# Machine Learning

313704071 _HW7

1. **Explain the concept of score matching and describe how it is used in score-based (diffusion) generative models.**

Because it is difficult to learn the probability density function of a data distribution which is one of the goals of a generative model, we try to learn the score function.

$$S(x) \approx \nabla_x \log p(x)$$

Score matching minimizes the difference between the model's score $s_\theta(x)$ and the true data score:

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

$\mathbb{E}_{x\sim p(x)}$: expectation of sample data.

Since the true score $\nabla_x \log p(x;\theta)$ is usually intractable, **implicit forms** (via integration by parts) or **denoising versions** are used to make it computable.

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

Using derivation, we know that

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

Hence, $$arg \min_\theta L_{ESM} = arg \min_\theta L_{ISM}$$

### Denoising score matching (DSM)

Instead of modeling $p(x)$ directly, we model how data changes under noise and learn the score function of the *noisy data distributions* at different noise levels.

Let $x_0$: original data, $p_0(x)$: original data

$x$: perturbed data, and $x = x_0 + \epsilon_\sigma$; $p(x)$: pdf of perturbed data.

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

where, $p(x|x_0) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\left(\frac{||x-x_0||}{2\sigma^2}\right)^2}$

$$\nabla_x \log p(x|x_0) = -\frac{1}{\sigma^2}(x - x_0)$$

Add Gaussian noise to the data:

$x = x_0 + \sigma\epsilon,\ \epsilon \in N(0, I)$

$\quad = x_0 + \epsilon_\sigma\ ,\ \epsilon_\sigma \in N(0, \sigma^2 I)$

$$L_{DSM}(\theta) = \mathbb{E}_{x_0 \sim p_0(x)} \mathbb{E}_{(x|x_0) \sim p(x|x_0)} \left\| S_\sigma(x;\theta) + \frac{1}{\sigma^2}(x - x_0) \right\|^2$$
$$= \mathbb{E}_{x_0 \sim p_0(x)} \mathbb{E}_{(x|x_0) \sim p(x|x_0)} \frac{1}{\sigma^2} \|\sigma S_\sigma(x;\theta) + \epsilon\|^2$$

We can conclude that, unlike ESM, DSM does not require the true gradient of the data log-density, and unlike ISM, it avoids complex integration by parts by formulating a computable “denoising” objective. It allows the model to efficiently learn stable score functions across multiple noise levels and has become the foundation of score-based (diffusion) generative models, where the learned scores guide the reverse process that gradually transforms pure noise back into realistic data.

2. Unanswered questions

將資料加噪，訓練 score function 去近似 $\nabla_x \log p(x)$，最後再反向擴散把純噪聲逐步還原成資料。那 DSM 主要有什麼用途? 我能想到的是通常衛星遙測拍攝的相片都會相當模糊，但透過一些除雜訊的技術，就能得到相對清晰的地球表面影像。而 GPT 告訴我更多應用，像是 MRI 影像復原、語音生成、文字生成、分子結構生成等等…