Week 7 assignment

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

（一）、 score matching 的概念

在 generative model 中，其中一個目標是想找到數據的分布情形，也就是去學習機率密度函數(PDF)，假設

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

其中$z(\theta)$是用來讓$\int_{-\infty}^{\infty} p(x)dx = 1$的常數，但在實際應用中$z(\theta)$可能因為維度高或是函數的複雜性導致非常難以計算，因此為了瞭解數據的分布狀況又避免去計算$z(\theta)$，就關注在$q(x;\theta)$函數上，先對$p(x;\theta)$取log會得到$\log p(x;\theta) = q(x;\theta) - \log z(\theta)$，因為$z(\theta)$只與$\theta$有關，所以對$x$做梯度計算後就能將$z(\theta)$項去掉，也得到所謂的 score function:

$$S(x) = \nabla_{\mathrm{x}} \log q(x;\theta)$$

（二）、 score matching 的原理

如果我們知道完整的機率密度函數$p(x)$的話，可以使用 Explicit score matching(ESM):

$$L_{ESM}(\theta) = \mathbb{E}_{x\sim p(x)}\|S(x;\theta) - \nabla_{\mathrm{x}} \log p(x)\|^2,$$

但在實際情況基本上不會知道$p(x)$，這時候就出現兩種解決辦法

1. Implicit Score Matching(ISM)搭配 Sliced Score Matching(SSM):

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

透過計算我們能知道

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

$$= \mathbb{E}_{x\sim p(x)}[\|S(x;\theta)\|^2 + 2\nabla_x \cdot S(x;\theta)] + \mathbb{E}_{x\sim p(x)}[\|\nabla_{\mathrm{x}} \log p(x)\|^2],$$

而$\mathbb{E}_{x\sim p(x)}[\|\nabla_{\mathrm{x}} \log p(x)\|^2]$這一項只與$p(x)$有關，與$\theta$無關，所以針對$L_{ESM}$與$L_{ISM}$的$\theta$進行最小化是等價的。

但另一個問題則是$L_{ISM}(\theta)$中的$\nabla_x \cdot S(x;\theta)$可以改寫成$tr\big(\nabla_x S(x;\theta)\big)$，但這一項也是非常難以計算的，而經過計算證明後會發現$tr\big(\nabla_x S(x;\theta)\big) = \mathbb{E}_{v\sim p(v)}\big[v^T \nabla_x \big(v^T S(x;\theta)\big)\big]$，所以$L_{ISM}(\theta)$就可以被改寫成

$$L_{SSM}(\theta)$$

$$= \mathbb{E}_{x\sim p(x)}[\|S(x;\theta)\|^2] + \mathbb{E}_{x\sim p(x)}\mathbb{E}_{v\sim p(v)}\big[2v^T \nabla_x \big(v^T S(x;\theta)\big)\big],$$

上述的$L_{SSM}(\theta)$在最小化的計算上就只需要通過 forward 和 back propagation 就能達成，因此在計算上就會方便很多，而因為 score function 可以知道數據分布的梯度值，所以透過最小化$L_{SSM}(\theta)$訓練出的$S(x)$，就可以用來引導一個數據跑到機率最大的預測值。

2. Denoising Score Matching (DSM):

假設我們有一個原始的數據分布$p_0(x_0)$，透過已知的雜訊去擾亂原始數據，得到帶噪的數據分布$p_\sigma(x|x_0)$，經過計算可以得到 L_{ESM} of noisy score function:

$$L_{ESM}(\theta)\ for\ noisy\ score\ function$$
$$= \mathbb{E}_{x\sim p_\sigma(x)}\|S_\sigma(x;\theta) - \nabla_x \log p_\sigma(x)\|^2$$
$$= \mathbb{E}_{x_0\sim p_0(x_0)}\mathbb{E}_{x|x_0\sim p(x|x_0)}[\|S_\sigma(x) - \nabla_x \log p(x|x_0)\|^2] + C$$
$$= L_{DSM}(\theta)$$

所以為了找到 noisy score function，最小化$L_{DSM}(\theta)$、$L_{ESM}(\theta)$與$L_{ISM}(\theta)$是相同的，而如果我們的加噪方式是利用 gaussian noise 來達成，那$\nabla_x p_\sigma(x|x_0) = -\frac{1}{\sigma^2}\epsilon_\sigma$，而$L_{DSM}(\theta)$就可以寫成

$$\mathbb{E}_{x\sim p(x)}\mathbb{E}_{v\sim p(v)}\left\|S_\sigma(x;\theta) + \frac{x - x_0}{\sigma^2}\right\|^2,$$

這樣的寫法就會讓$L_{DSM}(\theta)$的計算方便很多，只要透過 forward 和 back propagation 就能達成，所以就能輕鬆的訓練去噪模型，而之後給任何的噪音數據，都能夠通過去噪模型來預測出原始的乾淨數據。

二、Unanswered questions

score-based generative models 是透過 Score Function 來引導生成預測數據。但目前方法類似於梯度上升，每一步都採用預設的、保守的小步長，有沒有也找到或訓練出曲率資訊，來增加預測效率。