

# Notes on analytic gradient derivation

Artyom Bondartsov

November 28, 2019

## Abstract

The document illustrates the process of analytic gradient derivation for a neural network devised for the notorious iris dataset. It shows how neural network diagrams may be of great help when it comes to application of the chain rule correctly. Vectorized forms are also stated so that the derived gradients could be more efficiently implemented using Matlab or a linear algebra package Numpy from Python.

## Neural network

Having the iris dataset as input, we let  $Sx$  be a set of 4-dimensional input vectors,

$$Sx = \left\{ ..., \begin{bmatrix} x_1 \\ \vdots \\ x_4 \end{bmatrix}, ... \right\}.$$

We denote  $|Sx|$  to be the total number of such vectors in the set  $Sx$ .

We also have a set  $Sy$  of *one-hot* vectors transformed from the iris dataset's result variable with three classes,

$$Sy = \left\{ ..., \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, ... \right\}.$$

Please note that although  $|Sy|$  is never used in the derivation that follows, yet for the sake of clarity  $|Sx| = |Sy|$ .

And finally we define a neural network as depicted in the figure (1) below

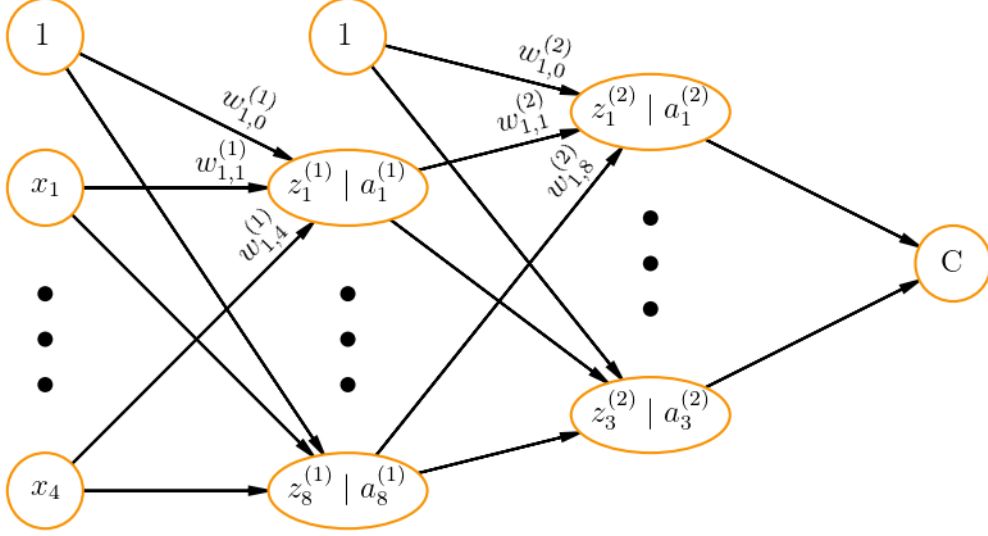


Figure 1: A neural network for the iris dataset has one input layer with 4 inputs (+bias), one hidden layer of 8 neurons (+bias) and an output layer of 3 neurons by the number of iris' classes

In addition we define neural network's functions as below. Note that these functions are pretty standard for neural networks and given here only as a reminder.

A sum of weighted inputs from the zero (input) layer that enters an arbitrary neuron

$$z_i^{(1)} = w_{i,0}^{(1)} + w_{i,1}^{(1)} x_1 + \dots + w_{i,4}^{(1)} x_4, i \in [1, 8].$$

Using dot product, in matrix notation for the entire layer it transforms into

$$\begin{aligned} \mathbf{z}^{(1)} &= \begin{bmatrix} w_{1,0}^{(1)} & w_{1,1}^{(1)} & \dots & w_{1,4}^{(1)} \\ \dots & \dots & \dots & \dots \\ w_{8,0}^{(1)} & w_{8,1}^{(1)} & \dots & w_{8,4}^{(1)} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_4 \end{bmatrix} \\ &= \mathbf{w}^{(1)\text{T}} \cdot \mathbf{x}. \end{aligned}$$

A sigmoid function of an arbitrary neuron of the first layer

$$a_i^{(1)} = \frac{1}{1 + e^{-z_i^{(1)}}}, i \in [1, 8].$$

A sum of weighted inputs from the first (hidden) layer that enters an arbitrary neuron

$$z_i^{(2)} = w_{i,0}^{(2)} + w_{i,1}^{(2)} a_1^{(1)} + \dots + w_{i,8}^{(2)} a_8^{(1)}, i \in [1, 3].$$

Using dot product, in matrix notation for the entire layer it transforms into

$$\begin{aligned} \mathbf{z}^{(2)} &= \begin{bmatrix} w_{1,0}^{(2)} & w_{1,1}^{(2)} & \dots & w_{1,8}^{(2)} \\ \dots & \dots & \dots & \dots \\ w_{3,0}^{(2)} & w_{3,1}^{(2)} & \dots & w_{3,8}^{(2)} \end{bmatrix} \begin{bmatrix} 1 \\ a_1^{(1)} \\ \vdots \\ a_8^{(1)} \end{bmatrix} \\ &= \mathbf{w}^{(2)\text{T}} \cdot \mathbf{a}^{(1)}. \end{aligned}$$

A sigmoid function of an arbitrary neuron of the second layer

$$a_i^{(2)} = \frac{1}{1 + e^{-z_i^{(2)}}}, i \in [1, 3].$$

And a cross entropy loss function

$$C = -\frac{1}{|Sx|} \sum_{Sx} \sum_{j=1}^3 y_j \ln a_j^{(2)} + (1 - y_j) \ln(1 - a_j^{(2)}). \quad (1)$$

Finally, we are interested in obtaining a gradient of (1) with respect to the weights. In particular we will be deriving

$$\nabla_{w^{(2)}} C = \begin{bmatrix} \frac{\partial C}{\partial w_{1,0}^{(2)}} & \frac{\partial C}{\partial w_{1,1}^{(2)}} & \cdots & \frac{\partial C}{\partial w_{1,8}^{(2)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial C}{\partial w_{3,0}^{(2)}} & \frac{\partial C}{\partial w_{3,1}^{(2)}} & \cdots & \frac{\partial C}{\partial w_{3,8}^{(2)}} \end{bmatrix} \quad (2)$$

$$\nabla_{w^{(1)}} C = \begin{bmatrix} \frac{\partial C}{\partial w_{1,0}^{(1)}} & \frac{\partial C}{\partial w_{1,1}^{(1)}} & \cdots & \frac{\partial C}{\partial w_{1,4}^{(1)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial C}{\partial w_{8,0}^{(1)}} & \frac{\partial C}{\partial w_{8,1}^{(1)}} & \cdots & \frac{\partial C}{\partial w_{8,4}^{(1)}} \end{bmatrix}. \quad (3)$$

## Gradient with respect to $\mathbf{w}^{(2)}$

We will start from the formula (2). Yet instead of calculating the gradient matrix in its entirety we are going to derive a partial derivative in general form first, namely  $\frac{\partial C}{\partial w_{i,j}^{(2)}}$ .

In order to spot all parts of (1) that depend on some arbitrary  $w_{i,j}^{(2)}$  we will use a figure (2) that purposefully depicts only those dependencies

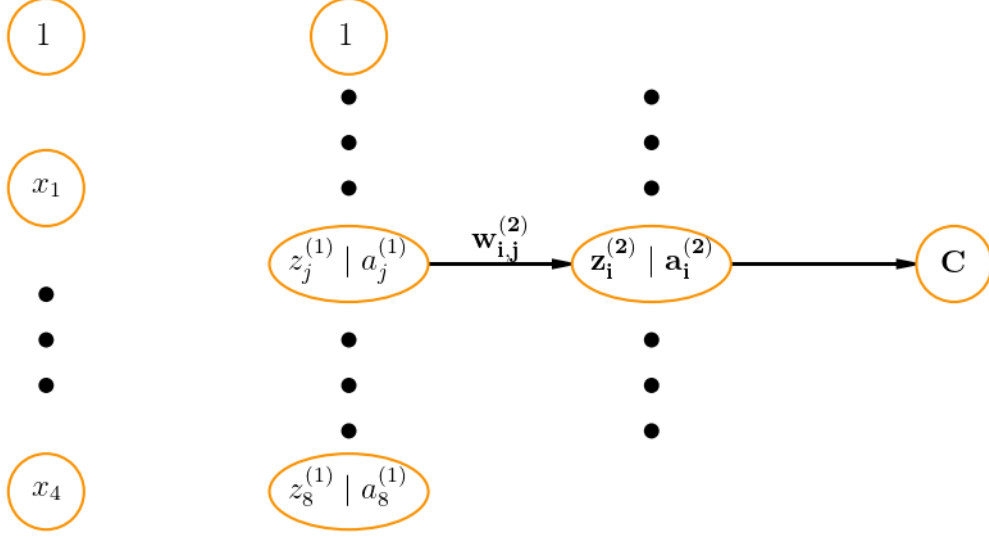


Figure 2: The diagram contains no arrows but those connecting the loss function  $C$  to a particular weight  $w_{i,j}^{(2)}$

Now, having all the dependencies before the very eyes it is easy to write down the chain rule formula following backwards from the loss function  $C$  to some arbitrary weight  $w_{i,j}^{(2)}$  in the picture

$$\frac{\partial C}{\partial w_{i,j}^{(2)}} = \frac{\partial C}{\partial a_i^{(2)}} \frac{\partial a_i^{(2)}}{\partial z_i^{(2)}} \frac{\partial z_i^{(2)}}{\partial w_{i,j}^{(2)}}. \quad (4)$$

Taking those partial derivatives independently we get

$$\begin{aligned}
\frac{\partial C}{\partial a_i^{(2)}} &= \frac{\partial}{\partial a_i^{(2)}} \left( -\frac{1}{|Sx|} \sum_{Sx} \sum_{j=1}^3 y_j \ln a_j^{(2)} + (1 - y_j) \ln(1 - a_j^{(2)}) \right) \\
&= -\frac{1}{|Sx|} \sum_{Sx} y_i \frac{1}{a_i^{(2)}} + (1 - y_i) \frac{1}{1 - a_i^{(2)}} (-1) \\
&= \frac{1}{|Sx|} \sum_{Sx} \frac{a_i^{(2)} - y_i}{a_i^{(2)}(1 - a_i^{(2)})}. \tag{5a}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial a_i^{(2)}}{\partial z_i^{(2)}} &= \frac{\partial}{\partial z_i^{(2)}} \frac{1}{1 + e^{-z_i^{(2)}}} \\
&= \frac{\partial}{\partial z_i^{(2)}} (1 + e^{-z_i^{(2)}})^{-1} \\
&= -(1 + e^{-z_i^{(2)}})^{-2} (-1) e^{-z_i^{(2)}} \\
&= \frac{e^{-z_i^{(2)}}}{(1 + e^{-z_i^{(2)}})^2} \\
&= \frac{1 + e^{-z_i^{(2)}} - 1}{(1 + e^{-z_i^{(2)}})^2} \\
&= \frac{\cancel{1 + e^{-z_i^{(2)}}} - 1}{(1 + e^{-z_i^{(2)}})^2} - \frac{1}{(1 + e^{-z_i^{(2)}})^2} \\
&= \frac{1}{1 + e^{-z_i^{(2)}}} - \frac{1}{(1 + e^{-z_i^{(2)}})^2} \\
&= a_i^{(2)} - (a_i^{(2)})^2 \\
&= a_i^{(2)}(1 - a_i^{(2)}). \tag{5b}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial z_i^{(2)}}{\partial w_{i,j}^{(2)}} &= \frac{\partial}{\partial w_{i,j}^{(2)}} (w_{i,0}^{(2)} + w_{i,1}^{(2)} a_1^{(1)} + \dots + w_{i,j}^{(2)} a_j^{(1)} + \dots + w_{i,8}^{(2)} a_8^{(1)}) \\
&= \begin{cases} 1 & , j = 0 \\ a_j^{(1)} & , j \neq 0. \end{cases} \tag{5c}
\end{aligned}$$

And finally combining (5a), (5b) and (5c) together we obtain

$$\begin{aligned}
\frac{\partial C}{\partial w_{i,j}^{(2)}} &= \frac{\partial C}{\partial a_i^{(2)}} \frac{\partial a_i^{(2)}}{\partial z_i^{(2)}} \frac{\partial z_i^{(2)}}{\partial w_{i,j}^{(2)}} \\
&= \frac{1}{|Sx|} \sum_{Sx} \frac{a_i^{(2)} - y_i}{a_i^{(2)}(1 - a_i^{(2)})} \cancel{a_i^{(2)}(1 - a_i^{(2)})} \frac{\partial z_i^{(2)}}{\partial w_{i,j}^{(2)}} \\
&= \frac{1}{|Sx|} \sum_{Sx} (a_i^{(2)} - y_i) \begin{cases} 1 & , j = 0 \\ a_j^{(1)} & , j \neq 0 \end{cases} \\
&= \begin{cases} \frac{1}{|Sx|} \sum_{Sx} (a_i^{(2)} - y_i) & , j = 0 \\ \frac{1}{|Sx|} \sum_{Sx} (a_i^{(2)} - y_i) a_j^{(1)} & , j \neq 0. \end{cases} \tag{6}
\end{aligned}$$

Using (6) our matrix gradient (2) now takes the form

$$\begin{aligned}
\nabla_{w^{(2)}} C &= \begin{bmatrix} \frac{\partial C}{\partial w_{1,0}^{(2)}} & \frac{\partial C}{\partial w_{1,1}^{(2)}} & \cdots & \frac{\partial C}{\partial w_{1,8}^{(2)}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial C}{\partial w_{3,0}^{(2)}} & \frac{\partial C}{\partial w_{3,1}^{(2)}} & \cdots & \frac{\partial C}{\partial w_{3,8}^{(2)}} \end{bmatrix} \\
&= \frac{1}{|Sx|} \begin{bmatrix} \sum_{Sx} (a_1^{(2)} - y_1) & \sum_{Sx} (a_1^{(2)} - y_1) a_1^{(1)} & \cdots & \sum_{Sx} (a_1^{(2)} - y_1) a_8^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{Sx} (a_3^{(2)} - y_3) & \sum_{Sx} (a_3^{(2)} - y_3) a_1^{(1)} & \cdots & \sum_{Sx} (a_3^{(2)} - y_3) a_8^{(1)} \end{bmatrix}.
\end{aligned} \tag{7}$$

In addition, (7) can also take a nice vectorized form using dot product

$$\begin{aligned}
\nabla_{w^{(2)}} C &= \frac{1}{|Sx|} \begin{bmatrix} a_{1,1}^{(2)} - y_{1,1} & \cdots & a_{1,|Sx|}^{(2)} - y_{1,|Sx|} \\ \cdots & \cdots & \cdots \\ a_{3,1}^{(2)} - y_{3,1} & \cdots & a_{3,|Sx|}^{(2)} - y_{3,|Sx|} \end{bmatrix} \begin{bmatrix} 1 & a_{1,1}^{(1)} & \cdots & a_{1,8}^{(1)} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{|Sx|,1}^{(1)} & \cdots & a_{|Sx|,8}^{(1)} \end{bmatrix} \\
&= \frac{1}{|Sx|} (\mathbf{a}^{(2)} - \mathbf{y}) \cdot \mathbf{a}^{(1)\mathbf{T}}.
\end{aligned} \tag{8}$$

Please note that  $\mathbf{a}^{(1)\mathbf{T}}$  includes a bias first column.

## Gradient with respect to $\mathbf{w}^{(1)}$

In order to obtain (3) we will use the same approach, i.e. we will derive a general formula of partial derivative, namely  $\frac{\partial C}{\partial w_{i,j}^{(1)}}$ .

Once again we are going to employ graphical approach to spot all parts of (1) that would get changed if some arbitrary  $w_{i,j}^{(1)}$  was wiggled.

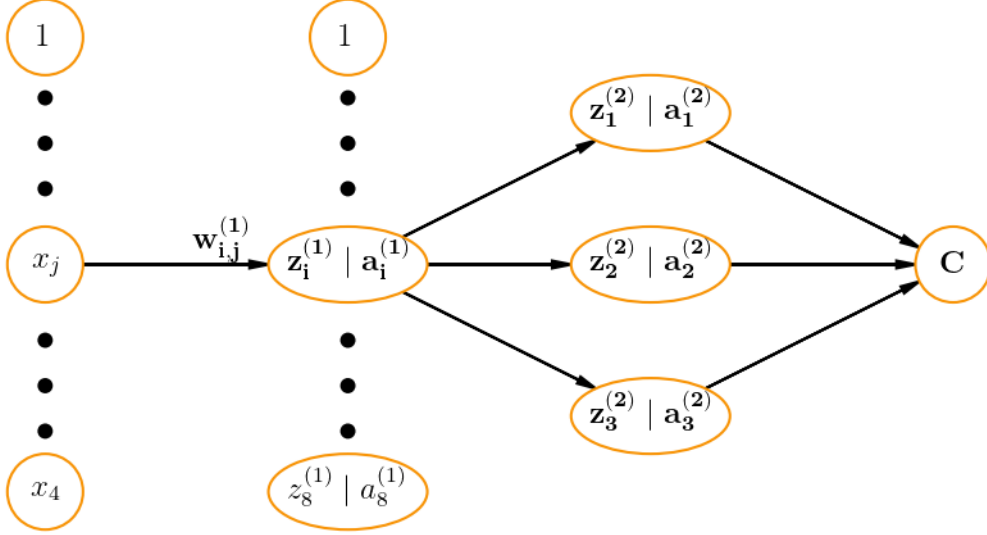


Figure 3: This time from the loss function  $C$  to a particular weight  $w_{i,j}^{(1)}$  leads multiple paths

Starting from  $C$  in figure (3) and going backwards to the arbitrary weight  $w_{i,j}^{(1)}$  we write down

$$\begin{aligned}
 \frac{\partial C}{\partial w_{i,j}^{(1)}} &= \frac{\partial C}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} \frac{\partial z_1^{(2)}}{\partial a_i^{(1)}} \frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} \frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}} + \frac{\partial C}{\partial a_2^{(2)}} \frac{\partial a_2^{(2)}}{\partial z_2^{(2)}} \frac{\partial z_2^{(2)}}{\partial a_i^{(1)}} \frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} \frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}} \\
 &\quad + \frac{\partial C}{\partial a_3^{(2)}} \frac{\partial a_3^{(2)}}{\partial z_3^{(2)}} \frac{\partial z_3^{(2)}}{\partial a_i^{(1)}} \frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} \frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}} \\
 &= \left( \sum_{k=1}^3 \frac{\partial C}{\partial a_k^{(2)}} \frac{\partial a_k^{(2)}}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial a_i^{(1)}} \right) \frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} \frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}}.
 \end{aligned} \tag{9}$$

Please note how the multiple paths in the picture have been transformed into summation in the formula. The chain rule (9) looks scary yet let us recall that we have already carried out derivations for some of these partial derivatives. Specifically,  $\frac{\partial C}{\partial a_k^{(2)}}$  is (5a) and  $\frac{\partial a_k^{(2)}}{\partial z_k^{(2)}}$  is (5b). Hence only 3 partial derivatives left, and one of them,  $\frac{\partial a_i^{(1)}}{\partial z_i^{(1)}}$ , is in fact also (5b) but with a different upper index. Thus, taking care of the rest of the derivatives

independently we get

$$\begin{aligned}\frac{\partial z_k^{(2)}}{\partial a_i^{(1)}} &= \frac{\partial}{\partial a_i^{(1)}} \left( w_{k,0}^{(2)} + w_{k,1}^{(2)} a_1^{(1)} + \dots + w_{k,i}^{(2)} a_i^{(1)} + \dots + w_{k,8}^{(2)} a_8^{(1)} \right) \\ &= w_{k,i}^{(2)}.\end{aligned}\tag{10a}$$

$$\begin{aligned}\frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}} &= \frac{\partial}{\partial w_{i,j}^{(1)}} \left( w_{i,0}^{(1)} + w_{i,1}^{(1)} x_1 + \dots + w_{i,j}^{(1)} x_j + \dots + w_{i,4}^{(1)} x_4 \right) \\ &= \begin{cases} 1 & , j = 0 \\ x_j & , j \neq 0 \end{cases}.\end{aligned}\tag{10b}$$

$$\frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} = a_i^{(1)}(1 - a_i^{(1)}).\tag{10c}$$

Now combining (5a), (5b), (10a), (10b) and (10c) we have

$$\begin{aligned}\frac{\partial C}{\partial w_{i,j}^{(1)}} &= \left( \sum_{k=1}^3 \frac{\partial C}{\partial a_k^{(2)}} \frac{\partial a_k^{(2)}}{\partial z_k^{(2)}} \frac{\partial z_k^{(2)}}{\partial a_i^{(1)}} \right) \frac{\partial a_i^{(1)}}{\partial z_i^{(1)}} \frac{\partial z_i^{(1)}}{\partial w_{i,j}^{(1)}} \\ &= \left( \sum_{k=1}^3 \frac{1}{|Sx|} \sum_{Sx} \frac{a_k^{(2)} - y_k}{a_k^{(2)}(1 - a_k^{(2)})} a_k^{(2)}(1 - a_k^{(2)}) w_{k,i}^{(2)} \right) a_i^{(1)}(1 - a_i^{(1)}) \begin{cases} 1 & , j = 0 \\ x_j & , j \neq 0 \end{cases} \\ &= \frac{1}{|Sx|} \sum_{Sx} \left( \sum_{k=1}^3 (a_k^{(2)} - y_k) w_{k,i}^{(2)} \right) a_i^{(1)}(1 - a_i^{(1)}) \begin{cases} 1 & , j = 0 \\ x_j & , j \neq 0 \end{cases} \\ &= \begin{cases} \frac{1}{|Sx|} \sum_{Sx} \left( \sum_{k=1}^3 (a_k^{(2)} - y_k) w_{k,i}^{(2)} \right) a_i^{(1)}(1 - a_i^{(1)}) & , j = 0 \\ \frac{1}{|Sx|} \sum_{Sx} \left( \sum_{k=1}^3 (a_k^{(2)} - y_k) w_{k,i}^{(2)} \right) a_i^{(1)}(1 - a_i^{(1)}) x_j & , j \neq 0 \end{cases}.\end{aligned}\tag{11}$$

Before proceeding to the matrix gradient (3) we are going to introduce some additional notation for the sake of simplification. We will denote

$$f_i^{(2)} = \left( \sum_{k=1}^3 (a_k^{(2)} - y_k) w_{k,i}^{(2)} \right), i \in [1, 8]\tag{12a}$$

$$g_i^{(1)} = a_i^{(1)}(1 - a_i^{(1)}), i \in [1, 8].\tag{12b}$$

Using (12a) and (12b) the matrix gradient (3) now takes the form

$$\begin{aligned}\nabla_{w^{(1)}} C &= \begin{bmatrix} \frac{\partial C}{\partial w_{1,0}^{(1)}} & \frac{\partial C}{\partial w_{1,1}^{(1)}} & \dots & \frac{\partial C}{\partial w_{1,4}^{(1)}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial C}{\partial w_{8,0}^{(1)}} & \frac{\partial C}{\partial w_{8,1}^{(1)}} & \dots & \frac{\partial C}{\partial w_{8,4}^{(1)}} \end{bmatrix} \\ &= \frac{1}{|Sx|} \begin{bmatrix} \sum_{Sx} g_1^{(1)} f_1^{(2)} & \sum_{Sx} g_1^{(1)} f_1^{(2)} x_1 & \dots & \sum_{Sx} g_1^{(1)} f_1^{(2)} x_4 \\ \dots & \dots & \dots & \dots \\ \sum_{Sx} g_8^{(1)} f_8^{(2)} & \sum_{Sx} g_8^{(1)} f_8^{(2)} x_1 & \dots & \sum_{Sx} g_8^{(1)} f_8^{(2)} x_4 \end{bmatrix}\end{aligned}\tag{13}$$



In addition, (13) can also take a nice vectorized form using dot product

$$\begin{aligned}\nabla_{w^{(1)}}C &= \frac{1}{|Sx|} \begin{bmatrix} g_{1,1}^{(1)}f_{1,1}^{(2)} & \cdots & g_{1,|Sx|}^{(1)}f_{1,|Sx|}^{(2)} \\ \cdots & \cdots & \cdots \\ g_{8,1}^{(1)}f_{8,1}^{(2)} & \cdots & g_{8,|Sx|}^{(1)}f_{8,|Sx|}^{(2)} \end{bmatrix} \cdot \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,4} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{|Sx|,1} & \cdots & x_{|Sx|,4} \end{bmatrix} \\ &= \frac{1}{|Sx|} (\mathbf{g}^{(1)} \odot \mathbf{f}^{(2)}) \cdot \mathbf{x}^T,\end{aligned}\tag{14}$$

where  $\odot$  is Hadamard (element-wise) product.

Please note that  $\mathbf{x}^T$  includes a bias first column. Heed that (12a) can also be vectorized if needed

$$\begin{aligned}\mathbf{f}^{(2)} &= \begin{bmatrix} a_{1,1}^{(2)} - y_{1,1} & \cdots & a_{1,3}^{(2)} - y_{1,3} \\ \cdots & \cdots & \cdots \\ a_{|Sx|,1}^{(2)} - y_{|Sx|,1} & \cdots & a_{|Sx|,3}^{(2)} - y_{|Sx|,3} \end{bmatrix} \begin{bmatrix} w_{1,1}^{(2)} & \cdots & w_{1,8}^{(2)} \\ \vdots & \vdots & \vdots \\ w_{3,1}^{(2)} & \cdots & w_{3,8}^{(2)} \end{bmatrix} \\ &= (\mathbf{a}^{(2)} - \mathbf{y})^T \cdot \mathbf{w}^{(2)}.\end{aligned}\tag{15}$$