Assignment 7

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

§ Score matching

- To train a model that learns the gradient of the data log-density rather than the density itself. $S(x) = \nabla_x \log p(x)$.
- The true score $\nabla_x \log p(x)$ is unknown and cannot be directly computed.
- Train a neural network $s_\theta(x)$ to approximate by minimizing

(i) Explicit score matching (ESM)

$$L_{ESM}(\theta) = E_{x\sim p(x)}[\| S(x;\theta) - \nabla_x \log p(x)\|^2]$$

(ii) Implicit score matching (ISM)

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

$$L_{ESM}(\theta) = E_{x\sim p(x)}[\| S(x;\theta) - \nabla_x \log p(x)\|^2]$$

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

$$= L_{ISM}(\theta) + C, \text{ where } C \text{ is a constant indep. of } \theta.$$

Hence minimizing L_{ESM} and L_{ISM} are equivalent.

• Denoising score matching (DSM)

Add Gaussian noise to the data x_0 to get $x = x_0 + \varepsilon$

Learn the score of the noisy distribution:

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

Goal: To find the noisy score function:

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

The DSM objective:

$$L_{DSM}(\theta) = E_{x_0 \sim p_0(x_0)}\, E_{x|x_0 \sim p(x|x_0)} \left[\| s_\sigma(x;\theta) - \nabla_x \log p(x|x_0) \|^2 \right]$$

Consider ESM objective to $s_\sigma(x;\theta)$:

$$L_{ESM} = E_{x \sim p_\sigma(x)} \left[\| s_\sigma(x;\theta) - \nabla_x \log p_\sigma(x) \|^2 \right]$$

$$= E_{x_0 \sim p_0(x_0)}\, E_{x|x_0 \sim p(x|x_0)} \left[\| s_\sigma(x) - \nabla_x \log p(x|x_0) \|^2 \right] + C .$$

$$= L_{DSM}(\theta) + C, \text{ where } C \text{ is a constant indep. of } \theta.$$

[Conclusion: To find the noisy score function, DSM, ESM, and ISM are equivalent.]

- Sliced score matching (SSM)

$L_{ISM}(\theta) = E_{x\sim p(x)}[\|s(x;\theta)\|^2 + 2\nabla_x \cdot s(x;\theta)]$

Hutchinson's trace estimator: for a random vector $v \in \mathbb{R}^d$ s.t. $E[vv^T] = I \Rightarrow tr(A) = E_v[v^T A v]$

So $tr(\nabla_x s(x;\theta)) = E_v[v^T \nabla_x (v^T s(x;\theta))]$

$\Rightarrow L_{SSM}(\theta) = E_{x\sim p(x)} \|s(x;\theta)\|^2$

$+ E_{x\sim p(x)} E_{v\sim p(v)}[2 v^T \nabla_x (v^T s(x;\theta))]$

§ DDPM (Denoising Diffusion Probabilistic Models)

- A generative model that adds Gaussian noise to data and learns to reverse the process. It is a practical realization of DSM.
- Graphical model: $x_T \to x_{T-1} \to \cdots \to x_0$

  forward: add noise ($q(x_t|x_{t-1})$)

  reverse: learn to denoise ($p_\theta(x_{t-1}|x_t)$).
- The network $\epsilon_\theta(x_t, t)$ predicts noise, equivalently estimating the score $s_\theta(x_t, t) \approx \nabla_{x_t} \log p_t(x_t)$. Hence DDPM implements DSM across multiple noise levels.

2. Unanswered Questions
   There are unanswered questions from the lecture, and there are likely more questions we haven't covered.

- Take a moment to think about these questions.
- Write down the ones you find important, confusing, or interesting.
- You do **not** need to answer them—just state them clearly.

How does loss function of score matching derived, and what is the intuitive meaning behind it?
(Where this form comes from and why minimizing the gradient difference is equivalent to learning the data distribution?)