Machine Learning — Homework 7

Student ID: 111652015

Score Let $p : \mathbb{R}^d \to [0, \infty)$ be a continuous probability density function. The *score* (*score function*) of p at $x \in \mathbb{R}^d$ is the gradient of its log-density (whenever $p(x) > 0$):

$$\mathbf{s}_p(x) \coloneqq \nabla_x \log p(x).$$

1 Score Matching

Score Matching - ESM We use a parametric model $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to directly approximate the true score

$$s_p(x) \;=\; \nabla_x \log p(x).$$

The ideal learning objective minimizes the mean–squared error between the model score and the true score:

$$\mathcal{L}_{\text{ESM}}(\theta) \;=\; \frac{1}{2}\,\mathbb{E}_{x\sim p}\Big[\big\|\, S(x;\theta) - \nabla_x \log p(x)\,\big\|^2\Big],$$

where the expectation is taken over the data distribution p_{data}, and $\|\cdot\|_2$ denotes the Euclidean norm.

Implicit Score Matching Since $\nabla_x \log p_0(x)$ is not available. Hyvärinen proposed implicit score matching (ISM), using integration by parts to rewrite the cross term as

$$\mathbb{E}_p\big[S(x;\theta)\cdot\nabla\log p\big] = \int S(x;\theta)\cdot\nabla p\,dx = \int p\,S(x;\theta)\cdot n\,dS \;-\; \mathbb{E}_p\big[\nabla\cdot S(x;\theta)\big].$$

when the boundary term can be neglected (e.g., the density decays at infinity, or the support is bounded with appropriate boundary conditions), we have

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

which is *equivalent* to the explicit score-matching loss $\mathcal{L}_{\text{ESM}}(\theta)$.

DSM Add Gaussian noise to a sample $x_0 \sim p_0$:

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

Define the noisy score and noisy data distribution:

$$S_\sigma(x;\theta) = \nabla_x \log p_\sigma(x), \qquad p_\sigma(x) = \int_{\mathbb{R}^d} p(x \mid x_0)\,p_0(x_0)\,dx_0.$$

Since

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

then

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

The DSM can be written as

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

$$= \mathbb{E}_{x_0 \sim p_0(x_0)}\ \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\frac{1}{\sigma^2} \left\|\, \sigma\, S_\sigma(x_0 + \sigma\epsilon;\theta) + \epsilon \,\right\|^2 \right].$$

2 Score-based generative models

Forward Noising (DDPM) Let $\{\alpha_t\}_{t=1}^T \subset (0,1]$ and $\bar{\alpha}_t \coloneqq \prod_{i=1}^t \alpha_i$. For $t = 1, \dots, T$, the forward process is

$$x_t \;=\; \sqrt{\bar{\alpha}_t}\, x_0 \;+\; \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \epsilon \sim (0, I).$$

Let $p_t(x)$ denote the noised data distribution at time t.

Noise-Conditional Score Learning (DSM) We train a noise–conditioned score network $S(x,t;\theta)$ to approximate the true score

$$S(x,t;\theta) \;\approx\; \nabla_x \log p_t(x).$$

Under Gaussian corruption in the DDPM setting,

$$\nabla_{x_t} \log p(x_t \,|\, x_0) \;=\; -\frac{x_t - \sqrt{\bar{\alpha}_t}\, x_0}{1 - \bar{\alpha}_t} \;=\; -\frac{\epsilon}{\sqrt{1-\bar{\alpha}_t}}.$$

A common DDPM-equivalent training objective predicts the noise directly:

$$\mathcal{L}_{\text{DDPM}}(\theta) \;=\; {}_{t,\,x_0,\,\epsilon}\!\left[\,\left\|\, \epsilon - \epsilon_\theta(x_t,t) \,\right\|^2 \,\right].$$

This is proportional to DSM and implies the correspondence (up to parameterization-dependent constants)

$$S(x_t,t;\theta) \;\approx\; -\frac{\epsilon_\theta(x_t,t)}{\sqrt{1-\bar{\alpha}_t}}.$$

Reverse SDE Given $S(x,t;\theta) \approx \nabla_x \log p_t(x)$, the reverse-time SDE (Anderson time reversal) is

$$\mathrm{d}x \;=\; \Big[f(x,t) \;-\; g(t)^2\, S(x,t;\theta) \Big] \mathrm{d}t \;+\; g(t)\, \mathrm{d}\bar{W}_t,$$

which is integrated from large noise to small noise with an appropriate schedule.

Probability Flow ODE The corresponding deterministic ODE is

$$\frac{\mathrm{d}x}{\mathrm{d}t} \;=\; f(x,t) \;-\; \tfrac{1}{2}\,g(t)^2\,S(x,t;\theta),$$

which enables tractable log-likelihood via integrating the divergence along the trajectory.

Annealed Langevin Dynamics (ALD) With a noise ladder $\sigma_L > \cdots > \sigma_0$, perform per-level Langevin steps

$$x \;\leftarrow\; x \;+\; \eta\,S(x,\sigma;\theta) \;+\; \sqrt{2\eta}\,z, \qquad z \sim (0, I),$$

then decrease σ and repeat. In practice, $\eta \propto \sigma^2$ and only a few steps per level are needed.