TD - Gradients

1 Gradients

Question 1 (Chain rule)**.**
We say that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at t if $\lim_{h\to 0} \frac{1}{h}\big(f(t+h) - f(t)\big)$ exists. In that case, we denote

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

Equivalently, we can write: there exists a function ϵ_f^t such that

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

and $\lim_{h\to 0} \epsilon_f^t(h) = 0$.
Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. Show that

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

Question 2 (Jacobian matrix)**.**
We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at x if there exists a vector $\nabla f(x) \in \mathbb{R}^n$ and a function ϵ_f^x such that

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

where $\lim_{h\to 0} \epsilon_f^x(h) = 0$.
The coordinates of $\nabla f(x)$ can be written in several ways:

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

In fact $\nabla_i f(x)$ is equal to the i^{th} directional derivative:

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

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a function with vectorial values, which means that $F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$.
We say that F is differentiable at x iffor all $i \in \{1, \dots, m\}$, f_i is differentiable at x:

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

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

The Jacobian matrix of F at x is the matrix that concatenates all the gradients of the f_i's, that is

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

Check that with this notation, we have

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

Question 3 (Gradient calculus)**.**

- $f_1(x) = \frac{1}{2}\|Ax - b\|_2^2$, A matrix of size $m \times n$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. Compute the gradient of f_1 at x.

- $f_2(x) = Bx + c$, B matrix of size $p \times n$, $c \in \mathbb{R}^p$, $x \in \mathbb{R}^n$. Compute the Jacobian of f_2 at x.

- $f_3(P, Q) = \frac{1}{2}\|M - PQ\|_F^2$, M matrix of size $m \times n$, P matrix of size $m \times k$ and Q matrix of size $k \times n$. Compute the gradient of f_3 at (P, Q).

Question 4. Let $F : \mathbb{R}^m \to \mathbb{R}^p$ and $G : \mathbb{R}^n \to \mathbb{R}^m$ be two differentiable functions. Show that for all i, j,

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

and that this formula is equivalent to

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

2 Backpropagation in neural networks

Let us consider the following 1-layer neural network:

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

In this formula:

- $x_1, \dots, x_N$ are the observations.

- y is the output of the model.

- The integer H is the number of neurons.

- σ and $v_1, \dots, v_H$ are given functions called activation functions. We suppose that these functions are differentiable. A classical choice is $\sigma(z) = v_i(z) = \tanh(z)$.

- $w_1, \dots, w_H, w_{1,1}, \dots, w_{1,N}, w_{2,1}, \dots, w_{H,1}, \dots, w_{H,N}$ are the parameters of the model. There are $N \times H + H$ of them.

The goal of this exercise is to find a formula to compute the graduebt of f with respect to w, which is the first step to implement a gradient method. This formula is the basis of softwares for neural network training like Tensorflow or 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.