Machine Learning Assignment 07

313652008 黃睿帆

October 17, 2025

Score Matching and Its Role in Score-Based (Diffusion) Generative Models

Diffusion model we talked in the class this Wednesday is an generative model, where a **generative model** aims to learn a probability distribution $p_{\text{data}}(x)$ for a given data $\{x\}$, such that we can later sample new data points that look realistic.

Ideally, we want to learn a parametric model $p(x;\theta)$ such that

$$p(x;\theta) \approx p_{\text{data}}(x),$$

and we could train it by **maximum likelihood estimation (MLE)**:

$$\max_{\theta} \mathbb{E}_{x\sim p_{\text{data}}}[\log p(x;\theta)].$$

However, in many models, $p(x;\theta)$ is *intractable* because it contains a **partition function** (Ausatz):

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp(q(x;\theta)),$$

where $Z(\theta) = \int \exp(q(x;\theta))dx$ is extremely hard to compute or differentiate.

For example, for MNIST data set (graph has size $28 \times 28 = 784$), that is, for the given data $\{x\}$ (graph), $x \in \mathbb{R}^{784}$, to find probability density function $p(x) : \mathbb{R}^{784} \to \mathbb{R}^1$ such that both $p(x) \geq 0$ and $\int_{\mathbb{R}^{784}} p(x)dx = 1$ are our difficulties. Thus, we seek a quantity that avoids explicit normalization.

Learning the Score Function

The **score function** is defined as

$$S(x;\theta) = \nabla_x \log p(x;\theta).$$

Notice that

$$\log p(x;\theta) = q(x;\theta) - \log Z(\theta),$$

and since $\log Z(\theta)$ depends only on θ, not on x,

$$\nabla_x \log p(x;\theta) = \nabla_x q(x;\theta).$$

Hence, we can learn the gradient of the log-density without needing to compute the normalizing constant.

1. Explicit Score Matching (ESM)

If we somehow knew the true score $\nabla_x \log p_{\text{data}}(x) = \nabla_x \log p_(x)$, the ideal training objective would be

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

However, $\nabla_x \log p(x)$ is unknown because $p(x)$ is not available explicitly.

2. Implicit Score Matching (ISM)

To obtain a computable loss, we manipulate L_{ESM} algebraically. Start from

$$\begin{aligned} L_{\text{ESM}} &= \mathbb{E}_{p(x)}[\|S(x) - \nabla_x \log p(x)\|^2] \\ &= \mathbb{E}_{p(x)}[\|S(x)\|^2] - 2\mathbb{E}_{p(x)}[S(x) \cdot \nabla_x \log p(x)] + \mathbb{E}_{p(x)}[\|\nabla_x \log p(x)\|^2]. \end{aligned}$$

First we see the middle term:

$$\begin{aligned} \mathbb{E}_{p(x)}[S(x) \cdot \nabla_x \log p(x)] &= \int S(x) \cdot \nabla_x \log p(x)\, p(x)\, dx \\ &= \int S(x) \cdot \nabla_x p(x)\, dx. \end{aligned}$$

Using **integration by parts** (assuming boundary terms vanish, like we did in the class):

$$\int S(x) \cdot \nabla_x p(x)\, dx = -\int (\nabla_x \cdot S(x)) p(x)\, dx.$$

Substituting this result into original equation gives:

$$L_{\text{ESM}} = \mathbb{E}_{p(x)}[\|S(x)\|^2] + 2\mathbb{E}_{p(x)}[\nabla_x \cdot S(x)] + \mathbb{E}_{p(x)}[\|\nabla_x \log p(x)\|^2].$$

The final term does not depend on θ, so it can be omitted for optimization purposes. Thus, the **Implicit Score Matching (ISM)** loss is

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

Hence, minimizing L_{ESM} and L_{ISM} are equivalent. In the case of $S(x) = \nabla_x \log p(x)$, we have $L_{\text{ESM}} = 0$, and $L_{\text{ISM}} \leq 0$. Or we say the optimal $L_{\text{ISM}} \leq 0$.

Note: The score function $S(x) = \nabla_x \log p(x)$ points toward regions of higher density. Score matching aligns the model’s score field $S(x;\theta)$ with that of the data distribution $\nabla_x \log p_(x)$. If they coincide everywhere, the model and data distributions have identical density contours.

Motivation for DSE: Instability for High-Dimensional Data

For complex or high-dimensional data (e.g. images), estimating $\nabla_x \log p(x)$ directly is unstable. To overcome this, we instead consider a *noisy version* of the data and learn the score of this smoothed (noisy) distribution —this leads to **Denoising Score Matching (DSM)**. (Idea: If probability density function (pdf) has a little difference, then same is score function (Depend on **noise**)).

Denoising Score Matching (DSM)–Setup and Notation

Let:

- x_0: original (clean) data sample;
- $p_0(x_0)$: data distribution;
- x: noisy version of x_0;
- $p(x|x_0)$: conditional (noise) distribution;
- $p_\sigma(x) = \int_{\mathbb{R}^d} p(x|x_0)p_0(x_0)\, dx_0$: marginal noisy data distribution.

Our goal is to learn the **noisy score function**:

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

3. Denoise Score Matching (DSM)

The **denoising score matching** loss is defined as:

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

This form is practical because $\nabla_x \log p(x|x_0)$ is known analytically for many noise models (e.g., Gaussian).

3.1 Derivation of DSM from ESM

We begin from the expectation under the noisy distribution $p_\sigma(x)$:

$$\begin{aligned}
\mathbb{E}_{x\sim p_\sigma(x)} \langle S_\sigma(x), \nabla_x \log p_\sigma(x) \rangle &= \int_{\mathbb{R}^d} S_\sigma(x) \cdot \nabla_x p_\sigma(x)\, dx \\
&= \int_{\mathbb{R}^d} S_\sigma(x) \cdot \nabla_x \left[\int_{\mathbb{R}^d} p(x|x_0) p_0(x_0)\, dx_0 \right] dx \\
&= \int_{\mathbb{R}^d} p_0(x_0) \left[\int_{\mathbb{R}^d} S_\sigma(x) \cdot (\nabla_x \log p(x|x_0))\, p(x|x_0)\, dx \right] dx_0 \\
&= \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p(x|x_0)} \langle S_\sigma(x), \nabla_x \log p(x|x_0) \rangle.
\end{aligned}$$

Similarly,

$$\mathbb{E}_{x\sim p_\sigma(x)} \|S_\sigma(x)\|^2 = \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p(x|x_0)} \|S_\sigma(x)\|^2.$$

Now consider the explicit score matching objective for the noisy score:

$$\begin{aligned}
&\mathbb{E}_{x\sim p_\sigma(x)} \left[\|S_\sigma(x;\theta) - \nabla_x \log p_\sigma(x)\|^2 \right] \\
=&\mathbb{E}_{x\sim p_\sigma(x)} \left[\|S_\sigma(x)\|^2 - 2 S_\sigma(x) \cdot \nabla_x \log p_\sigma(x) + \|\nabla_x \log p_\sigma(x)\|^2 \right].
\end{aligned}$$

Substituting the previous identities gives:

$$\begin{aligned}
&= \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p(x|x_0)} \|S_\sigma(x)\|^2 - 2\mathbb{E}_{x_0, x|x_0} \langle S_\sigma(x), \nabla_x \log p(x|x_0) \rangle + \mathbb{E}_{x\sim p_\sigma(x)} \|\nabla_x \log p_\sigma(x)\|^2 \\
&= \mathbb{E}_{x_0, x|x_0} \left[\|S_\sigma(x) - \nabla_x \log p(x|x_0)\|^2 \right] + \mathbb{E}_{x\sim p_\sigma(x)} \|\nabla_x \log p_\sigma(x)\|^2 - \mathbb{E}_{x_0, x|x_0} \|\nabla_x \log p(x|x_0)\|^2.
\end{aligned}$$

The last two terms do not depend on θ, so they form a constant C. Hence,

$$\boxed{L_{\text{ESM}}(\theta) = \mathbb{E}_{x_0, x|x_0} \left[\|S_\sigma(x;\theta) - \nabla_x \log p(x|x_0)\|^2 \right] + C.}$$

Therefore, minimizing L_{DSM} is equivalent to minimizing the noisy ESM (and ISM) objectives, up to an additive constant.

3.2 DSM with Gaussian Noise

The denoising score matching (DSM) loss aims to learn the noisy score function $S_\sigma(x;\theta) = \nabla_x \log p_\sigma(x)$ by minimizing

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

In practice, the conditional distribution $p(x|x_0)$ is chosen to be a Gaussian perturbation:

$$\begin{aligned}
x &= x_0 + \epsilon_\sigma, \quad \epsilon_\sigma \sim \mathcal{N}(0, \sigma^2 I), \\
&= x_0 + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).
\end{aligned}$$

Thus,

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

and its gradient with respect to x is

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

Substituting this into the DSM objective gives:

$$
\begin{aligned}
L_{DSM}(\theta) &= \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p_\sigma(x|x_0)} \left\| S_\sigma(x;\theta) + \frac{x - x_0}{\sigma^2} \right\|^2 \\
&= \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{x|x_0 \sim p_\sigma(x|x_0)} \frac{1}{\sigma^4} \left\| \left(\sigma^2 S_\sigma(x;\theta) + x \right) - x_0 \right\|^2 \\
&= \mathbb{E}_{x_0 \sim p_0(x_0)} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \frac{1}{\sigma^2} \left\| \sigma S_\sigma(x_0 + \sigma\epsilon;\theta) + \epsilon \right\|^2 .
\end{aligned}
$$

This final form is widely used in score-based and diffusion generative models. It shows that training the score network $S_\sigma(x;\theta)$ is equivalent to predicting the negative of the added noise $-\epsilon$, which corresponds to denoising the perturbed sample $x = x_0 + \sigma\epsilon$.

3.3 DSM and Diffusion Models

Score-based (or diffusion) generative models train a network $S_\theta(x,t)$ to estimate the score of a progressively noised data distribution $p_t(x)$. Once trained, samples can be generated by simulating the **reverse diffusion process**, guided by the learned score field $S_\theta(x,t)$, effectively denoising pure noise step-by-step back into realistic data.

To do comparison, **Score Matching** learns gradients of log densities rather than normalized densities. **Denoising Score Matching (DSM)** learns the score of a smoothed (noisy) version of the data. DSM connects directly to diffusion models: learning scores for multiple noise levels yields the foundation of the **score-based generative framework**. We also use Gaussian DSM reduces to a simple loss involving the known score of the Gaussian conditional $p(x|x_0)$.

UNANSWERED QUESTIONS Week 07

- How does the choice of noise scales (the -schedule) and the per-scale weighting in the DSM loss affect consistency and sampling quality?
- What happens to DSM-trained score estimators in low-density regions and near the data manifold —can the estimated score blow up or be poorly behaved?