

# Théorie mathématique pour les forwardpropagation et backpropagation

Alexandre AZOR

30 mai 2022

## 1 Notation

### 1.1 Entiers

- $p$  : Nombre de couches (Autres que la couche d'entrée)
- $k, l$  : Indices de couche,  $k \in \llbracket 0, p \rrbracket$ ,  $l \in \llbracket 0, p \rrbracket$ 
  - Couche d'entrée :  $k = 0$
  - Couche de sortie :  $k = p$
- $s_k$  : Nombre de neurones de la couche  $k$

### 1.2 Réel

- $\alpha \in \mathbb{R}^{+*}$  : Learning rate

### 1.3 Matrices

- $A^T$  : La transposée de la matrice  $A$
- $0_{n,m}$  : La matrice  $\mathbb{R}^{n \times m}$  ( $n$  lignes et  $m$  colonnes) ne contenant que des 0

- $E_{i,j}^{n \times m}$  : La matrice de  $\mathbb{R}^{n \times m}$  dont le coefficient ligne  $i$  colonne  $j$  est un 1, les autres étant des 0
  - $w^{(k)} \in \mathbb{R}^{s_{k+1} \times s_k}, k \in \llbracket 0, p-1 \rrbracket$  : Poids pour passer de la couche  $k$  à la couche  $k+1$   
 $w_{i,j}^{(k)}$  : Coefficient ligne  $i$  colonne  $j$
- On note  $AB$  le produit matriciel usuel entre 2 matrices ou vecteurs  $A$  et  $B$  de tailles compatibles
- $[a_{i,j}]_{\substack{i \in \llbracket 1, n \rrbracket \\ j \in \llbracket 1, m \rrbracket}}$  désigne la matrice de  $n$  lignes et  $m$  colonnes dont le coefficient ligne  $i$  colonne  $j$  est  $a_{i,j}$

## 1.4 Vecteurs

- $0_n$  : Le vecteur de  $\mathbb{R}^n$  n'ayant que des 0
- $E_i^n$  : Le vecteur de  $\mathbb{R}^n$  ayant un 1 sur la  $i$ -ième ligne et des 0 partout ailleurs
- $x \in \mathbb{R}^{s_0}$  : Valeur d'entrée
- $y \in \mathbb{R}^{s_p}$  : Valeur attendue
- $y_{\text{pred}} \in \mathbb{R}^{s_p}$  : Valeur de sortie
- $b^{(k)} \in \mathbb{R}^{s_{k+1}}, k \in \llbracket 0, p-1 \rrbracket$  : Biais pour passer de la couche  $k$  à la couche  $k+1$   
 $b_i^{(k)}$  : Coefficient ligne  $i$
- $z^{(k)}, k \in \llbracket 1, p \rrbracket$  : Valeurs des différents neurones de la couche  $k$  avant activation  
 $z_i^{(k)}$  : Coefficient ligne  $i$ , Valeur du  $i$ -ième neurone de la couche  $k$  avant activation  
 $z^{(0)}$  n'est pas défini  
 $z^{(p)} = y_{\text{pred}}$

- $a^{(k)}, k \in \llbracket 0, p-1 \rrbracket$  : Valeurs des différents neurones de la couche  $k$  après activation  
 $a_i^{(k)}$  : Coefficient ligne  $i$ , Valeur du  $i$ -ième neurone de la couche  $k$  après activation  
 $a^{(0)} = x$   
 $a^{(p)}$  n'est pas défini

## 1.5 Fonctions

- $diag$  : Fonction qui à un vecteur  $v$  de  $\mathbb{R}^n$  associe la matrice de  $\mathbb{R}^{n \times n}$  dont le coefficient ligne  $i$  colonne  $i$  est  $v_i$  pour  $i \in \llbracket 1, n \rrbracket$  et tous les autres sont nuls
- $g : \mathbb{R} \rightarrow \mathbb{R}$  : Fonction d'activation

On désigne également par  $g$  l'application qui à un vecteur  $u$  de  $\mathbb{R}^n$  associe  $[g(u_i)]_{i \in \llbracket 1, n \rrbracket}$   
De même,  $g'(u) = [g'(u_i)]_{i \in \llbracket 1, n \rrbracket}$

$$h_k : \begin{cases} \mathbb{R}^{s_k} \times \mathbb{R}^{s_{k+1} \times s_k} \times \mathbb{R}^{s_k} & \longrightarrow & \mathbb{R}^{s_{k+1}} \\ (b^{(k)}, w^{(k)}, a^{(k)}) & \longmapsto & b^{(k)} + w^{(k)} a^{(k)} \end{cases}$$

( $k \in \llbracket 0, p-1 \rrbracket$ )

Fonction permettant de passer de  $a^{(k)}$  à  $z^{(k+1)}$

- $L : \mathbb{R}^{s_p} \rightarrow (\mathbb{R}^{s_p} \rightarrow \mathbb{R}^+)$  : Fonction coût ayant pour paramètres la valeur attendue et la valeur prédite  
 $L(y)$  est la fonction mesurant l'écart avec la valeur  $y$
- $\tilde{L}_k(k \in \llbracket 0, p \rrbracket)$  : Fonction coût ayant pour paramètres :
  - ◇  $y$  : la valeur attendue
  - ◇  $(b_i^{(l)})_{\substack{l \in \llbracket k, p-1 \rrbracket \\ i \in \llbracket l, s_{k+1} \rrbracket}}$

$$\begin{aligned} & \diamond \left( w_{i,j}^{(l)} \right)_{\substack{l \in \llbracket k, p-1 \rrbracket \\ i \in \llbracket 1, s_{k+1} \rrbracket \\ j \in \llbracket 1, s_k \rrbracket}} \\ & \diamond z^{(k)} \text{ si } k > 0, x \text{ sinon} \end{aligned}$$

Pour alléger les notations, ces paramètres seront simplement désignés par (...)

$$\tilde{L}_p(y, y_{\text{pred}}) = L(y)(y_{\text{pred}})$$

$$\tilde{L}_{p-1}(\dots) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g\left(z^{(p-1)}\right)\right)\right)$$

$$\tilde{L}_k(\dots) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g \circ h_{p-2}\left(b^{(p-2)}, w^{(p-2)}, g \circ h_{p-3}\left(\dots b^{(k)}, w^{(k)}, g\left(z^{(k)}\right)\right)\right)\right)\right)$$

$$\tilde{L}_0(\dots) = L\left(y, h_{p-1}\left(b^{(p-1)}, w^{(p-1)}, g \circ h_{p-2}\left(b^{(p-2)}, w^{(p-2)}, g \circ h_{p-3}\left(\dots b^{(0)}, w^{(0)}, x\right)\right)\right)\right)$$

## 2 Forwardpropagation

On note  $x \in \mathbb{R}^{s_0}$  le vecteur donné en entrée du réseau

$$\begin{aligned} a^{(0)} & \leftarrow x \\ z^{(1)} & \leftarrow w^{(0)} a^{(0)} + b^{(0)} \end{aligned}$$

Pour  $k$  allant de 1 à  $p-1$  inclus :

$$\begin{aligned} a^{(k)} & \leftarrow g\left(z^{(k)}\right) \\ z^{(k+1)} & \leftarrow w^{(k)} a^{(k)} + b^{(k)} \end{aligned}$$

On renvoie alors  $z^{(p)}$

## 3 Backpropagation

### 3.1 Objectif et principe

On veut minimiser la valeur moyenne de la fonction coût  $\tilde{L}_0$  appliquée avec les couples  $(x, y)$  de la base d'apprentissage.

On utilise une méthode de descente de gradient : Pour chaque couple  $(x, y)$  de la base d'apprentissage :

- On calcule  $y_{\text{pred}}$  en enregistrant les valeurs des  $a^{(k)}$
- $\forall k \in \llbracket 0, p-1 \rrbracket$

$$\forall (i, j) \in \llbracket 1, s_{k+1} \rrbracket \times \llbracket 1, s_k \rrbracket \quad w_{i,j}^{(k)} \leftarrow w_{i,j}^{(k)} - \alpha \frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}}$$

$$\forall i \in \llbracket 1, s_{k+1} \rrbracket \quad b_i^{(k)} \leftarrow b_i^{(k)} - \alpha \frac{\partial \tilde{L}_0}{\partial b_i^{(k)}}$$

En posant

$$dW^{(k)} \triangleq \left[ \frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}} \right]_{(i,j) \in \llbracket 1, s_{k+1} \rrbracket \times \llbracket 1, s_k \rrbracket}$$

$$dB^{(k)} \triangleq \left[ \frac{\partial \tilde{L}_0}{\partial b_i^{(k)}} \right]_{i \in \llbracket 1, s_k \rrbracket}$$

les formules précédents se réécrivent :

$$w^{(k)} \leftarrow w^{(k)} - \alpha dW$$

$$b^{(k)} \leftarrow b^{(k)} - \alpha dB$$

### 3.2 $dL^{(k)}$

On pose  $dL^{(k)}$ , le vecteur de  $\mathbb{R}^{s_k}$  tel que :

$$D\tilde{L}_k(\dots)(0, \dots, 0, u) = dL^{(k)} \cdot u \Leftrightarrow \frac{\partial \tilde{L}_k}{\partial u} = dL^{(k)} \cdot u$$

On construit la suite  $\left( dL^{(k)} \left( z^{(k)} \right) \right)_{k \in \llbracket 1, p \rrbracket}$  par ordre décroissant des indices.

- $k = p$   
 $dL^{(p)}$  est l'unique vecteur vérifiant pour tout  $u \in \mathbb{R}^{s_p}$  :

$$D\tilde{L}_p(y, z^{(p)})(0, u) = dL^{(p)T} u$$

$$\begin{aligned} \tilde{L}_p(y, u) &= L(y)(u) \\ \Rightarrow D\tilde{L}_p(y, z^{(p)})(0, u) &= D(L(y))(z^{(p)})(u) = \vec{\nabla}(L(y))(z^{(p)})^T u \\ &\Rightarrow \boxed{dL^{(p)} = \vec{\nabla}L(y)(z^{(p)})} \end{aligned}$$

- $0 < k < p$   
On suppose  $dL^{(k+1)}$  connu avec  $z^{(k+1)} = h_k(b^{(k)}, w^{(k)}, g(z^{(k)}))$ .

$$\tilde{L}_k(\dots, b^{(k)}, w^{(k)}, z^{(k)}) = \tilde{L}_{k+1}(\dots, h(b^{(k)}, w^{(k)}, g(z^{(k)})))$$

Avec la règle de la chaîne, on obtient pour  $u \in \mathbb{R}^{s_k}$  :

$$\begin{aligned} &D\tilde{L}_k(\dots, b^{(k)}, w^{(k)}, z^{(k)})(\dots, 0_{s_k}, 0_{s_{k+1} \times s_k}, u) \\ &= D\tilde{L}_{k+1}(\dots)(\dots, Dh(b^{(k)}, w^{(k)}, g(z^{(k)}))(0_{s_k}, 0_{s_{k+1} \times s_k}, Dg(z^{(k)}(u)))) \\ &Dg(v)(u) = \text{diag}(g'(v))u \\ &Dh(b^{(k)}, w^{(k)}, g(z^{(k)}))(0_{s_k}, 0_{s_{k+1} \times s_k}, q) = w^{(k)}q \\ &D\tilde{L}_{k+1}(\dots, h(b^{(k)}, w^{(k)}, z^{(k)}))(\dots, r) = dL^{(k+1)T} r \\ &\Rightarrow \frac{\partial \tilde{L}_k}{\partial u}(\dots) = dL^{(k+1)T} w^{(k)} \text{diag}(g'(u))u \\ &\Rightarrow \boxed{dL^{(k)}(z^{(k)}) = \text{diag}(g'(z^{(k)}))w^{(k)T} dL^{(k+1)}} \end{aligned}$$

### 3.3 $dB^{(k)}$

Pour  $k$  allant de  $p-1$  à  $0$

$$\tilde{L}_0 (... , b^{(k)}, w^{(k)}, ...) = \tilde{L}_k (... , b^{(k)}, w^{(k)}, z^{(k)})$$

$$\text{Avec } z^{(k)} = h_{k-1} (b^{(k-1)}, w^{(k-1)}, g \circ h_{k-2} (...))$$

$$\tilde{L}_k (... , b^{(k)}, w^{(k)}, z^{(k)}) = \tilde{L}_{k+1} (... , h (b^{(k)}, w^{(k)}, g (z^{(k)})))$$

$$\Rightarrow \frac{\partial \tilde{L}_0}{\partial b_i^{(k)}} = \frac{\partial \tilde{L}_k}{\partial b_i^{(k)}} =$$

$$= D\tilde{L}_{k+1} (...) (... , Dh (b^{(k)}, w^{(k)}, g (z^{(k)}))) (E_i^{s_{k+1}}, 0_{s_{k+1} \times s_k}, 0_{s_k})$$

$$= dL^{(k+1) \ T} E_i^{s_{k+1}}$$

$$= (dL^{(k+1)})_i$$

$$\Rightarrow \boxed{dB^{(k)} = dL^{(k+1)}}$$

### 3.4 $dW^{(k)}$

— Pour  $k$  allant de  $p-1$  à  $1$

$$\tilde{L}_0 (... , b^{(k)}, w^{(k)}, ...) = \tilde{L}_k (... , b^{(k)}, w^{(k)}, z^{(k)})$$

$$\text{Avec } z^{(k)} = h_{k-1} (b^{(k-1)}, w^{(k-1)}, g \circ h_{k-2} (...))$$

$$\tilde{L}_k (... , b^{(k)}, w^{(k)}, z^{(k)}) = \tilde{L}_{k+1} (... , h (b^{(k)}, w^{(k)}, g (z^{(k)})))$$

$$\Rightarrow \frac{\partial \tilde{L}_0}{\partial w_{i,j}^{(k)}} = \frac{\partial \tilde{L}_k}{\partial w_{i,j}^{(k)}} =$$

$$= D\tilde{L}_{k+1}(\dots) \left( \dots, Dh \left( b^{(k)}, w^{(k)}, g \left( z^{(k)} \right) \right) \left( 0_{s_{k+1}}, E_{i,j}^{s_{k+1} \times s_k}, 0_{s_k} \right) \right)$$

$$= dL^{(k+1) \ T} g \left( z_j^{(k)} \right) E_i^{s_{k+1}}$$

$$= g \left( z_j^{(k)} \right) \left( dL^{(k+1)} \right)_i$$

$$= \left[ dL^{(k+1)} g \left( z^{(k)} \right)^T \right]_{i,j}$$

$$\Rightarrow \boxed{dW^{(k)} = \left[ dL^{(k+1)} g \left( z^{(k)} \right)^T \right]}$$

— Pour  $k = 0$

$$\Rightarrow \boxed{dW^{(0)} = \left[ dL^{(k+1)} x^T \right]}$$