Machine Learning Assignment 08

313652008 黃睿帆

October 25, 2025

Problem 1–Sliced Score Matching (SSM)

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

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

Background: Recall we did in HW07, in generative modeling, our goal is to learn the *score function*

$$S(x;\theta) = \nabla_x \log p(x;\theta),$$

which is the gradient of the log-density function. However, instead of maximizing likelihood directly, *score matching* learns $S(x;\theta)$ by minimizing the difference between the model score and the true score $\nabla_x \log p(x)$, since $p(x;\theta)$ is not available explicitly.

The *Implicit Score Matching* loss is defined as

$$L_{\text{ISM}}(\theta) = \mathbb{E}_{x\sim p(x)}\left[\|S(x;\theta)\|^2 + 2\,\nabla_x\cdot S(x;\theta)\right].$$

Here $S(x;\theta)\in\mathbb{R}^d$ is the model score, and

$$\nabla_x\cdot S(x;\theta) = \sum_{i=1}^{d}\frac{\partial S_i(x;\theta)}{\partial x_i}$$

is the divergence of the score function.

Rewriting the Divergence as a Trace: In the class, we once mentioned that the divergence can be written more compactly as a matrix trace:

$$\nabla_x\cdot S(x;\theta) = \text{tr}(\nabla_x S(x;\theta)),$$

where $\nabla_x S(x;\theta)$ is the Jacobian matrix of S:

$$[\nabla_x S(x;\theta)]_{ij} = \frac{\partial S_i(x;\theta)}{\partial x_j}.$$

The trace operator simply sums the diagonal entries of this Jacobian, which equals the divergence.

Hutchinson’s Trace Estimator: Since directly to calculate trace is still too hard. We using *Hutchinson’s trace estimator* to simplify. Let $v\in\mathbb{R}^d$ be a random vector with zero mean and identity covariance, i.e.

$$\mathbb{E}_v[vv^\top] = I.$$

Then for any matrix $A\in\mathbb{R}^{d\times d}$,

$$\text{tr}(A) = \mathbb{E}_v[v^\top Av].$$

Applying Hutchinson's Estimator to the Divergence Term: Using this estimator, the divergence term in L_{ISM} can be written as

$$\operatorname{tr}(\nabla_x S(x;\theta)) = \mathbb{E}_v \left[v^\top (\nabla_x S(x;\theta))v\right].$$

By the chain rule, this expression can be equivalently written as

$$v^\top \nabla_x S(x;\theta) v = v^\top \nabla_x \big(v^\top S(x;\theta)\big),$$

which is computationally simpler to implement.

Substituting this stochastic estimate into the ISM loss yields the *Sliced Score Matching (SSM)* loss:

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

Proof of Problem 1.

So to reach our conclusion. We starting from the definition of the Sliced Score Matching (SSM) loss:

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

where $v \in \mathbb{R}^d$ is a random vector satisfying $\mathbb{E}_v[vv^\top] = I$

First we **rewrite the first term (Expectation of v).** Observe that

$$\|S(x;\theta)\|^2 = S(x;\theta)^\top S(x;\theta) = S(x;\theta)^\top \mathbb{E}_v[vv^\top] S(x;\theta) = \mathbb{E}_v\left[S(x;\theta)^\top vv^\top S(x;\theta)\right].$$

Hence,

$$\|S(x;\theta)\|^2 = \mathbb{E}_v\left[(v^\top S(x;\theta))^2\right].$$

Next we substitute the above result back into L_{SSM}. Since expectation (function) is linear, so we have

$$L_{\text{SSM}}(\theta) = \mathbb{E}_{x\sim p(x)} \mathbb{E}_{v\sim p(v)} \Big[(v^\top S(x;\theta))^2 + 2\, v^\top \nabla_x \big(v^\top S(x;\theta)\big)\Big]. \tag{1}$$

Combining the expectations, the equivalent compact form of the SSM loss is:

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

Problem 2– Explanation of Stochastic Differential Equation (SDE)

1. Definition

A **stochastic differential equation (SDE)** describes the evolution of a random process:

$$dx_t = \underbrace{f(x_t,t)}_{\text{drift}}\, dt + \underbrace{G(x_t,t)}_{\text{diffusion}}\, dW_t, \quad x(0) = x_0,$$

where:

- $x_t \in \mathbb{R}^d$ is the stochastic process (the unknown)/(Wiener Process).
- $f(x_t,t) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is the **drift term** —the deterministic part.
- $G(x_t,t) : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d\times d}$ is the **diffusion term** —the random or noisy part.
- W_t : standard **Brownian motion (Wiener process)**.

In words: the infinitesimal change dx_t consists of a deterministic part $f\, dt$ and a random part $G\, dW_t$.

2. Integral Form (Ito Integral Equation)

The Ito integral form of the SDE is:

$$x_t = x_0 + \int_0^t f(x_s, s)\, ds + \int_0^t G(x_s, s)\, dW_s.$$

The first integral is deterministic, while the second one is a stochastic (Ito) integral. A process x_t satisfying this is called an **Ito process**.

For existence and uniqueness of solutions of an differential equations, f and G should at least satisfying the following:

$$\|f(x,t) - f(y,t)\| + \|G(x,t) - G(y,t)\| \le L\|x - y\|, \quad \text{(Lipschitz)}$$
$$\|f(x,t)\|^2 + \|G(x,t)\|^2 \le C(1 + \|x\|^2)$$

We have two special Cases:

- **Pure drift:** $G \equiv 0, dx_t = f(x_t, t)\, dt$ —deterministic ODE.
- **Pure diffusion:** $f \equiv 0, dx_t = G(x_t, t)\, dW_t$ —random motion with zero drift.

3. Stochastic Process

A stochastic process is a parametrized collection of random variables

$$\{x_t\}_{t \in T},$$

where T is the index set (time), e.g., $T = \{1, 2, 3, ...\}$ or $T \in [0, \infty)$. For each fixed t, x_t is a random variable; for each outcome ω, the mapping $t \mapsto x_t(\omega)$ is called a **path** or **realization**. T is defined in a probability space and takes value in $\mathbb{R}^d$. (比一般傳統的 ODE 多了 diffusion terms.)

4. Wiener Process (Brownian Motion)

A d-dimensional Wiener process W_t is continuous stochastic process that satisfies:

1. $W_0 = 0$.
2. Stationary Gaussian increments: $W_{t+u} - W_t \sim \mathcal{N}(0, uI)$.
3. Independent increments: increments over disjoint time intervals are independent, i.e., $0 = t_0 < t_1 < .... < t_n = T$ and $W_{t_1} - W_{t_0}, W_{t_2} - W_{t_1}, ..., W_{t_n} - W_{t_{n-1}}$ is independent.
4. Continuous paths: $t \mapsto W_t$ is continuous almost surely.

Properties:

$$\mathbb{E}[W_t] = 0, \quad \mathrm{Var}(W_t) = t.$$

With probability 1, W_t is nowhere differentiable. Formally, its "derivative" is called **white noise**:

$$h(t) = \frac{dW_t}{dt}.$$

5. White Noise

A white noise process $h(t) \in \mathbb{R}^d$ satisfies:

$$\mathbb{E}[h(t)] = 0,$$
$$\mathbb{E}[h(t)h(s)^T] = \delta(t - s)I.$$

The paths of white noise are discontinuous and unbounded. Brownian motion can be regarded as the time integral of white noise:

$$W_t = \int_0^t h(s)\, ds.$$

The Ito Integral: The stochastic integral is defined as the mean-square limit:

$$\int_0^t G(x_s, s)\, dW_s = \lim_{n\to\infty} \sum_{k=0}^{n-1} G(x(t_k), t_k)\, [W(t_{k+1}) - W(t_k)],$$

where the increments $W(t_{k+1}) - W(t_k) \sim \mathcal{N}(0, (t_{k+1} - t_k)I)$.
The simplest case:

$$\int_0^t dW_s = W_t - W_0 = W_t \sim \mathcal{N}(0, tI).$$

6. Euler–Maruyama Method

We consider a stochastic differential equation (SDE) of the form

$$dX_t = f(X_t, t)\, dt + G(X_t, t)\, dW_t, \quad X_0 = x_0,$$

where W_t denotes a standard Wiener process. The **Euler–Maruyama method** provides a simple numerical approximation for such SDEs, following a strategy similar to the forward Euler method for deterministic ODEs.

Algorithm introduced in the class

1. Partition the time interval $[0, T]$ into N equal subintervals with step size

$$\Delta t = \frac{T}{N} > 0, \quad t_k = k\Delta t, \quad k = 0, 1, \ldots, N.$$

2. Initialize with $X_0 = x_0$.

3. For each step, update using

$$X_{n+1} = X_n + f(X_n, t_n)\, \Delta t + G(X_n, t_n)\, \Delta W(t_n),$$

where $\Delta W(t_n) = W(t_{n+1}) - W(t_n)$.
Since $\Delta W(t_n) \sim \mathcal{N}(0, \Delta t)$, it can equivalently be written as

$$X_{n+1} = X_n + f(X_n, t_n)\, \Delta t + G(X_n, t_n)\sqrt{\Delta t}\, Z_n,$$

where $\{Z_n\}$ are independent standard normal random variables, $Z_n \sim \mathcal{N}(0, 1)$.

Three Examples in 1D

Example 1: Pure Diffusion Process. Consider the SDE

$$dx_t = \sigma\, dW_t, \quad x(0) = x_0,$$

where $\sigma > 0$ and x_0 is a constant.
The exact solution is obtained by direct integration:

$$x(t) = x_0 + \sigma \int_0^t dW_s = x_0 + \sigma W_t.$$

Since $W_t \sim \mathcal{N}(0, t)$, it follows that

$$x(t) \sim \mathcal{N}(x_0, \sigma^2 t).$$

Thus, we have

$$\mathbb{E}[x(t)] = x_0, \qquad \mathrm{Var}[x(t)] = \sigma^2 t.$$

The corresponding probability density function is

$$p(x, t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{(x - x_0)^2}{2\sigma^2 t}\right),$$

and $p(x, t)$ satisfies the diffusion (heat) equation:

$$\frac{\partial p}{\partial t} = \frac{\sigma^2}{2} \frac{\partial^2 p}{\partial x^2}.$$

Example 2: Constant Drift and Diffusion. Consider the SDE

$$dx_t = \mu\, dt + \sigma\, dW_t, \quad x(0) = x_0,$$

where $\mu > 0$, $\sigma > 0$, and x_0 is a constant.
Integrating gives the exact solution

$$x(t) = x_0 + \int_0^t \mu\, ds + \sigma \int_0^t dW_s = x_0 + \mu t + \sigma W_t.$$

Hence,

$$x(t) \sim \mathcal{N}(x_0 + \mu t,\, \sigma^2 t),$$

with mean and variance

$$\mathbb{E}[x(t)] = x_0 + \mu t, \qquad \mathrm{Var}[x(t)] = \sigma^2 t.$$

Example 3: Ornstein–Uhlenbeck (OU) Process. Consider the OU process defined by

$$dx_t = -\beta x_t\, dt + \sigma\, dW_t, \quad x(0) = x_0,$$

where $\beta > 0$, $\sigma > 0$, and x_0 is a constant.
This SDE describes a mean-reverting process, often used in physics and finance.
The exact solution is known to be

$$x(t) = x_0 e^{-\beta t} + \sigma \int_0^t e^{-\beta(t-s)}\, dW_s.$$

The process is Gaussian with

$$\mathbb{E}[x(t)] = x_0 e^{-\beta t}, \qquad \mathrm{Var}[x(t)] = \frac{\sigma^2}{2\beta}\left(1 - e^{-2\beta t}\right).$$

Unanswered Questions

- How does numerical stability differ between deterministic methods (like Euler’s method) and stochastic methods (like Euler—Maruyama)?
- Is it possible that an SDE approximation diverge even when the drift and diffusion terms are well-behaved?
- Under what condition can every SDE uniquely define a probability density evolution?