

# Assignment 7

2025 年 11 月 25 日

## 1 Score Matching 的概念與其在分數基礎生成模型中的應用

### 1.1 Score Matching 概念

#### 1.1.1 目標

在生成模型中，我們希望學習資料的機率密度函數 (PDF)  $p(x)$ ，通常表示為能：

$$p(x) = \frac{1}{Z(\theta)} e^{q(x;\theta)}$$

但由於歸一化常數  $Z(\theta) = \int e^{q(x;\theta)} dx$  非常難以計算，因此我們轉而使用其他方法。

#### 1.1.2 Score Function (分數函數)

對  $p(x)$  取對數並計算關於  $x$  的梯度，得到：

$$\nabla_x \log p(x; \theta) = \nabla_x (q(x; \theta) - \log Z(\theta)) = \nabla_x q(x; \theta)$$

我們定義 Score function  $S(x)$  為：

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

分數函數  $S(x)$  指出了在資料空間中每個點  $x$  上，機率密度函數  $p(x)$  上升最快的方向。

#### 1.1.3 Explicit Score Matching (ESM)

目標是訓練模型  $S(x; \theta)$  近似真實分數  $\nabla_x \log p(x)$ 。定義 ESM 損失函數為均方誤差：

$$L_{\text{ESM}}(\theta) = \mathbb{E}_{x \sim p(x)} \|S(x; \theta) - \nabla_x \log p(x)\|^2$$

挑戰：在實際應用中，我們並不知道真實的梯度  $\nabla_x \log p(x)$ ，因此無法直接計算此損失。

#### 1.1.4 Implicit Score Matching (ISM)

為了解決 ESM 依賴真實梯度的問題，ISM 利用了積分恆等式 (Integration by Parts)，將損失函數改寫為：

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

其中  $\nabla_x \cdot S(x; \theta)$  是  $S(x; \theta)$  的散度 (Divergence, 或稱為 Trace of the Jacobian  $\text{tr}(\nabla_x S(x; \theta))$ )。使用 ISM 就不需要計算  $\nabla_x \log p(x)$ 。

關係：透過數學推導可以證明：

$$\mathbb{E}_{x \sim p(x)} [\|S(x; \theta) - \nabla_x \log p(x)\|^2] = \mathbb{E}_{x \sim p(x)} [\|S(x; \theta)\|^2 + 2\nabla_x \cdot S(x; \theta)] + \mathbb{E}_{x \sim p(x)} [\|\nabla_x \log p(x)\|^2]$$

因此，最小化  $L_{\text{ISM}}(\theta)$  等價於最小化  $L_{\text{ESM}}(\theta)$  (因為最後一項  $\mathbb{E}[\|\nabla_x \log p(x)\|^2]$  與  $\theta$  無關)。

### 1.1.5 Denoising Score Matching (DSM)

ISM 在實作中仍可能涉及高維度散度的複雜計算。DSM 提出了更簡化的方法：

1. 加噪處理：先對資料  $x_0$  加上噪音，得到  $x \sim p(x|x_0)$ 。
2. 學習 Noisy Score：希望模型  $S_\sigma(x; \theta)$  學習 noisy 數據分佈  $p_\sigma(x)$  的分數：

$$S_\sigma(x; \theta) = \nabla_x \log p_\sigma(x) \quad , \quad p_\sigma(x) = \int_{\mathbb{R}^d} p(x|x_0)p_0(x_0) dx_0$$

DSM 的損失函數定義為：

$$L_{\text{DSM}}(\theta) = \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p(x|x_0)} [\|S_\sigma(x; \theta) - \nabla_x \log p(x|x_0)\|^2]$$

**Why Use DSM (簡化計算)：** 若我們使用等方差 (Isotropic) 的高斯噪音：

$$x = x_0 + \epsilon_\sigma, \quad \epsilon_\sigma \sim \mathcal{N}(0, \sigma^2 I) \quad \text{或} \quad x = x_0 + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

此時，條件機率的梯度有一個簡單的解析解：

$$\nabla_x \log p(x|x_0) = -\frac{(x - x_0)}{\sigma^2}$$

將此結果代入  $L_{\text{DSM}}(\theta)$  得到：

$$L_{\text{DSM}}(\theta) = \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p(x|x_0)} \left[ \left\| S_\sigma(x; \theta) + \frac{(x - x_0)}{\sigma^2} \right\|^2 \right]$$

這樣，目標變成了訓練模型  $S_\sigma(x; \theta)$  來預測施加的噪音  $\epsilon = \frac{x - x_0}{\sigma}$  (乘上  $-\frac{1}{\sigma}$ )，極大地簡化了計算。

### 1.1.6 Sliced Score Matching (SSM)

在高維度空間中，計算 ISM 損失中的散度  $\nabla_x \cdot S(x; \theta) = \sum_{i=1}^d \frac{\partial S_i(x; \theta)}{\partial x_i}$  仍是一個挑戰，因為涉及計算巨大的 Jacobian 矩陣  $\nabla_x S(x; \theta)$  的 Trace。

SSM 使用 Hutchinson's Trace Estimator 來近似散度：當  $v$  是一個滿足  $\mathbb{E}[vv^T] = I$  的隨機向量 (例如標準高斯分佈)，則矩陣  $A$  的 Trace 可以近似為：

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

利用這個原理，ISM 損失可以改寫為：

$$\nabla_x \cdot S(x; \theta) = \mathbb{E}_{v \sim p(v)}[v^T \nabla_x (v^T S(x; \theta))]$$

最終的 SSM 損失函數為：

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

這使得高維度中的分數匹配訓練變得可行。

## 1.2 總結：Score Matching 在生成模型中的應用

- 核心理念：在生成模型中，直接學習資料的機率密度函數 (PDF)  $p(x; \theta)$  很困難，因此改為學習它的對數梯度，即 Score function  $S(x) = \nabla_x \log p(x; \theta)$ 。
- 應用：分數基礎模型，特別是基於擴散 (Diffusion-based) 的生成模型，通過訓練一個神經網路  $S_\theta(x)$  來估計  $S(x)$ 。
- 模型訓練：這些模型通常使用 Denoising Score Matching (DSM)\*\* 的多尺度版本 (如 Noise Conditional Score Network, NCSN) 或 Sliced Score Matching (SSM) 進行訓練。
- 樣本生成 (採樣)：訓練好的分數函數  $S_\theta(x)$  可以搭配採樣算法 (如 Langevin Dynamics) 來從學到的數據分佈中生成新的樣本。

## 2 Unanswered Questions (待討論問題)

1. DSM 中的噪音選擇：在操作 Denoising Score Matching (DSM) 時，我們應如何去選擇噪音  $\sigma$ ？選擇的準則應能讓我們方便計算，同時保證結果的準確性 (特別是對於多尺度  $\sigma$  的選擇，如何確保模型在所有尺度上都能準確預測分數)。
2. DSM 與 SSM 的結合：我們是否可以結合 DSM 與 SSM 的觀念？例如，在加噪後的分布  $p_\sigma(x)$  上，使用 SSM 的方法來估計散度  $\nabla_x \cdot S_\sigma(x; \theta)$ ，從而創造出效果可能更好的訓練方法？