Langevin Dynamics Sampling and Optimization

Kan Chen

December 12, 2019

Contents

1 Langevin Dynamics and Sampling

1.1 Overview of Langevin Dynamics

In physics, Langevin dynamics is an approach to the mathematical modeling of the dynamics of molecular systems. It was originally developed by French physicist **Paul Langevin** [6]. The approach is characterized by the use of simplified models while accounting for omitted degrees of freedom by the use of stochastic differential equations.

1.2 Overdamped Langevin Dynamics

We are adapting the form from [2]. We first consider the **Overdamped Langevin Dynamics**. Consider the following Stochastic Differential Equation:

$$d\mathbf{X}_t = -\nabla f(\mathbf{X}_t)dt + \sqrt{2}d\mathbf{B}_t \tag{1}$$

where $\{\mathbf{X}_t\}_t \in \mathbb{R}^p$ is a series of random variables, $\{\mathbf{B}_t\}_t$ is a series of standard Brownian process, and f is any twice continuously-differential function. Then the stationary distribution of X_t when $t \to \infty$ is

$$p(\mathbf{x}) \propto e^{-f(\mathbf{x})}$$

This can be proved by using the so called **Fokker-Planck Equation**, we will prove it in 1-D case in the following: consider the general stochastic differential equation in this form:

$$d\mathbf{X}_t = \mu(\mathbf{X}_t, t)dt + \sigma(\mathbf{X}_t, t)d\mathbf{W}_t \tag{2}$$

where the drift term is $\mu(\mathbf{X}_t, t)$, diffusion coefficient is $D(\mathbf{X}_t, t) = \sigma^2(\mathbf{X}_t, t)/2$. Then for 1-D case, the Fokker-Planck equation for probability density $p(\mathbf{x}, t)$ of random variable $\mathbf{X}_t$ is:

$$\frac{\partial}{\partial t}p(\mathbf{x}, t) = -\frac{\partial}{\partial x}[\mu(\mathbf{X}_t, t)p(\mathbf{x}, t)] + \frac{\partial^2}{\partial x^2}[D(\mathbf{X}_t, t)p(\mathbf{x}, t)] \tag{3}$$

Here, we abused our notation a bit. $\mathbf{X}_t = \mathbf{x}$. If we plug $\mu(\mathbf{X}_t, t) = -\nabla f(\mathbf{X}_t)$ and $\sigma(\mathbf{X}_t, t) = \sqrt{2}$ into (3), then we get

$$\frac{\partial}{\partial t}p(\mathbf{x}, t) = \frac{\partial}{\partial x}[f'(\mathbf{X}_t)p(\mathbf{x}, t)] + \frac{\partial^2}{\partial x^2}p(\mathbf{x}, t) \tag{4}$$

Since we are looking for stationary distribution, (i.e., the density would not change by time), (4) can be rewritten as:

$$0 = \frac{d}{dx}[f'(\mathbf{x})p(\mathbf{x})] + \frac{d^2}{dx^2}p(\mathbf{x}) \tag{5}$$

Notice that since we are looking for stationary distribution, we can get rid of time variable t. Integrate both side of (5) with respect to $\mathbf{x}$, we will get:

$$C = \frac{d}{dx}p(\mathbf{x}) + f'(\mathbf{x})p(\mathbf{x}) \tag{6}$$

where C is some constant. This is a first order linear ordinary differential equation which can be solved by **Integrating Factor** method. And the solution is

$$p(\mathbf{x}) \propto e^{-f(\mathbf{x})}$$

This result holds for higher dimensional case and it can be proved by higher dimensional Fokker-Planck equation.

1.3 Underdamped Langevin Dynamics

Next, we study the continuous time **Underdamped Langevin Dynamics** represented by the following stochastic differential equation [1]:

$$\begin{aligned} d\mathbf{v}_t &= -\gamma \mathbf{v}_t dt - u\nabla f(\mathbf{x}_t)dt + \sqrt{2\gamma u}d\mathbf{B}_t \\ d\mathbf{x}_t &= \mathbf{v}_t dt \end{aligned} \tag{7}$$

where $(\mathbf{x}_t, \mathbf{v}_t) \in \mathbb{R}^{2p}$, $\{\mathbf{B}_t\}_t$ is a series of standard Brownian process, and f is any twice continuously-differential function. Under fairly mild conditions, it can be shown that the stationary distribution of the continuous-time process (7) is proportional to $\exp(-(f(\mathbf{x}) + \|v\|_2^2/2u)))$, hence the marginal distribution of $\mathbf{x}$ is proportional to $\exp(-(f(\mathbf{x})))$.

Underdamped Langevin diffusion is particularly interesting because it contains a Hamiltonian component, and its discretization can be viewed as a form of Hamiltonian MCMC [3] which has been empirically observed to converge faster to the stationary distribution than overdamped Langevin diffusion.

1.4 Langevin Dynamics and Sampling

Then here comes our problem. Our goal is to generate a series of points $\{\mathbf{x}_i\}$ such that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i) = p(\mathbf{x})$$

given density $p(\mathbf{x})$. We solve this problem by solving *overdamped* Langevin diffusion and *underdamped* Langevin diffusion SDEs. For *overdamped* Langevin diffusion, we use the following Euler–Maruyama method to solve the SDE:

$$\mathbf{X}_{t+1} - \mathbf{X}_t = -\nabla f(\mathbf{X}_t)h + \sqrt{2h}\xi_t \tag{8}$$

where ξ_t is a series of Gaussian vector and h is the step-size (learning rate). For *underdamped* Langevin diffusion SDE, we first write (7) into vector form:

$$d\begin{pmatrix}\mathbf{v}_t \\ \mathbf{x}_t\end{pmatrix} = \begin{pmatrix}-u\nabla f(\mathbf{x}_t) - \gamma \mathbf{v}_t \\ \mathbf{x}_t\end{pmatrix} dt + \begin{pmatrix}\sqrt{2\gamma u} \\ 0\end{pmatrix} d\mathbf{B}_t \tag{9}$$

denote $\mathbf{z}_t = \begin{pmatrix}\mathbf{v}_t \\ \mathbf{x}_t\end{pmatrix}$, $\mu(\mathbf{z}_t) = \begin{pmatrix}-u\nabla f(\mathbf{x}_t) - \gamma \mathbf{v}_t \\ \mathbf{x}_t\end{pmatrix}$ and $\sigma(\mathbf{z}_t) = \begin{pmatrix}\sqrt{2\gamma u} \\ 0\end{pmatrix}$. Hence, the discretization of (9) would be:

$$\mathbf{z}_{t+1} - \mathbf{z}_t = \mu(\mathbf{z}_t)\Delta t + \sigma(\mathbf{z}_t)\Delta \mathbf{W}_t \tag{10}$$

where Δt step size, and $\Delta \mathbf{W}_t$ are independent identical random variables with mean 0 variance Δt. We can simultaneously solve N SDEs at the same time to get N points. Then when $t \to \infty$ and $N \to \infty$, we have:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i) = p(\mathbf{x})$$

2 Langevin Dynamics and Optimization

2.1 Sampling and Optimization

Minimizing non-convex and high-dimensional objective functions is challenging, especially when training modern deep neural networks. Currently, a novel approach is proposed which divides the training process into two consecutive phases to obtain better generalization performance: Bayesian sampling and stochastic optimization [5].

Gradient descent algorithm for non-convex optimization is very sensitive to the choice of initialization point. Often, the solution would be trapped by saddle points or local optimals. Hence, we separate the optimization process into two steps:

1. Use Bayesian sampling to get a good initialization.
2. Apply traditional gradient descent algorithm with the above initialization to do optimization.

The insight behind this is that the optimizers of $f(\mathbf{x})$ is the modes of $p(\mathbf{x}) \propto \exp(-f(\mathbf{x}))$. It is very similar to MAP (Maximum a posteriori) estimation. The further connection between sampling and optimization is discussed in [4].

2.2 Stochastic Gradient Langevin Dynamics

Stochastic gradient Langevin dynamics (SGLD), is an optimization technique composed of characteristics from Stochastic gradient descent, a Robbins-Monro optimization algorithm, and Langevin dynamics, a mathematical extension of molecular dynamics models. Like stochastic gradient descent, SGLD is an iterative optimization algorithm which introduces additional noise to the stochastic gradient estimator used in SGD to optimize a differentiable objective function. However, unlike traditional SGD, SGLD can be used for Bayesian learning, since the method produces samples from a posterior distribution of parameters based on available data.

Given some parameter vector θ, its prior distribution $p(\theta)$, and a set of data points $X = \{x_i\}_{i=1}^{N}$, Stochastic Gradient Langevin dynamics samples from the posterior distribution $p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$ by updating the chain:

$$\theta_{t+1} - \theta_t = \frac{\epsilon_t}{2} \left(\nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{N} \nabla \log p(x_{t_i}|\theta_t) \right) + \eta_t \tag{11}$$

where $n < N$ is a positive integer, $\eta_t \sim N(0, \epsilon_t)$ is Gaussian noise, $p(x|\theta)$ is the likelihood of the data given the parameter vector θ, and the step size ϵ_t

satistfies:

$$\sum_{t=1}^{\infty} \epsilon_t = \infty, \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \tag{12}$$

3 Future Direction

3.1 Sampling from Marginal Distribution

Efficiently sampling from marginal distribution is an open problem. To sample points from marginal distribution, people usually sample points from joint distribution then keep the desired components of points. This is computational inefficient. In [2], Dalalyan provides the theoretical guarantees for approximate sampling from smooth and log-concave densities using *Overdamped* Langevin Dynamics. He gives a non-asymptotic connection between the error bound and number of iterations. This error bound is the total variation distance between target distribution and approximate probability distribution:

$$\epsilon = \|p_t(\mathbf{x}) - p(\mathbf{x})\|_{\mathrm{TV}} \tag{13}$$

where $p_t(\mathbf{x})$ is the approximate target distribution at t-th iteration, $p(\mathbf{x})$ is the target distribution. We can see that ϵ is a trivial bound of the total variation distance between approximate target marginal distribution and target marginal distribution denoted by:

$$\epsilon' = \|p'_t(\mathbf{x}) - p'(\mathbf{x})\|_{\mathrm{TV}} \tag{14}$$

A potential research is to derive the non-asymptotic connection between ϵ' and number of iterations. However, this require the analysis from IPS (interactive particle system) and mean field approximation which is technically a very hard topic.

3.2 Simulated Annealing and Langevin Dynamics

Simulated Annealing(SA) is a probabilistic technique for approximating the global optimum of a given function. Specifically, it is a metaheuristic to approximate global optimization in a large search space for an optimization problem. It often being compared with **Hill Climbing** algorithm.

In Hill Climbing algorithm, once we have $f(x_{t+1}) < f(x_t)$, we stop searching and return x_t as our maximizer. However, in Simulated Annealing, if we have $f(x_{t+1}) < f(x_t)$, then we calculate a probability using $p(\Delta E) \propto \exp(\Delta E/(k_B T))$ where $\Delta E = f(x_{t+1}) - f(x_t)$, T is the current temperature, and k_B is some constant. We genearte a number u from uniform distribution $U(0,1)$. If $u > p$, then we accept the update, otherwise we reject. The idea is very similar to Metropolis-Hastings algorithm. However, Metropolis-Hastings algorithm is fixed temperture T while in Simulated Annealing, T decreases in each iteration by rate $r \in (0,1)$. Combining Langevin diffusion and Simulated Annealing is an interesting problem [5]:

$$d\mathbf{X}_t = -\nabla f(\mathbf{X}_t)dt + \sqrt{2\beta^{-1}}d\mathbf{B}_t \tag{15}$$

where $\beta^{-1} = k_B T(t)$ and $T(t) = c/\log(2+t)$. k_B, c are some constants. However, logarithmic annealing is extremely slow. How to design a different type of annealing process to speed up this process is also an interesting open problem.

References

[1] Xiang Cheng et al. "Underdamped Langevin MCMC: A non-asymptotic analysis". In: *arXiv preprint arXiv:1707.03663* (2017).

[2] Arnak S Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[3] Paul Adrien Maurice Dirac. "Generalized hamiltonian dynamics". In: *Canadian journal of mathematics* 2 (1950), pp. 129–148.

[4] Yi-An Ma et al. "Sampling can be faster than optimization". In: *arXiv preprint arXiv:1811.08413* (2018).

[5] Nanyang Ye, Zhanxing Zhu, and Rafal K Mantiuk. "Langevin dynamics with continuous tempering for training deep neural networks". In: *arXiv preprint arXiv:1703.04379* (2017).

[6] Robert Zwanzig. "Nonlinear generalized Langevin equations". In: *Journal of Statistical Physics* 9.3 (1973), pp. 215–220.