[AI2613 Lecture 13] Diffusions, Langevin Dynamics

June 9, 2023

1 Diffusions

1.1 The Definition of Diffusion

A continuous stochastic process with Markov property is called a diffusion.[1] In other words, a diffusion can be viewed as a Markov process in continuous time with continuous sample paths.

[1] This is an informal definition of diffusions and it is enough for this course.

Actually, diffusions can be built up from local Brownian motions in the same way as differentiable functions being built up from local linear functions. Imagine that we want to draw the image of a function f with knowing $f'(t) = e^t$ and $f(0) = 1$. How to do this if you are not allowed to integrate $f'(t)$. A natural idea is to approximate f using segmented linear functions:

- Select a step length h;
- Draw a segment on $[0, h]$ which starts from $(0, f(0))$ with slope $f'(0) = 1$;
- Draw a segment on $[h, 2h]$ which starts from $(h, h + f(0))$ with slope $f'(h) = e^h$;
- ....

When $h \to 0$, our drawing is exactly the image of f. This gives an intuition that a differentiable function can be locally approximated as linear functions.

A diffusion $\{X(t)\}_{t\geq 0}$ is the stochastic analog of above process. That is, if we are currently at the position $X(t) = X_t$ and consider the small time interval $[t, t+h]$, the process acts as a $\left(\mu(X_t), \sigma^2(X_t)\right)$ Brownian motion where μ and σ^2 are functions of the position X_t. Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian. We can break the process into segments and use these normal random variables to simulate the diffusion:

- $X_h = X_0 + \mu(X_0)h + \sigma(X_0)\sqrt{h} \cdot Z_1$;
- $X_{2h} = X_h + \mu(X_h)h + \sigma(X_h)\sqrt{h} \cdot Z_2$;
- ...,

where each Z_i are independent standard Gaussian. Then for any $k \in \mathbb{N}$, $X_{(k+1)h} - X_{kh} \sim \mathcal{N}\left(\mu(X_{kh})h, \sigma^2(X_{kh})h\right)$.

Thus, when $h \to 0$, we can naturally develop a specification of diffusion: A time homogeneous diffusion can be specified by two functions $\mu(x)$ and $\sigma^2(x)$ which satisfies:

- $\forall t$, $\mathbf{E}\left[X(t+h)-X(t) \mid X(t)=x\right]=\mu(x)h+o(h)$;
- $\forall t$, $\mathbf{Var}\left[X(t+h)-X(t) \mid X(t)=x\right]=\sigma^2(x)h+o(h)$;
- $\forall t$, $\mathbf{E}\left[|X(t+h)-X(t)|^p \,\middle|\, X(t)=x\right]=o(h)$ for $p>2$.

Note that

$$\begin{aligned}&\mathbf{Var}\left[X(t+h)-X(t) \mid X(t)=x\right]\\ &=\mathbf{E}\left[(X(t+h)-X(t))^2 \,\middle|\, X(t)=x\right]-\left(\mathbf{E}\left[X(t+h)-X(t) \mid X(t)=x\right]\right)^2\\ &=\mathbf{E}\left[(X(t+h)-X(t))^2 \,\middle|\, X(t)=x\right]-\left(\mu(x)h+o(h)\right)^2.\end{aligned}$$

Thus

$$\mathbf{Var}\left[X(t+h)-X(t) \mid X(t)=x\right]=\sigma^2(x)h+o(h)$$

is equivalent to

$$\mathbf{E}\left[(X(t+h)-X(t))^2 \,\middle|\, X(t)=x\right]=\sigma^2(x)h+o(h).$$

Recall that in the analog of differentiable functions, we have $\mathrm{d}f(t)=g(t)\,\mathrm{d}t$ where $g(t)$ is the derivative of f. Similarly, for a diffusion $\{X(t)\}_{t\geq 0}$ specified by $\mu(x)$ and $\sigma^2(x)$, we can write it as

$$\mathrm{d}X(t)=\mu(X(t))\,\mathrm{d}t+\sigma(X(t))\,\mathrm{d}W(t),$$

where $\{W(t)\}$ is the standard Brownian motion and $\mathrm{d}W(t)$ can be understood as $\lim_{h\to 0} W(t+h)-W(t)$.

This expression is informal as we haven't give a mathematical meaning to the notation $dW(t)$. We will do this in the next lecture, but for now, we can sloppily understand it as a Brownian motion in an infinitisimal time.

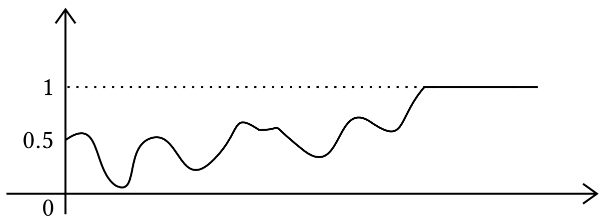
Example 1 (Ornstein-Uhlenbeck Process). *Consider a diffusion $\{X(t)\}_{t\geq 0}$ specified by $\sigma(x)=-x$ and $\sigma^2(x)=2$ with $X(0)=0$. This diffusion always has a tendency to 0 since if $X(t)$ is large, $\mu(X(t))$ is also large towards the reverse direction which acts as a spring intuitively. We can write this process as*

$$\mathrm{d}X(t)=-X(t)\,\mathrm{d}t+\sqrt{2}\,\mathrm{d}W(t).$$

The process can be used to model the discrete Ehrenfest chain. Suppose we have two boxes with a balls in the first box and b balls in the second box in the initial state. In each round, we choose a ball uniformly at random among the $a+b$ balls and put the chosen ball into the other box. It is more likely to choose the balls in the box with more balls. Thus, this discrete Markov process tends to the equilibrium state where each box has $\frac{a+b}{2}$ balls.

Example 2 (Wright-Fisher Process). *Next we consider a stochastic random walk with absorbing boundaries. Let $\mu(x)=0$, $\sigma^2(x)=x(1-x)$ and $X(0)=\frac{1}{2}$. Then this diffusion is jittery around $\frac{1}{2}$ and is more steady around the boundaries.*

The process can be used to model the following model of racial reproduction. Assume the total population is N which is invariant over time. At the t-th generation, there is X_t black people and $N-X_t$ white people where X_t is a non-negative random variable. Assume that there is no interracial marriage and



the child's race is the same with his or her parents. At the $t+1$-th generation, each person is white w.p. $1-\frac{X_t}{N}$ and is black w.p. $\frac{X_t}{N}$. Assume the race of each individual is independent with other people. If it starts with half white and half black, then we want to ask: Will there be genocide after a long period of time or will the two races tend to keep a balance?

The continuous version of the model is the Wright-Fisher process we just introduced. It is equivalent to ask whether the process tends to keep jittery or be absorbed. Since it seems to be "lazier" when it comes closer to the boundary, the answer of this question is not obvious. In fact, however, after a sufficiently long time, it does reach the boundary.

1.2 Geometric Brownian Motion

Let $\{X(t)\}$ be a (μ, σ^2) Brownian motion, that is, $\mathrm{d}X(t) = \mu\,\mathrm{d}t + \sigma dW(t)$. Define $Y(t) = e^{X(t)}$. Then $\{Y(t)\}$ is called a geometric Brownian motion. Geometric Brownian motion is widely applied to model the stock prices in finance. In fact, we can consider a more generalized situation that $\{Y(t)\}$ is defined by $Y_t = f(X_t)$ where f is strictly monotone and twice differentiable. Then we have the following proposition.

Proposition 1. *Suppose $\{X(t)\}$ is a diffusion specified by $\mu_X(x)$ and $\sigma_X^2(x)$. Let f be a strictly monotone and twice differentiable function. Define $Y(t) = f(X(t))$. Then $\{Y(t)\}$ is a diffusion specified by $\mu_Y(y)$ and $\sigma_Y^2(y)$ which satisfy*

$$\mu_Y(y) = \mu_X(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x) \quad \text{and} \quad \sigma_Y^2(y) = \left(f'(x)\right)^2\sigma_X^2(x)$$

where $x = f^{-1}(y)$.

Proof. For a small h, we have

$$\begin{aligned}
&\mathbf{E}\left[Y(t+h) - Y(t) \mid Y(t) = y\right] \\
=&\mathbf{E}\left[f(X(t+h)) - f(X(t)) \mid X(t) = x\right] \\
=&\mathbf{E}\left[f'(X(t))(X(t+h) - X(t)) + \frac{1}{2}f''(X(t))(X(t+h) - X(t))^2 \,\middle|\, X(t) = x\right] \\
&+ o(h) \\
=&\mu_X(x)f'(x)h + \frac{1}{2}\sigma^2(x)f''(x)h + o(h),
\end{aligned}$$

so that

$$\mu_Y(y) = \lim_{h\to 0} \frac{\mathbf{E}\left[Y(t+h) - Y(t) \mid Y(t) = y\right]}{h} = \mu_X(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x).$$

Similarly, we have

$$\begin{aligned}
&\mathbf{E}\left[(Y(t+h) - Y(t))^2 \,\middle|\, Y(t) = y\right] \\
=&\mathbf{E}\left[(f(X(t+h)) - f(X(t)))^2 \,\middle|\, X(t) = x\right] \\
=&\mathbf{E}\left[(f'(X(t))(X(t+h)) - X(t))^2 \,\middle|\, X(t) = x\right] + o(h) \\
=&\left(f'(x)\right)^2 \sigma_X^2(x)h + o(h),
\end{aligned}$$

so that

$$\sigma_Y^2(y) = \lim_{h\to 0} \frac{\mathbf{E}\left[(Y(t+h) - Y(t))^2 \,\middle|\, Y(t) = y\right]}{h} = \left(f'(x)\right)^2 \sigma_X^2(x).$$

□

2 The Langevin Dynamics

2.1 Fokker-Planck Equation

The diffusion

$$\mathrm{d}Y(t) = \mu(y)\,\mathrm{d}t + \sigma^2(y)\,\mathrm{d}W_t$$

is a continuous time Markov chain. If we use $g(t, y)$ to denote the density of $Y(t)$, then it satisfies the following PDE, called *Kolmogorov forward equation* or *Fokker-Plank equation*.

$$\frac{\partial}{\partial t} g(t, y) = -\frac{\partial}{\partial y}\left(\mu(y) \cdot p(t, y)\right) + \frac{\partial^2 a}{\partial y^2}\left(\frac{1}{2}\sigma^2(y) \cdot p(t, y)\right).$$

In applications, it is more useful to study diffusions in high dimension. Let $n \geq 1$ be an integer. An n-dimensional diffusion $\mathbf{X}_t$ can be described by

$$\mathrm{d}\mathbf{X}_t = \mu(t, \mathbf{X}_t)\mathrm{d}t + \Sigma(t, \mathbf{X}_t)\mathrm{d}\mathbf{W}_t, \tag{1}$$

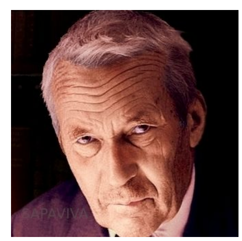


Figure 1: Andrey Kolmogorov

where $\mathbf{X}_t, \mu(t, \mathbf{x}) \in \mathbb{R}^n$, $\Sigma(t, \mathbf{x}) \in \mathbb{R}^{n\times m}$ and $\mathbf{W}_t$ is an m-dimensional standard Brownian motion. Its density $p(t, \mathbf{x})$ satisfies the following n-dimensional Fokker-Planck equation:

$$\frac{\partial}{\partial t} p(t, \mathbf{x}) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\mu_i(t, \mathbf{x}) p(t, \mathbf{x})\right) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j}\left(\left(\Sigma\Sigma^\top\right)_{ij} \cdot p(t, \mathbf{x})\right). \tag{2}$$

The derivation of Equation (2) is beyond the scope of this lecture and you can find a proof in e.g. [SS19].

3 *The Langevin Dynamics*

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The *Langevin dynamics* is the following diffusion:

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\cdot dW_t, \tag{3}$$

where $X_t \in \mathbb{R}^n$ and $W_t \in \mathbb{R}$ is the standard n-dimensional Brownian motion (that is, each coordinate is an independent standard Brownian motion).

It is easy to verify that $g(t,\mathbf{x}) = p(\mathbf{x}) = e^{-f(\mathbf{x})}$ satisfies Equation (2) and therefore it is a stationary distribution of the Langevin dynamics. As a result, Langevin dynamics can be used to in MCMC algorithm to sample from a target distribution $p(\mathbf{x}) = e^{-f()}$.

Are you familiar with this expression? Let's remove the random part, and it becomes to the ODE

$$dX_t = -\nabla f(X_t)dt.$$

The discrete version of above is

$$X_{t+1} - X_t = -\nabla f(X_t).$$

Yes, Langevin dynamics is simply gradient descent with some additional *Gaussian white noises.*

We will study the convergence of eq. (3) for *strongly convex* f. Recall a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is m-strongly convex if for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, it holds that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{m}{2}\cdot \|\mathbf{y}-\mathbf{x}\|_2^2. \tag{4}$$

Swapping $\mathbf{x}$ and $\mathbf{y}$, we obtain

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{m}{2}\cdot \|\mathbf{y}-\mathbf{x}\|_2^2. \tag{5}$$

Adding eq. (5) and eq. (4), we obtain

$$\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geq m \cdot \|\mathbf{x}-\mathbf{y}\|_2^2. \tag{6}$$

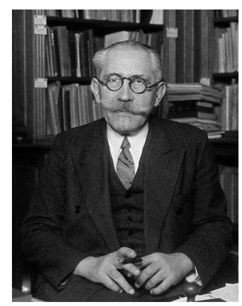


Figure 2: Paul Langevin

Let π be the stationary distribution of the Langevin dynamics. Formally, $\pi(dx) = e^{-f(x)}dx$. We want to study the *convergence rate* of X_t to its stationary distribution π. To this end, we let Y_t be another instance of eq. (3); that is,

$$dY_t = -\nabla f(Y_t)\,dt + \sqrt{2}\cdot dW_t,$$

and we let $Y_0 \sim \pi$.

We now *couple* X_t and Y_t using the *same Brownian motion* W_t and prove that the distance $\|X_t - Y_t\|_2^2$ decays exponentially with respect to t. For every $\mathbf{x}$, define $\Phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$. Then

$$\frac{d}{dt}\Phi(\mathbf{x}) = \langle X_t - Y_t, \nabla f(Y_t) - \nabla f(X_t)\rangle.$$

Plugging in eq. (6), we obtain

$$\frac{d}{dt}\Phi(X_t - Y_t) \leq -2m\cdot \Phi(X_t - Y_t).$$

This implies that

$$\|X_t - Y_t\|_2^2 \leq \|X_0 - Y_0\|^2 \cdot e^{-2mt}.$$

References

[SS19] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019. 4