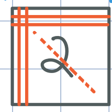
Generative model

Goal ① → given data to find $p(x)$ ② To sample from $p(x)$

Difficulty: ① $p(x)>0$ ② $\int_{\mathbb{R}^d} p(x)\,dx = 1$

Ansatz: $p(x;\theta) = \dfrac{e^{g(x;\theta)}}{Z(\theta)} \rightarrow \int p\,dx = 1 \Rightarrow \int \dfrac{e^{g(x)}}{Z}\,dx = \dfrac{1}{Z}\int e^{g(x)}\,dx$

※ MNIST Data



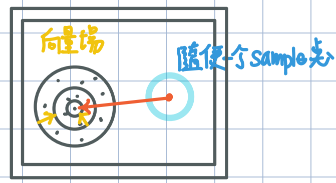
$28\times 28 = 784$

$X \in \mathbb{R}^{784}$, to find $p(x): \mathbb{R}^{784} \rightarrow \mathbb{R}^1$

$S(x): \mathbb{R}^{784} \rightarrow \mathbb{R}^{784}$

▷ Score function: $\boxed{S(x)} = \nabla \log p(x)$ ($\mathbb{R}^d \rightarrow \mathbb{R}^d$) (where pdf: $p(x): \mathbb{R}^d \rightarrow \mathbb{R}^1$)

Sampling: Gradient $x_{n+1} = x_n + \eta S(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$



▷ Score-matching

→ Explicit score matching (ESM)

$\mathcal{L}_{ESM}(\theta) = E_{x\sim p(x)} \| S(x;\theta) - \nabla_x \log p(x) \|^2$ ✲ Suppose $p(x)$ or $\nabla_x \log p(x)$ is known

↑ 照著pdf去選点

→ Implicit score matching (ISM)

$\mathcal{L}_{ISM}(\theta) = E_{x\sim p(x)} \left[\| S(x;\theta) \|^2 + 2\nabla_x \cdot S(x;\theta) \right]$ ✲ $\theta^* = \arg\min_\theta \mathcal{L}_{ESM} = \arg\min_\theta \mathcal{L}_{ISM}$

<example>

$p(x,y): \mathbb{R}^2 \rightarrow \mathbb{R}^1$, $S(x) = \nabla \log p = \begin{bmatrix} \partial_x \log p \\ \partial_y \log p \end{bmatrix} = \begin{bmatrix} \frac{1}{p} p_x \\ \frac{1}{p} p_y \end{bmatrix} = \frac{1}{p} \begin{bmatrix} p_x \\ p_y \end{bmatrix} = \frac{1}{p} \nabla p$

<pf. ✲>

$\mathcal{L}_{ESM} = E_{x\sim p(x)} \left[\| S(x;\theta) \|^2 - 2S(x;\theta)\cdot \nabla_x \log p + \| \nabla_x \log p \|^2 \right]$ (for $\|\vec{u}-\vec{v}\|^2 = \|\vec{u}\|^2 - 2\langle \vec{u}, \vec{v}\rangle + \|\vec{v}\|^2$)

$E_{x\sim p(x)} [S(x;\theta)\cdot \nabla_x \log p(x)] = \int (S(x)\cdot \nabla_x \log p)\, p(x)\,dx = \int (S(x)\cdot \frac{1}{\cancel{p}} \nabla_x p)\, \cancel{p(x)}\,dx$

Integration by part.

$= \int -(\nabla_x \cdot S_x)\, p(x)\,dx + 0 = E_{x\sim p(x)} - (\nabla \cdot S(x))$

Then, $\mathcal{L}_{ESM} = E_{x\sim p(x)} \left[\| S(x;\theta) \|^2 + 2\nabla_x \cdot S(x;\theta) \right] + E_{x\sim p(x)} \| \nabla_x \log p(x) \|^2$ independent of θ

$= \boxed{\mathcal{L}_{ISM}} + \underbrace{E_{x\sim p(x)} \| \nabla_x \log p(x) \|^2}_{>0}$

Therefore, $\arg\min_\theta \mathcal{L}_{ESM} = \arg\min_\theta \mathcal{L}_{ISM}$ ▨

Remark: The optimal $\mathcal{L}_{ISM} < 0$

▷ Denoising score matching (x_0: original data ; $p_0(x_0)$: original p.d.f)

x: perturbed data ($x = x_0 + \varepsilon_\sigma$) , $p(x)$: pdf of perturbed data

$\mathcal{L}_{DSM} = E_{x_0\sim p_0(x_0)} E_{(x|x_0)\sim p(x|x_0)} \| S_\sigma(x;\theta) - \nabla_x \log p(x|x_0) \|^2$

→ We add Gaussian noise, $x = x_0 + \varepsilon_\sigma$, $\varepsilon_\sigma \sim (0, \sigma^2 I)$ ($x = x_0 + \sigma\varepsilon$, $\varepsilon \sim \mathcal{N}(0, I) \Rightarrow x - x_0 = \sigma\varepsilon$)

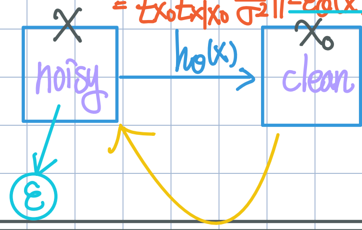
$p(x|x_0) = \dfrac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{\|x-x_0\|^2}{2\sigma^2}} \longrightarrow \nabla_x \log p(x|x_0) = \nabla_x \left(C - \dfrac{\|x-x_0\|^2}{2\sigma^2} \right) = \dfrac{-1}{\sigma^2}(x - x_0)$

✲ $\mathcal{L}_{DSM} = E_{x_0\sim p_0} E_{(x|x_0)\sim p(x|x_0)} \| S_\sigma(x) + \frac{1}{\sigma^2}(x - x_0) \|^2$

$= E_{x_0\sim p_0} E_{(x|x_0)\sim p(x|x_0)} \frac{1}{\sigma^4} \| \sigma^2 S_\sigma(x;\theta) + x - x_0 \|^2 \longrightarrow = E_{x_0} E_{x|x_0} \cdot \frac{1}{\sigma^2} \| \sigma S_\sigma(x) + \varepsilon \|^2$

$\updownarrow$ 和score有关.

$= E_{x_0} E_{x|x_0} \frac{1}{\sigma^2} \| -\epsilon_\theta(x) + \epsilon \|^2$



Denoising autoencoder : $\mathcal{L}oss = E_{x_0} E_{x|x_0} \| h_\theta(x) - x_0 \|^2$

→ <pf.>

By $\mathcal{L}_{MSE}$,

$p(x) = \int p(x|x_0) p(x_0) \, dx_0$, then

①

$$E_{x\sim p(x)} [\| S_\sigma(x;\theta) \|^2] = \int \|S_\sigma\|^2 p(x) \, dx = \int \|S_\sigma(x)\|^2 \left[\int p(x|x_0) p(x_0) \, dx_0 \right] dx$$

$$= \int \underline{\left[\int \|S_\sigma(x)\|^2 p(x|x_0) \, dx \right]} \boxed{p(x_0) \, dx_0}$$

$$= E_{x_0 \sim p(x_0)} E_{x|x_0 \sim p(x|x_0)} \|S_\sigma(x;\theta)\|^2$$

②

$$E_{x\sim p(x)} \langle S_\sigma(x), \nabla_x \log p(x) \rangle = \int \langle S_\sigma, \frac{1}{p} \nabla_x p \rangle p \, dx = \int S_\sigma(x) \cdot \nabla_x \left(\int p(x|x_0) p(x_0) \, dx_0 \right) dx$$

$$= \iint S_\sigma(x) \cdot \nabla_x p(x|x_0) p(x_0) \, dx_0 \, dx = E_{x_0 \sim p(x_0)} \left[\int S_\sigma(x) \cdot \nabla_x p(x|x_0) \, dx \right]$$

$$= E_{x_0} \int S_\sigma(x) \cdot \nabla_x \log p(x|x_0) p(x|x_0) \, dx$$

$$= E_{x_0 \sim p(x_0)} E_{x|x_0 \sim p(x|x_0)} \langle S_\sigma(x;\theta), \nabla_x \log p(x|x_0) \rangle$$

$$\mathcal{L}_{ESM} = E_{x\sim p(x)} \| S_\sigma(x;\theta) - \nabla_x \log p(x) \|^2 = E_{x_0 \sim p_0(x_0)} E_{x|x_0 \sim p(x|x_0)} \| S_\sigma(x;\theta) \|^2 - 2 \langle S_\sigma(x;\theta), \nabla_x \log p(x|x_0) \rangle + E_{x\sim p(x)} \| \nabla_x \log p(x) \|^2$$

$$= E_{x_0 \sim p_0(x_0)} E_{x|x_0 \sim p(x|x_0)} \| S_\sigma(x;\theta) - \nabla_x \log p(x|x_0) \|^2 \underbrace{- E E_x \| \nabla_x \log p(x|x_0) \|^2 + E_x \| \nabla_x \log p \|^2}_{\text{Independent of } S}$$

$$\nabla \cdot S(x) = \begin{bmatrix} \partial x_1 \\ \partial x_2 \\ \vdots \\ \partial x_d \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_d \end{bmatrix} = \partial_{x_1} S_1 + \partial_{x_2} S_2 + \cdots + \partial_{x_d} S_d$$ ← 超難算

▷ Sliced score matching (SSM, 2019)

First , $\nabla \cdot S(x) = \begin{bmatrix} \partial x_1 \\ \partial x_2 \\ \vdots \\ \partial x_d \end{bmatrix} \cdot [S_1 \; S_2 \; \cdots \; S_d] = \begin{bmatrix} \partial_{x_1} S_1 & \partial_{x_1} S_2 & \cdots & \partial_{x_1} S_d \\ \partial_{x_2} S_1 & \partial_{x_2} S_2 & & \\ \vdots & & \ddots & \\ \partial_{x_d} S_1 & & & \partial_{x_d} S_d \end{bmatrix} = \underbrace{\begin{bmatrix} \partial_{x_1}^2 \log p & \partial_{x_2}\partial_{x_1} \log p & \\ \vdots & \ddots & \\ \partial_{x_d}\partial_{x_1} \log p & & \end{bmatrix}}_{\log p \text{ 的 Hessian matrix.}}$

then, we know that $\nabla \cdot S = tr(\nabla S)$

$tr(\nabla S) = E_{v\sim p(v)} (v^T (\nabla S) v) = E_{v\sim p(v)} (v^T \nabla (v^T S))$

Thus, our $E_{x\sim p(x)} = \{ \| S(x;\theta) \|^2 + 2 \nabla \cdot S(x;\theta) \}$

▷ Hutchinson's trace estimator :

Let $v \in \mathbb{R}^d$ is a random vector , s.t. $E[vv^T] = I$, then $tr(A) = E_{v\sim p(v)} (v^T A v)$

Hence, our sliced score matching loss: $E_{x\sim p(x)} \| S(x;\theta) \|^2 + E_{x\sim p(x)} E_{v\sim p(v)} (v^T \nabla (v^T S))$

Remark : $\vec{v} \in \mathcal{N}(0, I) \Rightarrow E[vv^T] = I$

→ <pf.> $E_{v\sim p(v)} (v^T A v) = E_{v\sim p(v)} (tr(v^T A v)) = E_{v\sim p(v)} (tr(vv^T A)) = tr((E_{v\sim p(v)} vv^T) A) = tr(A).$

2. Unanswered Questions

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

Q: If in 3D image, how to use Denoising Score matching?