Introduction au Deep LEarning

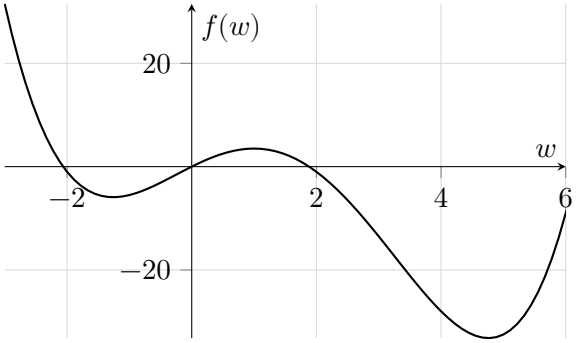
Réseaux de neurones et rétropropagation

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

2025-2026

Exercice 1 - *Rappel sur la descente de gradient*

On souhaite minimiser la fonction $f(w) = 0.25w^4 - 1.5w^3 - 1.25w^2 + 6 * w$ par descente de gradient.



Question 1 - *Descente de gradient*

On part de $w^{(0)} = 2$ et on utilise la règle de mise à jour : $w^{(t+1)} = w^{(t)} - \alpha \times f'(w^{(t)})$

avec un learning rate $\alpha = 0.05$.

a. Donner la dérivée f' de la fonction f.
b. Calculer les 3 premières itérations : $w^{(1)}$, $w^{(2)}$, $w^{(3)}$.
c. Vérifier la convergence en calculant les valeurs successives de la fonction.

Question 2 - *Influence du learning rate*

a. Avec $\alpha = 0.5$, calculer $w^{(1)}$ et $w^{(2)}$. Que se passe-t-il ?
b. Avec $\alpha = 0.01$, calculer $w^{(1)}$. Comment évolue la vitesse de convergence ?
c. Comment choisir le learning rate ?

Question 3 - *Influence de l'initialisation*

Sans faire les calculs, en analysant la courbe.

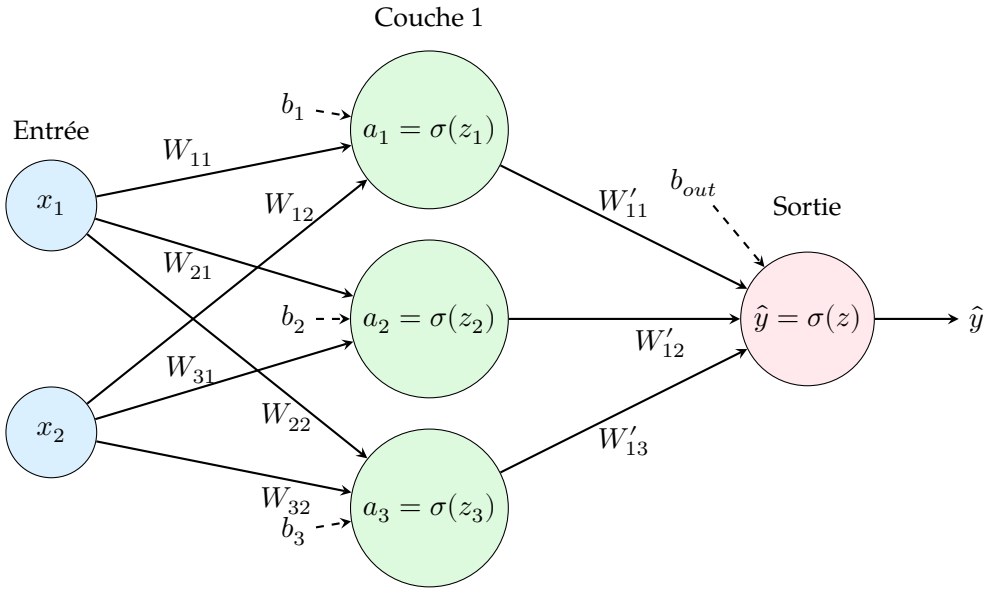
a. Si on initialise à $w^{(0)} = 6$, vers quel minimum converge-t-on avec un petit learning rate ? Même question avec un grand ?
b. Même question pour $w^{(0)} = -2$
c. Conclure pour les réseaux de neurones.

Exercice 2 - *Forward et backward avec une couche linéaire*

Soit un réseau de neurones à 2 couches :

Architecture :

- Entrée : $\mathbf{x} = [x_1, x_2]^T$ (vecteur colonne 2×1)
- Couche cachée : 3 neurones avec biais et activation sigmoïde
- Sortie : 1 neurone avec biais et activation sigmoïde
- Fonction d'activation : $\sigma(z) = \frac{1}{1+e^{-z}}$ (appliquée élément par élément)
- Fonction de perte : $L = \frac{1}{2}(\hat{y} - y)^2$ (MSE)



Question 4 - *Passe forward*

a. Écrire les équations de la passe avant **sans notation matricielle** en développant pour chaque neurone :
   - Calcul des $z_i^{(1)}$ (pré-activations de la couche cachée)
   - Calcul des $a_i^{(1)} = \sigma(z_i^{(1)})$ (activations de la couche cachée)
   - Calcul de $z^{(2)}$ (pré-activation de sortie)
   - Calcul de $\hat{y} = \sigma(z^{(2)})$ (sortie du réseau)
   - Calcul de la loss $L = \frac{1}{2}(\hat{y} - y)^2$

b. Écrire les mêmes équations sous forme matricielle.

Question 5 - *Paramètres du réseau*

a. Identifier les hyperparamètres et les paramètres de ce réseau.
b. Compter le nombre total de paramètres apprenables.
c. Donner une formule générale du nombre de paramètres pour un MLP à p couches cachées.

Question 6 - *Passe backward : couche de sortie*

a. Calculer la dérivée de la loss par rapport à la sortie du réseau :

$$\frac{\partial L}{\partial \hat{y}}$$

b. En déduire la dérivée par rapport à la pré-activation de sortie en appliquant la règle de la chaîne :

$$\frac{\partial L}{\partial z^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{(2)}}$$

Remarque : $\frac{\partial \hat{y}}{\partial z^{(2)}} = \sigma'(z^{(2)}) = \sigma(z^{(2)})(1 - \sigma(z^{(2)})) = \hat{y}(1 - \hat{y})$

c. En déduire les gradients par rapport aux paramètres de la couche de sortie :

— $\frac{\partial L}{\partial b'}$ (gradient du biais de sortie)
— $\frac{\partial L}{\partial \mathbf{W}'}$ (gradient des poids de sortie)

Expliciter chaque composante de $\frac{\partial L}{\partial \mathbf{W}'} = \begin{bmatrix} \frac{\partial L}{\partial W'_1} & \frac{\partial L}{\partial W'_2} & \frac{\partial L}{\partial W'_3} \end{bmatrix}$

Question 7 - *Passe backward : couche cachée*

a. Propager le gradient vers les activations de la couche cachée :

$$\frac{\partial L}{\partial \mathbf{a}^{(1)}}$$

Expliciter chaque composante $\frac{\partial L}{\partial a_j^{(1)}} = \frac{\partial L}{\partial z^{(2)}} \cdot W'_j$

b. En déduire le gradient par rapport aux pré-activations de la couche cachée :

$$\frac{\partial L}{\partial \mathbf{z}^{(1)}}$$

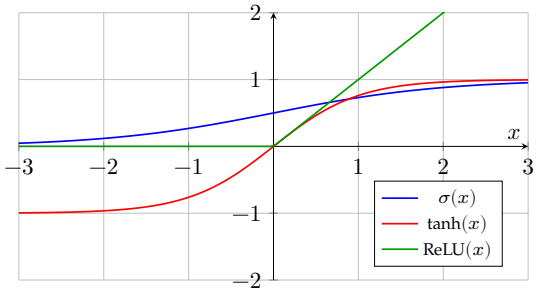
c. En déduire les gradients par rapport aux paramètres de la couche cachée :

— $\frac{\partial L}{\partial \mathbf{b}^{(1)}}$ (gradient des biais)
— $\frac{\partial L}{\partial \mathbf{W}^{(1)}}$ (gradient des poids)

Expliciter chaque composante de la matrice $\frac{\partial L}{\partial \mathbf{W}^{(1)}} \in \mathbb{R}^{3\times 2}$

Exercice 3 - *Fonctions d'activation*

Les fonctions d'activation introduisent de la non-linéarité dans les réseaux de neurones. Étudions trois fonctions classiques :



Définitions :

— **Sigmoid :** $\sigma(x) = \dfrac{1}{1+e^{-x}}$ (sortie dans $[0, 1]$)
— **Tanh :** $\tanh(x) = \dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ (sortie dans $[-1, 1]$)
— **ReLU :** $\mathrm{ReLU}(x) = \max(0, x)$ (sortie dans $[0, +\infty[$)

Question 8 - *Importance des activations*

a. Dans l'architecture précédente, que devient la passe avant si l'on supprime la fonction d'activation de la couche cachée ?
b. Quel est l'intérêt de l'activation finale ?
c. Conclure sur l'intérêt des activations.

Question 9 - *Comportement*

a. Pour chaque fonction d'activation, donner :
   — Image (ensemble des valeurs possibles)
   — Dérivée
   — Comportement à l'infini
b. Pourquoi la saturation pose-t-elle problème dans les réseaux profonds lors de la rétropropagation ?

Question 10 - *Vanishing gradient*

On considère un réseau à K couches cachées avec activation sigmoid.

a. Si toutes les activations d'une couche sont proches de 0 (saturation à gauche), que vaut la dérivée $\sigma'(z) = \sigma(z)(1 - \sigma(z))$?
b. Que devient le gradient après K couches si chaque dérivée locale est de l'ordre de 10^{-3} ? Conclure.
c. La fonction ReLU a-t-elle le même problème ? Donner les valeurs possibles de sa dérivée.

Exercice 4 - *Autograd et graphe de calcul*

Considérons l'expression : $L = (w_1 x + w_2 y)^2$

Question 11 - *Graphe de calcul*

a. Dessiner le graphe de calcul avec les opérations élémentaires :
   — $a = w_1 x$
   — $b = w_2 y$
   — $c = a + b$
   — $L = c^2$
b. En utilisant la règle de dérivation en chaîne, calculer $\frac{\partial L}{\partial w_1}$ et $\frac{\partial L}{\partial w_2}$.

Question 12 - *Autograd*

a. Expliquer comment calculer les dérivées automatiquement.
b. Conclure sur l'importance de ce mécanisme.