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 (Partie théorique) :* ***Alexandra Volokhova*** *(IFT6135B) et* ***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 $\boldsymbol{x}$. Cette distribution est entraînée pour correspondre à la vraie postériorité en maximisant la limite inférieure de l'évidence (ELBO) :

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

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 $\boldsymbol{z}$: $D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x}))$.

Réponse :

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) &= \log p_\theta(\boldsymbol{x}) + \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] - \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && (1)\\
&= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] - \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x})\right] && (2)\\
&= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] - \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} - \log p_\theta(\boldsymbol{x})\right] && (3)\\
&= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] - \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})p_\theta(\boldsymbol{x})}\right] && (4)\\
&= \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && (5)\\
\log p_\theta(\boldsymbol{x}) &= \mathcal{L}_{ELBO}(\theta, \phi; \boldsymbol{x}) + D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right] && (6)
\end{aligned}$$

À (2) on peut faire l'espérance de $p_\theta(\boldsymbol{x})$ sur $p_\theta(\boldsymbol{z}|\boldsymbol{x})$, car $\mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x})\right] = \int p_\theta(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}) d\boldsymbol{z} = \log p_\theta(\boldsymbol{x}) \int p_\theta(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} = \log p_\theta(\boldsymbol{x})$. À (5) on a utilisé $p_\theta(\boldsymbol{x}, \boldsymbol{z}) = p_\theta(\boldsymbol{z}|\boldsymbol{x})p_\theta(\boldsymbol{x})$. Enfin, on a le résultat final que la log-vraisemblance des données peut être décomposée en une somme d'ELBO et de divergence de KL entre les postérités variationnelles et réelles sur $\boldsymbol{z}$: $D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]$.

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.

Réponse : En réarangeant l'expression précédente on trouve,

$$\mathcal{L}_{ELBO}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x})\right] - D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right] \quad (7)$$

La maximisation de ELBO par rapport à ϕ est,

$$\max_{\phi}\{\mathcal{L}_{ELBO}(\theta,\phi;\boldsymbol{x})\} = \max_{\phi}\{\mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x})\right] - D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]\} \tag{8}$$

En réarrangeant $\mathcal{L}_{ELBO}(\theta,\phi;\boldsymbol{x})$, on obtient,

$$\mathcal{L}_{ELBO}(\theta,\phi;\boldsymbol{x}) = \mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] - D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right] \tag{9}$$

$$D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z} \mid \boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right] = -\mathcal{L}_{ELBO}(\theta,\phi;\boldsymbol{x}) + \mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] \tag{10}$$

Ainsi, la maximisation de ELBO par rapport à ϕ est équivalente à la minimisation de la divergence de KL,

$$\min_{\phi}\{D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}|\boldsymbol{x})\right]\} = \min_{\phi}\{-\mathcal{L}_{ELBO}(\theta,\phi;\boldsymbol{x}) + \mathbb{E}_{q_\phi}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right]\}. \tag{11}$$

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 ?
Réponse :

$$q^* = \arg\max_{q_\phi \in \mathcal{Q}} \sum_{i=1}^{n} \mathcal{L}(\theta,\phi;\boldsymbol{x}_i) \tag{12}$$

$$= \arg\min_{q_\phi \in \mathcal{Q}} \sum_{i=1}^{n} D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x}_i)||p_\phi(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{13}$$

$$q^* = \arg\min_{q_\phi \in \mathcal{Q}} \sum_{i=1}^{n} \mathbb{E}_{q_\phi}\left[\log q_\phi(\boldsymbol{z}|\boldsymbol{x}_i) - \log p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{14}$$

$$q_i^* = \arg\max_{q_\phi \in \mathcal{Q}} \mathcal{L}(\theta,\phi;\boldsymbol{x}_i) \tag{15}$$

$$= \arg\min_{q_\phi \in \mathcal{Q}} D_{\mathrm{KL}}\left[q_i(\boldsymbol{z}|\boldsymbol{x}_i)||p_\phi(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{16}$$

$$q_i^* = \arg\min_{q_\phi \in \mathcal{Q}} \mathbb{E}_{q_i}\left[\log q_i(\boldsymbol{z}|\boldsymbol{x}_i) - \log p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{17}$$

Dans les deux cas $\log p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ ne change pas. Puisque $\log q_i(\boldsymbol{z}|\boldsymbol{x}_i)$ est un modèle ajusté pour une donnée $\boldsymbol{x}_i$ on peut s'attendre à ce que la différence $\log q_i(\boldsymbol{z}|\boldsymbol{x}_i) - \log p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$ soit plus petite que $\log q_\phi(\boldsymbol{z}|\boldsymbol{x}_i) - \log p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)$. Alors, $D_{\mathrm{KL}}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x}_i)||p_\phi(\boldsymbol{z}|\boldsymbol{x}_i)\right] \geq D_{\mathrm{KL}}\left[q_i(\boldsymbol{z}|\boldsymbol{x}_i)||p_\phi(\boldsymbol{z}|\boldsymbol{x}_i)\right]$.
Autrement,

$$\log p_\theta(\boldsymbol{x}_i) = \mathcal{L}(\theta,\phi;\boldsymbol{x}_i) + D_{\mathrm{KL}}\left[q^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{18}$$

Ainsi que,

$$\log p_\theta(\boldsymbol{x}_i) = \mathcal{L}(\theta,\phi;\boldsymbol{x}_i) + D_{\mathrm{KL}}\left[q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \tag{19}$$

Ces deux équation ensemble donnent,

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}_i) + D_{\mathrm{KL}}\left[q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] = \mathcal{L}(\theta, \phi; \boldsymbol{x}_i) + D_{\mathrm{KL}}\left[q^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \quad (20)$$

$$D_{\mathrm{KL}}\left[q^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] = D_{\mathrm{KL}}\left[q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] + (\mathcal{L}(q^*; \boldsymbol{x}_i) - \mathcal{L}(q_i^*; \boldsymbol{x}_i)) \quad (21)$$

$(\mathcal{L}(q^*; \boldsymbol{x}_i) - \mathcal{L}(q_i^*; \boldsymbol{x}_i)) > 0$, car $\mathcal{L}(q_i^*; \boldsymbol{x}_i)$ maximise q_i^* de $\mathcal{L}(q_i^*; \boldsymbol{x}_i)$ et donc plus petit que $\mathcal{L}(q^*; \boldsymbol{x}_i)$. Ainsi,

$$D_{\mathrm{KL}}\left[q^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right] \geq D_{\mathrm{KL}}\left[q_i^*(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i)\right]. \quad (22)$$

C.Q.F.D.

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 ?

Réponse : Pour l'estimation de la vraisemblance marginale via ELBO empirique, l'approche d'inférence variationnelle amortie est généralement meilleure que l'approche traditionnelle. On peut comprendre cela par le concept de compromis variance-biais, l'approche traditionnelle optimise une distribution pour chaque donnée et donc souffre de peu de biais, mais de plus de variance, alors que l'approche d'inférence variationnelle amortie est biaisée mais de faible variance puisque le modèle est optimisé sur l'ensemble des données.

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

Réponse : En termes d'efficacité de calcul, l'approche d'inférence variationnelle amortie est généralement meilleure que l'approche traditionnelle. On peut comprendre cela parce que dans l'approche traditionnelle, on optimise une distribution variationnelle distincte pour chaque point de données, ce qui peut être computationnellement coûteux avec beaucoup de données. À l'opposé, l'approche amortie ne nécessite d'optimiser qu'une seule distribution variationnelle pour l'ensemble des données, ce qui est beaucoup plus rapide.

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

Réponse : En termes d'utilisation de la mémoire, l'approche d'inférence variationnelle amortie est généralement meilleure que l'approche traditionnelle. On peut comprendre cela parce que pour l'approche traditionnelle, n ensembles de paramètres des distributions variationnelles de chaque point de donnée sont stockés. À l'opposé, pour l'approche amortie, on stocke les paramètres d'une seule distribution variationnelle (pour l'ensemble de données).

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 "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

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}|\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t I) \quad (23)$$

où

$$\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{24}$$

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

$$\alpha_t = 1-\beta_t$$

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

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

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I) \tag{25}$$

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

Réponse :

Avec la règle de Bayes et la propriété markovienne du processus encodeur,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}. \tag{26}$$

Qu'on réécrit avec des gaussiennes comme,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\mathcal{N}(\boldsymbol{x}_t|\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I)\mathcal{N}(\boldsymbol{x}_{t-1}|\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})I)}{\mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I)} \tag{27}$$

Sous la forme explicite de Gausienne,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \exp\left(-\frac{(\boldsymbol{x}_t - \sqrt{1-\beta_t}\boldsymbol{x}_{t-1})^2}{2\beta_t} - \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} + \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right) \tag{28}$$

$$= \exp\left(-\frac{1}{2}\left(-\frac{2\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}\boldsymbol{x}_t}{\beta_t} + \frac{(1-\beta_t)\boldsymbol{x}_{t-1}^2}{\beta_t} + \frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0\boldsymbol{x}_{t-1}}{1-\bar{\alpha}_{t-1}} + K(\boldsymbol{x}_t, \boldsymbol{x}_0)\right)\right) \tag{29}$$

où on a placé les termes en $\boldsymbol{x}_0 et \boldsymbol{x}_t$ dans une variables $K(\boldsymbol{x}_t, \boldsymbol{x}_0)$.

On factorise les termes en $\boldsymbol{x}_{t-1}$ et $\boldsymbol{x}_{t-1}^2$,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \exp\left(-\frac{1}{2}\left(\left[\frac{1-\beta_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right]\boldsymbol{x}_{t-1}^2 - 2\left[\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{\beta_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\boldsymbol{x}_{t-1} + K(\boldsymbol{x}_t, \boldsymbol{x}_0)\right)\right) \tag{30}$$

Le terme se simplifie à $\frac{1-\beta_t}{\beta_t} + \frac{1}{1-\bar{\alpha}_{t-1}} = \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_{t-1})\beta_t}$, on met en évidence ce terme pour obtenir,

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \exp\left(-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_{t-1})\beta_t}\right]\left(\boldsymbol{x}_{t-1}^2 - 2\left[\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}} + \frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{\beta_t}\right]\left[\frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}\right]\boldsymbol{x}_{t-1} + K(\boldsymbol{x}_t, \boldsymbol{x}_0\right.\right.\right. \tag{31}$$

$$= \exp\left(-\frac{1}{2}\left[\frac{1}{\frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}}\right]\left(\boldsymbol{x}_{t-1}^2 - 2\left[\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\boldsymbol{x}_0}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t}{1-\bar{\alpha}_t}\right]\boldsymbol{x}_{t-1} + K(\boldsymbol{x}_t, \boldsymbol{x}_0)\right)\right) \tag{32}$$

$$= \mathcal{N}\left(\boldsymbol{x}_{t-1} \mid \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t\boldsymbol{x}_0}{1-\bar{\alpha}_t} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t}{1-\bar{\alpha}_t}, \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}\right) \tag{33}$$

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1} \mid \tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t I) \tag{34}$$

où $\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$ et $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.
Ainsi, $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1} \mid \tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t I)$.
C.Q.F.D.

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.

Réponse :

Nous devons encore entraîner le processus inverse pour générer des images car le processus de débruitage du *ground truth* nécessite la connaissance de la donnée initiale $\boldsymbol{x}_0$ et qu'en pratique, l'accès à la donnée initiale $\boldsymbol{x}_0$ lors de la génération n'est pas disponible.

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 :

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

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

Réponse : En utilisant les propriétés des logs

$$\log p_\theta(\boldsymbol{x}_{0:T}) = \log \left[p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \right] = \log p(\boldsymbol{x}_T) + \sum_{t=1}^{T} \log p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

,

$$\log p(x_0) \geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right] \tag{35}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p_\theta(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] + \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\sum_{t=1}^{T} \log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}\right] \tag{36}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\sum_{t=1}^{T} \log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)} \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p_\theta(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] \tag{37}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \tag{38}$$

$$+ \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\sum_{t=2}^{T} \log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] + \mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\right] \tag{39}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \tag{40}$$

$$- \sum_{t=2}^{T} \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\right]\right] - D_{\mathrm{KL}}\left[q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)\right] \tag{41}$$

$$\mathcal{L}_{DDPM}(\theta;\boldsymbol{x}_0) = -L_0(\boldsymbol{x}_0) - \sum_{t=2}^{T} L_{t-1}(\boldsymbol{x}_0) - L_T(\boldsymbol{x}_0) \tag{42}$$

où à (37) on a utilisé $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}$.

À (38) sorti le premier terme $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]$. À (39) sorti le dernier terme $\log \frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}$. C.Q.F.D.

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

Réponse :

Le terme $L_T(\boldsymbol{x}_0)$, $L_T(\boldsymbol{x}_0)$ représente la divergence KL entre la distribution $q(\boldsymbol{x}_T|\boldsymbol{x}_0)$ et $p(\boldsymbol{x}_T)$, qui sont des distributions de bruit gaussien de moyenne 0 et de covariance 1. Ainsi, cette distributions ne dépend pas des paramètres appris θ, le terme $L_T(\boldsymbol{x}_0)$ agit comme une constante.

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) ?

Réponse : La ELBO du modèle VAE est composé des termes de reconstruction et de régularisation, alors que la ELBO du modèle DDPM est composé des termes de reconstruction et de *denoising matching*. Ainsi, la différence entre les ELBO des modèle VAE et DDPM sont les termes *denoising matching* et de régularisation.

La différence en termes de paramètres est que le VAE a deux ensembles de paramètres θ et ϕ pour les réseaux d'encodeur et de décodeur, alors que le DDPM n'a qu'un seul ensemble de paramètres θ pour le processus inverse.

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

- En utilisant l'éq. 24 et 25, montrer que $\tilde{\mu}_t(\boldsymbol{x}_t,\boldsymbol{x}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$, où $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0,I)$

  Réponse : Avec l'astuce de reparamétrisation, on peut réécrire,

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)I) \tag{43}$$

$$= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{44}$$

Et donc, $\boldsymbol{x}_t \sim \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, en réarrangent,

$$\boldsymbol{x}_0 = \frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}} \tag{45}$$

En remplaçant dans $\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t$,

$$\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\frac{(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon})}{\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \tag{46}$$

Avec $\bar{\alpha}_t = \alpha_t\bar{\alpha}_{t-1}$, et quelques manipulations on obtient,

$$\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \left(\frac{\beta_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}} + \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}\right)\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon} \tag{47}$$

$$= \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon} \tag{48}$$

$$\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) \tag{49}$$

C.Q.F.D.

- 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

$$\mathbb{E}_{q(\boldsymbol{x}_0)}L_{t-1}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\Big[\lambda_t||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\big)||^2\Big] + const \tag{50}$$

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.

Réponse : L'équation de la divergence KL entre des distributions normales multivariées est donnée par,

$$D_{\mathrm{KL}}\left[\mathcal{N}(x|\mu_1, \sigma_1)||\mathcal{N}(x|\mu_2, \sigma_2)\right] = \frac{1}{2\sigma_1^2}\left[||\mu_1 - \mu_2||_2^2\right] \tag{51}$$

On obtient pour $L_{t-1}(\boldsymbol{x}_0)$ avec (51),

$$L_{t-1}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left[q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\right]\right] \tag{52}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left[\mathcal{N}(\boldsymbol{x}_{t-1}|\tilde{\mu}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \sigma_t)||\mathcal{N}(\boldsymbol{x}_{t-1}|\mu_\theta(\boldsymbol{x}_t, t), \sigma_t)\right]\right] \tag{53}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\frac{1}{2\sigma_t^2}\left|\left|\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) - \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right)\right|\right|_2^2\right] \tag{54}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\frac{1}{\sqrt{\alpha_t}}\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||_2^2\right] \tag{55}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\lambda_t\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right] \tag{56}$$

où à (53) on a utilisé les définitions de $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ et $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ sous forme de gausienne, à (54) la définition de la divergence KL entre des distributions normales multivariées. Et à (55) quelques manipulations algébriques simples.
Enfin, avec $\boldsymbol{x}_t \sim \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$,

$$\mathbb{E}_{q(\boldsymbol{x}_0)} L_{t-1}(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\lambda_t ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||_2^2\right] \tag{57}$$

C.Q.F.D.

- Montrer que $\mathbb{E}_{q(\boldsymbol{x}_0)} L_0(\boldsymbol{x}_0)$ peut être écrit de la même manière que l'éq. 50
  Réponse : À $t = 1$, $L_{t-1} \to L_0$, ainsi,

$$\mathbb{E}_{q(\boldsymbol{x}_0)} L_0(\boldsymbol{x}_0) = \mathbb{E}_{q(\boldsymbol{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\lambda_1 ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_1}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_1}\boldsymbol{\epsilon}, 1)||_2^2\right] \tag{58}$$

  En effet, rien ne nous aurait empêcher de ne pas sortir le premier terme de $\mathcal{L}_{DDPM}$ au problème 2.3 (mais en considérant les intervalles des termes).
  C.Q.F.D.

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

$$\begin{aligned}
\mathcal{L}_{DDPM}(\theta; \boldsymbol{x}_0) &= -L_0(\boldsymbol{x}_0) - \sum_{t=2}^{T} L_{t-1}(\boldsymbol{x}_0) - L_T(\boldsymbol{x}_0) && (59)\\
&= \lambda_1 ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_1}\boldsymbol{x}_0 + \sqrt{1-\alpha_1}\boldsymbol{\epsilon}, 1)||_2^2 && (60)\\
&\quad - \sum_{t=2} \lambda_t ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||_2^2 && (61)\\
&\quad - \lambda_T ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_T}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_T}\boldsymbol{\epsilon}, T)||_2^2 && (62)\\
&= -\sum_{t=1} \lambda_t ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||_2^2 && (63)
\end{aligned}$$

En partique on entraine de manière répété pour un step aléatoirement selectioné dans l'intervalle 1 à T, ainsi,

$$\mathcal{L}_{DDPM}(\theta; \boldsymbol{x}_0)_t = -\lambda_t ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||_2^2 \tag{64}$$

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 $\mu = \frac{1}{2}(p_0 + p_1)$.
Réponse : Pour trouver la solution optimale D^*, on écrit

$$\begin{aligned}
G(D(x)) &= \mathbb{E}_{\boldsymbol{x}\sim p_1}[\log D(x)] + \mathbb{E}_{\boldsymbol{x}\sim p_0}[\log(1 - D(x))] && (65)\\
&= \int f_1(x) \log D(x) dx + \int f_0(x) \log(1 - D(x)) dx && (66)\\
G(D(x)) &= \int f_1(x) \log D(x) + f_0(x) \log(1 - D(x)) dx && (67)
\end{aligned}$$

L'intégrant est optimal lorsque la dérivé fonctionnelle $\frac{\delta G(D(x)))}{\delta D(x)} = 0$, en utilisant les propriétés des dérivés fonctionnelles $\frac{\delta}{\delta f(x)} \int g\left(f(x), x\right) dx = \frac{\partial}{\partial y} g(f(x), x)$, on obtient,

$$\frac{\delta G(D(x))}{\delta D(x)} = \frac{\delta}{\delta D(x)} \left[\int f_1(x) \log D(x) + f_0(x) \log(1 - D(x)) dx \right] \tag{68}$$

$$= \frac{f_1(x)}{D(x)} - \frac{f_0(x)}{1 - D(x)} \tag{69}$$

$$\tag{70}$$

À $\frac{\delta G(D(x)))}{\delta D(x)} = 0$, on trouve le discriminateur optimal D^*,

$$\frac{\delta G(D(x))}{\delta D(x)} = \frac{f_1(x)}{D(x)} - \frac{f_0(x)}{1 - D(x)} \tag{71}$$

$$0 = \frac{f_1(x)}{D^*} - \frac{f_0(x)}{1 - D^*} \tag{72}$$

$$\frac{f_1(x)}{D^*} = \frac{f_0(x)}{1 - D^*} \tag{73}$$

$$D^* = \frac{f_1(x)}{f_0(x) + f_1(x)} \tag{74}$$

On réarrange $\text{JSD}(p_0, p_1)$ pour obtenir des termes en $\frac{p_1}{p_0 + p_1}$,

$$\text{JSD}(p_0, p_1) = \frac{1}{2} KL(p_0 \| \mu) + \frac{1}{2} KL(p_1 \| \mu) \tag{75}$$

$$= \mathbb{E} \left[\frac{1}{2} \log \frac{2p_0}{p_0 + p_1} + \frac{1}{2} \log \frac{2p_1}{p_0 + p_1} \right] \tag{76}$$

$$= \mathbb{E}_{\boldsymbol{x} \sim p_0} \left[\log \frac{p_0}{p_0 + p_1} \right] + \mathbb{E}_{\boldsymbol{x} \sim p_1} \left[\log \frac{p_1}{p_0 + p_1} \right] + \log 2 \tag{77}$$

$$\text{JSD}(p_0, p_1) = \mathbb{E}_{\boldsymbol{x} \sim p_0} \left[\log \left(1 - \frac{p_1}{p_0 + p_1} \right) \right] + \mathbb{E}_{\boldsymbol{x} \sim p_1} \left[\log \frac{p_1}{p_0 + p_1} \right] + \log 2 \tag{78}$$

$$\tag{79}$$

Tirer N échantillons des distributions p_0 et $p_1 \implies$

$$\text{JSD}(p_0, p_1) = \frac{1}{2N} \sum_{i=1}^{N} \log \left(1 - \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})} \right) + \log \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})} + \log 2 \tag{80}$$

Ainsi, on obtient une expression permettant d'estimer la divergence JSD d'un discriminateur entraîné,

$$\text{JSD}(p_0, p_1) = \frac{1}{2N} \sum_{i=1}^{N} \log(1 - D^*) + \log D^* + \log 2 \tag{81}$$

C.Q.F.D.

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 $f_0(\boldsymbol{x})$. 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 $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1-D^*(\boldsymbol{x}))$.

Réponse : À partir de $D^* = \frac{f_1(x)}{f_0(x)+f_1(x)} \implies \frac{f_1(x)}{D^*} = \frac{f_0(x)}{1-D^*}$ (voir la question précédente) et après quelques manipulation, il suit l'identité

$$f_1(\boldsymbol{x}) = \frac{f_0(\boldsymbol{x})D^*(\boldsymbol{x})}{1-D^*(\boldsymbol{x})} \tag{82}$$

Et donc, on peut estimer $f_1(\boldsymbol{x})$ par

$$f_1(\boldsymbol{x}) = \frac{f_0(\boldsymbol{x})D(\boldsymbol{x})}{1-D(\boldsymbol{x})} \tag{83}$$

C.Q.F.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 $W_p(t)$ 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 :

$$J(W, W_p) = \frac{1}{2}\mathbb{E}_{x_1,x_2}\left[\|W_p\boldsymbol{f}_1 - \text{Stop-Grad}(\boldsymbol{f}_2)\|_2^2\right], \tag{84}$$

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

$$J(W, W_p) = \frac{1}{2}\left[tr(W_p^\intercal W_p F_1) - tr(W_p F_{12}) - tr(F_{12}W_p) + tr(F_2))\right], \tag{85}$$

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.

Réponse :

$$J(W, W_p) = \frac{1}{2}\mathbb{E}_{x_1,x_2}\|W_p\boldsymbol{f}_1 - \text{Stop-Grad}(\boldsymbol{f}_2)\|_2^2 \tag{86}$$
$$= \frac{1}{2}\mathbb{E}_{x_1,x_2}(W_p\boldsymbol{f}_1 - \boldsymbol{f}_2)^T(W_p\boldsymbol{f}_1 - \boldsymbol{f}_2) \tag{87}$$
$$= \frac{1}{2}tr\left[\mathbb{E}_{x_1,x_2}\left[\boldsymbol{f}_1^T W_p^T W_p\boldsymbol{f}_1 - \boldsymbol{f}_1^T W_p^T\boldsymbol{f}_2 - \boldsymbol{f}_2^T W_p\boldsymbol{f}_1 + \boldsymbol{f}_2^T\boldsymbol{f}_2\right]\right] \tag{88}$$
$$= \frac{1}{2}\left[\mathbb{E}\left[tr\boldsymbol{f}_1^T W_p^T W_p\boldsymbol{f}_1\right] - \mathbb{E}\left[tr\boldsymbol{f}_1^T W_p^T\boldsymbol{f}_2\right] - \mathbb{E}\left[tr\boldsymbol{f}_2^T W_p\boldsymbol{f}_1\right] + \mathbb{E}\left[tr\boldsymbol{f}_2^T\boldsymbol{f}_2\right]\right] \tag{89}$$
$$= \frac{1}{2}\left[\mathbb{E}\left[trW_p^T W_p\boldsymbol{f}_1\boldsymbol{f}_1^T\right] - \mathbb{E}\left[tr\boldsymbol{f}_2\boldsymbol{f}_1^T W_p^T\right] - \mathbb{E}\left[trW_p\boldsymbol{f}_1\boldsymbol{f}_2^T\right] + \mathbb{E}\left[tr\boldsymbol{f}_2\boldsymbol{f}_2^T\right]\right] \tag{90}$$
$$J(W, W_p) = \frac{1}{2}\left[trW_p^T W_p F_1 - trF_{12}W_p^T - trW_p F_{12} + trF_2\right] \tag{91}$$

où on a utilisé les propriété cyclique des trace des matrices à (90) et les définitions de F_1, F_2, F_{12}, F_{21} à l'étape (91).
C.Q.F.D.

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 $\dot{W}_p = -\frac{\partial J}{\partial W_p}$ (la dérivée de la fonction d'objectif par rapport aux paramètres W_p.)

Réponse :

$$\dot{W}_p = -\frac{\partial J}{\partial W_p} \tag{92}$$
$$\dot{W}_p = -\frac{\partial}{\partial W_p}\left[\frac{1}{2}\left[trW_p^T W_p F_1 - trF_{12}W_p^T - trW_p F_{12} + trF_2\right]\right] \tag{93}$$
$$= -\frac{1}{2}\left[2W_p F_1 - 2F_{12}^T\right] \tag{94}$$
$$\dot{W}_p = -W_p F_1 + F_{12}^T \tag{95}$$

C.Q.F.D.

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 $\dot{W}(t) = -\frac{\partial J}{\partial W(t)}$ 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, $\tilde{W}_p = (W_p - I_{n_2})$, et "vec(W)" font références à la *vectorisation* de la matrice W [1]. Par simplicité, nous ne considérons pas le weight decay ici [2].

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

2. 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.

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 $\boldsymbol{w}(t) \to 0$.

Réponse :

$$\dot{W} = -\frac{\partial J}{\partial W} \tag{96}$$

$$= -\frac{\partial}{\partial W}\left[\frac{1}{2}\left[trW_p^T W_p W(X+X')W^T - trF_W XW^T W_p^T - trW_p WXW^T + trW(X+X')W^T\right]\right] \tag{97}$$

$$= -W_p^T W_p W(X+X') + (W_p^T + W_p)WX - W(X+X') \tag{98}$$

$$= -(W_p^T W_p + I_{n_2})WX' - (W_p^T W_p - W_p^T - W_p^T - W_p + I_{n_2})WX \tag{99}$$

$$= -(W_p^T W_p + I_{n_2})WX' - (W_p - I_{n_2})^T(W_p - I_{n_2})WX \tag{100}$$

$$\dot{W} = -(W_p^T W_p + I_{n_2})WX' - \tilde{W}_p^T\tilde{W}_p WX \tag{101}$$

$$\tag{102}$$

La définition du *vec trick* est donnée par,

$$\text{vec}(AXB) = (B^T \otimes A)\,\text{vec}(X). \tag{103}$$

Alors,

$$\frac{\mathrm{d}}{\mathrm{d}t}\text{vec}(W) = -\left[X' \otimes \left(W_p^T W_p + I_{n_2}\right) + X \otimes \tilde{W}_p^T\tilde{W}_p + \eta I_{n_1 n_2}\right]\text{vec}(W) \tag{104}$$

Enfin, si la valeur propre minimal $\lambda_{min}(H(t)) = \lambda_0$ est bornée loin de zéro, alors $\frac{d}{dt}\text{vec}(W(t)) = -H(t)\text{vec}(W(t))$ satisfait la contrainte suivante,

$$\|\text{vec}(W(t))\|_2 = e^{-\lambda_0 t}\|\text{vec}(W(0))\|_2, \tag{105}$$

impliquant que $\text{vec}(W(t)) \to 0$.

C.Q.F.D.

Notons que dans ce cas, sans stop-grad, il est plus impossible d'apprendre les features puisque $W(t) \to 0$.

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.

Réponse : Si le prédicteur est enlevés alors $W_p = I$. En partant de

$$\dot{W}(t) = -H(t)\text{vec}(W) = \left[X' \otimes (W_p^T W_p + I) + X \otimes (\tilde{W}_p^T\tilde{W}_p)\right]\text{vec}(W) \tag{106}$$

On obtient en remplaçant $W_p = I$:

$$\dot{W}(t) = \left[X' \otimes (I^T I + I) + X \otimes ((I - I)^T (I - I))\right] \text{vec}(W) \tag{107}$$

$$\dot{W}(t) = [X' \otimes I] \, \text{vec}(W) \tag{108}$$

Puisque X' est definite positive, ses valeurs propre positive et bornée loin de zéro. Avec un argument similaire à ce qui est fait à la question précédente on obtient que $W(t) \to 0$.

C.Q.F.D.

Notons que dans ce cas, sans stop-grad et sans prédicteur, il est plus difficile d'apprendre les features puisque $W(t) \to 0$ et sans prédicateur la représentation est triviale.

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

Réponse : Comme mentionner aux questions précédente $W(t) \to 0$, ce qui rend difficile l'apprentissage des features. Stop-gradient permet de découpler $\boldsymbol{f}_1$ et $\boldsymbol{f}_2$, tous deux fonction des mêmes paramètres, stabilisant la fonction d'objectif.