1

a)

$\mu = f(z) \leftarrow$ decoder خروجی

$$P(x|z) = \frac{1}{2\pi^{d/2} |\Sigma|^{d/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Σ: ماتریس قطری و برابر ثابت σ^2

$$\log P(x|z) = \underbrace{-\frac{d}{2} \log 2\pi |\Sigma|}_{\text{ثابت}} - \frac{1}{2\sigma^2} \|x-\mu\|^2$$

$$\Rightarrow \text{Reconstruct}: \frac{1}{n} \sum_{i=1}^{n} \frac{-1}{2\sigma^2} \|x^i - \mu^i\|^2$$

b)

$y = f_\theta$ (decoder خروجی)

$$P(x|z) = \prod_{k=1}^{d} y_k^{x_k} (1-y_k)^{1-x_k}$$

$$\log P(x|z) = \sum_{k=1}^{d} x_k \log y_k + (1-x_k) \log(1-y_k)$$

$$\Rightarrow \text{Reconstruct} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} x_k^i \log y_k^i + (1-x_k^i) \log(1-y_k^i)$$

c)

درصورت استفاده از توزیع برنولی چند متغیره ، عبارت Reconstruct به‌صورت BCE خواهد بود. این توزیع گسسته است و شاید باعث کاهش ظرفیت مدل شود. (فرض پیروی داده‌ها از توزیع برنولی سخت‌گیرانه و ممکن است نادقیق باشد)

درصورت استفاده از توزیع گاوسی چند متغیره عبارت Reconstruct به‌صورت MSE است و چون این توزیع پیوسته است و داده‌ها را به‌صورت توزیع نرمال مدل می‌کند احتمالاً مدل بهتر می‌تواند تحمل کند.

d) فرض کنید داده‌ها حول یک sin باشند () حال اگر شبکه برای این دادگان فضای نهانی که بدست می‌آورد به‌صورت () باشد با اینکه Reconstruct Error ضعیف است اما اگر در بعضی نقاط این منحنی نمونه برداریم و دادگان جدید تولید کنیم، شبیه داده‌های آموزش نیست. عبارت KL در تابع هزینه باعث می‌شود توزیع بدست آمده در فضای نهان شبیه توزیع پایه‌ای که برای z در نظر گرفته‌ایم شود. درنتیجه موقع نمونه برداری از فضای نهان و تولید داده جدید، شبیه دادگان آموزش باشد

$$P(x,y) = \begin{cases} 1 & 0 \leq y \leq 1, \ x_1 = 0 \\ 0 & o.w \end{cases}$$

② (a

$$Q(x,y) = \begin{cases} 1 & 0 \leq y < 1, \ x_1 = \theta \\ 0 & o.w \end{cases}$$

$$KL(P \| Q) = \begin{cases} 0 & \text{if } \theta = 0 \Rightarrow P = Q \Rightarrow KL = 0 \\ \int_{y=0}^{1} P(0,y) \log \frac{P(0,y)}{Q(0,y)} dy & \text{if } \theta \neq 0 \\ = \log \frac{1}{0} = \infty & \end{cases}$$

$$KL(Q \| P) = \begin{cases} 0 & \text{if } \theta = 0 \\ \int_{y=0}^{1} Q(\theta,) \log \frac{Q(\theta,y)}{P(\theta,y)} dy & \text{if } \theta \neq 0 \end{cases}$$

$$= \log \frac{1}{0} = \infty$$

$$JSD(P,Q) = \frac{1}{2} KL\left(P \| \frac{P+Q}{2}\right) + \frac{1}{2} KL\left(Q \| \frac{P+Q}{2}\right)$$

$$\begin{cases} 0 & \text{if } \theta = 0 \Rightarrow P = Q \Rightarrow P = Q = \frac{P+Q}{2} \Rightarrow JSD(P,Q) = 0 \\ \frac{1}{2}\left(\int_{y=0}^{1} P(0,y) \log \frac{P(0,y)}{\frac{P(0,y)+Q(0,y)}{2}} + \int_{y=0}^{1} Q(\theta,y) \log \frac{Q(\theta,y)}{\frac{P(\theta,y)+Q(\theta,y)}{2}} dy \right) & \text{if } \theta \neq 0 \end{cases}$$

$$\frac{1}{2}(\log 2 + \log 2) = \log 2$$

$$W(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} E_{(x,y)\sim\gamma} \|x-y\| = \inf_{\gamma \in \Gamma(P,Q)} E_{(x_1,y_1,x_2,y_2)\sim\gamma} \left[|x_1 - x_2| + |y_1 - y_2| \right]$$

$$\gamma(x_1, y_1, x_2, y_2) = \begin{cases} \beta(y_1, y_2) & \text{if } x_1 = 0, \ x_2 = \theta \\ 0 & o.w \end{cases}$$

$$W(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} E_{(x_1,y_1,x_2,y_2)\sim\gamma}\left[|x_1 - x_2| + |y_1 - y_2|\right]$$

$$= \inf_{\gamma \in \Gamma(P,Q)} E_{(y_1,y_2)\sim\beta}\left[|0-\theta| + |y_1 - y_2|\right]$$

$$= \theta \cdot \inf_{\beta \sim \Gamma(\text{uniform}(0,1),\ \text{uniform}(0,1))} E_{(y_1,y_2)\sim\beta}\left(|y_1 - y_2|\right)$$

$$\approx \theta$$

(b

$$P(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} & x_1 = 0 \\ 0 & \text{o.w} \end{cases}$$

$$Q(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} & x_1 = \theta \\ 0 & \text{o.w} \end{cases}$$

$$JSD(P,Q) = \frac{1}{2} KL\left(P \| \frac{P+Q}{2}\right) + \frac{1}{2} KL\left(Q \| \frac{P+Q}{2}\right)$$

$$= \begin{cases} 0 & \text{if } \theta = 0 \\ \frac{1}{2}\left(\int_y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \log \frac{\frac{1}{\sqrt{2\pi}} e^{\frac{y^2}{2}}}{\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} + 0}{2}} dy + \int_y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \log \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}}{\frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} + 0}{2}} dy\right) & \text{if } \theta \neq 0 \end{cases}$$

$$= \begin{cases} 0 & \text{if } \theta = 0 \\ \log 2 & \text{if } \theta \neq 0 \end{cases} \Rightarrow \frac{\partial JSD(P,Q)}{\partial \theta}$$ قابل محاسبه نیست

c) با توجه به اینکه JSD نسبت به پارامتر های توزیع P و Q می‌تواند مشتق پذیر نباشد
در نتیجه همگرایی شبکه GAN می‌تواند با مشکل روبرو شود.

3

(a

$$L_D(\phi;\theta) = -E_{x\sim P_{data}(x)}\left[\log D_\phi(x)\right] - E_{x\sim P_\theta(x)}\left[\log(1-D_\phi(x))\right]$$

$$= -\int P_{data}(x)\log D_\phi(x) + P_\theta(x)\log(1-D_\phi(x))\,dx$$

$$\xRightarrow{\text{مشتق}} \quad -\frac{P_{data}(x)}{D_\phi(x)} + \frac{P_\theta(x)}{1-D_\phi(x)} = 0$$

$$-P_{data}(x)(1-D_\phi(x)) + P_\theta(x)D_\phi(x) = 0$$

$$D^*(x) = \frac{P_{data}(x)}{P_\theta(x)+P_{data}(x)}$$

(b

$$\left.\begin{aligned} D^* &= \frac{P_{data}(x)}{P_\theta(x)+P_{data}(x)} = \frac{1}{1+\frac{P_\theta(x)}{P_{data}(x)}} \\ D^* &= \sigma(h_\phi(x)) = \frac{1}{1+e^{-h_\phi(x)}} \end{aligned}\right\} \rightarrow e^{-h_\phi(x)} = \frac{P_\theta(x)}{P_{data}(x)}$$

$$-h_\phi(x) = \log\frac{P_\theta(x)}{P_{data}(x)}$$

$$h_\phi(x) = \log\frac{P_{data}(x)}{P_\theta(x)}$$

$$L_G(\theta;\phi) = E_{x\sim P_\theta(x)}\left[\log(1-D_\phi(x))\right] - E_{x\sim P_\theta(x)}\left[\log D_\phi(x)\right]$$ (c

$$= E_{x\sim P_\theta(x)}\left[\log\frac{1-D_\phi(x)}{D_\phi(x)}\right]$$

$$= E_{x\sim P_\theta(x)}\left[-\log D_\phi(x) + \cancelto{0}{\log 1}\right]$$

$$= E_{x\sim P_\theta(x)}\left[-\log\frac{P_{data}(x)}{P_\theta(x)+P_{data}(x)}\right] =$$

$$= E_{x\sim P_\theta(x)}\left[\log\frac{P_\theta(x)}{P_{data}(x)} + \cancelto{0}{\log 1}\right] = E_{x\sim P_\theta(x)}\left[\log\frac{P_\theta(x)}{P_{data}(x)}\right]$$

$$= KL\left(P_\theta(x) \,\|\, P_{data}(x)\right)$$

c) a

$$-E_{x\sim P_{data}(x)}\left[\log P_\theta(x)\right] = -E_{x\sim P_{data}(x)}\left[\log P_\theta(x) + \log P_{data}(x) - \log P_\theta(x)\right]$$

$$= -E_{x\sim P_{data}(x)}\left[\log \frac{P_\theta(x)}{P_{data}(x)} + \log P_{data}(x)\right]$$

$$= E_{x\sim P_{data}(x)}\left[\log \frac{P_{data}(x)}{P_\theta(x)}\right] + E_{x\sim P_{data}(x)}\left[\log P_{data}(x)\right]$$

$$= KL\left(P_{data}(x) \,\|\, P_\theta(x)\right) + \underbrace{E_{x\sim P_{data}(x)}\left[\log P_{data}(x)\right]}_{\text{برحسب } \theta \text{ ثابت است.}}$$

$$= KL\left(P_{data}(x) \,\|\, P_\theta(x)\right) + const$$

از آنجایی که $KL\left(P_{data}(x) \,\|\, P_\theta(x)\right) \neq KL\left(P_\theta(x) \,\|\, P_{data}(x)\right)$ پس تابع منفی ELBO

و تابع ضرر L ها مشابه هم نیستند.

4

a)

$$q(x_t \mid x_{t-1}) = N\left(x_t \mid \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right)$$

$$\alpha_t = (1-\beta_t), \quad \bar\alpha = \prod_{i=1}^{t}(1-\beta_i) = \prod_{i=1}^{t}\alpha_i$$

$$q(x_t \mid x_{t-1}) = N\left(x_t \mid \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right)$$

$$= N\left(x_t \mid \sqrt{(1-\beta_t)(1-\beta_{t-1})}\, x_{t-2}, 1-(1-\beta_t)(1-\beta_{t-1}) I\right)$$

$$\vdots \quad \vdots$$

$$= N\left(x_t \mid \sqrt{(1-\beta_t)(1-\beta_{t-1})\cdots(1-\beta_1)}\, x_0, 1-\left((1-\beta_t)\cdots(1-\beta_1)\right) I\right)$$

$$= N\left(x_t \mid \sqrt{\bar\alpha_t}\, x_0, 1-\bar\alpha_t I\right)$$

b) اگر $\bar\alpha_t \to 0$ باشد $x_t \sim N(0,I)$ می‌شود.

c)

$$q(x_{t-1}|x_t) \propto q(x_{t-1})\, q(x_t|x_{t-1}) \rightarrow \text{interactable}$$ است.

محاسبه marginal distribution

$q(x_{t-1})$ قابل محاسبه نیست.

d)

$$q(x_t|x_0) \sim \mathcal{N}(x_t, \sqrt{\bar\alpha_t}\,x_0, (1-\bar\alpha)I)$$

$$q(x_t|x_{t-1}, x_0) \sim \mathcal{N}(x_t, \sqrt{1-\beta_t}\,x_{t-1}, \beta_t I) \quad , \quad q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}|\tilde\mu(x_t,x_0), \tilde\beta_t I)$$

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)\, q(x_{t-1}|x_0)}{q(x_t|x_0)}$$

$$= e^{-\frac{1}{2}\frac{(x_t-\sqrt{1-\beta_t}\,x_{t-1})^2}{\beta_t}} \cdot e^{-\frac{1}{2}\left(\frac{(x_{t-1}-\sqrt{\bar\alpha_{t-1}}\,x_0)^2}{1-\bar\alpha_{t-1}}\right)} \cdot e^{\frac{1}{2}\left(\frac{(x_t-\sqrt{\bar\alpha_t}\,x_0)^2}{1-\bar\alpha_t}\right)^2}$$

$$= e^{-\frac{1}{2}\left(\frac{(x_t-\sqrt{1-\beta_t}\,x_{t-1})^2}{\beta_t} + \frac{(x_{t-1}-\sqrt{\bar\alpha_{t-1}}\,x_0)^2}{1-\bar\alpha_{t-1}} - \frac{(x_t-\sqrt{\bar\alpha_t}\,x_0)^2}{1-\bar\alpha_t}\right)}$$

عبارت مستقل از x_{t-1}

$$= e^{-\frac{1}{2}\left(\left(\frac{-2\sqrt{1-\beta_t}\,x_t}{\beta_t} - \frac{2\sqrt{\bar\alpha_t}\,x_0}{1-\bar\alpha_t}\right)x_{t-1} + \left(\frac{(1-\beta_t)}{\beta_t} + \frac{1}{1-\bar\alpha_t}\right)x_{t-1}^2 + (\cdot)\right)}$$

عبارت مستقل از x_{t-1}

$$= e^{-\frac{1}{2}\left(\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar\alpha_{t-1}}\right)x_{t-1}^2 - \left(\frac{2\sqrt{\alpha_t}}{\beta_t}x_t + \frac{2\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}x_0\right)x_{t-1} + (\cdot)\right)}$$

$$\Rightarrow \hat\beta_t = \frac{1}{\left(\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar\alpha_{t-1}}\right)} = \frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t$$

$$\hat\mu(x_t, x_0) = \frac{\left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}x_0\right)}{\frac{\alpha_t}{\beta_t} + \frac{1}{1-\bar\alpha_{t-1}}} = \left(\frac{\sqrt{\alpha_t}}{\beta_t}x_t + \frac{\sqrt{\bar\alpha_{t-1}}}{1-\bar\alpha_{t-1}}x_0\right)\frac{1-\bar\alpha_{t-1}}{1-\bar\alpha_t}\cdot\beta_t$$

$$= \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t}x_t + \frac{\sqrt{\bar\alpha_{t-1}}\,\beta_t}{1-\bar\alpha_t}x_0$$

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_t\right) \Longrightarrow \tilde{\mu}_t = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_t\right)$$

علت چشم‌پوشی از L_T: چون L_T ثابت است و q هیچ پارامتر قابل یادگیری ندارد.

ه)

$$L_t = E_{x_0,\varepsilon}\left[\frac{1}{2\|\Sigma_\theta(x_t,t)\|_2^2}\left\|\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)\right\|^2\right]$$

$$= E_{x_0,\varepsilon}\left[\frac{1}{2\|\Sigma_\theta\|_2^2}\left\|\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_t\right) - \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_\theta(x_t,t)\right)\right\|^2\right]$$

$$= E_{x_0,\varepsilon}\left[\frac{1}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2}\left\|\varepsilon_t - \varepsilon_\theta(x_t,t)\right\|^2\right]$$

$$= E_{x_0,\varepsilon}\left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2}\left\|\varepsilon_t - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}\,x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon_t, t\right)\right\|^2\right]$$

و)

Lengevin dynamics برای نمونه‌گیری نقاط از توزیع احتمال، از $\nabla_x \log q(x)$ استفاده می‌کند. با توجه به اینکه اکثر داده‌ها در یک manifold با ابعاد پایین متمرکز می‌شوند و به دلیل اینکه نقاط نمی‌توانند کل فضا را پوشش دهند، بر تخمین تأثیر منفی دارد. در مناطق با چگالی کم، تخمین ممکن است نادقیق باشد. پس از اضافه کردن یک نویز گاوسی کوچک داده‌ها کل فضا را پوشش می‌دهند. که باعث می‌شود تخمین stable تر شود.