ML HW8

Date: 10/22/2025
Week: 8
Author: Alvin B. Lin
Student ID: 112652040

Problem 1: Form Transformation of SSM

Show that the sliced score matching (SSM) loss can also be written as

$$L_{\text{SSM}}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left[\left\|\mathbf{v}^\top S(\mathbf{x};\theta)\right\|^2 + 2\mathbf{v}^\top \nabla_{\mathbf{x}}(\mathbf{v}^\top S(\mathbf{x};\theta))\right].$$

Solution:

Given the sliced score matching loss is defined as (from the note[1])

$$L_{\text{SSM}}(\theta) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left\|S(\mathbf{x};\theta)\right\|^2 + \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left[2\mathbf{v}^\top \nabla_{\mathbf{x}}(\mathbf{v}^\top S(\mathbf{x};\theta))\right].$$

Meaning that we only need to show

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left\|\mathbf{v}^\top S(\mathbf{x};\theta)\right\|^2 = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left\|S(\mathbf{x};\theta)\right\|^2$$

Proof. First since $\mathbf{x}$ is independent of $\mathbf{v}$, thus the iterated expectation can be written as

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left\|\mathbf{v}^\top S(\mathbf{x};\theta)\right\|^2 = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left(\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left\|\mathbf{v}^\top S(\mathbf{x};\theta)\right\|^2\right) \tag{1}$$

Lemma 1

Given $\mathbf{v} \sim \mathcal{N}(\mathbf{0},\ I)$, we have

$$\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left(\mathbf{v}\mathbf{v}^\top\right) = I$$

Proof. We know that if $\mathbf{v} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, then

$$\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left((\mathbf{v}-\mu)(\mathbf{v}-\mu)^\top\right) = \mathbf{\Sigma}$$

Take $\mu = \mathbf{0}$ and $\mathbf{\Sigma} = I$, we get

$$\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left(\mathbf{v}\mathbf{v}^\top\right) = I$$

□

Equation (1) can also be written as

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left(S(\mathbf{x};\theta)^\top\left(\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\mathbf{v}\mathbf{v}^\top\right)S(\mathbf{x};\theta)\right) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left(S(\mathbf{x};\theta)^\top S(\mathbf{x};\theta)\right) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left\|S(\mathbf{x};\theta)\right\|^2,$$

which completes the proof. □

Still a non-trivial conversion from ISM occurs at:

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left(2\nabla_{\mathbf{x}}\cdot S(\mathbf{x};\theta)\right)=\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\left(2\mathbf{v}^{\top}\nabla_{\mathbf{x}}(\mathbf{v}^{\top}S(\mathbf{x};\theta))\right)$$

Equivalently, we will show:

$$\nabla_{\mathbf{x}}\cdot S(\mathbf{x};\theta)=\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\mathbf{v}^{\top}\nabla_{\mathbf{x}}(\mathbf{v}^{\top}S(\mathbf{x};\theta))$$

Proof. Recall we've defined

$$S(\mathbf{x};\theta)=\nabla_{\mathbf{x}}\log p(\mathbf{x})$$

Then

$$\nabla_{\mathbf{x}}\cdot S(\mathbf{x};\theta)=\Delta S(\mathbf{x};\theta)=\sum_{i=1}^{n}\frac{\partial^2 S}{\partial x_i^2}=\mathrm{Tr}(\mathbf{Hess}(S)),$$

where n is the dimension and $\mathbf{Hess}(\cdot)$ is the Hessian matrix. Here we need a theorem:

Theorem 1: Hutchinson's Trace Estimator

Given matrix $\mathbf{A}\in\mathbb{M}_{n\times n}(\mathbb{R})$, we have:

$$\mathrm{Tr}(\mathbf{A})=\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}(\mathbf{v}^{\top}\mathbf{A}\mathbf{v}),$$

for distribution of $\mathbf{v}$ such that $\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\mathbf{v}\mathbf{v}^{\top}=I$

Proof.

$$\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}(\mathbf{v}^{\top}\mathbf{A}\mathbf{v})=\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\mathrm{Tr}(\mathbf{v}^{\top}\mathbf{A}\mathbf{v})=\mathrm{Tr}\left(\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}(\mathbf{v}\mathbf{v}^{\top})\mathbf{A}\right)=\mathrm{Tr}(\mathbf{A})$$

□

We take $\mathbf{A}=\mathbf{Hess}(S)$, and for arbitrary $\mathbf{v}$ independent from $\mathbf{x}$,

$$(\nabla_{\mathbf{x}}\cdot S(\mathbf{x};\theta))\mathbf{v}=\left(\frac{\partial S}{\partial x_1}+\cdots+\frac{\partial S}{\partial x_n}\right)\mathbf{v}=\left(\frac{\partial S}{\partial x_1}\mathbf{v}+\cdots+\frac{\partial S}{\partial x_n}\mathbf{v}\right)=\nabla_{\mathbf{x}}\left(\mathbf{v}^{\top}S(\mathbf{x};\theta)\right)$$

Therefore we have:

$$\nabla_{\mathbf{x}}\cdot S(\mathbf{x};\theta)=\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})}\mathbf{v}^{\top}\nabla_{\mathbf{x}}(\mathbf{v}^{\top}S(\mathbf{x};\theta)),$$

which completes the proof. □

Problem 2: Understanding of SDE

Briefly explain SDE.

From Time-Series to Start

A better way to understand the **SDE** is from **time-series** (tutorial for beginner), a modeling way of time-related discrete data. We first look at an example of a time-series case, the stock price of Apple's from 2018 to 2024 (Figure 1).

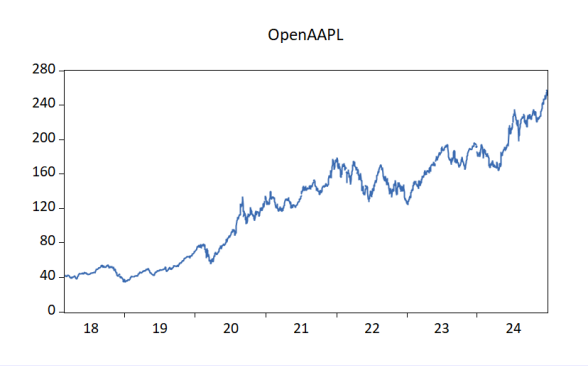

Figure 1: Stock price of Apple Inc (AAPL) from 2018 to 2024 (Data source: Yahoo Finance)

We can see from the graph that there is a trend that the price is rising with a linear relation, with a volatility/diffusion term. One can easily model the price $X(t)$ by AR(1), autoregressive model, plus a constant, we get:

$$X(t) = \mu + X(t-1) + \sigma\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0,1), \quad X(0) = x_0$$

Can also be arranged to:

$$\Delta X(t) = \mu\Delta t + \sigma\Delta\varepsilon_t$$

If we take $\Delta t \to 0$, we get:

$$\mathrm{d}X(t) = \mu\mathrm{d}t + \sigma\mathrm{d}W_t, \quad X(0) = \mathbf{x}_0 \tag{2}$$

which is the form of **SDE**, W_t be the Wiener process and $\mathbf{x}_0 = x_0$ **a.s.** ($\mathbb{P}(\mathbf{x}_0 = x_0) = 1$). (**Note:** Taking $\Delta t \to 0$ here is only intuitive, but not rigorous.)

Drift and Diffusion/Volatility

If we integrate equation (2) from 0 to T, then:

$$X(T) = X(0) + \mu T + \sigma W(T)$$

We can understand the formula by: $X(0)$ be the stock price at time 0 and μ be the expected return and $\sigma W(T)$ be the volatility of the stock price. Here we call such μ as drift term and σ as volatility term or diffusion term. And $X(T)$ is a random variable at time T.

Continuing the stock price for Apple case, if we set 1/1/2020 as $t = 0$, and use a simple SDE (equation 2) to model, we get the result (Figure 2), the volatility is proportional to $\sqrt{T}$; the actual stock price path can be seen as a possibility of the Stochastic process.

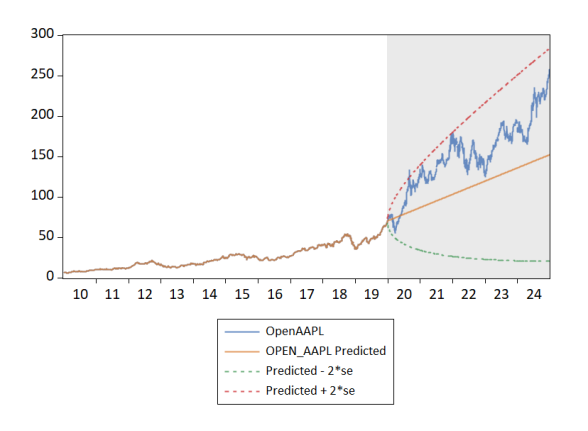

Figure 2: Prediction of the stock price by history data

More on SDE

Instead of putting only constant terms μ and σ on the drift and the volatility/diffusion terms, we can now have a more general form:

$$\mathrm{d}X(t) = f(X(t), t)\mathrm{d}t + g(X(t), t)\mathrm{d}W_t$$

One example is the log normal price process:

$$f(X(t)) = e^{X(t)}; \quad \mathrm{d}X(t) = \left(\mu - \frac{1}{2}\sigma^2\right)\mathrm{d}t + \sigma\mathrm{d}W_t; \quad X(0) = 0$$

One can easily show the price of underlying asset, under risk neutrality, $S(t)$ is:

$$S(t) = S(0)e^{(r-\frac{1}{2}\sigma^2)t+\sigma W(t)}$$

Moreover, with the help of **Itô calculus**, many things that have randomness involved can be modelled by SDE, and have some similar properties to the general calculus. One major use is the **Black-Scholes model**, a model that can estimate the option in the market.

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

The equation is very analogous to the ordinary partial differential equation, but with some special rules that:

$$(\mathrm{d}W_t)^2 = \mathrm{d}t; \quad \mathrm{d}W_t \cdot \mathrm{d}t = (\mathrm{d}W_t)^n = 0, \quad n \geq 3$$

How Is It Different from ODE?

	ODE	SDE
Chain Rule	Yes	Partially (Based on Itô's Lemma)
Solution	Unique	Unique but with Diffusion
Integration	Yes	Yes (Based on Itô's Integral)
Numerical Method	Euler's Method, RK4...	Euler-Maruyama, Milstein Method...
Usage	System w/o Stochasticity	System w/ Stochasticity
Example	Predator Prey	Brownian Motion, Black-Scholes

My Question 1: SDE and Time Series

It is well known that both time series analysis and SDE are powerful tools in research and practical applications. In particular, both play significant roles in finance—for example, time series models are often used for stock price prediction, while SDEs are commonly applied to market pricing. Generally speaking, time series methods handle data in **discrete time**, whereas SDEs are designed for **continuous** processes grounded in calculus. Interestingly, despite SDEs being naturally suited for continuous phenomena, time series methods are used far more widely across various domains, such as statistics and weather forecasting. Could you try to explain the reason we opt for time series more? (Refined by `ChatGPT`)

References

[1] Tesheng Lin. 2025_ml_week_8. https://hackmd.io/@teshenglin/2025_ML_week_8_AS, 2025. Accessed: 27 October 2025.