1. Relations et dérivés de quelques fonction de base

#1. MQ $\text{sigmoid}(x) = \frac{1}{2}\left(\tanh\left(\frac{1}{2}\cdot x\right)+1\right)$

$$\text{sigmoid}(x) = \frac{1}{2}\left(\tanh\left(\frac{x}{2}\right)+1\right) \iff 2\,\text{sigmoid}(x) - 1 = \tanh\left(\frac{x}{2}\right)$$

$$2\,\text{sigmoid}(x) - 1 = \frac{2 - 1 - \exp(-x)}{1+\exp(-x)} = \frac{1-\exp(-x)}{1+\exp(-x)} = \frac{\exp(\frac{x}{2})}{\exp(\frac{x}{2})}\left(\frac{1-\exp(-x)}{1+\exp(-x)}\right)$$

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

#2. MQ $\ln \text{sigmoid}(x) = \ln\left(\frac{1}{1+\exp(-x)}\right) = -\ln\left(1+\exp(-x)\right) = -\text{softplus}(-x)$

$$\ln \text{sigmoid}(x) = \ln\left(\frac{1}{1+\exp(-x)}\right) = \ln(1) - \ln\left(1+\exp(-x)\right) = 0 + \ln(1+\exp(-x)) = -\text{softplus}(-x)$$

#3. MQ $\frac{d\,\text{sigmoid}(x)}{dx} = \text{sigmoid}(x)\left(1-\text{sigmoid}(x)\right)$

$$\frac{d\,\text{sigmoid}(x)}{dx} = \frac{d\left((1+\exp(-x))^{-1}\right)}{dx} = \frac{-1}{(1+e^{-x})^2}\left(-e^{-x}\right) = \left(\frac{1}{(1+e^{-x})}\right)\cdot\left(\frac{e^{-x}}{1+e^{-x}}\right)$$

$$= \text{sigmoid}(x)\cdot\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) = \text{sigmoid}(x)\cdot\left(1-\text{sigmoid}(x)\right)$$

#4. MQ $\tanh'(x) = 1 - \tanh^2(x)$

$$\frac{d}{dx}\left(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right) = \frac{(e^{x}-e^{-x})'(e^{x}+e^{-x}) - (e^{x}-e^{-x})(e^{x}+e^{-x})'}{(e^{x}+e^{-x})^2} = \frac{(e^{x}+e^{-x})^2 - (e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2}$$

$$= \frac{(e^{x}+e^{-x})^2}{(e^{x}+e^{-x})^2} - \frac{(e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2} = 1 - \tanh^2(x)$$

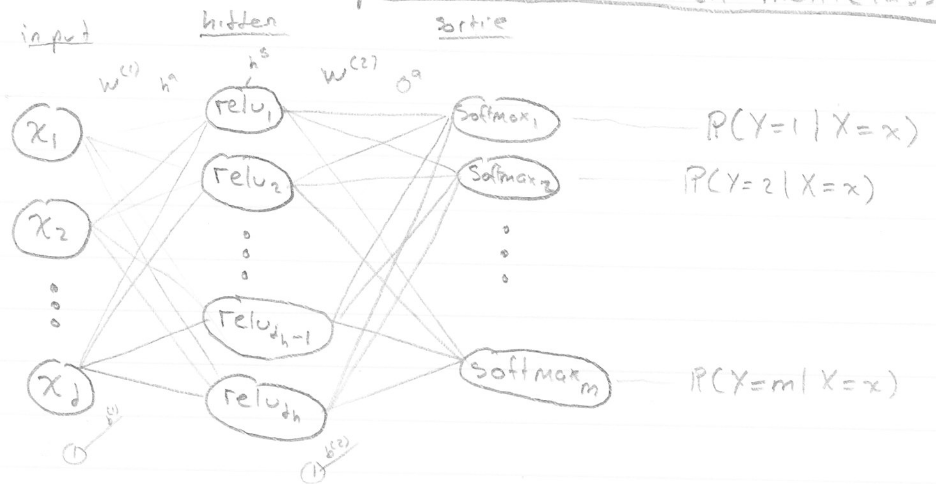
#8. $\|x\|_2^2 = \sum_i x_i^2$. Écrire $\frac{\partial \|x\|_2^2}{\partial x}$

$$\frac{\partial \sum_i x_i^2}{\partial x} = \begin{pmatrix} \frac{\partial(x_1^2 + x_2^2 + \ldots + x_n^2)}{\partial x_1} \\ \vdots \\ \frac{\partial(x_1^2 + x_2^2 + x_n^2)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{pmatrix} = 2 \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = 2x$$

#9. $\|x\|_1 = \sum_i |x_i|$. Écrire $\frac{\partial \|x\|_1}{\partial x}$

$$\frac{\partial \sum_i |x_i|}{\partial x} = \begin{pmatrix} \frac{\partial(|x_1| + |x_2| + \ldots + |x_n|)}{\partial x_1} \\ \vdots \\ \frac{\partial(|x_1| + |x_2| + \ldots + |x_n|)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \operatorname{sign}(x_1) \\ \operatorname{sign}(x_2) \\ \vdots \\ \operatorname{sign}(x_n) \end{pmatrix} = \operatorname{sign}(x)$$

2. Calcul du gradient pour l'optimisation des paramètres d'un réseau de neurones pour la classification multiclasse



#1. $W^{(1)}$ $d_h \times d$, entre couche input et hidden. $b^{(1)} \in \mathbb{R}^?$
Donner la formule de calcul du vecteur de préactivations des neurones de la hidden layer h^a given x as input. first in a matrix form ($h^a = \ldots$). then how to compute one element ($h_j^a = \ldots$)
Écrire le vecteur des sorties des neurones de la hidden layer h^s en fonction de h^a

$b^{(1)} \in \mathbb{R}^{d_h}$

$h^a = b^{(1)} + W^{(1)}x$

$$h^a = \begin{pmatrix} b_1^{(1)} \\ \cdots \\ b_{d_h}^{(1)} \end{pmatrix} + \begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} & \cdots & w_{1d}^{(1)} \\ \vdots & & & \vdots \\ w_{d_h1}^{(1)} & w_{d_h2}^{(1)} & \cdots & w_{d_hd}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ \cdots \\ x_d \end{bmatrix} = \begin{bmatrix} b_1^{(1)} + w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2 + \ldots + w_{1d}^{(1)}x_d \\ \cdots \\ b_{d_h}^{(1)} + w_{d_h1}^{(1)}x_1 + w_{d_h2}^{(1)}x_2 + \cdots + w_{d_hd}^{(1)}x_d \end{bmatrix} \Rightarrow h_j^a = b_j^{(1)} + \sum_{i=1}^{d} w_{ji}^{(1)} x_i$$

$h^s = \text{rect}(h^a)$

#2 Donner dimension de $W^{(2)}$ et $b^{(2)}$ (entre hidden and output).
Donner la formule du vecteur d'activation des neurones de sortie O^a en fonction de h^s, puis O_k^a

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

$O^a = b^{(2)} + W^{(2)} h^s$

$$O^a = \begin{pmatrix} b_1^{(2)} \\ \cdots \\ b_m^{(2)} \end{pmatrix} + \begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & \cdots & w_{1d_h}^{(2)} \\ \vdots & & & \vdots \\ w_{m1}^{(2)} & w_{m2}^{(2)} & \cdots & w_{md_h}^{(2)} \end{bmatrix} \begin{bmatrix} h_1^s \\ \vdots \\ h_{d_h}^s \end{bmatrix} \Rightarrow O_k^a = b_k^{(2)} + \sum_{i=1}^{d_h} w_{ki}^{(2)} h_i^s$$

#3. La sortie des neurones de sortie est $o^s = \text{softmax}(o^a)$. Préciser o^s_k. Démontrer que les o^s_k sont positifs et somment à 1. Pourquoi est-ce important?

$$o^s_k = \text{softmax}(o^a)_k = \frac{\exp(o^a_k)}{\sum_{i=1}^m \exp(o^a_i)}$$

Les o^s_k sont positifs par définitions de $\exp(x)$ (i.e. $\forall x \in \mathbb{R}, \exp(x) > 0$).

$$\sum_{k=1}^m o^s_k = \sum_{k=1}^m \frac{\exp(o^a_k)}{\sum_{i=1}^m \exp(o^a_i)} = \frac{\sum_{k=1}^m \exp(o^a_k)}{\sum_{i=1}^m \exp(o^a_i)} = 1$$

Il est important que les o^s_k soient positif et qu'ils somment à 1, car cela permet d'interpréter o^s_k comme $\mathbb{P}(Y=k \mid X=x)$ (c'est-à-dire qu'on interprète o^s_k comme étant la probabilité que l'entrée x soit de la classe k).

#4. $L(x,y) = -\log(o^s_y(x))$. Préciser L en fonction de o^a

$$L(x,y) = -\log\left(\frac{\exp(o^a_y)}{\sum_{i=1}^m \exp(o^a_i)}\right) = \log\left(\sum_{i=1}^m \exp(o^a_i)\right) - \log(\exp(o^a_y))$$
$$= \log\left(\sum_{i=1}^m \exp(o^a_i)\right) - o^a_y$$

#5. Formuler $\hat{R}$. Indiquer précisement l'ensemble θ des paramètres du réseau. Indiquer à combien de paramètres scalaires n_θ cela correspond. Formuler le problème d'optimisation qui correspond à l'entraînement du réseau pour trouver une valeur optimal des paramètres.

$$\hat{R} = \frac{1}{n} \sum_{i=1}^n L(x^{(i)}, y^{(i)}) = \frac{1}{n} \sum_{i=1}^n \left(\log\left(\sum_{j=1}^m \exp(o^a_j(x^{(i)}))\right) - o^a_{y^{(i)}}(x^{(i)})\right)$$

$$\theta = \{W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\}$$

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

Le problème d'optimisation qui correspond à l'entraînement du réseau permettant de trouver une valeur optimale des paramètres est $\underset{\theta}{\text{argmin}}\, \hat{R}(\theta, D_{train})$

#6. Exprimer avec un bref pseudo-code la descente de gradient pour ce problème

initialize θ

for N itérations :

$$\theta \leftarrow \theta - \eta \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \left(\log \left(\sum_{j=1}^{m} \exp(o_j^a(x^{(i)})) \right) - o^a_{y^{(i)}}(x^{(i)}) \right) \right)$$

#7. MQ $\frac{\partial L}{\partial o^a} = o^s - \text{onehot}_m(y)$

Pour $k \neq y$,

$$\frac{\partial L(x,y)}{\partial o_k^a} = \frac{\partial \left(\log \left(\sum_{j=1}^{m} \exp(o_j^a) \right) - o_y^a \right)}{\partial o_k^a} = \frac{\partial \log \left(\sum_{j=1}^{m} \exp(o_j^a) \right)}{\partial o_k^a}$$

$$= \frac{1}{\sum_{j=1}^{m} \exp(o_j^a)} \cdot \frac{\partial \sum_{j=1}^{m} \exp(o_j^a)}{\partial o_k^a} = \frac{\exp(o_k^a)}{\sum_{j=1}^{m} \exp(o_j^a)} = o_k^s$$

$$\frac{\partial L(x,y)}{\partial o_y^a} = \frac{\partial \left(\log \left(\sum_{j=1}^{m} \exp(o_j^a) \right) - o_y^a \right)}{\partial o_y^a} = \frac{\exp(o_y^a)}{\sum_{j=1}^{m} \exp(o_j^a)} - 1 = o_y^s - 1$$

Alors, $$\frac{\partial L(x,y)}{\partial o^a} = \begin{pmatrix} \frac{\partial L}{\partial o_1^a} \\ \cdots \\ \frac{\partial L}{\partial o_m^a} \end{pmatrix} = \begin{pmatrix} o_1^s \\ \cdots \\ o_y^s \\ \cdots \\ o_m^s \end{pmatrix} - \begin{pmatrix} 0 \\ \cdots \\ 1 \\ \cdots \\ 0 \end{pmatrix} = o^s - \text{onehot}_m(y)$$

#8. Donner l'expression correspondante en numpy

grad_oa = o_s − np.eye(m)[y-1]

car $y \in \{1, \dots, m\}$ et le vecteur onehot_m à des index de 0 à $m-1$

#9. Calculer $\frac{\partial L}{\partial W^{(2)}}$ et $\frac{\partial L}{\partial b^{(2)}}$

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

Pour $k = y$, $\frac{\partial L(x,y)}{\partial W_{yj}^{(2)}} = (o_y^s - 1) \cdot \frac{\partial o_y^a}{\partial W_{yj}^{(2)}} = (o_y^s - 1) \cdot h_j^s = o_y^s h_j^s - h_j^s$

Pour $k \neq y$, $\frac{\partial L(x,y)}{\partial b_k^{(2)}} = o_k^s \cdot \frac{\partial \left(b_k^{(2)} + \sum_{i=1}^{d_h} W_{ki}^{(2)} h_i^s \right)}{\partial b_k^{(2)}} = o_k^s$

Pour $k = y$, $\frac{\partial L(x,y)}{\partial b_y^{(2)}} = o_k^s - 1$

Alors, $$\frac{\partial L}{\partial W^{(2)}} = \begin{bmatrix} o_1^s h_1^s & o_1^s h_2^s & \cdots & o_1^s h_{d_h}^s \\ \vdots & & & \vdots \\ o_m^s h_1^s & & \cdots & o_m^s h_{d_h}^s \end{bmatrix} - \begin{array}{cl} \begin{bmatrix} 0 & & \cdots & 0 \\ \vdots & & & \vdots \\ h_1^s & h_2^s & \cdots & h_{d_h}^s \\ \vdots & & & \vdots \\ 0 & & \cdots & 0 \end{bmatrix} & \begin{array}{l} \\ \\ \leftarrow \text{ligne } y \\ \\ \\ \end{array} \end{array}$$

$$\frac{\partial L}{\partial b^{(2)}} = \begin{pmatrix} o_1^s \\ o_2^s \\ \vdots \\ o_y^s \\ \vdots \\ o_m^s \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = o^s - \text{onehot}_m(y)$$

#10 Donner les expressions correspondantes en numpy

```
grad_b2 = os - np.eye(m)[y-1]
grad_W2 = np.outer(os, hs) - np.concatenate((np.zeros((y-1,dh)), hs.reshape(1,dh), np.zeros((m-y,dh))))
```

grad_b2 est $m \times 1$

grad_W2 est $m \times d_h$

car $\frac{\partial L}{\partial b^{(2)}} = o^s - \text{onehot}_m(y)$ et $$\frac{\partial L}{\partial W^{(2)}} = o^s h^{s^T} - \begin{array}{cl} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ h_1^s & h_2^s & \cdots & h_{d_h}^s \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} & \begin{array}{l} \\ \\ \leftarrow \text{ligne } y \\ \\ \\ \end{array} \end{array}$$

où o^s et $\text{onehot}_m(y)$ sont $m \times 1$, h^s est $d_h \times 1$ et la matrice contenant que des zeros et les éléments de h^s est $m \times d_h$

#11. Calculer $\frac{\partial L}{\partial h^s}$

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

$$= \sum_{k=1}^{m} \frac{\partial L}{\partial o_k^a} \cdot W_{kj}^{(2)} = o_1^s W_{1j}^{(2)} + o_2^s W_{2j}^{(2)} + \ldots + (o_y^s - 1) W_{yj}^{(2)} + \ldots + o_m^s W_{mj}^{(2)}$$

$$= \sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}$$

Alors, $$\frac{\partial L}{\partial h^s} = \begin{pmatrix} \sum_{k=1}^{m} o_k^s W_{k1}^{(2)} - W_{y1}^{(2)} \\ \sum_{k=1}^{m} o_k^s W_{k2}^{(2)} - W_{y2}^{(2)} \\ \cdots \\ \sum_{k=1}^{m} o_k^s W_{kd_h}^{(2)} - W_{yd_h}^{(2)} \end{pmatrix}$$

#12. Exprimer sous forme matricielle, Préciser les dimensions, en numpy grad_hs = ?

$$\frac{\partial L}{\partial h^s} = W^{(2)^T} \cdot o^s - W^{(2)}[y, :]^T$$

où $W^{(2)^T}$ est $d_h \times m$, o^s est $m \times 1$ et $W^{(2)}[y, :]^T$ est $d_h \times 1$

grad_hs = $W2^T$.os − W2[y-1, :]^T

#13. Calculer $\frac{\partial L}{\partial h^a}$

$$\frac{\partial L}{\partial h_j^a} = \frac{\partial L}{\partial h_j^s} \cdot \frac{\partial h_j^s}{\partial h_j^a} = \left(\sum_{k=1}^m o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \cdot \frac{\partial}{\partial h_j^a}\left(\text{rect}(h_j^a)\right) = \left(\sum_{k=1}^m o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)$$

Alors, $$\frac{\partial L}{\partial h^a} = \begin{pmatrix} \left(\sum_{k=1}^m o_k^s W_{k1}^{(2)} - W_{y1}^{(2)}\right) \mathbb{1}_{\{h_1^a > 0\}}(h_1^a) \\ \vdots \\ \left(\sum_{k=1}^m o_k^s W_{kd_h}^{(2)} - W_{yd_h}^{(2)}\right) \mathbb{1}_{\{h_{d_h}^a > 0\}}(h_{d_h}^a) \end{pmatrix}$$

#14. Exprimer sous forme matricielle, Préciser les informations, donner l'équivalent en numpy

$$\frac{\partial L}{\partial h^a} = \left(\frac{\partial L}{\partial h^s}\right) \odot \begin{pmatrix} \mathbb{1}_{\{h_1^a > 0\}}(h_1^a) \\ \cdots \\ \mathbb{1}_{\{h_{d_h}^a > 0\}}(h_{d_h}^a) \end{pmatrix}$$

où $\frac{\partial L}{\partial h^a}$, $\frac{\partial L}{\partial h^s}$ et le vecteur contenant les fonctions indicatrices sont $d_h \times 1$

```
vector_indicator = np.array([1 if e>0 else 0 for e in hs])
grad_ha = np.multiply(grad_hs, vector_indicator)
```

#15. Calculer $\frac{\partial L}{\partial W^{(1)}}$ et $\frac{\partial L}{\partial b^{(1)}}$

$$\frac{\partial L}{\partial W_{j\ell}^{(1)}} = \frac{\partial L}{\partial h_j^a} \cdot \frac{\partial h_j^a}{\partial W_{j\ell}^{(1)}} = \left(\sum_{k=1}^{m} O_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a) \cdot \frac{\partial \left(b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i\right)}{\partial W_{j\ell}^{(1)}}$$

$$= \left(\sum_{k=1}^{m} O_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)\, x_\ell = \frac{\partial L}{\partial h_j^a} \cdot x_\ell$$

$$\frac{\partial L}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a} \cdot \frac{\partial h_j^a}{\partial b_j^{(1)}} = \frac{\partial L}{\partial h_j^a} \cdot 1 = \frac{\partial L}{\partial h_j^a}$$

$$\frac{\partial L}{\partial W^{(1)}} = \begin{bmatrix} \frac{\partial L}{\partial h_1^a} \cdot x_1 & \frac{\partial L}{\partial h_1^a} \cdot x_2 & \cdots & \frac{\partial L}{\partial h_1^a} \cdot x_d \\ \frac{\partial L}{\partial h_2^a} \cdot x_1 & \frac{\partial L}{\partial h_2^a} \cdot x_2 & \cdots & \frac{\partial L}{\partial h_2^a} \cdot x_d \\ \cdots & \cdots & & \cdots \\ \frac{\partial L}{\partial h_{d_h}} \cdot x_1 & \frac{\partial L}{\partial h_{d_h}} \cdot x_2 & \cdots & \frac{\partial L}{\partial h_{d_h}} \cdot x_d \end{bmatrix}$$

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

#16. Exprimer sous forme matricielle, définir les dimensions, donner l'équivalent en numpy

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

$\frac{\partial L}{\partial W^{(1)}} = \frac{\partial L}{\partial h^a} \cdot x^T$ est $d_h \times d$ car $\frac{\partial L}{\partial h^a}$ est $d_h \times 1$ et x^T est $1 \times d$

grad_b1 = grad_ha

grad_W1 = np.outer(grad_ha, x)

#17 Calculer $\frac{\partial L}{\partial x}$

$$\frac{\partial L}{\partial x_\ell} = \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \frac{\partial h_j^a}{\partial x_\ell} = \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \frac{\partial \left(b_j^{(1)} + \sum_{i=1}^{d} W_{ji}^{(1)} x_i\right)}{\partial x_\ell}$$

$$= \sum_{j=1}^{d_h} \left(\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)\, W_{j\ell}^{(1)}$$

$$\text{Alors, } \frac{\partial L}{\partial x} = \begin{pmatrix} \sum_{j=1}^{d_h} \left(\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)\, W_{j1}^{(1)} \\ \vdots \\ \sum_{j=1}^{d_h} \left(\sum_{k=1}^{m} o_k^s W_{kj}^{(2)} - W_{yj}^{(2)}\right) \mathbb{1}_{\{h_j^a > 0\}}(h_j^a)\, W_{jd}^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \cdot W_{j1}^{(1)} \\ \vdots \\ \sum_{j=1}^{d_h} \frac{\partial L}{\partial h_j^a} \cdot W_{jd}^{(1)} \end{pmatrix}$$

#18. Comment minimiser $\tilde{R} = \hat{R} + \lambda_{11}\left(\sum_{i,j} |W_{ij}^{(1)}|\right) + \lambda_{12}\left(\sum_{i,j} (W_{ij}^{(1)})^2\right) + \lambda_{21}\left(\sum_{i,j} |W_{ij}^{(2)}|\right) + \lambda_{22}\left(\sum_{i,j} (W_{ij}^{(2)})^2\right)$ au lieu de $\hat{R}$ change le gradient par rapport aux différents paramètres?

Il n'y a pas de différence pour $b^{(1)}$ et $b^{(2)}$ car $\frac{\partial \mathcal{L}(\theta)}{\partial b^{(1)}} = \frac{\partial \mathcal{L}(\theta)}{\partial b^{(2)}} = 0$

$$\frac{\partial \mathcal{L}(\theta)}{\partial W^{(1)}} = \lambda_{11} \begin{bmatrix} \text{sign}(W_{11}^{(1)}) & \text{sign}(W_{12}^{(1)}) & \cdots & \text{sign}(W_{1d}^{(1)}) \\ \text{sign}(W_{21}^{(1)}) & \cdots & & \vdots \\ \vdots & & & \vdots \\ \text{sign}(W_{d_h 1}^{(1)}) & \text{sign}(W_{d_h 2}^{(1)}) & \cdots & \text{sign}(W_{d_h d}^{(1)}) \end{bmatrix} + \lambda_{12} \begin{bmatrix} 2W_{11}^{(1)} & 2W_{12}^{(1)} & \cdots & 2W_{1d}^{(1)} \\ \vdots & & & \\ \vdots & & & \\ 2W_{d_h 1}^{(1)} & & \cdots & 2W_{d_h d}^{(1)} \end{bmatrix}$$

$$= \lambda_{11}\, \text{sign}(W^{(1)}) + 2\lambda_{12} W^{(1)}$$

$$\text{Alors, } \frac{\partial \tilde{R}}{\partial W^{(1)}} = \frac{\partial \hat{R}}{\partial W^{(1)}} + \lambda_{11}\, \text{sign}(W^{(1)}) + 2\lambda_{12} W^{(1)}$$

et de la même façon, $\frac{\partial \tilde{R}}{\partial W^{(2)}} = \frac{\partial \hat{R}}{\partial W^{(2)}} + \lambda_{21}\, \text{sign}(W^{(2)}) + 2\lambda_{22} W^{(2)}$