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. $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$

$\mathbb{KL}[q(x)||p(x)] = \mathbb{E}_{q(x)}[\log \frac{q(x)}{p(x)}] = \int_{-\infty}^{\infty} q(x)[\log \frac{q(x)}{p(x)}]\, dx = \int q(x) \log q(x)\, dx - \int q(x) \log p(x)\, dx$

The first integral: $\int q(x) \log q(x)\, dx = -\frac{1}{2}(1 + \log 2\pi\sigma_2^2)$

The second integral: $-\int q(x) \log p(x)\, dx = -\int q(x) \log \frac{1}{(2\pi\sigma_1^2)^{(0.5)}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\, dx = \frac{1}{2}\log(2\pi\sigma_1^2) - \int q(x) \log e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}\, dx = \frac{1}{2}\log(2\pi\sigma_1^2) - \int q(x)(-\frac{(x-\mu_1)^2}{2\sigma_1^2})\, dx = \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{\int q(x)x^2\, dx - \int q(x)2x\mu_1\, dx + \int q(x)\mu_1^2\, dx}{2\sigma_1^2}$

letting <> denote expectation operator under q **and** $var(x) = \sigma_2^2 =< x^2 > - < x >^2 \rightarrow <x^2> = \sigma_2^2 + \mu_2^2$

The second integral simplifies to: $-\int q(x) \log p(x)\, dx = \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{<x^2> - 2<x>\mu_1 + \mu_1^2}{2\sigma_1^2} = \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{\sigma_2^2 + \mu_2^2 - 2\mu_2\mu_1 + \mu_1^2}{2\sigma_1^2} = \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{\sigma_2^2 + (\mu_2 - \mu_1)^2}{2\sigma_1^2}$

Finally: $\mathbb{KL}[q(x)||p(x)] = \int q(x) \log q(x)\, dx - \int q(x) \log p(x)\, dx = -\frac{1}{2}(1 + \log 2\pi\sigma_2^2) + \frac{1}{2}\log(2\pi\sigma_1^2) + \frac{\sigma_2^2 + (\mu_1 - \mu_2)^2}{2\sigma_1^2} = \log \frac{\sigma_1}{\sigma_2} + \frac{\sigma_2^2 + (\mu_2 - \mu_1)^2}{2\sigma_1^2} - \frac{1}{2}$

No, it is not the same thing. $\mathbb{KL}$-Divergence is not symmetric.

$\mathbb{KL}[p(x)||q(x)] = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$

The two above statements for $\mathbb{KL}[q(x)||p(x)]$ and $\mathbb{KL}[p(x)||q(x)]$ are not equal..

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 $g(z) = \sigma^{-1}(\sum_{i=1}^{N} w_i\sigma(a_iz + b_i))$, $0 < w_i < 1$, 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. 2.1:

$g(z) = af(bz + c)$ and $f(x) = max(0, x)$

$$g(z) = \begin{cases} a(bz + c) & z >= -\frac{c}{b} \\ 0 & z <= -\frac{c}{b} \end{cases}$$

A function is invertible only if each input has a unique output. Here, for all inputs that are $z <= -\frac{c}{b}$, the function g has the same output, which means that g is non-invertible.

2.2:

We know: $\sigma^{-1}(x) = \log(\frac{x}{1-x}) \to \frac{d\sigma^{-1}(x)}{dx} = \frac{1}{\ln 10}\frac{1-x}{x}\frac{1}{(1-x)^2} = \frac{1}{\ln 10}\frac{1}{x(1-x)}$

and $\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$

consider: p $= \sum_{i=1}^{N} w_i\sigma(a_iz + b_i) \to g(z) = \sigma^{-1}(p)$

g is the composition of two functions σ^{-1} and P. If each function is separately invertible in their domain, then g is also invertible.

$\frac{d(g(z))}{dz} = \frac{d\sigma^{-1}(x)}{dx}\frac{dp}{dz} = \frac{1}{\ln 10}\frac{1}{p(1-p)}(\sum_{i=1}^{N} a_i w_i \sigma(a_i z + b_i)(1 - \sigma(a_i z + b_i)))$

since we know $0 < \sigma < 1$, so $1 - \sigma(a_i z + b_i) > 0 \rightarrow (\sum_{i=1}^{N} a_i w_i \sigma(a_i z + b_i)(1 - \sigma(a_i z + b_i))) > 0$ is positive and thus invertible.

we just have to show that $\forall p \in [0, 1]$ (which is the domain of σ^{-1}), $\frac{d\sigma^{-1}(x)}{dx} = \frac{1}{\ln 10}\frac{1}{p(1-p)} > 0$, which it is.

All the terms in $\frac{d(g(z))}{dz}$ is positive $\rightarrow \frac{d(g(z))}{dz} > 0$ is positive. We know that if a function has a positive derivative over an interval, then it is strictly monotonically increasing over that interval, and hence injective and invertible there.

2.3:

To prove a function (g) is invertible, we have to prove that it is monotonously increasing and injective. So, we need to show that it's derivative is positive.

If $\frac{df}{dz} > -1$:

$\frac{dg}{dz} = 1 + \frac{df}{dz} > 1 + (-1) = 0 \rightarrow \frac{dg}{dz} > 0$

This means that the function g is invertible.

2.4:

We have $r = ||\boldsymbol{z} - \boldsymbol{z}_0||_2, \qquad h(\alpha, r) = 1/(\alpha + r), \qquad \boldsymbol{z} = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}}$

$g(z) = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}} + \frac{\beta}{\alpha+r}(\boldsymbol{z} - \boldsymbol{z}_0) = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}} + \beta\frac{r\tilde{\boldsymbol{z}}}{\alpha+r} \qquad \text{and} \qquad g(z) = y$

part 2: It is shown that given r and y, the equation has a unique solution to $\tilde{\boldsymbol{z}}$.

$\rightarrow \tilde{\boldsymbol{z}} = \frac{y - \boldsymbol{z}_0}{r(1+\frac{\beta}{\alpha+r})}$

part 1: $\rightarrow y - \boldsymbol{z}_0 = \tilde{\boldsymbol{z}} r(1 + \frac{\beta}{\alpha+r})$

In the next step, we can take the norm of both sides:

$|y - \boldsymbol{z}_0| = r(1 + \frac{\beta}{\alpha+r})$

To show that $\beta \geq -\alpha$ is a sufficient condition to derive the unique r from the equation, we have to show that the equation above is invertible. a sufficient condition for the equation above to be invertible is for $r(1+\frac{\beta}{\alpha+r})$ to be a non-decreasing function, which means its derivative with respective to r should be positive.

$\frac{\partial(r+r\frac{\beta}{\alpha+r})}{r} = 1 + \frac{\alpha\beta}{(\alpha+r)^2} \geq 0 \rightarrow \beta \geq -\frac{(r+\alpha)^2}{\alpha}$

since $r \geq 0$, it is sufficient to impose $\beta \geq -\frac{\alpha^2}{\alpha} = -\alpha \rightarrow \beta \geq -\alpha$

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

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

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

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

(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

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

(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: part 1.

$\log p_\phi(\boldsymbol{x}_0) \geq \log p_\phi(\boldsymbol{x}_0) - \mathbb{KL}(q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)||p_\phi(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0))$

$= \log p_\phi(\boldsymbol{x}_0) - \mathbb{E}_{\boldsymbol{x}_{1:T} \sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}[\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p_\phi(\boldsymbol{x}_{0:T})/p_\phi(\boldsymbol{x}_0)}]$

$= \log p_\phi(\boldsymbol{x}_0) - \mathbb{E}_q[\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p_\phi(\boldsymbol{x}_{0:T})} + \log p_\phi(\boldsymbol{x}_0)]$

$= \mathrm{E}_q \left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \right]$

part 2.

Considering:

$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and $p_\phi(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$

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

$\overset{\text{product becomes sum outside of log}}{=} \mathbb{E}\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$

b: $\mathcal{L}_{DDPM} = \mathrm{E}_q\left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]$

$= \mathrm{E}\left[\log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right]$

$= \mathrm{E}\left[\log p(\boldsymbol{x}_T) + \sum_{t=2}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} + \log \frac{p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\right]$

$= \mathrm{E}\left[\log p(\boldsymbol{x}_T) + \sum_{t=2}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} \cdot \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} + \log \frac{p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1)}{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\right]$

$= \mathrm{E}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)} + \sum_{t=2}^{T} \log \frac{p_\phi(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} + \log p_\phi(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]$

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

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

c: $\log p(\boldsymbol{x}_0, \boldsymbol{c}) = \log \int p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})\, dx_{1:T} dz$

$= \log \int p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z}) \frac{q(\boldsymbol{x}_{1:T}, z|\boldsymbol{x}_0, c)}{q(\boldsymbol{x}_{1:T}, z|\boldsymbol{x}_0, c)}\, dx_{1:T} dz$

$= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}\,,\,\boldsymbol{z})} \frac{p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{x}_{1:T}, z|\boldsymbol{x}_0, c)}$

$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}\,,\,\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}_{0:T}\,,\,\boldsymbol{c}\,,\,\boldsymbol{z})}{q(\boldsymbol{x}_{1:T}\,,\,\boldsymbol{z}|\boldsymbol{x}_0\,,\,\boldsymbol{c})}\right]$

This final inequality follows from Jensen's Inequality.

d: This posterior $q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)$ can be factorized into the following conditional distributions $q_\psi(\boldsymbol{z}|c, \boldsymbol{x}_0)$ and $q(\boldsymbol{x}_{1:T}|c, z, \boldsymbol{x}_0)$.

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

e: considering:

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

$\log p(\boldsymbol{x}_0, \boldsymbol{c}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z})} \left[\log \frac{p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{c})}\right]$

$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{z})p_\theta(\boldsymbol{c}|\boldsymbol{z})p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)q(\boldsymbol{x}_{1:T}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{x}_0)}\right]$

$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)} + \log p_\theta(\boldsymbol{c}|\boldsymbol{z}) + \log \frac{p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{x}_0)}\right]$

$\geq \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)}[\log \frac{p(\boldsymbol{z})}{q_\psi(\boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)} + \log p_\theta(\boldsymbol{c}|\boldsymbol{z})] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c},\boldsymbol{x}_0)}[\log \frac{p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{x}_0)}]$

$\geq \mathbb{E}_{q_\psi(\boldsymbol{z}|c,\boldsymbol{x}_0)}[p_\theta(c|\boldsymbol{z})] - \mathbb{KL}[q_\psi(\boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0) \,||\, p(\boldsymbol{z})] + \mathbb{E}_{z\sim q(\boldsymbol{z}|c,\boldsymbol{x}_0)}[\mathbb{E}_{q(\boldsymbol{x}_{1:T}|c,\boldsymbol{z},\boldsymbol{x}_0)}[\frac{p_\phi(\boldsymbol{x}_{0:T}|\boldsymbol{c},\boldsymbol{z})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{x}_0)}]]$

The first two terms correspond to L_{VAE} and the last term correspond to 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:

$$\begin{array}{l} \psi^{(k+1)} = \psi^{(k)} + \eta v_\psi(\psi^{(k)}, \theta^{(k)}) \\ \theta^{(k+1)} = \theta^{(k)} + \eta v_\theta(\psi^{(k+1)}, \theta^{(k)}) \end{array} \quad \xrightarrow{\eta \to 0^+} \quad \begin{array}{ll} \dot{\psi} = v_\psi(\psi, \theta) & v_\psi(\psi, \theta) := \nabla_\psi \mathcal{L}(\psi, \theta) \\ \dot{\theta} = v_\theta(\psi, \theta) & v_\theta(\psi, \theta) := -\nabla_\theta \mathcal{L}(\psi, \theta) \end{array}$$

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.

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.

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:

$$\begin{aligned} \dot{\psi} &= \bar{v}_\psi(\psi,\theta) \qquad \bar{v}_\psi(\psi,\theta) \coloneqq \nabla_\psi \mathcal{L}(\psi,\theta) - \tfrac{\gamma}{2}\nabla_\psi \mathcal{R}_1(\psi) \\ \dot{\theta} &= \bar{v}_\theta(\psi,\theta) \qquad \bar{v}_\theta(\psi,\theta) \coloneqq -\nabla_\theta \mathcal{L}(\psi,\theta) \end{aligned}$$

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

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. 4.1:

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

consider $f = -\log(\sigma)) = -\log(1 + exp(-t))$

Also, we know that $\int_{-\infty}^{\infty}\delta(x)\,dx = 1$ and $\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(x)\int_{-\infty}^{\infty}\delta(x)\,dx = f(0)$

So, $\mathbb{E}_{p_D}\log(\sigma(C_\psi(x))) = \int_{-\infty}^{\infty}\delta_0 f(\psi_0\theta + \psi_1) = f(\psi_0\theta+\psi_1)\int_{-\infty}^{\infty}\delta(x)\,dx = f(\psi_1)$

and $\mathbb{E}_{p_\theta}\log(\sigma(-C_\psi(x))) = f(-\psi_0\theta - \psi_1)\int_{-\infty}^{\infty}\delta(x-\theta)\,dx = f(-\psi_0\theta-\psi_1)$

Therefore

or $\mathcal{L}(\psi,\theta) = f(-\psi_0\theta - \psi_1) + f(\psi_1)$

$$v(\theta,\psi) = \begin{pmatrix} v_{\psi_0} \\ v_{\psi_1} \\ v_\theta \end{pmatrix} = \begin{pmatrix} \nabla_{\psi 0}\mathcal{L}(\theta,\psi) \\ \nabla_{\psi 1}\mathcal{L}(\theta,\psi) \\ -\nabla_\theta \mathcal{L}(\theta,\psi) \end{pmatrix} = \begin{pmatrix} -f'(-\psi_0\theta-\psi_1)\theta \\ -f'(-\psi_0\theta-\psi_1) + f'(\psi_1) \\ f'(-\psi_0\theta-\psi_1)\psi_0 \end{pmatrix}$$

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

$$v(\theta,\psi) = 0 \rightarrow \begin{pmatrix} -f'(-\psi_0\theta-\psi_1)\theta \\ -f'(-\psi_0\theta-\psi_1) + f'(\psi_1) \\ f'(-\psi_0\theta-\psi_1)\psi_0 \end{pmatrix} = 0 \rightarrow \psi_0 = 0, \psi_1 = 0, \theta = 0$$

4.2:

$$J^* = \begin{bmatrix} \partial v_{\psi_0}/\partial\psi_0 & \partial v_{\psi_0}/\partial\psi_1 & \partial v_{\psi_0}/\partial\theta \\ \partial v_{\psi_1}/\partial\psi_0 & \partial v_{\psi_1}/\partial\psi_1 & \partial v_{\psi_1}/\partial\theta \\ \partial v_\theta/\partial\psi_0 & \partial v_\theta/\partial\psi_1 & \partial v_\theta/\partial\theta \end{bmatrix} = \begin{bmatrix} \theta^2 f'' & \theta f' & -f' + \psi_0\theta f'' \\ \theta f'' & 2f'' & \psi_0 f'' \\ -\theta\psi_0 f'' + f' & -\psi_0 f'' & -\psi_0^2 f'' \end{bmatrix}$$

where $f'' = f''(-\psi_0\theta - \psi)$ and $f' = f'(-\psi_o\theta - \psi)$

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

and at stationary point $(\psi_0 = 0, \psi_1 = 0, \theta = 0)$:

$f' = \frac{exp(-t)}{1+exp(-t)}$ and $f'' = \frac{-exp(-t)}{(1+exp(-t))^2} \rightarrow f'(0) = 1/2$ and $f''(0) = -1/4$

$$J^* = \begin{bmatrix} 0 & 0 & -0.5 \\ 0 & -0.5 & 0 \\ 0.5 & 0 & 0 \end{bmatrix}$$

4.3:

if all eigenvalues of the Jacobian at a stationary point have negative real-part, the continuous system converges locally to the stationary points with linear convergence rate. On the other hand, if Jacobian has eigenvalues with positive real-part, the continuous system is not locally convergent. If all eigenvalues have zero real part, it can be convergent, divergent or neither, but if it is convergent, it will generally converge with a sublinear rate. For simultaneous gradient descent linear convergence can be achieved if and only if all eigenvalues of the Jacobian of the gradient vector field have negative real part.

The eigenvalues are obtained in this way:

$$det(J^* - \lambda I) = 0 \rightarrow det(\begin{bmatrix} -\lambda & 0 & -0.5 \\ 0 & -0.5-\lambda & 0 \\ 0.5 & 0 & -\lambda \end{bmatrix}) = 0 \rightarrow -\lambda^3 - \lambda^2/2 - \frac{1}{8} - \frac{1}{4}\lambda = 0 \rightarrow \lambda_1 = 0.5i, \lambda_2 = -0.5i, \lambda_3 = -0.5$$

All eigenvalues of the Jacobian at a stationary point have negative real-part, so the continuous system converges locally to the stationary points with linear convergence rate.

4.4:

$\mathcal{R}_1(\psi) = \psi_0^2$

$$v(\theta, \psi) = \begin{pmatrix} v_{\psi_0} \\ v_{\psi_1} \\ v_\theta \end{pmatrix} = \begin{pmatrix} \nabla_{\psi 0}\mathcal{L}(\theta, \psi) - \gamma\psi_0 \\ \nabla_{\psi 1}\mathcal{L}(\theta, \psi) \\ -\nabla_\theta \mathcal{L}(\theta, \psi) \end{pmatrix}$$

for stationary points, set $v = 0$

$$v(\theta, \psi) = 0 \rightarrow \begin{pmatrix} -f'(-\psi_0\theta - \psi_1)\theta - \gamma\psi_0 \\ -f'(-\psi_0\theta - \psi_1) + f'(\psi_1) \\ f'(-\psi_0\theta - \psi_1)\psi_0 \end{pmatrix} = 0 \rightarrow \psi_0 = 0, \psi_1 = 0, \theta = 0$$

4.5:

$$J^* = \begin{bmatrix} \partial v_{\psi_0}/\partial\psi_0 & \partial v_{\psi_0}/\partial\psi_1 & \partial v_{\psi_0}/\partial\theta \\ \partial v_{\psi_1}/\partial\psi_0 & \partial v_{\psi_1}/\partial\psi_1 & \partial v_{\psi_0}/\partial\theta \\ \partial v_\theta/\partial\psi_0 & \partial v_\theta/\partial\psi_1 & \partial v_\theta/\partial\theta \end{bmatrix} = \begin{bmatrix} \theta^2 f'' - \gamma & \theta f' & -f' + \psi_0\theta f'' \\ \theta f'' & 2f'' & \psi_0 f'' \\ -\theta\psi_0 f'' + f' & -\psi_0 f'' & -\psi_0^2 f'' \end{bmatrix}$$

and at stationary point $(\psi_0 = 0, \psi_1 = 0, \theta = 0)$:

$f' = \frac{exp(-t)}{1+exp(-t)}$ and $f'' = \frac{-exp(-t)}{(1+exp(-t))^2} \rightarrow f'(0) = 1/2$ and $f''(0) = -1/4$

$$J^* = \begin{bmatrix} -\gamma & 0 & -0.5 \\ 0 & -0.5 & 0 \\ 0.5 & 0 & 0 \end{bmatrix}$$

4.6:

The eigenvalues are obtained in this way:

$$det(J^* - \lambda I) = 0 \rightarrow det(\begin{bmatrix} -\gamma - \lambda & 0 & -0.5 \\ 0 & -0.5 - \lambda & 0 \\ 0.5 & 0 & -\lambda \end{bmatrix}) = 0 \rightarrow -\lambda^3 - \lambda^2(\gamma + 0.5) - \lambda(\gamma - \tfrac{1}{4}) - \tfrac{1}{8} = 0$$

$$\lambda = [-\tfrac{\gamma}{2} - \sqrt{\tfrac{\gamma^2}{4} - f'(0)^2}, -\tfrac{\gamma}{2} + \sqrt{\tfrac{\gamma^2}{4} - f'(0)^2}] = [-\tfrac{\gamma}{2} - \sqrt{\tfrac{\gamma^2}{4} - 0.25}, -\tfrac{\gamma}{2} + \sqrt{\tfrac{\gamma^2}{4} - 0.25}]$$

The real parts are negative, so the system is convergent.

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.

5.1: **(a)**SimCLR uses contrastive self-supervised learning to leverage unlabeled datasets for representation learning. **(b)** A major contribution is using a combination of data augmentations. In this paper, they find out that using a sequence of random cropping followed by resize back to its original size, random color distortions, and random Gaussian blur significantly improves the performance of the model. Another key contribution is the formulation of the NT-Xent loss (normalized temperature-scaled cross entropy loss). This specific loss function allows for similar images to be learned to be mapped closer together. **(c)** The SimCLR algorithm explores the composition of data augmentations that are used to form positive and negative pairs of images for performing contrastive self-supervised learning by passing an original data point through two separate data augmentation functions. Then, the pair of images are fed to a CNN backbone (ResNet50) in batches to create vector representations for each of the images. After that, the outputs of the ResNet are passed to a projection head, which is one hidden layer MLP in this paper. And then, the projection head outputs are compared using a contrastive loss function. By using the contrastive loss function, the model tries to minimize the distance between images containing the same object and maximize the distance between images containing different objects. For evaluation, they remove the MLP projection head and only use the ResNet backbone model. They have done a few experiments. They show that as a linear classifier on the ImageNet dataset, SimCLR outperforms all other self-supervised methods. For finetuning, they show that SimCLR either outperforms or performs on par with supervised learning on 10 out of 12 datasets. Moreover, they evaluate the performance of SimCLR on different image datasets against supervised learning with the same ResNet model and show that SimCLR outperforms the supervised methods on most of the datasets.

5.2: **(a)**They find out that using a projection space (even a simple one hidden layer MLP) between the backbone of the model and the contrastive loss significantly improves the quality of the learned representations, which is a new theoretical insight. **(b)** This paper shows the composition of data Augmentations can be very useful in contrastive learning compared to supervised learning. The use of NT-Xent (the normalized temperature-scaled cross entropy loss) loss function **(c)** They show that by adding simple components such as adding augmentation and an MLP projection head, their approach performs a lot better than previous works even though they are not using a

memory bank (which requires a lot of RAM and computational costs) or specialized architectures. **(d)** Yes, the codes and algorithms are released, which can be used in the next works. **(e)** Yes, the experiments are well organized. The experiments on augmentation were very comprehensive and explored many sequences of augmentation. They have shown that augmentation is much more useful for self-supervised learning approaches compared to supervised learning. They have also done some experiments on color distortion and showed that it is critical to use cropping with color distortion. Moreover, they have done some ablation studies on batch size and have shown that a batch size of 4096 is the best, and bigger batch sizes do not perform better anymore. **(f)** Yes, this paper is very useful for the community, since a lot of other papers were published using the findings of this paper.

5.3: **(a)**One problem that can be mentioned is that the output of contrastive loss is not informative enough. After 100 epochs, supervised learning methods reach acceptable performance. However, for SimCLR to reach the same levels of accuracy as supervised learning, it needs to be trained for a lot longer, which shows that the contrastive loss is not informative enough. In the next works, such as the BYOL paper, by using a mean teacher method which is a self-supervised learning approach, it is shown that they can reach the same performance with a lot less number of epochs. Moreover, they are only using a 2-layer MLP for the projection head and they have not done any ablation study on the number of layers for the projection head. Probably, the model could benefit from a more complicated projection head, as other works have shown it, such as the CVRL and DINO paper.**(b)** In this paper, they have only used the ImageNet dataset. Since self-supervised learning methods actually need a lot of data for better performance, they could have explored bigger datasets such as JFT-300M and even the bigger ImageNet dataset. Moreover, they have only applied their method for classification tasks and they have not explored object detection, image segmentation, or localization. **(c)** They have compared the projection head size in one of their figures and found out that changing the size of the projection head does not have much effect on accuracy. So, it is not clear why they choose 2048 as the size of the projection head in their design and not any other lower number. Moreover, it is not clear where they find the formula for this loss function and why they are using this. The writing of the paper was of high quality, which made the paper easy to read and understand. **(d)** Yes, the novelty is sufficient. They have used a novel loss function which has performed very well in SimCLR. Moreover, the composition of several techniques such as data augmentation, projection head, and NT-Xent loss function was never done before. **(e)** Yes, the problem is interesting, and even though newer methods such as BYOL have been proposed, many self-supervised papers still use the contrastive loss for their work. Moreover, the next works still use the proposed novel loss function and the same sequence of data augmentations whenever needed.

5.4: **(a)** We studied self-supervised learning in class, so the paper was related. **(b)** The next research direction could be adding to the number of projection head layers or using a memory network like MoCo with a moving average of weights for stabilization. **(c)** One idea could be using this method for video, as they are the extension of the images. Another possible new idea could be applying contrastive loss to the output of DINO and seeing if it results in better performance.