SDE笔记

WangFangyikang

2023 年 10 月 12 日

目录

1 Motivation

In this section, we first illustrate the motivation of studying SDE from a machine learning perspective.

1.1 Approximating SGD

First, we look at the SGD process:

$$x^{n+1} = x^n - \eta_k \nabla \mathcal{L}(x^n; \xi^n), \tag{SGD}$$

where the white noise ξ^n characterize the randomness of the surrogate gradient in SGD method. Denote $\Sigma(x) := \mathbb{E}_\xi \left[(\nabla \mathcal{L}(x;\xi) - \nabla \mathcal{L}(x)) (\nabla \mathcal{L}(x;\xi) - \nabla \mathcal{L}(x))^T \right]$ and SGD writes:

$$x^{n+1} = x^n - \eta_k \nabla \mathcal{L}(x^n) + \sqrt{\eta \Sigma}\sqrt{\eta}\mathcal{Z}^n, \mathcal{Z}^n \sim N(0, I_d).$$

If we take the limit $\eta \to 0$ and regard $\sqrt{\eta}\mathcal{Z}^n = \mathrm{d}W_t$, the SDE form of SGD is:

$$\mathrm{d}X(t) = -\nabla \mathcal{L}(X(t))\mathrm{d}t + \sqrt{\eta\Sigma}\mathrm{d}W_t. \tag{SDE-1}$$

Q:

- Is SDE-1 a good approximation of SGD?
- Good in what sense?
- Is there a better one?

A:

- SDE-1 is a first-order weak approximation of SGD.
- Good in sense of testing: $\forall\, |g(x)| < K(1+|x|)^K, |\mathbb{E}g(X(n\eta)) - g(X^n)| < C\eta^\alpha$
- There are higher order approximations!

For example, the second-order approximation of SGD writes:

$$\mathrm{d}X(t) = -\nabla \left(\mathcal{L}(X(t)) + \frac{\eta}{4} \|\nabla \mathcal{L}(X(t))\|^2 \right) \mathrm{d}t + \sqrt{\eta\Sigma}\mathrm{d}W_t. \tag{SDE-2}$$

And another formulation (1-d Xiang) writes:

$$\mathrm{d}X(t) = \frac{\log(1-\eta\mathcal{L}''(x))}{\eta\mathcal{L}''(x)}\mathcal{L}'(x)\mathrm{d}t + \sqrt{\frac{2\Sigma \cdot \log(1-\mathcal{L}''(x)\eta)}{\mathcal{L}''(x)(\mathcal{L}''(x)\eta - 2)}}\mathrm{d}W_t. \tag{SDE-Xiang-1-dim}$$

The d-dimensional Xiang-Formulation is still under developing. Another class of questions is follows:

Q:

- How are these more advanced flows derived?
- Why would the SDE approximation be useful?

, which will be answered in the following.

1.2 Langevin Dynamics

Our goal of Langevin Dynamics is sampling from a Gibbs measure $\frac{e^{-\frac{\mathcal{L}(x)}{\sigma}}}{\mathcal{Z}_{\mathcal{L},\sigma}}$, where $\mathcal{Z}_{\mathcal{L},\sigma}$ is the normalizing constant. The Langevin dynamics writes:

$$\mathrm{d}X(t) = -\nabla\mathcal{L}(X(t))\mathrm{d}t + \sqrt{2\sigma}\mathrm{d}W_t. \tag{LD}$$

Q:

- Why is this approach correct? I.e. why does LD have the correct equilibrium?
- How fast is the convergence?

The discrete-time version of LD writes:

$$X^{k+1} = X^k - \eta\nabla\mathcal{L}(X^k) + \sqrt{2\sigma}\sqrt{\eta}\mathcal{Z}^k, \mathcal{Z}^n \sim N(0, I_d) \tag{1}$$