بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۳-۰۲

مدرس: مهدیه سلیمانی

تمرین چهارم

- مهلت ارسال پاسخ تا ساعت **۲۳:۵۹** روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر تمرین‌های نظری بدون کسر نمره تا سقف ۵ روز و تمرین‌های عملی تا سقف ۱۰ روز وجود دارد. محل بارگزاری جواب تمرین‌های نظری بعد از ۳ روز و تمرین‌های عملی بعد از ۵ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد.
- هم‌فکری در انجام تمرین مانعی ندارد، فقط توجه داشته باشید که پاسخ تمرین حتما باید توسط خود شخص نوشته شده باشد. همچنین در صورت همفکری در هر تمرین، در ابتدای جواب تمرین نام افرادی که با آن‌ها هم‌فکری کرده اید را حتما ذکر کنید.
- برای پاسخ به سوالات نظری در صورتی که از برگه خود عکس تهیه می‌کنید، حتما توجه داشته باشید که تصویر کاملا واضح و خوانا باشد. درصورتی که خوانایی کافی را نداشته باشد، تصحیح نخواهد شد.
- 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 : آقایان جواهریان، حسینی، علیخانی، ثقفیان

بخش نظری (۶۵ نمره (+۱۵ امتیازی))

سوال اول: استنباط متغیر (۲۰ نمره)

در این تمرین پایه های تئوری مدل VAE و پیشرفت های حاصل شده بر روی آن مورد بررسی قرار می گیرند.
یک مدل احتمالاتی که به صورت مجموعه ای از متغیر های قابل مشاهده X و مجموعه متغیر های پنهان Z می باشد را در نظر بگیرید؛ در استنباط بیزی ما علاقه مند به محاسبه توزیع پسین Z بعد از مشاهده دادگان هستیم. چالش اصلی این روش، intractable بودن توزیع پسین می باشد. استنباط متغیر (Variational Inference) یکی از روش های حل این چالش می باشد. در این روش، محاسبه توزیع پسین را به یک مسئله بهینه سازی روی خانواده ای از توزیع ها تبدیل می کنیم:

$$q = \underset{q \in \mathbb{Q}}{\operatorname{argmin}} \, D_{\mathrm{KL}}\left(q(z) \parallel p_\theta(z|x)\right)$$

۱. با بسط دادن رابطه مطرح شده، به رابطه ELBO برسید و نشان دهید، ELBO کران پایینی برای $\log p_\theta(x)$ می باشد.

۲. فرض کنید مجموعه دادگان $D = \{x_i\}$ را در اختیار داریم؛ اگر بخواهیم تخمین MLE برای پارامتر های مدل ارائه دهیم؛ بایستی تابع زیر را بیشینه کنیم:

$$\log p(D|\theta) = \sum_{i=1}^{N} \log p(x = x_i|\theta)$$

اما محاسبه مستقیم $\log p(x = x_i|\theta)$ برایمان مقدور نیست. بنابراین به جای رابطه بالا، رابطه زیر را بیشینه می کنیم:

$$\sum_{i=1}^{N} \max_{q \in \mathbb{Q}} \text{ELBO}(q, x_i, \theta)$$

حال فرض کنید خانواده Q به کمک پارامتر ψ قابل پارامترایز کردن می باشد. به کمک الگوریتم کاهش گرادیان، یک الگوریتم برای انجام MLE ارائه دهید.

۳. رابطه قبلی از لحاظ محاسباتی سنگین می باشد. برای حل این چالش دو روش ارائه می شود:

(آ) Stochastic VI: در این روش به کمک تخمین مونته کارلو، جمع موجود در تابع هزینه، با یک جمع روی batch mini تخمین زده می شود:

$$\sum_{i=1}^{N} \max_{q \in \mathbb{Q}} \text{ELBO}(q, x_i, \theta) \simeq \frac{N}{B} \sum_{i=1}^{B} \max_{q \in \mathbb{Q}} \text{ELBO}(q, x_i, \theta)$$

(ب) Amortized VI: در این روش به جای پیدا کردن N پارامتر متغیر $(\psi_{1:N})$ ، یک شبکه عصبی با پارامتر ϕ بهینه می شود که با ورودی گرفتن x_i، پارامتر بهینه ψ_i را خروجی دهد:

$$q(z_n|\psi_n) = q(z_n|g_\phi(x_n)) = q_\phi(z_n|x_n)$$
$$\sum_{i=1}^{B} \max_{q \in \mathbb{Q}} \text{ELBO}(q, x_i, \theta) = \max_{\phi} \sum_{i=1}^{B} \text{ELBO}(q_\phi(.|x_i), x_i, \theta)$$

این دو روش را در الگوریتم طراحی شده برای قسمت قبل اعمال کنید و الگوریتم تغییر یافته را ارائه دهید.

۴. حال با در نظر گرفتن مفروضات زیر و به کمک reparameterization trick الگوریتم خود را برای این حالت خاص، مشخص کنید:

$$p_\theta(z) = \mathcal{N}(0, I) \tag{۱}$$
$$p_\theta(x|z) = \mathcal{N}(f_\theta(z), \sigma^2 I) \tag{۲}$$
$$q_\phi(.|x) = \mathcal{N}(\mu_\phi(x), \text{diag}(\sigma_\phi^2(x))) \tag{۳}$$

* توجه کنید که بایستی تابع هزینه خود را به ساده ترین شکل ممکن در بیاورید. (بایستی KL دو توزیع نرمال را ساده کنید)

۵. در چند سوال قبل به بررسی حالت خاصی از VI به نام fixed-form VI پرداختیم. می توان به جای انتخاب یک خانوداه پارامتری از توزیع های متغیر و انجام مسئله بهینه سازی به کمک کاهش گرادیان، انتخاب های دیگری نیز انجام داد. در این سوال به بررسی یک روش پر استفاده دیگر به نام Mean-Field VI می پردازیم.

فرض کنید متغیر نهان Z یک بردار J بعدی باشد و توزیع متغیر را به صورت زیر بنویسیم:

$$q_\psi(z) = \prod_{j=1}^{J} q_j(z_j)$$

حال فرض کنید تمامی q_i ها بغیر از q_j ثابت باشند. نشان دهید q_j که مقدار ELBO را بیشینه می کند به صورت زیر می باشد:

$$q_j(z_j) \propto \exp\left[\mathbb{E}_{z_{-j}}\left[\log p_\theta(x, z_j, z_{-j})\right]\right]$$

که در آن z_{-j} مجموعه تمامی z_i ها بغیر از z_j می باشد.

سوال دوم: Diffusion Models (۲۵ نمره)

در کلاس درس با خانوادهای از مدلهای دیفیوژنی به نام DDPMs (Denoising Diffusion Probabilistic Models) آشنا شدیم و در این سوال قصد داریم با خانواده دیگری از این مدلهای دیفیوژنی به نام VDMs (Variational Diffusion Models) آشنا بشویم. برعکس DDPMs که فضای نهان را به صورت گسسته در زمان در نظر میگرفت، در VDMs فضای نهان به صورت پیوسته در نظر گرفته میشود. در ادامه با برخی از مراحل این مدل VDMs بیشتر آشنا میشویم.

فرایند رو به جلو (Forward Process)

در این فرایند، قصد داریم به آرامی داده x را به یک نویز تصادفی تبدیل کنیم و این کار را نیز با استفاده از نویز گاوسی در هر مرحله انجام میدهیم. در اینصورت، نسخه نویزی شده x در گام t را متغیر نهان z_t مینامیم که مقادیر t در بازهٔ $t \in [0, 1]$ قرار دارند. به این معنا که برای $t = 0$، کمترین تصویر نویزی و برای $t = 1$، بیشترین تصویر نویزی را خواهیم داشت. همچنین، فرض میکنیم که توزیع شرطی z_t برحسب x به صورت زیر میباشد:

$$q(z_t|x) = \mathcal{N}(a_t x, \sigma_t^2 I)$$

که در عبارت بالا a_t و σ_t^2 هر کدام توابعی مثبت برحسب t میباشند.

(آ) یکی از شرایطی که ما دوست داریم در فرایند فروارد رعایت شود این مورد میباشد که واریانس متغیرهای نهان ما در طول فرایند عوض نشود. نشان دهید برای اینکه چنین شرطی رعایت شود باید رابطه زیر برقرار باشد. (فرض شود داده ورودی استاندارد شده میباشد). (۲ نمره)

$$a_t = \sqrt{1 - \sigma_t^2}$$

مشابه DDPM فرض کنید زنجیره تغییرات ما از x تا z_1 یک زنجیره مارکوفی (Markov chain) باشد. زنجیره مارکوفی به این معنا میباشد که برای بدست آوردن تصویر نویزی شده در زمان t ما فقط نیاز به آخرین تصویر کمتر نویزی بدست آمده داریم. در حالت گسسته در زمان این زنجیره مارکوف به صورت زیر میباشد.

$$z_1 \leftarrow z_{\left(\frac{T-1}{T}\right)} \leftarrow \cdots \leftarrow z_0 \leftarrow x$$

در حالت پیوسته $T \to \infty$ میل میکند، در نتیجه مقدار تغییر در هر گام بسیار ریز میباشد.

(ب) با استفاده از تعریف مارکوف بودن فرایند فروارد، نشان دهید که رابطه زیر برقرار می‌باشد. (۲ نمره)

$$q(z_s, z_t|x) = q(z_t|z_s)q(z_s|x)$$

(ج) نشان دهید توزیع $q(z_t|z_s)$ به ازای $t > s$ یک توزیع گاوسی به صورت زیر می‌باشد. (۴ نمره)

$$q(z_t|z_s) = \mathcal{N}(a_{t|s}z_s, \sigma^2_{t|s}I)$$

$$a_{t|s} = \frac{a_t}{a_s}, \sigma^2_{t|s} = \sigma^2_t - a^2_{t|s}\sigma^2_s$$

مجددا مشابه DDPM ما علاقه‌مند هستیم تا فرم بسته‌ای برای توزیع $q(z_s|z_t, x); t > s$ بدست بیاوریم زیرا از آن در فرایند رو به عقب (Reverse process) و ساده‌سازی تابع هزینه مدل استفاده می‌شود.

(د) حالا با استفاده از نتایج بدست آمده از قسمت (ب) و قسمت (ج) نشان دهید روابط زیر برقرار می‌باشند. (۸ نمره)

$$q(z_s|z_t, x) = \mathcal{N}(\mu_Q(z_t, x; s, t), \sigma^2_Q(s, t)I)$$

$$\mu_Q(z_t, x; s, t) = \frac{(a^2_{t|s}\sigma^2_s)}{\sigma^2_t}z_t + \frac{(a_s\sigma^2_{t|s})}{\sigma^2_t}x$$

$$\sigma^2_Q(s, t) = \frac{(\sigma^2_{t|s}\sigma^2_s)}{\sigma^2_t}$$

راهنمایی: ابتدا نشان دهید $q(z_s|z_t, x) \propto q(z_t|z_s)q(z_s|x)$ برقرار می‌باشد و سپس عبارت را باز کنید و همچنین زمانیکه دو گاوسی را در هم ضرب می‌کنید ضرایب را در نظر نگیرید.

فرایند رو به عقب و تابع هزینه مدل (Loss function & Reverse process)

فرض کنید ما دو مدل با نام‌های $\hat{x}_\theta(z_t; t)$ و $p_\theta(z_s|z_t)$ آموزش داده‌ایم که در ادامه عملکرد هر کدام را توضیح می‌دهیم.
مدل $\hat{x}_\theta(z_t; t)$: ورودی آن تصویر نویزی و زمان مربوطه می‌باشد و خروجی آن تصویر بدون نویز x می‌باشد.
مدل $p_\theta(z_s|z_t)$: شبکه‌ای هست که سعی می‌کند تا رفتار $q(z_s|z_t, x)$ را تقلید بکند به این معنی که اگر تصویر نویزی شده را به شبکه بدهیم، مدل به ما تصویر کمتر نویزی شده در لحظه قبلی را می‌دهد و برای انجام این کار نیز x را از مدل $\hat{x}_\theta(z_t; t)$ بدست می‌آورد و به صورت زیر تعریف می‌شود.

$$p_\theta(z_s|z_t) = \mathcal{N}(\mu_\theta(z_t; s, t), \sigma^2_Q(s, t)I)$$

$$\mu_\theta(z_t; s, t) = \frac{(a_{t|s}\sigma^2_s)}{\sigma^2_t}z_t + \frac{(a_s\sigma^2_{t|s})}{\sigma^2_t}\hat{x}_\theta(z_t; t)$$

(ه) نشان دهید که divergence KL بین دو توزیع $p_\theta(z_s|z_t)$ و $q(z_s|z_t, x)$ به صورت زیر می‌باشد. (۲ نمره)

$$D_{\text{KL}}(q(z_s|z_t, x) \parallel p_\theta(z_s|z_t)) = \frac{1}{2\sigma^2_Q(s, t)}\|\mu_Q(z_t, x; s, t) - \mu_\theta(z_t; s, t)\|^2_2$$

راهنمایی: از رابطه زیر در ساده‌سازی استفاده کنید.

$$D_{\text{KL}}(N_0 \parallel N_1) = \frac{1}{2}\left(\text{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T\Sigma_1^{-1}(\mu_1 - \mu_0) - d + \log(\frac{\det \Sigma_1}{\det \Sigma_0})\right)$$

همچنین SNR (نسبت سیگنال به نویز) نیز به صورت زیر تعریف می‌شود:

$$\text{SNR}(t) = \frac{a_t}{\sigma^2_t}$$

(و) نشان دهید که می‌توان رابطه بدست آمده در قسمت (ه) را نیز بیشتر ساده‌سازی کرد و به عبارت زیر رسید. (۳ نمره)

$$D_{\mathrm{KL}}(q(z_s|z_t,x) \parallel p_\theta(z_s|z_t)) = \frac{1}{2}(\mathrm{SNR}(s) - \mathrm{SNR}(t))\|x - \hat{x}_\theta(z_t;t)\|_2^2$$

راهنمایی: ابتدا مشابه کاری که در DDPM برای ساده‌سازی انجام میدادید را بر روی نتیجه قسمت (ه) اعمال کنید تا به عبارتی به فرم زیر برسید و سپس ضریب بدست آمده را براساس SNR بازنویسی بکنید تا به حکم مسئله برسید.

$$D_{\mathrm{KL}}(q(z_s|z_t,x) \parallel p_\theta(z_s|z_t)) = \frac{1}{2}\gamma\|x - \hat{x}_\theta(z_t;t)\|_2^2$$

حالا ما برای اینکه تابع هزینه را بنویسیم نیاز داریم تا مقداری مسئله را سبک کنیم، فرض کنید ما بازه‌ی $[0,1]$ را به T قسمت با اندازه برابر $l = 1/T$ تقسیم کرده باشیم. در اینصورت مسئله ما به همان حالت گسسته در زمان DDPM تبدیل می‌شود در اینصورت می‌توان تابع هزینه VDM را نوشت:

$$L_T(x) = \sum_{i=1}^{T} \mathbb{E}_{q(z_{t(i)}|x)}\left[D_{\mathrm{KL}}\left(q(z_{s(i)}|z_{t(i)},x) \parallel p_\theta(z_{s(i)}|z_{t(i)})\right)\right]$$

در عبارت بالا $s(i) = \frac{i-1}{T}$ و $t(i) = \frac{i}{T}$ می‌باشند. می‌توان با نتایج بدست آمده از قسمت‌های قبل و اندکی ساده‌سازی تابع هزینه بالا را به صورت زیر بازنویسی کرد.

$$L_T(x) = \frac{T}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),\ i\sim\mathrm{U}\{1,T\}}\left[(\mathrm{SNR}(s(i)) - \mathrm{SNR}(t(i)))\|x - \hat{x}_\theta(z_{t(i)};t(i))\|^2\right]$$

اما ما دوست داریم که فضای نهان پیوسته داشته باشیم، یعنی می‌خواهیم $T \to \infty$ میل بکند تا بتوان $L_\infty(x)$ را بدست بیاوریم که براساس آن مدل را آموزش بدهیم.

(ر) نشان دهید $L_\infty(x)$ به صورت زیر می‌باشد. (۴ نمره)

$$L_\infty(x) = \lim_{T\to\infty} L_T(x) = -\frac{1}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),\ t\sim\mathrm{U}(0,1)}\left[\mathrm{SNR}'(t)\|x - \hat{x}_\theta(z_t;t)\|^2\right]$$

راهنمایی: برای اینکه بتوان به عبارت بالا رسید باید ابتدا s را براساس t بازنویسی بکنید و سپس از تعریف مشتق استفاده بکنید.

سوال سوم: Score Matching (۲۰ نمره)

در این سوال به بررسی روش score matching در تولید نمونه جدید خواهیم پرداخت. همانطور که در کلاس درس نیز بیان شد، روش‌های مبتنی بر امتیاز یکی دیگر از روش‌هایی هستند که به منظور تولید نمونه جدید برای توزیع مورد نظر خود مورد استفاده قرار می‌گیرند. این روش ها از دو جزء اصلی تشکیل شده اند: Langevin dynamics و score-matching که در ادامه با هر کدام از این قسمت‌ها بیشتر آشنا خواهیم شد.

۱. Langevin Dynamics

فرض کنید که توزیع $p(x)$ در دسترس است و قصد داریم از این توزیع نمونه برداری کنیم. Legevin dynamics یک فرآیند تکرار شونده[1] است که در انجام این کار به ما کمک می‌کند و به صورت عبارت زیر تعریف می‌شود:

$$x_{t+1} = x_t + \delta\nabla_x \log p\left(x_t\right) + \sqrt{2\delta}\epsilon; \epsilon \sim N(0,I) \quad (۴)$$

که در عبارت بالا x_1 نقطه شروع اولیه و δ اندازه گام است. همچنین t در بازه $[1...T]$ قرار دارد.

[1] Iterative

(آ) فرض کنید در عبارت ۴ ترمی که مربوط به نویز است وجود نداشته باشد، در اینصورت نشان دهید که عبارت حاصل معادل شروع از یک نقطه تصادفی در فضای R^d و رسیدن رسید به یکی از قله‌های توزیع $p(x)$ است. همچنین در مورد تفاوت عبارت حاصل با maximum likelihood نیز به صورت مختصر توضیح دهید. (۳ نمره)

(ب) بر اساس نتیجه بدست آمده در قسمت آ می‌توان گفت که عبارت بدون ترم نویز، می‌تواند به ما در رسیدن به یکی از قله‌های توزیع $p(x)$ کمک کند، یا به عبارت دیگر داده بدست آمده احتمال بالایی خواهد داشت، در اینصورت دلیل اضافه کردن نویز در عبارت ۴ چیست؟ همچنین رفتار Langevin dynamic را در زمانی که قله یافت شده تیز یا صاف باشد را توصیف کنید. (۲ نمره)

۲. Score Matching Techniques

یکی از چالش‌های تولید نمونه جدید، در درسترس نبودن توزیع نمونه‌ها یا همان $p(X)$ و عدم امکان استفاده از Langevin dynamic برای نمونه برداری از آن است. به منظور استفاده از این روش و رفع مشکل مذکور، سعی می‌کنند عبارت $\nabla_x \log p_\theta(x)$ را با استفاده از یک شبکه عصبی تخمین بزنند.

$$s_\theta(x) = \nabla_x \log p_\theta(x) \tag{۵}$$

به عبارت ۵ تابع امتیاز[۲] نیز گفته می‌شود. روش های مختلفی برای بدست آوردن تابع امتیاز وجود دارد که تفاوت آنها در تابع هزینه مدل است. فرض کنید دادگان آموزشی به صورت $X = \{x^{(1)}, ..., x^{(m)}\}$ باشد و تابع هزینه نیز به صورت زیر تعریف شده باشد:

$$J_1(\theta) = E_{q(x)}\left[\frac{1}{2}\left\|s_\theta(x) - \nabla_x \log q(x)\right\|^2\right] \tag{۶}$$

یک راحل این است که سعی کنیم از تخمین توزیع با استفاده از کرنل بهره ببریم:

$$q(x) = \frac{1}{M}\sum_{i=1}^{M} K\left(x \mid x^{(i)}\right) \tag{۷}$$

در عبارت ۷ $K(.)$ تابع کرنل است. یکی از کرنل‌هایی که می‌توان برای این کار مورد استفاده قرار داد، کرنل گاوسی[۳] است.

$$K\left(x \mid x^{(i)}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(\frac{-\left\|x - x^{(i)}\right\|^2}{2\sigma^2}\right) \tag{۸}$$

(آ) نشان دهید که با استفاده از کرنل عبارت ۸ و همچنین استفاده از توزیع تخمین زده شده $q(x)$ می‌توان به رابطه زیر رسید و با استفاده از این رابطه می‌توان مدل را آموزش داد. (۴ نمره)

$$\nabla_x \log q(x) = \frac{\sum_{i=1}^{M} \frac{1}{\sigma^2}\left(x^{(i)} - x\right) K\left(x \mid x^{(i)}\right)}{\frac{1}{M}\sum_{i=1}^{M} K\left(x \mid x^{(i)}\right)}$$

(ب) با وجود اینکه می‌توان با استفاده از کرنل قسمت ج شبکه را آموزش داد، معایت این روش را بیان کنید. (۱ نمره)

یک راه‌حل دیگر برای حل مسئله بالا استفاده از denoising score matching است. در این روش تابع هزینه به صورت زیر تعریف می‌شود:

$$J_2(\theta) = E_{q(x,x_0)}\left[\frac{1}{2}\left\|s_\theta(x) - \nabla_x \log q\left(x \mid x_0\right)\right\|^2\right] \tag{۹}$$

[2] Score function
[3] Gaussian kernel

(ج) نشان دهید بین دو تابع هزینه توشته شده در عبارت های ۶ و ۹ رابطه زیر برقرار است . (۸ نمره)

$$J_2(\theta) = J_1(\theta) + C$$

(د) حال که نشان داده شد تابع هزینه denoising score matching معادل با تابع هزینه اصلی است، با استفاده از فرض زیر ابتدا تابع هزینه denoising را دوباره بازنویسی کنید و سپس فرآیند آموزش را توضیح دهید. در نهایت با استفاده از Langevin dynamics برای مدل بدست آمده نمونه جدید تولید کنید. (۲ نمره)

$$q\left(x \mid x_0\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^d \exp\left(\frac{-\left\|x - x_0\right\|^2}{2\sigma^2}\right)$$

سوال چهارم (امتیازی): آنالیز پایداری GAN (۱۵ نمره)

می دانیم آموزش GAN ها در عمل چالش هایی به همراه دارد. یکی از روش هایی که برای تحلیل فرایند آموزش GAN ها استفاده می شود، تحلیل پایداری (Stability Analysis) می باشد. در تحلیل پایداری به بررسی همگرایی موضعی دینامیک آموزش به نقاط تعادل (Nash-equilibrium) تابع هزینه می پردازیم. به صورت کلی می توان تابع هزینه GAN ها را به صورت زیر نوشت:

$$\mathcal{L}(\theta, \psi) = \mathrm{E}_{p(z)}\left[f(D_\psi(G_\theta(z)))\right] + \mathrm{E}_{p_D(x)}\left[f(-D_\psi(x))\right]$$

که در آن θ پارامتر شبکه مولد، ψ پارامتر شبکه ممیز و f یک تابع مشتق پذیر با مشتق همواره ناصفر است.
آنالیز پایداری غالبا به دو صورت انجام می شود:

۱. بررسی فرآیند آموزش به صورت یک سیستم دینامیکی گسسته:

$$F_h(\theta, \psi) = (\theta, \psi) + h\, v(\theta, \psi)$$

که در آن h نرخ یادگیری و v بردار گرادیان (نسبت به هر دو پارامتر) می باشد.

۲. بررسی فرآیند آموزش به صورت گسسته شده ی یک سیستم دینامیکی پیوسته:

$$\begin{pmatrix} \dot{\theta}(t) \\ \dot{\psi}(t) \end{pmatrix} = v(\theta, \psi)$$

در هر دو روش، تحلیل پایداری بر اساس مقادیر ویژه ژاکوبین اپراتور (F_h و v) در نقطه تعادل انجام می شود.

۱. برای حالت گسسته:

(آ) اگر مقادیر ویژه همگی داخل دایره واحد (در فضای مختلط) باشند، الگوریتم با نرخ خطی به نقطه تعادل همگرا می شود.

(ب) اگر یکی از مقادیر ویژه بیرون دایره واحد باشد، الگوریتم همگرا نخواهد بود.

(ج) اگر همه ی مقادیر ویژه روی دایره واحد باشند، به طور کلی در مورد همگرایی نظری نمی توان داد، ولی اگر الگوریتم همگرا باشد، عموما با نرخ زیرخطی همگرا است.

۲. برای حالت پیوسته:

(آ) اگر همه ی مقادیر ویژه مقدار حقیقی منفی داشته باشند، الگوریتم به صورت خطی همگرا می شود.

(ب) اگر یکی از مقادیر ویژه، مقدار حقیقی مثبت داشته باشد، الگوریتم همگرا نخواهد بود.

(ج) اگر همه ی مقادیر ویژه موهومی باشند، به طور کلی در مورد همگرایی نظری نمی توان داد، ولی اگر الگوریتم همگرا باشد، عموما با نرخ زیر خطی همگرا است.

(د) اگر مقادیر ویژه بسیار نزدیک به محور موهومی باشند، برای همگرایی نیاز به نرخ یادگیری بسیار کوچک داریم.

در این سوال از شما میخواهیم با توجه به موارد ذکر شده، برای GAN زیر، تحلیل پایداری انحام بدهید:
توزیع واقعی دادگان $\delta(x)$ و توزیع مولد $\delta(x-\theta)$ می باشد. همجنین ممیز به صورت $D(x) = \psi x$ است.

۱. بخش اول: تحلیل پایداری پیوسته

(آ) اثبات کنید $\theta = \psi = 0$ تنها نقطه تعادل GAN مذکور می باشد. همچنین در این نقطه هر دو مقدار ویژه ماتریس ژاکوبین، موهومی می باشند. (۱.۵ نمره)

(ب) نشان دهید، صرف نظر از نقطه ی شروع، در طول مسیر آموزش، همواره مقدار $\theta^2 + \psi^2$ ثابت می ماند. (۱ نمره)

(ج) با توجه به دو سوال قبل، تحلیل پایداری پیوسته چه نظری در مورد همگرایی موضعه ای به سمت نقطه تعادل می دهد؟ (۰.۵ نمره)

۲. بخش دوم: تحلیل پایداری گسسته

(آ) نشان دهید در الگوریتم simultaneous gradient descent هر دو مقدار ویژه بیرون دایره واحد قرار می گیرند. همچنین نشان دهید صرف نظر از نقطه شروع، $\theta^2 + \psi^2$ در طول آموزش به صورت یکنوا افزایش می یابد. (۵ نمره)

(ب) نشان دهید در الگوریتم alternating gradient descent (یک مرحله آپدیت مولد، سپس یک مرحله آپدیت ممیز) مقادیر ویژه برابرند با:

$$\lambda_{1/2} = 1 - \frac{(hf'(0))^2}{2} \pm \sqrt{\left(1 - \frac{(hf'(0))^2}{2}\right)^2 - 1}.$$

(۴.۵ نمره)

(ج) با توجه به دو سوال قبل، تحلیل پایداری گسسته چه نظری در مورد همگرایی موضعه ای به سمت نقطه تعادل می دهد؟ (۰.۵ نمره)

۳. بخش سوم: بررسی instance noise

(آ) یکی از روش های ساده برای تثبیت آموزش GAN ها اضافه کردن نویز سفید $(\epsilon \sim \mathcal{N}(0,\sigma^2))$ به دادگان (هم دادگان واقعی و هم دادگان تولید شده توسط مولد) می باشد. در این حالت، مقادیر ویژه را (برای تحلیل پایداری پیوسته) محاسبه کنید. (۱.۵ نمره)

(ب) با فرض $(f(t) = -\log(1+\exp(-t)))$ نشان دهید در این حالت، فرایند آموزش با توجه به تحلیل پایداری پیوسته، همگرا است. (۰.۵ نمره)

بخش عملی (۴۸ نمره (+۷ امتیازی))

توجه: لطفا در کلیه سوال‌های عملی نوت بوک تکمیل شده خود را به همراه سایر موارد در کوئرا بارگذاری کنید و از ارسال لینک و به اشتراک گذاری نوت بوک خودداری فرمایید.

GAN-VAE (۲۰ نمره)

در این سوال به پیاده سازی و مقایسه مدل های مولد VAE و GAN می‌پردازیم. برای مقایسه نتایج نهایی می‌توانید از discriminator آموزش‌داده شده در GAN یا یک discriminator جدید استفاده کنید. بدین منظور لطفا نوت بوک GAN-VAE.ipynb را تکمیل کرده و فایل نهایی خود را ارسال کنید.

DDPM (۲۸ نمره (+۷ امتیازی))

کپچا یک راهکار اولیه برای تشخیص ربات از انسان در سیستم های کامپیوتری است که در دنیای امروزه کاربردهای بسیاری اعم از جلوگیری از دسترسی ربات ها به سرویس ها از طریق حملات Brute-Force را دارد. تولید کپچا های متنی عموما توسط الگوریتم های نسبتا قطعی با تصادف کم تولید می شوند و خلاقیت و تنوع کمی را دارا می باشند. در این تمرین تولید کپچا های خلاقانه با استفاده از مدل های مولد را مورد بررسی قرار خواهیم داد. به طور خاص یک DDPM[4] برای تولید تصاویر کپچای تصادفی طراحی نموده و در ادامه با شرطی کردن با هدف تولید تصویر مرتبط با متن امکان استفاده از آن در عمل را بررسی خواهیم کرد.با توجه به توضیحات فوق، لطفا نوت DDPM.ipynb را با توجه به دستور العمل های موجود در آن تکمیل کرده و فایل نهایی خود را ارسال کنید. لازم به توضیح است که کلیه سلول ها باید امکان اجرای مجدد را داشته و نتایج قابل باز تولید باشند.

[4]Denoising Diffusion Probabilistic Models