Date remise: Le 5 décembre, 23h

Instructions

- *Montrez votre démarche pour toutes les questions !*
- *Utilisez LaTeX et le modèle que nous vous fournissons pour rédiger vos réponses. Vous pouvez réutiliser la plupart des raccourcis de notation, des équations et/ou des tableaux. SVP voir la politique des devoirs sur le site web du cours pour plus de détails.*
- *Vous devez soumettre toutes vos réponses sur la page Gradescope du cours*
- *Les TAs pour ce devoir sont* ***Arian Khorasani et Sarthak Mittal***

Question 1 (5)**. (Divergence de Kullback-Leibler)**

Étant donné deux distributions gaussiennes univariées, $p(x) = \mathcal{N}(\mu_1, \sigma_1^2)$ et $q(x) = \mathcal{N}(\mu_2, \sigma_2^2)$, trouver la $\mathbb{KL}$-Divergence entre la distribution q et la distribution p. En particulier, dériver l'expression de forme fermée pour

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

C'est le même que $\mathbb{KL}[p(x)||q(x)]$?

Answer 1. Q.1.

$$\mathbb{KL}[q(x)\|p(x)] = -\int q(x)\log p(x)dx + \int q(x)\log q(x)dx$$

Et on a ,

$$\begin{aligned}\int q(x)\log q(x)dx &= \int \left[-\frac{1}{2}log(2\pi\sigma_2^2) - \frac{(x-\mu_2)^2}{2\sigma_2^2}\right] q(x)dx \\ &= -\frac{1}{2}log(2\pi\sigma_2^2) - \mathbb{E}_{q(x)}\left[\frac{(x-\mu_2)^2}{2\sigma_2^2}\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{q(x)}\left[(x-\mu_2)^2\right] &= \mathbb{E}_{q(x)}(x^2 - 2x\mu_2 + \mu_2^2) \\ &= \mathbb{V}_{q(x)}(x) - \mathbb{E}_{q(x)}(x)^2 - 2\mu_2\mathbb{E}_{q(x)}(x) + \mu_2^2 \\ &= \sigma_2^2 - \mu_2^2 - 2\mu_2^2 + \mu_2^2 \\ &= \sigma_2^2\end{aligned}$$

Ainsi,

$$\int q(x)\log p(x)dx = -\frac{1}{2}log(2\pi\sigma_2^2) - \frac{1}{2}$$

Et,

$$\begin{aligned}\int q(x)\log p(x)dx &= \int \left[-\frac{1}{2}log(2\pi\sigma_1^2) - \frac{(x-\mu_1)^2}{2\sigma_1^2}\right] q(x)dx \\ &= -\frac{1}{2}log(2\pi\sigma_1^2) - \mathbb{E}_{q(x)}\left[\frac{(x-\mu_1)^2}{2\sigma_1^2}\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{q(x)}\left[(x-\mu_1)^2\right] &= \mathbb{E}_{q(x)}(x^2 - 2x\mu_1 + \mu_1^2) \\ &= \mathbb{V}_{q(x)}(x) + \mathbb{E}_{q(x)}(x)^2 - 2\mu_1\mathbb{E}_{q(x)}(x) + \mu_1^2 \\ &= \sigma_2^2 + \mu_2^2 - 2\mu_2\mu_1 + \mu_1^2 \\ &= \sigma_2^2 + (\mu_2 - \mu_1)^2\end{aligned}$$

Ainsi,

$$\int q(x)\log p(x)dx = -\frac{1}{2}log(2\pi\sigma_1^2) - \frac{\sigma_2^2 + (\mu_2-\mu_1)^2}{2\sigma_1^2}$$

Conclusion,

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

Pour $\mu_1 = \mu_2 = 0$ et $\sigma_1 = 1$ et $\sigma_2 = 2$, on a $\mathbb{KL}[q(x)\|p(x)] = -log(2) + 4/2 - 1/2 = 1.2$ et $\mathbb{KL}[p(x)\|q(x)] = log(2) + 1/8 - 1/2 = -0.74$

Donc

$$\mathbb{KL}[q(x)\|p(x)] \neq \mathbb{KL}[p(x)\|q(x)]$$

Question 2 (5-5-5-6)**. (Les flux normalisants)**

Les flux normalisants (*normalizing flows*) sont des transformations inversibles et expressives des lois de probabilités.

Dans cet exercice, nous allons explorer la possibilité d'inverser quelques transformations. Dans les 3 premières questions, nous considérons une fonction déterministe $g : z \in \mathbb{R} \to \mathbb{R}$.

1. Soient $g(z) = af\,(bz + c)$ et f est un redresseur (*rectified linear unit*) $f(x) = \max(0, x)$. Montrez que la fonction g n'est pas inversible.

2. Soit $g : z \in \mathbb{R} \mapsto \mathbb{R}$, où $g(z) = \sigma^{-1}(\sum_{i=1}^{N} w_i \sigma(a_i z + b_i))$, $0 < w_i < 1$, $\sum_i w_i = 1$, et $a_i > 0$. Montrez que g est *strictement croissante* sur son domaine de définition $(-\infty, \infty)$ [1]
3. Soit $g(z) = z + f(z)$ et $df/dz > -1$. Montrez que g est inversible.
4. Considérez la transformation suivante

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

où $\boldsymbol{z}_0 \in \mathbb{R}^D$, $\alpha \in \mathbb{R}^+$, $\beta \in \mathbb{R}$, et $r = ||\boldsymbol{z} - \boldsymbol{z}_0||_2$, $h(\alpha, r) = 1/(\alpha + r)$. Considérez également la décomposition suivante $\boldsymbol{z} = \boldsymbol{z}_0 + r\tilde{\boldsymbol{z}}$. (i) Étant donné $\boldsymbol{y} = g(\boldsymbol{z})$, montrez que $\beta \geq -\alpha$ est une condition suffisante pour obtenir une valeur unique de r à partir de l'équation (1). (ii) Étant donnés r et $\boldsymbol{y}$, montrez que l'équation (1) possède une unique solution $\tilde{\boldsymbol{z}}$.

Answer 2. Q.2.1. Si $b = 0$: donc g est constance ainsi non inversible, car non injective.

Si $b > 0$: On a tout $z < -\frac{c}{b}$ donne $g(z) = 0$, donc g n'est pas injective et ensuite elle ne sera pas inversible.

Si $b < 0$: On a tout $z > -\frac{c}{b}$ donne $g(z) = 0$, donc g n'est pas injective et ensuite elle ne sera pas inversible.

Q.2.2.

Posons $h(z) = \sum_{i=1}^{N} w_i \sigma\left(a_i z + b_i\right)$ ainsi, $g(z) = \sigma^{-1}(h(z))$:

$$\begin{aligned}\frac{dg}{dz}(z) &= h'(z).(\sigma^{-1})'(h(z)) \\ &= h'(z).\frac{1}{\sigma(\sigma^{-1}(h(z))(1-\sigma(\sigma^{-1}(h(z)))} \\ &= \frac{h'(z)}{h(z)(1-h(z))}\end{aligned}$$

Or , on a $0 < w_i \sigma\left(a_i z + b_i\right) < w_i$, ainsi $0 < h(z) < \sum_{i=1}^{N} w_i = 1$.

Donc le signe de $\frac{dg}{dz}(z)$ est le même que celui de $h'(z)$

Et

$$h'(z) = \sum_{i=1}^{N} w_i a_i \sigma\left(a_i z + b_i\right) * \left(1 - \sigma\left(a_i z + b_i\right)\right.$$

et $a_i > 0$, $w_i > 0$ et $0 < \sigma(x) < 1$ pour tout x

1. Pour écrire votre réponse à cette question, rappelez vous que si une fonction f est *strictement croissante*, alors elle est injective sur son domaine de définition. De plus, si T est l'image de la fonction f, alors f possède une fonction inverse f^{-1} sur T. Vous pouvez considérer une fonction *strictement croissante*, *i.e.* $df(x)/dx > 0$, par exemple.

Donc $h'(z) > 0$, alors $\frac{dg}{dz}(z) > 0$

Donc g est strictement croissante sur son domaine de définition R

Q.2.3.I. On a $dg/dz = df/dz + 1 > 0$, ainsi g est strictement croissante, elle est donc injective.

De plus, vu que f est continue car dérivable donc g l'est aussi. Et selon le théorème des valeurs intermédiaires, l'image d'un intervalle, en l'ocurrence $(-\infty, +\infty)$, est aussi un intervalle d'arrivée I_g. Et g sera surjective.

Q.2.4. On a $g(z) = z + \beta\frac{z-z_0}{\alpha+\|z-z_0\|_2}$ et on veut résoudre $g(z) = y$

Et avec la transformation suggérée, on obtient

$$\begin{aligned} g(r) &= z_0 + r\tilde{z} + \beta\frac{r\tilde{z}}{\alpha + r} \\ &= z_0 + \tilde{z}(r + \frac{r\beta}{\alpha + r}) \end{aligned}$$

Posons $l(r) = r + \frac{r\beta}{\alpha+r}$

On a

$$\begin{aligned} \frac{dl}{dr} &= 1 + \frac{\alpha\beta}{(\alpha + r)^2} \\ &= \frac{\alpha(\alpha + \beta) + r^2 + 2\alpha\beta}{(\alpha + r)^2} \end{aligned}$$

Supposons que $\beta \geq -\alpha$,

donc $\frac{dl}{dr}$ est strictement positive pour $r \neq 0$.

Et vu que l est **continue** et ne s'annule potentiellement que sur $r = 0$. Donc on peut dire qu'elle est strictement monotone donc inversible.

Ainsi il existe une unique solution de $y = g(z)$

Q.2.4.ii.

On a $g(r) = z_0 + (r + \beta\frac{r}{\alpha+r})\tilde{z}$ Ainsi, $y - z_0 = (r + \beta\frac{r}{\alpha+r})\tilde{z}$

Si le r donné est nul ou égal à $\alpha + \gamma$ donc $\tilde{z} = y$

Sinon $\tilde{z} = \frac{y-z_0}{(r+\beta\frac{r}{\alpha+r})}$

Ainsi, pour r et y donnés, on une unique solution pour $\tilde{z}$

Question 3 (3-5-2-2-3)**.** (**Combiné Auto-encodeur variationnel et modèle de diffusion**) DDPMs sont une classe de modèles génératifs qui s'appuient sur un processus de diffusion directe connu. $q(x_t|x_{t-1})$, qui détruit progressivement la structure des données jusqu'à ce qu'elle converge au bruit non structuré, par exemple. $\mathcal{N}(0, I)$ et un processus paramétré appris (par un réseau neuronal !) en arrière $p_\theta(x_{t-1}|x_t)$ qui élimine le bruit de façon itérative jusqu'à ce que vous ayez obtenu un échantillon de la distribution de données.

Soit $\boldsymbol{x}_0$ être un échantillon de la distribution des données ; et laisser le processus de diffusion directe (processus bruyant) être défini en utilisant

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

et le processus de diffusion inverse (processus de dénotation) étant un processus appris suivant

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

(a) Montre ça $\log p_\phi(\boldsymbol{x}_0) \geq \underbrace{\mathbb{E}_q \left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right]}_{\mathcal{L}_{DDPM}}$ et puis, montre ça

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

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

(c) Considérons maintenant un échantillon de données du formulaire $(\boldsymbol{x}_0, \boldsymbol{c})$, où $\boldsymbol{c}$ est maintenant des données auxiliaires supplémentaires qui vous ont été fournies (en particulier ; $\boldsymbol{c}$ peut être le même que $\boldsymbol{x}_0$ aussi bien). Supposons que nous modélisons les données comme suit $p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})$ où nos variables latentes sont maintenant $\boldsymbol{z}$ et $\boldsymbol{x}_{1:T}$. Puisque le postérieur sera intraitable, essayons de l'approcher avec $q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)$. Pouvez-vous maintenant re-dériver l'ELBO, qui est la limite inférieure de la probabilité logarithmique, comme $\log p(\boldsymbol{x}_0, \boldsymbol{c}) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}, \boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}_{0:T}, \boldsymbol{c}, \boldsymbol{z})}{q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{x}_0, \boldsymbol{c})}\right]$?

(d) Supposons maintenant que vous modélisez ce problème avec une combinaison de VAE et Modèle probabiliste de diffusion de débruitage, où l'encodeur du VAE a les paramètres ψ, le décodeur θ et

le modèle de débruitage ϕ. Dans ce cas, la distribution générative se factorise comme

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

En outre, supposons que nous voulons maintenant modéliser $\boldsymbol{z}$ utilisant un encodeur VAE avec paramètres ψ, et ensuite les autres variables latentes $\boldsymbol{x}_{1:T}$ subordonnée à $\boldsymbol{z}$ par le processus de diffusion directe. Pouvez-vous fournir une factorisation de $q(\boldsymbol{x}_{1:T}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{x}_0)$ qui respecte cela ?

(e) Par la manipulation arithmétique de l'ELBO vous avez dérivé ci-dessus ainsi que la factorisation que vous avez fournies, Pouvez-vous maintenant décomposer l'objectif en une composante VAE et une composante DDPM ?

Answer 3. Q3.a.

$$\begin{aligned} \log p_\phi(\boldsymbol{x}_0) - \mathbb{E}_q\left[\log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] &= \mathbb{E}_q\left[\log p_\phi(\boldsymbol{x}_0) - log \frac{p_\phi(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] \\ &= \mathbb{E}_q\left[-log \frac{p_\phi(\boldsymbol{x}_{0:T}) / p_\phi(\boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] \\ &= \mathbb{E}_q\left[log \frac{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{p_\phi(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] \\ &= \mathbb{KL}\left[q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) \| p_\phi(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)\right] \geq 0 \end{aligned}$$

Puisque la divergence KL et positive donc

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

Et,

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

Q.3.b.

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

Or pour $t > 1$:

$$\begin{aligned}q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) &= q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0\right) \\ &= \frac{q\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{x_0}\right)}{q\left(\boldsymbol{x}_{t-1}, \boldsymbol{x}_0\right)} \\ &= \frac{q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) q(\boldsymbol{x_0})}{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x_0}) q(\boldsymbol{x_0})}\end{aligned}$$

Ainsi, et en constatant le produit telescopique entre $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x_0})$ et (

$$\begin{aligned}\mathcal{L}_{DDPM} &= \mathbb{E}_q\left[\log p\left(\boldsymbol{x}_T\right) - \sum_{t=2}^{T} \log \frac{q\left(\boldsymbol{x}_l \mid \boldsymbol{x}_{l-1}\right)}{p_\phi\left(\boldsymbol{x}_{l-1} \mid \boldsymbol{x}_\ell\right)} - \log \frac{q\left(\boldsymbol{x}_1 \mid \boldsymbol{x}_0\right)}{p_\phi\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)}\right] \\ &= \mathbb{E}_q\left[\log p\left(\boldsymbol{x}_T\right) - \sum_{t=2}^{T} \log \frac{q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)}{p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)} + \log \frac{q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right)}{q\left(\boldsymbol{x}_1 \mid \boldsymbol{x}_0\right)} - \log \frac{q\left(\boldsymbol{x}_1 \mid \boldsymbol{x}_0\right)}{p_\phi\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)}\right] \\ &= \mathbb{E}_q\left[-\log \frac{q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right)}{p_\theta\left(\boldsymbol{x}_T\right)} - \sum_{t=2}^{T} \log \frac{q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right)}{p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)} + \log p_\phi\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)\right] \\ &= \mathbb{E}_q\left[\log p_\phi\left(x_0 \mid x_1\right) - \sum_{t=2}^{T} \mathbb{KL}\left[q\left(x_{t-1} \mid x_t, x_0\right) \| p_\phi\left(x_{t-1} \mid x_t\right)\right] - \mathbb{KL}\left[q\left(x_T \mid x_0\right) \| p\left(x_T\right)\right]\right]\end{aligned}$$

Question 4 (1-2-1-1-2-3). (**Réseaux antagonistes génératifs**)

Dans cette question, nous souhaitons analyser la dynamique de l'entrainement de GAN sous l'ascension-descente de gradient. On dénote les paramètres du critique et du générateur respectivement par ψ et θ. La fonction objectif considérée est la Jensen-Shannon (standard):

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

où σ est la fonction logistique. Pour simplifier l'exposition du problème, nous considérerons le système à temps continu qui résulte du système à temps discret (alternant) quand le taux d'apprentissage, $\eta > 0$, approche zéro:

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

Le but est d'étudier la stabilité de l'algorithme d'entrainement. Pour cette raison, nous utiliserons un cadre simple: les données d'entrainement et générées ont le même support sur $\mathbb{R}$. De plus, $p_D = \delta_0$ et $p_\theta = \delta_\theta$. Cela signifie que les deux sont des distributions de Dirac [2] qui sont centrées en $x = 0$, pour les vraies données, et en $x = \theta$, pour les données générées.

Le critique, $C_\psi : \mathbb{R} \to \mathbb{R}$, est $C_\psi(x) = \psi_0 x + \psi_1$.

4.1 Dérivez les expressions pour le champ de "vélocité", v, du système dynamique dans l'espace de paramètres (ψ_0, ψ_1, θ), et trouvez les points stationnaires $(\psi_0^*, \psi_1^*, \theta^*)$. [3]

4.2 Dérivez J^*, la jacobienne (3×3) de v au point $(\psi_0^*, \psi_1^*, \theta^*)$.

Pour qu'un système à temps continu soit localement asymptotiquement stable, il suffit que toutes les valeurs propres de J^* aient des parties réelles négatives. Si ce n'est pas le cas, l'étude nécessaire pour conclure est plus complexe et malheureusement, la vitesse de convergence sera au mieux sublinéaire.

4.3 Trouvez les valeurs propres de J^* et discutez de la stabilité locale autour des points stationnaires.

Maintenant, introduisons une pénalité du gradient à la fonction de perte du critique, $\mathcal{R}_1(\psi) = \mathbb{E}_{p_D} \|\nabla_x C_\psi(x)\|^2$. Le système régularisé devient:

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

pour $\gamma > 0$. Répétez les étapes 1-2-3 pour ce système modifié et comparez la stabilité des deux systèmes.

4.4 Dérivez les expressions pour le champ de "vélocité", v, du système dynamique dans l'espace de paramètres (ψ_0, ψ_1, θ), et trouvez les points stationnaires $(\psi_0^*, \psi_1^*, \theta^*)$. [4]

4.5 Dérivez J^*, la jacobienne (3×3) de v au point $(\psi_0^*, \psi_1^*, \theta^*)$.

4.6 Trouvez les valeurs propres de J^* et discutez de la stabilité locale autour des points stationnaires.

Dans la partie pratique du devoir, vous pourrez vérifier empiriquement vos conclusions.

Answer 4. Q.4.1. Nous avons que ,

$$\begin{aligned}\mathcal{L}(\psi, \theta) &= \log\left(\sigma\left(C_\psi(0)\right)\right) + \log\left(\sigma\left(-C_\psi(\theta)\right)\right) \\ &= \log(\sigma(\psi_1)) + \log(\sigma(-\psi_0\theta - \psi_1))\end{aligned}$$

2. Si $p_X = \delta_z$, alors $p(X = z) = 1$.
3. Pour trouver les points stationnaires, fixez $v = 0$ et résolvez les équations pour chaque paramètre.
4. Pour trouver les points stationnaires, fixez $v = 0$ et résolvez les équations pour chaque paramètre.

Et on a que $v_\psi(\psi,\theta) = \nabla_\psi \mathcal{L}(\psi,\theta)$ et $v_\theta(\psi,\theta) = -\nabla_\theta \mathcal{L}(\psi,\theta)$

$$\begin{cases} v_{\psi_0} = -\theta(1-\sigma(-\psi_0\theta-\psi_1)) \\ v_{\psi_1} = \sigma(-\psi_0\theta-\psi_1) - \sigma(\psi_1) \\ v_\theta = \psi_0(1-\sigma(-\psi_0\theta-\psi_1)) \end{cases}$$

Ainsi pour $v = 0$; on obtient

$$\begin{cases} \theta^* = 0 \\ 0 = \sigma(-\psi_1^*) - \sigma(\psi_1^*) \\ \psi_0^* = 0 \end{cases}$$

Donc, enfin on obtient :

$$\begin{cases} \theta^* = 0 \\ \psi_1^* = 0 \\ \psi_0^* = 0 \end{cases}$$

Q.4.2.

Calcul du Jacobien J^*:

$$J^* = \begin{bmatrix} \frac{\partial v_\theta}{\partial \theta} & \frac{\partial v_\theta}{\partial \psi_0} & \frac{\partial v_\theta}{\partial \psi_1} \\ \frac{\partial v_{\psi_0}}{\partial \theta} & \frac{\partial v_{\psi_0}}{\partial \psi_0} & \frac{\partial v_{\psi_0}}{\partial \psi_1} \\ \frac{\partial v_{\psi_1}}{\partial \theta} & \frac{\partial v_{\psi_1}}{\partial \psi_0} & \frac{\partial v_{\psi_1}}{\partial \psi_1} \end{bmatrix} (\psi_0^*, \psi_1^*, \theta^*)$$

$$J^* = \begin{bmatrix} 0 * \frac{\partial v_\theta}{\partial \theta}(1-\sigma(-\psi_0\theta-\psi_1)) & 1-\sigma(0) + 0 * \frac{\partial v_\theta}{\partial \psi_0}(1-\sigma(-\psi_0\theta-\psi_1)) & 0 * \frac{\partial v_\theta}{\partial \psi_1}(1-\sigma(-\psi_0\theta-\psi_1)) \\ -(1-\sigma(0)) - 0 * \frac{\partial v_{\psi_0}}{\partial \theta}(1-\sigma(-\psi_0\theta-\psi_1)) & -0 * \frac{\partial v_{\psi_0}}{\partial \psi_0}(1-\sigma(-\psi_0\theta-\psi_1)) & -0 * \frac{\partial v_{\psi_0}}{\partial \psi_1}(1-\sigma(-\psi_0\theta-\psi_1)) \\ -0 * \sigma'(-\psi_0\theta-\psi_1) & -0 * \sigma'(-\psi_0\theta-\psi_1) & -\sigma'(0) - \sigma'(0) \end{bmatrix}$$

$$J^* = \begin{bmatrix} 0 & 1/2 & 0 \\ -1/2 & 0 & 0 \\ 0 & 0 & -1/2 \end{bmatrix}$$

Q.4.3. On a pour

$$v_1 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} -i & 1 & 0 \end{bmatrix}$$

$$v_3 = \begin{bmatrix} i & 1 & 0 \end{bmatrix}$$

On a

$$J^* v_1 = -1/2 v_1$$
$$J^* v_2 = i/2 v_2$$
$$J^* v_3 = -i/2 v_3$$

Ainsi les valeurs propres sont -0.5, $0.5i$ et $-0.5i$ avec des parties réelles négatives ou nulles.

Donc le système n'est pas asymptotiquement stable.

Q.4.4.

En rajoutant la régularisation on trouve les expressions suivantes de la vélocité.

$$\begin{cases} v_{\psi_0} = -\theta\left(1 - \sigma\left(-\psi_0\theta - \psi_1\right)\right) - \gamma\psi_0 \\ v_{\psi_1} = \sigma\left(-\psi_0\theta - \psi_1\right) - \sigma\left(\psi_1\right) \\ v_0 = \psi_0\left(1 - \sigma\left(-\psi_0\theta - \psi_1\right)\right) \end{cases}$$

pour $v = 0$, on obtient

$$\begin{cases} \theta^* = 0 \\ 0 = \sigma\left(-\psi_1^*\right) - \sigma\left(\psi_1^*\right) \\ \psi_0^* = 0 \end{cases}$$

On obtient donc le même point stable.

$$\begin{cases} \theta^* = 0 \\ \psi_1^* = 0 \\ \psi_0^* = 0 \end{cases}$$

Q.4.5.

On obtient pour la matrice jacobienne:

$$J^* = \begin{bmatrix} 0 * \frac{\partial v_\theta}{\partial \theta}(1 - \sigma(-\psi_0\theta - \psi_1)) & 1 - \sigma(0) + 0 * \frac{\partial v_\theta}{\partial \psi_0}(1 - \sigma(-\psi_0\theta - \psi_1)) & 0 * \frac{\partial v_\theta}{\partial \psi_1}(1 - \sigma(-\psi_0\theta - \psi_1)) \\ -(1 - \sigma(0)) - 0 * \frac{\partial v_{\psi_0}}{\partial \theta}(1 - \sigma(-\psi_0\theta - \psi_1)) & -0 * \frac{\partial v_{\psi_0}}{\partial \psi_0}(1 - \sigma(-\psi_0\theta - \psi_1)) - \gamma & -0 * \frac{\partial v_{\psi_0}}{\partial \psi_1}(1 - \sigma(-\psi_0\theta - \psi_1)) \\ -0 * \sigma'(-\psi_0\theta - \psi_1) & -0 * \sigma'(-\psi_0\theta - \psi_1) & -\sigma'(0) - \sigma'(0) \end{bmatrix}$$

$$J^* = \begin{bmatrix} 0 & 1/2 & 0 \\ -1/2 & -\gamma & 0 \\ 0 & 0 & -1/2 \end{bmatrix}$$

Q.4.6.

On a

$$det(J^* - \lambda\mathbb{I}) = -\frac{(2\lambda+1)\,(4\lambda^2+4\gamma\lambda+1)}{8}$$

On résout :

$$\left(\lambda^2+\gamma\lambda+1/4\right)=0$$

On a

$$\Delta = \gamma^2 - 1$$

Si $\gamma \geq 1$ donc les solutions sont :

$$\begin{cases} \lambda_1 = -1/2 \\ \lambda_2 = \frac{-\gamma-\sqrt{\Delta}}{2} \\ \lambda_3 = \frac{-\gamma+\sqrt{\Delta}}{2} \end{cases}$$

Donc dans ce cas le systeme est asymptotiquement stable car toutes les valeurs propres sont réelles et négatives

Si $\gamma < 1$ donc les solutions sont : Si $\gamma \geq 1$ donc les solutions sont :

$$\begin{cases} \lambda_1 = -1/2 \\ \lambda_2 = \frac{-\gamma-i\sqrt{\Delta}}{2} \\ \lambda_3 = \frac{-\gamma+i\sqrt{\Delta}}{2} \end{cases}$$

Encore une fois les valeurs propres ont une partie réelle négative. Ainsi, le système est asymptotiquement stable.

Question 5 (20 pts)**.** (**Revue de papier**)

dans cette question, vous allez écrire **une page** révision de Apprentissage auto-supervisé. Veuillez structurer votre examen dans les sections suivantes:

1. **Sommaire [5 pts]**:
   (a) De quoi parle cet article ?
   (b) Quelle est la principale contribution ?
   (c) Décrivez l'approche principale et les résultats. Seulement des faits, pas encore d'opinions.
2. **Forces [5 pts]**:
   (a) Y a-t-il un nouvel aperçu théorique ?
   (b) Ou un progrès empirique important ? Ont-ils résolu un problème permanent ?
   (c) Ou une bonne formulation pour un nouveau problème ?
   (d) Des résultats concrets (code, algorithme, etc.) ?
   (e) Les expériences sont-elles bien exécutées ?
   (f) Utile pour la communauté en général ?

3. **Faiblesses [5 pts]** :
   (a) Que peut-on faire de mieux ?
   (b) Des bases de référence manquantes ? Des ensembles de données manquants ?
   (c) Des choix de conception bizarres dans l'algorithme ne sont-ils pas bien expliqués ? Qualité de l'écriture ?
   (d) Est-ce qu'il y a suffisamment de nouveauté dans ce qu'ils proposent ? Variation mineure de travaux antérieurs ?
   (e) Pourquoi devrait-on s'en soucier ? Le problème est-il intéressant et important ?
4. **Reflets [5 pts]**:
   (a) Quel est le lien avec d'autres concepts que vous avez vus dans la classe ?
   (b) Quelles sont les prochaines orientations de recherche dans ce domaine ?
   (c) Quelles nouvelles idées (directement ou indirectement liées) ce document vous a-t-il donné ? Qu'est-ce que tu voudrais essayer ?

The analysis is written in English due to the key words.

Answer 5. 4.1. The paper talks about SimCLR as self-supervised algorithm by the introduction of **the use of data augementation**. The images used as input go through different transformations or augmentation like cropping, resizing rotating or adding noise to it. The first step of the algorithm is to pass through a batch of the data which creates from each image two augmented version. These new images are then passed through a ResNet model that serves as an encoder. After the encoder, the images are then projected by the use of an MLP with one hidden layer. Therefore, a new set of representation vectors are received. And now, the aim of the learning algorithm is to bring every two vector representation of the augmentation of the **the same image** closer, by minimizing the loss between images referring to the same object (positive pairs) . **The results** show that SimCLR outperforms the regular supervised learning models for the small datasets such as CIFAR 10 CIFAR 100, but has almost the same performance for the big datasets such as Caltech-101.The model also outperforms the other self-supevised models. The evaluation of the self-supervised models is used on applying its new data reprsentation on supervised s tasks.

4.2 **Strengths** The new theritical aspect that has been used is the **the data augmentation** to the model the loss function also is a new one that allows this particular self-supervised learning model to achieve great performances. In fact, the model is really empirically impressive for the datasets classification CIFAR10 and CIFAR100. **This paper solves the problem** of how getting features out from a non labeled data using self-supervised-learning. This kind of models already exist. But SimCLR explaines the contrastive loss that allows the encoder to focus on the global features but also on the local ones. So it actually is a new formulation of self supervised representation learning algorithms **The practical outcome** would be both the introduction of data augmentation to make the model learn better hidden local and global features. Also **the contrastive loss** can be used in other hybrid models that might even outperform this one. **The experiments** were well executed on different datasets. And it is very useful to the community since it introduces this new loss and according to google scholar this article was cited 7483 times, and was described by many in the community to be one of the best AI papers of 2020.

4.3. **Weaknesses** I think the best thing that can be added to the paper is how the choice of the data augmentation is made. What data augmentation technique to use for what desired task from the model. The paper actually does provide all the datasets and baselines. The datasets are public. However, the used datasets are small or medium no large datasets were used. The algorithm is quite well explained the evaluation was only on classification but not image recognition for example which is one of the most useful task in Computer Vision. The choices in the algorithm are well explained and the writting is clean and easy to understand. **The new added value** would be the use of data augmentation since contrastive loss has already been introduced in 2005 by Yann Le Cunn. Hence, the paper just adds the data augmentation method to models that already exist MLP, ResNEt. Nonetheless, this paper is quite important since it uses the "right loss" function even though it was invented before. by the combination between this loss function and the data augmentation given intersting results.

4.4.This paper falls whithin the self-supervised learning that has been discussed in class. I think the next research aspects would be to rethink this model in its implementations for tasks of NLP for example or graph analysis. It actually originated in the creation of a new model called DGI. The

DGI (Deep Graph InfoMax) is a similar graph that operates the data augmentation on graphs by randomizing the vertexes and the nodes features. I actually am familiar with SimCLR and DGI, that I have implemented while working on a pharmaceutical company. This kind of algorithms are very much used in cancer recognition in MRI images. I think the most interesting idea I got from this paper would actually be the data augmenation that I would use in detecting hidden local features in other types of tasks. NLP (mixing words together, using synonyms, etc.)