Due Date: April 5th 23:59, 2019

Instructions

- *Montrez votre démarche pour toutes les questions !*
- *Les questions marquées avec une * sont **difficiles**, et non **bonus**.*
- *Utilisez un logiciel de traitement de texte comme LaTeX, sauf indication contraire.*
- *Sauf indication contraire, supposez que les notations et définitions pour chaque question sont indépendantes.*
- *Vous devez soumettre toutes vos réponses sur la page Gradescope du cours.*
- ***Les TAs pour ce devoir sont Shawn Tan, Samuel Lavoie, et Chin-Wei Huang.***

Ce travail couvre les techniques mathématiques et algorithmiques de trois modèles génératifs très en vogue ; soit les *variational autoencoders* (VAEs – Question 1-3), les modèles autoregressifs (Question 4) et les modèles génératifs adversariels (GANs – Question 5-7).

Question 1 (8-8)**.** La technique de reparamétrisation est une technique standard qui rend les échantillons d'une variable aléatoire différentiables. Considérez un vecteur aléatoire $Z \in \mathbb{R}^K$ avec une fonction de densité $q(\boldsymbol{z}; \phi)$. Nous voulons trouver une fonction déterministe $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$ qui dépend de ϕ, pour transformer la variable aléatoire Z_0 ayant une fonction de densité $q(\boldsymbol{z}_0)$ indépendante de ϕ, telle que $\boldsymbol{g}(Z_0)$ a la même densité que Z. Rappelez vous du changement de densité pour une fonction bijective et différentiable $\boldsymbol{g}$:

$$q(\boldsymbol{g}(\boldsymbol{z}_0)) = q(\boldsymbol{z}_0) \left| \det \left(\frac{\partial \boldsymbol{g}(\boldsymbol{z}_0)}{\partial \boldsymbol{z}_0} \right) \right|^{-1} \tag{1}$$

1. Supposez $q(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ et $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \sigma \odot \boldsymbol{z}_0$, où $\mu \in \mathbb{R}^K$ et $\sigma \in \mathbb{R}^K_{>0}$. Montrez que $\boldsymbol{g}(\boldsymbol{z}_0)$ suit la distribution $\mathcal{N}(\mu, \text{diag}(\sigma^2))$ en utilisant l'Equation (1).
2. Supposez plutôt que $\boldsymbol{g}(\boldsymbol{z}_0) = \mu + \boldsymbol{S}\boldsymbol{z}_0$, où $\boldsymbol{S}$ est une matrice non-singulière $K \times K$. Derivez la densité de $\boldsymbol{g}(\boldsymbol{z}_0)$ en utilisant l'Equation (1).

Answer 1.

1.

Question 2 (5-5-6)**.** Considerez un modèle à variables latentes $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_K)$ où $\boldsymbol{z} \in \mathbb{R}^K$ et $\boldsymbol{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{z})$. Le réseau encodeur (aka "recognition model") du variational autoencoder, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, est utilisé pour produire une approximation (variationelle) de la distribution à posteriori sur les variables latentes $\boldsymbol{z}$ pour n'importe quel point de donnée $\boldsymbol{x}$.[1] Cette distribution est entraînée pour correspondre à la vraie postérieure en maximisant le *evidence lower bound* (ELBO):

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

Soit $\mathcal{Q}$ la famille de distributions variationelles avec un ensemble de paramètres réalisables $\mathcal{P}$; i.e. $\mathcal{Q} = \{q(\boldsymbol{z}; \pi) : \pi \in \mathcal{P}\}$; par example π peut être la moyenne et l'écart type d'une distribution normale. Nous supposons que q_ϕ est paramétrisée par un réseau de neurones (avec paramètres ϕ) dont la sortie sont les paramètres, $\pi_\phi(\boldsymbol{x})$, de la distribution $q \in \mathcal{Q}$, i.e. $q_\phi(\boldsymbol{z}|\boldsymbol{x}) := q(\boldsymbol{z}; \pi_\phi(\boldsymbol{x}))$.

1. Utiliser un *recognition model* est aussi appelé "amortized inference" ; Ceci fait contraste avec les approches traditionelles d'inférence variationelles (voir p.ex., Chapitre 10 de Bishop's *Pattern Recognition an Machine Learning*), qui ajuste la postérieure variationelle indépendamment de chaque nouveau point de donné.

1. Montrez que maximiser l'espérance complète de la log-likelihood des données

$$\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$$

pour un $q(\boldsymbol{z}|\boldsymbol{x})$ fixe, par rapport aux paramètres du modèle θ donne le maximiseur du log likelihood marginal biaisé: $\arg\max_\theta\{\log p_\theta(\boldsymbol{x}) + B(\theta)\}$, où $B(\theta)$ est non-positif. Trouvez $B(\theta)$.

2. Considérez un ensemble d'entraînement fini $\{\boldsymbol{x}_i : i \in \{1, ..., n\}\}$, n étant la taille des données d'entraînement. Soit ϕ^* le maximiseur de $\sum_{i=1}^n \mathcal{L}(\theta, \phi; \boldsymbol{x}_i)$ avec θ fixé. En plus, pour chaque $\boldsymbol{x}_i$, soit $q_i \in \mathcal{Q}$ une distribution variationelle dépendant de l'instance, et dénotez par q_i^* le maximiseur du ELBO correspondant. Comparez $D_{\mathrm{KL}}(q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x}_i)||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$ et $D_{\mathrm{KL}}(q_i^*(\boldsymbol{z})||p_\theta(\boldsymbol{z}|\boldsymbol{x}_i))$. Le quel est le plus grand ?

3. Suivant la question précédente, comparez les deux approches de la deuxième sous-question (question 2.2)
   (a) En termes de biais de l'estimation de la likelihood marginale via le ELBO, dans le meilleur cas (i.e. quand les deux approches sont optimales au sein de leur propres familles)
   (b) D'un point de vue d'efficacité de calcul
   (c) En termes de mémoire (espace de rangement des paramètres)

Answer 2.

1.

Question 3 (6-6)**.** Puisque l'inférence variationelle nous donne une borne inférieure sur la log-likelihood marginale des données, cela nous donne une estimée biaisée de la likelihood marginale. Ainsi, les méthodes pour rendre la borne plus serrée (i.e. trouver une borne inférieure plus grande) peuvent être désirables.

Considérez un modèle à variables latentes avec la distribution jointe $p(\boldsymbol{x}, \boldsymbol{h})$ où $\boldsymbol{x}$ et $\boldsymbol{h}$ sont les variables aléatoires observées et non-observées respectivement. Soit $q(\boldsymbol{h})$ une approximation variationelle de $p(\boldsymbol{h}|\boldsymbol{x})$. Nous définissons

$$\mathcal{L}_K = \mathbb{E}_{\boldsymbol{h}_j \sim q(\boldsymbol{h})}\left[\log \frac{1}{K}\sum_{j=1}^K \frac{p(\boldsymbol{x}, \boldsymbol{h}_j)}{q(\boldsymbol{h}_j)}\right]$$

Notez que $\mathcal{L}_1$ est équivalent au evidence lower bound (ELBO).

1. Montrez que $\mathcal{L}_K$ est une borne inférieure de la log-likelihood marginale $\log p(\boldsymbol{x})$.
2. Montrez que $\mathcal{L}_K \geq \mathcal{L}_1$; i.e. $\mathcal{L}_K$ est une famille de bornes inférieures plus serrée que le ELBO.

Answer 3.

1.

Question 4 (5-5-5-5)**.** Une façon de forcer un conditionnement autoregressif est d'appliquer un masque sur les paramètres de poids. [2] Considérez un réseau de neurones convolutionnel à deux couches sans retournement de noyau (kernel flipping), avec une taille de noyau de 3×3 et taille du

2. Un exemple de ceci est l'utilisation d'un masque dans l'architecture du transformeur (Problème 3 du TP2 partie pratique).

padding de 1 sur chaque côté (de telle sorte qu'un feature map en entrée de taille 5×5 demeure de taille 5×5 après la convolution). Définissez le masque de type A et le masque de type B comme

$$(\boldsymbol{M}^A)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j < 2 \\ 0 & \text{ailleurs} \end{cases} \qquad (\boldsymbol{M}^B)_{::ij} := \begin{cases} 1 & \text{if } i < 2 \\ 1 & \text{if } i = 2 \text{ and } j \leq 2 \\ 0 & \text{ailleurs} \end{cases}$$

où les indexes commencent à 1. Le masque est obtenue en multipliant par element le noyaux avec le masque binaire (elementwise multiplication). Spécifiez le *receptive field* du pixel à la sortie à la troisième ligne et troisième colonne (indexe 33 de la Figure 1) dans chacun des 4 cas suivants:

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 1 – 5×5 convolutional feature map.

1. Si nous utilisons $\boldsymbol{M}^A$ à la première couche et $\boldsymbol{M}^A$ à la deuxième couche.
2. Si nous utilisons $\boldsymbol{M}^A$ à la première couche et $\boldsymbol{M}^B$ à la deuxième couche.
3. Si nous utilisons $\boldsymbol{M}^B$ à la première couche et $\boldsymbol{M}^A$ à la deuxième couche.
4. Si nous utilisons $\boldsymbol{M}^B$ à la première couche et $\boldsymbol{M}^B$ à la seconde couche.

Answer 4. See Figure 2 for an example answer

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

11	12	13	14	15
21	22	23	24	25
31	32	33	34	35
41	42	43	44	45
51	52	53	54	55

FIGURE 2 – Receptive field under different masking schemes.

Question 5 (10)**.** Soit P_1 et P_0 deux distributions de probabilité de densités f_0 et f_1 respectivement. Ce problème démontre que le discriminateur optimal d'un GAN (i.e. qui est capable de distinguer les examples venant de P_0 et P_1 avec NLL minimale) peut être utilisé pour exprimer la probabilité de densité d'un point de donnée $\boldsymbol{x}$ sous f_1, $f_1(\boldsymbol{x})$ en terme de $f_0(\boldsymbol{x})$.

Supposons que f_0 et f_1 ont le même support. Montrez que $f_1(\boldsymbol{x})$ peut être estimée par $f_0(\boldsymbol{x})D(\boldsymbol{x})/(1 - D(\boldsymbol{x}))$ en établissant l'identité $f_1(\boldsymbol{x}) = f_0(\boldsymbol{x})D^*(\boldsymbol{x})/(1 - D^*(\boldsymbol{x}))$, où

$$D^* := \arg\max_D \mathbb{E}_{\boldsymbol{x}\sim P_1}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim P_0}[\log(1 - D(\boldsymbol{x}))]$$

Answer 5.

Question 6 (5-5-6)**.** Alors que les GANs ont originalement été formulés comme minimisant la Jensen-Shannon (JS)-divergence, ce cadre peut être généralisé pour utiliser d'autres divergences, telles que la Kullback-Leibler (KL)-divergence. Dans cet exercices, nous observons comment la KL peut être approximé (bornée inférieurement) via la fonction $T : \mathcal{X} \to \mathbb{R}$ (i.e. le discriminateur). Soient q et p des fonctions de densité de probabilité. On rappelle la définition de la KL divergence $D_{\mathrm{KL}}(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$.

*1. Soit $R_1[T] := \mathbb{E}_p[T(x)] - E_q[e^{T(x)-1}]$.

(a) La conjuguée de la fonction $f(u)$ est défini comme $f^*(t) = \sup_{u \in \mathrm{dom} f} ut - f(u)$. Montrez que la conjuguée de $f(u) = u \log u$ is $f^*(t) = e^{t-1}$, et sa biconjuguée [3], i.e. La conjuguée de sa conjuguée est $f^{**}(u) := (f^*)^*(u) = u \log u$.

(b) Utilisez le fait trouvé à la sous-question précédente et montrez que $D_{\mathrm{KL}}(p||q) = \sup_T R_1[T]$, où le supremum est pris sur l'ensemble de toutes les fonction mesurables $\mathcal{X} \to \mathbb{R}$. Commencez à l'aide de l'étape suivante

$$\sup_{T(x)} \int p(x)T(x) - q(x)e^{T(x)-1}dx = \int \sup_{t \in \mathbb{R}} p(x)t - q(x)e^{t-1}dx$$

que vous n'avez pas à prouver.

*2. Soit $r(x) = e^{T(x)}/\mathbb{E}_q[e^{T(x)}]$ et $R_2[T] := \mathbb{E}_p[T(x)] - \log \mathbb{E}_q[e^{T(x)}]$.

(a) Vérifiez que rq est une fonction de densité de probabilité, i.e. d'intégrale 1.

(b) Montrez que $D_{\mathrm{KL}}(p||q) \geq R_2[T]$, avec égalité si et seulement si $T(x) = \log(p(x)/q(x)) + c$ où c est une constante indépendante de x.

3. Comparez les deux représentations de la KL divergence. Pour $T(x)$, $p(x)$ et $q(x)$ fixés, lequel de $R_1[T]$ ou $R_2[T]$ est plus grand ou égal à l'autre ?

Answer 6.

1.

Question 7 (10)**.** Soit $q, p : \mathcal{X} \to [0, \infty)$ des fonction de probabilité avec support disjoint (i.e. sans chevauchement) ; plus formellement, $\{x \in \mathcal{X} : p(x) > 0 \text{ and } q(x) > 0\} = \varnothing$. Quel est le Jensen Shannon Divergence (JSD) entre p et q ? Le JSD est défini comme $D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||r) + \frac{1}{2}D_{KL}(q||r)$ où $r(x) = \dfrac{p(x) + q(x)}{2}$.

Answer 7.

3. Plus généralement, la biconjuguée de f est égale a elle-même si f est convexe et semi-continue inférieurement (ceci est aussi connu comme le **Fenchel-Monreau Theorem**).