1.1

$$\hat{q} = \arg\min_q D_{KL}\Big(q(z) \,\Big\|\, p_\theta(z|x)\Big) = \arg\min_q \mathbb{E}\left[\lg \frac{q(z)}{p_\theta(z|x)}\right]$$

$$= \arg\min_q \mathbb{E}\big[\lg q(z)\big] - \mathbb{E}\Big[\lg\big(p_\theta(x|z)\, p_\theta(z)\big)\Big] + \underbrace{\mathbb{E}\big[\lg p_\theta(x)\big]}_{\lg p_\theta(x) \longrightarrow \text{const}}$$

$$\longrightarrow \hat{q} = \arg\max_q \mathbb{E}\Big[\lg\big(p_\theta(x|z)\, p_\theta(z)\big) - \lg q(z)\Big]$$

$$= \arg\max_q \mathbb{E}\big[\lg p_\theta(x|z)\big] - D_{KL}\Big(q(z) \,\Big\|\, p_\theta(z)\Big) = \arg\max_q ELBO$$

$$\lg p_\theta(x) = \lg \int p_\theta(x,z)\, dz = \lg \underbrace{\int p_\theta(x,z) \frac{q(z)}{q(z)}\, dz}_{\mathbb{E}_{q(z)}\left[\frac{p(x,z)}{q(z)}\right]}$$

$$\xrightarrow[\text{inequality}]{\text{Jensen's}} \lg p_\theta(x) \geqslant \mathbb{E}_{q(z)}\left[\lg \frac{p(x,z)}{q(z)}\right] = \mathbb{E}_{q(z)}\left[\lg \frac{p(x|z)\,p(z)}{q(z)}\right] = \underbrace{\mathbb{E}_{q(z)}\big[\lg p_\theta(x|z)\big] - D_{KL}\Big(q(z) \,\Big\|\, p_\theta(z)\Big)}_{ELBO}$$

1.2

ابتدا لاس را به صورت مقابل در نظر می‌گیریم $L(x_i, \psi, \theta) = -ELBO(x_i, q, \theta)$ و توسط الگوریتم زیر آنرا کمینه می‌کنیم

$\theta, \psi \longleftarrow$ random init

while $L(x_i; \psi, \theta)$ has not converged :

    for $i \in 1 \dots N$:

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

    calc $L(x; \psi, \theta)$

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

return ψ, θ

1.3.a

stochastic VI

اگر D را معادل dataloader با $\text{batch size} = M$ بدانیم و B یک باشد داریم:

$$I(\psi_i) \triangleq \mathbb{E}_q\left[\nabla_{\psi_i} \lg q(z|x_i;\psi_i)\ \nabla_{\psi_i} \lg q(z|x_i;\psi_i)^T\right]$$

$$I(\theta) \triangleq \mathbb{E}_{p(x_i, z;\theta)}\left[\nabla_\theta \lg p(x_i, z;\theta)\ \nabla_\theta \lg p(x_i, z;\theta)^T\right]$$

⊛ منظور از ψ همان $\psi_{1:N}$ است

- $\theta, \psi \leftarrow$ random init
- while $L(x_i;\psi,\theta)$ has not converged:
  - for $B \in D$:
    - for $x_j \in B$:
      - $l_j \leftarrow$ calc $L(x_j, \psi_j; \theta)$
      - $g_j \leftarrow$ calc $\nabla_{\psi_j} L(x_j, \psi_j; \theta)$
      - $h_j \leftarrow$ calc $\nabla_\theta L(x_j, \psi_j, \theta)$
    - $\hat{g} = \frac{1}{M}\sum_{j=1}^{M} g_j$, $\hat{h} = \frac{1}{M}\sum_{j=1}^{M} h_j$
    - for $x_j \in B$:
      - $\tilde{g} = [I(\psi_j)]^{-1}\hat{g}$
      - $\psi_j \leftarrow$ update using $\tilde{g}$
    - $\bar{h} = [I(\theta)]^{-1}\hat{h}$
    - $\theta \leftarrow$ update using $\bar{h}$
    - $\hat{L}(B;\psi,\theta) = \sum_{j=1}^{M} l_j$
  - $L(x;\psi,\theta) \simeq \frac{N}{M}\hat{L}(B,\psi,\theta)$
- return ψ, θ

$\boxed{1.3.b}$

(gradient ascent) Amortized VI

$\theta, \phi \leftarrow$ random init

while $L(x;\phi,\theta)$ has not converged:

  for $B \in D$:

    for $x_j \in B$:

      $l_j \leftarrow$ calc $L(x_j;\phi,\theta)$

      $g_j \leftarrow$ calc $\nabla_\phi L(x_j;\phi,\theta)$

      $h_j \leftarrow$ calc $\nabla_\theta L(x_j;\phi,\theta)$

    $\hat{g} = \frac{1}{M}\sum_{j=1}^{M} g_j \quad , \hat{h} = \frac{1}{M}\sum_{j=1}^{M} h_j$

    $\phi \leftarrow$ update using $\hat{g}$

    $\theta \leftarrow$ update using $\hat{h}$

    $\hat{L}(B;\phi,\theta) = \sum_{j=1}^{M} l_j$

  $L(x;\phi,\theta) \simeq \frac{N}{M}\hat{L}(B;\phi,\theta)$

return ϕ, θ

$\boxed{1.4}$

$$ELBO = \mathbb{E}_{q_\phi(z)}\left[\lg \frac{1}{(2\pi)^{\frac{k}{2}}\ \sigma^k} - \frac{1}{2\sigma^2}\left\|x - f_\theta(z)\right\|^2 \right]$$

$$- \frac{1}{2}\left[k\,\sigma_\phi^2(x) + \left\|\mu_\phi(x)\right\|^2 - k + \underbrace{\lg \frac{1}{\sigma_\phi^{2k}(x)}}_{-2k \lg \sigma_\phi(x)} \right]$$

$$loss = -ELBO$$

$$\rightarrow \quad \nabla_\theta\, loss = \nabla_\theta\, \mathbb{E}_{q_\phi(z)}\left[\frac{-1}{2\sigma^2}\left\|x - f_\theta(z)\right\|^2 \right] = \frac{-1}{2\sigma^2}\, \mathbb{E}\Big[\underbrace{\nabla_\theta \left\|x - f_\theta(z)\right\|^2}_{2\left(f_\theta(z) - x\right) f_\theta'(z)} \Big]$$

برای ∇_ϕ اما نمی‌توان گرادیان را داخل $\mathbb{E}$ برد. بنابراین از reparameterization استفاده می‌کنیم:

$$\nabla_\phi\, loss = \nabla_\phi\, \mathbb{E}_{q_\phi(z)}\left[\frac{-1}{2\sigma^2}\left\|x - f_\theta(z)\right\|^2 \right] - \frac{k}{2}\, \nabla_\phi\, \sigma_\phi^2(x) + \nabla_\phi \left\|\mu_\phi(x)\right\|^2 - 2k\, \nabla_\phi \lg \sigma_\phi(x)$$

$$\underset{\varepsilon \sim q(\varepsilon)}{\overset{\hat{z} = \mu + \sigma\varepsilon}{=\!=\!=}} \frac{-1}{2\sigma^2}\, \mathbb{E}_{q(\varepsilon)}\Big[\underbrace{\nabla_\phi \left\|x - f_\theta(\hat{z})\right\|^2}_{2\left(f_\theta(\hat{z}) - x\right) f_\theta(\hat{z})\, \hat{z}'} \Big] - k\,\sigma_\phi(x)\,\sigma_\phi'(x) + 2\mu_\phi(x)\,\mu_\phi'(x) - 2k\, \frac{\sigma_\phi'(x)}{\sigma_\phi(x)}$$

ε مستقل از ϕ است

با توجه به مشتق های به دست آمده می‌توان این مقادیر را در الگوریتم های قبلی جایگذاری کرد
یا اگر توابع را داشته باشیم شاید بتوان مشتق ها را برابر 0 گذاشت و با یک گام به جواب رسید

$\boxed{2.a}$ $$q(z_t|x) = \mathcal{N}(a_t x, \sigma_t^2 I) \longrightarrow z_t = a_t x + \sigma_t \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, I)$$

$$\longrightarrow \operatorname{Var}[z_t] = \underbrace{\operatorname{Var}[a_t x]}_{a_t^2 I} + \underbrace{\operatorname{Var}[\sigma_t \varepsilon]}_{\sigma_t^2 I} = (1-\sigma_t^2) I + \sigma_t^2 I = I$$

بنابراین اگر $a_t^2 = 1 - \sigma_t^2$ واریانس ثابت می‌ماند

$\boxed{2.b}$

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

چون مارکو است با داشتن z_s .
z_t از x مستقل است

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

$\boxed{2.c}$

طبق بخش $\boxed{2.a}$

$$z_t = a_t x + \sigma_t \varepsilon$$

$$\underbrace{z_s = a_s x + \sigma_s \varepsilon}_{x = \frac{z_s - \sigma_s \varepsilon}{a_s}}$$

$$\longrightarrow z_t | z_s = a_t \frac{z_s - \sigma_s \varepsilon}{a_s} + \sigma_t \varepsilon$$

$$\mathbb{E}[z_t | z_s] = \frac{a_t}{a_s} z_s$$

$$\operatorname{Var}[z_t | z_s] = \left(\frac{a_t}{a_s}\right)^2 \underbrace{\operatorname{Var}[z_s]}_{0} - \left(\frac{a_t}{a_s}\right)^2 \sigma_s^2 I + \sigma_t^2 I$$

$$= \left(\sigma_t^2 - a_{t|s}^2 \sigma_s^2\right) I$$

چون z_s را به صورت given داریم پس واریانس
آن صفر است (deterministicly آنرا می‌دانیم)

بنابراین $q(z_t|z_s)$ یک گاوسی است که میانگین و واریانس آنرا به‌دست آوردیم

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

$\boxed{2.d}$

$$q(z_s|z_t,x)=\frac{q(z_t|z_s,\overset{\text{مارکو}}{\cancel{x}})\,q(z_s|x)}{q(z_t|x)}\propto q(z_t|z_s)\,q(z_s|x)$$

این دو گاوسی را از بخش های قبل داریم. می دانیم که برای ضرب دو گاوسی فرمول زیر برقرار است :

$$\overset{q(z_s|x)}{\mathcal{N}(\mu_A,\sigma_A^2)}\times\overset{q(z_t|z_s)}{\mathcal{N}(ax,\sigma_B^2)}=\mathcal{N}(\tilde{\mu},\tilde{\sigma}^2 I)$$

$$\tilde{\sigma}^{-2}=\sigma_A^{-2}+a^2\sigma_B^{-2}$$

$$\tilde{\mu}=\tilde{\sigma}^2\left(\sigma_A^{-2}\mu_A+a\,\sigma_B^{-2}\overset{z_t}{y}\right)$$

بنابراین برای مسئله مان داریم :

$$\underbrace{\tilde{\sigma}^{-2}}_{\sigma_Q^{-2}}=\sigma_s^{-2}+a_{t|s}^2\sigma_{t|s}^{-2}=\frac{1}{\sigma_s^2}+\frac{a_{t|s}^2}{\sigma_{t|s}^2}\longrightarrow\tilde{\sigma}^2=\frac{\sigma_{t|s}^2\,\sigma_s^2}{\underbrace{\sigma_{t|s}^2}_{\sigma_t^2-\cancel{a_{t|s}^2\sigma_s^2}}+\cancel{\sigma_s^2 a_{t|s}^2}}=\boxed{\frac{\sigma_{t|s}^2\,\sigma_s^2}{\sigma_t^2}}$$

$$\underbrace{\tilde{\mu}}_{\mu_Q}=\frac{\sigma_{t|s}^2\,\sigma_s^2}{\sigma_t^2}\left(\frac{a_s x}{\sigma_s^2}+\frac{a_{t|s}z_t}{\sigma_{t|s}^2}\right)=\boxed{\frac{a_{t|s}\,\sigma_s^2}{\sigma_t^2}z_t+\frac{a_s\,\sigma_{t|s}^2}{\sigma_t^2}x}$$

$\boxed{2.e}$

طبق تعریف صورت سوال واریانس $p_\theta(z_s|z_t)$ برابر $\sigma_Q^2 I$ است پس $\Sigma_0=\Sigma_1$

$$D_{KL}\left(q(z_s|z_t,x)\,\|\,p_\theta(z_s|z_t)\right)=\frac{1}{2}\left(\underbrace{tr\left[\underbrace{(\sigma_Q^2 I)^{-1}\sigma_Q^2 I}_{I}\right]}_{d}+(\mu_\theta-\mu_Q)^T(\sigma_Q^2 I)^{-1}(\mu_\theta-\mu_Q)-d+\underbrace{\lg\frac{\det(\sigma_Q^2 I)}{\det(\sigma_Q^2 I)}}_{0}\right)$$

$$=\frac{1}{2\sigma_Q^2}\left((\mu_\theta-\mu_Q)^T(\mu_\theta-\mu_Q)\right)=\frac{1}{2\sigma_Q^2}\left\|\mu_Q-\mu_\theta\right\|_2^2$$

$\boxed{2.f}$

$$D_{KL} = \frac{1}{2\sigma_Q^2}\left\| \cancel{\frac{a_{t|s}\,\sigma_s^2}{\sigma_t^2}} z_t + \frac{a_s\,\sigma_{t|s}^2}{\sigma_t^2} x - \cancel{\frac{a_{t|s}\,\sigma_s^2}{\sigma_t^2}} z_t - \frac{a_s\,\sigma_{t|s}^2}{\sigma_t^2}\hat{x}_\theta \right\|_2^2$$

$$= \frac{a_s^2\,\sigma_{t|s}^4}{2\,\sigma_Q^2\,\sigma_t^4}\left\| x - \hat{x}_\theta \right\|_2^2 = \frac{a_s^2\left(\sigma_t^2 - a_{t|s}^2\,\sigma_s^2\right)}{2\sigma_s^2\,\sigma_t^2}\left\| x - \hat{x}_\theta \right\|_2^2 = \frac{1}{2}\Big(\underbrace{\frac{a_s^2}{\sigma_s^2}}_{SNR(s)} - \underbrace{\frac{a_t^2}{\sigma_t^2}}_{SNR(t)}\Big)\left\| x - \hat{x}_\theta \right\|_2^2$$

$\boxed{2.g}$

$$\begin{aligned} s(i) &= \frac{i-1}{T} \\ t(i) &= \frac{i}{T} \end{aligned} \longrightarrow s = t - \frac{1}{T}$$

$$L_T(x) = \frac{T}{2}\,\mathbb{E}\left[\left(SNR\left(t-\tfrac{1}{T}\right) - SNR(t)\right)\left\| x - \hat{x}_\theta \right\|_2^2\right]$$

$$= \frac{-1}{2}\,\mathbb{E}\left[\frac{SNR\left(t-\frac{1}{T}\right) - SNR(t)}{\frac{-1}{T}}\left\| x - \hat{x}_\theta \right\|_2^2\right]$$

$$\lim_{T\to\infty} L_T(x) = \frac{-1}{2}\,\mathbb{E}\left[SNR'(t)\left\| x - \hat{x}_\theta \right\|_2^2\right]$$

3.1.a

اگر عبارت نویز نباشد داریم: $x_{t+1} = x_t + \delta \nabla_x \lg p(x_t)$

همان طور که مشاهده می‌شود داریم بر روی $p(x)$ عملیات gradient ascent را انجام می‌دهیم (در فضای $x \in \mathbb{R}^d$ توزیع را بیشینه می‌کنیم)

بنابراین به یکی از قله‌های $p(x)$ خواهیم رسید.

در MLE مشتق توزیع را باید حساب کنیم (نسبت به پارامترهایش) تا بتوانیم $\hat{\theta}_{MLE}$ را به دست آوریم و به توزیع مد نظر برسیم. همچنین فرض MLE آن است که خانواده توزیع را می‌دانیم. به طور کلی ممکن است MLE جواب نداشته باشد زیرا ممکن است نتوانیم مشتق را به صورت فرم بسته به دست آوریم.

در عبارت Langevin Dynamics اما مشتق را نسبت به x می‌گیریم نه θ و ضمنا اینکه فرایند iterative را حساب می‌کنیم و gradient ascent می‌زنیم تا به بیشینه محلی برسیم.

3.1.b

دلیل استفاده از نویز ϵ ایجاد randomness در فرایند generation و بنابراین ایجاد تنوع در سمپل گیری است.

به این صورت حتی اگر از نقطه شروع یکسانی نیز شروع به حرکت کنیم، هر دفعه به سمپل‌های متفاوتی می‌رسیم.

اگر قله یافت شده تیز باشد با توجه به شیبی که در راستاهای مختلف آن نقطه وجود دارد ممکن است حرکتی انجام شود اما اگر flat باشد در آن نقطه گیر می‌کند

3.2.a

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

$$= \frac{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^d \left[-2\times\frac{1}{2\sigma^2}\,(x-x^{(i)})\exp\left(\frac{-\|x-x^{(i)}\|^2}{2\sigma^2}\right)\right]}{\frac{1}{m}\sum_{i=1}^{m} k(x\,|\,x^{(i)})}$$

$$= \frac{\sum_{i=1}^{m}\frac{1}{\sigma^2}(x^{(i)}-x)\,k(x\,|\,x^{(i)})}{\sum_{i=1}^{m}k(x\,|\,x^{(i)})}$$

3.2.b

در بخش قبل $q(x)$ را با کرنل گاوسی تخمین زدیم (GMM) که لزوما تخمین خوبی از توزیع نیست. به طور کلی در این روش به صورت explicit به دنبال یافتن توزیع رفتیم اما روش implicit که در denoising score matching استفاده می‌شود بهتر است

3.2.d

$$\nabla_x \lg q(x\,|\,x_0) = \nabla_x\left[-d\lg(\sqrt{2\pi}\,\sigma) - \frac{\|x-x_0\|^2}{2\sigma^2}\right] = \frac{-1}{2\sigma^2}\,2(x-x_0) = \frac{x_0-x}{\sigma^2}$$

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

می‌دانیم که در هر مرحله x نسبت به مراحل قبل واریانس بیشتری دارد.
یعنی نویز بیشتری به تصویر اضافه شده یا به اصطلاح manifold بیشتر پف کرده. فرض می‌کنیم که داریم: $q(x\,|\,x_0) = \mathcal{N}(x_0, \sigma^2 I)$

$$\longrightarrow J_2(\theta) = \mathbb{E}_{\substack{q(x,x_0)\\ \varepsilon}}\left[\frac{1}{2}\Big\|\underbrace{s_\theta(x)}_{-\frac{\varepsilon_\theta(x)}{\sigma}}+\frac{(x_0+\sigma\varepsilon)-x_0}{\sigma^2}\Big\|^2\right] = \frac{1}{2\sigma^2}\mathbb{E}_{\substack{q(x_0)\\ \varepsilon}}\left[\left\|\varepsilon-\varepsilon_\theta(x_0+\sigma\varepsilon)\right\|^2\right]$$

بنابراین در فرایند آموزش می‌خواهیم فاصله پیش‌بینی denoiser و نویز اصلی را کم کنیم. ضمناً در فرایند آموزش $\mathbb{E}$ حذف شده و به ازای یک batch میانگین تجربی می‌گیریم.

برای sampling هم کافی است یک $x_T \sim \mathcal{N}(0,I)$ را توسط فرمول $x_{t-1} = x_t + \frac{\delta}{2}s_\theta(x_t) + \sqrt{\delta}\cdot z_t$ در طی T مرحله denoise کنیم

$z_t \sim \mathcal{N}(0, I)$

$\boxed{3.2.c}$

score matching $(J_1) \longrightarrow \mathbb{E}\left[\frac{1}{2}\| s_\theta(x) - \nabla_x \lg q(x)\|^2\right]$

$$= \underbrace{\mathbb{E}\left[\frac{1}{2}\| s_\theta(x)\|^2\right]}_{c_1} + \underbrace{\mathbb{E}\left[\frac{1}{2}\| \nabla_x \lg q(x)\|^2\right]}_{c_2} - \underbrace{\mathbb{E}\left[\langle s_\theta(x), \nabla_x \lg q(x)\rangle\right]}_{g(\theta)}$$

$$g(\theta) = \int_x \cancel{q(x)} \left\langle s_\theta(x), \frac{\nabla_x q(x)}{\cancel{q(x)}} \right\rangle dx = \int_x \left\langle s_\theta(x), \underbrace{\nabla_x \int_{x_0} q(x|x_0)\, q(x_0)\, dx_0}_{\int_{x_0} q(x_0) \underbrace{\nabla_x q(x|x_0)}_{q(x|x_0)\, \nabla_x \lg q(x|x_0)} dx_0} \right\rangle dx$$

$$= \int_x \int_{x_0} \overbrace{q(x_0)\, q(x|x_0)}^{q(x, x_0)} \langle s_\theta(x), \nabla_x \lg q(x|x_0) \rangle \, dx_0 \, dx$$

$$= \mathbb{E}_{q(x, x_0)}\left[\langle s_\theta(x), \nabla_x \lg q(x|x_0)\rangle\right]$$

$\longrightarrow$ score matching $(J_1) = c_1 + c_2 + g(\theta) = \overbrace{\mathbb{E}\left[\frac{1}{2}\| s_\theta(x) - \nabla_x \lg q(x|x_0)\|^2\right]}^{J_2}$

$$- \mathbb{E}\left[\frac{1}{2}\| \nabla_x \lg q(x|x_0)\|^2\right] + \mathbb{E}\left[\frac{1}{2}\| \nabla_x \lg q(x)\|^2\right]$$

$\longrightarrow J_2(\theta) = J_1(\theta) + \underbrace{\mathbb{E}\left[\frac{1}{2}\| \nabla_x \lg q(x|x_0)\|^2\right] - \mathbb{E}\left[\frac{1}{2}\| \nabla_x \lg q(x)\|^2\right]}_{C}$

$\boxed{4.1.a}$

می دانیم که لاس برابر است با $L(\theta,\psi) = f(\theta\psi) + c$ و بنابراین $V(\theta,\psi) = \begin{bmatrix} \dot\theta(t) \\ \dot\psi(t) \end{bmatrix} = \begin{bmatrix} -f'(\theta\psi)\psi \\ f'(\theta\psi)\theta \end{bmatrix}$

دقت شود که مشتق نسبت به θ علامت منفی دارد زیرا لاس generator منفی لاس discriminator است.

حال چون $L(\theta,0) = L(0,\psi)$ و برابر c است پس $(\theta,\psi) = (0,0)$ قطعا یکی از نقاط تعادل است. حال می خواهیم نشان دهیم که این نقطه، تنها نقطه ی تعادل است. فرض مسئله درباره f آن است که مشتق آن همواره مثبت است پس تنها حالتی که $V(\theta,\psi)$ بتواند 0 شود آن است که جفت θ ، ψ صفر شوند ← تنها نقطه تعادل $(\theta,\psi) = (0,0)$ است.

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

$$\longrightarrow V'(0,0) = \begin{bmatrix} 0 & -f'(0) \\ f'(0) & 0 \end{bmatrix} \longrightarrow \text{eig vals} = \pm f'(0)\, i$$

مشاهده می شود که هر دو مقدار ویژه ماتریس ژاکوبین موهومی است

$\boxed{4.1.b}$

اگر تعریف کنیم $V_1(\theta,\psi) = \begin{bmatrix} -\nabla_\theta L(\theta,\psi) \\ 0 \end{bmatrix}$ و $V_2(\theta,\psi) = \begin{bmatrix} 0 \\ \nabla_\psi L(\theta,\psi) \end{bmatrix}$ یعنی در حقیقت V را به V_1 و V_2 بشکنیم ، آنگاه داریم :

$$\frac{d}{dt}\, \theta(t)^2 + \psi(t)^2 = 2\theta(t)\, V_1(\theta,\psi) + 2\psi(t)\, V_2(\theta,\psi) = 0 \qquad \forall t \in [0,\infty)$$

بنابراین چون مشتق $\theta^2 + \psi^2$ صفر شود پس خودش const است (به ازای هر $t \in [0,\infty)$)

$\boxed{4.1.c}$

طبق بخش $\boxed{4.1.a}$ هر دو مقدار ویژه موهومی هستند پس به طور کلی درباره همگرایی نمی توان نظر داد ولی اگر همگرا باشد با نرخ زیر خطی همگرا است

$\boxed{4.2.a}$

برای simultenious gradient descent داریم $F_h(\theta,\psi) = (\theta,\psi) + hV(\theta,\psi)$

$$\longrightarrow F_h'(\theta^*,\psi^*) = I + hV'(\theta^*,\psi^*)$$

بنابراین اگر eig val های $V'(\theta^*,\psi^*)$ را μ بنامیم آنگاه $\lambda = 1 + h\mu$

$$\xrightarrow[a>0]{\mu = -a+ib} |\lambda|^2 = \left|1 + h(-a+ib)\right|^2 = (1-ha)^2 + h^2b^2 = (a^2+b^2)h^2 - 2ah + 1$$

$$\xrightarrow{\text{if } |\lambda|>1} (a^2+b^2)h^2 - 2ah + 1 > 1 \longrightarrow (a^2+b^2)h - 2a > 0 \longrightarrow h > \frac{2a}{a^2+b^2}$$

بنابراین به ازای نرخ یادگیری ذکر شده ، مقادیر ویژه بیرون دایره واحد قرار می گیرند

$\boxed{4.2.a}$ ادامه

در حالت گسسته $\left(\theta_{k+1}, \psi_{k+1}\right) = \left(\theta_k, \psi_k\right) + hV\left(\theta_k, \psi_k\right)$

$$\rightarrow \theta_{k+1}^2 + \psi_{k+1}^2 = \left[\theta_k + hV_1(\theta_k, \psi_k)\right]^2 + \left[\psi_k + hV_2(\theta_k, \psi_k)\right]^2$$

$$= \theta_k^2 + h^2 V_1(\theta_k, \psi_k)^2 + \cancel{2\theta_k h \underbrace{V_1(\theta_k, \psi_k)}_{-f'(\theta_k\psi_k)\psi_k}} + \psi_k^2 + h^2 V_2(\theta_k, \psi_k)^2 + \cancel{2\psi_k h \underbrace{V_2(\theta_k, \psi_k)}_{f'(\theta_k\psi_k)\theta_k}}$$

بنابراین مشاهده می‌شود که تمامی عبارات توان 2 دارند و بنابراین تفاضل دو گام بزرگتر مساوی صفر است. این کار مشابه مثبت شدن مشتق در حالت پیوسته است و نشان می‌دهد $\theta^2 + \psi^2$ به صورت یکنوا افزایش می‌یابد

$\boxed{4.2.b}$

برای alternating gradient descent داریم:

$$\text{update operators} \rightarrow \begin{cases} F_1(\theta, \psi) = \begin{bmatrix} \theta - hf'(\theta\psi)\psi \\ \psi \end{bmatrix} \\ F_2(\theta, \psi) = \begin{bmatrix} \theta \\ \psi + hf'(\theta\psi)\theta \end{bmatrix} \end{cases}$$

$$\rightarrow \begin{cases} F_1'(0,0) = \begin{bmatrix} 1 & -hf'(0) \\ 0 & 1 \end{bmatrix} \\ F_2'(0,0) = \begin{bmatrix} 1 & 0 \\ hf'(0) & 1 \end{bmatrix} \end{cases} \xrightarrow[\text{update operator}]{\text{Jacobian of the combined}} F_1'(0,0) \circ F_2'(0,0) = \begin{bmatrix} 1 & -hf'(0) \\ hf'(0) & 1 - h^2 f'(0)^2 \end{bmatrix}$$

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

$\boxed{4.2.c}$

برای قسمت $\boxed{4.2.b}$ مقادیر ویژه را به دست آوردیم و مشاهده می‌شود که به ازای $hf'(0) \leqslant 2$ روی دایره واحد می‌افتیم. در این صورت در مورد همگرایی نمی‌توان نظر داد اما اگر الگوریتم همگرا باشد با نرخ زیرخطی همگرا است

به ازای $hf'(0) > 2$ نیز مقدار بیرون دایره واحد خواهیم داشت که الگوریتم همگرا نخواهد بود

برای قسمت $\boxed{4.2.a}$ نیز در حالتی که مدنظر سوال است هر 2 مقدار ویژه خارج دایره واحد قرار دارند و بنابراین الگوریتم همگرا نیست

$\boxed{4.3.a}$

در صورت استفاده از نویز تابع هدف GAN به صورت زیر می‌شود:

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

$$\longrightarrow \tilde{V}(\theta,\psi) = \mathbb{E}_{\tilde{\theta},x}\begin{bmatrix} -\psi f'(\tilde{\theta}\psi) \\ \tilde{\theta} f'(\tilde{\theta}\psi) - x f'(-x\psi) \end{bmatrix}$$

$$\longrightarrow \tilde{V}'(\theta,\psi) = \mathbb{E}_{\tilde{\theta},x}\begin{bmatrix} -f''(\tilde{\theta}\psi)\psi^2 & -f'(\tilde{\theta}\psi) - f''(\tilde{\theta}\psi)\tilde{\theta}\psi \\ f'(\tilde{\theta}\psi) + f''(\tilde{\theta}\psi)\tilde{\theta}\psi & f''(\tilde{\theta}\psi)\tilde{\theta}^2 + x^2 f(-x\psi) \end{bmatrix}$$

$$\longrightarrow \tilde{V}'(0,0) = \begin{bmatrix} 0 & -f'(0) \\ f'(0) & 2f''(0)\sigma^2 \end{bmatrix} \longrightarrow \text{eig vals} = \lambda_{\frac{1}{2}} = f''(0)\sigma^2 \pm \sqrt{f''(0)^2\sigma^4 - f'(0)^2}$$

$\boxed{4.3.b}$

$$f'(t) = -\frac{-e^{-t}}{1+e^{-t}} \longrightarrow f'(0) = \frac{1}{2}$$

برای حالت پیوسته در بخش $\boxed{4.1.a}$ نشان دادیم که مقادیر ویژه $\pm f'(0)\,i$ هستند

بنابراین به ازای $f(t)$ داده شده مقادیر ویژه $\pm\frac{1}{2}\,i$ می‌شوند که با نرخ زیرخطی همگراست

در برخی سوال ها با علی باباییگ همفکری داشتم