Instructions

- *Pour toutes les questions, montrez votre travail !*
- *Utiliser un système de rédaction de documents tel que LaTeX.*
- *Soumettez vos réponses par voie électronique via le système de notation du cours*
- *les TAs pour ce devoir sont (theoretical part) :* ***Alexandra Volokhova*** *(IFT6135B) and* ***Ghait Boukachab*** *(IFT6135A).*

Question 1 (2-2-4-2)**.** Considérons un modèle de variable latente $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})dz$, où $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ et $\boldsymbol{z} \in \mathbb{R}^K$.Le réseau encodeur (également appelé "modèle de reconnaissance") de l'autoencodeur variationnel, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, est utilisé pour produire une distribution postérieure approximative (variationnelle) sur les variables latentes $\boldsymbol{z}$ pour tout point de données d'entrée $\boldsymbol{x}$.[1] Cette distribution est entraînée pour correspondre à la vraie postériorité en maximisant la limite inférieure de l'évidence (ELBO) :

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x})||p(\boldsymbol{z}))$$

On suppose que $q_\phi \in \mathcal{Q}$ où $\mathcal{Q}$ est une famille paramétrique, où nous indiquons ϕ pour spécifier quel membre de la famille nous utilisons.

1.1 Montrer que la log-vraisemblance des données $\log p_\theta(\boldsymbol{x})$ peut être décomposée en une somme d'ELBO et de divergence de KL entre les postérités variationnelles et réelles sur $\boldsymbol{z}$: $D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.

1.2 Montrer que la maximisation de ELBO par rapport à ϕ est équivalente à la minimisation de la divergence de KL entre les postérités variationnelles et réelles sur $\boldsymbol{z}$ par rapport à ϕ.

1.3 Dans cette sous-question et dans la suivante, l'objectif est de comparer l'inférence variationnelle armotisée (lorsque q_ϕ est optimisé pour l'ensemble des données) avec l'inférence variationnelle traditionnelle (lorsque q_ϕ est optimisé individuellement pour chaque x). Considérons un ensemble d'apprentissage fini $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n étant la taille des données d'apprentissage. Fixons θ pour plus de simplicité. Soit $q^* = \arg\max_{q_\phi \in \mathcal{Q}} \sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ (c'est-à-dire que q^* est la distribution variationnelle optimale dans la famille $\mathcal{Q}$ pour un θ et un ensemble d'apprentissage donnés). En outre, pour chaque $\boldsymbol{x}_i$, soit $q_i^* = \arg\max_{q_\phi \in \mathcal{Q}} \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$. Comparez $D_{\mathrm{KL}}(q^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ et $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Quel est le plus grand ?

1.4 Suite à la question précédente, comparez les deux approches dans la deuxième sous-question (justifiez les réponses).

(a) Quelle est la meilleure approche pour estimer la vraisemblance marginale via ELBO empirique ?

(b) Laquelle est la plus efficace en temps de calcul

(c) Laquelle est la plus efficace en termes de capacité de mémoire (stockage des paramètres)

Question 2 (3*-2-7-2-2-5-2)**.** Dans cette question, nous allons creuser plus profondément dans les mathématiques des modèles de diffusion. Considérons un modèle probabiliste de diffusion à dispersion (DDPM) avec le processus encodeur donné par un modèle gaussien linéaire : $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I)$, où $\beta_t \in (0, 1)$ est un schéma fixe de bruits. Le processus de diffusion

1. L'utilisation d'un modèle de reconnaissance de cette manière est connue sous le nom "d'inférence amortie" ; on peut l'opposer aux approches traditionnelles d'inférence variationnelle (voir, par exemple, le chapitre 10 de l'ouvrage de Bishop *Pattern Recognition an Machine Learning*), qui ajustent une postériorité variationnelle de manière indépendante pour chaque nouveau point de données.

"forward" commence par l'image initiale $\boldsymbol{x}_0$ de l'ensemble de données et se termine à $\boldsymbol{x}_T$ (T est un nombre fixe d'étapes). Nous supposons que $\boldsymbol{x}_T \sim p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T|0, I)$. L'objectif de l'entraînement est d'apprendre un processus inversé (processus de débruitage) $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, qui permettra de générer une image $\boldsymbol{x}_0$ à partir d'un bruit gaussien $\boldsymbol{x}_T$.

2.1 Etant donné l'équation du processus encodeur linéaire gauassien $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, montrez que le processus de débruitage de la "ground truth" est

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}|\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t I) \tag{1}$$

où

$$\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t \tag{2}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

$$\alpha_t = 1 - \beta_t$$

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$$

Si nécessaire, vous pouvez utiliser l'équation suivante sans la prouver :

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I) \tag{3}$$

Indice : utiliser la règle de Bayes et la propriété markovienne du processus encodeur.

2.2 Comme nous l'avons vu dans la tâche 2.1, il est possible d'inverser le processus de diffusion analytiquement, sans entraînement. Expliquer pourquoi nous avons toujours besoin d'entraîner le processus inverse $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ pour générer des images.

2.3 Montrons maintenant la fonction objective du DDPM. Essentiellement, le DDPM est un autoencodeur variationnel hiérarchique (avec des variables latentes $\{x_1, \ldots, x_T\}$) et son objectif est une limite d'évidence (ELBO) pour $\log p(x_0)$. Montrer que

$$\log p(x_0) \geq \mathcal{L}_{DDPM}(\theta; \boldsymbol{x}_0) = -L_0(\boldsymbol{x}_0) - \sum_{t=2}^{T} L_{t-1}(\boldsymbol{x}_0) - L_T(\boldsymbol{x}_0)$$

où

- (terme de reconstruction) $L_0(\boldsymbol{x}_0) = -\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)$
- (terme de correspondance pour le débruitage) $L_{t-1}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))$
- (terme de correspondance préalable) $L_T(\boldsymbol{x}_0) = D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T))$

Les équations suivantes peuvent être utiles pour les dérivations :

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

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$$

2.4 Quel terme de $\mathcal{L}_{DDPM}$ n'affecte pas l'optimisation des paramètres et peut donc être exclu de la fonction objective ?

2.5 Comparez ELBO pour vanila VAE (voir question précédente) et ELBO pour DDPM. Quelle est la principale différence entre eux (en termes de paramètres entraînables) ?

2.6 Analysons $L_{t-1}(\boldsymbol{x}_0)$ et $L_0(x_0)$ plus en détail.

- En utilisant l'éq. 2 et 3, montrer que $\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\big)$, où $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, I)$
- Une paramétrisation courante du processus de débruitage est la suivante $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\mu_\theta(\boldsymbol{x}_t, t), \sigma_t^2 I)$, où la moyenne de la gaussienne $\mu_\theta(\boldsymbol{x}_t, t)$ est entraînable (nous considérons ici que σ_t^2 est fixe pour des raisons de simplicité, alors qu'en pratique il est possible de l'entraîner).Cependant, au lieu d'entraîner un modèle pour prédire directement le $\mu_\theta(\boldsymbol{x}_t, t)$, un choix courant consiste à entraîner un réseau de neurones $\boldsymbol{\epsilon}_\theta$ (également appelé "débruiteur") pour prédire uniquement le terme de bruit : $\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\big)$. Montrer que

$$\mathbb{E}_{q(\boldsymbol{x}_0)} L_{t-1}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_0), \boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\Big[\lambda_t ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\big)||^2\Big] + const \tag{4}$$

où $\lambda_t = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}$ et $q(\boldsymbol{x}_0)$ est la distribution des données "groundtruth". Astuce : vous pouvez utiliser l'équation de la divergence KL entre des distributions normales multivariées sans la déterminer.

- Montrer que $\mathbb{E}_{q(\boldsymbol{x}_0)} L_0(\boldsymbol{x}_0)$ peut être écrit de la même manière que l'éq. 4

2.7 Enfin, rassemblez les équations pour les termes ELBO et obtenez la fonction de coût DDPM.

Question 3 (3-7)**.** Soit p_0 et p_1 deux distributions de probabilités avec les densités f_0 et f_1 (respectivement). Nous souhaitons explorer ce qu'on peut faire avec le discriminateur d'un GAN entraîné. Un discriminateur entraîné est considéré comme "proche" d'un discriminateur optimal :

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(\boldsymbol{x}))].$$

3.1 Pour la première partie de ce problème, dérivez une expression permettant d'estimer la divergence de Jensen-Shannon (JSD) d'un discriminateur entraîné. Comme rappel, la JSD est $\text{JSD}(p_0, p_1) = \frac{1}{2}\big(KL(p_0\|\mu) + KL(p_1\|\mu)\big)$, où $\mu = \frac{1}{2}(p_0 + p_1)$.

3.2 Pour la seconde partie, nous souhaitons démontrer qu'un discriminateur optimal d'un GAN (c.-à-d. un discriminateur qui peut distinguer des exemples provenant de p_0 et de p_1 avec une perte NLL minimale) peut être utilisé pour exprimer la densité de probabilité d'un exemple $\boldsymbol{x}$ sous f_1, $f_1(\boldsymbol{x})$ en termes de $f_0(\boldsymbol{x})$ [2]. Assumez que f_0 et f_1 ont le même support. Montrez que $f_1(\boldsymbol{x})$ peut être estimé par $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1 - D(\boldsymbol{x}))$ en établissant l'identité $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$.

Astuce : Trouvez la solution analytique de D^*.

Question 4 (4-2-8-4-2)**.** Dans cette question, nous allons voir que le *stop-gradient* est un élément critique pour les méthodes d'auto-supervision non contrastive comme SimSiam and BYOL. Nous allons démontrer que d'enlever le stop-gradient donne une représentation trivial en utilisant la dynamic de SimSiam comme example.

Considérez une réseau SimSiam a deux couches avec une matrice évoluant dans le temps donnée par $W(t) \in \mathbb{R}^{n_2 \times n_1}$. Notez que $W(t)$ correspond au poids des réseaux online et target alors que

2. Il est possible que vous ayez à utiliser la *dérivée fonctionnelle* pour résoudre ce problème. Pour davantages d'information, voir "19.4.2 Calculus of Variations" du livre *Deep Learning* ou l'appendice "D. Calculus of Variations" du livre *Pattern Recognition and Machine Learning* de Bishop.

$W_p(t)$ dénote les pods du prédicteur. Soit $\boldsymbol{x} \in \mathbb{R}^{n_1}$ une donnée d'entrée et $\boldsymbol{x}_1, \boldsymbol{x}_2$ deux versions augmentées de $\boldsymbol{x}$. Notez aussi que dans certaines instances, la dépendences sur le temps (t) est omis pour simplifier la notation et que les matrics de poids sont référencées par W et W_p.

Soit $\boldsymbol{f}_1 = W\boldsymbol{x}_1$ la représentation online de $\boldsymbol{x}_1$ et $\boldsymbol{f}_2 = W\boldsymbol{x}_2$ la représentation target de $\boldsymbol{x}_2$. Les dynamiques d'apprentissage de W et W_p sont obtenues en minimisant la fonction d'objectif de SimSiam comme montrée ci-dessous :

$$J(W, W_p) = \frac{1}{2}\mathbb{E}_{x_1,x_2}\left[\|W_p\boldsymbol{f}_1 - \text{Stop-Grad}(\boldsymbol{f}_2)\|_2^2\right], \tag{5}$$

4.1 Montrez, avec preuve, que l'objectif ci-haut peut être simplfiée comme suit :

$$J(W, W_p) = \frac{1}{2}\left[tr(W_p^\intercal W_p F_1) - tr(W_p F_{12}) - tr(F_{12}W_p) + tr(F_2))\right], \tag{6}$$

où $F_1 = F_2 = \mathbb{E}\left[\boldsymbol{f}_1\boldsymbol{f}_1^\intercal\right] = W(X + X')W^\intercal$ et $F_{12} = F_{21} = \mathbb{E}\left[\boldsymbol{f}_1\boldsymbol{f}_2^\intercal\right] = WXW^\intercal$. Ici, X est la vue augmentée moyenne d'un point de donnée $\boldsymbol{x}$ et X' est la matrice de covariance de vues augmentées $\boldsymbol{x}'$ conditionnées sur $\boldsymbol{x}$ et ensuite moyennée sur des données $\boldsymbol{x}$ et lr est l'opération de trace [3].

4.2 En vous basant sur l'expression ci-dessus pour $J(W, W_p)$, trouvez l'expression de mise-à-jour du gradient pour W_p (soit le réseau prédicteur). Autrement dit, obtenez une expression pour $\dot{W}_p = -\dfrac{\partial J}{\partial W_p}$ (la dérivée de la fonction d'objectif par rapport aux paramètres W_p.)

4.3 Considérez le cas où le Stop-Grad est enlevé. Le gradient de la fonction d'objectif $J(W, W_p)$ par rapport aux paramètres W (autrement dit $\dot{W}(t) = -\dfrac{\partial J}{\partial W(t)}$ est donné par) :

$$\dot{W}(t) = \frac{d}{dt}\text{vec}(W) = -H(t)\text{vec}(W),$$

où $H(t)$ est une matrice semi-definie positive definie changeant dans le temps défini par

$$H(t) = X' \otimes (W_p^\intercal W_p + I_{n_2}) + X \otimes (\tilde{W}_p^\intercal \tilde{W}_p).$$

Ici, $\otimes$ est le produt de Kronecker [4] , $\tilde{W}_p = (W_p - I_{n_2})$, et "vec(W)" font références à la *vectorisation* de la matrice W [5]. Par simplicité, nous ne considérons pas le weight decay ici [6].

Si la valeur propre minimal $\lambda_{min}(H(t))$ est bornée loin de zéro, i.e. $\inf_{t\geq 0}\lambda_{min}(H(t)) \geq \lambda_0 > 0$, alors **provez que** $W(t) \to 0$.

Note : Pour prouvez la question ci-dessus, la propriété suivante doit être utilisé :

Pour une matrice definite positive varyant dans le temps $H(t)$ dont la valeur propre minimal est bornée loin de 0, la dynamique montrée ci-dessous :

$$\frac{d}{dt}\boldsymbol{w}(t)) = -H(t)\boldsymbol{w}(t),$$

satisfait la contrainte $\|\boldsymbol{w}(t)\|_2 = e^{-\lambda_0 t}\|\boldsymbol{w}(0)\|_2$, impliquant que $\boldsymbol{w}(t) \to 0$.

3. https ://en.wikipedia.org/wiki/Trace_(linear_algebra)*https* : *//en.wikipedia.org/wiki/Trace_(linear_algebra)*

4. Pour plus d'information, voyez https ://en.wikipedia.org/wiki/Kronecker_product#Matrix_equationshttps ://en.wikipedia

5. Aussi connu comme le "vec trick", il est obtenue en empilant toutes les colonnes de la matrice A dans un seul vecteur.

6. Nous devons noter que l'utilisation du weight decay est importante. Il a été montré, un pratique, que weight decay mène à une apprentissage stable.

4.4 Considérez le cas où le Stop-Gradient **et** le prédicteur sont enlevés. Montrez que la représentation converge à la solution trivial, soit $W(t) \to 0$. Assumez que X' est una matrice définie positif.

4.5 Spéculez (en 1-2 phrases) pourquoi le stop-gradient et le prédicteur sont nécéssaires pour éviter une représentation triviale.