

## 1 Question 1

Calculer la dérivée partielle

$$\frac{\partial(L \circ h)}{\partial \mathbf{x}_i}(\mathbf{x})$$

en fonction de

$$\frac{\partial h_j}{\partial \mathbf{x}_i}(\mathbf{x}) \quad \text{et} \quad (\nabla L)_j = \frac{\partial L}{\partial \mathbf{y}_j}(h(\mathbf{x})),$$

où  $(\nabla L)_j$  permet de simplifier la notation.

Pour une fonction  $h$  connue (ce qui est toujours le cas pour nous), on peut calculer explicitement sa dérivée par rapport à  $x_i$ .

**Answer** On note

$$h(\mathbf{x}) = \mathbf{y}$$

Donc quand on va chercher la dérivée de  $L \circ h$  par rapport à  $x_i$ , il faut d'abord voir comment  $L$  varie quand on fait varier  $h$ , puis comment  $h$  varie quand on fait varier  $x_i$ .

Explicitement,  $L \circ h$  peut s'écrire,

$$L(h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_m(\mathbf{x})).$$

Donc la dérivée partielle de  $L \circ h$  par rapport à la coordonnée  $x_i$  s'écrit,

$$\frac{\partial(L \circ h)}{\partial x_i}(\mathbf{x}) = \sum_{j=1}^m \left( \frac{\partial L}{\partial y_j}(\mathbf{y}) \frac{\partial h_j}{\partial x_i}(\mathbf{x}) \right) = \sum_{j=1}^m \left( (\nabla L)_j \frac{\partial h_j}{\partial x_i}(\mathbf{x}) \right)$$

## 2 Question 2

Appliquer le résultat trouvé dans la question 1 pour calculer:

Les dérivées pour la fonction de coût MSE

$$\frac{\partial(L \circ \text{mse})}{\partial y}(\hat{y}, y) \quad \text{et} \quad \frac{\partial(L \circ \text{mse})}{\partial \hat{y}}(\hat{y}, y)$$

Les dérivées de la fonction linéaire  $f$  par rapport à ses entrées  $\mathbf{x}$  ( $\mathbf{w}$  et  $b$  sont obtenus de manière similaire)

$$\frac{\partial(L \circ f)}{\partial \mathbf{x}_j}(\mathbf{x}, \mathbf{w}, b)$$

La fonction  $L$  introduite ici est générique et est présente uniquement pour bien montrer le fonctionnement du chaînage des dérivées. Par rapport au problème qui nous intéresse, dans le premier cas comme la fonction  $\text{mse}$  est la dernière de étape de calcul, on peut considérer que  $L$  correspond à l'identité ; dans le deuxième cas il s'agit bien sûr de la fonction  $\text{mse}(\cdot, \hat{y})$ .

**Answer** D'abord on calcule la dérivée de mse par rapport à ses entrées  $\hat{y}$  et  $y$ .

$$\begin{aligned}\frac{\partial \text{mse}}{\partial y_i}(\hat{y}, y) &= \frac{\partial}{\partial y_i} ((\hat{y}_j - y_j)^2) = -2(\hat{y}_i - y_i) \\ \frac{\partial \text{mse}}{\partial \hat{y}_i}(\hat{y}, y) &= \frac{\partial}{\partial \hat{y}_i} ((\hat{y}_j - y_j)^2) = 2(\hat{y}_i - y_i)\end{aligned}$$

Donc maintenant on peut appliquer la formule de la question 1.

$$\frac{\partial(L \circ f)}{\partial \mathbf{x}_j}(\mathbf{x}, \mathbf{w}, b) = \sum_{i=1}^m \frac{\partial L}{\partial f_i} \frac{\partial f_i}{\partial \mathbf{x}_j}$$

Ensuite, la dérivée de  $f$  par rapport à  $\mathbf{x}_j$  est simplement la dérivée de la couche lineaire (uniquement dans le cas scalaire):

$$f(\mathbf{x}) = \mathbf{W}\mathbf{x} + b$$

Soit,

$$\frac{\partial f_i}{\partial \mathbf{x}_j}(\mathbf{x}) = \mathbf{W}_j$$

Donc on a:

$$\frac{\partial(L \circ f)}{\partial \mathbf{x}_j}(\mathbf{x}, \mathbf{w}, b) = \frac{\partial L}{\partial f_i} \mathbf{W}_j$$

On peut donc étendre la formule pour:

$$\frac{\partial f}{\partial \mathbf{x}_j}(\mathbf{x}, \mathbf{w}, b) \frac{\partial L}{\partial \hat{y}}(\hat{y}) = \mathbf{W}_j \frac{\partial L}{\partial \hat{y}}(\hat{y}) = 2\mathbf{W}_j(\hat{y} - y)$$

### 3 Question 3

Étendre les résultats obtenus dans la question précédente pour calculer:  
Les dérivées pour la fonction de coût MSE:

$$\frac{\partial(L \circ \text{mse})}{\partial \mathbf{Y}_{ij}}(\hat{Y}, Y)$$

et

$$\frac{\partial(L \circ \text{mse})}{\partial \hat{Y}_{ij}}(\hat{Y}, Y)$$

Les dérivées de la fonction linéaire  $f$  par rapport à ses entrées:

$$\frac{\partial(L \circ f)}{\partial \mathbf{X}_{ij}}(\mathbf{X}, \mathbf{W}, \mathbf{b}), \frac{\partial(L \circ f)}{\partial \mathbf{W}_{ij}}(\mathbf{X}, \mathbf{W}, \mathbf{b}), \text{ et } \frac{\partial(L \circ f)}{\partial \mathbf{b}_{ij}}(\mathbf{X}, \mathbf{W}, \mathbf{b})$$

**Answer** Rappel des dimensions:

$$\mathbf{X} \in \mathbb{R}^{q \times n}, \quad \mathbf{W} \in \mathbb{R}^{n \times p}, \quad \mathbf{b} \in \mathbb{R}^{1 \times p},$$

On reprend les mêmes calculs que dans la question 2, mais cette fois pour des matrices pour Pytorch. Rappel:

$$C(\mathbf{X}, \mathbf{W}, \mathbf{b}, \mathbf{Y}) = \text{mse}(f(\mathbf{X}, \mathbf{W}, \mathbf{b}), \mathbf{Y})$$

Donc,

$$\begin{aligned} \frac{\partial(L \circ \text{mse})}{\partial Y_{ij}}(\hat{Y}, Y) &= -\frac{2}{q}(\hat{Y}_{ij} - Y_{ij}) \\ \frac{\partial(L \circ \text{mse})}{\partial \hat{Y}_{ij}}(\hat{Y}, Y) &= \frac{2}{q}(\hat{Y}_{ij} - Y_{ij}) \end{aligned}$$

Ensuite pour les dérivées de f, mais avant petit rappel que chaque composante s'écrit:

$$f_{ik}(\mathbf{X}) = \sum_{j=1}^n \mathbf{X}_{ij} \mathbf{W}_{jk} + b_k$$

Pour une couche linéaire :

$$\frac{\partial(L \circ f)}{\partial X_{ij}} = \sum_{k=1}^p \sum_{l=1}^n \frac{\partial L}{\partial f_{kl}} \frac{\partial f_{kl}}{\partial X_{ij}} = \sum_{k=1}^p \sum_{l=1}^n \frac{\partial L}{\partial f_{kl}} \frac{\partial}{\partial X_{ij}} \sum_u (\mathbf{X}_{ku} \mathbf{W}_{ul} + b_u)$$

et ici on utilise le fait que  $(\mathbf{X}_{ku} \mathbf{W}_{ul} + b_u)$  est égal à 0 si  $k \neq i$  ou  $l \neq j$ , donc on a:

$$\frac{\partial L}{\partial f_{ij}} \sum_u \mathbf{W}_{uj}$$

Pour les deux autres on utilise le même raisonnement:

$$\frac{\partial(L \circ f)}{\partial W_{ij}} = \text{TODO}$$

$$\frac{\partial(L \circ f)}{\partial b_{ij}} = \text{TODO}$$

## 4 Question 4

**Answer** Pour  $\hat{Y}, Y \in \mathbb{R}^{q \times p}$ , la fonction de coût MSE est :

$$\text{MSE}(\hat{Y}, Y) = \frac{1}{p} \sum_{j=1}^p (\hat{Y}_{ij} - Y_{ij})^2$$

Les gradients matriciels sont :

$$\frac{\partial(L \circ \text{mse})}{\partial \hat{Y}} = \frac{2}{p}(\hat{Y} - Y)$$

$$\frac{\partial(L \circ \text{mse})}{\partial Y} = -\frac{2}{p}(\hat{Y} - Y)$$

— Pour une couche linéaire  $\mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{W} + \mathbf{b}$

$$\frac{\partial(L \circ f)}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{f}} \mathbf{W}^\top \in \mathbb{R}^{q \times n}$$

$$\frac{\partial(L \circ f)}{\partial \mathbf{W}} = \mathbf{X}^\top \frac{\partial L}{\partial \mathbf{f}} \in \mathbb{R}^{n \times p}$$

$$\frac{\partial(L \circ f)}{\partial \mathbf{b}} = \mathbf{1}_q^\top \frac{\partial L}{\partial \mathbf{f}} \in \mathbb{R}^{1 \times p}$$

où  $\mathbf{1}_q \in \mathbb{R}^{q \times 1}$  est un vecteur colonne de 1, permettant de sommer sur tous les exemples.

—

## 5 Question 5

**Answer** Rappel:

$$C(\mathbf{X}, \mathbf{W}, \mathbf{b}, \mathbf{Y}) = \text{mse}(f(\mathbf{X}, \mathbf{W}, \mathbf{b}), \mathbf{Y})$$

Soit,

$$\text{mse}(f(\mathbf{X}, \mathbf{W}, \mathbf{b}), \mathbf{Y}) = \text{mse}(\mathbf{X}\mathbf{W} + \mathbf{b}, \mathbf{Y}) = \frac{1}{qp} \sum_{i=1}^q \sum_{j=1}^p ((\mathbf{X}\mathbf{W} + \mathbf{b})_{ij} - \mathbf{Y}_{ij})^2$$

Donc,

$$\frac{\partial C}{\partial \mathbf{W}} = \frac{2}{qp} \mathbf{X}^\top (\mathbf{X}\mathbf{W} + \mathbf{b} - \mathbf{Y})$$

et pour  $\mathbf{b}$ ,

$$\frac{\partial C}{\partial \mathbf{b}} = \frac{2}{qp} \mathbf{1}_q^\top (\mathbf{X}\mathbf{W} + \mathbf{b} - \mathbf{Y})$$

Remarque, ici on pourrait remplacer  $\mathbf{X}\mathbf{W} + \mathbf{b}$  par juste  $\hat{Y}$ .