Unit 11: Extensions of Langevin Monte Carlo Sampling

11.1 Scaled Langevin Diffusions

Let X_t be a d-dimensional diffusion driven by a d-dimensional Brownian motion, given by

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

Define $a(x) = \sigma(x)\sigma(x)^\top$, and assume its entries are sufficiently smooth. As in the last unit, we will use the notation $b(x) = [b_i(x)]_{i=1}^d$ and $a(x) = [a_{ij}(x)]_{1\le i,j\le d}$. Recall that the stationary distribution π, if it exists, should satisfy the forward equation

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

We now give more examples of diffusions that satisfy this equation, where the drift coefficient $b(x)$ still involves $\nabla \log \pi(x)$ but the diffusion coefficient $\sigma(x)$ is no longer constant.

Example 11.1. We first give a general solution. Let $a(x)$ be given, and we define

$$b_i(x) = \frac{1}{2}\left\{\sum_{j=1}^d a_{ij}(x)\frac{\partial \log \pi(x)}{\partial x_j} + \sum_{j=1}^d \frac{\partial a_{ij}(x)}{\partial x_j}\right\}. \tag{3}$$

Then, the resulting diffusion has π as the stationary distribution (assuming regularity conditions hold) and is reversible. This was used in [14] to design a class of MALA algorithms.

Exercise 11.1. Verify that $a(x), b(x)$ given in Example 11.1 solve (2).

Example 11.2. Let $d = 1$, in which case (3) simplifies to

$$b(x) = \frac{1}{2}\left\{a(x)\frac{\mathrm{d}\log\pi(x)}{\mathrm{d}x} + \frac{\mathrm{d}a(x)}{\mathrm{d}x}\right\}.$$

The property of the resulting diffusion process was analyzed in [12].

Example 11.3 (Langevin tempered diffusion)**.** Let $a(x) = \pi(x)^{-2\gamma} I$ with $\gamma \in [0, 1/2]$. Then $b(x)$ given in (3) can be expressed by

$$b(x) = \frac{1-2\gamma}{2} a(x)\nabla \log \pi(x).$$

Such diffusions are called Langevin tempered diffusions and analyzed in [11].

Example 11.4. A special case of the Langevin tempered diffusion was studied in [6]. Assume that $\pi(x) \propto g(x)^{-\beta}$ for some $\beta > 0$, and let $\gamma = 1/(2\beta)$ in Example 11.3. Then,

$$a(x) = g(x)I, \quad b(x) = -\frac{\beta - 1}{2}\nabla g(x).$$

11.2 Non-reversible Langevin Diffusion

Consider SDE (1) and fix

$$\sigma(X) = \sqrt{2}I,$$

where I is the identity matrix. In this case, X_t is reversible with respect to π if and only if $b(x) = \nabla \log \pi(x)$. For $d \geq 2$, we can construct a non-reversible Langevin diffusion with stationary distribution π by letting

$$b(x) = (I + S)\nabla \log \pi(x), \tag{4}$$

where $S \in \mathbb{R}^{d\times d}$ is a fixed skew-symmetric matrix; see the definition below. Note that the diagonal elements of S must be zeros.

Definition 11.1. We say S is a skew-symmetric matrix if $S = -S^\top$.

Exercise 11.2. Show that (2) is satisfied if $a(x) = 2$ and $b(x)$ is given by (4), with S being an arbitrary skew-symmetric matrix.

Example 11.5. Let $d = 2$ and π be the standard normal distribution $N(0, I)$. Then $\nabla \log \pi(x) = -x$, and

$$b(x) = \begin{bmatrix} 1 & s \\ -s & 1 \end{bmatrix} \begin{bmatrix} -x_1 \\ -x_2 \end{bmatrix} = \begin{bmatrix} -x_1 - sx_2 \\ sx_1 - x_2 \end{bmatrix},$$

where $s \in \mathbb{R}$ can be any constant. See

More generally, let π be the d-dimensional normal distribution $N(0, \Sigma)$ for some $d > 1$. Then, $\nabla \log \pi(x) = -\Sigma^{-1}x$, and $b(x) = -(I + S)\Sigma^{-1}x$. It was shown in [7] that as long as Σ does not have identical eigenvalues, using any skew-symmetric matrix S will improve the convergence of X_t.

For more general results about the diffusion X_t beyond the Gaussian case, see, e.g. [8, 10]. Discretizing the diffusion (e.g. by the Euler-Maruyama scheme) yields a practical algorithm for sampling from π, and the convergence rates of such algorithms have been studied in the more recent literature; see [5] among many others.

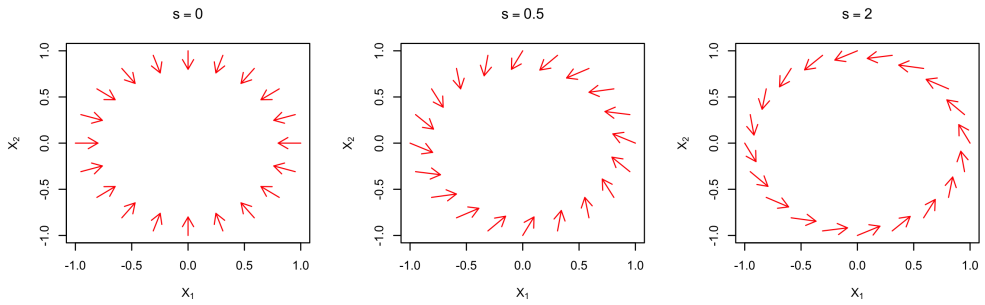

Figure 1: Directions of $b(x)$ in Example 11.5 with $d = 2$.

11.3 Underdamped Langevin Diffusion

Let $\pi(x) \propto e^{f(x)}$. A more complicated non-reversible generalization of the Langevin diffusion, known as underdamped Langevin diffusion, involves an auxiliary d-dimensional diffusion process V_t. The system (X_t, V_t) evolves by

$$\begin{aligned} \mathrm{d}X_t &= V_t \mathrm{d}t, \\ \mathrm{d}V_t &= \phi \nabla f(X_t)\mathrm{d}t - \lambda V_t \mathrm{d}t + \sqrt{2\lambda\phi}\,\mathrm{d}B_t, \end{aligned}$$

where $\lambda, \phi > 0$ are parameters (λ is often known as the friction coefficient). The joint stationary distribution is given by

$$\pi(x, v) \propto \exp\left\{ f(x) - \frac{1}{2\phi}\|v\|_2^2 \right\}. \tag{5}$$

Exercise 11.3. Show that the stationary distribution $\pi(x, v)$ given in (5) satisfies the forward equation (2) for the $2d$-dimensional process (X_t, V_t).

To simulate the underdamped Langevin diffusion, we can still use Euler-Maruyama discretization, which gives

$$\begin{aligned} \hat{V}_{(n+1)h} &= \hat{V}_{nh} + h\left(\phi\,\nabla f(\hat{X}_{nh}) - \lambda \hat{V}_{nh}\right) + \sqrt{2\lambda\phi}\,Z_{n+1}, \\ \hat{X}_{(n+1)h} &= \hat{X}_{nh} + h\hat{V}_{nh}. \end{aligned}$$

where h is the time increment size and $Z_1, Z_2, \ldots$ are i.i.d. standard normal random variables. A more accurate discretization can be obtained with minimum additional computational cost. We need to use the following lemma for linear SDE [9] (proof is omitted).

Lemma 11.1. *Let $Y_t \in \mathbb{R}^d$ be the solution to the linear SDE*

$$\mathrm{d}Y_t = (A\,Y_t + C)\mathrm{d}t + D\mathrm{d}B_t \tag{6}$$

where $A \in \mathbb{R}^{d\times d}, C \in \mathbb{R}^d, D \in \mathbb{R}^{d\times d}$ are fixed. Fix $Y_0 = y \in \mathbb{R}^d$. Then, Y_t is normally distributed with

$$\begin{aligned} \mathbb{E}[Y_t] &= e^{tA}y + \int_0^t e^{(t-s)A} C \mathrm{d}s, \\ \mathrm{Var}(Y_t) &= \int_0^t e^{(t-s)A} D D^\top e^{(t-s)A^\top} \mathrm{d}s. \end{aligned}$$

We now modify the naive discretization scheme by still usinng $\nabla f(X_t) \approx \nabla f(\hat{X}_{nh})$ over the time interval $[nh, (n+1)h)$ but integrating over the path of Brownian motion.

Lemma 11.2. *Fix $x, v \in \mathbb{R}^d$. Let $\tilde{X}_t, \tilde{V}_t$ be the solution to*

$$\begin{aligned} \mathrm{d}\tilde{X}_t &= \tilde{V}_t \mathrm{d}t, \\ \mathrm{d}\tilde{V}_t &= \phi \nabla f(x)\mathrm{d}t - \lambda \tilde{V}_t \mathrm{d}t + \sqrt{2\lambda\phi}\,\mathrm{d}B_t, \end{aligned}$$

with $\tilde{X}_0 = x, \tilde{V}_0 = v$. Then, $(\tilde{X}_t, \tilde{V}_t)$ is normally distributed with

$$\begin{aligned}
\mathbb{E}[\tilde{X}_t] &= x + \lambda^{-1}(1 - e^{-\lambda t})v - \phi\lambda^{-2}(1 - e^{-\lambda t} - \lambda t)\nabla f(x), \\
\mathbb{E}[\tilde{V}_t] &= e^{-\lambda t}v + \phi\lambda^{-1}(1 - e^{-\lambda t})\nabla f(x), \\
\mathrm{Var}(\tilde{X}_t) &= \phi\left\{\frac{2t}{\lambda} - \frac{4}{\lambda^2}(1 - e^{-\lambda t}) + \frac{1}{\lambda^2}(1 - e^{-2\lambda t})\right\} I, \\
\mathrm{Var}(\tilde{X}_t, \tilde{V}_t) &= \frac{\phi}{\lambda}\left(1 - 2e^{-\lambda t} + e^{-2\lambda t}\right) I, \\
\mathrm{Var}(\tilde{V}_t) &= \phi(1 - e^{-2\lambda t})I.
\end{aligned}$$

Denote this normal distribution by $F_t(x, v)$.

Proof. Write $\tilde{Y}_t = (\tilde{X}_t, \tilde{V}_t)$. Then, it satisfies the linear SDE (6) with

$$A = \begin{bmatrix} 0 & I \\ 0 & -\lambda I \end{bmatrix}, \quad C = \begin{bmatrix} 0 \\ \phi\nabla f(x) \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2\lambda\phi}I \end{bmatrix}.$$

Note that for $k \geq 1$,

$$A^k = \begin{bmatrix} 0 & (-\lambda)^{k-1}I \\ 0 & (-\lambda)^k I \end{bmatrix}.$$

Then a routine calculation using Lemma 11.1 yields the results. □

Hence, we can simulate the underdamped Langevin diffusion as follows. Let $h, \tilde{X}_0, \tilde{V}_0$ be given. For $n = 1, 2, \ldots$ we sample $(\tilde{X}_{nh}, \tilde{V}_{nh})$ from the distribution $F_h(\tilde{X}_{(n-1)h}, \tilde{V}_{(n-1)h})$, where $F_h(x, v)$ is defined in Lemma 11.2. See, e.g., [1] for the convergence analysis of this sampler. The dynamics of this sampler is very similar to that of Hamiltonian Monte Carlo.

11.4 Tempering of Langevin Diffusion

Let $\pi(x) \propto e^{f(x)}$ and $\pi_\beta(x) \propto e^{\beta f(x)}$. Observe that $\nabla \log \pi_\beta(x) = \beta \nabla \log f(x)$. Hence, the diffusion process X_t with dynamics

$$\mathrm{d}X_t = \nabla \log f(X_t) + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

has stationary distribution π_β. Its Euler-Maruyama discretization yields the sequence $(\hat{X}_{nh})_{n\geq 0}$ such that

$$\hat{X}_{nh} = \hat{X}_{(n-1)h} + \nabla \log f(\hat{X}_{(n-1)h}) + \sqrt{2h\beta^{-1}}Z_n,$$

where $Z_1, Z_2, \ldots$ are i.i.d. standard normal random variables. Compared with the Langevin Monte Carlo algorithm targeting π, the only difference is that the variance of the Gaussian noise has changed to $2h/\beta$. A smaller β (i.e., higher temperature) corresponds to larger variance of the noise. Observe that as $\beta \to \infty$, we actually recover the gradient descent algorithm, an optimization algorithm for finding $x^* = \arg\max_x \pi(x)$, which is not surprising since π_β converges towards the Dirac measure at x^*.

Simulated tempering and parallel tempering techniques can also be combined with Langevin diffusion. For example, pick two inverse temperatures $\beta_1, \beta_2 > 0$ and define two diffusions X^1, X^2 by

$$\mathrm{d}X_t^1 = \nabla \log f(X_t^1) + \sqrt{2\beta_1^{-1}}\mathrm{d}B_t^1,$$
$$\mathrm{d}X_t^2 = \nabla \log f(X_t^2) + \sqrt{2\beta_2^{-1}}\mathrm{d}B_t^2,$$

where B^1, B^2 denote two independent Brownian motions. The joint stationary distribution is given by

$$\pi(x_1, x_2) \propto \exp\left\{\beta_1 f(x_1) + \beta_2 f(x_2)\right\}.$$

Then, we simulate swapping times $\tau_1, \tau_2, \ldots$ by drawing $(\tau_k - \tau_{k-1})_{k\geq 1}$ independently from an exponential distribution with rate $\kappa > 0$. At each τ_k, we propose swapping the states of two diffusions. Explicitly, letting $\tau = \tau_k, x_1 = X_\tau^1, x_2 = X_\tau^2$, we propose setting $(X_{\tau+}^1, X_{\tau+}^2) = (x_2, x_1)$. According to the Metropolis–Hastings rule, we can accept this proposal with probability

$$\alpha(x_1, x_2) = \min\left\{1, \frac{\pi(x_2, x_1)}{\pi(x_1, x_2)}\right\},$$

which leaves the joint stationary distribution of the two diffusions unchanged. See, e.g., [4, 3] for the theoretical analysis of the process (X_t^1, X_t^2), which is often known as the replica-exchange Langevin diffusion.

11.5 Stochastic Gradient Langevin Dynamics

Suppose that $\pi(\theta)$ is a posterior distribution with $\theta \in \mathbb{R}^d$ being the parameter of some Bayesian statistical model. Let $\pi_0(\theta)$ denote the prior distribution and assume that we have n i.i.d. observations $y_1, \ldots, y_n$ with density $f(y_i \,|\, \theta)$. Then, we can express π by

$$\pi(\theta) \propto \pi_0(\theta) \prod_{i=1}^n f(y_i \,|\, \theta),$$

which yields

$$U(\theta) \coloneqq \nabla \log \pi(\theta) = \nabla \log \pi_0(\theta) + \sum_{i=1}^n \nabla \log f(y_i \,|\, \theta)$$

where ∇ always denotes the gradient with respect to θ. We can unbiasedly estimate $U(\theta)$ by drawing a simple random sample $\mathcal{S} \subset \{1, 2, \ldots, n\}$ of size m and computing

$$\hat{U}(\theta) = \nabla \log \pi_0(\theta) + \frac{n}{m} \sum_{i\in\mathcal{S}} \nabla \log f(y_i \,|\, \theta). \tag{7}$$

Note that $\mathcal{S}$ can be generated with or without replacement. This subsampling technique is very useful when n is huge (indeed, similar techniques are commonly used in training neural networks). Naturally, it leads to the following generalization of the Langevin Monte Carlo algorithm, known as stochastic gradient Langevin dynamics (SGLD).

Algorithm 11.1 (stochastic gradient Langevin dynamics)**.** Let $\theta_0 \in \mathbb{R}^d, h > 0$ be given. For $n = 1, 2, \ldots,$

(i) draw a subset $\mathcal{S} \subset \{1, 2, \ldots, n\}$ containing m samples and calculate $\hat{U}(\theta_{n-1})$ by (7);

(ii) draw $Z_n \sim N(0, 1)$ and set

$$\theta_n = \theta_{n-1} + h\hat{U}(\theta_{n-1}) + \sqrt{2h}Z_n.$$

SGLD was proposed by [13] and has become very popular in the machine learning community. In [13], a sequence of varying step sizes $(h_n)_{n\geq 1}$ was considered. Here, for simplicity, we fix the step size so that SGLD can be viewed as a noisy version of the Langevin Monte Carlo algorithm discussed in the last unit. Essentially the same coupling argument can be used to analyze the convergence of SGLD; see [2].

References

[1] Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on learning theory*, pages 300–323. PMLR, 2018.

[2] 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.

[3] Jing Dong and Xin T Tong. Spectral gap of replica exchange Langevin diffusion on mixture distributions. *Stochastic Processes and their Applications*, 151:451–489, 2022.

[4] Paul Dupuis, Yufei Liu, Nuria Plattner, and Jimmie D Doll. On the infinite swapping limit for parallel tempering. *Multiscale Modeling & Simulation*, 10(3):986–1022, 2012.

[5] Xuefeng Gao, Mert Gurbuzbalaban, and Lingjiong Zhu. Breaking reversibility accelerates Langevin dynamics for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:17850–17862, 2020.

[6] Ye He, Tyler Farghly, Krishnakumar Balasubramanian, and Murat A Erdogdu. Mean-square analysis of discretized Itô diffusions for heavy-tailed sampling. *Journal of Machine Learning Research*, 25(43):1–44, 2024.

[7] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, pages 897–913, 1993.

[8] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating diffusions. *The Annals of Applied Probability*, 15(2), 2005.

[9] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

[10] Tony Lelievre, Francis Nier, and Grigorios A Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

[11] Gareth O Roberts and Osnat Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357, 2002.

[12] O Stramer and RL Tweedie. Langevin-type models i: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, 1:283–306, 1999.

[13] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[14] Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.