



یادگیری عمیق

نیم سال دوم ۰۵-۰۴
مدرس: مهدیه سلیمانی

ددلاین تمرین : ۹ تیر

تمرین چهارم

- 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 تیر 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.
- از تحویل تمرین نظری به شکل تایپ شده (فایل Word-Latex یا هر نوع تایپ شده دیگری) بپرهیزید و حتما پاسخ دست نویس تحویل دهید.
- 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 تمرین تئوری در یک فایل pdf با نام `HW4_[First-Name]_[Last-Name]_[Student-Id].pdf` بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوثرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین: رضا قربانی، مبینا پولائی، مزدک تیموریان، مهشید دهقانی، حسین شریفی، مریم ولی، امیرحسین ایزدی

بخش نظری (۱۰۰ نمره)

پرسش ۱. Hierarchical VAEs (۲۰ نمره)

Ladder Variational Autoencoders (LVAE) متعلق به خانواده Hierarchical VAEs هستند، یعنی مدل‌های مولد که در آن‌ها پشته‌ای از متغیرهای نهان تصادفی (stochastic latent variables) به صورت $z = (z_1, z_2, \dots, z_L)$ در یک Markov chain از سطح بالا چیده شده‌اند. توزیع توأم به صورت زیر بیان می‌شود:

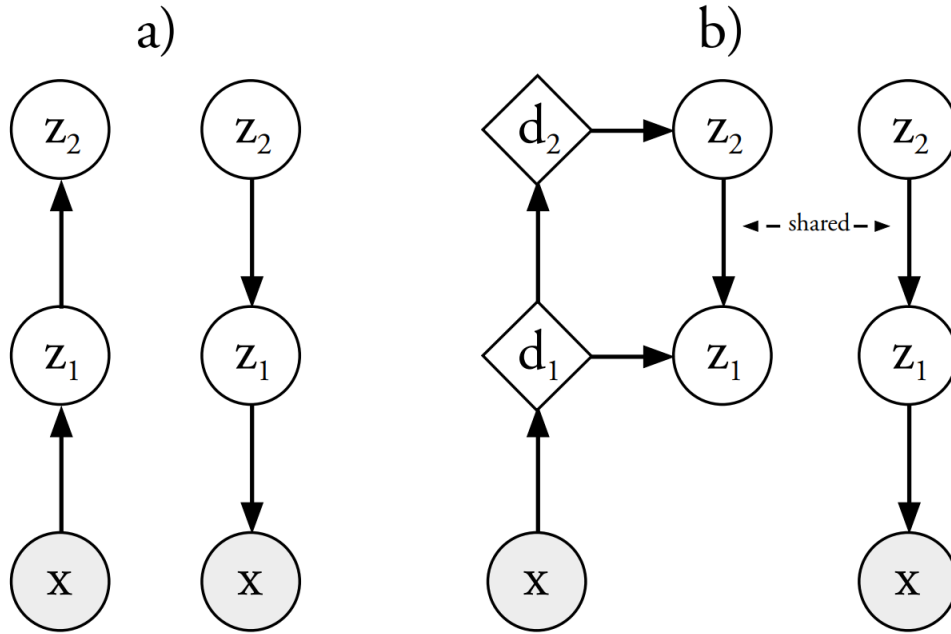
$$p_{\theta}(x, z) = p_{\theta}(x | z_1) p_{\theta}(z_L) \prod_{i=1}^{L-1} p_{\theta}(z_i | z_{i+1}).$$

در یک VAE استاندارد، مدل استنتاج (encoder) کاملاً پایین‌به‌بالا (bottom-up) است: با شروع از داده x ، یک شبکهٔ deterministic پارامترهای هر $q_{\phi}(z_i | z_{i-1})$ را تولید می‌کند. اما LVAE به جای آن از استنتاج بالا-به-پایین (top-down inference) استفاده می‌کند که اطلاعات را از دو منبع ادغام می‌نماید:

- یک approximate likelihood پایین-به-بالا $(\hat{\mu}_{q,i}, \hat{\sigma}_{q,i}^2)$ که توسط یک گذر deterministic به سمت بالا به دست می‌آید.

- prior بالا-به-پایین $(\mu_{p,i}, \sigma_{p,i}^2)$ که از ساختار سلسله‌مراتبی مدل مولد (generative model) می‌آید.

این تلفیق با یک ترکیب precision-weighted انجام می‌شود، دقیقاً همان‌طور که یک likelihood گاوسی و یک prior گاوسی را ترکیب می‌کنیم. شکل ۱ جریان اطلاعات را در یک VAE استاندارد و در LVAE مقایسه می‌کند.



شکل ۱: ساختار VAE (a) و LVAE (b)

در شکل ۱، دایره‌ها متغیرهای stochastic، لوزی‌ها متغیرهای deterministic هستند. در VAE استنتاج کاملاً bottom-up است؛ در LVAE استنتاج از همان وابستگی top-down مشابه مدل مولد استفاده می‌کند و به prior مولد اجازه می‌دهد تا به صورت بازگشتی توسط یک approximate likelihood وابسته به داده تصحیح شود.

(الف)

کران پایینی شواهد (evidence lower bound - ELBO) را برای مدل متغیر نهان سلسله‌مراتبی (hierarchical latent variable model) که در بالا توضیح داده شد، بنویسید. فرض کنید که posterior تقریبی factorised شده به صورت

$$q_{\phi}(z | x) = \prod_{i=1}^L q_{\phi}(z_i | z_{i+1}, x)$$

است (که در آن $z_{L+1} \equiv \emptyset$ تعریف می‌شود). جمله بازسازی (reconstruction term) و عبارت‌های منظم‌سازی (regularisation) مبتنی بر KL-divergence را مشخص کنید. چگونه عبارت منظم‌سازی، تعیین خودکار ارتباط (automatic relevance determination) واحدهای نهان را اعمال می‌کند؟

(ب)

در LVAE، posterior تقریبی لایه i به صورت یک Gaussian با میانگین $\mu_{q,i}$ و واریانس $\sigma_{q,i}^2$ ساخته می‌شود که به شکل زیر داده می‌شود:

$$\sigma_{q,i} = \frac{1}{\hat{\sigma}_{q,i}^{-2} + \sigma_{p,i}^{-2}}, \quad \mu_{q,i} = \frac{\hat{\mu}_{q,i} \hat{\sigma}_{q,i}^{-2} + \mu_{p,i} \sigma_{p,i}^{-2}}{\hat{\sigma}_{q,i}^{-2} + \sigma_{p,i}^{-2}}. \quad (19)$$

این فرمول‌ها را با ترکیب صریح یک Gaussian likelihood $\mathcal{N}(z_i | \hat{\mu}_{q,i}, \hat{\sigma}_{q,i}^2)$ با یک Gaussian prior $\mathcal{N}(z_i | \mu_{p,i}, \sigma_{p,i}^2)$ از طریق قواعد Bayes استخراج کنید. چرا می‌توان $\hat{\mu}_{q,i}, \hat{\sigma}_{q,i}^2$ را به عنوان پارامترهای یک approximate Gaussian likelihood تفسیر کرد؟

(ج)

مدل استنتاج LVAE مجدداً از همان ساختار top down مشابه مدل مولد استفاده می‌کند. توضیح دهید که چگونه پارامترهای استنتاج (به طور خاص $\mu_{p,i}, \sigma_{p,i}$) از مسیر مولد (generative path) به دست می‌آیند و چرا این اشتراک پارامتر صریح، مسئله بهینه‌سازی را نسبت به یک inference network purely bottom up که هیچ دانشی از توزیع مولد (generative distribution) ندارد، آسان‌تر می‌کند.

(د)

هدف آموزش (training objective) که در طول warm-up استفاده می‌شود به صورت زیر است:

$$\mathcal{L}_T(\theta, \phi; x) = -\beta KL(q_\phi(z | x) || p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x | z)], \quad (20)$$

که در آن β به طور خطی از ۰ به ۱ طی N_t دوره (epoch) افزایش می‌یابد.

(i) نشان دهید که برای $\beta = 0$ مدل به یک خودرمزگذار قطعی (deterministic autoencoder) استاندارد تبدیل می‌شود.

(ii) توضیح دهید چرا شروع با $\beta = 0$ از collapse واحدهای نهان (latent units) در مراحل اولیه آموزش جلوگیری می‌کند.

(iii) اگر β از ابتدا برابر ۱ قرار داده شود چه اتفاقی می‌افتد؟

پرسش ۲. از GAN تا فروپاشی مد: واگرایی‌ها، پایداری، و معیارهای کیفیت (۲۰ نمره)

پیش‌زمینه: مدل‌های مولد تخصصی (GAN) با معرفی گودفلو و همکاران (۲۰۱۴) انقلابی در یادگیری عمیق ایجاد کردند. اما به سرعت مشخص شد که GAN‌های استاندارد مبتنی بر واگرایی جنسن-شانون (JS) دچار یک نقص ساختاری هستند: **فروپاشی مد (mode collapse)** و **ناپدید شدن گرادیان**. در این پرسش، ریشه‌ی این مشکلات را در انتخاب واگرایی، سپس راه‌حل‌ها (واگرایی واترشتاین) و در نهایت معیارهای ارزیابی مدرن (precision/recall) بررسی می‌کنیم.

الف) واگرایی‌های f و رفتار بهینه‌ساز

ایده‌ی کلیدی f-GAN (نوزین و همکاران، ۲۰۱۶) این است که هر واگرایی f بین دو توزیع P و Q یک کران پایین تغییراتی از طریق مزدوج فنچل f^* دارد:

$$D_f(P||Q) \geq \sup_{T \in \mathcal{T}} \left[\mathbb{E}_{x \sim P} [T(x)] - \mathbb{E}_{x \sim Q} [f^*(T(x))] \right].$$

(آ) برای واگرایی KL با $f(u) = u \log u$ ، مزدوج فنچل $f^*(v)$ را محاسبه کنید و نشان دهید $D_f(P||Q) = D_{KL}(P||Q)$. سپس برای دو گاوسی هم‌واریانس $P = (\mu_1, \sigma^2 I)$ و $Q = (\mu_2, \sigma^2 I)$ ثابت کنید:

$$D_{KL}(P||Q) = \frac{\|\mu_1 - \mu_2\|^2}{2\sigma^2}.$$

(ب) چارچوب f-GAN نشان می‌دهد که انتخاب f تعیین می‌کند مدل «میانگین‌گرا» (mean-seeking) باشد یا «مدگرا» (mode-seeking). فرض کنید $P = \frac{1}{K} \sum_{k=1}^K (\mu_k, \sigma^2 I)$ یک مخلوط K -مدی و $Q = (\mu, \sigma^2 I)$ یک گاوسی تک‌مدی باشد.

(آ) کدام یک از واگرایی‌های $D_{KL}(P||Q)$ یا $D_{KL}(Q||P)$ باعث می‌شود Q در میانگین $\bar{\mu} = \frac{1}{K} \sum_k \mu_k$ متمرکز شود؟ μ را محاسبه کنید.

(ب) کدام یک باعث می‌شود Q به یکی از مدهای μ_k فروپاشد؟ چرا GAN استاندارد (مبتنی بر JS که میانگینی از دو جهت KL است) بیشتر تمایل به مدگرا دارد؟

(ج) جدول زیر از مقاله‌ی f-GAN (نوزین و همکاران، ۲۰۱۶، جداول ۱ و ۲) اقتباس شده است. با استفاده از این جدول توضیح دهید:

(الف) چرا برای واگرایی (JS) GAN دامنه‌ی f^* محدود به $t < 0$ است؟ (راهنمایی: با مشتق‌گیری از $f(u)$ در جدول، نشان دهید $T^*(x) = f'(p/q) = \log \frac{p}{p+q}$ و استدلال کنید چرا این مقدار همواره منفی است.) این محدودیت چه پیامدی برای پایداری آموزش دارد؟

(ب) برای واگرایی دلخواهی از جدول (غیر از KL)، با استفاده از $T^*(x) = f'(p/q)$ نشان دهید که چرا مقادیر $f'(u)$ در جدول با ستون «دامنه» سازگار است.

دامنه $\text{dom}(f^*)$	مزدوج $f^*(t)$	مولد $f(u)$	واگرایی
$\mathbb{R}$	$\exp(t-1)$	$u \log u$	KL
$\mathbb{R}^-$	$-1 - \log(-t)$	$-\log u$	KL معکوس
$\mathbb{R}$	$\frac{1}{4}t^2 + t$	$(u-1)^2$	پیرسن χ^2
$t < 0$	$-\log(1-e^t)$	$u \log u - (u+1) \log(u+1)$	GAN (JS)

(ب) تفکیک گر بهینه، اشباع JS، و واگرایی واترشتاین

(آ) (۳ نمره) برای GAN استاندارد با خطای آنتروپی متقاطع دودویی:

$$L_D =_{x \sim P} [\log D(x)] +_{x \sim Q} [\log(1 - D(x))],$$

نشان دهید تفکیک گر بهینه (برای Q ثابت) برابر است با:

$$D^*(x) = \frac{p(x)}{p(x) + q(x)}.$$

سپس ثابت کنید در بهینگی:

$$L_D^* = -\log 4 + 2 D_{JS}(P||Q).$$

چرا جمله‌ی $-\log 4$ برای درک اشباع (saturation) مهم است؟

(ب) فرض کنید $P = \delta_{x_1}$ و $Q = \delta_{x_2}$ با $x_1 \neq x_2$ دو دلتای دیراک باشند.

(آ) نشان دهید $D_{JS}(P||Q) = \log 2$ صرف‌نظر از $\|x_1 - x_2\|$.

(ب) نشان دهید فاصله‌ی واترشتاین-۱ برابر است با $W_1(P, Q) = \|x_1 - x_2\|$.

(ج) تعمیم دهید: اگر $\text{supp}(P) \cap \text{supp}(Q) = \emptyset$ ، چرا D_{JS} گرادیان صفر تولید می‌کند (اشباع) ولی W_1 گرادیان معنادار دارد؟

(ج) WGAN با استفاده از دوگان کانتورویچ-روبینشتاین، W_1 را تقریب می‌زند:

$$W_1(P, Q) = \sup_{\|f\|_L \leq 1} P[f(x)] - Q[f(x)].$$

نشان دهید برای $P = (\mu_P, I)$ و $Q = (\mu_Q, I)$ ، نگاشت حمل بهینه $T(x) = x + (\mu_P - \mu_Q)$ است و گرادیان به‌روزرسانی مولد در همه‌ی نقاط فضا یکسان و غیرصفر است. چرا این ویژگی آموزش WGAN را پایدارتر از GAN استاندارد می‌کند؟

(ج) دقت، پوشش، و چرا مدل‌های انتشار بر GAN غلبه کردند

(آ) دوازیوال و نیکول (۲۰۲۱) معیارهای دقت (precision: کیفیت نمونه‌ها) و پوشش (recall: تنوع مدهای پوشش داده‌شده) را معرفی کردند. با استفاده از ویژگی‌های واگرایی JS که در بخش‌های قبلی بررسی کردید، توضیح دهید:

(آ) چرا GAN‌ها معمولاً دقت بالا ولی پوشش پایین دارند.

(ب) چرا مدل‌های انتشار (Diffusion) معمولاً هر دو را متعادل نگه می‌دارند.

(ج) آیا این یک ویژگی ذاتی واگرایی JS است یا صرفاً یک مشکل بهینه‌سازی؟

(ب) یک شبکه‌ی GAN آموزش‌دیده روی ImageNet ممکن است نمونه‌های با کیفیتی از نژاد golden retriever تولید کند اما هرگز سگ نژاد basenji تولید نکند. مدل انتشار با همان تعداد پارامتر چگونه از این مشکل اجتناب می‌کند؟ یک استدلال سه‌مرحله‌ای ساختاریافته ارائه دهید:

(آ) GAN‌ها نسبت چگالی $p(x)/q(x)$ را تخمین می‌زنند — یک طبقه‌بندی که در نواحی کم‌چگالی اشباع می‌شود.

(ب) مدل‌های انتشار تابع امتیاز $\nabla \log p_t(x)$ را یاد می‌گیرند — یک رگرسیون که در همه‌جای فضا جهت رشد چگالی را نشان می‌دهد.

(ج) بنابراین وقتی مدل انتشار یک مد را از دست می‌دهد، امتیاز هنوز آن را به سمت مد داده هل می‌دهد، در حالی که گرادیان GAN صفر می‌شود.

(د) طراحی: انتخاب معیار و تابع هدف مناسب

(آ) دو سناریوی قبلی را کنار هم بگذارید تا یک مقایسه‌ی ساختاری بین W و D_{JS} انجام دهید:

(الف) برای دو گاوسی هم‌پوشان $P = (0, \sigma^2 I)$ و $Q_\theta = (\theta, \sigma^2 I)$ در بخش الف محاسبه کردید که $D_{KL}(P \| Q_\theta) = \frac{\|\theta\|^2}{2\sigma^2}$. نشان دهید $W_2^2(P, Q_\theta) = \|\theta\|^2$ دقیقاً و توضیح دهید چرا هر دو تابع «خوش‌رفتار» هستند وقتی توزیع‌ها هم‌پوشانی دارند (یعنی هم D_{KL} و هم W_2 به صورت $\propto \|\theta\|^2$ رشد می‌کنند و گرادیان غیرصفر در $\theta \neq 0$ دارند).

(ب) برای دو دلتای دیراک مجزا $P = \delta_{x_1}$ و $Q = \delta_{x_2}$ با $x_1 \neq x_2$ در بخش ب محاسبه کردید که $D_{JS}(P \| Q) = \log 2$ (ثابت، گرادیان صفر) و $W_1(P, Q) = \|x_1 - x_2\|$ (خطی، گرادیان معنادار). توضیح دهید چرا این تفاوت برای آموزش GAN روی داده‌های با پشتیبان کم‌بعد (low-dimensional manifold) حیاتی است و چگونه WGAN این مشکل را حل می‌کند.

(ج) با توجه به دو مورد بالا، چرا چشم‌انداز خطا (loss landscape) برای GAN مبتنی بر JS در نواحی دور از بهینگی «مسطح» (flat) است در حالی که برای WGAN شیب‌دار باقی می‌ماند؟

(ب) فرض کنید می‌خواهید یک سیستم تولید تصویر بسازید که باید در عین حال کیفیت بالا و پوشش کامل مدهای داده را داشته باشد. با توجه به آنچه در این پرسش آموختید:

(آ) اگر مجبور به استفاده از GAN باشید، چه تغییری در تابع هدف (واگرایی) پیشنهاد می‌دهید تا فروپاشی مد کاهش یابد؟

(ب) اگر محدودیتی در محاسبه نداشته باشید، چرا مدل انتشار گزینه‌ی بهتری است؟

(ج) برای ارزیابی مدل، چرا معیار FID به تنهایی کافی نیست و باید از Precision/Recall استفاده کرد؟

پرسش ۳. استخراج کران پایین شواهدی^۱ و تابع هدف DDPM (۲۰ نمره)

بخش اول

می‌دانیم در فرآیند روبه‌جلو DDPM (Forward) یک زنجیره مارکوفی داریم که به تدریج به تصویر ورودی نویز اضافه می‌کند:

$$q(x_{1:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}) \quad (1)$$

که در آن هر گام روبه‌جلو را یک توزیع گوسی در نظر می‌گیریم:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I) \quad (2)$$

الف) اما در فرآیند معکوس (Reverse) که به دنبال تخمین توزیع $q(x_{t-1}|x_t)$ هستیم، نمی‌توان به‌طور قطعی گفت که این توزیع نیز گوسی است. چرا؟ با در نظر گرفتن چه شرطی می‌توان این توزیع را نیز گوسی فرض کرد؟

ب) با اینکه تحت شرایط خاص، می‌توان فرض کرد فرآیند معکوس نیز از یک توزیع گوسی پیروی می‌کند، اما این شرط به تنهایی برای بدست آوردن این توزیع و نمونه‌برداری از آن کافی نیست. چرا؟ **راهنمایی:** اثبات کنید توزیع $q(x_{t-1}|x_t)$ Intractable است.

ج) با چه تغییری می‌شود یک توزیع tracktable ایجاد کرد؟ روابط مورد نیاز را نوشته و توضیح دهید.

بخش دوم

همانطور که در قسمت قبل دیدید، توزیع فرآیند معکوس DDPM باید به‌صورت رابطه زیر بازنویسی شود:

$$q(x_{t-1} | x_t, x_0) \quad (3)$$

الف) نشان دهید که این عبارت یک توزیع گاوسی با بردار میانگین $\tilde{\mu}_t(x_t, x_0)$ و ماتریس کوواریانس $\tilde{\beta}_t \mathbf{I}$ است که روابط صریح آن‌ها به صورت زیر تعریف می‌شوند:

$$\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0$$

^۱Evidence Lower Bound (ELBO)

به صورت کامل و گام به گام مسیر رسیدن به این روابط صریح را به دست آورید.

ب) در ادامه از کران پایین شواهدی (ELBO) برای $\log p_\theta(x_0)$ شروع کنید و نشان دهید که تابع هدف آموزش^۲ به صورت مجموعی از جمله‌های واگرایی KL^۳ تجزیه می‌شود. سپس ثابت کنید که وقتی واریانس فرایند معکوس برابر با β_t در نظر گرفته شود، کمینه کردن این واگرایی KL معادل با کمینه کردن خطای مجذور میانگین وزنی بین $\mu_\theta(x_t, t)$ و $\tilde{\mu}_t(x_t, x_0)$ خواهد بود. هر یک از جملات عبارت به دست آمده را تحلیل کنید و نقش آن را در تابع هدف آموزش توضیح دهید

ج) فرض کنید μ_θ از طریق یک noise predictor به صورت $\varepsilon_\theta(x_t, t)$ به شکل زیر پارامتریزه شده است:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \varepsilon_\theta(x_t, t) \right)$$

نشان دهید که تابع هدف KL در هر گام به صورت زیر درمی‌آیند:

$$\mathbb{E}_{x_0, \varepsilon, t} \left[\lambda_t \|\varepsilon - \varepsilon_\theta(x_t, t)\|^2 \right] \quad (۴)$$

به طور دقیق ضریب وابسته به گام زمانی λ_t را استخراج کنید.

د) همان طور که در قسمت قبل نشان دادید، تابع هدف اصلی که از کران پایین شواهدی (ELBO) به دست می‌آید، دارای یک ضریب وزن دهی وابسته به زمان است. اما نویسندگان مقاله اصلی DDPM نشان دادند که استفاده از یک تابع هدف ساده شده که این ضریب در آن نادیده گرفته شده و معادل ۱ فرض می‌شود ($\lambda_t = 1$)، نتایج بهتری به همراه دارد:

$$L_{simple} = \mathbb{E}_{x_0, \varepsilon, t} \left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2 \right] \quad (۵)$$

با توجه به رابطه ریاضی ضریب λ_t که در قسمت قبل استخراج کردید، به سوالات زیر پاسخ دهید:

۱. قرار دادن $\lambda_t = 1$ در تابع هدف ساده شده، چگونه تمرکز تابع هزینه را در گام‌های زمانی مختلف (مراحل با نویز کم یا t کوچک در برابر مراحل با نویز زیاد یا t بزرگ) نسبت به تابع هدف اصلی (ELBO) تغییر می‌دهد؟
۲. با توجه به مفهوم مراحل مختلف نویزگذاری در شکل‌گیری تصاویر، توضیح دهید که چرا این ساده‌سازی و تغییر تمرکز، در عمل منجر به تولید نمونه‌هایی با کیفیت بهتر توسط شبکه عصبی می‌شود؟

پرسش ۴. از تابع امتیاز تا انتشار: رگرسیون، شار احتمال، و نمونه برداری سریع (۲۰ نمره)

پیش زمینه: پرسش پیشین نشان داد که GANها به دلیل طبقه بندی تخاصمی مستعد فروپاشی مد هستند. چارچوب جایگزین — **تطبیق تابع امتیاز** (Score Matching) و **مدل های انتشار** (Diffusion) — از رگرسیون به جای طبقه بندی استفاده می‌کند. این پرسش نشان می‌دهد چرا این تغییر ظاهراً ساده، ریشه‌ی مشکل فروپاشی مد را حذف می‌کند، و چگونه می‌توان نمونه برداری از انتشار را تسریع بخشید.

Objective Function^۲
Kullback–Leibler Divergence^۳

الف) تابع امتیاز، فرمول تویدی، و پایداری در نواحی کم چگالی

(آ) وینسنت (۲۰۱۱) نشان داد که **تطبیق تابع امتیاز با نویز** (Denoising Score Matching) معادل آموزش یک بازسازی‌کننده‌ی نویز است. فرض کنید $x = x_0 + \sigma\epsilon$ با $x_0 \sim p$ و $\epsilon \sim (0, I)$. با استفاده از فرمول تویدی:

$$[x_0 | x] = x + \sigma^2 \nabla \log p(x),$$

نشان دهید تابع امتیاز بهینه برابر است با:

$$s^*(x) = \nabla \log p(x) = \frac{[x_0 | x] - x}{\sigma^2}.$$

این یعنی تابع امتیاز تنها با یک تفاضل و یک مقیاس‌دهی از بازسازی‌کننده‌ی نویز به دست می‌آید.

(ب) هدف آموزش بازسازی‌کننده‌ی نویز $\min_{\theta} \mathbb{E}_{x_0, \epsilon} [\|D_{\theta}(x_0 + \sigma\epsilon) - x_0\|^2]$ یک **مسئله‌ی رگرسیون** است. مقایسه کنید با هدف تفکیک‌گر GAN:

$$\max_D [\log D(x)] + \mathbb{E}_Q [\log(1 - D(x))],$$

که یک **مسئله‌ی طبقه‌بندی** است. چرا زمانی که $D^*(x) \approx 1$ (تفکیک‌گر بسیار مطمئن)، گرادیان طبقه‌بندی ناپدید می‌شود، در حالی که حتی اگر بازسازی‌کننده به طور کامل داده را بازسازی کند، گرادیان رگرسیون همچنان معنادار است؟

(ج) فرض کنید مولد در ناحیه‌ای از فضا قرار دارد که داده‌ی آموزشی تقریباً وجود ندارد ($p(x) \approx 0$). نشان دهید که برای تابع امتیاز، حتی در این نواحی $\nabla \log p(x)$ همچنان **جهت رشد چگالی** را نشان می‌دهد — مانند یک **قطب‌نما** به سمت داده. در مقابل، واگرایی JS در این ناحیه گرادیان صفر تولید می‌کند. چرا تابع امتیاز «همه‌جا زنده» است ولی تفکیک‌گر GAN «در بیابان می‌میرد»؟

ب) از ELBO تا پیش‌بینی نویز: مشتق‌گیری ساختاریافته

مدل انتشار یک مدل متغیر پنهان سلسله‌مراتبی است که فرآیند پیشروی (افزودن نویز) ثابت و فرآیند وارونه (حذف نویز) آموزش‌دینی است. هدف، بیشینه‌سازی درست‌نمایی حاشیه‌ای $p_{\theta}(x_0)$ است که از طریق یک کران پایین تغییراتی (ELBO) بهینه می‌شود.

(آ) فرآیند پیشروی در گام t به صورت $q(x_t | x_{t-1}) = (x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$ تعریف می‌شود. با استفاده از بازنویسی $\alpha_t := 1 - \beta_t$ و $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$ ، نشان دهید چگالی حاشیه‌ی شرطی بر حسب x_0 برابر است با:

$$q(x_t | x_0) = (x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I).$$

(ب) با استفاده از قاعده‌ی بیز، چگالی وارونه‌ی شرطی $q(x_{t-1} | x_t, x_0)$ نیز گاوسی است:

$$q(x_{t-1} | x_t, x_0) = (x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I),$$

که در آن:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t, \quad \tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t.$$

مدل وارونه‌ی آموزش‌دینی $p_{\theta}(x_{t-1} | x_t) = (x_{t-1}; \mu_{\theta}(x_t, t), \sigma_t^2 I)$ را در نظر بگیرید. نشان دهید هزینه‌ی $\mathcal{L}_{t-1} = \mathbb{E}_{x_0, \epsilon} [D_{\text{KL}}(q(x_{t-1} | x_t, x_0) || p_{\theta}(x_{t-1} | x_t))]$ معادل است با:

$$\mathcal{L}_{t-1} = \mathbb{E}_{x_0, \epsilon} \left[\frac{1}{2\sigma_t^2} \|\tilde{\mu}_t(x_t(x_0, \epsilon), x_0) - \mu_{\theta}(x_t(x_0, \epsilon), t)\|^2 \right] + C.$$

(ج) حال با استفاده از پارامتری سازی مجدد $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ و انتخاب:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right),$$

نشان دهید هدف آموزشی ساده شده به فرم زیر درمی آید:

$$\mathcal{L}_{\text{simple}} =_{x_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t) \right\|^2 \right].$$

چرا این هدف ساده تر از پیش بینی مستقیم x_0 یا میانگین $\bar{\mu}_t$ است؟ (راهنمایی: به مقیاس نویز و پایداری گرادیان فکر کنید.)

(ج) DDIM و نمونه برداری غیرمارکفی سریع

یک نقص بزرگ DDPM نیاز به $T = 1000$ گام تصادفی برای نمونه برداری است. سانگ و همکاران (۲۰۲۱) با معرفی DDIM نشان دادند که می توان همان مدل آموزش دیده را با گام های بسیار کمتر و به صورت **قطعی** نمونه برد.

(آ) در DDPM، فرآیند پیشروی یک زنجیره ی مارکفی است و فرآیند وارونه نیز برای تقریب از مارکوف استفاده می کند:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, \quad z \sim (0, I).$$

DDIM یک خانواده از فرآیندهای **غیرمارکفی** تعریف می کند که همان **حاشیه های** $q(x_t | x_0)$ را حفظ می کنند. معادله ی نمونه برداری DDIM عبارت است از:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(x_t, t) + \sigma_t z,$$

که در آن $\hat{x}_0(x_t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}$ برای حالت خاص $\sigma_t = 0$:

(آ) نشان دهید نمونه برداری **قطعی** (deterministic) می شود.

(ب) توضیح دهید چرا با وجود غیرمارکفی بودن فرآیند، **تابع هدف آموزش** تغییر نمی کند. (راهنمایی: $\mathcal{L}_{\text{simple}}$ فقط به حاشیه ی $q(x_t | x_0)$ وابسته است، نه به شرطی کامل $q(x_{1:T} | x_0)$.)

(ب) سانگ و همکاران (۲۰۲۱) در چارچوب Score-Based Generative Modeling through SDEs نشان دادند که فرآیند پیشروی انتشار را می توان به صورت یک معادله ی دیفرانسیل تصادفی (SDE) نوشت:

$$dx = f(x, t) dt + g(t) dW_t.$$

چگالی حاشیه ای $p_t(x)$ در معادله ی فوکر-پلانک صدق می کند. وارون زمانی این SDE عبارت است از:

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)] dt + g(t) d\bar{W}_t.$$

اگر به جای نویز تصادفی، مسیر **ODE** **شار احتمال** (Probability Flow ODE) را دنبال کنیم، چگونه عبارت بالا ساده می شود؟ چرا ODE (که قطعی است) برای محاسبه ی **درست نمایی** مناسب تر است، در حالی که SDE برای **نمونه برداری** انعطاف پذیرتر است؟

(د) جمع‌بندی: چرا رگرسیون از طبقه‌بندی بهتر است

(آ) یک استدلال سه‌مرحله‌ای ساختاریافته ارائه دهید که چرا مدل‌های انتشار از فروپاشی مد اجتناب می‌کنند:

(آ) GAN ها نسبت چگالی $p(x)/q(x)$ را تخمین می‌زنند — یک طبقه‌بندی که در نواحی کم‌چگالی اشباع می‌شود.

(ب) مدل‌های انتشار تابع امتیاز $\nabla \log p_t(x)$ را یاد می‌گیرند — یک رگرسیون که در همه‌جای فضا جهت رشد چگالی را نشان می‌دهد.

(ج) بنابراین وقتی مدل انتشار یک مد را از دست می‌دهد، امتیاز هنوز آن را به سمت مد داده هل می‌دهد، در حالی که گرادین GAN صفر می‌شود.

این استدلال را با مثال مخلوط دو گاوسی $P = \frac{1}{2}(-\mu, \sigma^2) + \frac{1}{2}(\mu, \sigma^2)$ و یک مولد تک‌مدی $Q_\theta = (\theta, \sigma^2)$ تشریح کنید.

(ب) **طراحی:** فرض کنید می‌خواهید از یک مدل انتشار برای تولید تصویر استفاده کنید. برای هر یک از سناریوهای زیر توضیح دهید کدام روش نمونه‌برداری (DDPM تصادفی در مقابل DDIM قطعی) و چه تعداد گام مناسب است:

(آ) تنوع بالا با اولویت اول (مثلاً برای augmentation).

(ب) کیفیت بسیار بالا با زمان محاسبه نامحدود (مثلاً تولید اثر هنری نهایی).

(ج) تعادل کیفیت-سرعت (مثلاً نمونه‌برداری تعاملی).

پرسش ۵. نمونه‌برداری، هدایت شرطی، و کنترل در مدل‌های انتشار (۲۰ نمره)

پیش‌زمینه: پرسش پیشین نشان داد که مدل‌های انتشار با پیش‌بینی نویز، تابع امتیاز $\nabla \log p_t(x)$ را یاد می‌گیرند. اما چگونه از این تابع امتیاز برای نمونه‌برداری استفاده کنیم؟ همچنین، چگونه می‌توان اطلاعات شرطی (مثل کلاس تصویر) را در تولید داده کنترل کرد؟ این پرسش سه موضوع را بررسی می‌کند: (۱) نمونه‌برداری Langevin، (۲) هدایت با دسته‌بند (Classifier Guidance)، و (۳) هدایت بدون دسته‌بند (Classifier-Free Guidance).

الف) نمونه‌برداری Langevin و میانگین متحرک نویز

(آ) برای نمونه‌برداری از یک توزیع $p(x)$ با استفاده از تابع امتیاز $\nabla \log p(x)$ ، از Stochastic Gradient Langevin Dynamics (SGLD) استفاده می‌کنیم:

$$x_{k+1} = x_k + \eta \nabla \log p(x_k) + \sqrt{2\eta} z_k, \quad z_k \sim (0, I).$$

نشان دهید که در حد $\eta \rightarrow 0$ و با تعداد گام کافی، توزیع x_k به $p(x)$ همگرا می‌شود. (راهنمایی: این یک فرآیند مارکوف با تابع پتانسیل $-\log p(x)$ است که با معادله‌ی فوکر-پلانک توصیف می‌شود.)

(ب) در عمل، تابع امتیاز در نواحی با چگالی کم دقت کافی ندارد — چون در این نواحی داده‌ی کمی برای یادگیری وجود دارد. راه‌حل: **افزودن نویز گاوسی** در چندین مقیاس. فرض کنید نویز $\epsilon \sim (0, \sigma^2 I)$ اضافه می‌کنیم و تابع امتیاز را روی داده‌ی نویزی تخمین می‌زنیم: $x_\sigma = x_0 + \sigma \epsilon$. نشان دهید:

$$\nabla_{x_\sigma} \log p_\sigma(x_\sigma) \approx -\frac{\epsilon_\theta(x_\sigma)}{\sigma},$$

که ϵ_θ شبکه‌ی پیش‌بینی نویز است. چرا با افزایش σ ، ناحیه‌ی بیشتری از فضا پوشش داده می‌شود؟

(ج) در **Langevin** با میانگین متحرک (Annealed Langevin Dynamics)، ابتدا با نویز زیاد شروع می‌کنیم و به تدریج نویز را کم می‌کنیم. توضیح دهید چرا این روش بهتر از Langevin با نویز ثابت کار می‌کند. چرا باید از نویز زیاد به نویز کم حرکت کنیم (و نه برعکس)؟

(ب) هدایت با دسته‌بند: کنترل از طریق گرادیان

(آ) فرض کنید یک دسته‌بند $p(c | x)$ آموزش دیده داریم. با استفاده از قاعده‌ی بیز، نشان دهید تابع امتیاز شرطی برابر است با:

$$\nabla_x \log p(x | c) = \nabla_x \log p(x) + \nabla_x \log p(c | x).$$

این نتیجه را با تصویر ساده توضیح دهید: اگر $\nabla_x \log p(x)$ جهت «بهبود کیفیت تصویر» باشد و $\nabla_x \log p(c | x)$ جهت «شبیه‌تر شدن به کلاس c »، ترکیب آن‌ها چه جهتی می‌دهد؟

(ب) در هدایت با دسته‌بند (Classifier Guidance)، نمونه‌برداری با اضافه کردن گرادیان دسته‌بند به تابع امتیاز انجام می‌شود:

$$\hat{\nabla} \log p(x_t | c) = \nabla \log p(x_t) + w \cdot \nabla_{x_t} \log p(c | x_t),$$

که w مقیاس هدایت است. وقتی w افزایش می‌یابد، چه اتفاقی برای کیفیت و تنوع نمونه‌ها می‌افتد؟ بررسی کنید: $w \rightarrow \infty$ چه پیامدی دارد؟

(ج) مشکل اصلی هدایت با دسته‌بند چیست؟ (راهنمایی: دسته‌بند روی تصاویر تمیز آموزش دیده، ولی در طول نمونه‌برداری، تصاویر نویزی هستند.) چرا این ناسازگاری می‌تواند باعث **نمونه‌های خصمانه** (adversarial) شود؟

(ج) هدایت بدون دسته‌بند: هوشمندی در سادگی

(آ) برای حذف نیاز به دسته‌بند جداگانه، از Classifier-Free Guidance (CFG) استفاده می‌کنیم. فرض کنید یک مدل انتشار شرطی $\epsilon_\theta(x_t, t, c)$ آموزش می‌دهیم که 10–20% مواقع شرط c را حذف می‌کند (جایگزین با یک توکن خالی $\emptyset$). نشان دهید که با استفاده از ترکیب خطی:

$$\hat{\epsilon}(x_t, t, c) = \epsilon_\theta(x_t, t, \emptyset) + w \cdot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset)),$$

در فضای تابع امتیاز معادل حرکت در جهت $\nabla_{x_t} \log p(c | x_t)$ هستیم. (راهنمایی: از رابطه‌ی بیز استفاده کنید و $\nabla \log p(x_t | c) - \nabla \log p(x_t)$ را محاسبه کنید.)

(ب) نشان دهید وقتی $w \rightarrow \infty$ ، نمونه‌برداری $\hat{\epsilon}$ به $p(c | x_t)$ میل می‌کند (نه نمونه‌برداری تصادفی از توزیع شرطی). این یعنی CFG با w بزرگ دچار **پختگی بیش از حد** (overcooked) می‌شود: تصاویر شبیه کلاس هستند ولی تنوع کم است. سپس بحث کنید:

(آ) چرا CFG یک «هک» هوشمندانه است؟ (پاسخ: چون مدل مولد را مجبور می‌کنیم هم شرطی و هم غیرشرطی یاد بگیرد.)

(ب) چرا CFG در عمل بهتر از هدایت با دسته‌بند کار می‌کند؟ (دو دلیل: نیاز به دسته‌بند جداگانه ندارد، و گرادیان‌های خصمانه تولید نمی‌کند.)

(د) مقایسه‌ی روش‌های هدایت و طراحی سیستم

(آ) جدول زیر را تکمیل کنید و ذیل جدول برای هر ویژگی یک جمله توضیح دهید:

Classifier-Free Guidance Classifier Guidance ویژگی

.....		نیاز به دسته‌بند جداگانه؟
.....		دسته‌بند باید روی تصاویر نویزی آموزش ببیند؟
.....		تعداد مدل‌های مورد نیاز
.....		خطر گرادیان خصمانه؟

(ب) فرض کنید می‌خواهید یک سیستم تولید تصویر با کنترل شرطی بسازید:

(آ) اگر کیفیت بالا با قیمت تنوع کم مطلوب باشد (مثل تولید تصویر دقیق از یک کلاس)، w را چقدر انتخاب می‌کنید و چرا؟

(ب) اگر تنوع بالا با کیفیت متوسط مطلوب باشد (مثل تولید تصاویر متنوع برای w augmentation)، چقدر انتخاب می‌کنید و چرا؟

(ج) اگر هم کیفیت و هم تنوع مهم باشد، چه راهکاری پیشنهاد می‌دهید؟ (راهنمایی: به ترکیب نمونه‌برداری DDIM و DDPM فکر کنید.)