HW7

112652011 廖晨鈞

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

The concept of score matching

The goal of a generative model is to learn a data distribution $p(x)$. However, directly modeling $p(x)$ is often intractable due to a difficult-to-compute normalization constant (e.g., $p(x) = \tilde{p}(x)/Z$).

Score matching avoids this by instead learning the **score function**, which is the gradient of the log-probability density with respect to the data x:

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

Intuition: The score function is a vector field that points in the direction of the steepest ascent in data density. It tells you how to modify a data point x to make it more likely under the distribution $p(x)$.

Key Advantage: The score function is independent of the normalization constant Z.

$$\nabla_x \log p(x) = \nabla_x (\log \tilde{p}(x) - \log Z) = \nabla_x \log \tilde{p}(x)$$

This makes it possible to learn without ever computing Z.

Denoising Score Matching (DSM)

We want to train a model $S(x; \theta)$ to approximate the true score $\nabla_x \log p(x)$. A naive loss function would be:

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

This is unusable because the true score $\nabla_x \log p(x)$ is unknown.

The Solution: Denoising Score Matching (DSM)

Instead of matching the score of clean data, we intentionally add known noise to the data and train the model to learn the score of the *noisy* data distribution.

1. **Perturb Data:** Take a clean data point x_0 and add Gaussian noise to get a noisy sample x:

$$x = x_0 + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

The conditional distribution $p(x|x_0)$ is a Gaussian centered at x_0.

2. **Calculate the Known Score:** The score of this conditional distribution is easy to compute:

$$\nabla_x \log p(x|x_0) = \nabla_x \log \left(\frac{1}{\sqrt{2\pi\sigma^2}^d} \exp\left(-\frac{\|x - x_0\|^2}{2\sigma^2} \right) \right) = -\frac{x - x_0}{\sigma^2} = -\frac{\epsilon}{\sigma^2}$$

3. **The DSM Loss:** It can be proven that minimizing the intractable score matching loss on the noisy data is equivalent to minimizing the following simple, tractable objective:

$$L_{DSM}(\theta) = \mathbb{E}_{x_0 \sim p(x_0), \epsilon \sim \mathcal{N}(0, \sigma^2 I)} \|S(x_0 + \epsilon; \theta) - \nabla_x \log p(x|x_0)\|^2$$

Substituting the known score, we get the final training objective:

$$L_{DSM}(\theta) = \mathbb{E}_{x_0, \epsilon} \left\| S(x_0 + \epsilon; \theta) + \frac{\epsilon}{\sigma^2} \right\|^2$$

Intuition: The model $S(x;\theta)$ is trained to take a noisy sample x and predict the noise ϵ that was added to it (up to a scaling factor). This transforms a complex density estimation problem into a simple noise prediction (denoising) task.

Usage in Score-Based (Diffusion) Models

Score-based models use this principle to generate new data from noise.

1. **Forward Process (Diffusion):** A sequence of increasing Gaussian noise is gradually added to the clean data over many timesteps $t = 0, \ldots, T$. This creates a series of noisy distributions $p_t(x)$, starting from the data distribution $p_0(x)$ and ending at a simple prior distribution (e.g., pure Gaussian noise) $p_T(x)$.
2. **Training:** A single time-dependent score model $S(x_t, t; \theta)$ is trained using the DSM loss to estimate the score $\nabla_{x_t} \log p_t(x_t)$ for all noise levels $t \in [0, T]$.
3. **Reverse Process (Generation):** To generate a new sample, we start with a sample from the prior distribution, $x_T \sim \mathcal{N}(0, I)$, and reverse the diffusion process. This is achieved by iteratively using the trained score model to guide the sample towards regions of higher data density.