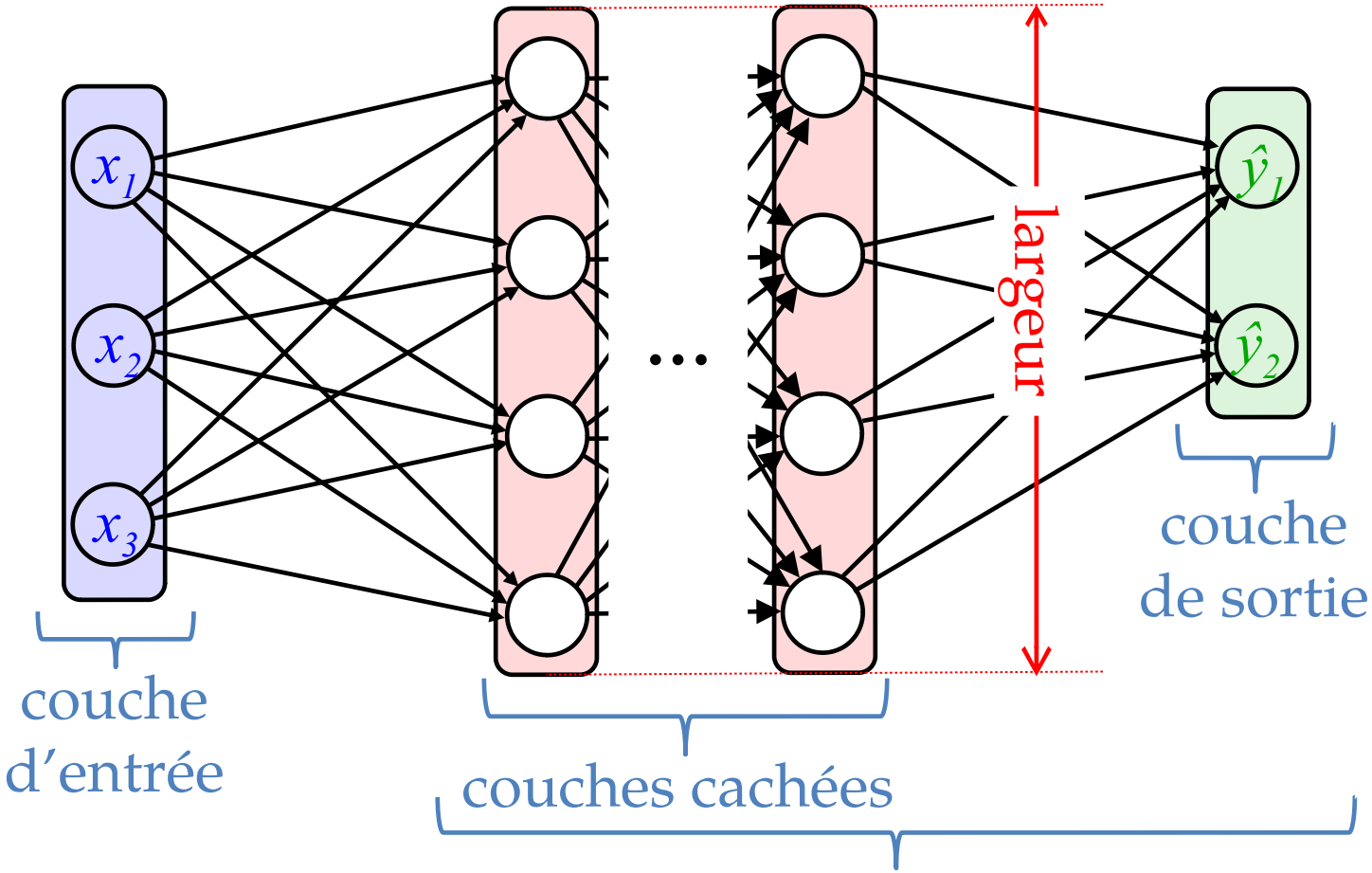
INTRODUCTION AUX RÉSEAUX DE NEURONES

Rétropropagation du gradient

Pascal Germain*, 2019

* Merci spécial à Philippe Giguère pour m'avoir permis de réutiliser une partie de ces transparents.

Illustration et nomenclature



$$\hat{y} = f^{(3)}\left(f^{(2)}\left(f^{(1)}(x)\right)\right)$$

Choix à faire

- Architecture
  - # couches
  - # neurones (cachés) par couche
  - type de couche
- Forme de la sortie & fonction de sortie
- Fonction de perte
- Optimiseur
  - et autres « *détails* »

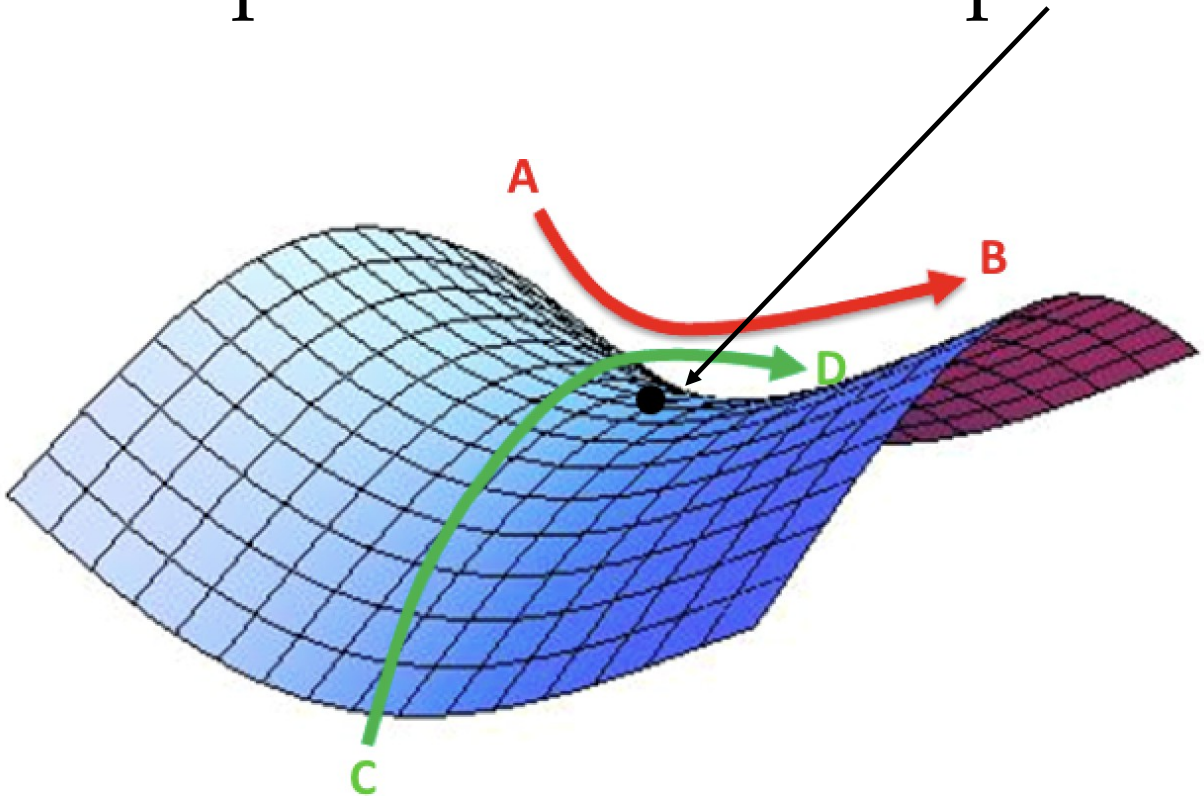
Comparaison avec les méthodes «classiques»

- Beaucoup de méthodes d'apprentissage sont convexes
  - Moindre carrés
  - Regression logistique
  - SVM
- Les réseaux de neurones sont non-convexes
  - Abandon de garanties théoriques
  - Le résultat varie selon l'initialisation de la descente en gradient.
  - On doit accepter que les minimums locaux peuvent être de bonnes solutions.
  - La recherche montre que les solutions sont en fait souvent points de selle
    - ratio (points de selle)/(minimum locaux) augmente exponentiellement avec nombre paramètres à optimiser

Voir Goodfellow et al. Section 8.2.3

Exemple point de selle

- Dérivées partielles nulles au point de selle



https://www.safaribooksonline.com/library/view/fundamentals-of-deep/9781491925607/ch04.html

Profil de la fonction de perte

Adapté de cs231n

Algorithme de rétropropagation («*backprop*»)

Règle de dérivation en chaîne

$$\frac{\partial f(h(x))}{\partial x} = \frac{\partial f(h(x))}{\partial h(x)} \frac{\partial h(x)}{\partial x}$$
$$= \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

Par exemple: $F(x) = (2x+3)^2$

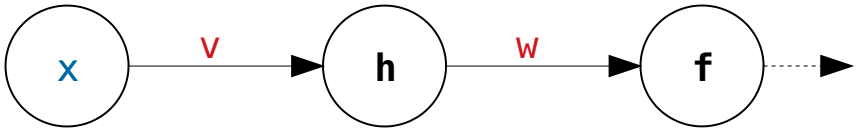
$= f(2x+3)$ où $f(x) = x^2$

$= f(h(x))$ où $h(x) = 2x+3.$

Donc : $\frac{\partial F(x)}{\partial x} = \frac{\partial f(h(x))}{\partial h(x)} \frac{\partial h(x)}{\partial x}$

$= 2\,h(x) \times 2$

$= 4(2x+3)$



$$R_{v,w}(x) = f(w \cdot h(v \cdot x))$$

$$\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

$$\frac{\partial L(R_{v,w}(x), y)}{\partial w} = \left[\frac{\partial L(r, y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \frac{\partial R_{v,w}(x)}{\partial w}$$

Règle de dérivation en chaîne

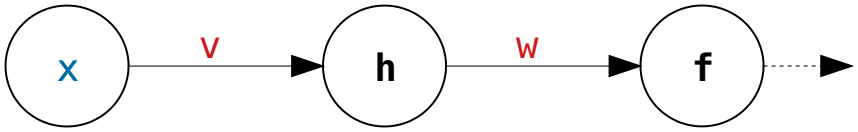
$$\frac{\partial f(h(x))}{\partial x} = \frac{\partial f(h(x))}{\partial h(x)} \frac{\partial h(x)}{\partial x}$$
$$= \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

On écrit aussi:

$$(f \circ h)' = (f' \circ h) \times h'$$

$$(\ f(h(x))\ )' = f'(h(x)) \times h'(x)$$

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h} \frac{\partial h}{\partial x}$$

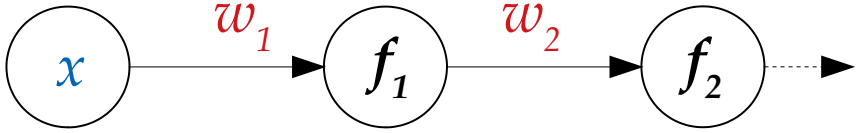


$$R_{v,w}(x) = f(w \cdot h(v \cdot x))$$

$$\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

$$\begin{aligned}
\frac{\partial L(R_{v,w}(x), y)}{\partial w} &= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \frac{\partial R_{v,w}(x)}{\partial w} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot \frac{\partial w \cdot h(v \cdot x)}{\partial w} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot h(v \cdot x)
\end{aligned}$$

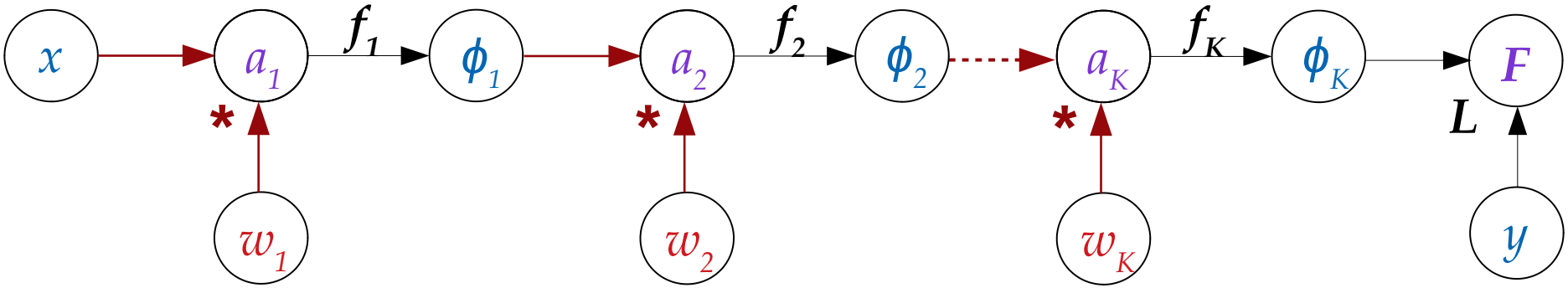
$$\begin{aligned}
\frac{\partial L(R_{v,w}(x), y)}{\partial v} &= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \frac{\partial R_{v,w}(x)}{\partial v} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot \frac{\partial w \cdot h(v \cdot x)}{\partial v} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot w \cdot \frac{\partial h(v \cdot x)}{\partial v} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot w \cdot \left[\frac{\partial h(b)}{\partial b}\right]_{b=v\cdot x} \cdot \frac{\partial v \cdot x}{\partial v} \\
&= \left[\frac{\partial L(r,y)}{\partial r}\right]_{r=R_{v,w}(x,y)} \cdot \left[\frac{\partial f(a)}{\partial a}\right]_{a=w\cdot h(v\cdot x)} \cdot w \cdot \left[\frac{\partial h(b)}{\partial b}\right]_{b=v\cdot x} \cdot x
\end{aligned}$$



$$\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

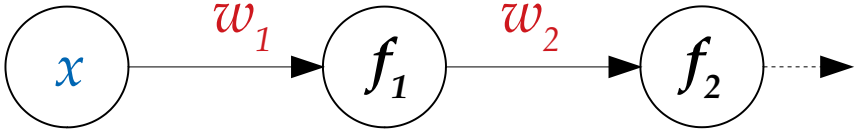
$$R(x) = f_2(w_2 \cdot f_1(w_1 \cdot x))$$

$$F = L\Big(R(x), y\Big)$$



Étape 1: Propagation avant

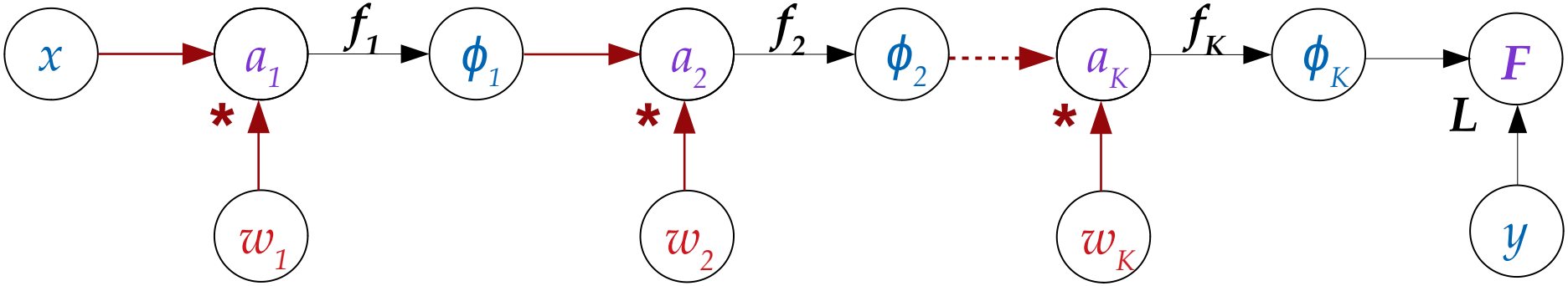
Étape 2: Rétropropagation du gradient



$$\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$

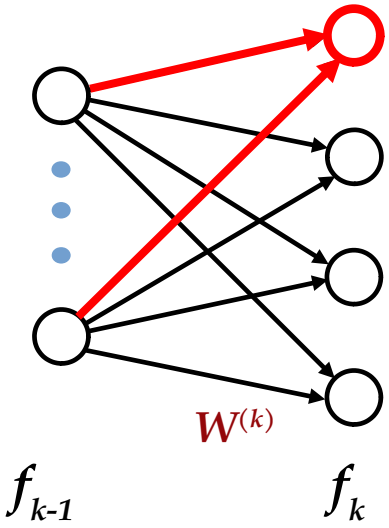
$$R(x) = f_2(w_2 \cdot f_1(w_1 \cdot x))$$

$$F = L\Big(R(x), y\Big)$$

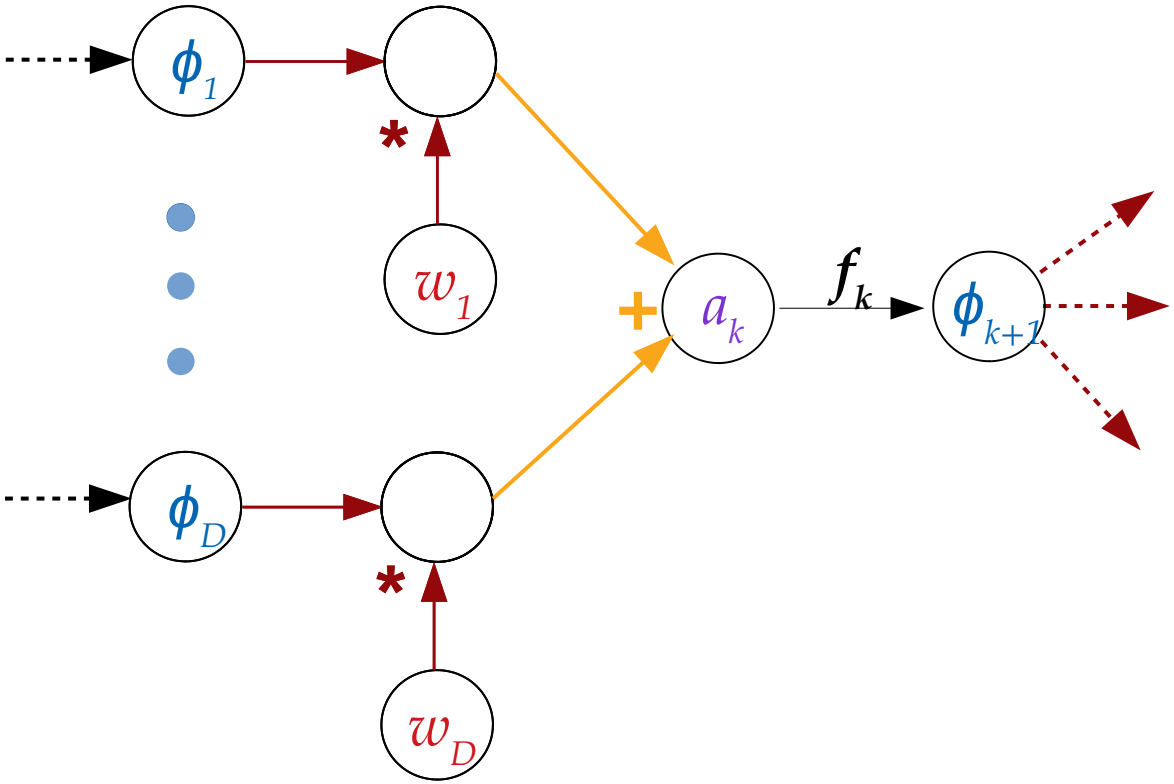


- $\phi_0 = x$
- Pour $k = 1, 2, \ldots K$:
  - $a_k = w_k \cdot \phi_{k-1}$
  - $\phi_k = f_k(a_k)$
- $F = L(\phi_K, y)$

- $\phi_K^\delta = \frac{\partial F}{\partial \phi_K} = L'(\phi_k, y)$
- Pour $k = K, K-1, \ldots, 1$:
  - $a_k^\delta = \frac{\partial F}{\partial \phi_k}\frac{\partial \phi_k}{\partial a_k} = \phi_k^\delta \, f_k'(a_k)$
  - $w_k^\delta = \frac{\partial F}{\partial a_k}\frac{\partial a_k}{\partial w_k} = a_k^\delta \, \phi_{k-1}$
  - $\phi_{k-1}^\delta = \frac{\partial F}{\partial a_k}\frac{\partial a_k}{\partial \phi_{k-1}} = a_k^\delta \, w_k$



$$\frac{\partial f(h(x))}{\partial x} = \left[\frac{\partial f(a)}{\partial a}\right]_{a=h(x)} \frac{\partial h(x)}{\partial x}$$



Soit un réseau R de K couches:

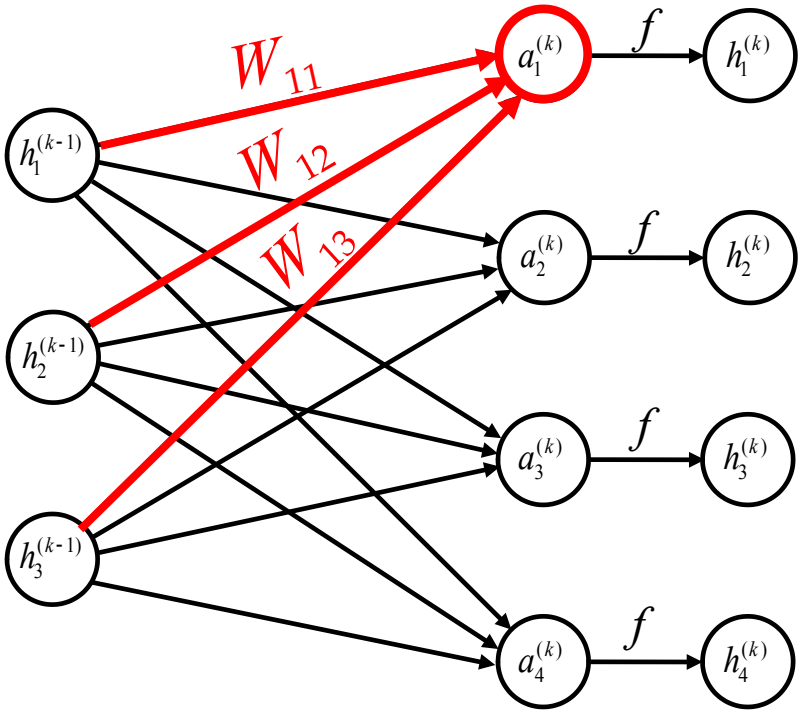
- Fonctions d'activations : $f_1, \dots f_K$
- Matrices de poids: $\mathbf{W}^{(1)}, \dots \mathbf{W}^{(K)}$
  - Chaque matrice $\mathbf{W}^{(k)}$ est de taille $d_k \times d_{k-1}$
  - d_k est le nombre de neurones sur la couche k
  - $k = 0$ correspond à la couche d'entrée: $\mathbf{x} \in \mathbb{R}^{d_0}$

$$\mathbf{W}^{(k)} = \begin{bmatrix} W_{11} & \cdots & W_{1d_{K-1}} \\ \vdots & \ddots & \vdots \\ W_{d_K 1} & \cdots & W_{d_K d_{K-1}} \end{bmatrix}$$

Algorithme de propagation avant.
ENTRÉES: Réseau R, Observation $\mathbf{x}$

- $\mathbf{h}[0] \leftarrow \mathbf{x}$
- Pour k de 1 à K:
  - $\mathbf{a}[k] \leftarrow \mathbf{W}^{(k)}\mathbf{h}^{(k-1)}$
  - $\mathbf{h}[k] \leftarrow f_k\big(\mathbf{a}[k]\big)$

SORTIE: $\mathbf{h}[K]$



Algorithme de propagation avant.
ENTRÉES: Réseau R, Observation $\mathbf{x}$

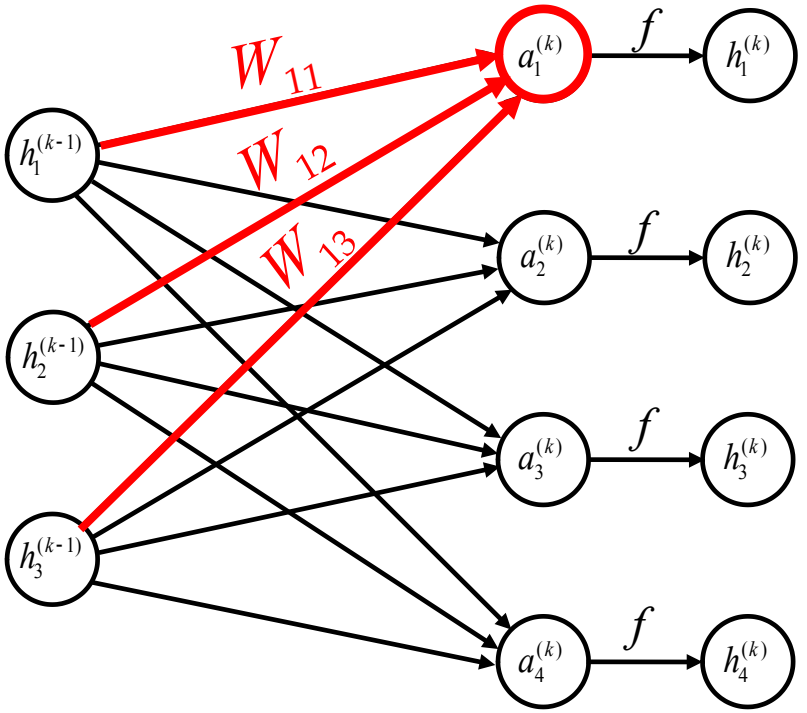
- $\mathbf{h}[0] \leftarrow \mathbf{x}$
- Pour k de 1 à K:
  - $\mathbf{a}[k] \leftarrow \mathbf{W}^{(k)}\mathbf{h}^{(k-1)}$
  - $\mathbf{h}[k] \leftarrow f_k\big(\mathbf{a}[k]\big)$

SORTIE: $\mathbf{h}[K]$

Algorithme de retropropagation.
ENTRÉES: Réseau R, Perte L, Observation $\mathbf{x}$, Sortie attendue $\mathbf{y}$

- $\mathbf{g} \leftarrow L'\big(h[K], \mathbf{y}\big)$
- Pour k décroissant de K à 1:
  - $\mathbf{g} \leftarrow \mathbf{g} \odot f_k'(\mathbf{a}[k])$
  - $\nabla \mathbf{w}[k] \leftarrow \mathbf{g}\,\mathbf{h}[k]^T$
  - $\mathbf{g} \leftarrow \mathbf{W}^{(k)T}\mathbf{g}$

SORTIE: $\nabla \mathbf{w}$



Différentiation automatique dans un graphe de calcul

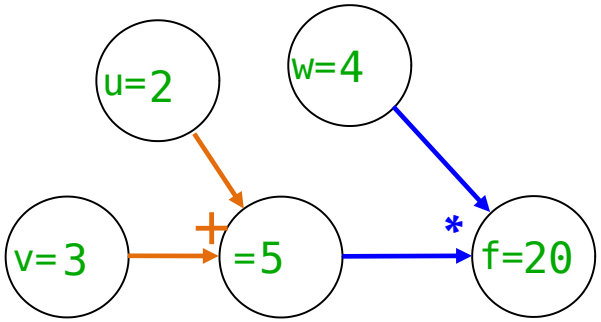
«Backprop» et différentiation automatique

- Algorithme qui calcule tous les gradients dans un graphe de calcul quelconque.
- **N'est pas l'algo d'optimisation!**
- Mais tous les algos d'optimisation des réseaux de neurones utilisent les gradients calculés par *backprop*.
- Basé sur la règle de dérivation en chaîne
- Les librairies modernes de réseau de neurones effectuent le calcul des dérivés automatiquement (comme *pyTorch* et *TensorFlow*).

Exemple sur un graphe de calcul simple

$$f = (u+v)w$$

nœud : variable
arête : opération

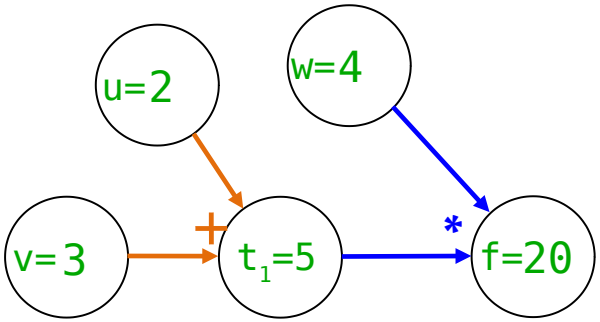


Instanciation des les variables:
u=2
v=3
w=4

Évalue le graphe pour avoir f : **Propagation avant** (forward pass)

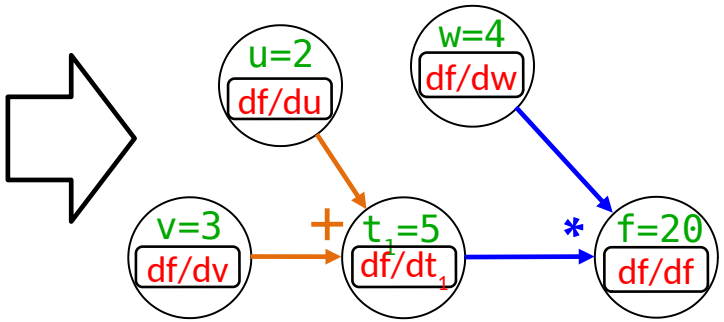
Exemple sur un graphe de calcul simple

À partir d'un graphe de calcul évalué :



$$f=(u+v)w$$

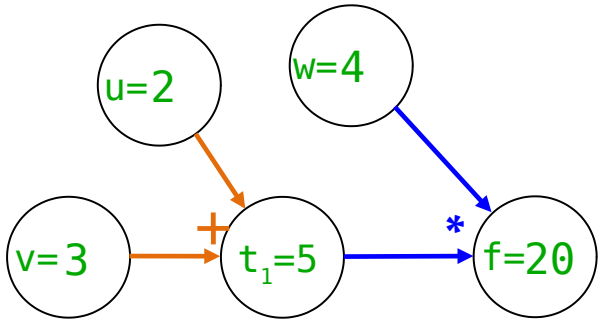
On ajoute une variable pour stocker les gradients :



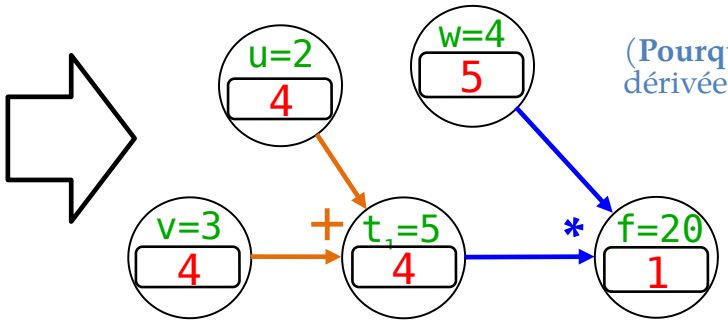
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial h}\frac{\partial h}{\partial x}$$

Exemple sur un graphe de calcul simple

À partir d'un graphe de calcul évalué :



On ajoute une variable pour stocker les gradients :



(**Pourquoi `/df`?** On cherche les dérivées partielle p.r. à la sortie, `f`)

$$f=(u+v)w$$
$$f=t_1 w,\ t_1=u+v$$

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial h}\frac{\partial h}{\partial x}$$

$$\frac{\partial f}{\partial f}=1$$

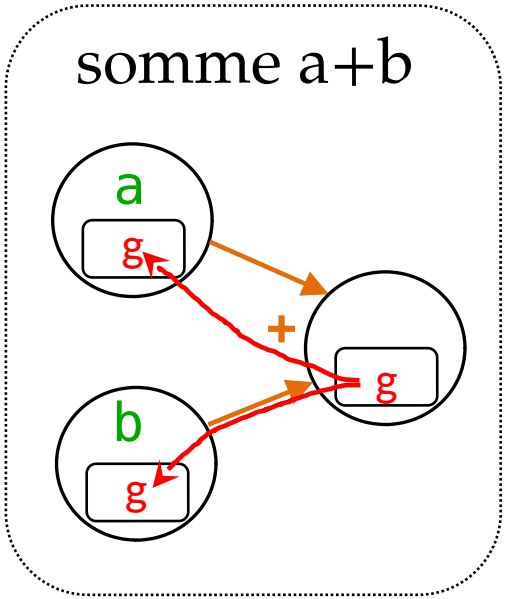
$$\frac{\partial f}{\partial w}=\frac{\partial}{\partial w}(t_1 w)\frac{\partial f}{\partial f}=t_1\cdot 1$$

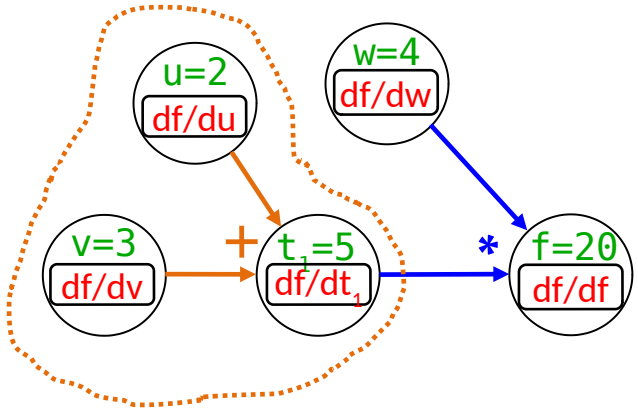
$$\frac{\partial f}{\partial t_1}=\frac{\partial}{\partial t_1}(t_1 w)\frac{\partial f}{\partial f}=w\cdot 1$$

$$\frac{\partial f}{\partial u}=\frac{\partial t_1}{\partial u}\frac{\partial f}{\partial t_1}=1\cdot\frac{\partial f}{\partial t_1}=4$$

$$\frac{\partial f}{\partial v}=\frac{\partial t_1}{\partial v}\frac{\partial f}{\partial t_1}=1\cdot\frac{\partial f}{\partial t_1}=4$$

Déduisons les règles de base





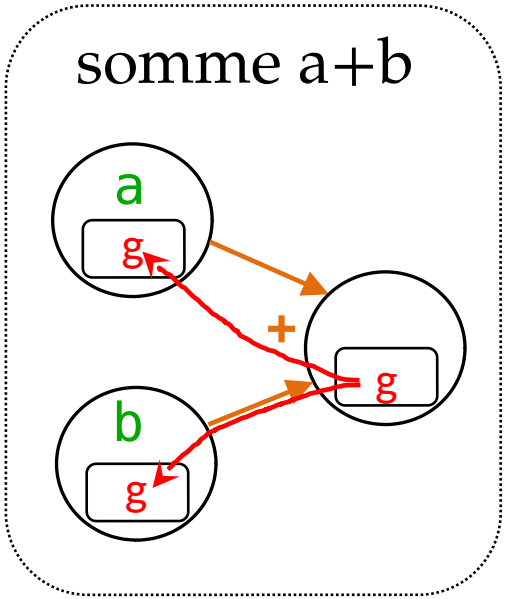
$$\frac{\partial f}{\partial w} = \frac{\partial}{\partial w}(t_1 w)\frac{\partial f}{\partial f} = t_1 \cdot 1$$

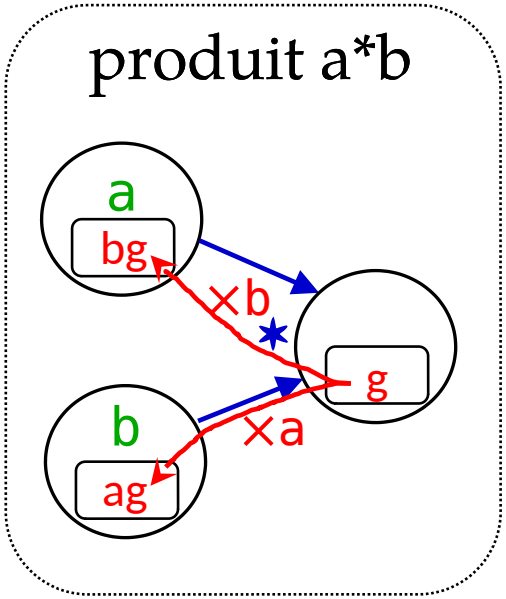
$$\frac{\partial f}{\partial t_1} = \frac{\partial}{\partial t_1}(t_1 w)\frac{\partial f}{\partial f} = w \cdot 1$$

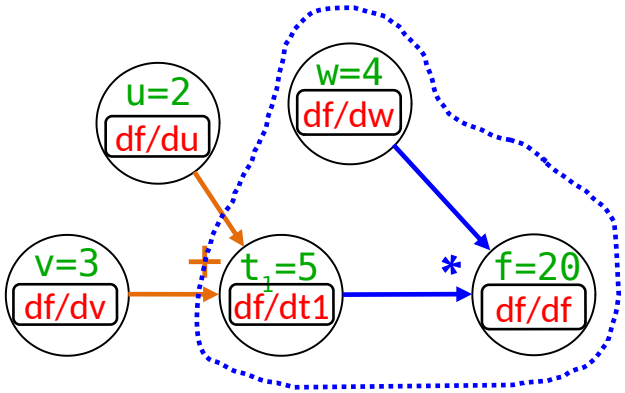
$$\frac{\partial f}{\partial u} = \frac{\partial t_1}{\partial u}\frac{\partial f}{\partial t_1} = 1 \cdot \frac{\partial f}{\partial t_1} = 4$$

$$\frac{\partial f}{\partial v} = \frac{\partial t_1}{\partial v}\frac{\partial f}{\partial t_1} = 1 \cdot \frac{\partial f}{\partial t_1} = 4$$

Déduisons les règles de base







$$\frac{\partial f}{\partial w} = \frac{\partial}{\partial w}(t_1 w)\frac{\partial f}{\partial f} = t_1 \frac{\partial f}{\partial f}$$

$$\frac{\partial f}{\partial u} = \frac{\partial t_1}{\partial u}\frac{\partial f}{\partial t_1} = 1 \cdot \frac{\partial f}{\partial t_1} = 4$$

$$\frac{\partial f}{\partial t_1} = \frac{\partial}{\partial t_1}(t_1 w)\frac{\partial f}{\partial f} = w \frac{\partial f}{\partial f}$$

$$\frac{\partial f}{\partial v} = \frac{\partial t_1}{\partial v}\frac{\partial f}{\partial t_1} = 1 \cdot \frac{\partial f}{\partial t_1} = 4$$

Dérivées des fonctions de d’activation et de perte *classiques*

Dérivées de fonctions de perte

Perte quadratique.

$$L_{\text{quad}}(\hat{y}, y) = (\hat{y} - y)^2$$

$$\begin{aligned} L'_{\text{quad}}(\hat{y}, y) &= \frac{\partial L_{\text{quad}}(\hat{y}, y)}{\partial \hat{y}} \\ &= 2(\hat{y} - y) \end{aligned}$$

Dérivées de fonctions de perte

Perte quadratique.

$$L_{\text{quad}}(\hat{y}, y) = (\hat{y} - y)^2$$

$$\begin{aligned} L'_{\text{quad}}(\hat{y}, y) &= \frac{\partial L_{\text{quad}}(\hat{y}, y)}{\partial \hat{y}} \\ &= 2(\hat{y} - y) \end{aligned}$$

Perte négatif log vraisemblance.

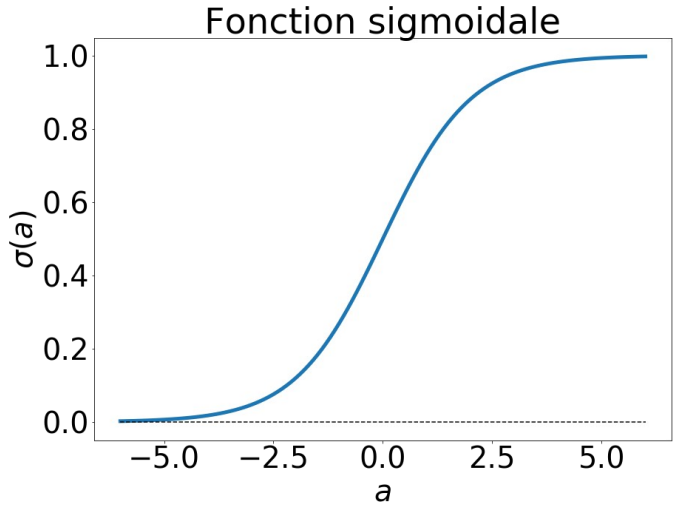
$$L_{\text{nlv}}(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

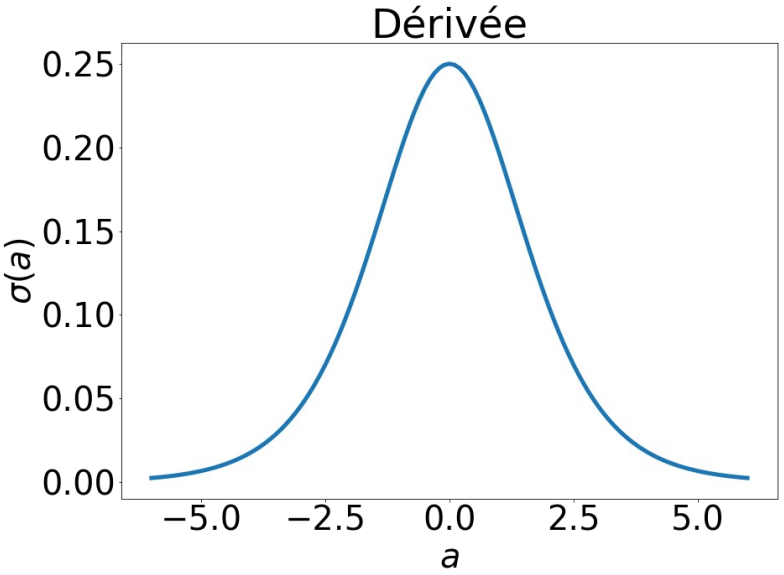
$$\begin{aligned} L'_{\text{nlv}}(\hat{y}, y) &= \frac{\partial L_{\text{nlv}}(\hat{y}, y)}{\partial \hat{y}} \\ &= -\frac{y}{\hat{y}} - \frac{1 - y}{1 - \hat{y}} \end{aligned}$$

Dérivées de fonctions d'activation

$$\sigma(a) = \frac{a}{1 + e^{-a}}$$

$$\begin{aligned}
\sigma'(a) &= \frac{\partial}{\partial a}(1 + e^{-a})^{-1} \\
&= -(1 + e^{-a})^{-2} \frac{\partial}{\partial a}(1 + e^{-a}) \\
&= -\frac{1}{(1 + e^{-a})^2} \left[-\frac{\partial}{\partial a} e^{a} \right] \\
&= \frac{e^{a}}{(1 + e^{-a})^2}
\end{aligned}$$

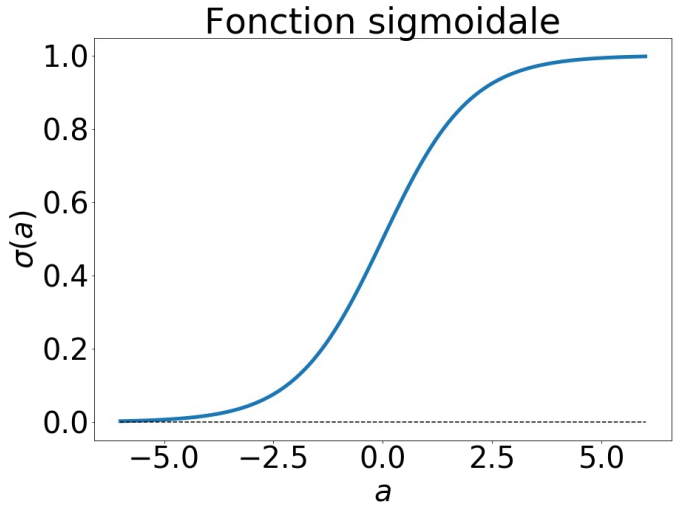


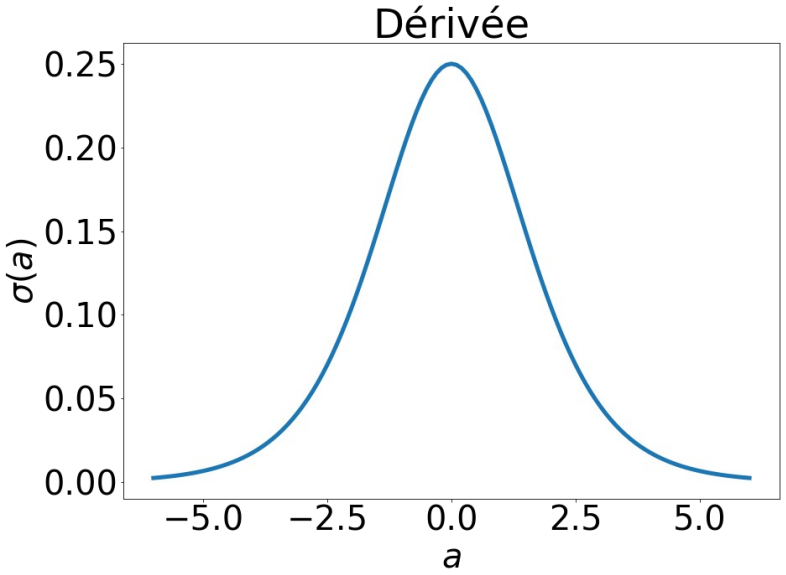


Dérivées de fonctions d'activation

$$\sigma(a) = \frac{a}{1+e^{-a}}$$

$$\begin{aligned}
\sigma'(a) &= \frac{\partial}{\partial a}(1+e^{-a})^{-1} \\
&= -(1+e^{-a})^{-2}\frac{\partial}{\partial a}(1+e^{-a}) \\
&= -\frac{1}{(1+e^{-a})^2}\left[-\frac{\partial}{\partial a}e^{a}\right] \\
&= \frac{e^{a}}{(1+e^{-a})^2} \\
&= \frac{1+e^{a}}{(1+e^{-a})^2} - \frac{1}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}} - \left(\frac{1}{1+e^{-a}}\right)^2 \\
&= \sigma(a) - \Big(\sigma(a)\Big)^2 \\
&= \sigma(a)\Big(1-\sigma(a)\Big)
\end{aligned}$$

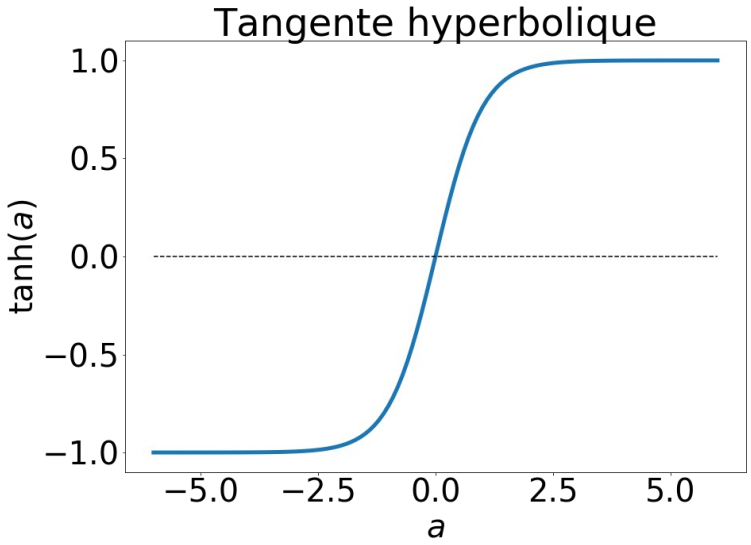


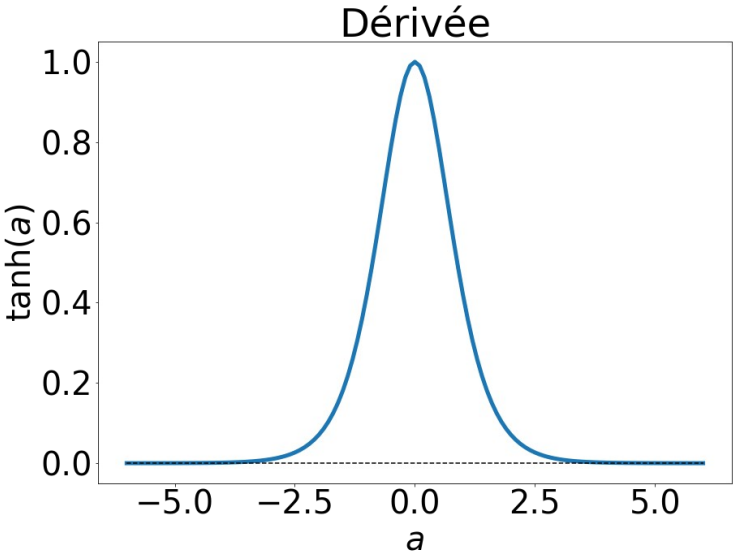


Dérivées de fonctions d'activation

$$\begin{aligned}\tanh(a) &= \frac{e^{2a}-1}{e^{2a}+1} \\ &= 2\sigma(2a) - 1\end{aligned}$$

$$\begin{aligned}\tanh'(a) &= \frac{\partial \tanh(a)}{\partial a} \\ &= 4\,\sigma'(2a) \\ &= 1 - \Big(\tanh(a)\Big)^2\end{aligned}$$





Dérivées de fonctions d'activation

$$\mathrm{relu}(a) = \max(0, a)$$

$$\begin{aligned}\mathrm{relu}'(a) &= \frac{\partial\,\mathrm{relu}(a)}{\partial a} \\ &= \mathbb{1}_{a>0}\end{aligned}$$