به نام خدا

یادگیری ژرف
نیم‌سال دوم 03 -02
مدرس: دکتر مهدیه سلیمانی

سید امیر کسائی- 402212214- همفکری با: امیر محمد عزتی

سوال اول: استنباط متغیر

$$q = \underset{q\epsilon\mathbb{Q}}{argmin}\, D_{KL}\big(q(z) \parallel p_\theta(z|x)\big)$$

1. $D_{KL}\big(q(z) \parallel p_\theta(z|x)\big) = \sum q(z) log\left(\frac{q(z)}{p_\theta(z|x)}\right) = \mathbb{E}_{z\sim q} log\left(\frac{q(z)}{p_\theta(z|x)}\right) = \mathbb{E}_{z\sim q}\log q(z) - \mathbb{E}_{z\sim q}\log p_\theta(z|x) = \mathbb{E}_{z\sim q}\log q(z) - \mathbb{E}_{z\sim q}\log p_\theta(z,x) + \log p_\theta(x)$

$$-ELBO = \mathbb{E}_{z\sim q}\log q(z) - \mathbb{E}_{z\sim q}\log p_\theta(z,x)$$
$$D_{KL}\big(q(z) \parallel p_\theta(z|x)\big) = -ELBO + \log p_\theta(x)$$
$$ELBO = \log p_\theta(x) - D_{KL}\big(q(z) \parallel p_\theta(z|x)\big)$$
$$Lower\ bound\ proof:\ \log p_\theta(x) = \log\int_z p(x,z) = log\int_z p(x,z)\frac{q(z)}{q(z)}$$
$$= log\left(E_q\left[\frac{p(x,z)}{q(z)}\right]\right) \geq E_q[\log p(x,z)] - E_q[\log q(z)]$$

$$\longrightarrow log\left(E_q\left[\frac{p(x,z)}{q(z)}\right]\right) \geq ELBO$$

2.

۱.در گام اول پارامتر های θ,ψ باید بصورت تصادفی مقدار دهی شوند.
۲. سپس باید عبارت زیر را maximum کنیم:

$$\sum_{i=1}^{N} maxELBO(q, x_i, \theta)$$

۳. تا زمانی که ELBO بهینه شود و converge کند (Using while loop) ، ψ_i را از طریق فرمول زیر محاسبه می‌کنیم:

$$\psi_i = argmax_{\psi_i}\mathcal{L}(\psi_i, \theta, x_i)$$

به این شکل N تا ψ_i بدست میاوریم.
۴. در همان حلقه‌ای که داشتیم، $\mathcal{L}(x_i, \psi_{1toN}, \theta)$ را محاسبه می‌کنیم. سپس با استفاده از مقادیر بدست آمده θ را از طریق فرمول زیر بدست میاوریم و این روند را تا همگرا شدن $\mathcal{L}$ ادامه می‌دهیم.

$$\theta = argmax_\theta \sum_{i=1}^{N} \mathcal{L}(x_i, \psi_i, \theta)$$

3.

آ) Stochastic VI: در این حالت هر batch دارای B داده است و این مرحله باید به الگوریتم بخش قبل اضافه شود. حال، به ازای هر نمونه، گرادیان های $\mathcal{L}(x_j,\psi_j,\theta)$ را نسبت به ψ_j,θ محاسبه می‌کنیم. که در واقع گرادیان نسبت به ψ_j,θ، میانگین گرادیان های B تا نمونه است. پس خواهیم داشت:

$$\hat{\nabla}_{\psi^B}\frac{1}{B}\sum_j \nabla_{\psi_j}\mathcal{L}(x_j,\psi_j,\theta), \hat{\nabla}_\theta = \frac{1}{B}\sum_j \nabla_\theta(x_j,\psi_j,\theta)$$

$$Updating\ \psi: \bar{\nabla}_{\psi_j}\mathcal{L}(x^B,\psi^B,\theta) \triangleq \left[I(\psi_j)^{-1}\right]\hat{\nabla}_{\psi^B}\mathcal{L}(x^B,\psi^B,\theta)$$

$$update\ \bar{\nabla}_{\psi_B} with\ \psi_j$$

$$Updating\ \theta: \bar{\nabla}_\theta\,\mathcal{L}(x^B,\psi^B,\theta) \triangleq [I(\theta)^{-1}]\hat{\nabla}_\theta\mathcal{L}(x^B,\psi^B,\theta)$$

$$update\ \bar{\nabla}_\theta with\ \theta$$

$$ELBO\ calculations:$$

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

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

ب) ابتدا θ و ϕ مقدار دهی می‌شوند. مثل بخش قبل گرادیان هر داده نسبت به دو متغیر θ و ϕ در هر batch محاسبه می‌شود. میانگین گرادیان ها روی B تا نمونه، گرادیان نهایی است.

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

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

4.

$$p_\theta(z) = \mathcal{N}(0,I)$$

$$p_\theta(x|z) = \mathcal{N}(f_\theta(z),\sigma^2 I)$$

$$q_\phi(z|x) = \mathcal{N}\left(\mu_\phi(x), diag\left(\sigma_\phi^2(x)\right)\right)$$

ELBO:

$$\log p_\theta(x) \geq \mathbb{E}_{q_{\phi(Z|X)}}[\log p_\theta(x|z)] - KL(q_\phi\,(z|x)||p_\theta(z))$$

ELBO Decomposition:

$$\mathcal{L}(\theta,\phi;x) = \mathbb{E}_{q_{\phi(Z|X)}}[\log p_\theta(x|z)] - KL(q_\phi\,(z|x)||p_\theta(z))$$

Using reparameterization trick:

For $q_\phi(z|x) = \mathcal{N}\left(\mu_\phi(x), diag\left(\sigma_\phi^2(x)\right)\right)$:

$$z = \mu_\phi(x) + \sigma_\phi(x)\odot\epsilon, \epsilon\sim\mathcal{N}(0,I)$$

۱.در مرحله اول θ,ϕ را مقدار دهی می‌کنیم.

۲. برای هر mini_batch از داده داریم:

$$\mu_\phi\left(x^{(i)}\right), \sigma_\phi\left(x^{(i)}\right) \leftarrow Encoder\left(x^{(i)}\right)$$
$$z^{(i)} = \mu_\phi\left(x^{(i)}\right) + \sigma_\phi\left(x^{(i)}\right) \odot \epsilon^{(i)}$$
$$\hat{x}^{(i)} = f_\theta\left(z^{(i)}\right)$$
$$\log p_\theta\left(x^{(i)} | z^{(i)}\right) = -\frac{1}{2\sigma^2}\left\|x^{(i)} - \hat{x}^{(i)}\right\|^2 - \frac{D}{2}\log(2\pi\sigma^2)$$
$$KL(q_\phi\left(z \middle| x^{(i)} \parallel p_\theta(z)\right) = \frac{1}{2}\sum_{j=1}^{D}\left(\mu_{\phi,j}\left(x^{(i)}\right)^2 + \sigma_{\phi,j}\left(x^{(i)}\right)^2 - log\sigma_{\phi,j}\left(x^{(i)}\right)^2 - 1\right)$$

ELBO:

$$\mathcal{L}(\theta, \phi; x) = \log p_\theta\left(x^{(i)} | z^{(i)}\right) - KL(q_\phi\left(z \middle| x^{(i)} \parallel p_\theta(z)\right)$$

سوال دوم: Diffusion Models

آ)

$$x_1 = ax_0 + \sigma_1\varepsilon_1$$
$$var(x_1) = a^2 + \sigma_1^2$$

اگر واریانس x1 را برابر ۱ بگذاریم، می‌توانیم واریانس را بدون تغییر نگه داریم.

$$a^2 + \sigma_1^2 = 1 \rightarrow a = \sqrt{1 - \sigma_1^2}$$

که این برای هر time step صدق می‌کند. پس:

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

ب)

$$q(z_s, z_t | x) = q(z_t | z_s) q(z_s | x)$$
$$q(z_s, z_t | x) = \frac{q(z_s, z_t, x)}{q(x)}$$
$$q(z_s, z_t, x) = q(z_t | z_s, x).q(z_s, x) \xrightarrow{z_t only\ depends\ on\ z_s} q(z_t | z_s, x) \approx q(z_t | z_s)$$
$$q(z_s, z_t | x) = q(z_t | z_s).q(z_s, x)$$

ج)

$$We\ have\ q(z_t | x) = \mathcal{N}(a_t x, \sigma_t^2 I)$$
$$\xrightarrow{t>s} q(z_t | z_s) = \mathcal{N}\left(a_{t|s} z_s, \sigma_{t|s}^2 I\right)$$
$$a_{t|s} = \frac{a_t}{a_s}, \qquad \sigma_{t|s}^2 = \sigma_t^2 - a_{t|s}^2 \sigma_s^2$$

توزیع مشترک z_t و z_s را می‌توان به صورت زیر نوشت:

$$\binom{z_s}{z_t} \sim \mathcal{N}\left(\binom{0}{0}, \begin{pmatrix} \sigma_s^2 I & C(z_s, z_t) \\ C(z_s, z_t)^T & \sigma_t^2 I \end{pmatrix}\right)$$

$$Mean = A_{z_s}, Covariance = \sum$$

$$A = C(\sigma_s^2 I)^{-1} = \frac{C^T}{\sigma_s^2}, \sum = \sigma_t^2 I - C^T(\sigma_s^2 I)^{-1} C = \sigma_t^2 I - \frac{C^T C}{\sigma_s^2}$$

$$C = a_t \sigma_s^2 I$$

$$A = \frac{(a_t \sigma_s^2 I)}{\sigma_s^2} = a_t I$$

$$\sum = \sigma_t^2 I - \frac{(a_t \sigma_s^2 I)(a_t \sigma_s^2 I)}{\sigma_s^2} = \sigma_t^2 I - a_t^2 \sigma_s^2 I = (\sigma_t^2 - a_t^2 \sigma_s^2) I$$

$$A = a_t I = \left(\frac{a_t}{a_s} a_s\right) I = a_{t|s} a_s I$$

$$A_{z_s} = a_{t|s} a_s z_s$$

د) به طور کلی وقتی یک prior گاوسی به شکل $p(x) = \mathcal{N}(\mu_A, \sigma_A^2)$ و یک احتمال خطی ـ گاوسی به شکل $p(y|x) = \mathcal{N}(ax, \sigma)$ داریم، راه حل کلی برای posterior به شکل $p(x|y) = \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ است. که $\tilde{\sigma}^{-2} = \sigma_A^{-2} + a^2 \sigma_B^{-2}$ و $\tilde{\mu} = \tilde{\sigma}^{-2}(\sigma_A^{-2} \mu_A + a \sigma_B^{-2} y)$.

با اعمال کردن $q(z_s|z_t, x) = \mathcal{N}(a_s x, \sigma_S^2 I)$ و همچنین احتمال خطی ـ گاوسی $q(z_t|z_s) = \mathcal{N}(a_{t|s} z_s, \sigma_{t|s}^2 I)$ در این معادله کلی به posterior زیر می‌رسیم.

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

$$where\ \sigma_Q^{-2}(s,t) = \sigma_S^{-2} + a_{t|s}^2 \sigma_{t|s}^{-2} = \frac{\sigma_t^2}{\sigma_{t|s}^2 \sigma_S^2}$$

$$\sigma_Q^2 = \frac{\sigma_{t|s}^2 \sigma_S^2}{\sigma_t^2}$$

$$and\ \mu_Q(z_t, x; s, t) = \frac{\sigma_{t|s}^2 \sigma_S^2}{\sigma_t^2} z_t + \frac{a_s \sigma_{t|s}^2}{\sigma_t^2} x$$

۵)

$$D_{KL}(P||Q) = \int p(x) \log \frac{q(x)}{p(x)} d_x = E_{p(x)}[\log \frac{q(x)}{p(x)}]$$

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

$$D_{KL} = E_p[\log \frac{q}{p_\theta}] = E_p[\frac{-1}{2\sigma_Q^2}(z^2 + \mu_Q^2 - 2z\mu_Q - (z^2 + \mu_\theta^2 - 2z\mu_\theta))]$$
$$= E_p[\frac{-1}{2\sigma_Q^2}(\mu_Q^2 - \mu_\theta^2 + 2z\big(\mu_\theta - \mu_Q\big)]$$
$$D_{KL} = \frac{-1}{2\sigma_Q^2}[2\big(\mu_\theta - \mu_Q\big)E_p[z] + \mu_Q^2 - \mu_\theta^2]$$
$$D_{KL} = \frac{-1}{2\sigma_Q^2}\left|\left|\mu_Q - \mu_\theta\right|\right|_2^2$$

و)

$$D_{KL}(q(z_s|z_t, x)||p(z_s|z_t)) = \frac{1}{2\sigma_Q^2(s,t)}\left\|\mu_Q - \mu_\theta\right\|_2^2$$
$$= \frac{\sigma_t^2}{2\sigma_{t|s}^2\sigma_S^2}\frac{\sigma_{t|s}^4 a_S^2}{\sigma_t^4}\|x - \hat{x}_\theta(z_t;t)\|_2^2$$
$$= \frac{1}{2\sigma_S^2}\frac{a_S^2\sigma_{t|s}^2}{\sigma_t^2}\|x - \hat{x}_\theta(z_t;t)\|_2^2$$
$$= \frac{1}{2\sigma_S^2}\frac{a_S^2\big(\sigma_t^2 - a_{t|s}^2\sigma_S^2\big)}{\sigma_t^2}\ \|x - \hat{x}_\theta(z_t;t)\|_2^2$$
$$= \frac{1}{2\sigma_S^2}\frac{\frac{a_S^2\sigma_t^2}{\sigma_S^2} - a_t^2}{\sigma_t^2}\|x - \hat{x}_\theta(z_t;t)\|_2^2$$
$$= \frac{1}{2}\left(\frac{a_S^2}{\sigma_S^2} - \frac{a_t^2}{\sigma_t^2}\right)\|x - \hat{x}_\theta(z_t;t)\|_2^2$$
$$= \frac{1}{2}\big(SNR(s) - SNR(t)\big)\|x - \hat{x}_\theta(z_t;t)\|_2^2$$

ز)

$$\mathcal{L}_T(x) = \frac{T}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I), i\sim U\{1,T\}}\big[\big(SNR(s) - SNR(t)\big)\|x - \hat{x}_\theta(z_t;t)\|_2^2\big]$$
$$s = \frac{(i-1)}{T}, t = \frac{i}{T}, z_t = a_t x + \sigma_t\epsilon$$
$$\mathcal{L}_T(x) = \frac{1}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I), i\sim U\{1,T\}}\left[\frac{SNR(t-\tau) - SNR(t)}{\tau}\|x - \hat{x}_\theta(z_t;t)\|_2^2\right]$$
$$As\ \tau \to 0, T \to \infty$$
$$\mathcal{L}_\infty(x) = -\frac{1}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I), t\sim U[0,1]}[SNR'(t)\|x - \hat{x}_\theta(z_t;t)\|_2^2]$$
$$= -\frac{1}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\int_0^1 SNR'(t)\|x - \hat{x}_\theta(z_t;t)\|_2^2 dt$$

سوال سوم: Score Matching

Langevin Dynamics

آ)

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

اگر ترم مربوط به نویز وجود نداشته باشد:

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

این معادله فرآیندی را توصیف می‌کند که در آن نمونه x_t به صورت تکراری در جهت گرادیان تابع log p(x) تنظیم می‌شود.

با یک نقطه تصادفی x_0 در R^d شروع می‌کنیم. در هر تکرار داریم:

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

که $\nabla_x \log p(x_t)$ گرادیان تابع چگالی احتمال در x_t است و به تند ترین صعود $\log p(x_t)$ اشاره می‌کند. δ نیز اندازه گام است که مقدار بزرگی update را کنترل می‌کند. قانون به روز رسانی به طور موثر عمل صعود گرادیان را در $\log p(x_t)$ انجام می‌دهد. با شروع از یک نقطه تصادفی x_0، نمونه x_t به طور مکرر در جهتی حرکت می‌کند که $p(x_t)$ با بیشترین سرعت افزایش پیدا کند. این روند تا زمانی ادامه دارد که x_t به ماکسیمم محلی $\log p(x_t)$ برسد که مربوط به پیک $p(x_t)$ است.

از آن جایی که نقطه اولیه x_0 به طور تصادفی در R^d انتخاب می‌شود، ممکن است در هر نقطه از فضای تعریف شده توسط $p(x_t)$ باشد. هر به روز رسانی x_t را در جهت افزایش $\log p(x_t)$ حرکت می‌دهد. در طی چندین تکرار تضمین می‌شود که x_t از گرادیان بالا می‌رود و در نهایت به اوج می‌رسد. به طور خلاصه، معادله بدست آمده معادل شروع از یک نقطه تصادفی در فضای R^d و انجام gradient ascent رو تابع چگالی احتمال است. این فرآیند منجر به همگرایی در یکی از قله های $p(x_t)$ می‌شود. جایی که چگالی احتمال بالاترین است.

در MLE به دنبال پیدا کردن پارامتر های توزیع هستیم. بطوریکه داده ها به خوبی fit شوند. در واقع می‌خواهیم پارامتر های توزیع p(x) را طوری تخمین بزنیم که با توجه به داده ها، احتمال ماکسیمم شود. اما در این حالت از یک نقطه تصادفی شروع می‌کنیم و بعد به سمتی حرکت می‌کنیم که توزیع به حداکثر برسد.

ب) یکی از دلایل استفاده از نویز نرمال این است که به عنوان regularization عمل می‌کند و از overfit شدن score function و عدم داشتن یک generalization خوب جلوگیری می‌کند. همچنین باعث اطمینان پیدا کردن از این است که فرآیند sampling در نقاط بهینه محلی گیر نمی‌کند و کل فضای توزیع را جستجو می‌کند. بعلاوه از داشتن نمونه های یکسان جلوگیری می‌کند و باعث می‌شود نمونه های متنوعی داشته باشیم.

Langevin dynamics در زمان برخورد با قله های تیز، بخاطر وجود نویز به اطراف حرکت می‌کند و در برخورد با قله های صاف نیز اطراف آن را بررسی می‌کند تا از گیر کردن در بهینه های محلی جلوگیری کند.

آ)

$$q(x) = \frac{1}{M}\sum_{i=1}^{M} K\left(x \middle| x^{(i)}\right)$$

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

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

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

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

ب) کرنل گاوسی به مقدار σ حساسیت بالایی دارد و انتخاب نادرست آن می‌تواند باعث under_smoothing و یا over_smoothing شود و بعلاوه ممکن است دقت خوبی در تخمین چگالی توزیع نداشته باشد. همچنین استفاده از آن پیچیدگی محاسباتی را بالا می‌برد و این افزایش پیچیدگی با افزایش داده ها ارتباط مستقیم دارد. بعلاوه اگر بعد فضایی داده زیاد شود، عملکرد کرنل دچار مشکل می‌شود. زیرا دقت تخمین چگالی هنگامی که نقاط بیشتر در فواصل دور و یکسان در ابعاد بالاتر قرار بگیرد، کاهش پیدا می‌کند.و در آخر، به دلیل فراتر بودن کرنل از محدوده حقیقی داده ها، امکان وجود انحراف در نقاط مرزی داده ها وجود دارد.

ج)

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

$$J_1(\theta) = E_{q(x)}\left[\frac{1}{2}\|s_\theta(x)\|^2\right] - K(\theta) + C_1$$

$$C_2 = E_{q(x)}\left[\frac{1}{2}\|\nabla_x \log q(x)\|^2\right]$$

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

$$= \int_x q(x)\,\langle s_\theta(x), \frac{\nabla_x q(x)}{q(x)}\rangle\, d_x$$

$$= \int_x \langle s_\theta(x), \nabla_x q(x)\rangle\, d_x = \int_x \langle s_\theta(x), \frac{\partial}{\partial_x}\int_{x_0} q_0(x_0) q(x|x_0)\, d_{x_0}\rangle\, d_x$$

$$= \int_x \langle s_\theta(x), \int_{x_0} q_0(x_0) \frac{\partial q(x|x_0)}{\partial_x}\ d_{x_0} \rangle\, d_x$$

$$= \int_x \langle s_\theta(x), \int_{x_0} q_0(x_0) q(x|x_0) \frac{\partial \log q(x|x_0)}{\partial_x}\ d_{x_0} \rangle\, d_x$$

$$= \int_x \int_{x_0} q_0(x_0) q(x|x_0) \langle s_\theta(x), \frac{\partial \log q(x|x_0)}{\partial_x} \rangle\ d_{x_0} d_x$$

$$= \int_x \int_{x_0} q(x, x_0) \langle s_\theta(x), \frac{\partial \log q(x|x_0)}{\partial_x} \rangle\ d_{x_0} d_x$$

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

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

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

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

$$J_1(\theta) - J_2(\theta) = C_1 - C_2$$

$$\rightarrow J_2(\theta) = J_1(\theta) - C_1 + C_2 \rightarrow J_2(\theta) = J_1(\theta) + C$$

د)

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

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

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

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

ابتدا از توزیع x_0 نمونه می‌گیریم و سپس از x در توزیع شرطی به شرط داشتن x_0. در واقع از توزیع $q(x|x_0)$، سمپل می‌گیریم. در ادامه $\nabla_x \log q(x|x_0)$ را محاسبه می‌کنیم و سپس score function (s_θ) را بدست می‌آوریم.

سوال چهارم: آنالیز پایداری GAN

بخش اول: تحلیل پایداری پیوسته

آ)

$$GAN\ training\ objective: L(\theta,\psi) = f(\psi\theta) + f(0)$$
$$Loss: L(\theta,\psi) = f(\theta\psi) + const$$
$$Gradient\ vector\ field: v(\theta,\psi) = \begin{pmatrix} -f'(\theta\psi)\psi \\ f'(\theta\psi)\theta \end{pmatrix}$$

چون $L(\theta,0) = L(0,\psi) = const$ برای همه $\theta,\psi\epsilon\mathbb{R}$، $(\theta,\psi) = (0,0)$ حتما نقطه تعادل است زیرا ما در نظر می‌گیریم که برای همه $t\epsilon\mathbb{R}$، $f'(t) \neq 0$ و داریم $v(\theta,\psi) = 0$ اگر و تنها اگر $(\theta,\psi) =$ $(0,0)$ که نشان می‌دهد این نقطه قطعا نقطه تعادل یکتا است.

بعلاوه ژاکوبین $v'(\theta,\psi) of v$ بصورت زیر است:

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

با در نظر گرفتن $(\theta,\psi) = (0,0)$ خواهیم داشت:

$$v'(0,0) = \begin{pmatrix} 0 & -f'(0) \\ f'(0) & 0 \end{pmatrix}\ which\ has\ the\ eigenvalues\ \pm f'(0)i$$

ب)

در نظر بگیریم که $R(\theta,\psi) := \frac{1}{2}(\theta^2 + \psi^2)$ آنگاه:

$$\frac{d}{dt}R\big(\theta(t),\psi(t)\big) = \theta(t)v_1\big(\theta(t),\psi(t)\big) + \psi(t)v_2\big(\theta(t),\psi(t)\big) = 0$$

که نشان می‌دهد $R(\theta,\psi)$ برای همه $t\epsilon[0,\infty)$ ثابت است.

ج) چون مقادیر ویژه ژاکوبین عملگر F_h بزرگ تر از یک هستند (از نظر قدر مطلقی)، و با توجه به نتیجه گیری بخش های قبل مشخص است که هر نرخ یادگیری محلی همگرا نیست و همچنین، گرادیان کاهشی همزمان GAN غیر اشباع می‌باشد. دانامیک آموزش پیوسته همچنان این امکان را دارد که با نرخ همگرایی sublinear، همگرا شود اما در حالت پیوسته، همگرایی خطی به سمت نقطه تعادل رد می‌شود.

بخش دوم: تحلیل پایداری گسسته

آ) مقادیر ویژه ژاکوبین به روز رسانی عملگر ها برای simultaneous gradient descent با $\lambda = 1 +$ $h\mu$ حساب می‌شوند. در نظر بگیریم که $v'(\theta^*,\psi^*)$ فقط دارای مقادیر ویژه با قسمت حقیقی منفی می‌باشد. مقادیر ویژه ژاکوبین به روز رسانی عملگر F_h برای simultaneous gradient descent همه در دایره واحد هستند اگر و تنها اگر:

$$h < \frac{1}{|Re(\lambda)|}\frac{2}{1 + \left(\frac{Im(\lambda)}{Re(\lambda)}\right)^2}$$
$$for\ all\ eigenvalues\ \lambda\ of\ v'(\theta^*,\psi^*)$$

اثبات:

برای simultaneous gradient descent داریم:

$$F_h(\theta, \psi) = (\theta, \psi) + hv(\theta, \psi)$$

از این رو $F_h'(\theta^*, \psi^*) = I + hv'(\theta^*, \psi^*)$. بنابراین مقادیر ویژه با $\lambda = 1 + h$ حساب می‌شوند.

برای اینکه ببینیم چه زمانی $|\lambda| < 1$ می‌گوییم $\mu = -a + ib$ با $a, b \epsilon \mathbb{R}$ و $a > 0$. پس:

$$|\lambda|^2 = (1 - ha)^2 + h^2 b^2$$

که کوچک تر از ۱ است اگر و تنها اگر:

$$h < \frac{2a}{a^2 + b^2}$$

برای اثبات افزایش یکنوای $\theta^2 + \psi^2$:

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

ب)

$Update\ operators\ for\ alternating\ gradient\ descent$:

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

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

$The\ Jacobian\ of\ these\ operators\ at\ 0\ are\ given\ by$:

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

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

$Jacobian\ of\ the\ combined\ update\ operator$:

$$\left(F_2^{n_d} \circ F_1^{n_g}\right)'(0,0) = F_2'(0,0)^{n_d} \cdot F_1'(0,0)^{n_g}$$

$$= \begin{pmatrix} 1 & -n_g hf'(0) \\ n_d hf'(0) & -n_g n_d h^2 f'(0)^2 + 1 \end{pmatrix}$$

$Eigenvalues\ of\ the\ given\ matrix\ are$:

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

ج) مقادیر ویژه خارج از دایره ←همگرایی محلی به نقطه تعادل مشکل دار می‌شود.

اگر $hf'(0) \leq 2$ ← همه مقادیر ویژه روی دایره واحد

اگر $hf'(0) > 2$ ← همه مقادیر ویژه خارج دایره واحد

بخش سوم: instance noise

آ)

موقع استفاده از instance noise داریم:

$GAN\ training\ objective$:

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

$Corresponding\ gradient\ vector\ field\ is\ given\ by$:

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

$$Jacobian\ \tilde{v}'(\theta,\psi): E_{\tilde{\theta},x}\begin{pmatrix} -f''\left(\tilde{\theta}\psi\right)\psi^2 & -f'\left(\tilde{\theta}\psi\right) - f''\left(\tilde{\theta}\psi\right)\tilde{\theta}\psi \\ f'\left(\tilde{\theta}\psi\right) + f''\left(\tilde{\theta}\psi\right)\tilde{\theta}\psi & f''\left(\tilde{\theta}\psi\right)\tilde{\theta}^2 + x^2 f(-x\psi) \end{pmatrix}$$

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

$Eigenvalues\ are\ given\ by$:

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