TD - Gradients

1 Gradients

Question 1 (Dérivation des fonctions composées)**.**
On dit qu'une fonction $f : \mathbb{R} \to \mathbb{R}$ est dérivable en t si $\lim_{h\to 0} \frac{1}{h}\big(f(t+h) - f(t)\big)$ existe. Dans ce cas on note

$$f'(t) = \lim_{h\to 0} \frac{1}{h}\big(f(t+h) - f(t)\big) \ .$$

De manière équivalente, on peut écrire : il existe une fonction ϵ_f^x telle que

$$f(t+h) = f(t) + f'(t)h + h\epsilon_f^x(h) \ .$$

et $\lim_{h\to 0} \epsilon_f^x(h) = 0$.
Soit $f : \mathbb{R} \to \mathbb{R}$ et $g : \mathbb{R} \to \mathbb{R}$. Montrer que

$$(f \circ g)'(t) = f'(g(t)) \times g'(t)$$

Question 2 (Matrice jacobienne)**.**
On dit qu'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dérivable en x si il existe un vecteur $\nabla f(x) \in \mathbb{R}^n$ et une fonction ϵ_f^x tels que

$$f(x+h) = f(x) + \nabla f(x)^\top h + \|h\|\epsilon_f^x(h)$$

où $\lim_{h\to 0} \epsilon_f^x(h) = 0$.
On note les coordonnées de $\nabla f(x)$ de plusieurs manières :

$$(\nabla f(x))_i = \nabla_i f(x) = \frac{\partial f}{\partial x_i}(x) \ .$$

Il se trouve que $\nabla_i f(x)$ est égale à la i^{me} dérivée directionnelle :

$$\nabla_i f(x) = \lim_{t\to 0} \frac{f(x + te_i) - f(x)}{t} \ .$$

Soit $F : \mathbb{R}^n \to \mathbb{R}^m$ une fonction à valeurs vectorielles, c'est à dire que $F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$.
On dit que F est dérivable en x si pour tout $i \in \{1, \dots, m\}$, f_i est dérivable en x :

$$f_i(x+h) = f_i(x) + \nabla f_i(x)^\top h + \|h\|\epsilon_{f_i}^x(h)$$

où $\lim_{h\to 0} \epsilon_{f_i}^x(h) = 0$.

On appelle matrice jacobienne de F en x la matrice qui concatène tous les gradients des f_i, c'est à dire

$$J_F(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \dots & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix} .$$

Vérifier qu'avec cette notation, on a

$$F(x+h) = F(x) + J_F(x)h + o(\|h\|) .$$

Question 3 (Calculs de gradients)**.**

- $f_1(x) = \frac{1}{2}\|Ax-b\|_2^2$, A matrice de taille $m \times n$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. Calculer le gradient de f_1 en x.
- $f_2(x) = Bx + c$, B matrice de taille $p \times n$, $c \in \mathbb{R}^p$, $x \in \mathbb{R}^n$. Calculer la Jacobienne de f_2 en x.
- $f_3(P,Q) = \frac{1}{2}\|M - PQ\|_F^2$, M matrice de taille $m \times n$, P matrice de taille $m \times k$ et Q matrice de taille $k \times n$. Calculer le gradient de f_3 en (P,Q).

Question 4. Soient $F : \mathbb{R}^m \to \mathbb{R}^p$ et $G : \mathbb{R}^n \to \mathbb{R}^m$ deux fonctions dérivables. Montrer que pour tout i, j,

$$\frac{\partial (F \circ G)_j}{\partial x_i}(x) = \sum_{l=1}^{m} \frac{\partial F_j}{\partial y_l}(G(x)) \frac{\partial G_l}{\partial x_i}(x) ,$$

et que cette formule est équivalente à

$$J_{F \circ G}(x) = J_F(G(x)) J_G(x) .$$

2 Rétropropagation dans les réseaux de neurones

Considérons le modèle de réseau de neurones à 1 couche suivant :

$$y = f(w,x) = \sigma\Big(\sum_{i=1}^{H} w_i v_i \big(\sum_{j=1}^{N} w_{i,j} x_j \big) \Big) . \tag{1}$$

Dans cette formule :

- $x_1, \dots, x_N$ sont les observations.
- y est la sortie du modèle.
- Le nombre entier H est appelé nombre de neurones.
- σ et $v_1, \dots, v_H$ sont des fonctions fixées appelées fonctions d'activation. On supposera que ces fonctions sont dérivables. Un choix classique est $\sigma(z) = v_i(z) = \tanh(z)$.
- $w_1, \dots, w_H, w_{1,1}, \dots, w_{1,N}, w_{2,1}, \dots, w_{H,1}, \dots, w_{H,N}$ sont les paramètres du modèle. Il y en a $N \times H + H$.

Le but de cette partie du TD est de trouver une formule pour calculer le gradient de f par rapport à w, ce qui est la première étape pour implémenter une méthode de gradient. Cette formule est à la base des logiciels d'apprentissage de réseaux de neurones comme Tensorflow ou Keras.

Question 5. Écrire la fonction $f : \mathbb{R}^{NH+H} \times \mathbb{R}^N \to \mathbb{R}$ du modèle de réseau de neurones (1) comme une composition de fonctions plus simples de la forme suivante :

$$f(w, x) = \sigma \circ M\big(w, V \circ L(w, x)\big) \ .$$

Vous expliciterez les fonctions M, V et L en faisant attention à leur nombre de variables et à la dimension des images.

Question 6. Calculer les jacobiennes de chacune des fonctions en jeu.

Question 7. Montrer que le gradient de f par rapport à w, que l'on notera $\nabla_w f$ peut s'écrire comme produit matriciel et somme des jacobiennes calculées à question précédente.

Question 8. Évaluer le nombre d'opérations nécessaires pour calculer $\nabla_w f$ quand on commence par la couche d'entrée du réseau de neurones. On rappelle que pour calculer le produit matriciel $A \times B$ où A est de taille $n \times m$ et B de taille $m \times p$, il faut environ nmp opérations.

Question 9. Évaluer le nombre d'opérations nécessaires pour calculer $\nabla_w f$ quand on commence par la couches de sortie du réseau de neurones.