1. Concept of Score Matching

Score Matching 學習未歸一化機率模型（unnormalized models）的方法

MLE需要計算 $\log p_\theta(x)$ 的 **歸一化常數（partition function）**，但在能量模型或擴散模型中，這個常數往往無法直接求得

Score Matching 的想法是：與其學習整個機率密度，不如直接學習 **分布的梯度方向（score function）**：

$$\psi(x;\theta) = \nabla_x \log p_\theta(x)$$

這個梯度描述了資料點在特徵空間中指向高機率區域的方向
若模型的 score 與真實資料的 score 相符，代表模型分布與真實分布一致

---

2. Explicit vs. Implicit Score Matching

2.1 Explicit Score Matching (ESM)

ESM 的目標是直接最小化模型 score 與真實 score 的 L2 距離期望：

$$J_{ESM_q}(\theta) = \mathbb{E}_{q(x)}\left[\frac{1}{2}\|\psi(x;\theta) - \nabla_x \log q(x)\|^2\right]$$

然而，真實分布的 $\nabla_x \log q(x)$ 是未知的，導致此式 無法直接計算

2.2 Implicit Score Matching (ISM)

Hyvärinen（2005）提出 **Implicit Score Matching (ISM)**，透過積分轉換避免直接出現 $\nabla_x \log q(x)$，得到可計算的形式：

$$J_{ISM_q}(\theta) = \mathbb{E}_{q(x)}\left[\mathrm{tr}(\nabla_x \psi(x;\theta)) + \frac{1}{2}\|\psi(x;\theta)\|^2\right]$$

其中
$\mathrm{tr}(\nabla_x \psi)$ 為 Jacobian 的跡（trace of the Jacobian）
ISM 的優點是只需模型 $\psi(x;\theta)$ 的一階與二階導數，而不需真實資料分布的梯度

3. Proof of Equivalence (ESM ⇔ ISM)

為證明兩者等價，只需證明它們在 θ 上的差異僅為常數 C

Step 1. 展開 ESM

$$J_{ESM_q}(\theta) = \mathbb{E}_{q(x)}\left[\frac{1}{2}\|\psi(x;\theta)\|^2 - \psi(x;\theta)^\top \nabla_x \log q(x) + \frac{1}{2}\|\nabla_x \log q(x)\|^2\right]$$

Step 2. 拆分

- (a) $\mathbb{E}_{q(x)}\left[\frac{1}{2}\|\psi(x;\theta)\|^2\right]$
- (b) $-\mathbb{E}_{q(x)}\left[\psi(x;\theta)^\top \nabla_x \log q(x)\right]$
- © $\mathbb{E}_{q(x)}\left[\frac{1}{2}\|\nabla_x \log q(x)\|^2\right]$

其中 © 與模型參數 θ 無關，因此可略去

Step 3. 處理 (b)

$\nabla_x \log q(x) = \frac{\nabla_x q(x)}{q(x)}$：

$$(b) = -\int q(x)\psi(x;\theta)^\top \frac{\nabla_x q(x)}{q(x)} dx = -\int \psi(x;\theta)^\top \nabla_x q(x) dx$$

根據高斯散度定理（Gaussian divergence theorem）：

$$(b) = \mathbb{E}_{q(x)}[\mathrm{tr}(\nabla_x \psi(x;\theta))]$$

Step 4. 合併結果

$$J_{ESM_q}(\theta) = \mathbb{E}_{q(x)}\left[\mathrm{tr}(\nabla_x \psi(x;\theta)) + \frac{1}{2}\|\psi(x;\theta)\|^2\right] + C$$

即：

$$J_{ESM_q}(\theta) = J_{ISM_q}(\theta) + C$$

因此，最小化 J_{ESM_q} 等價於最小化 J_{ISM_q}

---

4. Denoising Score Matching (DSM)

Introduction

在實際應用中，ISM 雖然可行，但在高維度或資料稀疏時會有數值不穩定問題
Denoising Score Matching (DSM) 是 Vincent (2011) 提出的改進方法，透過對資料加上隨機噪聲後學習 denoising score

---

DSM Loss — Noisy Score Function

Notation:

- x_0: original (clean) data
- $p_0(x_0)$: data distribution
- x: noisy version of x_0
- $p(x|x_0)$: conditional noisy data distribution
- $p_\sigma(x)$: marginal noisy data distribution

且有

且有

$$p_\sigma(x) = \int_{\mathbb{R}^d} p(x|x_0) p_0(x_0)\, dx_0$$

DSM 要學習的 **noisy score function** 定義為：

$$S_\sigma(x;\theta) = \nabla_x \log p_\sigma(x)$$

---

5. Application in Diffusion (Score-based Generative Models)

在 **score-based diffusion models** 中，Score Matching 用來學習時間依賴的 score function：

$$s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$$

模型不直接生成資料，而是學習如何從噪聲中逐步「逆轉」擴散過程
Sampling 階段則從高斯分布開始，利用學得的 score 函數逐步去噪（denoise）還原資料

6. Unanswered or Open Questions

在 diffusion 過程中，時間依賴的 score 是否唯一？

---

7. Summary

方法	形式	是否需 $\nabla_x \log q(x)$	可計算性	應用
ESM	$\mathbb{E}_{q(x)}[\frac{1}{2}\lvert\psi - \nabla_x \log q\rvert^2]$	✅ 需要	❌ 無法直接計算	理論分析
ISM	$\mathbb{E}_{q(x)}[\mathrm{tr}(\nabla_x \psi) + \frac{1}{2}\lvert\psi\rvert^2]$	❌ 不需要	✅ 可計算	實際訓練
DSM	加噪後估計 score	❌	✅ 高維穩定	Diffusion models

本文件之文獻整理與內容編排由 GPT 協助完成