbb{E}[\left(\frac{1}{2}B(t)^2 - \frac{1}{2}t - \sum_{i=1}^n B(t_{i-1})(B(t_i)-B(t_{i-1}))\right)^2]\\
=&\mathbb{E}[\left(\frac{1}{2}B(t)^2 - \frac{1}{2}t - \frac{1}{2}\sum_{i=1}^n(B^2(t_i)-B^2(t_{i-1})) + \frac{1}{2}\sum_{i=1}^n(B(t_i)-B(t_{i-1}))^2\right)^2]\\
=&\mathbb{E}[\left(-\frac{1}{2}t + \frac{1}{2}\sum_{i=1}^n(B(t_i)-B(t_{i-1}))^2\right)^2] \to 0
\end{aligned}$$

Suppose $t_i = it/n$, then $B(t_i) - B(t_{i-1})$ is Gaussian with variance t/n. The Stratonovich integral can be verified similarly.

Theorem 1. *(Itô isometry) Suppose $f \in L^2_{ad}([a,b] \times \Omega)$. Then the Itô integral $I(f) = \int_a^b f(t)\, dB(t)$ is a random variable with $\mathbb{E}[I(f)] = 0$ and*

$$\mathbb{E}[I(f)^2] = \int_a^b \mathbb{E}[f^2(t)]\, dt.$$

For any $f, g \in L^2_{ad}([a,b] \times \Omega)$, the following equality holds:

$$\mathbb{E}\left[\int_a^b f(t)\, dB(t) \int_a^b g(t)\, dB(t)\right] = \int_a^b \mathbb{E}[f(t)g(t)]\, dt.$$

Remark: For deterministic f, the Itô integral is reduced to the Wiener integral defined above. In this case, the Wiener integral $\int_a^b f(t)\, dB(t)$ is Gaussian. We can show this starting from step functions, and then take the limit.

2.3 Itô's Formula

Recall the chain rule of Riemann integral. If f and g are differentiable, then $f(g(t))$ is also differentiable and has derivative

$$\frac{d}{dt} f(g(t)) = f'(g(t))g'(t)$$

And

$$f(g(t)) = f(g(t_0)) + \int_{t_0}^t f'(g(s))dg(s)$$

For g as a Brownian motion, the above equality doesn't hold. Actually $g'(t)$ makes no sense since the Brownian motion is nowhere differentiable.

Theorem 2. *(Itô formula in a simple form) If f and its first two derivatives are continuous on $\mathbb{R}$, then it holds with probability one (almost surely, a.s.) that*

$$f(B(t)) = f(B(t_0)) + \int_{t_0}^t f'(B(s))\, dB(s) + \frac{1}{2}\int_{t_0}^t f''(B(s))\, ds.$$

So we can compute the Itô integral $\int_{t_0}^t f'(B(s))\, dB(s)$ by

$$\int_{t_0}^t f'(B(s))\, dB(s) = f(B(t)) - f(B(t_0)) - \frac{1}{2}\int_{t_0}^t f''(B(s))\, ds.$$

Theorem 3. *(Itô formula in a general form) Let X_t be a stochastic process given by*

$$X_t = X_a + \int_a^t f(s)\, dB(s) + \int_a^t g(s)\, ds, \quad a \le t \le b.$$

Suppose $\theta(t,x)$ is a continuous function with continuous partial derivatives $\frac{\partial\theta}{\partial t}$, $\frac{\partial\theta}{\partial x}$, and $\frac{\partial^2\theta}{\partial x^2}$. Then $\theta(t,X_t)$ satisfies

$$\theta(t,X_t) = \theta(a,X_a) + \int_a^t \frac{\partial\theta}{\partial x}(s,X_s)f(s)\,dB(s)$$

$$+\int_a^t \left[\frac{\partial\theta}{\partial t}(s,X_s) + \frac{\partial\theta}{\partial x}(s,X_s)g(s) + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}(s,X_s)f(s)^2\right] ds. \tag{7.4.3}$$

The proof of Itô's formula is out of the scope of this course.

A good way to memorize the above integral equation is the "symbolic derivation of differential form". First, apply the Taylor expansion to get

$$d\theta(t,X_t) = \frac{\partial\theta}{\partial t}(t,X_t)\,dt + \frac{\partial\theta}{\partial x}(t,X_t)\,dX_t + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}(t,X_t)(dX_t)^2.$$

Then replace dX_t with $f(t)dB(t)+g(t)dt$. Also, $(dX_t)^2 = (f(t)dB(t)+g(t)dt)^2 \approx f^2(t)dt$. Therefore,

$$\begin{aligned} d\theta(t,X_t) &= \frac{\partial\theta}{\partial t}\,dt + \frac{\partial\theta}{\partial x}\left(f(t)\,dB(t)+g(t)\,dt\right) + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}f(t)^2\,dt \\ &= \frac{\partial\theta}{\partial x}f(t)\,dB(t) + \left(\frac{\partial\theta}{\partial t} + \frac{\partial\theta}{\partial x}g(t) + \frac{1}{2}\frac{\partial^2\theta}{\partial x^2}f(t)^2\right)dt. \end{aligned}$$

For computation we can always use this kind of symbolic derivation on stochastic differentials to get the results. However, we need to emphasize that this derivation is not a proof. It just happens to produce the correct results.

Example 1: $f(x) = x^2$

$$\int_a^b B(t)dB(t) = \frac{1}{2}(B^2(b) - B^2(a) - (b-a))$$

This is consistent with the example we have seen when introducing Itô integral.

Example 2: Langevin equation

$$X_t = x_0 + \alpha B(t) - \beta\int_0^t X_s ds$$

where $\alpha \in \mathbb{R}$ and $b > 0$. This "stochastic differential equation" can also be written as

$$dX_t = \alpha dB(t) - \beta X_t dt, \quad X_0 = x_0,$$

Let $\theta(t,x) = e^{\beta t}x$. Then

$$\frac{\partial\theta}{\partial t} = \beta e^{\beta t}x, \quad \frac{\partial\theta}{\partial x} = e^{\beta t}, \quad \text{and} \quad \frac{\partial^2\theta}{\partial x^2} = 0.$$

Hence by Itô's formula, we have

$$\begin{aligned} d(e^{\beta t}X_t) &= \beta e^{\beta t}X_t\,dt + e^{\beta t}\,dX(t) \\ &= \beta e^{\beta t}X_t\,dt + e^{\beta t}\,(\alpha\,dB(t) - \beta X_t\,dt) \\ &= \alpha e^{\beta t}\,dB(t). \end{aligned}$$

or the integral form:

$$e^{\beta t}X_t = X_0 + \int_0^t \alpha e^{\beta s}\,dB(s), \quad 0 \le t.$$

i.e.

$$X_t = e^{-\beta t}x_0 + \alpha\int_0^t e^{-\beta(t-s)}\,dB(s).$$

The solution X_t is called an Ornstein–Uhlenbeck process.

Question: What is the distribution of X_t at $t\ > 0$?

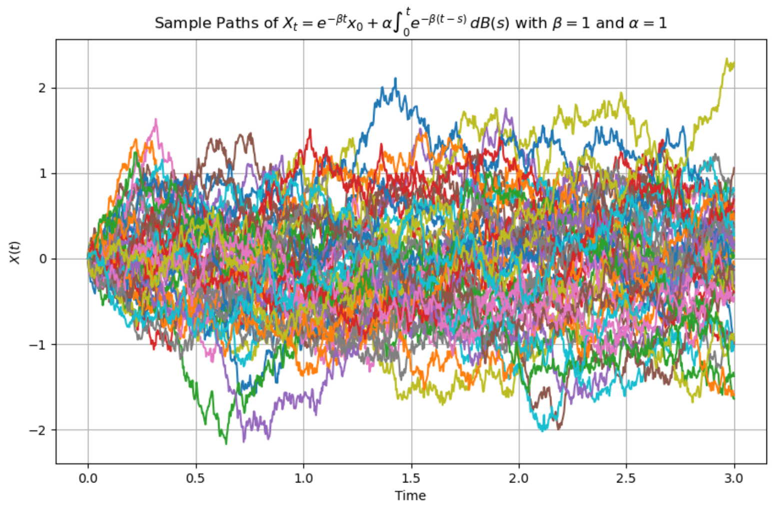

Figure 2: An illustration of sample paths of Ornstein–Uhlenbeck process

3 Stochastic Ordinary Differential Equations

3.1 Existence and Uniqueness

Definition 3. *Let X_t be a $\mathbb{R}^n$-valued stochastic process, $B(t)$ is $\mathbb{R}^d$-valued Brownian motion, $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0,T] \times \mathbb{R}^{n\times d} \to \mathbb{R}^n$. We say X_t is a strong solution of the stochastic ordinary differential equation*

$$X_0 = x$$

$$dX_t = f(t, X_t)dt + \sigma(t, X_t)dB(t), \quad \forall t \in [0, T]$$

if

1. *X_t is almost surely continuous and adapted to $B(t)$.*
2. *$f(t, X_t) \in L^1([0, T])$ almost surely.*
3. *$\sigma(t, X_t) \in L^2_{ad}([0, T] \times \Omega)$, so that the Itô's integral is well-defined.*
4. *The stochastic integral equation*

$$X_0 = x$$

$$X_t = x + \int_0^t f(s, X_s)ds + \int_0^t \sigma(s, X_s)dB(s), \quad \forall t \in [0, T]$$

is satisfied almost surely.

Remark: The condition for σ can be relaxed to weaker conditions, but here we keep it as above, so that it's consistent with the condition in the definition of Itô's integral we introduced.

Theorem 4. *Suppose there exists constant C and K, s.t. $\forall t \in [0, T], x, y \in \mathbb{R}^n$*

$$|f(t, x)| + |\sigma(t, x)|_F \le C(1 + |x|), \quad \textit{(linear growth condition)}$$

and

$$|f(t, x) - f(t, y)| + |\sigma(t, x) - \sigma(t, y)|_F \le K|x - y| \quad \textit{(Lipschitz condition)}$$

Assume furthermore that the initial condition x is a random variable independent of the Brownian motion W_t with

$$\mathbb{E}|x|^2 < \infty.$$

Then the SDE defined above has a unique strong solution X_t with

$$\mathbb{E}\left[\int_0^t |X_s|^2 \, ds\right] < \infty \tag{3.39}$$

for all $t > 0$.

Here $|\cdot|_F$ is the Frobenius norm. For a matrix $A \in \mathbb{R}^{m \times n}$, the Frobenius norm is defined as:

$$|A|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2\right)^{1/2}$$

By uniqueness of strong solutions we mean that, if X_t and Y_t are strong solutions, then $X_t = Y_t$ for all t almost surely.

3.2 Numerical Methods for SODEs

Numerical methods are essential for solving stochastic ordinary differential equations (SODEs) when analytical solutions are not available. This section introduces two widely used numerical schemes: the Euler-Maruyama scheme and the Milstein scheme. More details on this topic can be found in [2].

3.2.1 Euler-Maruyama Scheme

The Euler-Maruyama scheme is a simple and popular method for approximating solutions to SODEs. It is a straightforward extension of the Euler method for ordinary differential equations to the stochastic case.

Consider the SODE

$$dX_t = f(t, X_t)\, dt + \sigma(t, X_t)\, dB(t), \quad X_0 = x_0,$$

where X_t is the state variable, f is the drift coefficient, σ is the diffusion coefficient, and $B(t)$ is a Brownian motion.

The Euler-Maruyama scheme discretizes the time interval $[0, T]$ into N steps with a fixed step size $\Delta t = \frac{T}{N}$. The approximation X_n of X_{t_n} at the discrete times $t_n = n\Delta t$ is given by:

$$X_{n+1} = X_n + f(t_n, X_n)\Delta t + \sigma(t_n, X_n)\Delta B_n,$$

where $\Delta B_n = B(t_{n+1}) - B(t_n)$ are the increments of the Brownian motion, which are normally distributed with mean 0 and variance Δt, i.e., $\Delta B_n \sim \mathcal{N}(0, \Delta t)$.

The Euler-Maruyama scheme is easy to implement and computationally efficient. However, its accuracy is limited, especially for problems requiring high precision.

3.2.2 Milstein Scheme

The Milstein scheme improves upon the Euler-Maruyama scheme by including an additional term that accounts for the variation in the diffusion coefficient. It is particularly useful for SODEs where the diffusion term $\sigma(t, X_t)$ is not constant.

The Milstein scheme for the SODE

$$dX_t = f(t, X_t)\, dt + \sigma(t, X_t)\, dB(t), \quad X_0 = x_0,$$

is given by:

$$X_{n+1} = X_n + f(t_n, X_n)\Delta t + \sigma(t_n, X_n)\Delta B_n + \frac{1}{2}\sigma(t_n, X_n)\sigma'(t_n, X_n)\left((\Delta B_n)^2 - \Delta t\right),$$

where σ' denotes the derivative of σ with respect to X.

The Milstein scheme generally provides better accuracy than the Euler-Maruyama scheme, particularly for problems where the diffusion coefficient has significant variability.

4 Fokker-Planck Equation

Consider the stochastic differential equation (SODE)

$$dX_t = f(t, X_t)\, dt + \sigma(t, X_t)\, dB(t), \quad X_0 = x_0,$$

Informally, suppose f and σ satisfy the linear growth, Lipschitz condition, and additional differentiable regularity. Also, suppose X_t at t has the density $p(t, \cdot)$, then the density function $p(t, x)$ of the solution X_t satisfies the Fokker-Planck equation:

$$\frac{\partial p(t,x)}{\partial t} = -\frac{\partial}{\partial x}\left[f(t,x)p(t,x)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma^2(t,x)p(t,x)\right].$$

In higher dimensions, let X_t be a $\mathbb{R}^n$-valued stochastic process, $B(t)$ is $\mathbb{R}^d$-valued Brownian motion, $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0,T] \times \mathbb{R}^{n\times d} \to \mathbb{R}^n$. The Fokker-Planck equation is given by:

$$\frac{\partial p(t,\mathbf{x})}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}\left[f_i(t,\mathbf{x})p(t,\mathbf{x})\right] + \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2}{\partial x_i \partial x_j}\left[D_{ij}(t,\mathbf{x})p(t,\mathbf{x})\right],$$

where the diffusion tensor $\mathbf{D}$ is given by:

$$D_{ij}(t,\mathbf{x}) = \frac{1}{2}\sum_{k=1}^{d}\sigma_{ik}(t,\mathbf{x})\sigma_{jk}(t,\mathbf{x}).$$

The Fokker-Planck equation describes the time evolution of the probability density function of the solution to the SODE. The connection of the Fokker-Planck equation to the SODE is very important in understanding the behavior of stochastic processes and generative diffusion models.

Example 1: Brownian motion, $f(x) = 0$, $\sigma(x) = 1$. The Fokker-Planck equation is

$$\frac{\partial p(t,x)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}p(t,x).$$

If $X_0 = 0$, the solution is the Gaussian distribution

$$p(t,x) = \frac{1}{\sqrt{2\pi t}}\exp(-x^2/2t).$$

Example 2: Ornstein–Uhlenbeck process, $f(x) = -\beta x$, $\sigma(x) = \alpha$. The Fokker-Planck equation is

$$\frac{\partial p(t,x)}{\partial t} = \beta\frac{\partial}{\partial x}(xp(t,x)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\alpha^2 p(t,x)).$$

Recall the solution of Langevin equation we obtained

$$X_t = e^{-\beta t}x_0 + \alpha\int_0^t e^{-\beta(t-s)}\, dB(s).$$

According to Itô isometry, the distribution of X_t is Gaussian with mean $e^{-\beta t}x_0$ and variance

$$\alpha^2 \int_0^t (e^{-\beta(t-s)})^2 ds = \alpha^2(1-e^{-2\beta t})/(2\beta) := h^2(t)$$

i.e.

$$p(X_t = x|X_0 = x_0) = \frac{1}{\sqrt{2\pi h^2(t)}} \exp(-\frac{(x-e^{-\beta t}x_0)^2}{2h^2(t)})$$

You can verify that $p(X_t = x|X_0 = x_0)$ satisfies the Fokker-Planck equation.

Suppose the intial distribution of X_0 is $p_0(x)$, then the distribution density of X_t is

$$p(t,x) = \int_{\mathbb{R}} p(X_t = x|X_0 = y)p_0(y)dy.$$

which also satisfies the Fokker-Planck equation since the Fokker-Planck equation is linear.

Some properties of the Ornstein–Uhlenbeck process:

1. Start from any fixed x_0, the process X_t converges to the stationary distribution, which is Gaussian with zero mean and variance $\alpha^2/(2\beta)$.

2. Actually if we start from a bounded distribution, the process X_t will converge to the stationary distribution exponentially fast. In other words, for a sufficiently large t, the distribution of X_t is very close to the stationary distribution and "forgets" the initial distribution. As we will see later, this property is very important in generative models.

5 Diffusion Models and Score Matching

A diffusion model aims to build a function that maps a sample from a simple distribution (e.g., Gaussian) to a more complex distribution. Mapping from a complex distribution to a Gaussian distribution is easy. We just saw an example: the Ornstein–Uhlenbeck process. The reverse direction is much more challenging. But the idea is that, given a forward process with $p(t,x)$ satisfies the corresponding Fokker-Planck equation, can we define a backward process which follows the same Fokker-Planck equation but in the reverse direction? In other words, we want the density evolution of the backward process to be the same as the forward process but in the reverse direction.

5.1 Forward and Reverse SODE

Consider the forward SODE

$$dX_t = f(X_t,t)dt + g(t)dB_t$$

For simplicity, we assume g is independent of X_t.

The corresponding Fokker-Planck equation for density $q(x,t)$ is

$$\frac{\partial}{\partial t}q(x,t) = -\nabla \cdot [f(x,t)q(x,t)] + \frac{1}{2}g^2(t)\nabla^2 q(x,t)$$

Suppose the reverse SODE is

$$d\bar{X}_\tau = \bar{f}(\bar{X}_\tau, \tau)d\tau + \bar{g}(\tau)d\bar{B}_\tau,$$

where $\tau = T - t$ is the reverse time, $d\bar{B}_\tau$ is another Brownian motion. It has the corresponding Fokker-Planck equation

$$\frac{\partial}{\partial \tau}p(x,\tau) = -\nabla \cdot [\bar{f}(x,\tau)p(x,\tau)] + \frac{1}{2}\bar{g}^2(\tau)\nabla^2 p(x,\tau)$$

i.e.,

$$-\frac{\partial}{\partial t}p(x,T-t) = -\nabla \cdot [\bar{f}(x,T-t)p(x,T-t)] + \frac{1}{2}\bar{g}^2(T-t)\nabla^2 p(x,T-t)$$

Note the negative sign in LFH. Move it to RHS:

$$\frac{\partial}{\partial t}p(x,T-t) = \nabla \cdot [\bar{f}(x,T-t)p(x,T-t)] - \frac{1}{2}\bar{g}^2(T-t)\nabla^2 p(x,T-t)$$

Now compare the above equation with the Fokker-Planck equation of forward SODE. In order to have $p(x,T-t) = q(x,t)$, we only need

$$\bar{g}^2(T-t) = g(t)$$

$$\bar{f}(x,T-t) = -f(x,t) + g^2(t)\nabla \log q(x,t)$$

Note that the reverse dynamic is not unique. For example, we can also remove the diffusion term in the reverse process, i.e.

$$\bar{g}^2(T-t) = 0$$

$$\bar{f}(x,T-t) = -f(x,t) + \frac{1}{2}g^2(t)\nabla \log q(x,t)$$

Now the key challenge is to calculate/approximate $\nabla \log q(x,t)$. This function is the so-called score function.

5.2 Approximating Score Function

We need to approximate $\nabla_x \log q(x,t)$ (score function) with samples of $q(x,t)$, i.e., X_t, which can be obtained from the forward SODE simulation.

The key equation is:

$$\nabla_x \log p(x) = \arg\min_{h(x)} E_{p(x,y)}||h(x) - \nabla_x \log p(x|y)||^2$$

To show this, we take the variation of the right hand side w.r.t. $h(x)$ at $h(x) = \nabla_x \log p(x)$, we need

$$E_{p(x,y)}[\epsilon(x)\nabla_x \log p(x) - \epsilon(x)\nabla_x \log p(x|y)] = 0, \forall \epsilon(x)$$

i.e. we require

$$\nabla_x \log p(x) = E_{p(y|x)} \nabla_x \log p(x|y)$$

This is true, since

$$\begin{aligned} RHS =& E_{p(y|x)} \nabla_x \log p(y|x) + E_{p(y|x)} \nabla_x \log p(x) \\ =& \int \nabla_x p(y|x) dy + \nabla_x \log p(x) \\ =& \nabla_x \int p(y|x) dy + \nabla_x \log p(x) \\ =& \nabla_x \log p(x) \end{aligned}$$

If we set $x = x_t$, $y = x_0$, $p(x|y) = q(x_t|x_0)$, we have

$$\nabla_{x_t} \log q(x_t, t) = \arg \min_{h(x_t)} E_{q(x_0,x_t)} ||h(x_t) - \nabla_{x_t} \log q(x_t|x_0)||^2$$

This is exactly the score-matching loss in training diffusion models.

Now this loss function is tractable:

1. $\nabla_{x_t} \log q(x_t|x_0)$ can has analytical form if we know the forward SODE. For example, if the forward SODE is Ornstein–Uhlenbeck process, $q(x_t|x_0)$ is the Gaussian distribution, and

$$\nabla_{x_t} \log q(x_t|x_0) = \frac{x_t - e^{-\beta t} x_0}{\alpha^2 (1 - e^{-2\beta t})/(2\beta)}$$

2. Samples of $q(x_0, x_t)$ is accessible through forward SODE, so the expectation can be evaluated via Monte Carlo.

3. The score function doesn't need the normalization term $\nabla_{x_t} \log q(x_t, t)$, so we can use a neural network to represent it.

References

[1] H.H. Kuo. *Introduction to Stochastic Integration.* Universitext. Springer New York, 2006.

[2] Zhongqiang Zhang and George Em Karniadakis. *Numerical methods for stochastic partial differential equations with white noise*, volume 196. Springer, 2017.