Assignment 7

314652021羅瑋翔

Written Assignment

Q1. Explain the concept of score matching and describe how it is used in score-based (diffusion) generative models.

Concept of score matching

Motivation:
Learning the probability density function of generative model is difficult.

$$p_\theta(x) = \frac{e^{f_\theta(x)}}{Z_\theta}, Z_\theta = \int e^{f_\theta(x)} dx$$

where $f_\theta(x)$ denotes the probability density function while Z_θ denotes the normalizing constant such that $\int \frac{e^{f_\theta(x)}}{Z_\theta} dx = 1$.
Since the partition function is intractable, computing Z_θ is very hard. We use an alternative way to compute it, scoring matching.
Idea:
Since it is difficult to compute $p_\theta(x)$ directly.
Here, we compute $\nabla_x log\, p_\theta(x)$ instead.

$$log\, p_\theta(x) = \nabla_x f_\theta(x) - \nabla_x log Z_\theta = \nabla_x f_\theta(x)$$

it depicts the direction in which the PDF increases fasted near x.

Hence, we define $s_\theta(x)$ to be the score function and :

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

Optimize:
Then, we use **Fisher divergence** to compare two vertor fields of scores to be the loss function of **Explicit score matching** as follows:

$$L_{ESM}(\theta) = \mathbb{E}_{p_{data}}(x)[||s_\theta(x) - \nabla_x log\, p_\theta(x)||_2^2]$$

However, $\nabla_x log(p(x))$ is unknown, we cannot use it.
Hence, we use the methods **Integration by parts(Gauss' theorem)** to obtained another $\nabla_x log(p(x))$-free loss function of **Implicit scoring maching** as follows:

$$L_{ISM}(\theta) = \mathbb{E}_{p_{data}}(x)[||s_\theta(x)||_2^2 + 2trace(\nabla_x s_\theta(x))]$$

The derivation is as follows:

$$\begin{aligned}
&\mathbb{E}_{p_{data}}(x)\left[\|s_\theta(x) - \nabla_x \log p(x)\|^2\right] \\
&= \mathbb{E}_{p_{data}}(x)\left[\|s_\theta(x)\|^2 - 2s_\theta(x)\cdot\nabla_x \log p(x) + \|\nabla_x \log p(x)\|^2\right] \\
&= \mathbb{E}_{p_{data}}(x)\left[\|s_\theta(x)\|^2\right] - 2\int_{R^d}(s_\theta(x)\cdot\nabla_x \log p(x))p(x)dx + \mathbb{E}_{p_{data}}(x)\left[\|\nabla_x \log p(x)\|^2\right] \\
&= \mathbb{E}_{p_{data}}(x)\left[\|s_\theta(x)\|^2\right] - 2\int_{R^d} s_\theta(x)\cdot\nabla_x p(x)dx + \mathbb{E}_{p_{data}}(x)\left[\|\nabla_x \log p(x)\|^2\right] \\
&= \mathbb{E}_{p_{data}}(x)\left[\|s_\theta(x)\|^2\right] + 2\int_{R^d}(\nabla_x\cdot s_\theta(x))p(x)dx + \mathbb{E}_{p_{data}}(x)\left[\|\nabla_x \log p(x)\|^2\right] \\
&= \mathbb{E}_{p_{data}}(x)\left[\left(\|s_\theta(x)\|^2 + 2\nabla_x\cdot s_\theta(x)\right)\right] + \mathbb{E}_{p_{data}}(x)\left[\|\nabla_x \log p(x)\|^2\right].
\end{aligned}$$

Since the second term, $\mathbb{E}_{p_{data}}(x)\left[\|\nabla_x \log p(x)\|^2\right]$, of the last equation is constant, it does not affect the result of optimization, we can only focus on the first term of it.
With millions of dimensions, computing $trace(\nabla_x s_\theta(x))$ is not scalable.
Then we came up with an alternative way to do this.

How it is used in diffusion Generative model

Notation

- x: original data,
- $\tilde{x}$: noisy data(by perturbing the originaldata),
- $p(x)$: data distribution of original data,
- $\tilde{p}(\tilde{x})$: noisy data distribution,
- $p(\tilde{x}|x)$: coditional noisy data distrbution

In, diffusion generative model, score matching plays an important role. It is used to find the noisy score function $\tilde{s}_\theta(\tilde{x}) = \nabla_x log\,\tilde{p}(x)$.

Loss function

$$\begin{aligned} &L_{DSM}(\theta) \\ &= \mathbb{E}_{\tilde{p}(\tilde{x})}[||\nabla_{\tilde{x}} log\, \tilde{p}(\tilde{x}) - \tilde{s}_\theta(\tilde{x})||_2^2] \\ &= \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}[||\nabla_{\tilde{x}} log\, \tilde{p}(\tilde{x}|x) - \tilde{s}_\theta(\tilde{x})||_2^2] \end{aligned}$$

The derivation is as follow:

$$\begin{aligned} &\mathbb{E}_{\tilde{p}(\tilde{x})}\langle \tilde{s}_\theta(\tilde{x}), \nabla_x \log \tilde{p}(\tilde{x})\rangle \\ &= \int_{R^d} (\tilde{s}_\theta(\tilde{x}) \cdot \nabla_x \log \tilde{p}(\tilde{x}))\tilde{p}(\tilde{x})d\tilde{x} = \int_{R^d} \tilde{s}_\theta(\tilde{x}) \cdot \nabla_x \tilde{p}(\tilde{x})\, d\tilde{x} \\ &= \int_{R^d} \tilde{s}_\theta(\tilde{x}) \cdot \nabla_x \left[\int_{R^d} p(\tilde{x}|x)p(x)dx\right] d\tilde{x} \\ &= \int_{R^d} \tilde{s}_\theta(\tilde{x}) \cdot \left[\int_{R^d} \nabla_x p(\tilde{x}|x)p(x)dx\right] d\tilde{x} \\ &= \int_{R^d} p(x) \left[\int_{R^d} \tilde{s}_\theta(\tilde{x}) \cdot (\nabla_x \log p(\tilde{x}|x))p(\tilde{x}|x)d\tilde{x}\right] dx \\ &= \int_{R^d} p(x) \left[\mathbb{E}_{p(\tilde{x}|x)}\langle \tilde{s}_\theta(\tilde{x}), \nabla_x \log p(\tilde{x}|x)\rangle\right] dx \\ &= \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\langle \tilde{s}_\theta(\tilde{x}), \nabla_x \log p(\tilde{x}|x)\rangle. \end{aligned}$$

Simlilarly,

$$\mathbb{E}_{\tilde{p}(\tilde{x})}[||\tilde{s}_\theta(x)||^2] = \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}[||\tilde{s}_\theta(x)||^2].$$

And, we consider ESM objective to for the noisy score function to have:

$$\begin{aligned} &\mathbb{E}_{\tilde{p}(\tilde{x})}\left(||\tilde{s}_\theta(\tilde{x}) - \nabla_x \log\ \tilde{p}(\tilde{x})||^2\right) \\ =&\mathbb{E}_{\tilde{p}(\tilde{x})}\left(||\tilde{s}_\theta(\tilde{x})||^2 - 2\tilde{s}_\theta(\tilde{x}) \cdot \nabla_x \log\ \tilde{p}(\tilde{x}) + ||\nabla_x \log \tilde{p}(\tilde{x})||^2\right) \\ =&\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}||\tilde{s}_\theta(\tilde{x})||^2 + (-2)\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\langle \tilde{s}_\theta(\tilde{x}), \nabla_x \log\ p(\tilde{x}|x)\rangle \\ &+ \mathbb{E}_{\tilde{p}(\tilde{x})}\left[||\nabla_x \log\ \tilde{p}(\tilde{x})||^2\right] \\ =&\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\left[||\tilde{s}_\theta(\tilde{x}) - \nabla_x \log\ p(\tilde{x}|x)||^2\right] \\ &+ \mathbb{E}_{\tilde{p}(\tilde{x})}\left[||\nabla_x \log \tilde{p}(\tilde{x})||^2\right] - \mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\left[||\nabla_x \log\ p(\tilde{x}|x)||^2\right] \\ =&\mathbb{E}_{p(x)}\mathbb{E}_{p(\tilde{x}|x)}\left[||\tilde{s}_\theta(\tilde{x}) - \nabla_x \log\ p(\tilde{x}|x)||^2\right] + C, \end{aligned}$$

where C is a constant, independent of θ.

Overall, DSM, ESM and ISM are equivalent in finding the noisy score function.

Q2. Unanswered questions

1.

In diffusion generative model, we need to add Gaussian noise to the original data points to learn it.
Since we all know choose a very small ϵ seems good, but the variance of this objective will explode.
How do I choose an appropriate noise ϵ to learn the generative model well?

2.

In sliced score matching, we used the idea projection to simplify $\nabla_x s_\theta(x)$, since it is not scalable while $\nabla_x(v^T s_\theta(x))$ is scalable.
If we choose less projections per data points, it might loss the characristic while choose more projections per data points would increase computation.
Is there any relations between the number of chosed projections and accuracy?