Unit 10: Langevin Diffusion

10.1 Brownian Motion and Stochastic Differential Equations

Let $Z_1, Z_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. Define $S_n = Z_1 + \cdots + Z_n$ with $S_0 = 0$. Define $(X_t^{(n)})_{0 \leq t \leq 1}$ as the scaled linear interpolation of $(S_j)_{1 \leq j \leq n}$:

$$X_t^{(n)} = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor} + \frac{nt - \lfloor nt \rfloor}{\sqrt{n}} Z_{\lfloor nt \rfloor + 1}.$$

Note that the trajectory $(X_t^{(n)})_{0 \leq t \leq 1}$ is very easy to simulate and visualize. First, one generates $Z_1, \ldots, Z_n$ and calculates the partial sums $(S_k)_{0 \leq k \leq n}$. Then, one calculates $X_{k/n}^{(n)} = S_k / \sqrt{n}$ for each $k = 0, 1, \ldots, n$. Finally, connecting $(X_{k/n}^{(n)})_{0 \leq k \leq n}$ by straight lines yields the trajectory $(X_t^{(n)})_{0 \leq t \leq 1}$. An application of multivariate CLT yields the following result.

Exercise 10.1. Prove that

(i) for any fixed $t \in (0, 1]$, $X_t^{(n)} \overset{\mathrm{D}}{\to} N(0, t)$ as $n \to \infty$;

(ii) for any fixed $0 < s < t \leq 1$, $(X_s^{(n)}, X_t^{(n)} - X_s^{(n)})$ converges in distribution, as $n \to \infty$, to a bivariate normal random vector with independent coordinates.

Moreover, letting $X^{(n)} = (X_t^{(n)})_{0 \leq t \leq 1}$, we have that $X^{(n)}$ converges weakly to a stochastic process, $B = (B_t)_{0 \leq t \leq 1}$, which is called the (one-dimensional) Brownian motion or the Wiener process; see Figure 1 for a simulated trajectory. A lot of technical details are omitted here (e.g., why does this limit exist), since they are not important to the development of the sampling algorithms we will discuss. This construction can be extended to the time interval $t \in [0, \infty)$ straightforwardly, and it is clear from the construction that the resulting process $B = (B_t)_{t > 0}$ satisfies the following conditions:

(i) $B_0 = 0$;

(ii) $B_{t_0}, B_{t_1} - B_{t_0}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent for any $0 \leq t_0 < t_1 < \cdots < t_n < \infty$;

(iii) for any $s, t \geq 0$, $B_{s+t} - B_s \sim N(0, t)$;

(iv) sample paths of B are a.s. continuous.

Indeed, Brownian motion is *defined* as the stochastic process that satisfies the above four conditions. The following result (proof omitted) is known as the scaling property of the Brownian motion.

Theorem 10.1. *Let B_t be a Brownian motion. Then, for any $\lambda > 0$, $\tilde{B}_t = \lambda^{-1/2} B_{\lambda t}$ is also a Brownian motion.*

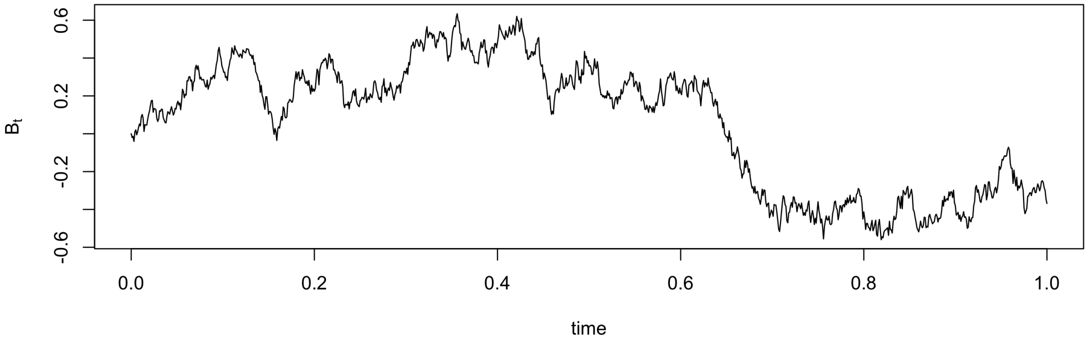

Figure 1: A simulated sample path of the Brownian motion.

Let ξ be a random variable independent of B_t. Consider the following equation:

$$\begin{aligned} X_0 &= \xi, \\ \mathrm{d}X_t &= b(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t, \end{aligned}$$

where $b\colon \mathbb{R} \times [0, \infty) \to \mathbb{R}$ and $\sigma\colon \mathbb{R} \times [0, \infty) \to \mathbb{R}$. This is called a stochastic differential equation (SDE), since the dynamics of X_t depends on the random process B_t. Under some conditions, there exists a unique solution to this SDE (the exact meaning of a "unique solution" here is beyond the scope of this course), and we call this solution, which is a stochastic process with a.s. continuous sample paths, a diffusion.

What will be more relevant to this course is how we numerically simulate a diffusion. This can be done in a way very similar to the finite difference method used for simulating deterministic differential equations.

Algorithm 10.1 (Euler-Maruyama Method)**.** Let h denote the size of each time increment, and N be the number of steps we will simulate. Define $t_n = nh$ for each $n = 0, 1, \dots, N$. Set $\hat{X}_0 = \xi$, where ξ is drawn from its distribution. For $n = 1, \dots, N$, set

$$\hat{X}_{t_n} = \hat{X}_{t_{n-1}} + b(\hat{X}_{t_{n-1}}, t_{n-1})h + \sigma(\hat{X}_{t_{n-1}}, t_{n-1})(B_{t_n} - B_{t_{n-1}}).$$

By the properties of Brownian motion, $(B_{t_n} - B_{t_{n-1}})_{n=1}^N$ are i.i.d. with distribution $N(0, h)$, and $B_{t_n} - B_{t_{n-1}}$ is independent of $\hat{X}_{t_{n-1}}$ for each n. That is, we can draw i.i.d. $Z_1, Z_2, \dots$ from $N(0, I)$ and set

$$\hat{X}_{t_n} = \hat{X}_{t_{n-1}} + b(\hat{X}_{t_{n-1}}, t_{n-1})h + h^{1/2}\sigma(\hat{X}_{t_{n-1}}, t_{n-1})Z_n.$$

We can also define Brownian motion and diffusions on $\mathbb{R}^d$ for $d > 1$. For simplicity, we still denote the d-dimensional Brownian motion by B_t, which has d independent coordinates, each being a one-dimensional Brownian motion. Note that for a d-dimensional diffusion process X_t satisfying $\mathrm{d}X_t = b(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t$, with B_t being a d'-dimensional Brownian motion, we have $b\colon \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^d$ and $\sigma\colon \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^{d \times d'}$.

10.2 Langevin Diffusion

Let π be a continuous probability distribution on $\mathbb{R}^d$; denote its density function by $\pi(x)$, which is assumed to be continuously differentiable. The diffusion X_t which evolves by

$$\mathrm{d}X_t = \nabla \log \pi(x)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

is called a Langevin diffusion. It is a reversible continuous-time process with stationary distribution π, provided that X_t does not explode; see Remarks 10.1 and 10.2 below. A sufficient condition that guarantees the non-explosion is [2]:

$$\langle \nabla \log \pi(x), x \rangle \leq a\|x\|_2^2 + b, \quad \forall x \text{ s.t. } \|x\|_2 \geq C,$$

where $a, b, C < \infty$ are fixed constants. Similar to the scaling property of Brownian motion, we can also re-scale the Langevin diffusion by considering

$$\mathrm{d}\tilde{X}_t = \lambda \nabla \log \pi(x)\mathrm{d}t + \sqrt{2\lambda}\,\mathrm{d}B_t,$$

for any $\lambda > 0$. Then, $\tilde{X}_t$ still has π has the stationary distribution.

Example 10.1. Let π be the multivariate normal distribution $N(0, \phi^{-1}I)$. Then, the resulting Langevin diffusion is given by

$$\mathrm{d}X_t = -\phi X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

This is known as the Ornstein-Uhlenbeck process.

Example 10.2. Let $d = 1$ and $\pi(x) \propto \exp(-\gamma |x|^\beta)$ for some $\gamma, \beta > 0$. Then,

$$\nabla \log \pi(x) = -\gamma\beta \,\mathrm{sgn}(x)\, |x|^{\beta - 1}.$$

It was shown in [2] that in this case, Langevin diffusion converges to π exponentially fast if and only if $\beta \geq 1$ (see [2] for the exact statement). The intuition behind this result is very important and applies to many MCMC algorithms: the convergence rate of a sampling algorithm largely depends on the tail decay rate of the target distribution.

Remark 10.1. Consider a d-dimensional diffusion X_t evolving by

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}B_t, \tag{1}$$

where B_t is d'-dimensional. When we say X_t has a stationary distribution π, it means that if $X_0 \sim \pi$, then $X_t \sim \pi$ for every $t > 0$. Not every diffusion process has a stationary distribution; for example, Brownian motion has no stationary distribution. Under certain conditions, the stationary distribution π exists and satisfies the forward Kolmogorov equation (also known as Fokker–Planck equation):

$$-\nabla \cdot (b(x)\pi(x)) + \frac{1}{2}\nabla \cdot \nabla \cdot (a(x)\pi(x)) = 0, \tag{2}$$

where $a(x) = \sigma(x)\sigma(x)^\top$ and $\nabla \cdot F$ denotes the divergence of the vector-valued function F. More explicitly, we denote functions b, a by $b(x) = (b_1(x), \dots, b_d(x))$ and $a(x) = (a_{ij}(x))_{1 \le i,j \le d}$, and then (2) can be written as

$$-\sum_{i=1}^{d} \frac{\partial [b_i(x)\pi(x)]}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2 [a_{ij}(x)\pi(x)]}{\partial x_i \partial x_j} = 0. \tag{3}$$

For the Langevin diffusion, we have $b(x) = \nabla \log \pi(x)$ and $a(x) = 2I$, and it is easy to verify that (3) is satisfied.

Remark 10.2. Here we give an informal justification for why Langevin diffusion has the desired stationary distribution. Consider (1) with $d = 1$. Fix some $h > 0$, and denote $X = X_0$ and

$$Y = X + h\, b(X) + \sigma(X)\sqrt{h} Z, \quad Z \sim N(0, 1).$$

So this is just one step of the Euler-Maruyama discretization. For a smooth function f,

$$f(y) = f(x) + (y - x)f'(x) + \frac{1}{2}(y - x)^2 f''(x) + o((y - x)^2)$$

by Taylor expansion. Now consider $\mathbb{E}[Y \mid X = x]$, which can be expressed by

$$\begin{aligned}\mathbb{E}[f(Y) \mid X = x] &\approx f(x) + f'(x)\mathbb{E}[(Y - x)] + \frac{1}{2}f''(x)\mathbb{E}[(Y - x)^2] \\ &= f(x) + h\, b(x)f'(x) + \frac{1}{2}\, h\, \sigma^2(x)f''(x) + o(h),\end{aligned}$$

where in the first equation we have omitted the remainder term. This shows that

$$\lim_{h \downarrow 0} \frac{\mathbb{E}[f(Y) \mid X = x] - f(x)}{h} = b(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x). \tag{4}$$

(Again this is an informal derivation; we need more regularity assumptions on b, σ, f so that this holds.) Now if π is the stationary distribution of X_t, and we let $X_0 \sim \pi$, then we expect that $\mathrm{d}\mathbb{E}[f(X_t)]/\mathrm{d}t = 0$ for any well-behaved function f. Using (4) and

$$\mathbb{E}[f(X_t)] - \mathbb{E}[f(X_0)] = \int_{\mathbb{R}} \{\mathbb{E}[f(X_t) \mid X_0 = x] - f(x)\}\, \pi(x)\mathrm{d}x,$$

assuming that we can interchange the order of limit and integral, we get

$$\int_{\mathbb{R}} \left\{ b(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x) \right\} \pi(x)\mathrm{d}x = 0.$$

If $b(x) = \nabla \log \pi(x) = \pi'(x)/\pi(x)$ and $\sigma^2(x) = 2$, the left-hand side is

$$\int_{\mathbb{R}} \{\pi'(x)f'(x) + \pi(x)f''(x)\}\, \mathrm{d}x = \pi(x)f'(x)\Big|_{-\infty}^{\infty},$$

which equals zero under certain regularity conditions (e.g., $\pi(x)$ vanishes at $\pm\infty$ and $f'(x)$ is bounded).

We can use Euler-Maruyama method to simulate the Langevin diffusion. This yields a discrete-time Markov chain $(\hat{X}_{t_n})_{n\geq 0}$, which is often called the unadjusted Langevin algorithm (ULA) or the Langevin Monte Carlo algorithm (LMC). Due to the time discretization, ULA does not have π as the stationary distribution. To correct for this bias, we can use the acceptance-rejection step in the Metropolis–Hastings algorithm, and the resulting algorithm is MALA (Metropolis-adjusted Langevin algorithm), which we have discussed in Unit 3. Note that to simulate the Langevin diffusion, we only need to be able to evaluate $\nabla \log \pi$, which does NOT require the knowledge of the normalizing constant.

10.3 Convergence of Langevin Monte Carlo Sampling

Now consider the Langevin Monte Carlo algorithm for sampling from the distribution π (i.e., the Euler-Maruyama discretization of Langevin diffusion without Metropolis–Hastings correction). To simplify the notation, we use $Y_0, Y_1, \dots$ to denote the samples generated from this algorithm, which satisfy $Y_0 \sim \nu_0$ and

$$Y_{n+1} = Y_n + h\nabla \log \pi(Y_n) + \sqrt{2h} Z_{n+1}, \tag{5}$$

for each $n \geq 0$, where $h > 0$ is the step size, and $Z_1, Z_2, \dots$ are independent standard normal random variables. Let ν_n denote the distribution of Y_n. Since $(Y_n)_{n\geq 0}$ is a discretization of the Langevin diffusion, if h is sufficiently small, we expect that the distribution ν_n will be sufficiently close to π as n increases. Such convergence analysis of the Langevin Monte Carlo algorithm (and its variants) has been an important research focus among the theoretical machine learning community. Here is a link to a free textbook on this topic.

In this lecture note, we give a brief review of the method used in [1] for analyzing the convergence rate of ν_n towards π. The main idea is coupling. Let $(X_t)_{t\geq 0}$ denote the Langevin diffusion with $X_0 \sim \pi$, which implies $X_t \sim \pi$ for every $t > 0$. Taking integral on both sides of the Langevin SDE, we get

$$X_t = X_0 + \int_0^t \nabla \log \pi(X_s) \mathrm{d}s + \sqrt{2} B_t. \tag{6}$$

Hence,

$$X_{(n+1)h} = X_{nh} + \int_{nh}^{(n+1)h} \nabla \log \pi(X_s) \mathrm{d}s + \sqrt{2}(B_{(n+1)h} - B_{nh}).$$

So far, the two chains $(Y_n)_{n\geq 0}$ and $(X_{nh})_{n\geq 0}$ are defined separately, and note that comparing the distance between ν_n and π is equivalent to comparing the distance between $\mathrm{Law}(Y_n)$ and $\mathrm{Law}(X_{nh})$. The "coupling" technique means that we construct $(Y_n)_{n\geq 0}$ and $(X_{nh})_{n\geq 0}$ jointly, in whatever way we want as long as the marginal dynamics of each chain is unchanged, so that Y_n and X_{nh} become as close to each other as possible. (For example, if we can somehow let Y_n and X_{nh} be always equal, then this implies that ν_n and π are the same.) In particular, by the definition of Wasserstein distance (see Remark 10.3), we have

$$W_2(\nu_n, \pi) \leq \left(\mathbb{E} \| X_{nh} - Y_n \|_2^2 \right)^{1/2},$$

where W_2 denotes the L^2-Wasserstein distance.

Here is the coupling strategy we use. Let the Brownian motion B_t in (6) be given, and then let $Z_1, Z_2, \dots$ in (5) be given by

$$\sqrt{h} Z_{n+1} = B_{(n+1)h} - B_{nh}. \tag{7}$$

We can do this because Brownian motion has independent normal increments; that is, this construction does not change the joint distribution of $(Z_1, Z_2, \dots)$. To simplify the notation, let us write $U(x) = \nabla \log \pi(x)$. Define $\Delta_n = X_{nh} - Y_n$, which satisfies

$$\begin{aligned}
\Delta_{n+1} &= X_{(n+1)h} - Y_{n+1} \\
&= X_{nh} + \int_{nh}^{(n+1)h} U(X_s) \mathrm{d}s - Y_n - h\, U(Y_n) \\
&= \Delta_n + \int_{nh}^{(n+1)h} \{U(X_s) - U(X_{nh})\} \,\mathrm{d}s + h\,\{U(X_{nh}) - U(Y_n)\} \\
&= \Delta_n + A_n + hW_n
\end{aligned}$$

where in the second step the normal increments have canceled out, and

$$A_n = \int_{nh}^{(n+1)h} \{U(X_s) - U(X_{nh})\} \,\mathrm{d}s, \qquad W_n = U(Y_n + \Delta_n) - U(Y_n).$$

If π is log-concave, then U is monotone decreasing (in every coordinate), and this implies that $\|\Delta_{n+1}\|_2 \leq \|\Delta_n + hW_n\|_2$. In [1], it is further assumed that U is strongly concave and has a Lipschitz continuous gradient (see the section "Proximal Sampling" in Unit 5 for definitions), and it is shown that $(\mathbb{E}\|\Delta_{n+1}\|_2^2)^{1/2} \leq \gamma (\mathbb{E}\|\Delta_n\|_2^2)^{1/2} + c$ for some $\gamma \in (0,1)$ and $c > 0$. A routine calculation then yields an upper bound on $\mathbb{E}\|\Delta_n\|_2^2$.

We conclude this unit with a numerical simulation illustrating this coupling idea. We let π be the standard univariate normal distribution, which has $U(x) = \nabla \log \pi(x) = -x$. Next we simulate the two processes X and Y over the time interval $[0, 2]$. For X_t, we generate X_0 from π and use Euler-Maruyama method with time step size $h_x = 10^{-4}$ (so that the discretization error is almost negligible). For Y_n, we fix $Y_0 = 0.2$ and use step size $h_y = 0.01$, with $Z_1, Z_2, \dots$ given by (7). Four simulated sample paths are shown in Figure 2, from which we can see a very clear tendency of $|\Delta_n|$ (the difference between the two lines) to decrease.

Remark 10.3. Let μ, ν be two probability distributions defined on the same space. The Wasserstein distance between μ, ν is defined by

$$W_{p,\rho}(\mu, \nu) = \inf \left\{ (\mathbb{E}[\rho(X,Y)^p])^{1/p} : \mathrm{Law}(X) = \mu,\ \mathrm{Law}(Y) = \nu \right\},$$

where ρ is a distance function. For example, if $X, Y \in \mathbb{R}^d$, we can let $\rho(x, y) = \|x - y\|_2$ be the Euclidean distance. Hence, any construction of (X, Y) with the given marginals yields an upper bound on the Wasserstein distance. So, for this coupling-based analysis, the choice of Wasserstein distance as the convergence metric is quite natural and simpifies the analysis.

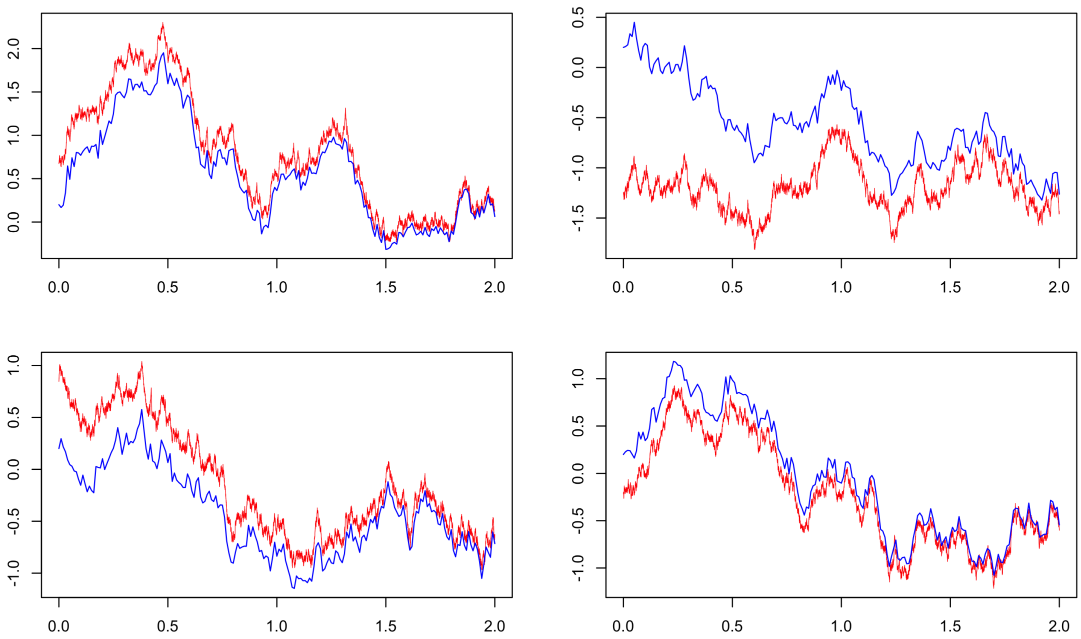

Figure 2: Coupling of ULA (blue) and Langevin diffusion (red).

References

[1] Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017.

[2] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.