Assignment_7

1. What is Score Matching?

Score matching is a method for training probabilistic models **without explicitly computing the likelihood**.
Instead of learning the probability density $p_\theta(x)$ directly, we learn its **score function**, i.e. the gradient of its log-density:

$$s_\theta(x) = \nabla_x \log p_\theta(x)$$

This vector field $s_\theta(x)$ called the **score** — points in the direction where the data density increases most rapidly.

2. Why use Score Matching?

For many models, computing the likelihood $p_\theta(x)$ or its gradient is **intractable** because of a partition function Z_θ:

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$$

where $E_\theta(x)$ is an energy function.
Since $Z_\theta = \int \exp(-E_\theta(x))\, dx$ is often impossible to compute, **maximum likelihood training** becomes infeasible.

Score matching avoids this problem:
Hyvärinen (2005) showed we can fit $s_\theta(x)$ **without needing** Z_θ using the following objective:

$$J(\theta) = \frac{1}{2} E_{p_{data}(x)}[||s_\theta(x) - \nabla_x \log p_{data}(x)||^2]$$

Since $\nabla_x \log p_{data}(x)$ is unknown, Hyvärinen derived a **tractable equivalent**:

$$J(\theta) = E_{p_{data}(x)}[\frac{1}{2}||s_\theta(x)||^2 + \nabla_x \cdot s_\theta(x)$$

which depends only on $s_\theta(x)$ and its divergence.

3. Connection to Diffusion (Score-Based) Models

In **score-based generative models** (Song & Ermon, 2019–2021), we extend score matching to **noisy data** and **continuous diffusion processes**.

(a) Noise Perturbation

Instead of matching scores only on clean data, we train the model to predict scores of **noisy versions** of data:

$$x_t = x_0 + \text{noise}$$

for varying noise levels t.
Then, the model $s_\theta(x_t, t)$ approximates:

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

where $p_t(x_t)$ is the data distribution after adding Gaussian noise of variance corresponding to t.

This is called **denoising score matching**.

---

4. Generating Samples

Once we've learned all these scores $s_\theta(x_t, t)$, we can **reverse the diffusion process** using stochastic differential equations (SDEs):

(a) Forward SDE (adds noise)

$$dx = f(x, t)dt + g(t)dW_t$$

(b) Reverse SDE (removes noise)

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\mathbb{w}_t$$

We replace the true score $\nabla_x \log p_t(x)$ with our learned $s_\theta(x, t)$ and **simulate backward** from pure Gaussian noise to produce realistic data.

---

5. Summary

Concept	**Description**
Score function	Gradient of log density: (\nabla_x \log p(x))
Score matching	Learn (s_\theta(x)) directly, without computing (p(x))
Denoising score matching	Learn scores of noisy data distributions
Diffusion models	Use learned scores to reverse a noise diffusion process and generate data