SDE笔记

WangFangyikang

2023 年 10 月 24 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:

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

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:

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

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:

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

And another formulation (1-d Xiang) writes:

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

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:

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

  Q:

- What is the convergence property?

- Can we accelerate the convergence?

2 Ordinary Differential Equations

To better understand the behavior of SDE, we can first take a look at its non-random counterpart, i.e., ODEs (Ordinary Differential Equations).

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t, X(0) = X_0 \qquad \text{(ODE)}$$

Example: Linear ODE, i.e., $f(t, X(t)) = LX(t)$

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} e^{-tL} X(t) &= -e^{-tL} L X(t) + e^{-tL} \frac{\mathrm{d}}{\mathrm{d}t} X(t) \\
&= e^{-tL} \left(\frac{\mathrm{d}}{\mathrm{d}t} X(t) - L X(t) \right) = 0 \\
&\Rightarrow e^{-tL} X(t) = e^{-tL} X(t)|_{t=0} = X_0 \\
&\Rightarrow X(t) = e^{tL} X_0
\end{aligned}
\qquad (1)
$$

Next we look at the conception of Principle Flow proposed in Rosca et al. (2023), consider minimizing the quadratic objective $f(x) = \frac{1}{2} x^T H x$, assuming that H is positive definite.

Consider the Gradient Descent Dynamic here:

$$X^{n+1} = X^n - \eta \nabla \mathcal{L}(x) = X^n - \eta H X^n = (1 - \eta H) X^n$$

Therefore $X^n = (1 - \eta H)^n X_0$, if we want to have $X^n = X(n\eta)$ for all n, we should have:

$$(1 - \eta H)^n = e^{n\eta L} \Rightarrow L = \frac{\log(1 - \eta H)}{\eta}$$

We obtain $\mathrm{d}X(t) = \frac{\log(1 - \eta H)}{\eta} X(t) \mathrm{d}t$, which is the Principle Flow in the quadratic case. This can be generalized to the non-linear case:

$$\mathrm{d}X(t) = \sum_{i=1}^{d} \frac{\log(1 - \eta \lambda_i)}{\eta \lambda_i} \nabla \mathcal{L}(X(t))^T u_i \cdot u_i,$$

where $\nabla^2 \mathcal{L}(X(t)) = \sum_{i=1}^{d} \lambda_i u_i u_i^T$ is the SVD of $\nabla^2 \mathcal{L}(X(t))$. This generalization is derived in the sense of "backwards error analysis":

$$
\begin{cases}
\dot{\theta} = -\nabla \mathcal{L}(\theta) + \eta f_1(\theta) + \cdots + \eta^n f_n(\theta), \\
\theta^{n+1} = \theta^n - \eta \nabla \mathcal{L}(\theta^n),
\end{cases}
\qquad (2)
$$

We want to have $\theta^{n+1} = \theta(n\eta + \eta)$ and $\theta^n = \theta(n\eta)$

2.1 Numerical Solvers for ODE

Write the ODE in the integral form:

$$X(t + \Delta t) = X(t) + \int_t^{t+\Delta t} f(X(\tau), \tau) \, \mathrm{d}\tau$$

Explicit Euler method:

$$\overline{X}(t + \Delta t) = X(t) + \int_t^{t+\Delta t} f(X(t), t) \, \mathrm{d}\tau$$

Implicit Euler method:

$$X(t + \Delta t) = X(t) + \int_t^{t+\Delta t} f(X(t + \Delta t), t + \Delta t) \, \mathrm{d}\tau$$

Heun method:

$$X(t + \Delta t) = X(t) + \frac{1}{2} \int_t^{t+\Delta t} f(X(t), t) + f(\overline{X}(t + \Delta t), t + \Delta t) \, \mathrm{d}\tau$$

Fourth order Runge-Kutta method:

$$\Delta X_k^1 = f(\widehat{X}(t_k), t_k) \Delta t$$
$$\Delta X_k^2 = f(\widehat{X}(t_k) + \Delta X_k^1/2, t_k + \Delta t/2) \Delta t$$
$$\Delta X_k^3 = f(\widehat{X}(t_k) + \Delta X_k^2/2, t_k + \Delta t/2) \Delta t$$
$$\Delta X_k^4 = f(\widehat{X}(t_k) + \Delta X_k^3, t_k + \Delta t) \Delta t$$
$$\widehat{X}(t_{k+1}) = \widehat{X}(t_k) + \frac{1}{6}(\Delta X_k^1 + 2\Delta X_k^2 + 2\Delta X_k^3 + \Delta X_k^4)$$

Order of approximation: $\left| \widehat{X}(t_M) - X(t_M) \le k\Delta t^P \right|, M = \frac{1}{\Delta t}$

2.2 Existence and Uniqueness of the solution to the ODE

Picard iteration: start from the initial guess $\varphi_0(t) = X_0$, recursively compute

$$\varphi_{n+1}(t) = X_0 + \int_{t_0}^t f(\varphi_n(\tau), \tau) \, \mathrm{d}\tau.$$

If f is continuous in both x and t and Lipschitz continuous in x, then:

$$\lim_{n \to \infty} \varphi_n(t) = X(t)$$

3 Heuristic Derivation of SDE

3.1 Linear SDE

For SDE, we assume $\mathrm{d}W_t \sim N(0, \mathrm{d}t)$ and SDE writes:

$$\mathrm{d}X_t = FX_t\mathrm{d}t + \sqrt{\widehat{\Sigma}}\mathrm{d}W_t$$

Then,

$$\begin{aligned}
&\mathrm{d}\exp(-Ft)X_t \\
&= -F \cdot \exp(-Ft)X_t\mathrm{d}t + \exp(-Ft)\mathrm{d}X_t \\
&= \exp(-Ft)\sqrt{\widehat{\Sigma}}\mathrm{d}W_t \\
&\Rightarrow \exp(-Ft)X_t = X_0 + \int_0^t \exp(-F\tau)\sqrt{\widehat{\Sigma}}\,\mathrm{d}W_\tau \\
&\Rightarrow X_t = \exp(Ft)X_0 + \exp(F(t-\tau))\sqrt{\widehat{\Sigma}}\,\mathrm{d}W_\tau
\end{aligned}$$

So we know that X_t remains Gaussian given $X_0 \sim N(m_0, P_0)$.

$$\begin{aligned}
m_t &= \mathbb{E}X_t = \exp(F_t)m_0 \\
P_t &= \mathbb{E}\left[(X_t - m_t)(X_t - m_t)^T\right] \\
&= \exp(Ft)P_0\exp(Ft)^T + \int_0^t \exp(F(t-\tau))\widehat{\Sigma}\exp(F(t-\tau))^T\,d\tau
\end{aligned}$$

, which reveals the property of OU process.

3.2 Informal derivation of Xiang's approach

Consider the stochastic dynamics $(d = 1)$:

$$X^{n+1} = X^n - \eta\left(HX^n + \sqrt{\Sigma}Z^n\right), Z^n \sim N(0, I)$$

so, X^n remains Gaussian.

$$\begin{aligned}
&X^{n+1} = (1 - \eta H)X^n - \eta\sqrt{\Sigma}Z^n \\
&\Rightarrow \frac{X^{n+1}}{(1-\eta H)^{n+1}} = \frac{X^n}{(1-\eta H)^n} - \frac{\eta\sqrt{\Sigma}Z^n}{(1-\eta H)^{n+1}} \\
&\Rightarrow \frac{X^n}{(1-\eta H)^n} = X^0 - \eta\sqrt{\Sigma}\sum_{i=1}^n \frac{Z^i}{(1-\eta H)^i} \\
&\Rightarrow X^n = (1-\eta H)^n X^0 - \eta\sqrt{\Sigma}\sum_{i=1}^n (1-\eta H)^{n-i} Z^i
\end{aligned}$$

We can further calculate its mean and variance:

$$m^n = \mathbb{E}X^n = (1 - \eta H)^n \mathbb{E}X^0$$

$$
\begin{aligned}
\mathbb{E}\left[(x^n - m^n)(x^n - m^n)^T\right] &= \eta^2 \Sigma \sum_{i=1}^{n} (1 - \eta H)^{2(n-i)} \\
&= \eta^2 \Sigma \sum_{i=0}^{n-1} (1 - \eta H)^{2i} \\
&= \eta^2 \Sigma \frac{1 - (1 - \eta H)^{2n}}{1 - (1 - \eta H)^2}
\end{aligned}
$$

Following the idea of principle flow, first we let the mean of two meet:

$$(1 - \eta H)^n = \exp(Fn\eta)$$
$$\Rightarrow F = \log(1 - \eta H)/\eta$$

then look at the variance:

$$\int_0^t \exp(F(t - \tau))\widehat{\Sigma}\exp(F(t - \tau))^T \, d\tau = \eta^2 \Sigma \frac{1 - (1 - \eta H)^{2n}}{2\eta H - (\eta H)^2}$$

For the left side, we have:

$$
\begin{aligned}
\int_0^t \exp(F(t - \tau))\widehat{\Sigma}\exp(F(t - \tau))^T \, d\tau &= \frac{1}{2F} \exp(2F\tau) \mid_0^t \widehat{\Sigma} \\
&= \frac{1}{2F} \widehat{\Sigma} \cdot (\exp(2Fn\eta) - 1) \\
&= \frac{(1 - \eta H)^{2n} - 1}{2F} \cdot \widehat{\Sigma}
\end{aligned}
$$

So,

$$\frac{(1 - \eta H)^{2n} - 1}{2F} \cdot \widehat{\Sigma} = \eta^2 \Sigma \frac{1 - (1 - \eta H)^{2n}}{2\eta H - (\eta H)^2}$$
$$\Rightarrow \widehat{\Sigma} = -\frac{2F \cdot \eta^2 \cdot \Sigma}{2\eta H - \eta^2 H^2} = \frac{2\Sigma \log(1 - \eta H)}{H(H\eta - 2)}$$

参考文献

Rosca, M.; Wu, Y.; Qin, C.; and Dherin, B. 2023. On a continuous time model of gradient descent dynamics and instability in deep learning. *Transactions on Machine Learning Research.*