Homework 5

Deep Learning 2025 Spring

Due on 2025/4/7

1 True or False

Problem 1. By adding noise to the embedding of a sequence of words and conditionally resample the perturbed sequence to generate a new sequence, we can use diffusion model to generate text.

2 Q&A

Problem 2. (DDPM objective)

In the diffusion model, we train a model $\boldsymbol{\epsilon}_\theta$ that takes $\mathbf{x}_t$ and step t as input to be the parameterization of $\boldsymbol{\mu}_\theta$ to predict $\tilde{\boldsymbol{\mu}}_t$(the mean value of $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$ and $\mathbf{x}_0$), which is used in the reverse process where

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right) \tag{1}$$

The forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1})=\mathcal{N}(\mathbf{x}_t;\sqrt{\alpha_t}\mathbf{x}_{t-1},(1-\alpha_t)\mathbf{I}), q(\mathbf{x}_{1:T}|\mathbf{x}_0)=\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$

use the reparameterization trick, we have

$$\mathbf{x}_t=\sqrt{\alpha_t}\mathbf{x}_{t-1}+\sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}, \text{where } \boldsymbol{\epsilon}_{t-1},\boldsymbol{\epsilon}_{t-2},\cdots\sim\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{3}$$

1. Prove that with reparameterization trick we have

   $$\mathbf{x}_t=\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \text{where } \boldsymbol{\epsilon}\cdots\sim\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{4}$$

   which means $q(\mathbf{x}_t|\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_t;\sqrt{\bar{\alpha}_t}\mathbf{x}_0,(1-\bar{\alpha}_t)\mathbf{I})$, where $\{a_t\}$ is a given array and $\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$.

2. Prove the conditional probability $q(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0)$ (which is the target of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$) is a Guassian distribution with mean

   $$\tilde{\boldsymbol{\mu}}_t=\frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\Big) \tag{5}$$

3. Since the KL divergence is always non-negative, we have

$$-\log p_\theta(\mathbf{x}_0) \leq -\log p_\theta(\mathbf{x}_0) + D_{\mathrm{KL}}(q(\mathbf{x}_{1:T}|\mathbf{x}_0)\|p_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0)) \tag{6}$$

Show that

$$\mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] \geq -\mathbb{E}_{q(\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0) \tag{7}$$

and

$$\mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] = \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_T))}_{L_T} + \sum_{t=2}^{T}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}] \tag{8}$$

4. In practise L_T is a constant and L_0 is often taken out for separate processing so here we consider the expression of $L_{1:T-1}$ and since $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ and $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ are gaussian distribution, we have

$$L_t = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[\frac{1}{2\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2\Big] \tag{9}$$

Prove that the formula above can be rewritten as:

$$L_t = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\boldsymbol{\Sigma}_\theta\|_2^2}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2\Big] \tag{10}$$

where $\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\Big)$ and $\boldsymbol{\mu}_\theta(\mathbf{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\Big)$

Problem 3. (Fisher divergence) Let $p_{\text{data}}(x)$ denote the data distribution (unknown) and $p_\theta(x) = \dfrac{e^{-E_\theta(x)}}{Z(\theta)}$ the model distribution, where $E_\theta(x)$ is the energy function and $Z(\theta)$ the partition function. The score function of a distribution $p(x)$ is defined as $\nabla_x \log p(x)$. The Fisher divergence between p_{data} and p_θ is given by:

$$F(p_{\text{data}}\|p_\theta) = \frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}\left[\|\nabla_x \log p_{\text{data}}(x) - \nabla_x \log p_\theta(x)\|_2^2\right].$$

Prove that the Fisher divergence can be rewritten as:

$$F(p_{\text{data}}\|p_\theta) = \mathbb{E}_{p_{\text{data}}}\left[\frac{1}{2}\|\nabla_x \log p_\theta(x)\|_2^2 + \mathrm{tr}(\nabla_x^2 \log p_\theta(x))\right] + \text{Const.},$$

where $\mathrm{tr}(\nabla_x^2 \log p_\theta(x))$ is the trace of the Hessian of $\log p_\theta(x)$.

Hint: Refer to the paper by Song, 2020[2].

Problem 4. (Denoising score matching) Prove that the objective in denoising score matching

$$\int q_\sigma(\tilde{\mathbf{x}})\nabla_{\tilde{\mathbf{x}}}\log q_\sigma(\tilde{\mathbf{x}})^\top \boldsymbol{s}_\theta(\tilde{\mathbf{x}})\mathrm{d}\tilde{\mathbf{x}} \tag{11}$$

can be rewritten as

$$E_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\tilde{\mathbf{x}}\sim q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}\left[\nabla_{\tilde{\mathbf{x}}}\log q_\sigma(\tilde{\mathbf{x}}\mid\mathbf{x})^\top \boldsymbol{s}_\theta(\tilde{\mathbf{x}})\right] \tag{12}$$

Problem 5. The schedule of increasing noise levels in the noise-conditioned score network (NCSN)[2] resembles the forward diffusion process in denoising diffusion probabilistic models (DDPM)[1]. Explain how the diffusion process in DDPM can be used to approximate the score function $\mathbf{s}_\theta\left(\mathbf{x}_t, t\right)$ in NCSN.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[2] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020.