تمرین سری چهارم Deep Learning — همفکری با سید امیر کسائی

امیدمحمد عزتی 402212269

سوال ①

1)

$$\log P_\theta(x) = \log \sum_z P_\theta(x, z)$$

$$= \log \sum_z q(z) \frac{P_\theta(x,z)}{q(z)} \geqslant \sum_z q(z) \log \frac{P_\theta(x,z)}{q(z)}$$

$$= \sum_z q(z) \frac{P_\theta(x|z)\, P_\theta(z)}{q(z)} = E$$

$$= \underbrace{E_{q(z)}\left[\log P_\theta(x|z)\right] - KL\left(q(z) \,\|\, P_\theta(z)\right)}_{ELBO}$$

2)

1) پارامترهای روبه‌رو را بصورت رندوم مقداردهی اولیه می‌کنیم: $\theta, \psi_{1:N}, N$

2) باید با ماکزیمم کردن ELBO، رابطه زیر را بهینه کنیم:

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

3) در یک while loop تا زمانی که ELBO به بهینه برسد و Converge کند ادامه می‌دهیم و به تعداد N تا داده ψ_i ها را بصورت زیر محاسبه می‌کنیم:

$$\psi_i \leftarrow \arg\max_{\psi_i} L(\psi_i, \theta; x_i)$$

4) در ادامه‌ی loop، مقدار $L(x, \psi_{1:N}, \theta)$ را محاسبه کرده و با استفاده از آنها θ را محاسبه می‌کنیم:

$$\theta \leftarrow \arg\max_\theta \sum_{i=1}^{N} L(x_i; \psi_i, \theta)$$

ادامه ۲) همینطور تا زمانی که ... همگرا شود و به بهینه برسیم حلقه while را ادامه می‌دهیم.

۳)

الف) مثل الگوریتم بخش قبل عمل می‌کنیم اما در اینجا B تا داده در هر بچ داریم سپس در loop ابتدا B تا سمپل برای Batch در نظر می‌گیریم.

سپس گرادیان‌های $L(x_j; \psi_j, \theta)$ را نسبت به ψ_j و θ به ازای هر سمپل محاسبه می‌کنیم. در نهایت گرادیان نسبت به ψ و θ میانگین گرادیان‌های B تا سمپل است یعنی $\frac{1}{B}\sum_j \nabla_{\psi_j} L(x_j; \psi_j, \theta)$ و $\frac{1}{B}\sum_j \nabla_\theta (x_j; \psi_j, \theta)$

به ترتیب این مقادیر را با $\hat{\nabla}_{\psi B}$ و $\hat{\nabla}_\theta$ نشان می‌دهیم.

در مرحله بعد برای آپدیت ψ و θ محاسبات زیر انجام می‌شود:

$$\begin{cases} \bar{\nabla}_{\psi_j} L(x^B; \psi^B, \theta) \triangleq [I(\psi_j)]^{-1} \hat{\nabla}_{\psi B} L(x^B; \psi^B, \theta) \\ \psi_j \leftarrow \quad \text{با } \bar{\nabla}_{\psi B} \text{ آپدیت می‌کنیم} \end{cases}$$

$$\begin{cases} \bar{\nabla}_\theta L(x^B; \psi^B, \theta) \triangleq [I(\theta)]^{-1} \hat{\nabla}_\theta (x^B; \psi^B, \theta) \\ \theta \leftarrow \quad \text{با } \bar{\nabla}_\theta \text{ آپدیت می‌کنیم} \end{cases}$$

سپس مراحل زیر برای محاسبه ELBO نهایی انجام می‌شود:

$$\hat{L}(x^B; \psi^B, \theta) = \sum_{i=1}^{B} L(x_i; \psi_i, \theta)$$

$$L(x; \psi_{1:N}, \theta) = \frac{N}{B} \hat{L}(x^B; \psi^B, \theta)$$

3) ب) در این حالت دیگر $\psi_{1:N}$ نداریم. پس در مرحله اول θ و ϕ مقداردهی اولیه می‌شوند.

سپس مثل حالت قبلی گرادیان هر داده در batch نسبت به ϕ و θ محاسبه می‌شود و گرادیان نهایی میانگین گرادیان‌ها روی B تا داده است.

سپس ϕ و θ را با استفاده از گرادیان میانگین $\hat{\nabla}_\theta$ و $\hat{\nabla}_\phi$ آپدیت می‌کنیم.

$$\hat{L}(x^B, \phi, \theta) = \sum_{i=1}^{B} L(x_i; \phi, \theta)$$

$$L(x, \phi, \theta) = \frac{N}{B} \hat{L}(x^B, \phi, \theta)$$

4)

سوال ②

آ) در مرحله اول x_0 را با a_1 اسکیل می‌کنیم و با $N(0,\sigma_1^2)$ جمع می‌کنیم پس:

$$x_1 = a_1 x_0 + \sigma_1 \varepsilon_1$$

برای اینکه واریانس متغیرهای نهان ما تغییر نکند باید $Var(x_1) = 1$ باشد پس:

$$Var(x_1) = a_1^2 + \sigma_1^2 = 1 \implies a_1 = \sqrt{1-\sigma_1^2}$$

پس بطور کلی $a_t = \sqrt{1-\sigma_t^2}$ است.

ب) طبق قاعده مارکوف:

$$q(z_t, z_s | x) = \frac{q(z_t, z_s, x)}{q(x)}$$

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

$$q(z_t, z_s | x) = \frac{q(z_t | z_s, x) \cdot q(z_s, x)}{q(x)} = q(z_t | z_s, x)\, q(z_s | x)$$

چون z_t فقط به z_s در مارکوف وابسته است پس: $q(z_t | z_s, x) = q(z_t | z_s)$

$$\Rightarrow q(z_t, z_s | x) = q(z_t | z_s) \cdot q(z_s | x)$$

ج)

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

$$\alpha_{t|s} = \frac{a_t}{a_s} \quad , \quad \sigma^2_{t|s} = \sigma_t^2 - a^2_{t|s} \sigma_s^2$$

$$\begin{cases} q(z_t|x) = N(a_t x, \sigma_t^2 I) \\ q(z_s|x) = N(a_s x, \sigma_s^2 I) \\ q(z_s, z_t|x) = q(z_t|z_s)\, q(z_s|x) \end{cases}$$

توزیع joint :

$$\begin{pmatrix} z_s \\ z_t \end{pmatrix} \Big| x \sim N\left(\begin{pmatrix} a_s x \\ a_t x \end{pmatrix}, \begin{pmatrix} \sigma_s^2 & \Sigma_{st} \\ \Sigma_{ts} & \sigma_t^2 \end{pmatrix} \right)$$

$$q(z_t|z_s, x) = N(\mu_{t|s}, \Sigma_{t|s})$$

$$\begin{cases} \mu_{t|s} = a_t x + \Sigma_{ts} \Sigma_s^{-1} (z_s - a_s x) \\ \Sigma_{t|s} = \sigma_t^2 I - \Sigma_{ts} \Sigma_s^{-1} \Sigma_{st} \end{cases}$$

$$\Sigma_{ts} = E\left[(z_t - a_t x)(z_s - a_s x)^T \right] = a_t a_s^T \sigma_x^2 I$$

$$\Sigma_s = \sigma_s^2 I$$

$$\Sigma_{ts} \Sigma_s^{-1} = a_t a_s^T \sigma_x^2 I \, (\sigma_s^2 I)^{-1} = \frac{a_t}{a_s}$$

$$\begin{cases} \mu_{t|s} = a_t x + \frac{a_t}{a_s}(z_s - a_s x) = \frac{a_t}{a_s} z_s \\ \Sigma_{t|s} = \sigma_t^2 I - \left(\frac{a_t}{a_s}\right)^2 \sigma_s^2 I = \left(\sigma_t^2 - \frac{a_t^2}{a_s^2} \sigma_s^2\right) I \end{cases}$$

$$q(z_t|z_s) = N\left(\frac{a_t}{a_s} z_s , \left(\sigma_t^2 - \left(\frac{a_t}{a_s}\right)^2 \sigma_s^2 \right) I \right)$$

$$\begin{cases} a_{t|s} = \dfrac{a_t}{a_s} \\ \sigma_{t|s}^2 = \sigma_t^2 - \left(\dfrac{a_t}{a_s}\right)^2 \sigma_s^2 \end{cases}$$

$$\Rightarrow \quad q(z_t|z_s) = N\left(a_{t|s} z_s , \sigma_{t|s}^2 I \right)$$

د) با توجه به بخش های قبل: $q(z_s, z_t|s) = q(z_s|x)\, q(z_t|z_s)$

همچنین $q(z_s|z_t, x)$ را می توانیم بطور زیر بنویسیم:

$$q(z_s|z_t,x) = \underbrace{q(z_s|x)}_{\text{prior}}\ \underbrace{q(z_t|z_s)}_{\text{likelihood}}$$

چون هر دو توزیع prior و likelihood گاوسی است پس $q(z_s|z_t,x)$ نیز گاوسی است.

در حالت کلی اگر یک prior گاوسی $p(x) = N(\mu_A, \sigma_A^2)$ و یک likelihood بشکل $p(y|x) = N(ax, \sigma_B^2)$ داشته باشیم آنگاه توزیع $p(x|y)$ خواهد بود:

$$p(x|y) = N(\tilde{\mu}, \tilde{\sigma}^2)$$

$$\begin{cases} \tilde{\mu} = \tilde{\sigma}^{-2}(\sigma_A^{-2}\mu_A + a\,\sigma_B^{-2}\, y) \\ \tilde{\sigma}^{-2} = \sigma_A^{-2} + a^2 \sigma_B^{-2} \end{cases}$$

اگر توزیع prior و likelihood را بصورت زیر بنویسیم:

$$q(z_s|x) = N(\alpha_s x, \sigma_s^2 I)$$

$$q(z_t|z_s) = N(\alpha_{t|s} z_s, \sigma_{t|s}^2)$$

آنگاه:

$$q(z_s|z_t,x) = N\left(\mu_Q(z_t, x; s,t),\ \sigma_Q^2(s,t) I\right)$$

$$\sigma_Q^{-2}(s,t) = \sigma_s^{-2} + \alpha_{t|s}^2\, \sigma_{t|s}^{-2} = \sigma_t^2 / \sigma_{t|s}^2\, \sigma_s^2$$

$$\sigma_Q^2(s,t) = \frac{\sigma_{t|s}^2\, \sigma_s^2}{\sigma_t^2} \qquad \boxed{\mu_Q(z_t, x; s,t) = \frac{\alpha_{t|s}\,\sigma_s^2}{\sigma_t^2} z_t + \frac{\alpha_s\, \sigma_{t|s}^2}{\sigma_t^2} x}$$

$$D_{KL}(P\|Q) = \int p(x)\log\frac{q(x)}{p(x)}\,dx = E_{p(x)}\left[\log\frac{q(x)}{p(x)}\right] \quad (d$$

$$\log\frac{q(z_s|z_t,x)}{p_\theta(z_s|z_t)} = \log\left[\frac{\frac{1}{\sqrt{2\pi}\,\sigma}e^{\left(-\frac{1}{2\sigma^2}(z-\mu_Q)\right)}}{\frac{1}{\sqrt{2\pi}\,\sigma}e^{\left(-\frac{1}{2\sigma^2}(z-\mu_\theta)\right)}}\right] =$$

$$= \frac{-1}{2\sigma_Q^2}\left[(z-\mu_Q)^2-(z-\mu_\theta)^2\right]$$

$$D_{KL} = E_p\left[\log\frac{q}{p_\theta}\right] = E_p\left[\frac{-1}{2\sigma_Q^2}\left(\mu_Q^2-\mu_\theta^2+2z(\mu_\theta-\mu_Q)\right)\right]$$

$$\Rightarrow D_{KL} = -\frac{1}{2\sigma_Q^2}\left[2(\mu_\theta-\mu_Q)\,E_p[z]+\mu_Q^2-\mu_\theta^2\right]$$

$$\Rightarrow D_{KL} = \frac{1}{2\sigma_Q^2}\|\mu_Q-\mu_\theta\|_2^2$$

$$D_{KL}\left(q(z_s|z_t,x)\,\|\,p_\theta(z_s|z_t)\right) = \frac{1}{2\sigma_Q^2(s,t)}\|\mu_Q-\mu_\theta\|_2^2 \quad (9$$

$$= \frac{\sigma_t^2}{2\sigma_{t|s}^2\,\sigma_s^2}\;\frac{\alpha_s^2\,\sigma_{t|s}^2}{\sigma_t^2}\,\|x-\hat{x}_\theta(z_t;t)\|_2^2$$

$$= \frac{1}{2\sigma_s^2}\;\frac{\alpha_s^2\,\sigma_{t|s}^2}{\sigma_t^2}\,\|x-\hat{x}_\theta(z_t;t)\|_2^2 =$$

$$= \frac{1}{2\sigma_s^2}\;\frac{\alpha_s^2(\sigma_t^2-\alpha_{t|s}^2\,\sigma_s^2)}{\sigma_t^2}\,\|x-\hat{x}_\theta(z_t;t)\|$$

$$= \frac{1}{2}\;\frac{\alpha_s^2\alpha_t^2/\sigma_s^2-\alpha_t^2}{\sigma_t^2}\,\|x-\hat{x}_\theta(z_t;t)\|_2^2$$

$$D_{KL}\left(q(z_s|z_t,x) \,||\, p_\theta(z_s|z_t)\right) =$$ ادامه و

$$\frac{1}{2}\left[\left(SNR(s) - SNR(t)\right) \left\| x - \hat{x}_\theta(z_t;t)\right\|_2^2\right.$$

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

با توجه به فرض سوال

$$\begin{cases} s(i) = \frac{i-1}{T} \\ t(i) = \frac{i}{T} \end{cases} \Rightarrow s(i) = t(i) - \frac{1}{T}$$

$$SNR(s(i)) = SNR\left(t(i) - \frac{1}{T}\right)$$

با بسط تیلور:

$$SNR\left(t(i) - \frac{1}{T}\right) \simeq SNR(t(i)) - \frac{1}{T} SNR'(t(i))$$

$$SNR\left(t(i) - \frac{1}{T}\right) - SNR(t(i)) \simeq -\frac{1}{T} SNR'(t(i))$$

$$SNR(s(i)) - SNR(t(i)) \simeq -\frac{1}{T} SNR'(t(i))$$

$$L_T(x) = \frac{T}{2} E_{\epsilon \sim N(0,I),\, i \sim U\{1,T\}}\left[\left(-\frac{1}{T} SNR'(t(i))\right) \left\| x - \hat{x}_\theta(z_t;t)\right\|^2\right]$$

با انتگرال و تغییر متغیر $U\{1,T\}$ به $t \to (0,1)$ ، فضای گسسته به پیوسته تبدیل می‌شود.

$$L_\infty(x) = \lim_{T\to\infty} L_T(x) = -\frac{1}{2} E_{\epsilon \sim N(0,I)} \int_0^1 SNR'(t) \left\| x - \hat{x}_\theta(z_t;t)\right\|^2 dt$$

$$L_\infty(x) = -\frac{1}{2} E_{\epsilon \sim N(0,I),\, t \sim U(0,1)}\left[SNR'(t) \left\| x - \hat{x}_\theta(z_t;t)\right\|_2^2\right]$$

سوال 3

1) آ) $x_{t+1} = x_t + \delta \nabla_x \log p(x_t) + \sqrt{2\delta}\, \epsilon \qquad \epsilon \sim N(0, I)$

بعد از حذف نویز:

$$x_{t+1} = x_t + \delta \nabla_x \log p(x_t)$$

این عبارت در واقع انجام عمل gradient ascent است و بخش $\nabla_x \log(p(x_t))$ گرادیان تابع چگالی احتمال را نشان می‌دهد. این بخش کمک می‌کند به سمت جایی از توزیع برویم که چگالی حضور داده‌ها در آن بیشتر است. در واقع این عمل باعث یافتن global maxima یا local maxima می‌شود. چون x_0 بطور تصادفی در فضای R^d پیدا می‌شود پس با توجه به شروع مختلف به قله‌های متفاوتی می‌توانیم برسیم. حرکت در جهت گرادیان ما را به قله‌های تابع توزیع می‌رساند.

در maximum likelihood ما به دنبال تخمین پارامترهای توزیع $p(x)$ هستیم بطوریکه با توجه به داده‌ها، احتمال ماکزیمم شود. اما در اینجا می‌خواهیم از یک نقطه رندوم شروع کنیم و به سمتی حرکت کنیم که توزیع حداکثر می‌شود و احتمال حضور سمپل‌ها بیشتر است. در واقع در اینجا به دنبال ماکزیمم کردن یک نقطه یا سمپل در توزیع موجود هستیم. اما در MLE بدنبال یافتن پارامترهای توزیع بطوریکه به داده‌هایمان به خوبی fit شود، هستیم.

ب) استفاده از نویز نرمال از جهات مختلفی مهم است. یکی از دلایل این است مطمئن شویم فرایند Sampling کل فضای توزیع را جستجو می‌کند و در نقاط بهینه محلی stuck نمی‌شود. یک دلیل دیگر وجود نویز، نقش regularization است و از اینکه score function اورفیت شود و generalization خوبی نداشته باشد جلوگیری می‌کند.

یک دلیل دیگر آن در Sampling این است که sample ها متنوعی از توزیع گرفته شود و یکسان نباشند.

langevin dynamics وقتی به قله‌های صاف می‌رسد بدلیل داشتن نویز اطراف قله رانیز چک می‌کند، تا اگر در بهینه های محلی قرار داشت، در آنها گیر نکند. اما زمانی که به قله‌های تیزتر می‌رسد نیز بدلیل وجود نویز به اطراف حرکت می‌کند و به بهترین نتیجه لزوما نمی‌رسد ولی اطراف بهینه حرکت می‌کند.

2) آ)

$$q(x) = \frac{1}{m} \sum_{i=1}^{m} K(x \mid x^{(i)})$$

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

$$\nabla_x \log q(x) = \frac{\nabla_x q(x)}{q(x)} = ?$$

$$\nabla_x q(x) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{d} \frac{-2(x - x^{(i)})}{2\sigma^2} e^{\left(\frac{-\|x - x^{(i)}\|^2}{2\sigma^2}\right)}$$

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

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

2) ب) معایب استفاده از کرنل گاوسی در Score matching

1- استفاده از کرنل پیچیدگی محاسباتی بالا دارد. اگر تعداد داده‌ها نیز بالا برود این پیچیدگی بسیار بیشتر می‌شود.

2- حساسیت بالای کرنل گاوسی به میزان σ دارد. اگر بسیار بزرگ یا کوچک انتخاب شود می‌تواند باعث over-smoothing و under-smoothing می‌شود و دقت خوبی در تخمین چگالی توزیع ممکن است نداشته باشد.

3- افزایش بُعد فضای داده باعث ضعیف شدن عملکرد کرنل گاوسی می‌شود. چون نقاط بیشتر در فاصله یکسان و دور از هم در ابعاد بالا قرار می‌گیرند و کیفیت تخمین چگالی را کاهش می‌دهند.

4- کرنل گاوسی ممکن است در نقاط مرزی داده‌ها دچار انحراف و مشکل باشد این بخاطر فراتر بودن کرنل گاوسی از محدوده واقعی داده‌هاست.

ج)

$$\begin{cases} J_1(\theta) = E_{q(x)}\left[\frac{1}{2}\left\| s_\theta(x) - \nabla_x \log q(x)\right\|^2\right] \\ J_2(\theta) = E_{q(x,x_0)}\left[\frac{1}{2}\left\| s_\theta(x) - \nabla_x \log q(x|x_0)\right\|^2\right] \end{cases}$$

$$J_1(\theta) = E_{q(x)}\left[\frac{1}{2}\left\| s_\theta(x)\right\|^2\right] - K(\theta) + c_1$$

$$c_2 = E_{q(x)}\left[\frac{1}{2}\left\| \nabla_x \log q(x)\right\|^2\right]$$ ثابت است و به θ وابسته نیست

$$\Rightarrow K(\theta) = E_{q(x)}\left[\left\langle s_\theta(x), \nabla_x \log q(x)\right\rangle\right]$$

$$= \int_x q(x)\left\langle s_\theta(x), \nabla_x \log q(x)\right\rangle dx$$

$$K(\theta) = \int_x q(x) \left\langle S_\theta(x), \frac{\nabla_x q(x)}{q(x)} \right\rangle dx$$

$$= \int_x \left\langle S_\theta(x), \nabla_x q(x) \right\rangle dx$$

$$= \int_x \left\langle S_\theta(x), \frac{\partial}{\partial x} \int_{x_0} q_0(x_0)\, q(x|x_0)\, dx_0 \right\rangle dx$$

$$= \int_x \left\langle S_\theta(x), \int_{x_0} q_0(x_0) \frac{\partial q(x|x_0)}{\partial x} dx_0 \right\rangle dx$$

$$= \int_x \left\langle S_\theta(x), \int_{x_0} q_0(x_0)\, q(x|x_0) \frac{\partial \log q(x|x_0)}{\partial x} dx_0 \right\rangle dx$$

$$= \int_x \int_{x_0} q_0(x_0)\, q(x|x_0) \left\langle S_\theta(x), \frac{\partial \log q(x|x_0)}{\partial x} \right\rangle dx_0\, dx$$

$$= \int_x \int_{x_0} q(x, x_0) \left\langle S_\theta(x), \frac{\partial \log q(x|x_0)}{\partial x} \right\rangle dx_0\, dx$$

$$= E_{q(x, x_0)} \left[\left\langle S_\theta(x), \frac{\partial \log q(x|x_0)}{\partial x} \right\rangle \right]$$

$$\Rightarrow J_1(\theta) = E_{q(x)} \left[\frac{1}{2} \| S_\theta(x) \|^2 \right] - E_{q(x, x_0)} \left[\left\langle S_\theta(x), \frac{\partial \log q(x|x_0)}{\partial x} \right\rangle \right] + C_1$$

$$J_2(\theta) = E_{q(x)}\left[\frac{1}{2}\|s_\theta(x)\|^2\right] - E_{q(x,x_0)}\left[\left\langle s_\theta(x), \frac{\partial \log q(x|x_0)}{\partial x}\right\rangle\right] + C_2$$

$$\Rightarrow J_1(\theta) - J_2(\theta) = C_1 - C_2 \Rightarrow \boxed{J_2(\theta) = J_1(\theta) + C_2 - C_1}$$

اگر $C_2 - C_1$ که ثابت است را برابر C بگیریم اثبات می‌شود که :

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

د)

$$q(x|x_0) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-\|x-x_0\|^2}{2\sigma^2}}$$

$$\log q(x|x_0) = -d\log(\sqrt{2\pi}\,\sigma) - \frac{\|x-x_0\|^2}{2\sigma^2}$$

$$\nabla_x \log q(x|x_0) = \frac{-1}{2\sigma^2}\nabla_x\|x-x_0\|^2 = -\frac{x-x_0}{\sigma^2}$$

$$J_2(\theta) = E_{q(x,x_0)}\left[\frac{1}{2}\left\|s_\theta(x) + \frac{x-x_0}{\sigma^2}\right\|^2\right]$$

برای آموزش این مدل بصورت زیر عمل می‌کنیم :

1. ابتدا از توزیع $q(x|x_0)$ سمپل می‌گیریم. این سمپل گیری ابتدا با سمپل گیری از توزیع x_0 و سپس با سمپل گیری از x در توزیع شرطی به شرط داشتن x_0 انجام می‌شود.

2. سپس بخش گرادیان $\nabla_x \log q(x|x_0)$ را محاسبه می‌کنیم

3. سپس Score function $s_\theta(x)$ را محاسبه می‌کنیم

4. از روی دو مرحله قبل تابع هزینه $J(\theta)$ را حساب می‌کنیم.

5. با استفاده الگوریتم gradient descent یا الگوریتم دیگر بهینه سازی، مقدار $J(\theta)$ را اپدیت می‌کنیم تا مینیمم شود.

- استفاده از Langevin Dynamics برای تولید نمونه

$$x_{t+1} = x_t + \delta \nabla_x \log p(x_t) + \sqrt{2\delta}\ \varepsilon \sim N(0, I)$$

1. از یک سمپل رندوم x_0 شروع می‌کنیم.

2. با استفاده از شبکه مقدار score $S_\theta(x_t) = \nabla_x \log p(x_t)$ را محاسبه و x را اپدیت می‌کنیم با توجه به فرمول Langevin dynamics.

3. این مراحل اپدیت را پشت سر هم تکرار می‌کنیم به تعداد کافی ای تا به توزیع هدف برسیم.

سوال 4

1. تحلیل پایداری پیوسته

آ) رابطه loss به شکل روبه رواست:

$$L(\theta, \psi) = f(\theta\psi) + \text{Constant}$$

بردار گرادیان ها برابراست با:

$$v(\theta, \psi) = \begin{pmatrix} -f'(\theta\psi)\,\psi \\ f'(\theta\psi)\,\theta \end{pmatrix}$$

چون $L(0,\psi) = L(\theta,0) = \text{Constant}$ است پس $(\theta,\psi) = (0,0)$ یک نقطه تعادل برای GAN است همچنین چون $f'(t) \neq 0$ for all $t \in \mathbb{R}$ پس زمانی $v(\theta,\psi) = 0$ است که $(\theta,\psi) = (0,0)$ باشد. پس نقطه $(0,0)$ تنها نقطه تعادل GAN می‌باشد.

سوال ④

ادامه ۱) آ) جاکوبین بردار گرادیان برابر است با:

$$v'(\theta, \psi) = \begin{pmatrix} f''(-\theta\psi)\,\psi^2 & -f'(-\theta\psi) + f''(-\theta\psi)\,\theta\psi \\ f'(\theta\psi) + f''(\theta\psi)\,\theta\psi & f''(\theta\psi)\,\theta^2 \end{pmatrix}$$

در نقطه $\theta = \psi = 0$ جاکوبین بشکل زیر است:

$$v'(0,0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & 0 \end{pmatrix}$$

با مقدار ویژه‌های $\lambda_{1,2} = \pm f'(0)\, i$ که موهومی هستند.

ب) با فرض روبرو:

$$R(\theta, \psi) = \frac{1}{2}\left(\theta^2 + \psi^2\right)$$

$$\frac{\partial R(\theta(t), \psi(t))}{\partial t} = -\theta(t)\,\psi(t)\, f'\big(\theta(t)\,\psi(t)\big) + \psi(t)\,\theta(t)\, f'\big(\theta(t)\,\psi(t)\big) = 0$$

چون گرادیان نسبت به t برابر با 0 است پس نتیجه می‌گیریم مقدار R در همه $t \in [0, \infty)$ ثابت است پس مقدار $\theta^2 + \psi^2 = \text{Constant}$ است.

ج) نتایج بخش بالا نشان می‌دهد که گرادیان کاهشی همزمان برای GAN غیراشباع و هر نرخ یادگیری ای $(h>0)$ بصورت محلی همگرا نیست. زیرا مقادیر ویژه جاکوبین عملگر بروزرسانی F_h از لحاظ قدرمطلق همگی بزرگتر از 1 هستند.

پس همگرایی خطی به سمت نقطه تعادل درحالت پیوسته رد می‌شود (برای $h \to 0$)

البته داینامیک آموزش پیوسته می‌تواند هنوز با نرخ همگرایی sublinear همگرا شود.

سوال 4

(2   (آ

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

$$v(\theta,\psi) = \begin{pmatrix} -f'(\theta\psi)\psi \\ f'(\theta\psi)\theta \end{pmatrix}$$

جاکوبین F برابراست با:

$$J = \begin{pmatrix} 1 - h f''(\theta,\psi)\psi^2 & -hf'(\theta,\psi) - h f''(\theta\psi)\theta\psi \\ hf'(\theta\psi) + hf''(\theta,\psi)\theta\psi & h f''(\theta\psi)\theta^2 \end{pmatrix}$$

در نقطه (0,0) برابراست با:

$$J(0,0) = \begin{pmatrix} 1 & -hf'(0) \\ hf'(0) & 1 \end{pmatrix}$$

مقادیر ویژه جاکوبین برابراست با:

$$\det(J - \lambda I) = 0 \Rightarrow (1-\lambda)^2 - (hf'(0))^2 = 0$$

$$1 - \lambda = \pm h f'(0)\, i$$

$$\boxed{\lambda_{1/2} = 1 \pm h f'(0)\, i}$$

چون اندازه $|\lambda_{1/2}| = \sqrt{1+h^2 f'^2(0)}$ است وبزرگتر از یک می‌باشد پس دو مقدار ویژه خارج دایره واحد قرار می‌گیرد.

$$\theta_{k+1}^2 + \psi_{k+1}^2 = \left(\theta_k - h f'(\theta_k\psi_k)\psi_k\right)^2 + \left(\psi_k + h f'(\theta_k\psi_k)\theta_k\right)^2$$

$$= \theta_k^2 + \psi_k^2 + h^2 f'(\theta_k\psi_k)^2(\theta_k^2 + \psi_k^2)$$

$$\geqslant \theta_k^2 + \psi_k^2$$

پس $\theta^2 + \psi^2$ بصورت یکنوا در طول آموزش افزایش می‌یابد.

2) ب)

$$F_1(\theta, \psi) = \begin{pmatrix} \theta - h f'(\theta\psi)\,\psi \\ \psi \end{pmatrix}$$

$$F_2(\theta, \psi) = \begin{pmatrix} \theta \\ \psi + h f'(\theta\psi)\,\theta \end{pmatrix}$$

جاکوبین نگاشت‌ها در $(0, 0)$ برابر است با:

$$F_1'(0,0) = \begin{pmatrix} 1 & -h f'(0) \\ 0 & 1 \end{pmatrix}$$

$$F_2'(0,0) = \begin{pmatrix} 1 & 0 \\ h f'(0) & 1 \end{pmatrix}$$

$$J = (F_2 \circ F_1)'(0,0) = F_2'(0,0) \cdot F_1'(0,0) = \begin{pmatrix} 1 & -h f'(0) \\ h f'(\theta) & 1 - h^2 f'(0)^2 \end{pmatrix}$$

$$\det(J - \lambda I) = 0 \Rightarrow 1 - h^2 f'(0)^2 - \lambda - \lambda + h^2 \lambda f'(0)^2 + \lambda^2 + h^2 f'(0)^2 = 0$$

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

ج) اگر $h f'(0) \leq 2$ باشد همه مقدار ویژه‌هایش روی دایره واحد قرار می‌گیرد

اگر $h f'(0) > 2$ باشد همه مقدار ویژه‌هایش خارج دایره واحد قرار می‌گیرد.

اگر مقادیر ویژه خارج دایره باشد همگرایی محلی به نقطه تعادل دچار مشکل می‌شود.

3) الف)

$$E_{\hat{\theta} \sim N(\theta, \sigma^2)} \left[f(\hat{\theta} \psi) \right] + E_{x \sim N(0, \sigma^2)} \left[- f(-x\psi) \right]$$

گرادیان را برای objective فوق محاسبه می‌کنیم:

$$\hat{v}(\theta, \psi) = E_{\hat{\theta}, x} \begin{pmatrix} - \psi f'(\hat{\theta} \psi) \\ \hat{\theta} f'(\hat{\theta} \psi) - x f'(-x\psi) \end{pmatrix}$$

جاکوبین گرادیان برابر است با:

$$\hat{v}'(\theta, \psi) = E_{\hat{\theta}, x} \begin{pmatrix} - f''(\hat{\theta} \psi) \psi^2 & - f'(\hat{\theta} \psi) - f''(\hat{\theta} \psi) \hat{\theta} \psi \\ f'(\hat{\theta} \psi) + f''(\hat{\theta} \psi) \hat{\theta} \psi & f''(\hat{\theta} \psi) \hat{\theta}^2 + x^2 f(-x\psi) \end{pmatrix}$$

در نقطه $(0,0)$ برابر است با:

$$\hat{v}'(0,0) = \begin{pmatrix} 0 & - f'(0) \\ f'(0) & 2 f''(0) \sigma^2 \end{pmatrix}$$

مقادیر ویژه را بصورت زیر بدست می‌آوریم:

$$\det(\hat{v}' - \lambda I) = 0 \Rightarrow -2 f''(0) \sigma^2 \lambda + \lambda^2 + f'(0)^2 = 0$$

$$\lambda_{1,2} = f''(0) \sigma^2 \pm \sqrt{f''(0)^2 \sigma^4 - f'(0)^2}$$

ب)