Notes on Langevin Dynamics

Zhenya Liu

In energy based model, we learn the energy function to train on the Contrastive Divergence, where p_θ is modeled by restricted Boltzmann machine (RBM),

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$$

After training we get E_θ, tradition MCMC sampling methods such as Gibb's sampling always have low efficiency because of long mixing time. To make the sampling more efficient, we can choose Langevin dynamics which utilize the energy function:

$$\begin{aligned} &Given\ \varepsilon_t \sim \mathcal{N}(0, I): \\ &\qquad x_{t+1} = x_t - \eta \nabla_x \log p_\theta(x_t) + \sqrt{2\eta}\varepsilon_t \end{aligned}$$

Consider the distribution of random variable X_t of each state t, we form one discrete Markov chain $\{X_{t \geq 0}\}$, and it has unique stationary distribution, $i.e\ \ p(X_t = x) \xrightarrow[t]{} p_\theta(x)$.

We can verify this result by proving the integral equality given a specific stochastic differential equation (SDE) and one desired stationary distribution. However, this approach cannot explain how to actually find that SDE given arbitrary $p(x)$.

Now we will try to show this result from a more general point of view.

Idea behind using Langevin Dynamics:

While directly sampling from p_θ is hard, but sampling from a Markov chain is easy. We can approximate samples using a Markov chain with stationary distribution p_θ. A convenient construction on $\mathcal{X} = \mathbb{R}^d$ is Langevin Dynamics, which is a continuous Markov process with dynamics given by the SDE:

$$dx_t = \frac{1}{2} \nabla_x \log p_\theta(x_t) dt + dW_t \tag{1}$$

By discretization,

$$x_{t+1} = x_t - \frac{\eta}{2} \nabla_x \log p_\theta(x_t) + \sqrt{\eta}\varepsilon_t$$

where η is the coefficient controlling the trade-off between mixing speed and approximation precision.

Recall $p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}$, (1) becomes:

$$dx_t = -\frac{1}{2}\nabla_x E_\theta(x_t)dt + dW_t \tag{2}$$

Now assume that (2) has one unique stationary distribution (under mild conditions), we **claim** that it is p_θ.

Proof:
Consider a general SDE with Weiner Process W_t,

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t \tag{3}$$

with drift $\mu(X_t, t)$ and diffusion coefficient $D(X_t, t) = \sigma^2(X_t, t) \,/\, 2$.

The **Fokker–Planck equation** for the probability density $p(x, t)$ is the generally unsolvable PDE

$$\frac{\partial}{\partial t} p(x,t) = -\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)] \tag{4}$$

When the SDE (3) satisfies mild conditions on the drift and diffusion terms. then it has one unique stationary distribution $p(x)$, such that $p(x,t) \xrightarrow[t]{} p(x)$. Since $p(x)$ is stationary, $\frac{\partial}{\partial t} p(x) = 0$. Take the limit of t in both sides of (4), then we get one ODE and solve it:

$$\begin{aligned}
0 &= -\frac{\partial}{\partial x}[\mu(x,t)p(x)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x)] \\
\frac{\partial}{\partial x}[\mu(x,t)p(x)] &= \frac{\partial^2}{\partial x^2}[D(x,t)p(x)] \\
\frac{\partial}{\partial x}[D(x,t)p(x)] \; &= \mu(x,t)p(x) \; + \; C \\
p(x)\frac{\partial}{\partial x}[D(x,t)] + D(x,t)\frac{\partial}{\partial x}p(x) &= \mu(x,t)p(x) \; + \; C \\
\frac{\partial}{\partial x}[D(x,t)] + D(x,t)\left[\frac{1}{p(x)}\frac{\partial}{\partial x}p(x)\right] &= \mu(x,t) \; + \; C
\end{aligned}$$

$$\mu(x,t) = \frac{\partial}{\partial x}[D(x,t)] + D(x,t)\frac{\partial}{\partial x}\log p(x) + C \tag{5}$$

Now given $p(x) = p_\theta(x) = \frac{1}{Z_\theta}e^{-E_\theta(x)}$,

as long as we find suitable $D(x,t), \mu(x,t), C$ which satisfies (5), the SDE (3) with defined $\sigma(x,t)$ and $\mu(x,t)$ must have the stationary distribution $p_\theta(x)$ by the **Fokker–Planck equation**.

For simplicity, we set $C = 0$, $D(x,t) = \frac{1}{2}$, so that $\sigma(x,t) \;=\; \sqrt{2D(x,t)} = 1$, and (5) becomes

$$\begin{aligned}\mu(x,t) &= \frac{\partial}{\partial x}[D(x,t)] + D(x,t)\frac{\partial}{\partial x}\log p(x) + C \\ &= 0 + \frac{1}{2}\cdot\left(\frac{\partial}{\partial x} - E_\theta(x)\right) + 0 \\ &= -\frac{1}{2}\nabla_x E_\theta(x)\end{aligned}$$

With calculated $\mu(x,t)$ and $\sigma(x,t)$, we can form our desired SDE

$$\begin{aligned}dX_t &= \mu(X_t,t)dt + \sigma(X_t,t)dW_t \\ dx_t &= -\frac{1}{2}\nabla_x E_\theta(x_t)dt + dW_t\end{aligned}$$

□

Remark:

Langevin Dynamics is also applied in score based models like noise conditional score networks (NCSN). Given the RBM $p_\theta(x) = \frac{1}{Z_\theta}e^{-E_\theta(x)}$, the score function is the gradient $\nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x)$.
In score based models, the score function is trained by one technique called score matching. With trained score function, we can directly apply it on Langevin dynamics for sampling.

We can find that the energy based model indirectly calculate the score function. The training on CD aims to maximize MLE. However, score based models are directly trained on score functions.