

# SDE笔记

WangFangyikang

2023 年 10 月 24 日

## 目录

<b>1</b>	<b>Motivation</b>	<b>2</b>
1.1	Approximating SGD . . . . .	2
1.2	Langevin Dynamics . . . . .	3
<b>2</b>	<b>Ordinary Differential Equations</b>	<b>4</b>
2.1	Numerical Solvers for ODE . . . . .	5
2.2	Existence and Uniqueness of the solution to the ODE . . . . .	5
<b>3</b>	<b>Heuristic Derivation of SDE</b>	<b>6</b>
3.1	Linear SDE . . . . .	6
3.2	Informal derivation of Xiang's approach . . . . .	6

# 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:

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

where the white noise  $\xi^n$  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  $\eta \rightarrow 0$  and regard  $\sqrt{\eta} \mathcal{Z}^n = dW_t$ , the SDE form of SGD is:

$$dX(t) = -\nabla \mathcal{L}(X(t))dt + \sqrt{\eta \Sigma} dW_t. \quad (\text{SDE-1})$$

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:

$$dX(t) = -\nabla \left( \mathcal{L}(X(t)) + \frac{\eta}{4} \|\nabla \mathcal{L}(X(t))\|^2 \right) dt + \sqrt{\eta \Sigma} dW_t. \quad (\text{SDE-2})$$

And another formulation (1-d Xiang) writes:

$$dX(t) = \frac{\log(1 - \eta \mathcal{L}''(x))}{\eta \mathcal{L}''(x)} \mathcal{L}'(x) dt + \sqrt{\frac{2\Sigma \cdot \log(1 - \mathcal{L}''(x)\eta)}{\mathcal{L}''(x)(\mathcal{L}''(x)\eta - 2)}} dW_t. \quad (\text{SDE-Xiang-1-dim})$$

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:

$$dX(t) = -\nabla\mathcal{L}(X(t))dt + \sqrt{2\sigma}dW_t. \quad (\text{LD})$$

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:

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

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).

$$dX(t) = f(t, X(t))dt, X(0) = X_0 \quad (\text{ODE})$$

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

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

Next we look at the conception of Principle Flow proposed in [Rosca et al. \(2023\)](#), consider minimizing the quadratic objective  $f(x) = \frac{1}{2}x^T Hx$ , 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  $dX(t) = \frac{\log(1 - \eta H)}{\eta} X(t)dt$ , which is the Principle Flow in the quadratic case. This can be generalized to the non-linear case:

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

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":

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

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:

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

Explicit Euler method:

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

Implicit Euler method:

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

Heun method:

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

Fourth order Runge-Kutta method:

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

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

## 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

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

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

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

### 3 Heuristic Derivation of SDE

#### 3.1 Linear SDE

For SDE, we assume  $dW_t \sim N(0, dt)$  and SDE writes:

$$dX_t = F X_t dt + \sqrt{\widehat{\Sigma}} dW_t$$

Then,

$$\begin{aligned} d \exp(-Ft) X_t &= -F \cdot \exp(-Ft) X_t dt + \exp(-Ft) dX_t \\ &= \exp(-Ft) \sqrt{\widehat{\Sigma}} dW_t \\ \Rightarrow \exp(-Ft) X_t &= X_0 + \int_0^t \exp(-F\tau) \sqrt{\widehat{\Sigma}} dW_\tau \\ \Rightarrow X_t &= \exp(Ft) X_0 + \exp(F(t-\tau)) \sqrt{\widehat{\Sigma}} dW_\tau \end{aligned}$$

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

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

, which reveals the property of OU process.

#### 3.2 Informal derivation of Xiang's approach

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

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

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:

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

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

## 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*.