TP8 - Réseaux de Neurones

1 Différentier des matrices ?

Soit $f : \mathbb{R}^{n_{out}} \to \mathbb{R}$ une fonction définie sur l'espace des matrices de taille $n_{out} \times n_{in}$ et à valeurs réelles. Soit X une valeur "input" fixée. Lors de la phase de backpropagation, on est typiquement amené à calculer le gradient de l'application

$$\phi(W, b) = f(WX + b),$$

où

- $W \in \mathbb{R}^{n_{out} \times n_{in}}$ est une matrice de poids.
- $b \in \mathbb{R}^{n_{out}}$ un vecteur de biais.

Rappelons que le produit scalaire usuel sur l'espace des matrices est donné par :

$$\langle A, B \rangle := \sum_{i,j=1}^{n} A_{ij} B_{ij} = \mathrm{Tr}(A^T B).$$

1.1 Calcul du gradient $\nabla_W \phi(W, b)$

La définition du gradient stipule que (c'est la formule d'approximation de Taylor au premier ordre appliquée à la fonction f au point $WX + b$) :

$$\phi(W + H, b) = f((W + H)X + b) \approx f(WX + b) + \langle \nabla f(WX + b), HX \rangle.$$

Or, par définition du gradient :

$$\begin{aligned} \langle \nabla f(WX + b), HX \rangle &= \mathrm{Tr}(\nabla f(WX + b)(HX)^T) \\ &= \mathrm{Tr}(\nabla f(WX + b) X^T H^T) = \langle \nabla f(WX + b) X^T, H \rangle. \end{aligned}$$

Ceci implique que

$$\nabla_W \phi(W) = \nabla f(WX + b) X^T.$$

1.2 Calcul du gradient $\nabla_b \phi(W, b)$

De la même manière que précédemment, on commence par appliquer la formulation d'approximation de Taylor au premier ordre appliquée à la fonction f au point $WX + b$:

$$\phi(W, b + h) = f(WX + b + h) \approx f(WX + b) + \langle \nabla f(WX + b), h \rangle.$$

Ceci permet d'identifier immédiatement le gradient cherché :

$$\nabla_b \phi(W) = \nabla f(WX + b).$$

1.3 Figure illustrative