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:

$$\int p_r(x)\log\frac{p_r(x)}{p_r(x)+p_g(x)}\,dx + \int p_g(x)\log\frac{p_g(x)}{p_r(x)+p_g(x)}\,dx$$

$$= D_{\mathrm{KL}}(p_r(x)\|p_r(x)+p_g(x)) + D_{\mathrm{KL}}(p_g(x)\|p_r(x)+p_g(x))$$

$$= 2D_{\mathrm{JS}}(p_r(x);p_g(x))$$

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

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

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

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

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

   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:

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

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

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

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

4. From the last question we can get:

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

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

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

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

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

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

5. According to Bayes' rule,

$$\log p_\theta(\boldsymbol{x},\delta|\boldsymbol{x}_{\text{known}}) = \log\frac{p(\boldsymbol{x}_{\text{known}}|\boldsymbol{x})p_\theta(\boldsymbol{x},\delta)}{p(\boldsymbol{x}_{\text{known}})}$$

$$= \log p(\boldsymbol{x}_{\text{known}}|\boldsymbol{x})+\log p_\theta(\boldsymbol{x},\delta)-\log p(\boldsymbol{x}_{\text{known}})$$

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