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

Question 2. (**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 $\tilde{\boldsymbol{z}}$.

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.

Question 3 (5). (**Mélanger 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 ?

Question 4 (6-2-6-6)**.** (**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.

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.

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:

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

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.

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 ?

4. Pour trouver les points stationnaires, fixez $v = 0$ et résolvez les équations pour chaque paramètre.