Due Date: December 5th, 2022 at 11:00 pm

Instructions

- *For all questions, show your work!*
- *Use LaTeX and the template we provide when writing your answers. You may reuse most of the notation shorthands, equations and/or tables. See the assignment policy on the course website for more details.*
- *Submit your answers electronically via Gradescope.*
- *TAs for this assignment are* ***Arian Khorasani and Sarthak Mittal****.*

Question 1 (5)**.** (**KL Divergence**)

Given two univariate gaussian distributions, $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$, find the $\mathbb{KL}$-Divergence between the distribution q with the distribution p. In particular, derive the closed form expression for

$$\mathbb{KL}[q(x)||p(x)] = \mathbb{E}_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

Is this the same as $\mathbb{KL}[p(x)||q(x)]$?

Answer 1. By the definition of KL divergence, we get

$$\begin{aligned}
\mathbb{KL}[q(x)||p(x)] &= \mathbb{E}_{q(x)}[\log q(x) - \log p(x)] \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_2,\sigma_2^2)}\left[\log \mathcal{N}(x|\mu_2,\sigma_2^2) - \log \mathcal{N}(x|\mu_1,\sigma_1^2)\right] \\
&= \mathbb{E}_{x\sim\mathcal{N}(\mu_2,\sigma_2^2)}\left[-\log\sigma_2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 + \log\sigma_1 + \frac{1}{2}\log 2\pi + \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \\
&= \log\frac{\sigma_1}{\sigma_2} - \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(\mu_2,\sigma_2^2)}\left[\frac{x^2+\mu_2^2-2x\mu_2}{\sigma_2^2}\right] + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(\mu_2,\sigma_2^2)}\left[\frac{x^2+\mu_1^2-2x\mu_1}{\sigma_1^2}\right] \\
&= \log\frac{\sigma_1}{\sigma_2} - \frac{\mu_2^2+\sigma_2^2+\mu_2^2-2\mu_2^2}{2\sigma_2^2} + \frac{\mu_2^2+\sigma_2^2+\mu_1^2-2\mu_1\mu_2}{2\sigma_1^2} \\
&= \log\frac{\sigma_1}{\sigma_2} - \frac{1}{2} + \frac{\sigma_2^2+(\mu_1-\mu_2)^2}{2\sigma_1^2}
\end{aligned}$$

It is straight-forward to see that it is not symmetric.

Question 2 (2-5-5-5-3)**.** (**Normalizing Flows**) Normalizing flows are expressive invertible transformations of probability distributions. In this exercise, we will see how to satisfy the invertibility constraint of some family of parameterizations. For the first 3 questions, we assume the function $g : \mathbb{R} \to \mathbb{R}$ maps from real space to real space.

1. Let $g(z) = af\,(bz + c)$ where f is the ReLU activation function $f(x) = \max(0, x)$. Show that g is non-invertible.

2. Let $g(z) = \sigma^{-1}(\sum_{i=1}^{N} w_i \sigma(a_i z + b_i))$, $0 < w_i < 1$, where $\sum_i w_i = 1$, $a_i > 0$, and $\sigma(x) = 1/(1+\exp(-x))$ is the logistic sigmoid activation function and σ^{-1} is its inverse. Show that g is *strictly monotonically increasing* on its domain $(-\infty, \infty)$, which implies invertiblity.
3. Consider a residual function of the form $g(z) = z + f(z)$. Show that $df/dz > -1$ implies g is invertible.
4. Consider the following transformation:

$$g(\boldsymbol{z}) = \boldsymbol{z} + \beta h(\alpha, r)(\boldsymbol{z} - \boldsymbol{z}_0) \tag{1}$$

where $\boldsymbol{z}_0 \in \mathbb{R}^D$, $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}$, and $r = ||\boldsymbol{z} - \boldsymbol{z}_0||_2$, $h(\alpha, r) = 1/(\alpha + r)$. Consider the following decomposition of $\boldsymbol{z} = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}}$. (i) Given $\boldsymbol{y} = g(\boldsymbol{z})$, show that $\beta \geq -\alpha$ is a sufficient condition to derive the unique r from equation (1). (ii) Given r and $\boldsymbol{y}$, show that equation (1) has a unique solution $\tilde{\boldsymbol{z}}$.

Answer 2. 1.

$$\frac{dg(z)}{dz} = \begin{cases} ab & \text{if } z \geq \frac{-c}{b} \\ 0 & \text{if } z < \frac{-c}{b} \end{cases}.$$

As g is not strictly monotonic for $z < \frac{-c}{b}$, g is not injective. Thus, its inversion doesn't exist.

2. (a) $\frac{d}{dx} \ln\left(\frac{x}{1-x}\right) = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)}$. We can show (1) $\lim_{x\to 0^+} \frac{1}{x} + \frac{1}{1-x} = \infty$, (2) $\lim_{x\to 1^-} \frac{1}{x} + \frac{1}{1-x} = \infty$, and (3) $\frac{1}{x} + \frac{1}{1-x} > 0$ for $x \in (0,1)$. Thus, $\sigma^{-1}(x)$ is strictly monotonically increasing for its domain $(0,1)$.

   (b) $\sigma(x)$ is strictly monotonically increasing for its domain $(-\infty, \infty)$, since $\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} > 0$ for $\forall x \in (-\infty, \infty)$.

   (c) Let $h(x) \doteq \sum_{i=1}^{N} w_i \sigma(a_i x + b_i)$, $0 < w_i < 1$, $\sum_i w_i = 1$, and $a_i > 0$. $h(x)$ is strictly monotonically increasing for its domain $(-\infty, \infty)$, and its range is $(0,1)$.

   Note that $a_i > 0$, $w_i > 0$, and $\sigma'(x) > 0$ for $x \in (-\infty, \infty)$. Thus, $\frac{dh(x)}{dx} = \sum_{i=1}^{N} w_i a_i \sigma'(a_i x + b_i) > 0$. Moreover, the range of $h(x)$ is $(0,1)$ because

$$\lim_{x\to\infty} h(x) = \sum_{i=1}^{N} w_i \lim_{x\to\infty} \sigma(a_i x + b_i) = \left(\sum_{i=1}^{N} w_i\right) \cdot 1 = 1.$$

$$\lim_{x\to-\infty} h(x) = \sum_{i=1}^{N} w_i \lim_{x\to-\infty} \sigma(a_i x + b_i) = \left(\sum_{i=1}^{N} w_i\right) \cdot 0 = 0.$$

   (d) Due to (a), (b), and (c), $g(z)$ is strictly monotonically increasing on its domain $(-\infty, \infty)$, and its range is $(-\infty, \infty)$. When a function f is a *strictly monotonic function*, then f is injective on its domain. Moveover, if T is the range of f, then there is an inverse function f^{-1} on T.

3. Note that $dg(z)/dz = 1 + df(z)/dz$. If $df/dz > -1$, then $dg/dz > 0$ on its domain. As g is strictly monotonically increasing on its domain, thus g is injective on its domain and there exists its inverse function on the range.

4. (i) We subtract $\boldsymbol{z}_0$ from both sides of equation (1) then we take the norm of both sides to obtain r. The given condition allows for $r(1 + \beta/(\alpha + r))$ to be invertible with respect to r.

   (ii) The substitution of the expression of $\boldsymbol{z}$ gives us the unique solution.

Question 3 (3-5-2-2-3)**.** (**Mixing VAEs and Diffusion Models**)

Variational Autoencoder. VAEs are a class of latent-variable generative models that work on optimizing the ELBO, which is defined as

$$ELBO(\theta, \phi) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z|x_i)}[logp_\theta(x_i|z)] + \mathbb{KL}[q_\phi(z|x_i)||p(z)]$$

where we are given a dataset $\{x_i\}_{i=1}^N$ and $p_\theta(x|z)$ is the conditional likelihood, $p(z)$ is the prior and $q_\phi(z|x)$ is the approximate variational distribution. Optimization is done by maximizing the ELBO, or minimizing the negative of it.

Denoising Diffusion Probabilistic Model. DDPMs are a class of generative models that rely on a known forward diffusion process $q(x_t|x_{t-1})$, which progressively destroys structure from the data until it converges to unstructured noise, eg. $\mathcal{N}(0, I)$ and a learned parameterized (by a Neural Network !) backward process $p_\theta(x_{t-1}|x_t)$ that iteratively removes noise until you have obtained a sample from the data distribution.

Let $\boldsymbol{x}_0$ be a sample from the data distribution ; and let the forward diffusion process (noising process) be defined using

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \qquad \text{where} \qquad q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t)$$

and the reverse diffusion process (denoising process) being a learned process following

$$p_\phi(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \qquad \text{where} \qquad p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_\phi(\boldsymbol{x}_t, t), _\phi(\boldsymbol{x}_t, t))$$

(a) Show that $\log p_\phi(\boldsymbol{x}_0) \geq \underbrace{\mathbb{E}_q\left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]}_{\mathcal{L}_{DDPM}}$ and further show that that $\mathbb{E}_q\left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] = \mathbb{E}\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$

(b) Show that $\mathcal{L}_{DDPM} = \mathbb{E}_q\left[\log p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1) - \sum_{t=2}^{T} \mathbb{KL}\left[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \;||\; p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\right] - \mathbb{KL}[q(\boldsymbol{x}_T|\boldsymbol{x}_0) \;||\; p(\boldsymbol{x}_T)]\right]$

(c) Now consider a data sample of the form $(\boldsymbol{x}_0, \boldsymbol{c})$, where $\boldsymbol{c}$ is now some additional auxiliary data that you have been provided (in particular ; $\boldsymbol{c}$ can be the same as $\boldsymbol{x}_0$ as well). Suppose we are modeling the data as $p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})$ where our latent variables now are $\boldsymbol{z}$ and $\boldsymbol{x}_{1:T}$. Since the posterior will be intractable, lets try to approximate it with $q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)$. Can you now re-derive the ELBO, which is the lower-bound on the log likelihood, as $\log p(\boldsymbol{x}_0, \boldsymbol{c}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{c})}\right]$?

(d) Suppose now you are modeling this problem with a combination of VAE and Denoising Diffusion Probabilistic Model, where the encoder of the VAE has the parameters ψ, the decoder θ and the denoising model ϕ. In this case, the generative distribution factorizes as

$$\begin{aligned} p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z}) &= p(\boldsymbol{z}) p_\theta(\boldsymbol{c}|\boldsymbol{z}) p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c}, \boldsymbol{z}) \\ &= p(\boldsymbol{z}) p_\theta(\boldsymbol{c}|\boldsymbol{z}) p(\boldsymbol{x}_T|\boldsymbol{c}, \boldsymbol{z}) \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t \boldsymbol{c}, \boldsymbol{z}) \end{aligned}$$

Further, suppose we now want to model $\boldsymbol{z}$ using a VAE's encoder with parameters ψ, and then the remaining latent variables $\boldsymbol{x}_{1:T}$ conditioned on $\boldsymbol{z}$ through the forward diffusion process. Can you provide a factorization of $q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)$ that respects this ?

(e) Through arithmetic manipulation of the ELBO you derived above as well as the factorization that you have provided, can you now decompose the objective into a VAE component and a DDPM component ?

Answer 3. (a)

$$\begin{aligned} \log p_\phi(\boldsymbol{x}_0) &= \int_{\boldsymbol{x}_0} \left(\log \int_{\boldsymbol{x}_{1:T}} p_\phi(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T} \right) d\boldsymbol{x}_0 \\ &= \int_{\boldsymbol{x}_0} \left(\log \int_{\boldsymbol{x}_{1:T}} p_\phi(\boldsymbol{x}_{0:T}) \cdot \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} d\boldsymbol{x}_{1:T} \right) d\boldsymbol{x}_0 \\ &= \int_{\boldsymbol{x}_0} \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] d\boldsymbol{x}_0 \\ &\geq \int_{\boldsymbol{x}_0} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] d\boldsymbol{x}_0 \qquad \text{(Jensen's Inequality)} \\ &= \mathbb{E}_q \left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] \end{aligned}$$

Now using the factorization provided above, we get

$$\begin{aligned} \mathbb{E}_q \left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right] &= \mathbb{E}_q \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \\ &= \mathbb{E}_q \left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \right] \end{aligned}$$

(b) Note here that we need to bring $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ in the picture. We can see that

$$\begin{aligned} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) \cdot q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\ &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \cdot q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\ \implies q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &= \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t.\boldsymbol{x}_0) \cdot q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)} \end{aligned}$$

Plugging this into the previous derived $\mathcal{L}_{DDPM}$, we get

$$\begin{aligned}\mathcal{L}_{DDPM} &= \mathbb{E}_q\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T}\log\frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\cdot q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\cdot q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\right]\\ &= \mathbb{E}_q\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T}\log\frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} + \sum_{t=1}^{T}\log\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\right]\\ &= \mathbb{E}_q\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T}\log\frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} - \log q(\boldsymbol{x}_T|\boldsymbol{x}_0)\right]\end{aligned}$$

This follows because

$$\begin{aligned}\sum_{t=1}^{T}\log\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} &= \sum_{t=1}^{T}\log q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) - \sum_{t=1}^{T}\log q(\boldsymbol{x}_t|\boldsymbol{x}_0)\\ &= -\log q(\boldsymbol{x}_T|\boldsymbol{x}_0)\end{aligned}$$

Thus, we get

$$\begin{aligned}\mathcal{L}_{DDPM} &= \mathbb{E}_q\left[\log p(\boldsymbol{x}_T) + \log p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1) + \sum_{t=2}^{T}\log\frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} - \log q(\boldsymbol{x}_T|\boldsymbol{x}_0)\right]\\ &= \mathbb{E}_q\left[\log p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1) - \log\frac{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}{p(\boldsymbol{x}_T)} - \sum_{t=2}^{T}\log\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}\right]\\ &= \mathbb{E}_q\left[\log p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1) - \mathbb{KL}[q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)] - \sum_{t=2}^{T}\mathbb{KL}[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)]\right]\end{aligned}$$

(c) If we follow the exact same steps as (a) but instead of the data $\boldsymbol{x}_0$, we plug in $(\boldsymbol{x}_0, \boldsymbol{c})$ for the data and instead of the unobserved variables $\boldsymbol{x}_{1:T}$, we now consider $(\boldsymbol{x}_{1:T}, \boldsymbol{z})$. The derivation follows trivially from this.

(d) Respecting the mentioned conditions, we can have the variational distribution factorization as

$$q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0) = q_\psi(\boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)\cdot q(\boldsymbol{x}_{1:T}|\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{x}_0)$$

(e)

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{x}_{1:T},\boldsymbol{z})}\left[\log\frac{p(\boldsymbol{x}_{0:T},\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T},\boldsymbol{z}|\boldsymbol{x}_0,\boldsymbol{c})}\right] &= \mathbb{E}_q\left[\log\frac{p(\boldsymbol{z})\cdot p_\theta(\boldsymbol{c}|\boldsymbol{z})\cdot p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)\cdot q(\boldsymbol{x}_{1:T}|\boldsymbol{z},\boldsymbol{c},\boldsymbol{x}_0)}\right]\\ &= \mathbb{E}_q\left[\log p_\theta(\boldsymbol{c}|\boldsymbol{z}) - \log\frac{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)}{p(\boldsymbol{z})} - \log\frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{z},\boldsymbol{c},\boldsymbol{x}_0)}{p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}\right]\end{aligned}$$

Writing this explicitly, we get

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T},\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}_{0:T},\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T},\boldsymbol{z}|\boldsymbol{x}_0,\boldsymbol{c})}\right] = \underbrace{\mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{x}_0,\boldsymbol{c})}\left[\log p_\theta(\boldsymbol{c}|\boldsymbol{z})\right] - \mathbb{KL}[q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)||p(\boldsymbol{z})]}_{\text{VAE Loss}} + \mathbb{E}_{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{x}_0)}\left[\log \frac{p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{z},\boldsymbol{c},\boldsymbol{x}_0)}\right]}_{\mathcal{L}_{DDPM}}$$

Question 4 (6-2-6-6). (**Generative Adversarial Network**)

In this question, we are concerned with analyzing the training dynamics of GANs under gradient ascent-descent. We denote the parameters of the critic and the generator by ψ and θ respectively. The objective function under consideration is the Jensen-Shannon (standard) GAN one:

$$\mathcal{L}(\psi,\theta) = \mathbb{E}_{p_D}\log(\sigma(C_\psi(x))) + \mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x)))$$

where σ is the logistic function. For ease of exposition, we will study the continuous-time system which results from the (alternating) discrete-time system when learning rate, $\eta > 0$, approaches zero:

$$\psi^{(k+1)} = \psi^{(k)} + \eta v_\psi(\psi^{(k)},\theta^{(k)}) \theta^{(k+1)} = \theta^{(k)} + \eta v_\theta(\psi^{(k+1)},\theta^{(k)}) \quad \eta \to 0^+ \quad \begin{matrix}\dot{\psi} = v_\psi(\psi,\theta)\\ \dot{\theta} = v_\theta(\psi,\theta)\end{matrix} \qquad v_\psi(\psi,\theta)\nabla_\psi\mathcal{L}(\psi,\theta) v_\theta(\psi$$

The purpose is to initiate a study on the stability of the training algorithm. For this reason, we will utilize the following simple setting: Both training and generated data have support on $\mathbb{R}$. In addition, $p_D = \delta_0$ and $p_\theta = \delta_\theta$. This means that both of them are Dirac distributions [1] which are centered at $x = 0$, for the real data, and at $x = \theta$, for the generated. The critic, $C_\psi : \mathbb{R} \to \mathbb{R}$, is $C_\psi(x) = \psi_0 x + \psi_1$.

4.1 Derive the expressions for the "velocity" field, v, of the dynamical system in the joint parameter space (ψ_0, ψ_1, θ), and find its stationary points $(\psi_0^*, \psi_1^*, \theta^*)$. [2]

4.2 Derive J^*, the (3×3) Jacobian of v at $(\psi_0^*, \psi_1^*, \theta^*)$.

For a continuous-time system to be locally asymptotically stable it suffices that all eigenvalues of J^* have negative real part. Otherwise, further study is needed to conclude. However, this case is not great news since the fastest achievable convergence is sublinear.

4.3 Find the eigenvalues of J^* and comment on the system's local stability around the stationary points.

Now we will include a gradient penalty, $\mathcal{R}_1(\psi) = \mathbb{E}_{p_D}\|\nabla_x C_\psi(x)\|^2$, to the critic's loss, so the regularized system becomes:

$$\begin{matrix}\dot{\psi} = \bar{v}_\psi(\psi,\theta)\\ \dot{\theta} = \bar{v}_\theta(\psi,\theta)\end{matrix} \qquad [l]\bar{v}_\psi(\psi,\theta)\nabla_\psi\mathcal{L}(\psi,\theta) - \frac{\gamma}{2}\nabla_\psi\mathcal{R}_1(\psi)\bar{v}_\theta(\psi,\theta) - \nabla_\theta\mathcal{L}(\psi,\theta)$$

for $\gamma > 0$. Repeat 1-2-3 for the modified system and compare the stability of the two.

1. If $p_X = \delta_z$, then $p(X = z) = 1$.
2. To find the stationary points, set $v = 0$ and solve for each of the parameters.

4.4 Derive the expressions for the "velocity" field, $\bar{v}$, of the dynamical system in the joint parameter space (ψ_0, ψ_1, θ), and find its stationary points $(\psi_0^*, \psi_1^*, \theta^*)$. [3]

4.5 Derive $\bar{J}^*$, the (3×3) Jacobian of $\bar{v}$ at $(\psi_0^*, \psi_1^*, \theta^*)$.

4.6 Find the eigenvalues of $\bar{J}^*$ and comment on the system's local stability around the stationary points.

Answer 4. Simplifying:

$$\mathcal{L}(\psi, \theta) = \log(\sigma(0\psi_0 + \psi_1)) + \log(\sigma(-\psi_0\theta - \psi_1))$$
$$\mathcal{R}_1(\psi) = \psi_0^2$$

4.1 Evaluate the partial derivatives of the loss:

$$\nabla_{\psi_0}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)(-\theta)$$
$$\nabla_{\psi_1}\mathcal{L} = \sigma(-\psi_1) - \sigma(\psi_0\theta + \psi_1)$$
$$\nabla_{\theta}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)(-\psi_0)$$

So the velocity field is:

$$v_{\psi_0}(\psi_0, \psi_1, \theta) = \sigma(\psi_0\theta + \psi_1)(-\theta)$$
$$v_{\psi_1}(\psi_0, \psi_1, \theta) = \sigma(-\psi_1) - \sigma(\psi_0\theta + \psi_1)$$
$$v_{\theta}(\psi_0, \psi_1, \theta) = -\sigma(\psi_0\theta + \psi_1)(-\psi_0)$$

We solve the equation $v(\psi_0^*, \psi_1^*, \theta^*) = 0$ to find that there is a single stationary point $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$.

4.2 To find the Jacobian of the velocity field, let's first calculate second partial derivatives for the loss:

$$\nabla_{\psi_0\psi_0}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)(-\theta^2)$$
$$\nabla_{\psi_0\psi_1}\mathcal{L} = \nabla_{\psi_1\psi_0}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)(-\theta)$$
$$\nabla_{\psi_1\psi_1}\mathcal{L} = \sigma(-\psi_1)\sigma(\psi_1)(-1) - \sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)$$
$$\nabla_{\psi_0\theta}\mathcal{L} = \nabla_{\theta\psi_0}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)(-\psi_0\theta) - \sigma(\psi_0\theta + \psi_1)$$
$$\nabla_{\psi_1\theta}\mathcal{L} = \nabla_{\theta\psi_1}\mathcal{L} = -\sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)\psi_0$$
$$\nabla_{\theta\theta}\mathcal{L} = \sigma(\psi_0\theta + \psi_1)\sigma(-\psi_0\theta - \psi_1)(-\psi_0^2)$$

Then formulate the Jacobian:

$$J(\psi_0, \psi_1, \theta) = \begin{bmatrix} \nabla_{\psi_0\psi_0}\mathcal{L} & \nabla_{\psi_1\psi_0}\mathcal{L} & \nabla_{\theta\psi_0}\mathcal{L} \\ \nabla_{\psi_0\psi_1}\mathcal{L} & \nabla_{\psi_1\psi_1}\mathcal{L} & \nabla_{\theta\psi_1}\mathcal{L} \\ -\nabla_{\psi_0\theta}\mathcal{L} & -\nabla_{\psi_1\theta}\mathcal{L} & -\nabla_{\theta\theta}\mathcal{L} \end{bmatrix}$$

And evaluate it at $(\psi_0^*, \psi_1^*, \theta^*)$. So for the JSD GAN: $J^* = \begin{bmatrix} 0 & 0 & -1/2 \\ 0 & -1/2 & 0 \\ +1/2 & 0 & 0 \end{bmatrix}$.

3. To find the stationary points, set $v = 0$ and solve for each of the parameters.

4.3 Solving the equation $\det(\lambda I - J^*) = 0$, we find that the eigenvalues are are $-1/2$, and $\pm 1/2i$. Since not all eigenvalues have negative real parts (0 is not negative), we cannot determine stability around the stationary point and in the best case we have sublinear (slow) local convergence - for the pure JSD system.

4.4 Evaluate the partial derivatives for $\mathcal{R}_1$ and reuse the derivatives for the loss:

$$\nabla_{\psi_0}\mathcal{R}_1 = 2\psi_0$$
$$\nabla_{\psi_1}\mathcal{R}_1 = 0$$

So the *modified* velocity field is:

$$\bar{v}_{\psi_0}(\psi_0, \psi_1, \theta) = \sigma(\psi_0\theta + \psi_1)(-\theta) - \gamma\psi_0$$
$$\bar{v}_{\psi_1}(\psi_0, \psi_1, \theta) = \sigma(-\psi_1) - \sigma(\psi_0\theta + \psi_1)$$
$$\bar{v}_{\theta}(\psi_0, \psi_1, \theta) = -\sigma(\psi_0\theta + \psi_1)(-\psi_0)$$

From this we derive that there is a single stationary point for the gradient penalized system: $(\psi_0^*, \psi_1^*, \theta^*) = (0, 0, 0)$.

4.5

$$\nabla_{\psi_0\psi_0}\mathcal{R}_1 = 2$$
$$\nabla_{\psi_0\psi_1}\mathcal{R}_1 = \nabla_{\psi_0\theta}\mathcal{R}_1 = 0$$

So for the gradient penalized system we have $\bar{J}^* = \begin{bmatrix} -\gamma & 0 & -1/2 \\ 0 & -1/2 & 0 \\ +1/2 & 0 & 0 \end{bmatrix}$.

4.6 Solving for the eigenvalues of $\bar{J}^*$, we get $-1/2$, and solutions of $x^2 + \gamma x + \frac{1}{4}$ which have negative real part in any case (one can identify the cases and solve for each of them, or quicker apply the Vieta's formulas: $x_1 + x_2 = -\gamma \implies 2\text{Re}(x) = -\gamma \implies \text{Re}(x) = -\gamma/2 < 0$ if roots are complex conjugates, else they are reals with same sign which is negative because $x_1 + x_2 < 0$ and $x_1 x_2 = \frac{1}{4} > 0$). As a result, the trajectories of the continuous-time modified JSD system are asymptotically stable locally around the equilibrium.

Conclusion: The $\mathcal{R}_1$ gradient penalty effectively stabilizes (locally) the gradient ascent-descent (JSD) GAN system.

Question 5 (5-5-5-5). (**Paper Review: A Simple Framework for Contrastive Learning of Visual Representations**)

In this question, you are going to write a **one page review** of the A Simple Framework for Contrastive Learning of Visual Representations paper. Please structure your review into the following sections:

(5.1) **Summary:**

(a) What is this paper about ?

(b) What is the main contribution ?

(c) Describe the main approach and results. Just facts, no opinions yet.

(5.2) **Strengths:**

(a) Is there a new theoretical insight ?

(b) Or a significant empirical advance ? Did they solve a standing open problem ?

(c) Or a good formulation for a new problem ?

(d) Any good practical outcome (code, algorithm, etc) ?

(e) Are the experiments well executed ?

(f) Useful for the community in general ?

(5.3) **Weaknesses:**

(a) What can be done better ?

(b) Any missing baselines ? Missing datasets ?

(c) Any odd design choices in the algorithm not explained well ? Quality of writing ?

(d) Is there sufficient novelty in what they propose ? Minor variation of previous work ?

(e) Why should anyone care ? Is the problem interesting and significant ?

(5.4) **Reflections:**

(a) How does this relate to other concepts you have seen in the class ?

(b) What are the next research directions in this line of work ?

(c) What (directly or indirectly related) new ideas did this paper give you ? What would you be curious to try ?

Answer 5. This question is subjective and so we will accept a variety of answers. You are expected to analyze the paper and offer your own perspective and ideas, beyond what the paper itself discusses.