# IFT3395 - Fondements de l'apprentissage machine

## Travail Pratique 2

**Nom**: Paul CHAFFANET
**Matricule**: 1009543

**Nom**: Émile Labbé
**Matricule**: 20019813

# I. PARTIE THÉORIQUE A : relations et dérivées de quelques fonctions de base.

## Question 1

***Montrez que*** $\operatorname{sigmoid}(x)=\frac{1}{2}\left(\tanh\left(\frac{1}{2}x\right)+1\right)$***.***

$$
\begin{aligned}
\frac{1}{2}\left(\tanh\left(\frac{1}{2}x\right)+1\right) &= \frac{1}{2}\left(\frac{\exp\left(\frac{1}{2}x\right)-\exp\left(-\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}+1\right)\\
&= \frac{1}{2}\left(\frac{\exp\left(\frac{1}{2}x\right)-\exp\left(-\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}+\frac{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}\right)\\
&= \frac{1}{2}\left(\frac{\exp\left(\frac{1}{2}x\right)-\exp\left(-\frac{1}{2}x\right)+\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}\right)\\
&= \frac{1}{2}\left(\frac{2\exp\left(\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}\right)\\
&= \frac{\exp\left(\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}\\
&= \frac{\exp\left(\frac{1}{2}x\right)}{\exp\left(\frac{1}{2}x\right)+\exp\left(-\frac{1}{2}x\right)}\bullet\frac{\exp\left(-\frac{1}{2}x\right)}{\exp\left(-\frac{1}{2}x\right)}\\
&= \frac{\exp(0)}{\exp(0)+\exp(-x)}\\
&= \frac{1}{1+\exp(-x)}\\
&= \operatorname{sigmoid}(x)
\end{aligned}
$$

## Question 2

***Montrez que*** $\ln \mathrm{sigmoid}(x) = -\mathrm{softplus}(-x)$ ***.***

$$
\begin{aligned}
&\ln \mathrm{sigmoid}(x) = \ln\left(\frac{1}{1+\exp(-x)}\right) \\
&= \ln 1 - \ln\left(1+\exp(x)\right) \\
&= -\ln\left(1+\exp(x)\right) \\
&= -\mathrm{softplus}(-x)
\end{aligned}
$$

## Question 3

***Montrez que la dérivée de*** $\mathrm{sigmoid}$ ***est :*** $\mathrm{sigmoid}'(x) = \frac{d\mathrm{sigmoid}}{dx}(x)$

$= \mathrm{sigmoid}(x)\left(1-\mathrm{sigmoid}(x)\right)$ ***.***

$$
\begin{aligned}
&\mathrm{sigmoid}(x)\left(1-\mathrm{sigmoid}(x)\right) = \frac{1}{1+\exp(-x)}\left(1-\frac{1}{1+\exp(-x)}\right) \\
&= \frac{1}{1+\exp(-x)} - \frac{1}{\left(1+\exp(-x)\right)^2} \\
&= \frac{1+\exp(-x)}{1+\exp(-x)} \bullet \frac{1}{1+\exp(-x)} - \frac{1}{\left(1+\exp(-x)\right)^2} \\
&= \frac{\exp(-x)}{\left(1+\exp(-x)\right)^2}
\end{aligned}
$$

Or, par la règle de la dérivée d'un quotient, $\mathrm{sigmoid}'(x) = \frac{u'v - uv'}{v^2}$ avec:

$$
\begin{aligned}
&u = 1 \quad v = 1+\exp(-x) \\
&u' = 0 \quad v' = -\exp(-x)
\end{aligned}
$$

Et donc :

$$\text{sigmoid}'(x) = \frac{0 \bullet (1+\exp(x)) - 1 \bullet (-\exp(-x))}{(1+\exp(x))^2}$$
$$= \frac{\exp(-x)}{(1+\exp(x))^2}$$
$$= \text{sigmoid}(x)(1-\text{sigmoid}(x))$$

---

## Question 4

***Montrez que la dérivée de*** $\tanh$ ***est :*** $\tanh'(x) = 1 - \tanh^2(x)$ ***.***

Par la règle de la dérivée d'un quotidient, $\tanh(x) = \frac{u'v - uv'}{v^2}$ avec:

$$u = \exp(x) - \exp(-x) \quad v = \exp(x) + \exp(-x)$$
$$u' = \exp(x) + \exp(-x) \quad v' = \exp(x) - \exp(-x)$$

Et donc :

$$\tanh'(x) = \frac{(\exp(x)+\exp(-x))^2 - (\exp(x)-\exp(-x))^2}{(\exp(x)+\exp(-x))^2}$$
$$= 1 - \left(\frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}\right)^2$$
$$= 1 - \tanh^2(x)$$

---

## Question 5

***Exprimez la fonction*** $\text{sign}$ ***en utilisant des fonctions indicatrices.***

$$\text{sign}(x) = \begin{cases} 1 \text{ si } x > 0 \\ -1 \text{ si } x < 0 \\ 0 \text{ sinon} \end{cases} = 2 \bullet 1_{\{x>0\}} + 2 \bullet 1_{\{x=0\}} - 1$$

## Question 6

**Écrivez la dérivée de la fonction valeur absolue** $abs(x) = |x|$ **.**

$$\text{abs}'(x) = \begin{cases} 1 \ \text{si} \ x > 0 \\ -1 \ \text{si} \ x < 0 \\ 0 \ \text{sinon} \end{cases} \text{ alors } \text{abs}'(x) = \text{sign}(x)$$

## Question 7

**Écrivez la dérivée de la fonction** $rect$ **.**

$$\text{rect}'(x) = \begin{cases} 1 \ \text{si} \ x > 0 \\ 0 \ \text{sinon} \end{cases} = 1_{\{x>0\}}$$

## Question 8

**Soit la norme** $L_2$ **d'un vecteur :** $\|x\|_2^2 = \sum_i x_i^2$ **. Écrivez le vecteur de gradient :** $\frac{\partial \|x\|_2^2}{\partial x}$ **.**

$$\frac{\partial \|x\|_2^2}{\partial x} = \begin{pmatrix} \frac{\partial \|x\|_2^2}{\partial x_1} \\ \vdots \\ \frac{\partial \|x\|_2^2}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial x_1^2}{\partial x_1} \\ \vdots \\ \frac{\partial x_n^2}{\partial x_n} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_n \end{pmatrix}$$

## Question 9

**Soit la norme** $L_2$ **d'un vecteur :** $\|x\|_1 = \sum_i |x_i|$ **. Écrivez le vecteur de gradient :** $\frac{\partial \|x\|_1}{\partial x}$ **.**

$$\frac{\partial \|x\|_1}{\partial x} = \begin{pmatrix} \frac{\partial \|x\|_1}{\partial x_1} \\ \vdots \\ \frac{\partial \|x\|_1}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial |x_1|}{\partial x_1} \\ \vdots \\ \frac{\partial |x_n|}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \operatorname{sign}(x_1) \\ \vdots \\ \operatorname{sign}(x_n) \end{pmatrix}$$

# II. PARTIE THÉORIQUE B : Calcul du gradient pour l'optimisation des paramètres d'un réseau de neurones pour la classification multiclasse

---

## Question 1

***a) Indiquer la dimension de*** $b^{(1)}$ ***.***

La dimension du vecteur de biais est $b^{(1)} \in \mathbb{R}^{d_h \times 1}$, où $d_h$ est le nombre de neurones de la couche cachée.

***b) Donnez la formule de calcul du vecteur d'activations (i.e. avant non-linéarité) des neurones de la couche cachée*** $h^a$ ***à partir d'une observation d'entrée*** $x$ ***sous la forme d'une expression de calcul matriciel.***

Le vecteur d'activations se définit par $h^a = W^{(1)}x + b^{(1)}$.

***c) Détaillez le calcul d'un élément*** $h_j^a$ ***.***

Le calcul d'un élément du vecteur d'activations de la couche cachée est :

$$h_j^a = W_j^{(1)}x + b_j^{(1)} = b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i \ .$$

***d) Exprimez le vecteur des sorties des neurones de la couche cachée*** $h^s$ ***en fonction de*** $h^a$ ***.***

Le vecteur des sorties des neurones de la couche cachée est défini par:

$$h^s = rect\left(h^a\right)$$

## Question 2

***a) Indiquer la dimension de $W^{(2)}$ et $b^{(2)}$.***

$W^{(2)} \in \mathbb{R}^{m\times d_h}$ et $b^{(2)} \in \mathbb{R}^{m\times 1}$.

***b) Donnez la formule de calcul du vecteur d'activations des neurones de la couche de sortie $o^a$ à partir de leurs entrées $h^s$ sous la forme d'une expression de calcul matriciel.***

Le vecteur d'activations se définit par $o^a = W^{(2)}h^s + b^{(2)}$.

***c) Détaillez le calcul de $o_k^a$.***

$o_k^a = b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s$ où $h_i^s = rect\left(h_i^a\right) = \left(1_{\{h_i^a > 0\}}\right) h_i^a$.

## Question 3

***a) La sortie des neurones de sortie est donnée par $o^s = \mathrm{softmax}\left(o^a\right)$. Précisez l'équation des $o_k^s$ en utilisant explicitement la formule du $\mathrm{softmax}$ (formule avec des exp).***

$o^s = \mathrm{softmax}\left(o^a\right) = \frac{1}{\sum_{i=1}^{m} \exp\left(o_i^a\right)} \bullet \begin{pmatrix} \exp\left(o_1^a\right) \\ \exp\left(o_2^a\right) \\ \vdots \\ \exp\left(o_m^a\right) \end{pmatrix}$ alors $o_k^s = \frac{\exp\left(o_k^a\right)}{\sum_{i=1}^{m} \exp\left(o_i^a\right)}$

***b) Démontrez que les $o_k^s$ sont positifs et somment à 1.***

Comme $\lim_{x \to -\infty} \exp(x) = 0$ et que $\lim_{x \to +\infty} \exp(x) = +\infty$, on peut déduire que:

- le numérateur est positif du fait de la fonction exponentielle;
- le dénominateur qui est une somme $\sum_{i=1}^{m} \exp(o_i^a)$ de valeurs positives est forcément de signe positif;
- le rapport entre deux valeurs positives donne une valeur positive.

Ainsi, $o_k^s$ est positive et somment à 1:

$$\begin{aligned}\sum_{k=1}^{m} o_k^s &= \frac{1}{\sum_{i=1}^{m} \exp(o_i^a)} \bullet \sum_{i=1}^{m} \exp(o_i^a) \\ &= 1\end{aligned}$$

***c) Pourquoi est-ce important ?***

Les sorties représentent les probabilités d'un exemple x d'appartenir à une certaine classe. Ainsi, c'est important que ces sorties somment à 1 afin d'avoir une probabilité d'univers à 1. Les sorties sont des évènements mutuellements exclusifs grâce à la fonction softmax qui construit un espace probabilisé.

## Question 4

***Précisez l'équation de $L(x,y) = -\log o_y^s(x)$ directement en fonction du vecteur $o^a$.***

$$\begin{aligned}L(x,y) &= -\log o_y^s(x) \\ &= -\log \frac{\exp(o_y^a)}{\sum_{i=1}^{m} \exp(o_i^a)} \\ &= -\log \exp(o_y^a) + \log \sum_{i=1}^{m} \exp(o_i^a) \\ &= -o_y^a + \log \sum_{i=1}^{m} \exp(o_i^a)\end{aligned}$$

## Question 5

***a) Formulez $\widehat{R}$ .***

$$\widehat{R}(f_\theta, D) = \frac{1}{|D|}\sum_{i=1}^{|D|} L\left(x^{(i)}, y^{(i)}\right) = -\frac{1}{|D|}\sum_{i=1}^{|D|} \log o^{s}_{y^{(i)}}\left(x^{(i)}; \theta\right)$$

***b) Indiquez précisément de quoi est constitué l'ensemble $\theta$ des paramètres du réseau. Indiquez à combien de paramètres scalaires $n_\theta$ cela correspond.***

$\theta = \left\{W^{(2)}, W^{(1)}, b^{(2)}, b^{(1)}\right\}$ constitue l'ensemble des paramètres du réseau.

Comme $W^{(1)} \in \mathbb{R}^{d_h \times d}$, $W^{(2)} \in \mathbb{R}^{m \times d_h}$, $b^{(1)} \in \mathbb{R}^{d_h \times 1}$ et $b^{(2)} \in \mathbb{R}^{m \times 1}$, alors on a :

$$n_\theta = m + d_h + d_h \times d + m \times d_h$$

***c) Formulez le problème d'optimisation qui correspond à l'entraînement du réseau permettant de trouver une valeur optimale des paramètres.***

$$\theta^* = \arg\min_\theta \widehat{R}(f_\theta, D)$$

## Question 6 (pas sûr du tout)

***Exprimez sous forme d'un bref pseudo-code la technique de descente de gradient (batch) pour ce problème.***

**procedure** descenteGradientBatch(batch, epoch) :
    $i \leftarrow 0$
    **while** $i <$ epoch :
        grad $\leftarrow 0$
        **for** exemple **in** batch :
            grad $\leftarrow$ grad + backprop(ex)
        $\theta \leftarrow \theta - \frac{\eta}{\text{batch\_size}} \cdot \text{grad}$

## Question 7

***Démontrez que*** $\frac{\partial L}{\partial o^a} = o^s - \mathrm{onehot}_m(y)$***.***

On commence par calculer $\frac{\partial L}{\partial o_k^a}$ tel que $k \neq y$:

$$\frac{\partial L}{\partial o_k^a} = \frac{\partial}{\partial o_k^a}\left(-o_y^a + \log\sum_{i=1}^{m}\exp\left(o_i^a\right)\right)$$

$$= 0 + \frac{\exp\left(o_k^a\right)}{\sum_{i=1}^{m}\exp\left(o_i^a\right)}$$

$$= o_k^s$$

Puis on calcule $\frac{\partial L}{\partial o_y^a}$ (cas où $k = y$):

$$\frac{\partial L}{\partial o_y^a} = \frac{\partial}{\partial o_y^a}\left(-o_y^a + \log\sum_{i=1}^{m}\exp\left(o_i^a\right)\right)$$

$$= -1 + \frac{o_y^a}{\sum_{i=1}^{m}\exp\left(o_i^a\right)}$$

$$= -1 + o_y^s$$

On peut conclure à cette étape que:

$$\frac{\partial L}{\partial o_k^a} = -1_{\{k=y\}} + \frac{o_k^a}{\sum_{i=1}^{m}\exp\left(o_i^a\right)}$$

On peut ainsi déduire $\dfrac{\partial L}{\partial o^a}$ :

$$
\begin{aligned}
\frac{\partial L}{\partial o^a} &= \begin{pmatrix} \dfrac{\partial L}{\partial o_1^a} \\ \vdots \\ \dfrac{\partial L}{\partial o_m^a} \end{pmatrix} \\
&= \begin{pmatrix} -1_{\{1=y\}} + \dfrac{o_1^a}{\sum_{i=1}^{m} \exp\left(o_i^a\right)} \\ \vdots \\ -1_{\{m=y\}} + \dfrac{o_m^a}{\sum_{i=1}^{m} \exp\left(o_i^a\right)} \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{o_1^a}{\sum_{i=1}^{m} \exp\left(o_i^a\right)} \\ \vdots \\ \dfrac{o_m^a}{\sum_{i=1}^{m} \exp\left(o_i^a\right)} \end{pmatrix} - \text{onehot}_m(y) \\
&= \begin{pmatrix} o_1^s \\ o_2^s \\ \vdots \\ o_m^s \end{pmatrix} - \text{onehot}_m(y) \\
&= o^s - \text{onehot}_m(y)
\end{aligned}
$$

## Question 8

***Donnez l'expression correspondante du gradient en numpy (possiblement en 2 opérations successives).***

```
grad_oa = os - numpy.array([0 if i != y - 1 else 1 for i in range(m)]).reshape(m,1)
```

## Question 9

***Calculez les gradients par rapport aux paramètres $W^{(2)}$ et $b^{(2)}$ de la couche de sortie.***

$$\frac{\partial L}{\partial W_{kj}^{(2)}} = \frac{\partial L}{\partial o_k^a}\frac{\partial o_k^a}{\partial W_{kj}^{(2)}}$$
$$= \frac{\partial L}{\partial o_k^a}\frac{\partial}{\partial W_{kj}^{(2)}}\left(b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s\right)$$
$$= \frac{\partial L}{\partial o_k^a} h_j^s$$

et

$$\frac{\partial L}{\partial b_k^{(2)}} = \frac{\partial L}{\partial o_k^a}\frac{\partial o_k^a}{\partial b_k^{(2)}}$$
$$= \frac{\partial L}{\partial o_k^a}\frac{\partial}{\partial b_k^{(2)}}\left(b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s\right)$$
$$= \frac{\partial L}{\partial o_k^a}$$

## Question 10

**a)** ***Exprimez le calcul du gradient de la question précédente sous forme d'une expression matricielle, en définissant la dimension de chacune des matrices ou vecteurs manipulés.***

$$\frac{\partial L}{\partial W^{(2)}} = \begin{pmatrix} \frac{\partial L}{\partial W_{11}^{(2)}} & \cdots & \frac{\partial L}{\partial W_{1d_h}^{(2)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial W_{m1}^{(2)}} & \cdots & \frac{\partial L}{\partial W_{md_h}^{(2)}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial L}{\partial o_1^a} h_1^s & \cdots & \frac{\partial L}{\partial o_1^a} h_{d_h}^s \\ \vdots & \ddots & \vdots \\ \frac{\partial L}{\partial o_m^a} h_1^s & \cdots & \frac{\partial L}{\partial o_m^a} h_{d_h}^s \end{pmatrix}$$

$$= \frac{\partial L}{\partial o^a} h^{s^T}$$

et:

$$\frac{\partial L}{\partial b^{(2)}} = \begin{pmatrix} \frac{\partial L}{\partial b_1^{(2)}} \\ \vdots \\ \frac{\partial L}{\partial b_m^{(2)}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\partial L}{\partial o_1^a} \\ \vdots \\ \frac{\partial L}{\partial o_m^a} \end{pmatrix}$$

$$= \frac{\partial L}{\partial o^a}$$

Les dimensions de chaque matrices ou vecteurs manipulés sont:

- $\frac{\partial L}{\partial W^{(2)}} \in \mathbb{R}^{m \times d_h}$
- $\frac{\partial L}{\partial o^a} \in \mathbb{R}^{m \times 1}$
- $h^s \in \mathbb{R}^{d_h \times 1}$ donc $h^{s^T} \in \mathbb{R}^{1 \times d_h}$
- $\frac{\partial L}{\partial b^{(2)}} \in \mathbb{R}^{m \times 1}$

***b) Donnez l'expression correspondante en numpy.***

```
grad_b2 = grad_oa
grad_W2 = np.dot(grad_oa, np.transpose(hs))
```

---

## Question 11

***Calculez les dérivées partielles du coût $L$ par rapport aux sorties des neurones de la couche cachée.***

$$\frac{\partial L}{\partial h_j^s} = \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} \frac{\partial o_k^a}{\partial h_j^s}$$
$$= \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} \frac{\partial}{\partial h_j^s} \left( b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s \right)$$
$$= \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} W_{kj}^{(2)}$$

## Question 12

**a)** ***Exprimez le calcul de la question précédente sous forme d'une expression matricielle, en définissant la dimension de chacune des matrices ou vecteurs manipulées.***

$$
\begin{aligned}
\frac{\partial L}{\partial h^s} &= \begin{pmatrix} \dfrac{\partial L}{\partial h_1^s} \\ \vdots \\ \dfrac{\partial L}{\partial h_{d_h}^s} \end{pmatrix} \\
&= \begin{pmatrix} \displaystyle\sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} W_{k1}^{(2)} \\ \vdots \\ \displaystyle\sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} W_{kd_h}^{(2)} \end{pmatrix} \\
&= W^{(2)^T} \frac{\partial L}{\partial o^a}
\end{aligned}
$$

Les dimensions de chaque matrices ou vecteurs manipulés sont:

- $\dfrac{\partial L}{\partial h^s} \in \mathbb{R}^{d_h \times 1}$
- $W^{(2)} \in \mathbb{R}^{m \times d_h}$ alors $W^{(2)^T} \in \mathbb{R}^{d_h \times m}$
- $\dfrac{\partial L}{\partial o^a} \in \mathbb{R}^{m \times 1}$

***b) Donnez l'expression correspondante en numpy.***

```
grad_hs = np.dot(np.transpose(W2), grad_oa)
```

## Question 13

***Calculez les dérivées partielles par rapport aux activations des neurones de la couche cachée.***

Par la réponse à la question A.7, on sait que: $\frac{\partial \text{rect}(z)}{\partial z} = 1_{\{z>0\}}$ .

Donc:

$$\begin{aligned}\frac{\partial L}{\partial h_j^a} &= \frac{\partial L}{\partial h_j^s}\frac{\partial \text{rect}\left(h_j^a\right)}{\partial h_j^a} \\ &= \frac{\partial L}{\partial h_j^s} 1_{\{h_j^a>0\}}\end{aligned}.$$

## Question 14

***a) Exprimez le calcul de la question précédente sous forme d'une expression matricielle, en définissant la dimension de chacune des matrices ou vecteurs manipulés.***

$$\begin{aligned}\frac{\partial L}{\partial h^a} &= \begin{pmatrix} \frac{\partial L}{\partial h_1^a} \\ \vdots \\ \frac{\partial L}{\partial h_{d_h}^a} \end{pmatrix} = \begin{pmatrix} \frac{\partial L}{\partial h_1^s} 1_{\{h_1^a>0\}} \\ \vdots \\ \frac{\partial L}{\partial h_{d_h}^s} 1_{\{h_{d_h}^a>0\}} \end{pmatrix} \\ &= \text{diag}\left(1_{\{h^a>0\}}\right)\frac{\partial L}{\partial h^s}\end{aligned}$$

avec la fonction $diag$ qui transforme en matrice diagonale un vecteur colonne passé en paramètre.

Les dimensions de chaque matrices ou vecteurs manipulés sont:

- $\frac{\partial L}{\partial h^a} \in \mathbb{R}^{d_h \times 1}$
- $\frac{\partial L}{\partial h^s} \in \mathbb{R}^{d_h \times 1}$
- $1_{\{h^a > 0\}} \in \mathbb{R}^{d_h \times d_h}$

***b) Donnez l'expression équivalente en numpy.***

grad_ha = grad_hs * np.where(ha > 0, [1], [0])

---

## Question 15

***Calculez les gradients par rapport aux éléments des paramètres $W^{(1)}$ et $b^{(1)}$ de la couche cachée.***

Pour $W^{(1)}$ :

$$\begin{aligned}
\frac{\partial L}{\partial W_{kj}^{(1)}} &= \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial W_{kj}^{(1)}} \\
&= \frac{\partial L}{\partial h_k^a} \frac{\partial}{\partial W_{kj}^{(1)}} \left( b_k^{(1)} + \sum_{i=1}^{d} W_{ki}^{(1)} x_i \right) \\
&= \frac{\partial L}{\partial h_k^a} x_j
\end{aligned}$$

Pour $b^{(1)}$ :

$$\begin{aligned}
\frac{\partial L}{\partial b_k^{(1)}} &= \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial b_k^{(1)}} \\
&= \frac{\partial L}{\partial h_k^a} \frac{\partial}{\partial b_k^{(1)}} \left( b_k^{(1)} + \sum_{i=1}^{d} W_{ki}^{(1)} x_i \right) \\
&= \frac{\partial L}{\partial h_k^a}
\end{aligned}$$

## Question 16

***a) Exprimez le calcul de la question précédente sous forme d'une expression matricielle, en définissant la dimension de chacune des matrices ou vecteurs manipulés.***

$$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial h^a} x^T$$

$$\frac{\partial L}{\partial b^{(1)}} = \frac{\partial L}{\partial h^a}$$

Les dimensions de chaque matrices ou vecteurs manipulés sont:

- $\frac{\partial L}{\partial W^{(1)}} \in \mathbb{R}^{d_h \times d}$
- $\frac{\partial L}{\partial b^{(1)}} \in \mathbb{R}^{d_h \times 1}$
- $\frac{\partial L}{\partial h^a} \in \mathbb{R}^{d_h \times 1}$
- $x^T \in \mathbb{R}^{1 \times d}$

***b) Donnez l'expression équivalente en numpy.***

```
grad_b1 = grad_ha
grad_W1 = np.dot(grad_ha, np.transpose(x))
```

## Question 17

***Calculez les dérivées partielles du coût $L$ par rapport au vecteur d'entrée x.***

$$\frac{\partial L}{\partial x_j} = \sum_{k=1}^{d} \frac{\partial L}{\partial h_k^a} \frac{\partial h_k^a}{\partial x_j} = \sum_{k=1}^{d} \frac{\partial L}{\partial h_k^a} \frac{\partial}{\partial x_j}\left( b_k^{(1)} + \sum_{i=1}^{d} W_{ki}^{(1)} x_i \right) = \sum_{k=1}^{d} \frac{\partial L}{\partial h_k^a} W_{kj}^{(1)}$$

donc $\frac{\partial L}{\partial x} = W^{(1)^T} \frac{\partial L}{\partial h^a}$

## Question 18

Le nouvel objectif à minimiser devient donc:

$$\begin{aligned}
&\widehat{R}\left(f_\theta, D\right)=\widehat{R}\left(f_\theta, D\right)+L(\theta) \\
&=\frac{1}{|D|} \sum_{i=1}^{|D|} L\left(x^{(i)}, y^{(i)}\right)+\lambda_{11}\left\|W^{(1)}\right\|_1+\lambda_{12}\left\|W^{(1)}\right\|_2^2+\lambda_{21}\left\|W^{(2)}\right\|_1+\lambda_{22}\left\|W^{(2)}\right\|_2^2 \\
&=\frac{1}{|D|} \sum_{i=1}^{|D|} L\left(x^{(i)}, y^{(i)}\right)+\lambda_{11} \sum_{i, j}\left|W_{ij}^{(1)}\right|+\lambda_{12} \sum_{i, j}\left(W_{ij}^{(1)}\right)^2+\lambda_{21} \sum_{i, j}\left|W_{ij}^{(2)}\right|+\lambda_{22} \sum_{i, j}\left(W_{ij}^{(2)}\right)^2
\end{aligned}$$

On connait déjà le gradient du terme $\frac{1}{|D|} \sum_{i=1}^{|D|} L\left(x^{(i)}, y^{(i)}\right)$. Il apparaît donc clairement que les changements apportés par la régularisation du risque empirique n'affecteront que les gradients par rapport aux paramètres $W^{(1)}$ et $W^{(2)}$.

$$\frac{\partial}{\partial W_{kj}^{(2)}}\left(\lambda_{21} \sum_{i, j}\left|W_{ij}^{(2)}\right|+\lambda_{22} \sum_{i, j}\left(W_{ij}^{(2)}\right)^2\right)=\lambda_{21}+2 \lambda_{22} W_{kj}^{(2)}$$

donc: $\frac{\partial \widehat{R}}{\partial W_{kj}^{(2)}}=\frac{\partial L}{\partial o_k^a} h_j^s+\lambda_{21}+2 \lambda_{22} W_{kj}^{(2)}$ et donc $\frac{\partial \widehat{R}}{\partial W^{(2)}}=\frac{\partial L}{\partial o^a} h^{s^T}+\lambda_{21}+2 \lambda_{22} W^{(2)}$

Dans la même idée pour $W^{(1)}$:

$$\frac{\partial \widehat{R}}{\partial W^{(1)}}=\frac{\partial L}{\partial h^a} x^T+\lambda_{11}+2 \lambda_{12} W^{(1)}$$