بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۴-۰۳

مدرس: مهدیه سلیمانی

تمرین چهارم | پاسخنامه

- طراحان: پرهام رضایی، امیرحسین حاجی‌محمدرضایی، علیرضا فرج‌تبریزی، محمدجواد احمدپور

بخش نظری (۱۰۰ نمره)

پرسش ۱. طراحی گام به گام VAE (۲۰ نمره)

در این تمرین قصد داریم به صورت گام به گام یک مدل VAE را توسعه دهیم. به دلیل وابستگی بین صورت سوال هر بخش به جواب بخش قبل، در صورت هرگونه ابهام یا سوال می‌توانید در تلگرام ابهامات خود را بر طرف نمایید (ایدی طراح Alireza_1197@). در ادامه پیش‌فرض ها مسئله آمده است:

فرضیات:

ما مدل خود را بر پایه‌ی یک معماری **encoder–decoder** با یک **متغیر نهان پیوسته** توسعه خواهیم داد. مؤلفه‌های کلیدی عبارتند از:

- **داده‌ها**: مجموعه‌ای از مشاهدات $\mathcal{D} = \{X^{(i)}\}_{i=1}^{N}$، که در آن هر داده‌ی $X^{(i)} \in \mathbb{R}^n$ است.
- **فضای نهان (Latent Space)**: یک متغیر نهان $Z \in \mathbb{R}^m$ که اطلاعات فشرده‌شده‌ای از ورودی را نمایش می‌دهد.
- **مدل مولد (Decoder)**: مدل توزیع خروجی را یک توزیع گوسی به شرط توابعی از متغیر نهان تعریف می‌کند:
$$p_\theta(X \mid Z) = \mathcal{N}(X \mid \mu_\theta(Z), \Sigma_\theta(Z))$$
که در آن Decoder نگاشت Z به پارامترهای خروجی را انجام می‌دهد:
$$\mu_\theta(Z) = \mathrm{NN}_\alpha(Z), \quad \Sigma_\theta(Z) = \mathrm{diag}(\sigma(\mathrm{NN}_\beta(Z)))$$
- **توزیع پیشین (Prior)**: متغیر پنهان از یک توزیع گاوسی چندمتغیره نمونه گیری می می‌شود:
$$p_\theta(Z) = \mathcal{N}(Z \mid \mu_z, \sigma_z^2 I)$$
- **پارامترهای مدل**: مجموعه پارامترهای قابل آموزش $\theta = \{\mu_z, \sigma_z^2, \alpha, \beta\}$، که α و β وزن‌های شبکه Decoder هستند.

گام ۱: برآورد درست‌نمایی (Likelihood Estimation)

در مدل‌های مولد، هدف بیشینه کردن درست‌نمایی داده‌های مشاهده‌شده در توزیع خروجی مدل است.
(الف) یک عبارت برای **درست‌نمایی حاشیه‌ای** $p_\theta(X)$ برای یک نمونه‌ی داده X استخراج کنید.
(ب) **چالش اصلی** در محاسبه یا بهینه‌سازی مستقیم $p_\theta(X)$ در عمل چیست؟
(ج) از آنجا که محاسبه‌ی $p_\theta(X)$ عملی نیست، یک **تقریب** برای آن پیشنهاد دهید.
راهنمایی: سعی کنید درست‌نمایی حاشیه‌ای را به صورت امید ریاضی بیان کنید.

گام ۲: کاهش واریانس با Importance Sampling

چالش اصلی در برآورد $p_\theta(X)$ از طریق نمونه‌گیری، **واریانس بالا** است، زیرا نمونه‌های Z ممکن است از نواحی‌ای بیایند که به داده‌ی مشاهده‌شده‌ی X ارتباطی ندارند.
(الف) برای حل این مسئله، از importance sampling استفاده می‌کنیم. $p_\theta(X)$ را با استفاده از یک توزیع پیشنهادی $q(Z)$ بازنویسی کنید.
(ب) یک **انتخاب شهودی** برای توزیع پیشنهادی $q(Z)$ معرفی کنید.

گام ۳: از Likelihood به ELBO

برای توجیه انتخاب توزیع پیشنهادی (Proposal Distribution)، دوباره به تخمین تابع هدف $p_\theta(X)$ بازمی‌گردیم.
(الف) چرا در یادگیری ماشین معمولاً با $\log p_\theta(X)$ کار می‌کنیم و نه با خود $p_\theta(X)$؟
(ب) چرا نمی‌توانیم $\log p_\theta(X)$ را مستقیماً برآورد کنیم؟ با استفاده از Jensen's inequality یک **کران پایین** برای آن استخراج کنید (این کران به عنوان Evidence Lower Bound یا ELBO شناخته می‌شود).
(ج) آیا بیشینه کردن این کران پایین در حین آموزش به‌تنهایی کافی است؟ چه مشکلاتی ممکن است ایجاد شود؟
(د) **تفاضل** بین $\log p_\theta(X)$ و کران پایین استخراج‌شده چیست؟
راهنمایی: این فاصله را به صورت KL divergence بنویسید.

گام ۴: پارامتری کردن و بهینه‌سازی Posterior Approximation

تا اینجا انتخاب توزیع پیشنهادی را توجیه کردیم، اما با یک چالش جدید مواجه هستیم: posterior واقعی $p_\theta(Z \mid X)$ غیرقابل محاسبه است. بنابراین نمی‌توانیم مستقیماً KL divergence را کمینه کنیم. برای رفع این مشکل باید فرم توزیع تقریبی $q(Z)$ را مشخص کنیم. یک انتخاب رایج، توزیع گاوسی است:

$$q_\lambda(Z) = \mathcal{N}(Z \mid \mu, \sigma^2 I), \quad \lambda = \{\mu, \sigma^2\}$$

(الف) چگونه می‌توان KL divergence را به صورت **غیرمستقیم** کمینه کرد؟
راهنمایی: به گرادیان $\nabla_\lambda \log p_\theta(X)$ فکر کنید.
(ب) گرادیان ∇_θ ELBO و تخمین Monte Carlo آن را محاسبه کنید.
(ج) آیا می‌توانید ∇_λ ELBO را نیز محاسبه کنید؟ چه چالش‌هایی در هنگام تخمین این گرادیان با نمونه‌گیری Monte Carlo وجود دارد و چگونه می‌توان آن‌ها را حل کرد؟
راهنمایی: بررسی کنید چگونه می‌توان نمونه‌گیری $Z \sim q_\lambda(Z)$ را طوری نوشت که تصادفی بودن از پارامترهای λ جدا شود.

گام ۵: تعمیم به کل مجموعه داده

تمام محاسبات قبلی بر اساس یک نمونه داده انجام شدند. اما اگر بخواهیم $\log p_\theta(X)$ را روی کل مجموعه داده بیشینه کنیم، چه تغییراتی ایجاد می‌شود؟
(الف) این کار چه تأثیری روی **تابع هدف** و **گرادیان‌ها** نسبت به θ و λ دارد؟

(ب) پارامترهای رمزگشا (θ) به‌سادگی قابل تعمیم به کل مجموعه هستند. اما λ، که پارامترهای توزیع تقریبی $q_\lambda(Z)$ است، برای هر داده به طور جدا تعریف شده. چگونه می‌توان λ را طوری مدل کرد که قابلیت تعمیم به همه داده‌ها را داشته باشد؟
راهنمایی: اینجا نقش شبکه encoder مطرح می‌شود (پارامترهای آن را با ϕ نام‌گذاری کنید).

گام ۶: ELBO نهایی و آموزش مدل

اکنون که کل هدف VAE استخراج شده است، فرآیند آموزش را توصیف کنید.
(الف) نشان دهید که ELBO را می‌توان به صورت زیر نوشت و مفهوم هر ترم از عبارات را توضیح دهید:

$$\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \mathrm{KL}\left(q_\phi(z \mid x) \,\|\, p_\theta(z)\right)$$

(ب) حال که مدل VAE کامل شده است، الگوریتم آموزش آن را بنویسید. در این الگوریتم نشان دهید:

- چگونه از encoder و decoder استفاده می‌شود.
- نمونه‌گیری چگونه انجام می‌شود.
- .تابع هدف چگونه بهینه می‌شود

پاسخ.

گام ۱: برآورد Likelihood

(الف) Marginal likelihood برای یک نمونه داده X به صورت زیر محاسبه می‌شود:

$$p_\theta(X) = \int p_\theta(X \mid Z)\, p_\theta(Z)\, dZ$$

(ب) چالش اصلی در محاسبه‌ی $p_\theta(X)$ این است که این انتگرال معمولاً به دلیل فضای نهان با بعد بالا و نگاشت غیرخطی decoder قابل محاسبه به‌صورت تحلیلی نیست.
(ج) می‌توان انتگرال بالا را به صورت امید ریاضی نسبت به توزیع prior بازنویسی کرد:

$$p_\theta(X) = \mathbb{E}_{Z\sim p_\theta(Z)}[p_\theta(X \mid Z)]$$

با توجه به اینکه محاسبه این امید ریاضی به‌صورت تحلیلی ممکن نیست، از تخمین Monte Carlo استفاده می‌کنیم. برای این کار، L نمونه $Z^{(l)} \sim p_\theta(Z)$ گرفته و تخمین می‌زنیم:

$$p_\theta(X) \approx \frac{1}{L}\sum_{l=1}^{L} p_\theta(X \mid Z^{(l)})$$

اما این تخمین واریانس بالایی دارد چرا که نمونه‌های $Z^{(l)}$ لزوما از نواحی مرتبط با X نیستند و نمونه‌های زیادی نیاز هست تا بتوان از نواحی مرتبط با این X خاص هم نمونه در تخمین داشته باشیم.

گام ۲: کاهش واریانس با Importance Sampling

(الف) برای کاهش واریانس، marginal likelihood را با استفاده از توزیع پیشنهادی $q(Z)$ بازنویسی می‌کنیم:

$$p_\theta(X) = \int \frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)} q(Z)\, dZ = \mathbb{E}_{q(Z)}\left[\frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)}\right]$$

(ب) یک انتخاب شهودی برای $q(Z)$، توزیع posterior واقعی است:

$$q(Z) = p_\theta(Z \mid X)$$

چون این توزیع Zهایی را پیشنهاد می‌دهد که بر اساس این X ،خاص احتمال بالایی دارند پس این نمونه‌ها اطلاعات بیشتری دارند و با تعداد یکسان نمونه از Z تخمین با واریانس کمتری نسبت به تخمین مونت کارلو اولیه ارائه می‌کنند.

گام ۳: از Likelihood به ELBO

(الف) ما معمولاً در یادگیری ماشین با $\log p_\theta(X)$ کار می‌کنیم و نه خود $p_\theta(X)$. دلایل اصلی این انتخاب عبارتند از:

- **پایداری عددی**: مقادیر احتمال می‌توانند بسیار کوچک باشند. استفاده از لگاریتم باعث می‌شود محاسبات پایدارتر و از نظر عددی در محدوده بهتری باشند.
- **ساده‌سازی ریاضی**: استفاده از لگاریتم باعث می‌شود توابع نمایی در مدل‌های Gaussian یا خانواده‌های نمایی (Exponential Family) حذف شوند و مشتق‌گیری و تعریف تابع زیان ساده‌تر شود.
- **تبدیل حاصل‌ضرب به جمع**: وقتی می‌خواهیم درست‌نمایی کل مجموعه داده را بیشینه کنیم، داریم:

$$\log p_\theta(\mathcal{D}) = \sum_{i=1}^{N} \log p_\theta(X^{(i)})$$

این فرم جمعی این مزیت را دارد که هنگام آموزش با mini-batch، سیگنال‌های گرادیانی متعددی به مدل وارد می‌شود که باعث یادگیری بهتر می‌گردد.

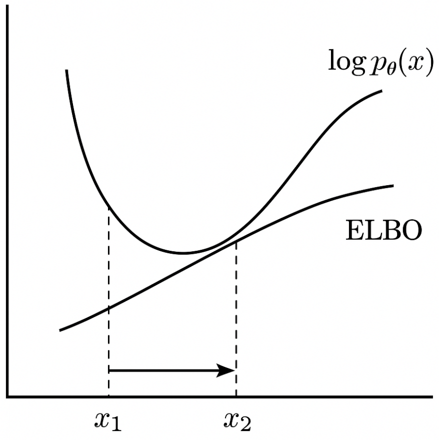
(ب) ابزارهای ریاضی ای که در اختیار داریم به ما در تخمین امید ریاضی کمک می‌کنند نه لگاریتم آن، لذا نیاز هست که اول آن را به فرم امید ریاضی دربیاوریم. با استفاده از نابرابری Jensen، یک کران پایین برای $\log p_\theta(X)$ به‌دست می‌آوریم:

$$\log p_\theta(X) = \log \int p_\theta(X \mid Z)p_\theta(Z)\, dZ = \log \mathbb{E}_{q(Z)}\left[\frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)}\right]$$

$$\geq \mathbb{E}_{q(Z)}\left[\log \frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)}\right] = \mathbb{E}_{q(Z)}\left[\log \frac{p_\theta(X, Z)}{q(Z)}\right] = \mathbb{E}_{q(Z)}\left[\log p_\theta(X, Z) - \log q(Z)\right]$$

که این عبارت همان ELBO است.
(ج) مشکل افزایش کران پایین این است که هیچ تضمینی وجود ندارد با افزایش کران پایین تابع هدف هم افزایش یابد. به شکل یک توجه کنید:
(د) تفاضل $\log p_\theta(X)$ و ELBO به صورت زیر محاسبه می‌شود:

$$\log p_\theta(X) - \mathbb{E}_{q(Z)}\left[\log \frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)}\right] = \mathbb{E}_{q(Z)}\left[\log p_\theta(X) - \log \frac{p_\theta(X \mid Z)p_\theta(Z)}{q(Z)}\right]$$



شکل ۱: ELBO vs $\log p_\theta(X)$

$$= \mathbb{E}_{q(Z)}\left[\log\left(\frac{p_\theta(X)q(Z)}{p_\theta(X \mid Z)p_\theta(Z)}\right)\right] = \mathbb{E}_{q(Z)}\left[\log\left(\frac{q(Z)}{p_\theta(Z \mid X)}\right)\right] = \mathrm{KL}(q(Z) \parallel p_\theta(Z \mid X))$$

در اینجا از این واقعیت استفاده شده است که $\log p_\theta(X)$ نسبت به $q(Z)$ مستقل است (از انتگرال خارج می‌شود)، بنابراین می‌توان آن را به صورت امید ریاضی نوشت:

$$\log p_\theta(X) = \mathbb{E}_{q(Z)}[\log p_\theta(X)]$$

گام ۴: پارامتری کردن و بهینه‌سازی تقریب Posterior

(الف) چگونه می‌توان KL divergence را به صورت غیرمستقیم کمینه کرد؟
هدف ما کمینه‌سازی عبارت زیر است:

$$\mathrm{KL}(q_\lambda(Z) \parallel p_\theta(Z \mid X))$$

اما چون $p_\theta(Z \mid X)$ ناشناخته و غیرقابل محاسبه است، نمی‌توان این عبارت را به‌صورت مستقیم کمینه کرد. بنابراین به جای آن، ما ELBO را به صورت زیر بیشینه می‌کنیم:

$$\mathcal{L}(X) = \mathbb{E}_{q_\lambda(Z)}[\log p_\theta(X, Z) - \log q_\lambda(Z)]$$

برای توجیه این کار، گرادیان درست‌نمایی را در نظر بگیرید:

$$\nabla_\lambda \log p_\theta(X) = \nabla_\lambda \left(\mathcal{L}(X) + \mathrm{KL}(q_\lambda(Z) \parallel p_\theta(Z \mid X))\right)$$

اما از آنجا که $\log p_\theta(X)$ به λ وابسته نیست، داریم:

$$\nabla_\lambda \log p_\theta(X) = 0$$

در نتیجه:

$$\nabla_\lambda \mathcal{L}(X) = -\nabla_\lambda \mathrm{KL}(q_\lambda(Z) \parallel p_\theta(Z \mid X))$$

این رابطه نشان می‌دهد که بیشینه‌سازی ELBO با توجه به λ، معادل کمینه‌سازی KL divergence با توزیع پسین واقعی است.
(ب) گرادیان ELBO نسبت به θ و تخمین Monte Carlo آن:

فرمول ELBO را به صورت زیر یادآوری می‌کنیم:

$$\mathcal{L}(X) = \mathbb{E}_{Z\sim q_\lambda(Z)}[\log p_\theta(X,Z) - \log q_\lambda(Z)]$$

از آنجا که ترم دوم به θ وابسته نیست، تنها گرادیان ترم اول موردنیاز است:

$$\nabla_\theta \mathcal{L}(X) = \nabla_\theta\, \mathbb{E}_{Z\sim q_\lambda(Z)}[\log p_\theta(X,Z) - \log q_\lambda(Z)]$$

می‌توان گرادیان را داخل امید ریاضی برد:

$$= \mathbb{E}_{Z\sim q_\lambda(Z)}[\nabla_\theta \log p_\theta(X,Z)]$$

زیرا توزیع نمونه‌گیری $q_\lambda(Z)$ به θ وابسته نیست. بنابراین Z در هنگام گرادیان‌گیری نسبت به θ، مقدار ثابتی محسوب می‌شود.
برای تخمین این امید ریاضی، از نمونه‌گیری Monte Carlo استفاده می‌کنیم. با گرفتن L نمونه $Z^{(l)} \sim q_\lambda(Z)$ خواهیم داشت:

$$\nabla_\theta \mathcal{L}(X) \approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\theta \log p_\theta(X, Z^{(l)})$$

این گرادیان در عمل با Backpropagation قابل محاسبه است.
(ج) چرا برای تخمین $\nabla_\lambda \mathcal{L}(X)$ نیاز به Reparameterization Trick داریم؟

$$\nabla_\lambda \mathbb{E}_{Z\sim q_\lambda(Z)}[\log p_\theta(X,Z) - \log q_\lambda(Z)] = \nabla_\lambda \int q_\lambda(Z)\left(\log p_\theta(X,Z) - \log q_\lambda(Z)\right) dZ$$

به‌صورت ایده‌آل، می‌خواهیم گرادیان را به داخل امید ریاضی منتقل کنیم:

$$= \int \nabla_\lambda \left(q_\lambda(Z)\left(\log p_\theta(X,Z) - \log q_\lambda(Z)\right)\right) dZ$$

اما این کار به‌راحتی ممکن نیست، چون $q_\lambda(Z)$ به λ وابسته است، و در نتیجه خود Z نیز به λ وابسته خواهد بود. این وابستگی باعث می‌شود که نتوان به‌سادگی گرادیان را منتقل کرد و تخمین آن ناپایدار و پرنوسان باشد.
راه‌حل: استفاده از Reparameterization Trick. اگر $q_\lambda(Z) = \mathcal{N}(Z \mid \mu, \sigma^2 I)$ باشد، می‌توان Z را به صورت زیر بازنویسی کرد:

$$Z = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

در این صورت، تابع هدف به شکل زیر بازنویسی می‌شود:

$$\mathbb{E}_{Z\sim q_\lambda(Z)}[\log p_\theta(X,Z) - \log q_\lambda(Z)] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\log p_\theta(X, \mu + \sigma\cdot\epsilon) - \log q_\lambda(\mu + \sigma\cdot\epsilon)]$$

حالا ϵ مستقل از λ است، و ما می‌توانیم گرادیان را بدون مشکل محاسبه کرده و Backpropagation را روی μ و σ اعمال کنیم.

گام ۵: تعمیم به کل مجموعه داده

(الف) هدف جدید برای کل داده‌ها به صورت مجموع ELBO برای تمام نمونه‌هاست:

$$\sum_{i=1}^{N} \log p_\theta(X^{(i)}) \geq \sum_{i=1}^{N} \mathcal{L}(X^{(i)})$$

(ب) برای تعمیم λ به کل داده‌ها، از یک شبکه encoder استفاده می‌کنیم که برای هر X پارامترهای $q(Z \mid X)$ را تولید می‌کند:

$$q_\phi(Z \mid X) = \mathcal{N}(Z \mid \mu_\phi(X), \sigma^2_\phi(X) I)$$

در نتیجه، $\lambda = \phi$ و قابل یادگیری برای کل داده‌ها خواهد بود.

گام ۶: ELBO نهایی و آموزش مدل

(الف) با استفاده از کران پایین به‌دست‌آمده از نابرابری Jensen، داریم:

$$\log p_\theta(X) \geq \mathbb{E}_{q_\phi(Z|X)}\left[\log \frac{p_\theta(X \mid Z)p_\theta(Z)}{q_\phi(Z \mid X)}\right]$$

می‌توان لگاریتم را به صورت جمعی از دو بخش بازنویسی کرد:

$$= \mathbb{E}_{q_\phi(Z|X)}[\log p_\theta(X \mid Z)] + \mathbb{E}_{q_\phi(Z|X)}[\log p_\theta(Z) - \log q_\phi(Z \mid X)]$$

عبارت دوم را می‌توان به صورت KL divergence نوشت:

$$= \mathbb{E}_{q_\phi(Z|X)}[\log p_\theta(X \mid Z)] - \mathrm{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z))$$

در نتیجه فرم نهایی ELBO برابر است با:

$$\mathcal{L}(X) = \mathbb{E}_{Z\sim q_\phi(Z|X)}[\log p_\theta(X \mid Z)] - \mathrm{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z))$$

- $\mathbb{E}_{Z\sim q_\phi(Z|X)}[\log p_\theta(X \mid Z)]$: این ترم بیان می‌کند که اگر Zهایی که توسط encoder (یعنی $q_\phi(Z \mid X)$) برای یک ورودی X پیشنهاد شده‌اند، به decoder داده شوند، تا چه اندازه احتمال دارد که دوباره X تولید شود. به عبارت دیگر، اگر encoder می‌گوید «این Zها خوب هستند»، آنگاه decoder نیز باید بتواند با آن‌ها دقیقاً X را بازسازی کند. این ترم به عنوان تابع هزینه **بازسازی** شناخته می‌شود و هرچه بزرگ‌تر باشد، یعنی decoder عملکرد بهتری داشته است.

- $\mathrm{KL}(q_\phi(Z \mid X) \parallel p_\theta(Z))$: این ترم به‌عنوان regularization عمل می‌کند. این بخش فاصله بین توزیع نهان تقریبی $q_\phi(Z \mid X)$ و توزیع پیشین $p_\theta(Z)$ را با استفاده از KL divergence اندازه‌گیری می‌کند. هدف آن این است که بازنمایی‌های نهان از داده‌ها، با ساختار توزیع پیشین هم‌راستا باشند و از پراکندگی بیش از حد جلوگیری شود.

(ب) الگوریتم آموزش VAE:

Variational Autoencoder Training Algorithm

Input: Dataset $\mathcal{D} = \{X^{(i)}\}_{i=1}^N$, encoder $q_\phi(Z \mid X)$, decoder $p_\theta(X \mid Z)$, prior $p_\theta(Z)$, number of training epochs, batch size B

Output: Trained encoder parameters ϕ and decoder parameters θ

1. Initialize parameters ϕ, θ
2. For each epoch:
   - 2.1 Shuffle dataset and divide into mini-batches of size B
   - 2.2 For each mini-batch $\{X^{(i)}\}_{i=1}^B$:
     - 2.2.1 Compute $\mu_\phi(X^{(i)}), \sigma_\phi(X^{(i)})$ using encoder
     - 2.2.2 Sample $\epsilon^{(i)} \sim \mathcal{N}(0, I)$
     - 2.2.3 Compute latent variable: $Z^{(i)} = \mu + \sigma \cdot \epsilon$
     - 2.2.4 Reconstruct $X^{(i)}$ with decoder
     - 2.2.5 Evaluate $\log p_\theta(X^{(i)} \mid Z^{(i)})$, $\mathrm{KL}(q_\phi(Z \mid X^{(i)}) \| p(Z))$
     - 2.2.6 Compute total ELBO loss: $\mathcal{L}(X^{(i)}) = \log p_\theta(X^{(i)} \mid Z^{(i)}) - \mathrm{KL}(\cdot)$
3. Average loss and update parameters using gradient descent

$\triangleright$

پرسش ۲. Hierarchical VAE (۲۰ نمره)

یک مدل Hierarchical Variational Autoencoder (HVAE) با دو لایه را در نظر بگیرید که دارای متغیرهای نهان z_1 و z_2 است، به‌طوری‌که z_1 لایه‌ی پایین و z_2 لایه‌ی بالا می‌باشد. مدل مولد به صورت زیر تعریف می‌شود:

$$p(x, z_1, z_2) = p(x \mid z_1)\, p(z_1 \mid z_2)\, p(z_2)$$

تقریب پسین (Encoder) به صورت زیر تعریف می‌شود:

$$q(z_1, z_2 \mid x) = q(z_1 \mid x)\, q(z_2 \mid z_1)$$

(الف)

کران پایین شواهد (ELBO) را به صورت $L(x)$ برای این مدل HVAE استخراج کنید، ELBO به شکل امید ریاضی روی توزیع تقریبی $q(z_1 \mid x)q(z_2 \mid z_1)$ است که از ۳ بخش تشکیل شده است، لگاریتم درست‌نمایی و KL divergence بین توزیع‌های مربوطه.

(ب)

معنای هر یک از اجزای موجود در ELBO به‌دست آمده در قسمت (الف) را توضیح دهید و تأثیر آن‌ها را بررسی کنید.

(ج)

در Hierarchical VAEها، هدف این است که هر لایه‌ی نهان، اطلاعات مفیدی را کدگذاری کند که در بازسازی داده نقش داشته باشد. با این حال، در عمل اغلب مشاهده می‌شود که متغیرهای نهان لایه‌های بالایی مانند z_2 معمولاً توسط مدل نادیده گرفته می‌شوند. به این معنا که خروجی encoder برای z_2، یعنی $q(z_2 \mid z_1)$، تقریباً مستقل از ورودی x شده و با توزیع پیشین $p(z_2)$ هم‌راستا می‌شود.

- توضیح دهید که چرا این پدیده که به آن «فروریزش توزیع پسین» (Posterior Collapse) یا «متغیرهای نهان غیرفعال» (Inactice Latent Variables) گفته می‌شود، در زمان بهینه‌سازی ELBO رخ می‌دهد.
- این رفتار چه چیزی را در مورد جریان اطلاعات در ساختار سلسله‌مراتبی مدل نشان می‌دهد؟
- حداقل دو تکنیک (اعم از فرایند آموزش مدل یا تغییر در معماری مدل) پیشنهاد دهید که می‌توانند به جلوگیری از این مسئله کمک کنند و توضیح دهید چرا این تکنیک‌ها مؤثر هستند.

پاسخ.

(الف) استخراج ELBO:

$$\log p(x) = \log \int\int p(x, z_1, z_2)\, dz_1\, dz_2$$

حال از ترفند ضرب و تقسیم در توزیع پسین استفاده می‌کنیم:

$$= \log \int\int q(z_1, z_2 \mid x) \frac{p(x, z_1, z_2)}{q(z_1, z_2 \mid x)}\, dz_1\, dz_2$$

با استفاده از نامساوی Jensen داریم:

$$\log p(x) \geq \mathbb{E}_{q(z_1,z_2|x)}\left[\log \frac{p(x,z_1,z_2)}{q(z_1,z_2 \mid x)}\right]$$

که آن را ELBO یا $L(x)$ می‌نامیم.
اکنون با جایگذاری فرم‌های تجزیه‌شده‌ی توزیع‌ها:

$$p(x,z_1,z_2) = p(x \mid z_1)\,p(z_1 \mid z_2)\,p(z_2) \quad \text{و} \quad q(z_1,z_2 \mid x) = q(z_1 \mid x)\,q(z_2 \mid z_1)$$

داریم:

$$L(x) = \mathbb{E}_{q(z_1|x)\,q(z_2|z_1)}\left[\log \frac{p(x \mid z_1)\,p(z_1 \mid z_2)\,p(z_2)}{q(z_1 \mid x)\,q(z_2 \mid z_1)}\right]$$

حالا عبارت داخل لگاریتم را به صورت جمعی از لگاریتم‌ها می‌نویسیم:

$$= \mathbb{E}_{q(z_1|x)\,q(z_2|z_1)}\left[\log p(x \mid z_1) + \log p(z_1 \mid z_2) + \log p(z_2) - \log q(z_1 \mid x) - \log q(z_2 \mid z_1)\right]$$

در این مرحله، برخی امیدهای ریاضی را می‌توان ساده‌سازی کرد چون برخی توزیع‌ها مستقل از متغیر مورد انتظار هستند:

- چون $p(x \mid z_1)$ مستقل از z_2 است، می‌توان آن را از امید ریاضی نسبت به $q(z_2 \mid z_1)$ بیرون کشید.
- همچنین در حل این سوال $\mathbb{E}_{q(z_2|z_1)}\left[D_{\mathrm{KL}}\left(q(z_1 \mid x) \,\|\, p(z_1 \mid z_2)\right)\right] = \mathbb{E}_{q(z_1|x)q(z_2|z_1)}\left[\log \frac{q(z_1|x)}{q(z_1|z_2)}\right]$ فرض شده بود، که تساوی نادرستی است و به دلیل ترتیب نمونه‌گیری z_1 ، z_2 ممکن نیست. اما دقت شود که علی رغم این مورد کماکان این ترم رفتار KL مانند دارد و اگر دو توزیع نزدیک باشند ۰ می‌شود

پس می‌نویسیم:

$$L(x) = \mathbb{E}_{q(z_1|x)}\left[\log p(x \mid z_1)\right] - \mathbb{E}_{q(z_2|z_1)}\left[D_{\mathrm{KL}}\left(q(z_1 \mid x) \,\|\, p(z_1 \mid z_2)\right)\right] - \mathbb{E}_{q(z_1|x)}\left[D_{\mathrm{KL}}\left(q(z_2 \mid z_1) \,\|\, p(z_2)\right)\right]$$

(ب) تفسیر اجزای ELBO

- **امید لگاریتم درست‌نمایی (ترم بازسازی):**

$$\mathbb{E}_{q(z_1|x)}\left[\log p(x \mid z_1)\right]$$

این ترم میزان توانایی مدل را در بازسازی داده‌ی ورودی x از روی متغیر نهان z_1 اندازه‌گیری می‌کند. مقدار بالاتر آن نشان‌دهنده‌ی بازسازی بهتر و یادگیری ویژگی‌های اساسی داده است.

- **امید فاصله KL بین $q(z_1 \mid x)$ و $p(z_1 \mid z_2)$:**

$$\mathbb{E}_{q(z_2|z_1)}\left[D_{\mathrm{KL}}\left(q(z_1 \mid x) \,\|\, p(z_1 \mid z_2)\right)\right]$$

این ترم باعث منظم‌سازی توزیع پسین تقریب‌یافته برای z_1 می‌شود و آن را به توزیع پیشین شرطی $p(z_1 \mid z_2)$ نزدیک می‌کند. این کار به ساخت نمایشی منظم و معنادار در فضای نهان کمک می‌کند. به عبارتی دیگر باعث می‌شود فرآیند آموزش مدل Encoder و Decoder به نحوی پیش برود که متغیر z_1ای که با دیدن z_2 پیشنهاد می‌شود با z_1ای که بعد از دیدن X .پیشنهاد می‌شود نزدیک باشند

- **امید فاصله KL بین $q(z_2 \mid z_1)$ و $p(z_2)$:**

$$\mathbb{E}_{q(z_1|x)}\left[D_{\mathrm{KL}}\left(q(z_2 \mid z_1) \,\|\, p(z_2)\right)\right]$$

این ترم متغیر نهان لایه‌ی بالا z_2 را منظم می‌کند و آن را به توزیع پیشین $p(z_2)$ نزدیک می‌سازد. به عبارتی دیگر روی تمام داده‌ها و z_1 متناظرشان تلاش می‌کند تا توزیع z_2 نمونه گرفته شده به شرط z_1 را به توزیع پیشین آن نزدیک کند. این امر از بیش‌برازش جلوگیری کرده و باعث می‌شود z_2 ویژگی‌های انتزاعی‌تری از داده را ثبت کند.

(ج) بررسی پدیده Posterior Collapse

- این مسئله زمانی رخ می‌دهد که مدل در حین بهینه‌سازی ELBO تلاش می‌کند ترم KL زیر را کاهش دهد:

$$\mathbb{E}_{q(z_1|x)}\left[D_{\mathrm{KL}}(q(z_2 \mid z_1) \,\|\, p(z_2))\right]$$

برای به حداقل رساندن این مقدار، مدل تمایل دارد $q(z_2 \mid z_1) \approx p(z_2)$ شود، که در نتیجه متغیر نهان لایه‌ی بالا z_2 تقریباً مستقل از ورودی x می‌شود. اگر decoder بتواند ورودی را تنها با z_1 به‌خوبی بازسازی کند، استفاده از z_2 ضرورتی ندارد و در نتیجه posterior collapse یا متغیر نهان غیرفعال رخ می‌دهد.

- این رفتار نشان می‌دهد که جریان اطلاعات از داده‌ی ورودی x به لایه‌های بالاتر نهان منتقل نمی‌شود. در نتیجه z_2 نتوانسته ویژگی‌های انتزاعی و معناداری را از داده ثبت کند و نقش آن به متغیری شبیه نویز کاهش می‌یابد. این مسئله ساختار سلسله‌مراتبی مدل را بی‌اثر می‌کند.

- دو راهکار مؤثر برای مقابله با این پدیده عبارت‌اند از:

۱- KL Annealing: افزایش تدریجی وزن ترم KL در طول آموزش. این کار به مدل اجازه می‌دهد ابتدا روی بازسازی تمرکز کرده و سپس منظم‌سازی را اعمال کند.

۲- Free Bits: تعیین حداقل مقدار برای KL در هر واحد نهان (مثلاً ۰.۵ nats)، تا از صفر شدن آن جلوگیری کرده و اطمینان حاصل شود که هر متغیر نهان مقداری از اطلاعات را حمل می‌کند.

راه‌حل‌های دیگر شامل استفاده از توزیع‌های پیشین پیچیده‌تر یا افزودن ویژگی‌های معماری (مثل skip connections) برای حفظ نقش لایه‌های بالایی می‌باشد.

$\triangleright$

پرسش ۳. GAN (۲۰ نمره)

در این سوال، قصد داریم بین مسئله تشخیص داده واقعی و داده‌های تولید شده و فاصله توزیع تصاویر واقعی و تولید شده ارتباطی برقرار کنیم.
توزیع تصاویر واقعی را با P_d و توزیع تصاویر تولید شده را با P_g نشان می‌دهیم. همچنین توزیع P که داده‌ای تصادفی را به ما می‌دهد، به‌صورت زیر تعریف می‌شود:

$$P(x) = 0.5P_g(x) + 0.5P_d(x)$$

بنابراین، هر داده با احتمال مساوی از یکی از دو توزیع P_d یا P_g می‌آید.
حال، مجموعه‌ای از نمونه‌ها (x, y) تشکیل می‌دهیم، به این ترتیب که:

- x را از توزیع P نمونه‌گیری می‌کنیم.

- اگر x از توزیع واقعی P_d باشد، برچسب آن را $y = 1$ قرار می‌دهیم، و در غیر این صورت (اگر از P_g آمده باشد) $y = -1$.

اکنون، $\mathcal{D}$ را مجموعه تمامی تمایزدهنده‌ها در نظر می‌گیریم. هدف ما یافتن بهترین تمایزدهنده در این مجموعه است. فرض می‌کنیم این مجموعه هیچ محدودیتی ندارد و تمایزدهنده‌ها دارای ظرفیت نامحدود هستند. منظور از ظرفیت نامحدود این است به ازای هر تابع $y(x)$ دلخواه، تابعی در $\mathcal{D}$ با چنین رفتاری وجود دارد. بنابراین اگر نقطه بهینه تابع برای هر x را بیابید میتوانید از وجود تمایزدهنده‌ای با این اتخاذها مطمئن باشید. بهترین تمایزدهنده D در مجموعه $\mathcal{D}$ تابع هزینه زیر را کمینه می‌کند:

$$R_l(D) = \mathbb{E}_{(x,y)\sim P}[l(yD(x))]$$

در اینجا، هزینه به‌صورت امیدریاضی روی داده‌های (x, y) تعریف شده است. اکنون مقدار بهینه این تابع را تعریف می‌کنیم:

$$R_l(\mathcal{D}) = \inf_{D\in\mathcal{D}} R_l(D)$$

بخش اول

ابتدا امیدریاضی موجود در رابطه بالا را به‌صورت انتگرال روی توزیع x بازنویسی کنید. عبارت شما باید مشابه فرم زیر باشد:

$$\inf_{D\in\mathcal{D}} \left\{ \int \left([\mathsf{sth}]\ p_d(x)\ + [\mathsf{sth}]\ p_g(x) \right)\ dx \right\}$$

سپس با توجه به ظرفیت نامحدود مجموعه $\mathcal{D}$، بهینه عبارت بدست آمده را از روی بهینه نقطه‌ای حساب کنید. توجه کنید که ظرفیت نامحدود چه تاثیری روی اینفیمم دارد.
نهایتا با جاگذاری بهینه و جابه‌جایی عبارت‌ها نشان دهید که این مقدار برابر است با:

$$-\frac{1}{2} \int f \left(\frac{p_d(x)}{p_g(x)} \right) p_g(x) dx$$

توجه کنید باید تابع f را به دست آورید. پس از رسیدن به این رابطه، بررسی کنید که تابع f به‌دست‌آمده بر حسب ورودی خود، محدب و نزولی باشد.
ازین به بعد عبارت بالا را به استفاده از مفهوم واگرایی f به صورت

$$-\frac{1}{2}\mathbb{I}_f(\mathbb{P}_d, \mathbb{P}_\mathfrak{g})$$

نمایش می‌دهیم.

بخش دوم

در بخش قبل، نشان دادیم که این مسئله معادل کمینه کردن یک واگرایی بین دو توزیع است. اکنون می‌خواهیم بررسی کنیم که با انتخاب توابع هزینه مختلف $l(x)$، چه نوع واگرایی‌هایی حاصل می‌شوند.
برای هر تابع هزینه $l(x)$:

۱. ابتدا محدب بودن و یکنوایی تابع f متناظر را بررسی کنید.

۲. سپس، تمایزدهنده بهینه، تابع f، و واگرایی متناظر را استخراج کنید.

حالت‌های مورد بررسی:

- **الف)** $l(x) = \mathbb{I}(x \leq 0)$
- **ب)** $l(x) = (1 - x)^2$
- **ج)** $l(x) = \log(1 + e^{-x})$

واگرایی های متناظر توابع هزینه به صورت نامرتب در ادامه آمده‌اند. می‌توانید جهت بررسی درستی حل خود از این بخش استفاده کنید.

- $R(\mathcal{D}) = \frac{1}{2}(1 - \mathbb{I}_{\mathsf{TV}}(\mathbb{P}_d, \mathbb{P}g))$ که منظور از TV تابع Total Variation است.
- $R(\mathcal{D}) = \log 2 - \mathbb{I}_{\mathsf{JS}}(\mathbb{P}_d, \mathbb{P}_g)$ که منظور از JS تابع Jensen Shannon است.

- $R(\mathcal{D}) = 1 - \mathbb{I}_g(\mathbb{P}_d, \mathbb{P}_g)$ که منظور از g تابع $\frac{-4t}{t+1}$ است.

پاسخ.

بخش اول.

$$
\begin{aligned}
R_\ell(\mathcal{D}) = \inf_{D\in\mathcal{D}} R_\ell(D) = \inf_{D\in\mathcal{D}} \mathbb{E}_{\mathbb{P}_{x,y}}\left[\ell(yD(x))\right] = \inf_{D\in\mathcal{D}} \sum_{y=-1}^{1} \int \ell(yD(x))p(x,y)\,dx \\
= \inf_{D\in\mathcal{D}} \left\{ \int \ell(D(x))p(x,1)\,dx + \int \ell(-D(x))p(x,-1)\,dx \right\} \\
= \frac{1}{2}\inf_{D\in\mathcal{D}} \left\{ \int \ell(D(x))p(x \mid y=1)\,dx + \int \ell(-D(x))p(x \mid y=-1)\,dx \right\} \\
= \frac{1}{2}\inf_{D\in\mathcal{D}} \left\{ \int \ell(D(x))p_d(x)\,dx + \int \ell(-D(x))p_g(x)\,dx \right\} \\
= \frac{1}{2}\inf_{D\in\mathcal{D}} \left\{ \int \left[\ell(D(x))p_d(x) + \ell(-D(x))p_g(x)\right] dx \right\} \\
= \frac{1}{2}\inf_{D\in\mathcal{D}} \left\{ \int \left[\ell(D(x))\frac{p_d(x)}{p_g(x)} + \ell(-D(x))\right] p_g(x)\,dx \right\} \\
R_\ell(\mathcal{D}) = \frac{1}{2}\int \inf_{\alpha}\left[\ell(\alpha)\frac{p_d(x)}{p_g(x)} + \ell(-\alpha)\right] p_g(x)\,dx \\
f(t) = -\inf_{\alpha}\left[\ell(\alpha)t + \ell(-\alpha)\right] \\
R_\ell(\mathcal{D}) = -\frac{1}{2}\int f\left(\frac{p_d(x)}{p_g(x)}\right) p_g(x)\,dx = -\frac{1}{2} I_f(\mathbb{P}_d \| \mathbb{P}_g)
\end{aligned}
$$

بخش دوم

با جایگزاری به دست می‌آید. به تصاویر زیر توجه نمایید.

3.3.1 0-1 Loss

This loss has the form $\ell(\alpha) = \mathbb{I}[\alpha \le 0]$, where $\mathbb{I}$ is the indicator function. From Eq. (7), the optimal discriminator takes the form of $D^*(\boldsymbol{x}) = \mathrm{sign}(p_g(\boldsymbol{x}) - p_d(\boldsymbol{x}))$ and the general loss takes the following form:

$$
\begin{aligned}
R_{0-1}(\mathfrak{D}) &= \frac{1}{2}\int \min\{p_d(\boldsymbol{x}), p_g(\boldsymbol{x})\}\,d\boldsymbol{x} = \frac{1}{2}\int\left[\frac{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}{2} - \frac{|p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})|}{2}\right] d\boldsymbol{x} \\
&= \frac{1}{2}\left(1 - \mathbb{I}_{TV}(\mathbb{P}_d \| \mathbb{P}_g)\right)
\end{aligned}
$$

where $\mathbb{I}_{TV}$ specifies the total variance distance between two distributions.

شکل ۲

▷

3.3.4 Least Square Loss

This loss has the form $\ell(\alpha) = (1-\alpha)^2$. From Eq. (7), the optimal discriminator takes the form of $D^*(\boldsymbol{x}) = \frac{p_d(\boldsymbol{x}) - p_g(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}$ and the general loss takes the following form:

$$
\begin{aligned}
R_{\mathrm{sqr}}(\mathfrak{D}) &= \frac{1}{2}\int \frac{4p_d(\boldsymbol{x})\,p_g(\boldsymbol{x})}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})} d\boldsymbol{x} = \frac{1}{2}\left[2 - \int \frac{(p_d(\boldsymbol{x}) - p_g(\boldsymbol{x}))^2}{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})} d\boldsymbol{x}\right] \\
&= 1 - \mathbb{I}_{\mathrm{f}}(\mathbb{P}_d \| \mathbb{P}_g)
\end{aligned}
$$

where $f(t) = \frac{-4t}{t+1}$ with $t \geq 0$. In addition, this f-divergence is known as the *triangular discrimination distance.*

شکل ۳

3.3.5 Logistic Loss

This loss has the form $\ell(\alpha) = \log(1 + \exp(-\alpha))$. From Eq. (7), the optimal discriminator takes the form of $D^*(\boldsymbol{x}) = \log \frac{p_d(\boldsymbol{x})}{p_g(\boldsymbol{x})}$ and the general loss takes the following form:

$$
\begin{aligned}
R_{\mathrm{sqr}}(\mathfrak{D}) &= \frac{1}{2}\int \left[p_d(\boldsymbol{x}) \log \frac{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}{p_d(\boldsymbol{x})} + p_g(\boldsymbol{x}) \log \frac{p_d(\boldsymbol{x}) + p_g(\boldsymbol{x})}{p_g(\boldsymbol{x})}\right] d\boldsymbol{x} \\
&= \frac{1}{2}\left[2\log 2 - \mathbb{I}_{\mathrm{KL}}\left(\mathbb{P}_d \| \frac{\mathbb{P}_d + \mathbb{P}_g}{2}\right) - \mathbb{I}_{\mathrm{KL}}\left(\mathbb{P}_g \| \frac{\mathbb{P}_d + \mathbb{P}_g}{2}\right)\right] \\
&= \log 2 - \mathbb{I}_{\mathrm{JS}}(\mathbb{P}_d \| \mathbb{P}_g)
\end{aligned}
$$

where $\mathbb{I}_{\mathrm{JS}}$ specifies the Jensen-Shannon divergence, which is a f-divergence with $f(t) = -t\log\frac{t+1}{t} - \log(t+1)$, $t \geq 0$.

شکل ۴

پرسش ۴. یادگیری تابع امتیاز در Diffusion (۲۰ نمره)

در اسلایدها دیدیم که هدف اصلی ما یادگیری تابع امتیاز است. یک راه ساده برای رسیدن به این هدف، تعریف تابع هزینه‌ای بین مقدار حقیقی تابع امتیاز و خروجی شبکه عصبی است:

$$
l_1(\theta) = \mathbb{E}_{q(x)}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log q(x)\|_2^2\right]
$$

اما مشکل اینجاست که به $\nabla_x \log q(x)$ دسترسی نداریم.

در ادامه یاد گرفتیم که اگر مدل بتواند نویز را تخمین بزند، می‌توان از آن برای تخمین تابع امتیاز استفاده کرد. قبل از پرداختن به آن، رابطه‌ای ساده‌تر را بررسی می‌کنیم.

بخش (الف): ساده‌سازی تابع هزینه

در این بخش می‌خواهیم مقدار ناشناخته تابع امتیاز را از تابع هزینه حذف کنیم. با استفاده از انتگرال جز به جز و حذف بخش‌های مستقل از θ نشان دهید که:

$$
l_1(\theta) = \mathbb{E}_{q(x)}\left[\frac{1}{2}\|s_\theta(x)\|_2^2 + \mathrm{Tr}(\nabla_x s_\theta(x))\right] + C_1
$$

که C_1 مستقل از θ است. در اثبات خود فرض کنید که وقتی $x \to \infty$ داریم که $q(x)s_\theta(x) \to 0$. در پایان، توضیح دهید چرا این تابع هزینه در عمل ممکن است برای آموزش مناسب نباشد؟

بخش (ب): ارتباط با تخمین نویز

اکنون هدف ما این است که نشان دهیم تابع هزینه بالا معادل با تابع هزینه‌ای است که بر اساس تخمین نویز تعریف می‌شود:

$$l_3(\theta) = \mathbb{E}_{x\sim q(x),\ \epsilon\sim\mathcal{N}(0,\mathcal{I})}\left[\frac{1}{2}\left\|s_\theta(\underbrace{x+\sigma\epsilon}_{\tilde{x}}) + \frac{\epsilon}{\sigma}\right\|^2\right]$$

جفت داده سالم و نویزی را $(x, \tilde{x})$ می‌نامیم. طبق رابطه زیر، برای کرنل گاوسی داریم:

$$\frac{\partial \log q(\tilde{x}|x)}{\partial \tilde{x}} = \frac{1}{\sigma^2}(x - \tilde{x})$$

در نتیجه:

$$l_3(\theta) = \mathbb{E}_{q(x,\tilde{x})}\left[\frac{1}{2}\left\|s_\theta(\tilde{x}) - \frac{\partial \log q(\tilde{x}|x)}{\partial \tilde{x}}\right\|^2\right]$$

ثابت کنید که:

$$l_3 = l + C$$

که C مستقل از θ است. اگر نیاز به راهنمایی دارید، به پیوست مقاله زیر مراجعه کنید:
A Connection Between Score Matching and Denoising Autoencoders
اما سعی کنید مراحل اثبات را خودتان کامل بنویسید و توضیح دهید.

پاسخ.

الف

$$l_1(\theta) = \int q(x)\left(\frac{1}{2}\|s_\theta(x)\|^2 + \frac{1}{2}\|\nabla_x \log q(x)\|^2 + s_\theta(x)^T \nabla_x \log q(x)\right) dx$$

جمله دوم در لاس هست پس کاری نداریم. جمله بدون تتا هم که میتوان از لاس حذف کرد. پس فقط جمله کراس بین دو عبارت باقی می‌ماند.

$$-\int q(x) s_\theta(x)^T \nabla_x \log q(x) dx$$

حال عبارت ضرب داخلی را باز میکنیم پس برای هر اندیس داریم که:

$$-\int q(x)\frac{\partial \log q(x)}{\partial x_i} s_{\theta i}(x)\, dx = -\int \frac{q(x)}{q(x)}\frac{\partial q(x)}{\partial x_i} s_{\theta i}(x)\, dx = -\int \frac{\partial q(x)}{\partial x_i} s_{\theta i}(x)\, dx = \int q(x)\frac{\partial s_{\theta i}(x)}{\partial x_i} dx$$

آخرین بخش از انتگرال جز به جز حاصل شده است.
چراکه

$$\begin{aligned}
-\int \frac{\partial q(x)}{\partial x_1} s_{\theta 1}(x)\, dx &= -\int \left[\int \frac{\partial q(x)}{\partial x_1} s_{\theta 1}(x)\, dx_1\right] d(x_2, \ldots, x_n) \\
&= -\int \Bigg[\lim_{a\to\infty,\, b\to-\infty} \Big(q(a, x_2, \ldots, x_n) s_{\theta 1}(a, x_2, \ldots, x_n) \\
&\quad - q(b, x_2, \ldots, x_n) s_{\theta 1}(b, x_2, \ldots, x_n)\Big) \\
&\quad - \int \frac{\partial s_{\theta 1}(x)}{\partial x_1} q(x)\, dx_1\Bigg] d(x_2, \ldots, x_n).
\end{aligned}$$

که با جمع برای اندیس ها همان حکم مسئله حاصل می‌شود.
عدم استفاده به دلیل دشواری محاسبه تریس هسیان شبکه عصبی است. البته می‌توان از تقریب‌هایی برای تریس استفاده کرد که محاسبه ساده تر شود ولی در عمل دیگه روی این خیلی کار نشد.

ب

$$J_{1_q}(\theta) = \mathbb{E}_{q(\tilde{x})}\left[\left\|\frac{1}{2}s_\theta(\tilde{x}) - \frac{\partial \log q(\tilde{x})}{\partial \tilde{x}}\right\|^2\right]$$

$$J_{1_q}(\theta) = \mathbb{E}_{q(\tilde{x})}\left[\frac{1}{2}\|s_\theta(\tilde{x})\|^2\right] - S(\theta) + C_2$$

$$C_2 = \mathbb{E}_{q(\tilde{x})}\left[\frac{1}{2}\left\|\frac{\partial \log q(\tilde{x})}{\partial \tilde{x}}\right\|^2\right]$$

$$\begin{aligned}
S(\theta) &= \mathbb{E}_{q(\tilde{x})}\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x})}{\partial \tilde{x}}\right\rangle \\
&= \int q(\tilde{x})\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x})}{\partial \tilde{x}}\right\rangle d\tilde{x} \\
&= \int q(\tilde{x})\left\langle s_\theta(\tilde{x}), \frac{1}{q(\tilde{x})}\frac{\partial q(\tilde{x})}{\partial \tilde{x}}\right\rangle d\tilde{x} \\
&= \int \left\langle s_\theta(\tilde{x}), \frac{\partial q(\tilde{x})}{\partial \tilde{x}}\right\rangle d\tilde{x} \\
&= \int \left\langle s_\theta(\tilde{x}), \frac{\partial}{\partial \tilde{x}}\int q(x)q(\tilde{x} \mid x)dx\right\rangle d\tilde{x} \\
&= \int \left\langle s_\theta(\tilde{x}), \int q(x)\frac{\partial q(\tilde{x} \mid x)}{\partial \tilde{x}}dx\right\rangle d\tilde{x} \\
&= \int \left\langle s_\theta(\tilde{x}), \int q(x)q(\tilde{x} \mid x)\frac{\partial \log q(\tilde{x} \mid x)}{\partial \tilde{x}}dx\right\rangle d\tilde{x} \\
&= \int\int q(x)q(\tilde{x} \mid x)\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x} \mid x)}{\partial \tilde{x}}\right\rangle dx d\tilde{x} \\
&= \int\int q(\tilde{x}, x)\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x} \mid x)}{\partial \tilde{x}}\right\rangle dx d\tilde{x} \\
&= \mathbb{E}_{q(x,\tilde{x})}\left[\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x} \mid x)}{\partial \tilde{x}}\right\rangle\right]
\end{aligned}$$

$$\begin{aligned}
J_{1_q}(\theta) = \mathbb{E}_{q(\tilde{x})}&\left[\frac{1}{2}\|s_\theta(\tilde{x})\|^2\right] \\
&- \mathbb{E}_{q(x,\tilde{x})}\left[\left\langle s_\theta(\tilde{x}), \frac{\partial \log q(\tilde{x} \mid x)}{\partial \tilde{x}}\right\rangle\right] + C_2.
\end{aligned}$$

$$J_{۲_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(x,\tilde{x})}\left[\left\|\frac{1}{2}s_\theta(\tilde{x}) - \frac{\partial \log q_\sigma(\tilde{x} \mid x)}{\partial \tilde{x}}\right\|^2\right],$$

$$J_{۲_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x})}\left[\frac{1}{2}\|s_\theta(\tilde{x})\|^2\right] - \mathbb{E}_{q_\sigma(x,\tilde{x})}\left[\left\langle s_\theta(\tilde{x}), \frac{\partial \log q_\sigma(\tilde{x} \mid x)}{\partial \tilde{x}}\right\rangle\right] + C_3$$

$$C_3 = \mathbb{E}_{q_\sigma(x,\tilde{x})}\left[\frac{1}{2}\left\|\frac{\partial \log q_\sigma(\tilde{x} \mid x)}{\partial \tilde{x}}\right\|^2\right]$$

$$J_{۱_{q_\sigma}}(\theta) = J_{۲_{q_\sigma}}(\theta) + C_2 - C_3.$$

$\triangleright$

پرسش ۵. آیا فرض مارکوف در Diffusion الزامی است؟ (۲۰ نمره)

در ساختار فروارد دیفیوژن، فرض کردیم که:

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$

این فرض مارکوف باعث می‌شود که برای تولید، نیاز به انجام T مرحله به صورت ترتیبی داشته باشیم. برای درستی فرض گاوسی بودن می‌دانیم، T باید بزرگ باشد که باعث هزینه زمانی زیاد می‌شود. آیا می‌توانیم بدون فرض مارکوف بودن هم به همان توزیع حاشیه $q(x_t \mid x_0)$ برسیم؟
دقت کنید که در فرم ابجکتیومان یعنی

$$L(\epsilon_\theta) := \sum_{t=1}^{T} \gamma_t \, \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0), \, \boldsymbol{\epsilon}_t \sim \mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon_\theta^{(t)}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t\right) - \boldsymbol{\epsilon}_t\right\|_2^2\right]$$

تنها این حاشیه تاثیر دارد و توزیع مشترک x_i ها تاثیری ندارد.
در مدل استاندارد، این مارجین به صورت زیر است:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

بنابراین، می‌توان ساختاری غیرمارکوفی تعریف کرد که همین مارجین را تولید کند.
برای یک بردار $\sigma \in \mathbb{R}^T_{\geq 0}$ توزیع اینفرنس را به‌صورت زیر تعریف می‌کنیم:

$$q_\sigma(x_{1:T} \mid x_0) := q_\sigma(x_T \mid x_0) \prod_{t=2}^{T} q_\sigma(x_{t-1} \mid x_t, x_0)$$

که در آن:

$$q_\sigma(x_T \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_T}x_0, (1-\bar{\alpha}_T)\mathbf{I})$$

و برای $t > 1$:

$$q_\sigma(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}, \sigma_t^2\mathbf{I}\right)$$

الف) اثبات برابری مارجین‌ها
با استفاده از خواص توزیع گاوسی و به کمک استقرا ثابت کنید:

$$q_\sigma(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

توجه کنید که فرایند فروارد ما دیگر مارکوف نیست، زیرا x_t به x_0 نیز وابسته است:

$$q_\sigma(x_t \mid x_{t-1}, x_0) = \frac{q_\sigma(x_{t-1} \mid x_t, x_0)\, q_\sigma(x_t \mid x_0)}{q_\sigma(x_{t-1} \mid x_0)}$$

مشاهده جالب: اگر $\sigma_t \to 0$، واریانس گاوسی به صفر میل می‌کند و x_{t-1} به‌صورت قطعی از x_0 و x_t به دست می‌آید. بنابراین می‌توان فرآیند جنریشن را به یک فرآیند قطعی تبدیل کرد که تنها منبع تصادفی آن نویز اولیه x_T است.

این دیدگاه امکان نوعی تناظر بین فضای پنهان و خروجی را فراهم می‌کند، و دقت کنید که توزیع حاشیه‌ای که حاصل می‌شود همان توزیع مدل دیفیوژن تصادفی خواهد بود.

فرآیند جنریشن

ابتدا با استفاده از عبارت زیر که مشابه مدل استاندارد است، نمونه‌ای نویزی ساخته و بدون دسترسی به x_0 نویز آن را تخمین می‌زنیم:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

با مدل تخمین نویز می‌توانیم نمونه‌ی بدون نویز را به صورت زیر بازسازی کنیم:

$$f_\theta^{(t)}(x_t) := \frac{x_t - \sqrt{1-\bar{\alpha}_t}\cdot \epsilon_\theta^{(t)}(x_t)}{\sqrt{\bar{\alpha}_t}}$$

بنابراین ابتدا داریم که

$$p_\theta(x_T) = \mathcal{N}(0, \mathbf{I})$$

و پس از آن

$$p_\theta^{(t)}(x_{t-1} \mid x_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(x_1), \sigma_1^2\mathbf{I}) & t = 1 \\ q_\sigma(x_{t-1} \mid x_t, f_\theta^{(t)}(x_t)) & \text{غیر این صورت} \end{cases}$$

حال با این مدل جنریشن، آموزش به چه صورتی می‌شود؟ ابتدا دقت می‌کنیم که هدف آموزش مدل ما چه بود؟ مشابه قبل، ما به دنبال پارامترهایی هستیم که عبارت زیر اتخاذ شود.

$$\arg\max_\theta \mathbb{E}_{x_0 \sim q}[\log p_\theta(x_0)]$$

بنابراین می‌توان همان استدلال‌های مدل استاندارد را برای یافت تابع هزینه در این مدل نیز کرد.

ب) تطابق آبجکتیو با دیفیوژن کلاسیک

با تکرار مراحل اسلایدهای ۵۱ تا ۵۵ و تحلیل ELBO نشان دهید که آبجکتیو این مدل جنرتیو با مدل بررسی‌شده در اسلایدها یکسان است (به جز یک ثابت مستقل از θ). در برابری فرض کنید که در آبجکتیو مدل استاندارد تابع هزینه t های مختلف، ضریب برابری دارند.

پرسش: به نظر شما چگونه می‌توان از این نگاه برای کاهش زمان طولانی جنریشن استفاده کرد؟

پاسخ.

بخش الف

ابتدا یک ریزالت معروفی که توی بیشاپ و امثالهم هست رو مرور میکنیم.

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \mu, \Lambda^{-1})$$
$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

آنگاه

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mu + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\Lambda^{-1}\mathbf{A}^{\top})$$
$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \Sigma\{\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \Lambda\mu\}, \Sigma)$$

که

$$\Sigma = (\Lambda + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}.$$

حالا با استفاده از این داریم:

می‌توانیم حکم این بخش را با استفاده از استدلال استقرا برای t از T تا ۱ اثبات کنیم، چراکه حالت پایه یعنی $t = T$ برقرار است.

ابتدا داریم:

$$q_\sigma(x_{t-1} \mid x_0) := \int_{x_t} q_\sigma(x_t \mid x_0)\, q_\sigma(x_{t-1} \mid x_t, x_0)\, dx_t$$

و

$$q_\sigma(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I)$$

$$q_\sigma(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1 - \bar{\alpha}_t}},\ \sigma_t^2 I\right)$$

حالا طبق لمی که اول گفتم داریم که

$$\begin{aligned}\mu_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{\sqrt{\bar{\alpha}_t}x_0 - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1 - \bar{\alpha}_t}} \\ &= \sqrt{\bar{\alpha}_{t-1}}x_0\end{aligned}$$

و

$$\begin{aligned}\Sigma_{t-1} &= \sigma_t^2 I + \frac{1 - \bar{\alpha}_{t-1} - \sigma_t^2}{1 - \bar{\alpha}_t}(1 - \bar{\alpha}_t)I \\ &= (1 - \bar{\alpha}_{t-1})I\end{aligned}$$

در نتیجه:

$$q_\sigma(x_{t-1} \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}x_0, (1 - \bar{\alpha}_{t-1})I)$$

که این به ما اجازه می‌دهد استقرا را ادامه دهیم.

بخش ب

مشابه اسلاید البو را مینویسیم.

$$\mathcal{J}_\sigma = \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[\log q_\sigma(x_T \mid x_0) + \sum_{t=2}^{T}\log q_\sigma(x_{t-1}\mid x_t, x_0) - \sum_{t=1}^{T}\log p_\theta^{(t)}(x_{t-1}\mid x_t) - \log p_\theta(x_T)\right]$$

$$\begin{aligned}\mathcal{J}_\sigma &= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[\sum_{t=2}^{T}\log q_\sigma(x_{t-1}\mid x_t, x_0) - \sum_{t=1}^{T}\log p_\theta^{(t)}(x_{t-1}\mid x_t)\right]\\ &= \mathbb{E}_{x_{1:T}\sim q(x_{1:T}|x_0)}\left[\sum_{t=2}^{T}\log \frac{q_\sigma(x_{t-1}\mid x_t, x_0)}{p_\theta^{(t)}(x_{t-1}\mid x_t)} - \log p_\theta^{(0)}(x_0\mid x_1)\right]\end{aligned}$$

$$= \sum_{t=2}^{T}\int q(x_{t-1}, x_t\mid x_0)\log\frac{q_\sigma(x_{t-1}\mid x_t, x_0)}{p_\theta^{(t)}(x_{t-1}\mid x_t)}dx_{t-1}dx_t - \int q(x_1\mid x_0)\log p_\theta^{(0)}(x_0\mid x_1)dx_1$$

$$= \sum_{t=2}^{T}\int q(x_t\mid x_0)\int q(x_{t-1}\mid x_t, x_0)\log\frac{q_\sigma(x_{t-1}\mid x_t, x_0)}{p_\theta^{(t)}(x_{t-1}\mid x_t)}dx_{t-1}dx_t - \mathbb{E}_{x_1|x_0}\log p_\theta^{(0)}(x_0\mid x_1)$$

$$= \sum_{t=2}^{T}\mathbb{E}_{x_t|x_0}\left[D_{KL}\left(q_\sigma(x_{t-1}\mid x_t, x_0)\,\|\,p_\theta^{(t)}(x_{t-1}\mid x_t)\right)\right] - \mathbb{E}_{x_1|x_0}\log p_\theta^{(0)}(x_0\mid x_1)$$

با توجه به اینکه:

$$q_\sigma(x_{t-1}\mid x_t, x_0) = \mathcal{N}\left(\sqrt{\bar\alpha_{t-1}}x_0 + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t - \sqrt{\bar\alpha_t}x_0}{\sqrt{1-\bar\alpha_t}},\ \sigma_t^2 I\right)$$

و

$$p_\theta(x_{t-1}\mid x_t) = \mathcal{N}\left(\sqrt{\bar\alpha_{t-1}}f_\theta(x_t) + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t - \sqrt{\bar\alpha_t}f_\theta(x_t)}{\sqrt{1-\bar\alpha_t}},\ \sigma_t^2 I\right)$$

داریم که

واگرایی بین توزیع‌های $q(x_{t-1}\mid x_t, x_0)$ و $p_\theta^{(t)}(x_{t-1}\mid x_t)$ برابر است با:

$$\begin{aligned}D_{KL}(q(x_{t-1}\mid x_t, x_0)\,\|\,p_\theta^{(t)}(x_{t-1}\mid x_t)) &= \frac{1}{2\sigma_t^2}(\mu_q - \mu_p)^2\\ &= \frac{1}{2\sigma_t^2}\Bigg(\left(\sqrt{\bar\alpha_{t-1}}x_0 + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t-\sqrt{\bar\alpha_t}x_0}{\sqrt{1-\bar\alpha_t}}\right)\\ &\quad - \left(\sqrt{\bar\alpha_{t-1}}f_\theta(x_t) + \sqrt{1-\bar\alpha_{t-1}-\sigma_t^2}\cdot\frac{x_t-\sqrt{\bar\alpha_t}f_\theta(x_t)}{\sqrt{1-\bar\alpha_t}}\right)\Bigg)^2\end{aligned}$$

با ساده‌سازی:

$$
\begin{aligned}
&= \frac{1}{2\sigma_t^2}\left(\sqrt{\bar{\alpha}_{t-1}}(x_0 - f_\theta(x_t)) - \sqrt{\frac{1-\bar{\alpha}_{t-1}-\sigma_t^2}{1-\bar{\alpha}_t}}\cdot\sqrt{\bar{\alpha}_t}(x_0 - f_\theta(x_t))\right)^2 \\
&= \frac{1}{2\sigma_t^2}\left(\sqrt{\bar{\alpha}_{t-1}} - \sqrt{\frac{(1-\bar{\alpha}_{t-1}-\sigma_t^2)\bar{\alpha}_t}{1-\bar{\alpha}_t}}\right)^2 (x_0 - f_\theta(x_t))^2 \\
&= \frac{1}{2\sigma_t^2}\,\gamma_t (x_0 - f_\theta(x_t))^2
\end{aligned}
$$

با استفاده از دو تعریف:

$$
\epsilon_\theta^{(t)}(x_t) = \frac{x_t - \sqrt{\bar{\alpha}_t} f_\theta(x_t)}{\sqrt{1-\bar{\alpha}_t}} \quad \text{و} \quad f_\theta^{(t)}(x_t) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\bar{\alpha}_t}}
$$

پس تابع هزینه معادل زیر است که همان تابع هزینه دیفیوژن عادی است.

$$
\frac{1}{2\sigma_t^2}\,\gamma_t\,\frac{\left\|\epsilon_t - \epsilon_\theta^{(t)}(x_t)\right\|^2}{\bar{\alpha}_t}
$$

پرسش انتهایی صرفا نظر است. (بدون نمره) یک نمونه از این می شود روشی که پیپر دی دی آی ام معرفی کرده. ◁