یادگیری ژرف

نیم‌سال دوم ۴۰۲-۱۴۰۱

به نام خدا

مدرس: دکتر مهدیه سلیمانی

دانشکده‌ی مهندسی کامپیوتر

تمرین سری چهارم (۲۵+۲۹۰ نمره) | **مدل‌های مولد** | زمان تحویل: ۲۹ اردیبهشت (نظری)، ۳۱ اردیبهشت (عملی)

لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- 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 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 اعتنا باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. مهلت تاخیر (مجاز و غیر مجاز) برای این تمرین، ۱۰ روز است.

سوال ۱: (نظری) خودکدگذار وردشی (۴۰ نمره)

می‌دانیم که تابع هدف خودکدگذار وردشی[1] (VAE) به صورت زیر است.

$$\mathcal{L}_{ELBO}(\phi,\theta) = \mathbb{E}_{x'\sim p_{\text{Data}}}\left[\mathbb{E}_{z'\sim q_\phi(z|x')}\left[\log p_\theta\left(x' \mid z'\right)\right] - KL\left(q_\phi\left(z \mid x'\right)\|p(z)\right)\right]$$
$$\approx \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{z'\sim q_\phi\left(z|x^{(i)}\right)}\left[\log p_\theta\left(x^{(i)} \mid z'\right)\right] - KL\left(q_\phi\left(z \mid x^{(i)}\right)\|p(z)\right)\right)$$

که در آن توزیع prior به شکل توزیع چند متغیره‌ی گاوسی استاندارد $\mathrm{N}(0,\mathrm{I})$ است. همان‌طور که واضح است هدف بیشینه کردن قسمت مربوط به reconstruction این تابع است.

آ) فرض کنید توزیع خروجی کدگشا گاوسی چند متغیره باشد. در این حالت عبارت reconstruct را به صورت دقیق‌تر بازنویسی کنید.

ب) این بار فرض کنید خروجی کدگشا به صورت باینری است و توزیع آن را برنولی چند متغیره در نظر بگیرید. در این حالت نیز عبارت reconstruct را به صورت دقیق‌تر بازنویسی کنید.

ج) با توجه به دو قسمت قبل توضیح دهید که انتخاب توزیع خروجی کدگشا چگونه بر شکل و ماهیت تابع هدف و فرایند بهینه‌سازی در VAE تأثیر می گذارد؟

د) توضیح دهید که پدیده‌ی بیش برازش در VAE ها به چه صورت اتفاق می‌افتد؟ هم‌چنین ارتباط عبارت KL Divergence موجود در ELBO را با این پدیده بررسی کنید.

سوال ۲: (نظری) تابع ضرر GAN (۴۰ نمره)

می‌دانیم که هدف GAN بهینه‌سازی تابع ضرری مبتنی بر فاصله‌ی توزیع داده p_d و توزیع مدل مولد p_g است. تابع ضررهای ذیل برای تخمین فاصله دو توزیع p و q به کار می‌روند.

- KL divergence:

$$KL(p\|q) = \int \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right)p(\boldsymbol{x})d\boldsymbol{x}$$

- JSD:

$$JSD(p,q) = \frac{1}{2}KL(p\|m) + \frac{1}{2}KL(q\|m)$$

که $m = \frac{1}{2}(p+q)$.

- Earth-Mover (EM) distance:

$$W(p,q) = \inf_{\gamma\in\Gamma(p,q)}\mathbb{E}_{(x,y)\sim\gamma}\|x-y\|$$

که $\Gamma(p,q)$ مجموعه تمام توزیع‌های $\gamma(x,y)$ که داریم:

$$\forall\gamma(x,y)\in\Gamma(p,q): \int_y\gamma(x,y)dy = p(x), \int_x\gamma(x,y)dx = q(y)$$

.

[1] Variational AutoEncoder

اکنون دو توزیع دوبعدی P و Q را درنظر بگیرید.

$$\forall(x,y) \sim P, x = 0, y \sim \text{Uniform}(0,1)$$

.

$$\forall(x,y) \sim Q, x = \theta(0 \leq \theta \leq 1), y \sim \text{Uniform}(0,1)$$

.

آ) $KL(P\|Q), KL(Q\|P), JSD(P,Q)$ و $W(P,Q)$ را محاسبه کنید.

ب) دو توزیع P و Q دیگری بسازید که $JSD(P,Q)$ نسبت به پارامترهای P و Q مشتق‌پذیر نباشد.

ج) از مثال بالا استفاده کنید و معایب استفاده از JSD در مدل GAN را شرح دهید.

سوال ۳: (نظری) بهینه‌سازی واگرایی (۵۰ نمره)

یک شبکه GAN را در نظر بگیرید که با تولید نویز z، سعی دارد $G_\theta(z)$ را به عنوان نمونه‌ای از توزیع اصلی داده‌ها p_{data} مطرح کند. توزیع $G_\theta(z)$ را $p_\theta(\boldsymbol{x})$ درنظر بگیرید. تابع ضرر discriminator به صورت زیر است:

$$L_D(\phi;\theta) = -\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\log D_\phi(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}\left[\log\left(1 - D_\phi(\boldsymbol{x})\right)\right]$$

آ) نشان دهید L_D هنگامی کمینه می‌شودکه داشته باشیم $D_\phi = D^*$ که:

$$D^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}) + p_{\text{data}}(\boldsymbol{x})}$$

ب) می‌دانیم $D_\phi(\boldsymbol{x}) = \sigma\left(h_\phi(\boldsymbol{x})\right)$ نشان دهید که $h_\phi(\boldsymbol{x})$ (که بردار logit های discriminator را نمایش می‌دهد.) لگاریتم درست‌نمایی توزیع اصلی x و توزیع تخمینی توسط مدل را تخمین می‌زند. به عبارت دیگر، نشان دهید اگر $D_\phi = D^*$ داریم:

$$h_\phi(\boldsymbol{x}) = \log \frac{p_{\text{data}}(\boldsymbol{x})}{p_\theta(\boldsymbol{x})}.$$

ج) اگر تابع ضرر generator به صورت ذیل باشد:

$$L_G(\theta;\phi) = \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}\left[\log\left(1 - D_\phi(\boldsymbol{x})\right)\right] - \mathbb{E}_{\boldsymbol{x}\sim p_\theta(\boldsymbol{x})}\left[\log D_\phi(\boldsymbol{x})\right]$$

نشان دهید اگر $D_\phi = D^*$، داریم:

$$L_G(\theta;\phi) = \text{KL}\left(p_\theta(\boldsymbol{x})\|p_{\text{data}}(\boldsymbol{x})\right)$$

د) می‌دانیم در زمان آموزش مدل‌های VAE، تابع منفی ELBO، حد بالای منفی تابع لگاریتم درست‌نمایی را کمینه می‌کنیم. نشان دهید منفی تابع لگاریتم درست‌نمایی یعنی $-\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x})\right]$ را می‌توان به عنوان یک تابع KL Divergence به علاوه‌ی یک عبارت ثابت برحسب θ نوشت. آیا این موضوع به این معنی است که کدگشای VAE که با ELBO آموزش دیده است و مولد GAN که با تابع ضرر L_G آموزش داده شده است، یک هدف را آموزش می‌بینند؟ توضیح دهید.

سوال ۴: (نظری) مروری بر مدل‌های دیفیوژنی (۶۰ نمره)

در این سوال قصد داریم نحوه‌ی استخراج زیان مدل‌های دیفیوژنی را یک بار دیگر با جزئیات مرور کنیم.
اگر توزیع تصاویر واقعی را $q(x)$ بنامیم، تصویر $x_0 \sim q(x)$ را در نظر بگیرید. منظور از فرآیند دیفیوژنی رو به جلو، این است که به تدریج و طی T گام، نویزهای گاوسی به x_0 اضافه شوند تا نهایتا یک الگوی کاملا تصادفی بدست آید. نویزها و تصاویر حاصل را به ترتیب با $\epsilon_0, ..., \epsilon_{T-1}$ و $x_1, ..., x_T$ نشان می‌دهیم. واریانس نویز گاوسی در گام t را با β_t نمایش می‌دهیم. همچنین برای راحتی کار، نمادهای $\alpha_t := 1 - \beta_t$ و $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$ را تعریف می‌کنیم. از جایی که فرض مارکوف بودن را در تولید تصاویر نویزی داریم، توزیع فرآیند رو به جلو، یعنی $q(x_{1:T}|x_0)$ را بصورت زیر تعریف می‌کنیم:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (۱)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \quad (۲)$$

(آ) حال با استفاده از تکنیک تغییر پارامتر [۲] به کمک ϵ_t ها، x_t را برحسب α_t ، x_{t-1} و ϵ_{t-1} بازنویسی کنید. سپس با بسط بازگشتی رابطه ثابت کنید

$$q(x_t|x_0) = \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \quad (۳)$$

(ب) با توجه به نتیجه‌ی حاصل، شرط لازم برای میل کردن x_T به نویز کامل گاوسی چیست؟

حال به فرآیند دیفیوژنی معکوس می‌پردازیم. فرض کنید $x_T \sim \mathcal{N}(0, I)$ را در اختیار داریم. اگر $q(x_{t-1}|x_t)$ها را از توزیع واقعی در اختیار داشته باشیم می‌توانیم x_0 را بازیابی کنیم.

(ج) چه مشکلی برای محاسبه‌ی مستقیم $q(x_{t-1}|x_t)$ وجود دارد؟

با توجه به مشکلاتی که ذکر کردید، به جای محاسبه‌ی مستقیم، آن را با توزیع گاوسی تخمین می‌زنند. پس اگر تعریف کنیم

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}|\mu_\theta(x,t), \Sigma_\theta(x,t)) \quad (۴)$$

کافیست با پارامترهای θ میانگین و واریانس را یاد بگیریم. توزیع فرآیند معکوس نیز بصورت زیر در می‌آید:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \quad (۵)$$

(د) نشان دهید که علیرغم بخش ج، $q(x_{t-1}|x_t, x_0)$ محاسبه‌پذیر است. به طور دقیق‌تر، با استفاده از قانون بیز نشان دهید که

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}|\tilde{\mu}(x_t, x_0), \tilde{\beta}_t I) \quad (۶)$$

$$\tilde{\mu}(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{1-\bar{\alpha}_t}\epsilon_t) \quad (۷)$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \quad (۸)$$

پس با یادگیری میانگین از روی x_0 و x_t و محاسبه‌ی $\tilde{\beta}_t$ می‌توان احتمال $q(x_{t-1}|x_t, x_0)$ را بدست آورد. با توجه به شباهت این قسمت با VAE به سراغ حد پایین وردشی (VLB) می‌رویم.
با استفاده از نامساوی Jensen می‌توان نشان داد

$$-\mathbb{E}_{q(x_0)}[\log p_\theta(x_0)] \leq \mathbb{E}_{q(x_{0:T})}[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})}] \triangleq L_{VLB} \quad (۹)$$

که در آن سمت چپ همان زیان لگاریتم درست‌نمایی است که تمایل داریم کمینه شود. لذا حد بالای آن را کمینه می‌کنیم. هم‌چنین همانطور که در اسلایدهای درس دیدید، می‌توان ثابت کرد که

$$L_{VLB} = \underbrace{D_{KL}(q(x_T|x_0)\|p_\theta(x_T))}_{L_T} + \quad (۱۰)$$

$$\sum_{t=2}^{T}\underbrace{D_{KL}(q(x_{t-1}|x_t, x_0)\|p_\theta(x_{t-1}|x_t))}_{L_t} + \quad (۱۱)$$

$$\underbrace{(-\log p_\theta(x_0|x_1))}_{L_0} \quad (۱۲)$$

تمام عبارات بصورت KL-Divergence نوشته شدند، بجز (۱۲) که می‌توان آن را نیز با یک دیکودر مجزا مدل کرد. از عبارت L_T نیز می‌توان چشم‌پوشی کرد (چرا؟).
گفتیم که با توجه به (۴) نیاز داریم با یک شبکه‌ی حاوی پارامترهای θ ، μ_θ و Σ_θ را یاد بگیریم. می‌توان μ_θ را چنان آموزش داد که $\tilde{\mu}_t$ را پیش‌بینی کند. از جایی که x_t در حین آموزش در دسترس است، می‌توان با تکنیک تغییر پارامتر $\mu_\theta(x_t, t)$ را بازنویسی کرد تا تنها $\epsilon_\theta(x_t, t)$ را یاد بگیرد؛ یعنی

$$\mu_\theta(x,t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \quad (۱۳)$$

(ه) با توجه به این موارد، ثابت کنید

$$L_t = \mathbb{E}_{x_0,\epsilon}[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2}\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|_2^2] \quad (۱۴)$$

[۲] reparameterization technique

در مقاله‌ی Ho et al. 2020 به این نتیجه رسیدند که با نادیده گرفتن ضریب $\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)\|\Sigma_\theta\|_2^2}$ می‌توان به زیان ساده‌تری رسید:

$$L_t^{simple} = \mathbb{E}_{t\sim T, x_0, \epsilon_t}[\|\epsilon_t - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, t)\|_2^2] + C \quad (۱۵)$$

که در آن C یک مقدار مستقل از θ است.

(و) دیدیم که یکی از مزایای معرفی ϵ_θ و تغییر پارامتر در تخمین μ_θ این است که تابع زیان ساده‌تر می‌شود. حال استدلال کنید که چگونه این کار در درک همگرایی الگوریتم کمک می‌کند (به بحث Langevin dynamics توجه شود).

سوال ۵: (عملی) مدل‌های مولد (۲۵+۱۰۰ نمره)

در این سوال به پیاده‌سازی و مقایسه‌ی مدل‌های مولد VAE ، GAN و DDPM می‌پردازیم. لطفا نوت‌بوک `Generative_Models.ipynb` را طبق توضیحات و با رعایت ساختار پیشنهادی تکمیل کنید. لطفا در نهایت خود فایل نوت‌بوک را ارسال کنید و از ارسال لینک یا به اشتراک‌گذاری در کولب و ... خودداری کنید. برای کاهش حجم نوت‌بوک و اطمینان از مشاهده‌پذیر بودن تصاویر، می‌توانید نمونه‌هایی از تصاویر تولیدشده در اپاک‌های آموزش را در پوشه‌های جداگانه‌ای ذخیره کنید.