MLP project formula

Anthony Fraga

April 2023

1 Préambule

Ce document est une trace des calculs effectué par ma part pour le projet MLP de Maths-Deep 2023. Pour ce qui est des notations, je réutilise les mêmes que celles de l'énoncé (je vous y renvoie donc). Bonne lecture !

2 Exercice 1

2.1 Introduction

La fonction C, qui est au premier abord définie en fonction de $a^{(l)}$ [1], puis dans un deuxième temps en fonction de $z_j^{(l)}$ [2]. Pour expliquer le lien entre $a^{(L)}$ et $z_j^{(l)}$, il faut bien comprendre deux choses :

- on peut passer de $a^{(l)}$ à $z^{(l)}, l \in \{2, \cdots, L\}$ en appliquant la fonction σ, comme il est expliqué dans l'énoncé de l'exercice ;

$$a^{(l)} = \sigma(z^{(l)})$$

- on peut aussi passer de $z^{(l)}$ à $z^{(l+1)}$ ainsi :

$$z^{(l+1)} = W^{(l+1)}\sigma(z^{(l)}) + b^{(l+1)}$$

.

Et donc en comprenant cela, on peut donc définir C en fonction de $z_j^{(l)}$,

[1] Dans l'écriture de $Loss_i(\lambda)$ principalement

[2] cf. la définition de $\delta_j^{(l)}$

de la manière suivante :

$$C(z_n^{(l-1)}) = \frac{1}{2}\sum_{j=1}^{n_L}\Bigg(y(x^i) - \sigma_L\Bigg(b_j^{(L)} + \sum_{k=1}^{n_{L-1}} w_{j,1}^{(L)}\ \sigma_{L-1}\Bigg(\cdots$$
$$\sum_{n=1}^{n_l}\sigma_l\big(b_n^{(l)} + \sum_{m=1}^{n_{l-1}} w_{n,m}^{(l)}\ z_m^{(l-1)})\big)\Bigg)\Bigg)^2$$

avec $n \in \{1, ..., n_{l-1}\}$.

De manière analogue, on pourrait définir C en fonction de $b_n^{(l)}$ ou $w_{n,m}^{(l)}$, avec $n \in \{1, ..., n_{l-1}\}$, $m \in \{1, ..., n_l\}$.

On peut, avec une définition de la sorte, mieux comprendre ce qu'est que la dérivée partielle de C. En effet, C est une fonction de composée de sigmoïdes, le tout au carré [3]. Calculer sa dérivée nous ramène donc à un calcul du style :

$$\big(y(x^i) - (\sigma_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_l)^2\big)' = 2\times\big(y(x^i) - \sigma_L \circ \sigma_2 \circ \cdots \circ \sigma_l\big)\times(\sigma_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_l)'$$

Or, la dérivée de la composée de sigmoïdes donne :

$$(\sigma_L \circ \sigma_{L-1} \circ \cdots \circ \sigma_l)' = (\sigma_l') \times (\sigma_{l+1}' \circ \sigma_l) \times \cdots (\sigma_L' \circ \sigma_{L-1} \circ \cdots \circ \sigma_l)$$

avec $\sigma_l = \sigma(W^{(l)}a^{(l-1)} + b^{(l)}) = z^{(l)}$, pour tout $2 \leq l \leq L$. Passons désormais aux calculs.

2.2 Formule 10

Par définition (comme expliquée dans l'introduction), on a

$$C(z^{(L)}) = \frac{1}{2}\left\|y(x^i) - \sigma(z^{(L)})\right\|_2^2 = \frac{1}{2}\left(\sum_{k=1}^{n^L}\Big(y(x^i) - \sigma(z_k^{(L)})\Big)^2\right)$$

Pour le calcul de $\delta^{(L)}$, on a besoin des dérivées partielles de C en fonction de $z_1^{(L)}$ et $z_2^{(L)}$, que l'on obtient ainsi :

$$\frac{\partial C}{\partial z^{(L)}} = \begin{pmatrix} \frac{\partial C}{\partial z_1^{(L)}} \\ \frac{\partial C}{\partial z_2^{(L)}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \times (-2)\sigma'(z_1^{(L)})\big(y(x^i) - \sigma(z_1^{(L)})\big) \\ \frac{1}{2} \times (-2)\sigma'(z_2^{(L)})\big(y(x^i) - \sigma(z_2^{(L)})\big) \end{pmatrix}$$
$$= \begin{pmatrix} \sigma'(z_1^{(L)})\big(\sigma(z_1^{(L)}) - y(x^i)\big) \\ \sigma'(z_2^{(L)})\big(\sigma(z_2^{(L)}) - y(x^i)\big) \end{pmatrix} = \sigma'(z^{(L)}) \circ (\sigma(z^{(L)}) - y(x^i))$$

[3] puisque chaque coefficient z,b ou w se retrouve une seule et unique fois dans la somme des termes, qui est lui au carré

Ainsi, d'après (9), on a

$$\delta^{(L)} = \sigma'(z^{(L)}) \circ (a^{(L)} - y(x^i))$$

Remarque : On peut généraliser cette formule pour $n_L = n$ un entier quelconque, avec $\delta^{(L)}$ un vecteur à n dimension.

2.3 Formule 11

Montrons ce résultat au rang $L-1$, puis dans un cas quelconque.

Avec la notation de l'introduction, on a :

$$C(z_n^{(L-1)}) = \frac{1}{2} \sum_{j=1}^{n_L=2} \left(y(x^i) - \sigma_L\Big(b^{(L)} + \sum_{i=1}^{n_{L-1}} w_{j,i}^{(L)} \ \sigma_{L-1}(z_i^{(L-1)})\Big)\right)^2$$

$n \in \{1, \cdots, n_{L-1}\}$. On peut remarquer que $z^{(L-1)}$ n'apparaît qu'une seule fois dans chaque carré, avec les coefficients $w_{1,n}$ et $w_{2,n}$, et cette information sera utile lors du calcul de la dérivée ... d'ailleurs, en se rappelant la formule de l'introduction, on a

$$\begin{aligned}
\frac{\partial C}{\partial z_n^{(L-1)}} = \frac{1}{2} \times 2 \times (a^{(L)} - y(x^i)) \times \sigma'_{L-1}(z_n^{(L-1)}) \times \\
\sum_{j=1}^{n_L} w_{i,n}^{(L)} \left(\sigma'_L\Big(b_j^{(L)} + \sum_{i=1}^{n_{L-1}} w_{j,i}^{(L)} \ \sigma_{L-1}(z_i^{(L-1)})\Big)\right) \\
= \sigma'_{L-1}(z_n^{(L-1)}) \times \sum_{i=1}^{n_L} w_{i,n}^{(L)} \sigma'_L(z^{(L)}) \times (a^{(L)} - y(x^i))
\end{aligned}$$

En faisant cela pour tout $n \in \{1, \cdots, n_L\}$, on obtient :

$$\begin{aligned}
\delta^{(L-1)} = \begin{pmatrix} \frac{\partial C}{\partial z_1^{(L-1)}} \\ \vdots \\ \frac{\partial C}{\partial z_{n_L-1}^{(L-1)}} \end{pmatrix} &= \begin{pmatrix} \sigma'_{L-1}(z_1^{(L-1)}) \times \sum_{i=1}^{n_L} w_{i,1}^{(L)} \sigma'_L(z^{(L)}) \\ \vdots \\ \sigma'_{L-1}(z_n^{(L-1)}) \times \sum_{i=1}^{n_L} w_{i,n}^{(L)} \sigma'_L(z^{(L)}) \\ \vdots \\ \sigma'_{L-1}(z_{n_L}^{(L-1)}) \times \sum_{i=1}^{n_L} w_{i,n_L}^{(L)} \sigma'_L(z^{(L)}) \end{pmatrix} \times (a^{(L)} - y(x^i)) \\
&= \begin{pmatrix} \sigma'_{L-1}(z_1^{(L-1)}) \\ \vdots \\ \sigma'_{L-1}(z_n^{(L-1)}) \\ \vdots \\ \sigma'_{L-1}(z_{n_L}^{(L-1)}) \end{pmatrix} \circ \begin{pmatrix} \sum_{i=1}^{n_L} w_{i,1}^{(L)} \\ \vdots \\ \sum_{i=1}^{n_L} w_{i,n}^{(L)} \\ \vdots \\ \sum_{i=1}^{n_L} w_{i,n_L}^{(L)} \end{pmatrix} \times \sigma'_L(z^{(L)}) \circ (a^{(L)} - y(x^i)) \\
&= \sigma'_{L-1}(z^{(L-1)}) \circ \ {}^tW^{(L)} \delta^{(L)}
\end{aligned}$$

Avec, évidement :

$$
{}^tW^{(L)} = \begin{pmatrix} w_{1,1} & w_{2,1} \\ \vdots & \vdots \\ w_{1,n} & w_{2,n} \\ \vdots & \vdots \\ w_{1,n_L} & w_{2,n_L} \end{pmatrix}
$$

Pour le rang quelconque l, la méthode reste la même, le calcul est juste plus complexe/long. On a[4] :

$$
\frac{\partial C}{\partial z_n^{(l)}} = \sigma_l'(z_n^{(l)}) \times \sum_{i=1}^{n_{l+1}} w_{i,n}^{(l+1)} \sigma_{l+1}'(z^{(l+1)}) \times {}^tW^{(l+2)} \sigma_{l+2}'(z^{(l+2)}) \times \cdots \times \sigma_L'(z^{(L)})(a^{(L)} - y(x^i))
$$

et ainsi, on obtient bien

$$
\delta^{(l)} = \begin{pmatrix} \frac{\partial C}{\partial z_1^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial z_n^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial z_1^{(l)}} \end{pmatrix} = \begin{pmatrix} \sigma_l'(z_1^{(l)}) \times \sum_{i=1}^{n_{l+1}} w_{i,1}^{(l+1)} \sigma_{l+1}'(z^{(l+1)}) \\ \vdots \\ \sigma_l'(z_1^{(l)}) \times \sum_{i=1}^{n_{l+1}} w_{i,1}^{(l+1)} \sigma_{l+1}'(z^{(l+1)}) \\ \vdots \\ \sigma_l'(z_{n_l}^{(l)}) \times \sum_{i=1}^{n_{l+1}} w_{i,n_l}^{(l+1)} \sigma_{l+1}'(z^{(l+1)}) \end{pmatrix}
$$
$$
\times \ {}^tW^{(l+2)} \sigma_{l+2}'(z^{(l+2)}) \times \cdots \times \sigma_L'(z^{(L)})(a^{(L)} - y(x^i))
$$

En simplifiant, on obtient bel et bien, pour $2 \leq l \leq L$:

$$
\begin{aligned}
\delta^{(l)} &= \sigma_l'(z^{(l)}) \circ \ {}^tW^{(l+1)} \sigma_{l+1}'(z^{(l+1)}) \times \ {}^tW^{(l+2)} \sigma_{l+2}'(z^{(l+2)}) \times \cdots \times \sigma_L'(z^{(L)})(a^{(L)} - y(x^i)) \\
&= \sigma_l'(z^{(l)}) \circ \ {}^tW^{(l+1)} \delta^{(l+1)}
\end{aligned}
$$

2.4 Formule 12

Montrons ce résultat au rang L, puis dans un cas quelconque. On a pour rappel :

$$
C(b_n^{(L)}) = \frac{1}{2} \sum_{j=1}^{n_L = 2} \left(y(x^i) - \sigma_L \Big(b_j^{(L)} + \sum_{i=1}^{n_{L-1}} w_{j,i}^{(L)} \ \sigma_{L-1}(z_i^{(L-1)}) \Big) \right)^2
$$

[4] A part que l'on a pas vraiment ${}^tW^{(l)}$ qui sort aussi rapidement, on a juste une somme de coefficients, mais lorsque l'on passe au calcul de $\delta^{(l)}$, la matrice se recompose. Il faut voir cela comme une simplification pour éviter les calculs trop longs et pas compréhensible.

avec $n \in \{1,2\}$. En remarquant que lorsque l'on dérivé par rapport à $b_j^{(L)}$, on tombe sur quelque chose de style $((y(x^i) - \sigma_L)^2)' = 2(y(x^i) - \sigma_L)\sigma'_L$, avec

$$\frac{\partial \sigma_L}{\partial b_n^{(L)}} = \sigma'_L\Big(b_j^{(L)} + \sum_{i=1}^{n_{L-1}} w_{j,i}^{(L)}\ \sigma_{L-1}(z_i^{(L-1)})\Big)$$

En faisant de même pour tout les biais du rang L, on obtient

$$\begin{aligned}\frac{\partial C}{\partial b^{(L)}} = \begin{pmatrix} \frac{\partial C}{\partial b_1^{(L)}} \\ \frac{\partial C}{\partial b_2^{(L)}} \end{pmatrix} &= \begin{pmatrix} (a^{(L)} - y(x^i))\ \sigma'_L\Big(b_1^{(L)} + \sum_{i=1}^{n_{L-1}} w_{1,i}^{(L)}\ \sigma_{L-1}(z_i^{(L-1)})\Big) \\ (a^{(L)} - y(x^i))\ \sigma'_L\Big(b_2^{(L)} + \sum_{i=1}^{n_{L-1}} w_{2,i}^{(L)}\ \sigma_{L-1}(z_i^{(L-1)})\Big) \end{pmatrix} \\ &= \begin{pmatrix} (a^{(L)} - y(x^i))\ \sigma'_L(z_1^{(L)}) \\ (a^{(L)} - y(x^i))\ \sigma'_L(z_2^{(L)}) \end{pmatrix} = \begin{pmatrix} \delta_1^{(L)} \\ \delta_2^{(L)} \end{pmatrix}\end{aligned}$$

d'où on peut en extraire la formule (12) ...

Pour un rang l quelconque, c'est la même chose avec plus de calculs - pour changer -, dû aux dérivées de composées. En bref, on a :

$$C(b_n^{(l)}) = \frac{1}{2}\sum_{j=1}^{n_L}\left(y(x^i) - \sigma_L\left(b_j^{(L)} + \sum_{k=1}^{n_{L-1}} w_{j,1}^{(L)}\ \sigma_{L-1}\left(\cdots \sum_{n=1}^{n_l} \sigma_l\Big(b_n^{(l)} + \sum_{m=1}^{n_{l-1}} w_{n,m}^{(l)}\ z_m^{(l-1)}\Big)\right)\right)\right)^2$$

Encore une fois, on peut dériver C en fonction de $b_n^{(l)}$, et comme vu précédemment, la dérivée est du style

$$\big((y - \sigma_L \circ \cdots \sigma_l)^2\big)' = 2(a^{(L)} - y)\sigma'_l \times \sigma_{l+1}(\sigma_l) \cdots \sigma'_L(\sigma_{L-1} \circ \cdots \sigma_l)$$

ce qui ressemble beaucoup à la formule (11) pour $z^{(l-1)}$, non ? Cela est du au fait que l'on a

$$\frac{\partial \sigma_l}{\partial b_n^{(l)}} = \sigma'_l(b_n + \sum_{i=1}^{n_{l-1}} w_{i,n}\ \sigma(z_i^{(L)})) = \sigma'_l(z_n^{(l)})$$

On peut en conclure - en généralisant la formule précédente pour tout les biais du rang l, que l'on a :

$$\frac{\partial C}{\partial b^{(l)}} = \begin{pmatrix} \frac{\partial C}{\partial b_1^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial b_n^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial b_{n_l}^{(l)}} \end{pmatrix} = \begin{pmatrix} \frac{\partial C}{\partial z_1^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial z_n^{(l)}} \\ \vdots \\ \frac{\partial C}{\partial z_{n_l}^{(l)}} \end{pmatrix} = \begin{pmatrix} \delta_1^{(l)} \\ \vdots \\ \delta_n^{(l)} \\ \vdots \\ \delta_{n_l}^{(l)} \end{pmatrix} = \delta^{(l)}$$

2.5 Formule 13

Montrons encore une fois tout d'abord ce résultat pour le rang L. On a, en définissant C en fonction de $w_{j,i}$:

$$C(w_{n,m}^{(L)}) = \frac{1}{2}\sum_{j=1}^{nL=2}\left(y - \sigma_L\Big(b_j^{(L)} + \sum_{i=1}^{n_{L-1}} w_{j,i}^{(L)}\ \sigma_{L-1}(z_i^{(L-1)})\Big)\right)^2$$

En particulier, on a $w_{n,m}^{(L)}\sigma(z_m^{(L)})$ qui est le seul terme qui dépend de $w_{n,m}^{(L)}$. Notons que dans σ_L, c'est $b_n + \sum_{i=1}^{n_{L-1}} w_{j,i}\ \sigma_{L-1}(z_i^{(L-1)})$ qui varie, et donc lors du calcul de la dérivée partielle, ce terme que l'on retrouvera, c'est à dire on a :

$$\frac{\partial C}{\partial w_{n,m}^{(L)}} = \sigma_L'(z_n^{(L)}) \times (a^{(L)} - y) \times \sigma_{L-1}(z_m^{(L-1)}) = \delta_j^{(L)} a_k^{(L-1)}$$

Pour un rang quelconque l, c'est la même logique, avec $w_{n,m}^{(l)}$ on aura le terme $\sigma_l(z_m^{(l)})$ qui sortira lors du calcul de la dérivée partielle :

$$\begin{aligned}\frac{\partial C}{\partial w_{n,m}^{(l)}} &= \left(\sigma'(z^{(l)})\ {}^tW^{(l)} \times \cdots \sigma_L'(z^{(L)})\right)_n \times (a^{(L)} - y) \times \sigma_{l-1}(z_m^{(L-1)}) \\ &= \delta_n \sigma_{l-1}(z_m^{(l-1)}) = \delta_n\ a_m^{(l-1)}\end{aligned}$$

Et c'est donc ainsi que l'on obtient la formule (15)

3 Exercice 2 - Formule 14

Introduisons donc maintenant la matrice $D^{(l)}$. C'est une matrice diagonale (donc carré) avec les termes fait des $\sigma'(z_i(l))$.

$$D^{(l)} = \begin{pmatrix} \sigma'(z_1^{(l)}) & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \sigma'(z_i^{(l)}) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma'(z_{n_l}^{(l)}) \end{pmatrix}$$

Avec cela, montrons donc la formule (14) par récurence.

Pour le rang L, on a :

$$\begin{aligned}
D^{(L)}\left(a^{(L)}-y\right) &= \begin{pmatrix} \sigma'(z_1^{(L)}) & 0 \\ 0 & \sigma'(z_2^{(L)}) \end{pmatrix} \begin{pmatrix} a_1^{(L)}-y \\ a_2^{(L)}-y \end{pmatrix} \\
&= \begin{pmatrix} \sigma'(z_1^{(L)})(a_1^{(L)}-y) \\ \sigma'(z_2^{(L)})(a_2^{(L)}-y) \end{pmatrix} = \delta^{(L)}
\end{aligned}$$

Supposons que la formule est vraie pour le rang $l+1$, montrons que cela reste vrai pour le rang l:

$$D^{(L)}\ {}^tW^{(l+1)}D^{(l+1)}\ {}^tW^{(l+1)}\cdots D^{(L-1)}\ {}^tW^{(L)}D^{(L)}(a^{(L)}-y)D^{(L)}\ {}^tW^{(l+1)}\delta^{(l+1)}$$

$$= D^{(L)}\ {}^tW^{(l+1)}\delta^{(l+1)}$$

$$= \begin{pmatrix} \sigma'(z_1^{(l)}) & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \sigma'(z_i^{(l)}) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma'(z_{n_l}^{(l)}) \end{pmatrix} \begin{pmatrix} w_{1,1} & \cdots & w_{i,1} & \cdots & w_{n_{l+1},1} \\ \vdots & \ddots & & & \vdots \\ w_{1,j} & \cdots & w_{i,j} & \cdots & w_{n_{l+1},j} \\ \vdots & & & \ddots & \vdots \\ w_{1,n_l} & \cdots & w_{i,n_l} & \cdots & w_{n_{l+1},n_l} \end{pmatrix} \begin{pmatrix} \delta_1^{(l+1)} \\ \vdots \\ \delta_j^{(l+1)} \\ \vdots \\ \delta_{n_{l+1}}^{(l+1)} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma'(z_1^{(l)}) & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \sigma'(z_i^{(l)}) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma'(z_{n_l}^{(l)}) \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_{l+1}} w_{i,1}\delta_i^{(l+1)} \\ \vdots \\ \sum_{j=1}^{n_{l+1}} w_{i,j}\delta_j^{(l+1)} \\ \vdots \\ \sum_{j=1}^{n_{l+1}} w_{i,n_l}\delta_{n_{l+1}}^{(l+1)} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma'(z_1^{(l)}) \times \sum_{j=1}^{n_{l+1}} w_{i,1}\delta_i^{(l+1)} \\ \vdots \\ \sigma'(z_j^{(l)}) \times \sum_{j=1}^{n_{l+1}} w_{i,j}\delta_j^{(l+1)} \\ \vdots \\ \sigma'(z_{n_l}^{(l)}) \times \sum_{j=1}^{n_{l+1}} w_{i,n_l}\delta_{n_{l+1}}^{(l+1)} \end{pmatrix} = \sigma'(z^{(l)}) \ \circ\ {}^tW^{(l+1)}\delta^{(l+1)} = \delta^{(l)}$$

C'est donc ainsi que s'achève l'écriture des calculs montrant les formules demandées. Si je peux me permettre un mot pour cet exercice, il est vrai qu'il m'a permis de beaucoup mieux comprendre comment cela fonctionnait, et ça m'as beaucoup éclaircies les idées sur le code qu'il me reste à faire. Enfin, je vous remercie pour votre attention à la suite de la lecture de ce document.