

1.

Score function

我們一開始的目標是學習數據分佈 $p_{\text{data}}(x)$ 的參數化 PDF: $p(x; \theta)$

PDF的性質 ① For any x , $p(x) \geq 0$. 所以令 $p(x; \theta) = e^{g(x; \theta)} \geq 0$

$$\textcircled{2} \int p(x) dx = 1 \text{ 所以令 } Z(\theta) = \int e^{g(x; \theta)} dx, p(x; \theta) = \frac{1}{Z(\theta)} e^{g(x; \theta)}$$

$$\text{s.t. } \int p(x; \theta) dx = \int \frac{1}{Z(\theta)} e^{g(x; \theta)} dx = 1$$

此時 $p(x; \theta) = \frac{1}{Z(\theta)} e^{g(x; \theta)}$ 已建構完成，這在低維時確實可行

但在高維度時 $Z(\theta) = \int e^{g(x; \theta)} dx$ 是沒辦法計算的

為了方便計算，我們在不丟失“特性”的同時 試著避開 $Z(\theta)$.

取對數並不影響高低值的相對順序

所以我們取對數 $\ln p(x; \theta) = g(x; \theta) - \underline{Z(\theta)} \rightarrow \nabla_x Z(\theta) = 0$

再取梯度 $\nabla_x \ln p(x; \theta) = \nabla_x g(x; \theta)$

稱之為 Score function: $S(x; \theta) = \nabla_x \ln p(x; \theta) = \nabla_x g(x; \theta)$

實際意義： $P_{\text{data}}(x)$ 是「真實資料分佈」，假設你在識別寫著數字 0 的圖片

$P_{\text{data}}(x)$ 就是你不知道長怎樣的函數，你餵給它一張圖 x ，它會告訴你這張圖有多像“0” ($P_{\text{data}}(x)$ 值越高表示越像)

$\Rightarrow \ln P_{\text{data}}(x)$ 不影響高低值的相對順序

$\Rightarrow \nabla_x P_{\text{data}}(x) = S(x)$ 是一個指向函數上升最快方向的向量

也就是對於任意一個 x ，往 $S(x)$ 這個方向修改 x ，它會變得更像典型的“0”

Score Matching

ESM, ISM

最小化 $L_{ESM}(\theta) = E_{x \sim p(x)} \|S(x; \theta) - \nabla_x \log p_{data}(x)\|^2$ 使 $S(x; \theta)$ 尽可能接近 $\nabla_x \log p_{data}(x)$

由於不知道 $p_{data}(x)$ 的樣子，我們無法做

引進 $L_{ISM}(\theta) = E_{x \sim p(x)} [\|S(x; \theta)\|^2 + 2 \nabla_x \cdot S(x; \theta)]$ ，這個我們能算

並且

$$\begin{aligned}
& E_{x \sim p(x)} [\|S(x) - \nabla_x \log p(x)\|^2] \\
&= E_{x \sim p(x)} [\|S(x)\|^2 - 2S(x) \cdot \nabla_x \log p(x) + \|\nabla_x \log p(x)\|^2] \\
&= E_{x \sim p(x)} [\|S(x)\|^2] - 2 \int_{R^d} (S(x) \cdot \nabla_x \log p(x)) p(x) dx + E_{x \sim p(x)} [\|\nabla_x \log p(x)\|^2] \\
&= E_{x \sim p(x)} [\|S(x)\|^2] - 2 \int_{R^d} S(x) \cdot \nabla_x p(x) dx + E_{x \sim p(x)} [\|\nabla_x \log p(x)\|^2] \\
&= E_{x \sim p(x)} [\|S(x)\|^2] + 2 \int_{R^d} (\nabla_x \cdot S(x)) p(x) dx + E_{x \sim p(x)} [\|\nabla_x \log p(x)\|^2] \\
&= E_{x \sim p(x)} [(\|S(x)\|^2 + 2 \nabla_x \cdot S(x))] + \underbrace{E_{x \sim p(x)} [\|\nabla_x \log p(x)\|^2]}_C = C
\end{aligned}$$

$$L_{ESM}(\theta) = L_{ISM}(\theta) + C$$

我們不能算，但其為定值

故最小化 $L_{ESM}(\theta)$ 和最小化 $L_{ISM}(\theta)$ 的功能一樣

DSM

• 符號定義 (Notation):

- x_0 : 原始的、乾淨的數據 (original data)。
- $p_0(x_0)$: 原始數據的真實分佈。
- x : 帶有噪聲的數據 (noisy data)，由 x_0 擾動而來。
- $p(x|x_0)$: 條件機率，即給定一個乾淨數據 x_0 ，產生噪聲數據 x 的機率 (這就是「加噪過程」)。
- $p_\sigma(x)$: 噪聲數據的邊際分佈。它可以通過對所有可能的 x_0 積分得到：

$$p_\sigma(x) = \int p(x|x_0)p_0(x_0)dx_0.$$

• DSM 的目標 (Goal):

DSM 的目標是訓練一個模型 $S_\sigma(x; \theta)$ 來估計噪聲數據分佈 $p_\sigma(x)$ 的分數 (score)，即

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

Let $L_{DSM}(\theta) = E_{x_0 \sim p_0(x_0)} E_{x|x_0 \sim p(x|x_0)} [\|S_\sigma(x; \theta) - \nabla_x \log p(x|x_0)\|^2].$

我們定義噪聲數據的 ESM 損失為：

$$\begin{aligned}
L_{ESM_noisy}(\theta) &= E_{x \sim p_\sigma(x)} [\|S_\sigma(x; \theta) - \nabla_x \log p_\sigma(x)\|^2] \\
&= E_{x \sim p_\sigma(x)} [\|S_\sigma(x)\|^2 - 2S_\sigma(x) \cdot \nabla_x \log p_\sigma(x) + \|\nabla_x \log p_\sigma(x)\|^2]
\end{aligned}$$

$L_{ESM_noisy}(\theta)$

$$= \mathbb{E}_{x \sim p_\sigma(x)} [\|S_\sigma(x)\|^2 - 2S_\sigma(x) \cdot \nabla_x \log p_\sigma(x) + \|\nabla_x \log p_\sigma(x)\|^2]$$

$$= \mathbb{E}_{x_0, x} [\|S_\sigma(x)\|^2] - 2\mathbb{E}_{x_0, x} [\langle S_\sigma(x), \nabla_x \log p(x|x_0) \rangle] + \mathbb{E}_{x \sim p_\sigma(x)} [\|\nabla_x \log p_\sigma(x)\|^2]$$

$$= E_{x_0, x} [\|S_\sigma(x)\|^2 - 2\langle S_\sigma(x), \nabla_x \log p(x|x_0) \rangle + \|\nabla_x \log p(x|x_0)\|^2]$$

$\hookrightarrow L_{DSM}(\theta)$

$$+ E_{x \sim p_\sigma(x)} [\|\nabla_x \log p_\sigma(x)\|^2] - E_{x_0, x} [\|\nabla_x \log p(x|x_0)\|^2]$$

$\hookrightarrow Constant$

$$L_{ESM_noisy}(\theta) = L_{DSM}(\theta) + C$$

所以最小化 $L_{ESM}(\theta)$ 和最小化 $L_{ISM}(\theta)$ 和最小化 $L_{DSM}(\theta)$ 是等價的

Score matching used in score-based generative models

*這部分使用AI工具輔助生成

擴散模型的「生成」過程是一個逐步「去噪」的過程，而要正確地去噪，模型在每一步都必須知道「噪音該往哪個方向減少」。這個「方向」就是分數(score)。

分數匹配(Score Matching)，特別是您圖片中提到的去噪分數匹配(Denoising Score Matching, DSM)，就是用來訓練一個神經網路，使其能夠準確估計這個「方向」的方法。

擴散模型的兩個過程

要理解這一點，我們需要看擴散模型的兩個過程：

1. 前向過程(Forward Process): 加噪

這是一個固定的、不需要學習的過程。我們從一張乾淨的圖片 x_0 開始，然後逐步(例如 $T = 1000$ 步)向其添加高斯噪聲，直到它在 T 時刻變成一張完全是噪聲的圖片 x_T 。

- $x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_T$

- 我們完全知道在任意 t 時刻，給定 x_0 , x_t 的分佈 $q(x_t|x_0)$ 是什麼。

2. 反向過程(Reverse Process): 去噪(生成)

這就是我們希望模型學習的過程。我們從純噪聲 x_T 開始，然後逐步反向運行，一步一步地去除噪聲，直到在 $t = 0$ 時刻得到一張乾淨的圖片 x_0 。

- $x_T \rightarrow x_{T-1} \rightarrow \dots \rightarrow x_1 \rightarrow x_0$

- **關鍵問題：**在 t 時刻，我們手上有一張噪聲圖片 x_t ，我們該如何計算出「上一步」的、更乾淨的圖片 x_{t-1} ？

分數匹配的作用：估計去噪方向

理論證明，要從 x_t 逆向推導出 x_{t-1} ，我們需要知道在 t 時刻的噪聲數據分佈 $p_t(x_t)$ 的分數(score)。

(score)，即 $\nabla_x \log p_t(x_t)$ 。

- $p_t(x_t)$: 在 t 時刻，所有可能的噪聲圖片(來自所有 x_0)的整體分佈。

- $\nabla_x \log p_t(x_t)$: 這個分佈的分數。這是一個向量場，它指向數據密度 $p_t(x_t)$ 增加最快的方

向。直觀地說，它指向「更像數據」的方向，也就是去噪的方向。

但這裡有一個大問題：

我們根本不知道 $p_t(x_t)$ 這個複雜的分佈是什麼(它涉及所有真實數據 p_0)，因此我們無法計算它的分數 $\nabla_x \log p_t(x_t)$ 。

解法：最小化 $L_{ESM}(\theta)$ 和最小化 $L_{ISM}(\theta)$ 和最小化 $L_{DSM}(\theta)$ 是等價的

實際的訓練和生成流程

訓練 (使用 DSM)

1. 從您的數據集 p_0 中隨機抽取一張乾淨圖片 x_0 。
2. 隨機選擇一個時間步 t (即一個噪聲水平 σ_t)。
3. 生成一個隨機高斯噪聲 $\epsilon \sim \mathcal{N}(0, I)$ 。
4. 計算 t 時刻的噪聲圖片 $x_t = x_0 + \sigma_t \epsilon$ 。
5. 將 x_t 和 t 輸入到您的神經網路 S_θ 中，得到預測的分數。
6. 計算損失： $L_{DSM} = \|S_\theta(x_t, t) - (-\frac{\epsilon}{\sigma_t})\|^2$ 。
7. 使用梯度下降更新神經網路 θ 。

注意：許多現代擴散模型（如 DDPM）不是直接預測分數 S_θ ，而是重新參數化，讓網路 $\epsilon_\theta(x_t, t)$ 直接預測噪聲 ϵ 。這在數學上是等價的，因為分數和噪聲只差一個常數縮放。

生成 (使用 S_θ)

1. 從一個純高斯噪聲 $x_T \sim \mathcal{N}(0, I)$ 開始。
2. 迭代 T 次（例如從 $t = 1000$ 到 1 ）：
 - a. 將當前的噪聲圖片 x_t 和時間 t 輸入我們訓練好的神經網路 S_θ 。
 - b. 網路輸出預測的分數 $\nabla_x \log p_t(x_t)$ 。
 - c. 使用這個分數，通過求解反向 SDE (隨機微分方程) 或反向馬可夫鏈，從 x_t 計算出上一步的、更乾淨的 x_{t-1} 。
3. 在 $t = 0$ 時， x_0 就是一張全新的、由模型生成的圖片。

2.

Q: 能說 $p_{\text{data}}(x)$ 靠近 $\overset{!}{\Rightarrow} \text{像}$ $\underset{!}{\Rightarrow} \text{不像}$ 嗎？ PDF $0 \leq p(x) \leq 1$ ？

A: No! $p(x)$ 有可能大於 1,

Why? $p(x)$ 是機率密度，不是機率

For example, the continuous uniform distribution on the interval $[0, \frac{1}{2}]$

has probability density function $p(x) = \begin{cases} 2 & \text{for } 0 \leq x \leq \frac{1}{2} \\ 0 & \text{o.w.} \end{cases}$

$$(\text{check } \int_{-\infty}^{\infty} p(x) dx = \int_0^{\frac{1}{2}} 2 dx = 1)$$