Due Date : February 22nd, 2023

Instructions

- *Montrez vos traces pour toutes les questions !*
- *Utilisez le modèle LaTeX que nous vous fournissons pour écrire vos réponses. Vous pouvez réutiliser les racourcis pour la notation, les équations et/ou les tables. Voir la politique sur les travaux pratique sur le site du cours pour plus de détails.*
- *Remettez vos questions électroniquement par Gradescope.*
- *TAs pour ce devoir sont :* ***Andjela Mladenovic*** *(Pour IFT6135B) et* ***Ghait Boukachab*** *(Pour IFT6135A).*

1. **Choix de la fonction d'activation (10 pts)** Nous allons comparer deux fonctions d'activation différentes dans cette question. Rappelons la définition de $\sigma(x) = \frac{1}{1+e^{-x}}$ et $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
   (a) (2 pts) Déterminer la dérivée de la fonction sigmoïde $\sigma'(x)$. Exprimer la dérivée en fonction de $\sigma(x)$.
   Réponse :

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}(1+e^{-x})^{-1} \tag{1}$$
$$= (1+e^{-x})^{-2}(1+e^{-x}-1) \tag{2}$$
$$\frac{d}{dx}\sigma(x) = (1-\sigma(x))\sigma(x) \tag{3}$$

   (b) (2 pts) Déterminer la dérivée de la fonction $\tanh'(x)$ et la présenter en fonction de $\tanh(x)$.
   Réponse :

$$\frac{d}{dx}\tanh(x) = \frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{4}$$
$$= \frac{e^x + e^{-x}}{e^x + e^{-x}} - \frac{(e^x - e^{-x})^2(e^x - e^{-x})}{(e^x + e^{-x})^2(e^x - e^{-x})} \tag{5}$$
$$\frac{d}{dx}\tanh(x) = 1 - \tanh(x)^2 \tag{6}$$

   (c) (2 pts) Borner (borne supérieure) la valeur de $\sigma'(x)$ avec une constante (vous pouvez utiliser l'inégalité AM-GM).
   Réponse : Avec $x_1 x_2 = (1-\sigma(x))\sigma(x)$ et $x_1 + x_2 = (1-\sigma(x)) + \sigma(x)$,

$$\sqrt{(1-\sigma(x))\sigma(x)} \leq \frac{(1-\sigma(x)) + \sigma(x)}{2} \tag{7}$$
$$\leq \frac{1}{2} \tag{8}$$
$$\frac{d}{dx}\sigma(x) \leq \frac{1}{4} \tag{9}$$

   (d) (2 pts) Borner (borne supérieure) la valeur de $\tanh'(x)$ avec une constante (vous pouvez utiliser l'inégalité GM-HM ou la propriété selon laquelle le carré d'un nombre réel est toujours non négatif).

Réponse :
Le domaine de $\tanh(x)$ prend des valeurs de minimale et maximale de -1 à 1 en passant par zéro. Alors $1 \geq \tanh(x)^2 \geq 0$. Ce qui implique que $1 - \tanh(x)^2$ est maximal lorsque $\tanh(x)^2$ est nul.

$$1 - \tanh(x)^2 \leq 1 - 0 \tag{10}$$
$$1 - \tanh(x)^2 \leq 1 \tag{11}$$
$$\frac{d}{dx}\tanh(x) \leq 1 \tag{12}$$

(e) (2 pts) Comparez les deux limites supérieures et expliquez quel serait l'impact de leur différence sur l'optimisation.
Réponse : La valeur de la limite supérieure $\tanh'(x)$ est plus élevé que pour $\sigma'(x)$, ainsi le gradient peux se propager plus rapidement i.e. il est possible d'obtenir de plus grande correction des paramètres lors de l'optimisation avec une fonction d'activation $\tanh(x)$.

Inégalités utiles pour cette question:

Inégalité des moyennes arithmétiques et géométriques (AM-GM)

$$\frac{x_1 + x_2 + \ldots x_n}{n} \geq \sqrt[n]{x_1 x_2 \ldots x_n} \tag{13}$$

Inégalité des moyennes géométriques et harmoniques (GM-HM)

$$\sqrt[n]{x_1 x_2 \ldots x_n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_n}} \tag{14}$$

Les inégalités ci-dessus sont valables pour tout nombre positif réel $x_1, x_2, \cdots x_n$ avec égalité si et seulement si $x_1 = x_2 = \cdots = x_n$.

2. **Propriétés d'entropie croisée (12 pts)**

La fonction de perte d'entropie croisée (une fonction de perte populaire) est donnée par:

$$\text{CE}(p, x) = -x \log(p) - (1 - x) \log(1 - p)$$

.

Veuillez vous référer à cette perte pour les parties (a) et (b).

(a) (2 pts) **Entropie croisée et maximum de vraisemblance** Pour cette dérivation, nous supposons que x est binaire, c'est-à-dire que $x \in \{0, 1\}$. Déterminez la fonction de coût de l'entropie croisée en utilisant le principe du maximum de vraisemblance pour $x \in \{0, 1\}$.
Réponse : La vraisemblance est donné par,

$$\mathcal{L}(\theta|x) = f(x|\theta) \tag{15}$$

ou bien,

$$\log \mathcal{L}(\theta|x) = \log f(x|\theta) \tag{16}$$

Le maximum de vraisemblance se trouve à

$$\frac{d\mathcal{L}(\theta|x)}{d\theta} = 0 \tag{17}$$

La distribution de Bernoulli est donné par,

$$P(X = x) = p^x(1-p)^{1-x} \tag{18}$$

Pour maximisé la vraisemblance, on minimise le négatif de la fonction de vraisemblance,

$$-\log \mathcal{L}(\theta|x) = -\log p^x(1-p)^{1-x} \tag{19}$$

$$\mathrm{CE}(p, x) = -x \log p - (1-x)\log(1-p) \tag{20}$$

(b) (2 pts) **Entropie croisée et divergence KL** Suggérer une interprétation probabiliste de la fonction de coût de l'entropie croisée lorsque $x \in (0, 1)$. (Hint: divergence de KL entre deux distributions)

Réponse : La fonction de coût de l'entropie croisée donne la distance entre deux distribution de données i.e. diminuer la distance entre la distribution des données de l'ensemble d'entraînement et la distribution donné par le modèle sur ces données.

(c) (4 pts) **Distribution discrète - Entropie maximale** Soit X une variable aléatoire qui prend n valeurs avec des probabilités $p_1, p_2, \ldots, p_n$ avec $p_i > 0, \forall i$. Quelle est la distribution qui maximise l'entropie $H(X) = -\sum_{i=1}^{n} p_i \log p_i$?. Déterminez la limite supérieure de l'entropie $H(X)$ exprimée en fonction de n. (Indice : utilisez l'inégalité de Jensen).

Réponse : Avec l'inégalité de Jensen :

$$\sum p_i f(x_i) \leq f\left(\sum p_i x_i\right) \tag{21}$$

On trouve,

$$H(X) = -\sum p_i \log p_i \tag{22}$$

$$= \sum p_i \log \frac{1}{p_i} \tag{23}$$

$$\leq \log\left(\sum p_i \frac{1}{p_i}\right) \tag{24}$$

$$H(X) \leq \log n \tag{25}$$

$H(X) \leq \log n$ implique une distribution uniforme avec chacune des classes de probabilité,

$$p_i = \frac{1}{n} \tag{26}$$

(d) (4 pts) **Distribution continue (moyenne μ et variance σ^2 connues) - Entropie maximale** Étant donné la moyenne μ et la variance σ^2, quelle est la distribution continue qui maximise l'entropie différentielle $h(X) = -\int_x f(x) \log f(x) dx$? Démontrez le.

Réponse : Avec les multiplicateurs de Lagrange,

$$\mathscr{L} = -\int f(x) \log f(x) dx - \lambda_0 \left(\int f(x) dx - 1\right) - \lambda_1 \left(\int (x-\mu)^2 f(x) dx - \sigma^2\right) \tag{27}$$

$$\frac{d\mathscr{L}}{df(x)} = 0 \rightarrow \log f(x) - 1 + \lambda_0 + \lambda_1(x-\mu)^2 = 0 \tag{28}$$

On trouve,

$$f(x) = e^{1-\lambda_0-\lambda_1(x-\mu)^2} \tag{29}$$

Et avec,

$$\int (x-\mu)^2 f(x)dx - \sigma^2 = 0 \;\text{ et }\; \int f(x)dx - 1 = 0 \tag{30}$$

On trouve avec la première équation,

$$e^{\lambda_0 - 1} = \sqrt{\frac{\pi}{\lambda_1}} \tag{31}$$

Et en l'insérant dans la deuxième,

$$\sqrt{\frac{\pi}{\lambda_1}} = 2\lambda_1\sigma^2 e^{\lambda_0-1} \tag{32}$$

Enfin, on trouve

$$\lambda_0 = \log\sqrt{2\sigma^2\pi} + 1 \tag{33}$$

$$\lambda_1 = \frac{1}{2\sigma^2} \tag{34}$$

Tout assemblé, on trouve une distribution normale,

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{35}$$

3. **Dimension de sortie et paramètres des couches de convolution (5 pts)**
Considérons un réseau de neurones de convolution à 3 couches cachées. Supposons que l'entrée est une image couleur d'une taille de 128×128 suivant la représentation RGB. La première couche convole 64 noyaux de 8×8 avec l'entrée, en utilisant un stride de 2 et un zero-padding de 4. La deuxième couche déséchantillonne la sortie de la première couche avec un pooling max non chevauchant de 2×2. La troisième couche convolue 128 noyaux de 4×4 avec un stride de 2 et un zero-padding de 2.

   (a) (3 pts) Quelle est la dimension de la sortie de la troisième couche ?

   Réponse : La première couche sort 64@65x65, la deuxième 64@32x32 (en coupant la dernière ligne/colonne), la troisième 128@17x17 (le résultat ne change pas selon si on coupe la dernière ligne/colonne ou ajoute une ligne/colonne de 0).

   (b) (2 pts) Sans compter les biais, combien de paramètres sont requis pour la dernière couche ?

   Réponse : Le nombre de paramètre dépend de la taille du kernel (de la dernière couche) et du nombre de filtre de la couche précédente et actuelle i.e. le nombre de paramètre est de 4x4x64x128=131 072

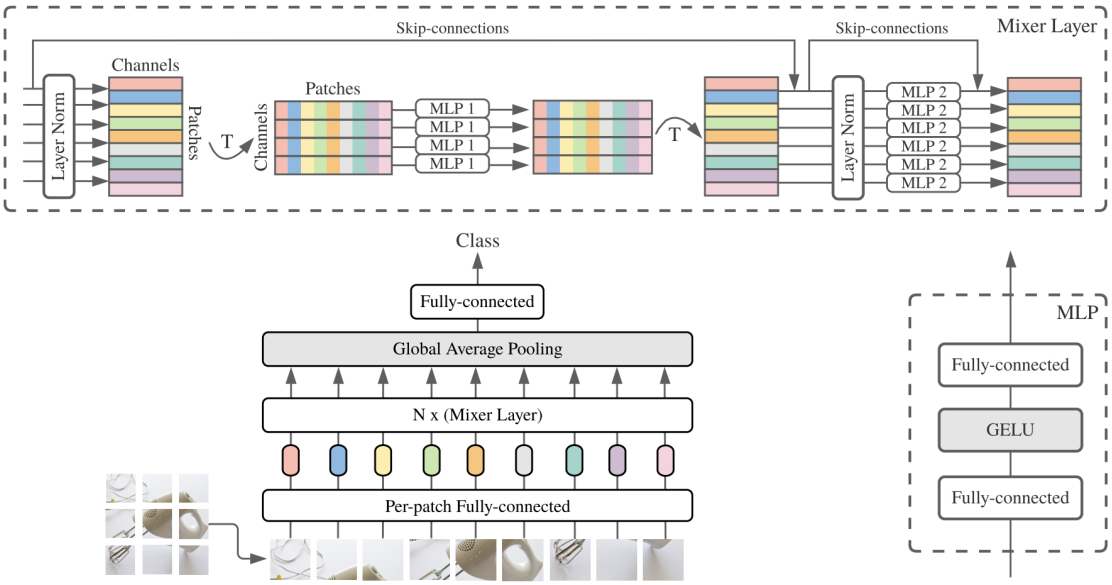

FIGURE 1 – Reproduite de l'article de MLPMixer. MLP-Mixer se compose des embeddings linéaires par patch, des couches Mixer et d'une tête de classification. Les couches de mixage contiennent un MLP de mixage de tokens et un MLP de mixage de chaînes, chacun étant composé de deux couches entièrement connectées et d'une non-linéarité GELU. couches entièrement connectées et d'une non-linéarité GELU. Les autres composants comprennent : des connexions de saut, le dropout, et la norme de couche sur les chaînes.

4. **MLP Mixer (16 pts)**

   (a) (2 pts) **Mixer Dimensions** Supposons que l'architecture du Mixer est utilisée pour une image d'entrée de taille 64×64. La sortie du mixeur est de taille 16×128. Déterminez la résolution de patch P, le nombre de patchs S, ainsi que la dimension cachée C(chaînes).
   Réponse : $S = I \times I/P^2 = 16$; $P = \sqrt{64 \times 64/16} = 16$; $C = 128$;

   (b) (2 pts) **Complexité du mixeur MLP** Montrez que la complexité de calcul du mixeur MLP est linéaire en fonction du nombre de patches d'entrée.
   Réponse : Chaque étape implique des opérations appliqué indépendamment sur chaque patch (ou channel) i.e. les opérations matriciel appliqué aux patch sont des transposé et des projections (tous des opérations linéaire en nombre de patch) et elle passent ensuite chacune indépendamment à travers les mêmes MLP (linéaire en nombre de patch) (MLP1 et MLP2). En augmentant le nombre de patch le nombre d'opération augmente linéairement.

   (c) (6 pts) **Transformation des entrées - Mixage des chaînes MLP** Considérons le scénario suivant : L'image d'entrée originale A a une taille de 9×9. Nous convertissons l'image d'entrée en patchs non chevauchants de taille 3×3, puis nous projetons linéairement tous les patchs avec la même matrice de projection. Le résultat de ces opérations est une matrice X de taille 9×6. Ensuite, nous appliquons le MLP de mixage de chaînes qui agit sur les lignes de X, et qui est partagé entre toutes les lignes. Le résultat de cette opération est la matrice U de taille 9×6. Considérons maintenant une image modifiée A

telle que $A_{\text{modified}} = PA$, où nous définissons la matrice P de la manière suivante:

$$P = \begin{bmatrix} e_{\pi(1)} \\ e_{\pi(2)} \\ \vdots \\ e_{\pi(9)} \end{bmatrix} \tag{36}$$

Ici e_k est le k$-ime$ vecteur de base et π représente la permutation des indices de $1 \dots 9$. Trouver tous les P possibles tels qu'en permutant les lignes de U_{modified} on puisse récupérer la matrice U.

Réponse :

On applique l'opérateur de projection linéaire $\hat{L}$ sur A, $\hat{L}A = X$.

Puis l'opérateur de mixing $\hat{M}$ sur X, $\hat{M}X = U$.

On cherche les matrices de permutation P tel qu'après avoir appliqué $PA = A_{\text{modified}}$, $PU_{\text{modified}} = U$.

Ce sont les matrices de permutation $P^2 = I$. Puisque les matrices de permutation sont orthogonale $P^{-1} = P^{\text{T}}$. On trouve alors les matrices symétrique soit des matrice tel que $P^T = P$.

(d) (6 pts) Sélectionnez une de vos solutions pour P et trouvez $P_{reverse}$ telle que $P_{\text{reverse}} U_{\text{modified}} = U$.

$$P = P_{\text{reverse}} = P^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

5. **Convergence par descente de gradient (12 pts)**

(a) (6 pts) **Convex Function Convergence** Considérons la fonction suivante : :

$$f(x) = \begin{cases} \frac{3}{4}(1-x)^2 - 2(1-x) & \text{if } x > 1 \\ \frac{3}{4}(1+x)^2 - 2(1+x) & \text{if } x < -1 \\ x^2 - 1 & \text{otherwise} \end{cases} \tag{37}$$

Montrez que f est une fonction convexe. Trouvez son unique minimiseur et son gradient. Considérons l'algorithme suivant : $x_t = x_{t-1} - \eta f'(x_{t-1})$ où $\eta = 1$. Cet algorithme convergera-t-il vers un point stationnaire s'il commence au point x_0, où $x_0 > 1$? Pourquoi ou pourquoi pas ?

Réponse : Il suffit de montrer que $\frac{d^2}{dx^2} f(x) > 0, \forall x$, la dérivé seconde est donné par,

$$\frac{d^2}{dx^2} f(x) = \begin{cases} \frac{3}{2} & \text{if } x > 1 \\ \frac{3}{2} & \text{if } x < -1 \\ 2 & \text{otherwise} \end{cases} \tag{38}$$

En effet $\frac{d^2}{dx^2}f(x) > 0, \forall x$, ainsi, la fonction est convexe.
Le minimum de la fonction se trouve à $\frac{d}{dx}f(x) = 0$,

$$\frac{d}{dx}f(x) = \begin{cases} -\frac{3}{2}(1-x)+2 & \text{if } x > 1 \\ \frac{3}{2}(1+x)^2 - 2 & \text{if } x < -1 \\ 2x & \text{otherwise} \end{cases} \tag{39}$$

$\frac{d}{dx}f(x) = 0$ lorsque $x = 0$ i.e. son unique minimiseur est $x = 0$.
L'algorithme $x_t = x_{t-1} - f'(x_{t-1})$ converge pour $x_0 > 1$. En calculant la limite à $x = 1$ de $\frac{d}{dx}f(x)$ on trouve,

$$\lim_{x\to 1}\frac{d}{dx}f(x) = \begin{cases} \lim_{x\to 1^+}\frac{d}{dx}f(x) = \lim_{x\to 1^+} -\frac{3}{2}(1-x)+2 & = 2 \\ \lim_{x\to 1^-}\frac{d}{dx}f(x) = \lim_{x\to 1^-} 2x & = 2 \end{cases} \tag{40}$$

$$\lim_{x\to -1}\frac{d}{dx}f(x) = \begin{cases} \lim_{x\to -1^-}\frac{d}{dx}f(x) = \lim_{x\to 1^+}\frac{3}{2}(1-x)+2 & = -2 \\ \lim_{x\to -1^+}\frac{d}{dx}f(x) = \lim_{x\to 1^-} 2x & = -2 \end{cases} \tag{41}$$

Ainsi la limite est définit à $x = \pm 1$, en passant par ce point la dérivé est définit, l'algorithme converge.

(b) (6 pts) **Prouver la convergence de la descente du gradient vers un point stationnaire dans le cas non-convexe** Supposons que nous essayons de minimiser la fonction $F(w)$ qui est L-lisse. Soit F_* la valeur minimale de la fonction (c'est-à-dire la valeur au niveau des minima globaux). En utilisant $\eta = \frac{1}{L}$, prouvez que la descente par gradient convergera "presque" vers un point stationnaire en un nombre borné (et polynomial) de pas. Précisément,

$$\min_{k<K}\|\nabla F(w^{(k)})\|^2 \leq \frac{2L}{K}(F(w^{(0)}) - F_*) \tag{42}$$

Réponse : Avec $\eta = \frac{1}{L}$,

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \frac{1}{L}\|\nabla F(w^{(k)})\|^2 + \frac{1}{2}\frac{1}{L}\|\nabla F(w^{(k)})\|^2 \tag{43}$$

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \frac{1}{2L}\|\nabla F(w^{(k)})\|^2 \tag{44}$$

$$\|\nabla F(w^{(k)})\|^2 \leq 2L(F(w^{(k)}) - F(w^{(k+1)})) \tag{45}$$

Sachant qu'à chaque itération $\|\nabla F(w^{(k)})\|^2$ le minimum d'une suite d'éléments est inférieur à la moyenne de la suite,

$$\sum_j^K \min_j\{\|\nabla F(w^{(k)})\|^2\} \leq 2L\sum_{k=0}^{K} F(w^{(k)}) - F(w^{(k+1)}) \tag{46}$$

$$\leq 2L\left(F(w^{(0)}) + \sum_{k=1}^{K-1}(F(w^{(k)}) - F(w^{(k+1)}))\right) \tag{47}$$

$$K\min_{k<K}\|\nabla F(w^{(k)})\|^2 \leq 2L\left(F(w^{(0)}) - F(w^{(K)})\right) \tag{48}$$

$$\min_{k<K}\|\nabla F(w^{(k)})\|^2 \leq \frac{2L}{K}\left(F(w^{(0)}) - F_*\right) \tag{49}$$

où $F(w^{(K)}) = F_*$.
Si on converge lorsque un certain seuil $\|\nabla F(w^{(k)})\|^2 < \epsilon$ est atteint, le nombre d'étape pour converger en dessous de ϵ est de l'ordre de $K \sim \mathcal{O}(1/\epsilon)$ étapes,

$$\frac{2L}{\epsilon}\left(F(w^{(0)}) - F_*\right) \leq K \tag{50}$$

Hints:

i. L-smoothness implique que:

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \eta\|\nabla F(w^{(k)})\|^2 + \frac{1}{2}\eta^2 L\|\nabla F(w^{(k)})\|^2 \tag{51}$$

Combinez ceci avec $\eta = \frac{1}{L}$.

ii. Utilisez le fait que le minimum d'une suite d'éléments est inférieur à la moyenne de la suite.