بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۴-۰۳

مدرس: مهدیه سلیمانی

تمرین چهارم | ددلاین تمرین : ۹ خرداد

- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. مهلت تاخیر (مجاز و غیر مجاز) برای این تمرین، ۷ روز است (یعنی حداکثر تاریخ ارسال تمرین ۱۶ خرداد است).
- در هر کدام از سوالات، اگر از منابع خارجی استفاده کرده‌اید باید آن را ذکر کنید. در صورت همفکری با افراد دیگر هم باید نام ایشان را در سوال مورد نظر ذکر نمایید.
- پاسخ تمرین باید ماحصل دانسته‌های خود شما باشد. در صورت رعایت این موضوع، استفاده از ابزارهای هوش مصنوعی با ذکر نحوه و مصداق استفاده بلامانع است.
- پاسخ ارسالی واضح و خوانا باشد. در غیر این صورت ممکن است منجر به از دست دادن نمره شود.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد. به اسکرین‌شات از منابع یا پاسخ افراد دیگر نمره‌ای تعلق نمی‌گیرد.
- در صورتی که بخشی از سوال‌ها را جای دیگری آپلود کرده و لینک آن را قرار داده باشید، حتما باید تاریخ آپلود مشخص و قابل اتکا باشد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمارین تئوری در یک فایل pdf با نام HW4_[First-Name]_[Last-Name]_[Student-Id].pdf و تمارین عملی نیز در یک فایل مجزای زیپ با نام HW4_[First-Name]_[Last-Name]_[Student-Id].zip بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین : پرهام رضایی، امیرحسین حاجی‌محمدرضایی، علیرضا فرج تبریزی، 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 پیش‌فرض‌های 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)**: 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 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؟

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.

پرسش ۳. GAN (۲۰ نمره)

در این سوال، قصد داریم بین مسئله‌ی "تشخیص داده واقعی و داده تولید شده" و مسئله‌ی "فاصله توزیع تصاویر واقعی و تصاویر 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 که 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

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

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

راهنمایی: 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.

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

در اسلایدهای درس 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، می‌توانید به پیوست مقاله زیر مراجعه کنید:
A Connection Between Score Matching and Denoising Autoencoders
اما سعی کنید مراحل اثبات را خودتان کامل نوشته و توضیح دهید.

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 تابع هدف (Objective) نهایی، 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 $\sigma \in \mathbb{R}_{\geq 0}^{T}$ 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 تابع هدف (Objective) با دیفیوژن کلاسیک با تکرار مراحل ذکر شده در اسلایدهای درس و تحلیل ELBO نشان دهید که تابع هدف این مدل مولد با مدل بررسی‌شده در اسلایدها یکسان است (به جز یک ثابت مستقل از θ). در برابری فرض کنید که در تابع هدف 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؟