

0 Instructions

Homework is due Tuesday, April 16, 2024 at 23:59pm Central Time. Please refer to <https://courses.grainger.illinois.edu/cs446/sp2024/homework/hw/index.html> for course policy on homeworks and submission instructions.

1 GAN: 5pts

1. The problem will be:

$$\max_{\mathcal{D}} \mathbb{E}_{x \sim p_r(x)} [\log \mathcal{D}(x)] + \mathbb{E}_{x \sim p_g(x)} [\log(1 - \mathcal{D}(x))]$$

which is equivalent to maximize:

$$\int p_r(x) \log \mathcal{D}(x) + p_g(x) \log(1 - \mathcal{D}(x)) dx$$

Hence, the optimal choice of $\mathcal{D}(x)$ is:

$$\mathcal{D}^*(x) = \frac{p_r(x)}{p_r(x) + p_g(x)}$$

2. Plugged in the optimal $\mathcal{D}(x)$, Eq. 1 will turn into:

$$\min_{\mathcal{G}} \mathbb{E}_{x \sim p_r(x)} \left[\log \frac{p_r(x)}{p_r(x) + p_g(x)} \right] + \mathbb{E}_{x \sim p_g(x)} \left[\log \frac{p_g(x)}{p_r(x) + p_g(x)} \right]$$

which is equivalent to minimize:

$$\begin{aligned} & \int p_r(x) \log \frac{p_r(x)}{p_r(x) + p_g(x)} dx + \int p_g(x) \log \frac{p_g(x)}{p_r(x) + p_g(x)} dx \\ &= D_{\text{KL}}(p_r(x) \| p_r(x) + p_g(x)) + D_{\text{KL}}(p_g(x) \| p_r(x) + p_g(x)) \\ &= 2D_{\text{JS}}(p_r(x); p_g(x)) \end{aligned}$$

Therefore, when \mathcal{D} reaches optimal, optimizing Eq. 1 is the same as minimizing $D_{\text{JS}}(p_r(x); p_g(x))$.

3. When \mathcal{D} perfectly classifies generated samples, the output of \mathcal{D} will saturate and the gradient of \mathcal{D} will be almost 0, which makes the gradient of \mathcal{G} almost 0 as well.

2 Diffusion model: 11pts

1.

$$\text{ELBO}_\theta(\mathbf{x}_0) = \sum_{t=1}^T \frac{1}{2\sigma^2} \frac{\beta_t(1 - \bar{\beta}_{t-1})}{\bar{\beta}_t^2} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} [\|\hat{\mathbf{x}}_\theta(\mathbf{x}_t) - \mathbf{x}_0\|_2^2]$$

where $\bar{\beta}_t := 1 - \prod_{i=1}^t (1 - \beta_i)$.

2. No, because $p_\theta(\cdot)$ represent the reconstruction process from random noise in diffusion models and thus cannot directly give the likelihood of an existing test sample.

3.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \prod_{i=1}^t q(\mathbf{x}_i|\mathbf{x}_{i-1}) = \prod_{i=1}^t \mathcal{N}(\mathbf{x}_i; \sqrt{1 - \beta_i} \mathbf{x}_{i-1}, \beta_i \mathbf{I})$$

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}_{t-1} = \sqrt{1 - \beta_t} \sqrt{1 - \beta_{t-1}} \mathbf{x}_{t-2} + \sqrt{\beta_t} \boldsymbol{\epsilon}_{t-1} + \sqrt{1 - \beta_t} \sqrt{\beta_{t-1}} \boldsymbol{\epsilon}_{t-2}$$

We can estimate covariance of the new Gaussian noise $\sqrt{\beta_t} \boldsymbol{\epsilon}_{t-1} + \sqrt{1 - \beta_t} \sqrt{\beta_{t-1}} \boldsymbol{\epsilon}_{t-2}$:

$$\boldsymbol{\sigma}_{t-2} = [(\sqrt{\beta_t})^2 + (\sqrt{1 - \beta_t} \sqrt{\beta_{t-1}})^2] \mathbf{I} = [\beta_t + \beta_{t-1} - \beta_t \beta_{t-1}] \mathbf{I} = [1 - (1 - \beta_t)(1 - \beta_{t-1})] \mathbf{I}$$

and thus:

$$\begin{aligned} \mathbf{x}_t &= \sqrt{(1 - \beta_t)(1 - \beta_{t-1})} \mathbf{x}_{t-2} + \sqrt{1 - (1 - \beta_t)(1 - \beta_{t-1})} \boldsymbol{\epsilon}_{t-2} \\ &= \sqrt{(1 - \beta_t)(1 - \beta_{t-1})(1 - \beta_{t-2})} \mathbf{x}_{t-3} + \sqrt{1 - (1 - \beta_t)(1 - \beta_{t-1})(1 - \beta_{t-2})} \boldsymbol{\epsilon}_{t-3} \\ &= \dots = \sqrt{\prod_{i=1}^t (1 - \beta_i)} \mathbf{x}_0 + \sqrt{1 - \prod_{i=1}^t (1 - \beta_i)} \boldsymbol{\epsilon}_0 \\ &= \sqrt{1 - \bar{\beta}_t} \mathbf{x}_0 + \sqrt{\bar{\beta}_t} \boldsymbol{\epsilon}_0 \end{aligned}$$

where $\bar{\beta}_t := 1 - \prod_{i=1}^t (1 - \beta_i)$. Hence, as $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$, we have:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t | \sqrt{1 - \bar{\beta}_t} \mathbf{x}_0, \bar{\beta}_t \mathbf{I})$$

$$\bar{\beta}_t := 1 - \prod_{i=1}^t (1 - \beta_i)$$

4. From the last question we can get:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}{q(\mathbf{x}_t | \mathbf{x}_0)} = \mathcal{N}(\mathbf{x}_t | \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \frac{\mathcal{N}(\mathbf{x}_{t-1} | \sqrt{1 - \bar{\beta}_{t-1}} \mathbf{x}_0, \bar{\beta}_{t-1} \mathbf{I})}{\mathcal{N}(\mathbf{x}_t | \sqrt{1 - \bar{\beta}_t} \mathbf{x}_0, \bar{\beta}_t \mathbf{I})}$$

$$\propto \exp \left(\frac{(\mathbf{x}_t - \sqrt{1 - \beta_t} \mathbf{x}_{t-1})^2}{2\beta_t} + \frac{(\mathbf{x}_{t-1} - \sqrt{1 - \bar{\beta}_{t-1}} \mathbf{x}_0)^2}{2\bar{\beta}_{t-1}} - \frac{(\mathbf{x}_t - \sqrt{1 - \bar{\beta}_t} \mathbf{x}_0)^2}{2\bar{\beta}_t} \right)$$

Denote the polynomial in the above exponential as $r(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_0)$. Since $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ is a Gaussian distribution, minimize r with respect to \mathbf{x}_{t-1} should lead to the mean $\mu_\theta(\mathbf{x}_t, \mathbf{x}_0)$. Hence, taking derivative of r with respect to \mathbf{x}_{t-1} :

$$\frac{\partial r}{\partial \mathbf{x}_{t-1}} = \frac{-\sqrt{1 - \beta_t} \mathbf{x}_t + (1 - \beta_t) \mathbf{x}_{t-1}}{\beta_t} + \frac{-\sqrt{1 - \bar{\beta}_{t-1}} \mathbf{x}_0 + \mathbf{x}_{t-1}}{\bar{\beta}_{t-1}} = 0$$

$$\Rightarrow \frac{\beta_t + \bar{\beta}_{t-1} - \beta_t \bar{\beta}_{t-1}}{\beta_t \bar{\beta}_{t-1}} \mathbf{x}_{t-1} = \left(\frac{\sqrt{1 - \beta_t} \mathbf{x}_t}{\beta_t} + \frac{\sqrt{1 - \bar{\beta}_{t-1}} \mathbf{x}_0}{\bar{\beta}_{t-1}} \right)$$

$$\Rightarrow \mu_\theta(\mathbf{x}_t, \mathbf{x}_0) = \mathbf{x}_{t-1} = \frac{\bar{\beta}_{t-1} \sqrt{1 - \beta_t} \mathbf{x}_t + \beta_t \sqrt{1 - \bar{\beta}_{t-1}} \mathbf{x}_0}{\beta_t + \bar{\beta}_{t-1} - \beta_t \bar{\beta}_{t-1}}$$

5. According to Bayes' rule,

$$\begin{aligned}\log p_\theta(\mathbf{x}, \delta | \mathbf{x}_{\text{known}}) &= \log \frac{p(\mathbf{x}_{\text{known}} | \mathbf{x}) p_\theta(\mathbf{x}, \delta)}{p(\mathbf{x}_{\text{known}})} \\ &= \log p(\mathbf{x}_{\text{known}} | \mathbf{x}) + \log p_\theta(\mathbf{x}, \delta) - \log p(\mathbf{x}_{\text{known}})\end{aligned}$$

Hence we have:

$$\begin{aligned}\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} | \mathbf{x}_{\text{known}}) &= \nabla_{\mathbf{x}} \log p(\mathbf{x}_{\text{known}} | \mathbf{x}) + \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}, \delta) - \nabla_{\mathbf{x}} \log p(\mathbf{x}_{\text{known}}) \\ &= \nabla_{\mathbf{x}} \log p(\mathbf{x}_{\text{known}} | \mathbf{x}) + \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}, \delta)\end{aligned}$$

Since $p(\mathbf{x}_{\text{known}} | \mathbf{x}) \propto \exp(-\|\mathbf{x} - \mathbf{x}_{\text{known}} \odot \mathbf{M}\|_2^2)$:

$$\begin{aligned}s_\theta(\mathbf{x}, \delta | \mathbf{x}_{\text{known}}) &= \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x} | \mathbf{x}_{\text{known}}) = \nabla_{\mathbf{x}} (-\|\mathbf{x} - \mathbf{x}_{\text{known}} \odot \mathbf{M}\|_2^2) + \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}, \delta) \\ &= s_\theta(\mathbf{x}, \delta) - \nabla_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{\text{known}} \odot \mathbf{M}\|_2^2\end{aligned}$$

3 Unsupervised learning / contrastive learning: 4 pts

1. True.
2. False. MAE is an approach for computer vision, and the mask-out rate can vary greatly.
3. True.
4. False. CLIP does enable zero-shot classification with contrastive pre-training.

4 Coding: GAN, 10pts

5 Coding: Diffusion model, 10pts