*remarques :* vous n'avez le droit à aucun documents. Certaines questions peuvent admettre plusieurs bonnes réponses.

# QCM

**Question 1 : Soit** $F : \mathbb{R}^{3\times2\times2} \mapsto \mathbb{R}^{4\times1}$

- ☐ $F$ admet 12 dérivées partielles
- ☐ $F$ admet 24 dérivées partielles
- ☑ $F$ admet 48 dérivées partielles
- ☐ $F$ admet 96 dérivées partielles
- ☐ $F$ admet 256 dérivées partielles

**Question 2 : La méthode des pivots de Gauss permet**

- ☐ d'inverser toutes les matrices carrées
- ☐ d'inverser toutes les matrices diagonales
- ☑ d'inverser toutes les matrices inversibles
- ☐ d'inverser toutes les matrices transposées

**Question 3 : Quel produit matriciel est valide**

- ☐ $A \in \mathbb{R}^{3\times2}$ et $B \in \mathbb{R}^{3\times3}$, on veut $A \times B$
- ☐ $A \in \mathbb{R}^{3\times2}$ et $B \in \mathbb{R}^{2\times2}$, on veut $B \times A$
- ☑ $A \in \mathbb{R}^{2\times2}$ et $B \in \mathbb{R}^{2\times3}$, on veut $A \times B$
- ☐ $A \in \mathbb{R}^{2\times2}$ et $B \in \mathbb{R}^{2\times3}$, on veut $B \times A$

**Question 4 : La fonction** $F : X \mapsto \sigma(X)$ **avec** $X \in \mathbb{R}^4$

- ☐ a 2 dérivées partielles
- ☐ a 4 dérivées partielles
- ☑ a 16 dérivées partielles
- ☐ le gradient de $F$ est un vecteur
- ☑ le gradient de $F$ est une matrice
- ☑ le gradient de $F$ est une matrice diagonale

**Question 5 : Sachant que** $\cosh' = \sinh$ **et** $\sinh' = \cosh$ **et** $\cosh^2 - \sinh^2 = 1$**, que vaut la dérivée de** $\tanh = \frac{\sinh}{\cosh}$

- ☐ $\frac{\sinh-\cosh}{\sinh^2}$
- ☐ $\frac{1}{\sinh^2}$
- ☐ $\frac{\sinh-\cosh}{\cosh^2}$
- ☑ $\frac{1}{\cosh^2}$

# Exercice 2 : Test d'hypothèses

On s'intéresse à une entreprise qui utilise un moteur de recherche : Bong. Les utilisateurs trouvent un résultat pertinent pour leur recherche dans 80% des cas en moyenne. Une jeune start-up française approche Bong avec une solution qui devrait augmenter la pertinence des résultats à 91%. Bong vous diligente pour tester cette affirmation. Vous mettez en œuvre une enquête d'opinion sur 49 personnes et obtenez le résultat suivant, avec l'utilisation de la solution de la start-up française : les utilisateurs estiment que le résultat est pertinent à 87%. Votre petit doigt vous dit que l'écart-type pour un utilisateur est de 28.

Pour l'ensemble de l'exercice, on notera que si X vaut 11% alors, on dira $X = 11$ (et pas $X = 0.11$).

**Question 1 : Pourra-t-on appliquer le TCL ?**

On note $X_n$ la variable aléatoire désignant l'avis d'un utilisateur. Notre étude comporte $N = 49 > 30$ individus, alors nous pourrons appliquer le TCL.

**Question 2 : Que peut-on dire de l'affirmation suivante "le résultat annoncé par la start-up ne peut pas être réfuté par un test d'hypothèses à 95%"**

| $\alpha$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.00 | $\infty$ | 2.576 | 2.326 | 2.170 | 2.054 | 1.960 | 1.881 | 1.812 | 1.751 | 1.695 |
| 0.10 | 1.645 | 1.598 | 1.555 | 1.514 | 1.476 | 1.440 | 1.405 | 1.372 | 1.341 | 1.311 |
| 0.20 | 1.282 | 1.254 | 1.227 | 1.200 | 1.175 | 1.150 | 1.126 | 1.103 | 1.080 | 1.058 |
| 0.30 | 1.036 | 1.015 | 0.994 | 0.974 | 0.954 | 0.935 | 0.915 | 0.896 | 0.878 | 0.860 |
| 0.40 | 0.842 | 0.824 | 0.806 | 0.789 | 0.772 | 0.755 | 0.739 | 0.722 | 0.706 | 0.690 |
| 0.50 | 0.674 | 0.659 | 0.643 | 0.628 | 0.613 | 0.598 | 0.583 | 0.568 | 0.553 | 0.539 |
| 0.60 | 0.524 | 0.510 | 0.496 | 0.482 | 0.468 | 0.454 | 0.440 | 0.426 | 0.412 | 0.399 |
| 0.70 | 0.385 | 0.372 | 0.358 | 0.345 | 0.332 | 0.319 | 0.305 | 0.292 | 0.279 | 0.266 |
| 0.80 | 0.253 | 0.240 | 0.228 | 0.215 | 0.202 | 0.189 | 0.176 | 0.164 | 0.151 | 0.138 |
| 0.90 | 0.126 | 0.113 | 0.100 | 0.088 | 0.075 | 0.063 | 0.050 | 0.038 | 0.025 | 0.013 |

| $\alpha$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ | $10^{-7}$ | $10^{-8}$ | $10^{-9}$ |
|---|---|---|---|---|---|---|---|
| $u_\alpha$ | 3.290 | 3.891 | 4.417 | 4.892 | 5.327 | 5.731 | 6.109 |

sous l'hypothèse que l'affirmation de la start-up est correcte, nous avons $\mu_{M_{49}} = 91$. D'après l'énoncé, on a $\sigma_{X_n} = 28$ et donc $\sigma_{M_{49}} = \sigma_{X_n}/\sqrt{49} = 4$. On applique le TCL, pour obtenir

$$\mathbb{P}\left(\frac{M_{49} - \mu_{M_{49}}}{\sigma_{M_{49}}} \in ]-1.96; 1.96[\right) = 0.95$$

ainsi

$$\mathbb{P}\left(M_{49} \in ]-1.96 \times 4 + 91; 1.96 \times 4 + 91[\right) = 0.95$$

Or $M_{49} = 87$ appartient clairement à l'intervalle $]-1.96 \times 4 + 91; 1.96 \times 4 + 91[$. En conséquence, le résultat annoncé par la start-up ne peut pas être réfuté par un test d'hypothèses à 95%.

**Question 3 : Que peut-on dire de l'affirmation suivante "l'affirmation selon laquelle la start-up n'a servi à rien ne peut être réfutée par un test d'hypothèses à 95%"**

sous l'hypothèse que l'aide de la start-up est inutile, nous avons $\mu_{M_{49}} = 80$. On applique le TCL, pour obtenir

$$\mathbb{P}\left(\frac{M_{49} - \mu_{M_{49}}}{\sigma_{M_{49}}} \in ]-1.96; 1.96[\right) = 0.95$$

ainsi

$$\mathbb{P}\left(M_{49} \in ]-1.96 \times 4 + 80; 1.96 \times 4 + 80[\right) = 0.95$$

Or $M_{49} = 87$ appartient clairement à l'intervalle $]-1.96 \times 4 + 80; 1.96 \times 4 + 80[$. En conséquence, l'affirmation selon laquelle la start-up n'a servi à rien ne peut être réfutée par un test d'hypothèses à 95%.

**Question 4 : Que recommandez-vous ?**

Plusieurs recommandations possibles :

— faire un test sur un échantillon plus grand

— ne pas s'engager avec la start-up

## Exercice 3 : Self-Attention

Les blocs d'auto attention sont au cœur des modèles dit "transformers" qui occupent le devant de la scène en deep learning depuis quelques années maintenant. Le principe est de calculer une corrélation entre les différentes entrées

du réseau et de s'en servir comme pondération. Ainsi un bloc d'attention $B : \mathbb{R}^n \to \mathbb{R}^n$, calcule une attention $A$, une projection $V$ et une recombinaison telles que

$$\begin{cases} A : X \mapsto (W_1 X^T) \times (X W_2), \text{ avec } W_1 \in \mathbb{R}^{n\times n}, W_2 \in \mathbb{R}^{n\times n} \\ V : X \mapsto W_3 X \\ B : X \mapsto W_4 \text{softmax}(A(X))V(X) \end{cases}$$

avec $X^T \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{1\times n} = \mathbb{R}^n$ et $R(X) \in \mathbb{R}^n$.

*remarque : il est recommandé de travailler avec les dérivées partielles plutôt que le gradient directement*

**Question 1 : Quelles sont les dimensions de $W_1$, $W_2$, $W_3$ et $W_4$ ?**

D'après l'énoncé, on a $W_1 \in \mathbb{R}^{n\times n}$ et $W_2 \in \mathbb{R}^{n\times n}$. Donc $A(X) \in \mathbb{R}^{n\times n}$ et par définition du softmax, qui ne change pas les dimensions, $\text{softmax}(A(X)) \in \mathbb{R}^{n\times n}$. Donc $W_3 X \in \mathbb{R}^n$, ce qui implique $W_3 \in \mathbb{R}^{n\times n}$. Enfin, on sait que $\text{softmax}(A(X))V(X) \in \mathbb{R}^n$ et $B(X) \in \mathbb{R}^n$, donc $W_4 \in \mathbb{R}^{n\times n}$.

**Question 2 : Calculez le gradient $\nabla_X A(X)$**

$A$ est de fonction de $\mathbb{R}^n$ dans $\mathbb{R}^{n\times n}$ et donc admet $n^3$ dérivées partielles. On cherche donc à calculer tous les $\frac{\partial A(X)_{i,j}}{\partial X_k}$. Pour cela, nous allons expliciter la formule de $A(X)_{i,j}$. En effet, $A(X)$ est le résultat d'un outer product entre deux vecteurs $(W_1 X^T)$ et $(XW_2)$. Ainsi, $A(X)_{i,j} = (W_1 X^T)_i \times (XW_2)_j$. On en déduit que

$$\frac{\partial A(X)_{i,j}}{\partial X_k} = W_{1i,k}(XW_2)_j + (W_1 X^T)_i W_{2k,j}$$

**Question 3 : Calculez le gradient $\nabla_X V(X)$**

$V$ est une fonction linéaire de $X$ que l'on connait bien, telle que $\nabla_X V(X) = W_3$.

Ici le softmax est un cas particulier. En effet, la sortie de $A$ est une matrice de taille $n \times n$ (à cause de la transposée). Le softmax est alors le suivant

$$\text{softmax}(M) = \begin{pmatrix} \frac{e^{M_{1,1}}}{\sum_{k=1}^n e^{M_{1,k}}} & \cdots & \frac{e^{M_{1,n}}}{\sum_{k=1}^n e^{M_{1,k}}} \\ \vdots & \ddots & \vdots \\ \frac{e^{M_{n,1}}}{\sum_{k=1}^n e^{M_{n,k}}} & \cdots & \frac{e^{M_{n,n}}}{\sum_{k=1}^n e^{M_{n,k}}} \end{pmatrix}$$

Autrement dit, on transforme chaque ligne de $M$ en une distribution de probabilités.

**Question 4 : Calculez le gradient $\nabla_M \textbf{softmax}(M)$**

Ici la fonction softmax est une fonction de $\mathbb{R}^{n\times n}$ dans $\mathbb{R}^{n\times n}$, pour un total de $n^4$ dérivées partielles. On cherche donc à calculer tous les $\frac{\partial \text{softmax}(M)_{i,j}}{\partial M_{k,l}}$. Si $k \neq i$ on a immédiatement $\frac{\partial \text{softmax}(M)_{i,j}}{\partial M_{k,l}} = 0$. Sinon, nous sommes dans le cas du softmax sur un vecteur $M_i$ et ainsi

$$\frac{\partial \text{softmax}(M)_{i,j}}{\partial M_{i,l}} = \begin{cases} \text{softmax}(M)_{i,j}(1 - \text{softmax}(M)_{i,j}) \text{ si } j = l \\ -\text{softmax}(M)_{i,j}\text{softmax}(M)_{i,l}) \text{ sinon} \end{cases}$$

Ainsi,

$$\frac{\partial \text{softmax}(M)_{i,j}}{\partial M_{k,l}} = \begin{cases} \text{softmax}(M)_{i,j}(1 - \text{softmax}(M)_{i,j}) \text{ si } j = l \text{ et } i = k \\ -\text{softmax}(M)_{i,j}\text{softmax}(M)_{i,l}) \text{ si } i = k \\ 0 \text{ sinon} \end{cases}$$

**Question 5 : Calculez le gradient $\nabla_X B(X)$**

Ok... il est temps de tout mettre bout à bout. $B$ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$ et admet donc $n^2$ dérivées partielles.

$$\frac{\partial B(X)_i}{\partial X_j} = \frac{\partial (W_4 \text{softmax}(A(X))V(X))_i}{\partial X_j} = W_4 \frac{\partial (\text{softmax}(A(X))V(X))_i}{\partial X_j} = W_4 \frac{\partial \left(\sum_{k=1}^n \text{softmax}(A(X))_{i,k} V(X)_k\right)}{\partial X_j}$$

On utilise la linéarité de la dérivation

$$\frac{\partial B(X)_i}{\partial X_j} = W_4 \sum_{k=1}^n \frac{\partial \left(\text{softmax}(A(X))_{i,k} V(X)_k\right)}{\partial X_j}$$

On utilise la propriété de la dérivation d'un produit

$$\frac{\partial B(X)_i}{\partial X_j} = W_4 \sum_{k=1}^{n} \frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial X_j} V(X)_k + \frac{\partial \left(V(X)_k\right)}{\partial X_j} \text{softmax}(A(X))_{i,k}$$

Or, on sait que $frac\partial \left(V(X)_k\right)\partial X_j = W_{3k,j}$, ainsi

$$\frac{\partial B(X)_i}{\partial X_j} = W_4 \sum_{k=1}^{n} \frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial X_j} V(X)_k + W_{3k,j} \times \text{softmax}(A(X))_{i,k}$$

Il nous manque juste $\frac{\partial(\text{softmax}(A(X))_{i,k})}{\partial X_j}$. Pour cette dérivée partielle, nous allons utiliser la chain rule.

$$\frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial X_j} = \frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial A(X)_i} \frac{\partial A(X)_i}{\partial X_j}$$

On reconnait ici un produit scalaire

$$\frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial X_j} = \sum_{l=1}^{n} \frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial A(X)_{i,l}} \frac{\partial A(X)_{i,l}}{\partial X_j}$$

On remplace toutes ses dérivées partielles et on trouve

$$\frac{\partial \left(\text{softmax}(A(X))_{i,k}\right)}{\partial X_j} = \sum_{l=1}^{n} \left(W_{1i,j}(XW_2)_l + (W_1 X^T)_i W_{2j,l}\right) \begin{cases} \text{softmax}(A(X))_{i,k}(1 - \text{softmax}(A(X))_{i,k}) \text{ si } k = l \\ -\text{softmax}(A(X))_{i,k}\text{softmax}(A(X))_{i,l}) \text{ sinon} \end{cases}$$

# Exercice 4 : (théorème d'approximation universelle)

Nous allons démontrer que les réseaux de neurones ReLU sont des approximateurs universels. Ce théorème est fondamental en deep learning.

## Partie I

Pour se mettre en jambes, nous allons montrer que les fonctions constantes par morceaux sont des approximateurs universels.

Nous allons utiliser le théorème de Heine :

**Theorem 1.** *Toute fonction $f$ continue de $[a;b]$ dans $\mathbb{R}$ est uniformément continue, i.e.*

$$\forall \epsilon > 0, \exists \delta > 0, \forall x, y \quad |x - y| < \delta \Rightarrow |f(x) - f(y)| < \epsilon$$

(la différence avec la continuité est le fait que $\delta$ ne dépend pas de $x$)

Soit $f$ une fonction continue sur $[a;b]$ mais pas uniformément continue.

**Question 1 : Montrer qu'il existe un réel $\epsilon > 0$ et deux suites $(x_n)$ et $(y_n)$ d'éléments de $[a;b]$ tels que pour tout entier $|x_n - y_n| < \frac{1}{n}$ et $|f(x_n) - f(y_n)| \geq \epsilon$**

Puisque $f$ n'est pas uniformément continue, il existe $\epsilon > 0$ tel que pour tout $\delta > 0$, il existe $x, y$ tels que $|x - y| < \delta \Rightarrow |f(x) - f(y)| \geq \epsilon$. On pose pour chaque $n \in \mathbb{N}$, $\delta = \frac{1}{n}$. On a alors l'existence de $x$ et $y$ que l'on notera $x_n$ et $y_n$ tels que $|x_n - y_n| < \frac{1}{n}$ et $|f(x_n) - f(y_n)| \geq \epsilon$.

Pour rappel, toute suite sur un fermé borné admet une sous-suite convergente.

**Question 2 : Montrer que les suites précédentes $(x_n)$ et $(y_n)$ admettent deux sous suites convergentes $(x_{\sigma(n)})$ et $(y_{\sigma(n)})$ telle que $|x_{\sigma(n)} - y_{\sigma(n)}| < \frac{1}{n}$ et $|f(x_{\sigma(n)}) - f(y_{\sigma(n)})| \geq \epsilon$**

Puisque les $x_n$ appartiennent à $[a;b]$, cette suite admet une sous-suite convergente que l'on notera $(x_{\phi(n)})$. Les éléments de la suite $(y_{\phi(n)})$ appartiennent à $[a;b]$, cette suite admet une sous-suite convergente que l'on notera $(y_{\sigma(n)})$. La sous-suite $(x_{\sigma(n)})$ est une sous-suite d'une suite convergente et est donc convergente.

**Question 3 : montrez que ces deux sous-suites convergent vers la même limite**

Par construction de $(x_n)$ et $(y_n)$ on a $\lim_{n\to\infty} |x_{\sigma(n)} - y_{\sigma(n)}| = 0$ donc les deux sous-suites convergentes convergent vers la même limite.

**Question 4 : conclure**

Par continuité de $f$, on a que $\lim_{n\to\infty} |f(x_{\sigma(n)}) - f(y_{\sigma(n)})| = 0$ et $|f(x_{\sigma(n)}) - f(y_{\sigma(n)})| \geq \epsilon > 0$ ce qui est absurde et achève la preuve.

Un approximateur universel est un type $T$ de fonction tel que, quelle que soit la fonction $f$ à approximer, on a

$$\forall \epsilon > 0, \exists t \in T, \quad \text{tel que } \max_x \{|f(x) - t(x)|\} < \epsilon$$

**Question 5 : En supposant le théorème de Heine, prouvez que les fonctions constantes par morceaux sont des approximateurs universels des fonctions continue sur $[a; b]$.**

Ce résultat découle immédiatement de $f$ en prenant $\epsilon$ (de l'approximation universelle) $= \epsilon$ (de l'uniforme continuité) et des morceaux de taille $\delta$ pour une constante égale a un $f(x)$.

## Partie II

Soit la fonction $\bar{\sigma} = \mathbb{1}_{\mathbb{R}_+}$ (la fonction heavyside) et $\sigma$ la relu.

**Question 1 : Montrer que toute fonction constante par morceau est un réseau $g : \mathbb{R} \to \mathbb{R}$ tel que $g(x) = W_2\bar{\sigma}(W_1 x + b_1) + b_2$ avec $W_1$ et $W_2$ des matrices et $b_1$ et $b_2$ sont des vecteurs.**

On commence par définir la fonction constante sur $[a; b]$ de valeur $c$ et qui vaut 0 partout ailleurs :

$$\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} \bar{\sigma}\left(\begin{pmatrix} 1 & -1 & 1 & -1 \end{pmatrix} x + \begin{pmatrix} -a & b & b & a \end{pmatrix}\right)$$

En suivant cette construction, on peut construire toutes les fonctions constantes par morceaux en concaténant les paramètres des fonctions constantes sur un morceau.