Logistic Regression Gradient Descent Detailed Derivative Derivations

1. Variables:

The following table provides an inventory of all the variables used:

#	Name	Description
1	m	Sample size
2	n_x	Number of input features
3	$x_j^{(i)}$	$j$the input feature of the ith sample
4	$\vec{x}^{(i)}$	Equal to $\left[x_1^{(i)}, x_2^{(i)}, \dots, x_{n_x}^{(i)}\right]^T$
5	X	Equal to $\left[\vec{x}^{(1)}, \vec{x}^{(2)}, \dots, \vec{x}^{(m)}\right]$
6	w_j	jth input feature weight
7	$\vec{w}$	Equal to $\left[w_1, w_2, \dots, w_{n_x}\right]^T$
8	b	Bias
9	$\mathbf{1}_{n\times p}$	Matrix of ones with n rows and p columns
10	$z^{(i)}$	Equal to $\vec{w}^T x^{(i)} + b\mathbf{1}_{m\times 1}$
11	$\vec{z}$	Equal to $\left[z^{(1)}, z^{(2)}, \dots, z^{(m)}\right]$
12	$g(x)$	Transformation function on variable x. Here, we set $g(x) = \frac{1}{1+e^{-x}}$, which is the sigmoid function
13	$a^{(i)}$	Activation value of the ith sample. Equal to $g\left(z^{(i)}\right)$
14	$g^*\left(\vec{v}_{1\times p}\right)$	For $\vec{v}_{1\times p} = \left[x_1, x_2, \dots, x_p\right]$, equal to $\left[g(x_1), g(x_2), \dots, g\left(x_p\right)\right]$
15	$\vec{a}$	Equal to $g^*(\vec{z})$
16	$y^{(i)}$	Output variable of the ith sample. Possible values $\in \{0,1\}$
17	Y	Equal to $\left[y^{(1)}, y^{(2)}, \dots, y^{(m)}\right]$
18	$\mathcal{L}\left(a^{(i)}, y^{(i)}\right)$	Loss value of the ith sample. Set equal to $-\left(y^{(i)} \ln\left(a^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - a^{(i)}\right)\right)$
19	$\mathcal{J}$	Loss value of the entire training dataset. Set equal to $\frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(a^{(i)}, y^{(i)}\right)$

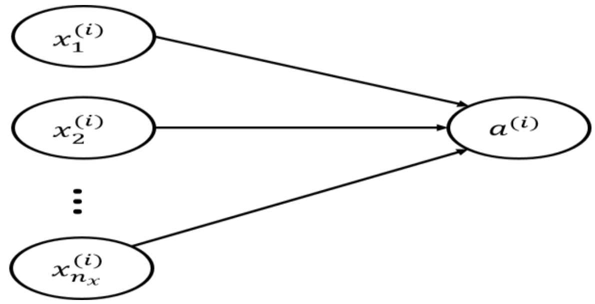
2. Vector Derivative Convention

We use the so-called numerator layout convention: let $\vec{v}_{n\times 1} = [v_1, v_2, \dots, v_n]^T$, $\vec{u}_{1\times p} = \left[u_1, u_2, \dots, u_p\right]$, $f_u\left(\vec{u}_{1\times p}\right)$ a function of the vector $\vec{u}_{1\times p}$ mapping onto $\mathbb{R}$. Furthermore, let $\vec{v}_{n\times 1} = h\left(\vec{u}_{1\times p}\right)$.We thus have the following results based on the chosen convention:

- $\frac{\partial h(\vec{u}_{1\times p})}{\partial u_i} = \frac{\partial \vec{v}_{n\times 1}}{\partial u_i} = \begin{bmatrix} \frac{\partial v_1}{\partial u_i} \\ \frac{\partial v_n}{\partial u_i} \\ \vdots \\ \frac{\partial v_n}{\partial u_i} \end{bmatrix}$
- $\frac{\partial \left(h(\vec{u}_{1\times p})\right)^T}{\partial u_i} = \frac{\partial (\vec{v}_{n\times 1})^T}{\partial u_i} = \begin{bmatrix} \frac{\partial v_1}{\partial u_i} \\ \frac{\partial v_n}{\partial u_i} \\ \vdots \\ \frac{\partial v_n}{\partial u_i} \end{bmatrix}^T$
-
- $\frac{\partial f_u(\vec{u}_{1\times p})}{\partial \vec{u}_{1\times p}} = \begin{bmatrix} \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_1} & \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_2} & \cdots & \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_p} \end{bmatrix}$
- $\frac{\partial f_u(\vec{u}_{1\times p})}{\partial (\vec{u}_{1\times p})^T} = \begin{bmatrix} \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_1} & \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_2} & \cdots & \frac{\partial f_u(\vec{u}_{1\times p})}{\partial u_p} \end{bmatrix}^T$
-
- $\frac{\partial h(\vec{u}_{1\times p})}{\partial \vec{u}_{1\times p}} = \begin{bmatrix} \frac{\partial h(\vec{u}_{1\times p})}{\partial u_1} & \frac{\partial h(\vec{u}_{1\times p})}{\partial u_2} & \cdots & \frac{\partial h(\vec{u}_{1\times p})}{\partial u_p} \end{bmatrix} = \begin{bmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_n}{\partial u_1} & \frac{\partial v_n}{\partial u_2} & \cdots & \frac{\partial v_n}{\partial u_p} \end{bmatrix}$
- $\frac{\partial h(\vec{u}_{1\times p})^T}{\partial \vec{u}_{1\times p}^T} = \begin{bmatrix} \frac{\partial h(\vec{u}_{1\times p})}{\partial u_1} & \frac{\partial h(\vec{u}_{1\times p})}{\partial u_2} & \cdots & \frac{\partial h(\vec{u}_{1\times p})}{\partial u_p} \end{bmatrix}^T = \begin{bmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_2}{\partial u_1} & \cdots & \frac{\partial v_n}{\partial u_1} \\ \frac{\partial v_1}{\partial u_2} & \frac{\partial v_2}{\partial u_2} & \cdots & \frac{\partial v_n}{\partial u_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_1}{\partial u_p} & \frac{\partial v_2}{\partial u_p} & \cdots & \frac{\partial v_n}{\partial u_p} \end{bmatrix} = \begin{bmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_1}{\partial u_2} & \cdots & \frac{\partial v_1}{\partial u_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial v_n}{\partial u_1} & \frac{\partial v_n}{\partial u_2} & \cdots & \frac{\partial v_n}{\partial u_p} \end{bmatrix}^T$
-

3. Model Setup

The following diagram provides a visual representation of the model for training sample i:



The following formulas describe the model behaviour:

- Individual sample i:
  - $a^{(i)} = g\left(z^{(i)}\right) = \frac{1}{1+e^{-z^{(i)}}}$
  - $z^{(i)} = \vec{w}^T \vec{x}^{(i)} + b$
  - $\mathcal{L}\left(a^{(i)}, y^{(i)}\right) = -\left(y^{(i)} \ln\left(a^{(i)}\right) + \left(1 - y^{(i)}\right) \ln\left(1 - a^{(i)}\right)\right)$
- Training dataset:
  - $\vec{a} = \left[a^{(1)}, a^{(2)}, \dots, a^{(m)}\right] = \left[g\left(z^{(1)}\right), g\left(z^{(2)}\right), \dots, g\left(z^{(m)}\right)\right] = g^*(\vec{z})$
  - $\vec{z} = \left[z^{(1)}, z^{(2)}, \dots, z^{(m)}\right] = \left[\vec{w}^T \vec{x}^{(1)} + b, \vec{w}^T \vec{x}^{(2)} + b, \dots, \vec{w}^T \vec{x}^{(m)} + b\right]$
  - $\mathcal{J} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(a^{(i)}, y^{(i)}\right)$

4. Goal

Find the parameter values $w_1, w_2, \dots, w_{n_x}, b$ that minimize $\mathcal{J}$

5. Method

Gradient descent: An iterative algorithm that recalibrates the parameter values systematically reducing $\mathcal{J}$. Formulas (square bracket superscript reflects iteration number):

- w_j's, where $j = 1,2, \dots, n_x$:
  - $w_j^{[0]} = random\ number\ (typically\ chosen\ to\ be\ close\ to\ zero)$ → Vector form: $\vec{w}^{[0]}$
  - $w_j^{[k+1]} = w_j^{[k]} - \alpha \times \frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}}$, where α is a set learning rate (e.g., 0.005), and $k = 0,1, \dots, N$, where N is the total number of iterations → Vector form: $\vec{w}^{[k+1]} = \vec{w}^{[k]} - \alpha \times \frac{\partial \mathcal{J}^{[k]}}{\partial \vec{w}^{[k]}}$
- b:
  - $b^{[0]} = random\ number\ (typically\ chosen\ to\ be\ zero)$
  - $b^{[k+1]} = b^{[k]} - \alpha \times \frac{\partial \mathcal{J}^{[k]}}{\partial b^{[k]}}$, where α is a set learning rate (e.g., 0.005), and $k = 0,1, \dots, N$, where N is the total number of iterations

Notes:

- $\mathcal{J}^{[k]} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(a^{(i)[k]}, y^{(i)}\right)$; $y^{(i)}$ doesn't have a superscript k since its value does not change as we iterate

- We will prove that $\mathcal{J}^{[k]}$ is convex with respect to w_j's (making one minor assumption to guarantee strict convexity) and b (strictly convex), justifying putting a minus sign next to the α parameters

5.1 Deriving $\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}}$ and $\frac{\partial \mathcal{J}^{[k]}}{\partial b^{[k]}}$

Via the chain rule, we have the following results:

- $\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}} = \frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}} \times \frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} \times \frac{\partial \vec{z}^{[k]^T}}{\partial w_j^{[k]}}$ for $j = 1,2, \ldots, n_x$ and $k = 0,1, \ldots, N$
- $\frac{\partial \mathcal{J}^{[k]}}{\partial b^{[k]}} = \frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}} \times \frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} \times \frac{\partial \vec{z}^{[k]^T}}{\partial b^{[k]}}$ for $k = 0,1, \ldots, N$

We now derive the mathematical expression of each component separately:

5.1.1 Deriving $\frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}}$

$$\frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]^T}} = \left[\frac{\partial \mathcal{J}^{[k]}}{\partial a^{(1)[k]}}, \frac{\partial \mathcal{J}^{[k]}}{\partial a^{(2)[k]}}, \ldots, \frac{\partial \mathcal{J}^{[k]}}{\partial a^{(m)[k]}}\right]$$

$$= \left[\frac{\partial \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(a^{(i)[k]}, y^{(i)}\right)}{\partial a^{(1)[k]}}, \frac{\partial \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(a^{(i)[k]}, y^{(i)}\right)}{\partial a^{(2)[k]}}, \ldots, \frac{\partial \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}\left(a^{(i)[k]}, y^{(i)}\right)}{\partial a^{(m)[k]}}\right]$$

$$= \frac{1}{m}\left[\frac{\partial \mathcal{L}\left(a^{(1)[k]}, y^{(1)}\right)}{\partial a^{(1)[k]}}, \frac{\partial \mathcal{L}\left(a^{(2)[k]}, y^{(2)}\right)}{\partial a^{(2)[k]}}, \ldots, \frac{\partial \mathcal{L}\left(a^{(m)[k]}, y^{(m)}\right)}{\partial a^{(m)[k]}}\right]$$

For $i = 1,2, \ldots, m$,

$$\frac{\partial \mathcal{L}\left(a^{(i)[k]}, y^{(i)}\right)}{\partial a^{(i)[k]}} = -\left(\frac{y^{(i)}}{a^{(i)[k]}} - \frac{1-y^{(i)}}{1-a^{(i)[k]}}\right)$$

$$= \frac{-\left(y^{(i)} \times \left(1-a^{(i)[k]}\right) - a^{(i)[k]} \times \left(1-y^{(i)}\right)\right)}{a^{(i)[k]} \times \left(1-a^{(i)[k]}\right)}$$

$$= \frac{-\left(y^{(i)} - y^{(i)}a^{(i)[k]} - a^{(i)[k]} + a^{(i)[k]}y^{(i)}\right)}{a^{(i)[k]} \times \left(1-a^{(i)[k]}\right)}$$

$$= \frac{a^{(i)[k]} - y^{(i)}}{a^{(i)[k]} \times (1-a^{(i)[k]})}$$

Thus, $\frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}} = \frac{1}{m}\left[\frac{a^{(1)[k]}-y^{(1)}}{a^{(1)[k]} \times \left(1-a^{(1)[k]}\right)}, \frac{a^{(2)[k]}-y^{(2)}}{a^{(2)[k]} \times \left(1-a^{(2)[k]}\right)}, \ldots, \frac{a^{(m)[k]}-y^{(m)}}{a^{(m)[k]} \times \left(1-a^{(m)[k]}\right)}\right]$

5.1.2 Deriving $\frac{\partial \vec{a}^{[k]^T}}{\partial \vec{z}^{[k]}}$

$$\frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} = \begin{bmatrix} \frac{\partial \vec{a}^{[k]}}{\partial z^{(1)[k]}} \\ \frac{\partial \vec{a}^{[k]}}{\partial z^{(2)[k]}} \\ \vdots \\ \frac{\partial \vec{a}^{[k]}}{\partial z^{(m)[k]}} \end{bmatrix} = \begin{bmatrix} \frac{\partial a^{(1)[k]}}{\partial z^{(1)[k]}} & \frac{\partial a^{(2)[k]}}{\partial z^{(1)[k]}} & \cdots & \frac{\partial a^{(m)[k]}}{\partial z^{(1)[k]}} \\ \frac{\partial a^{(1)[k]}}{\partial z^{(2)[k]}} & \frac{\partial a^{(2)[k]}}{\partial z^{(2)[k]}} & \cdots & \frac{\partial a^{(m)[k]}}{\partial z^{(2)[k]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a^{(1)[k]}}{\partial z^{(m)[k]}} & \frac{\partial a^{(2)[k]}}{\partial z^{(m)[k]}} & \cdots & \frac{\partial a^{(m)[k]}}{\partial z^{(m)[k]}} \end{bmatrix}$$

For $i = 1,2, \dots, m$, and $j = 1,2, \dots, m$, we have th following expression:

$$\frac{\partial a^{(i)[k]}}{\partial z^{(j)[k]}} = \frac{\partial \left(\frac{1}{1 + e^{-z^{(i)[k]}}} \right)}{\partial z^{(j)[k]}} = \begin{cases} \frac{e^{-z^{(i)[k]}}}{\left(1 + e^{-z^{(i)[k]}}\right)^2} \, if \ i = j \\ 0 \, if \ i \neq j \end{cases}$$

For $i = j$, $\frac{e^{-z^{(i)[k]}}}{\left(1+e^{-z^{(i)[k]}}\right)^2} = \frac{1}{1+e^{-z^{(i)[k]}}} \times \frac{e^{-z^{(i)[k]}}}{1+e^{-z^{(i)[k]}}} = \left(\frac{1}{1+e^{-z^{(i)[k]}}}\right) \times \left(1 - \frac{1}{1+e^{-z^{(i)[k]}}}\right)$, and thus, we can express the expression in terms of $a^{(i)[k]}$:

$$\frac{\partial a^{(i)[k]}}{\partial z^{(j)[k]}} = \begin{cases} a^{(i)[k]} \times \left(1 - a^{(i)[k]}\right) \, if \ i = j \\ 0 \, if \ i \neq j \end{cases}$$

Thus, $\frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} = \begin{bmatrix} a^{(1)[k]} \times \left(1 - a^{(1)[k]}\right) & 0 & \dots & 0 \\ 0 & a^{(2)[k]} \times \left(1 - a^{(2)[k]}\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a^{(m)[k]} \times \left(1 - a^{(m)[k]}\right) \end{bmatrix}$

5.1.3 Deriving $\frac{\partial \vec{z}^{[k]^T}}{\partial w_j^{[k]}}$ and $\frac{\partial \vec{z}^{[k]^T}}{\partial b^{[k]}}$

$$\frac{\partial \vec{z}^{[k]^T}}{\partial w_j^{[k]}} = \begin{bmatrix} \frac{\partial z^{[k](1)}}{\partial w_j^{[k]}} \\ \frac{\partial z^{[k](2)}}{\partial w_j^{[k]}} \\ \vdots \\ \frac{\partial z^{[k](m)}}{\partial w_j^{[k]}} \end{bmatrix}, \frac{\partial \vec{z}^{[k]^T}}{\partial b^{[k]}} = \begin{bmatrix} \frac{\partial z^{[k](1)}}{\partial b^{[k]}} \\ \frac{\partial z^{[k](2)}}{\partial b^{[k]}} \\ \vdots \\ \frac{\partial z^{[k](m)}}{\partial b^{[k]}} \end{bmatrix}$$

For $i = 1,2, \dots, m$, and $j = 1,2, \dots, n_x$, we have the following expressions:

$$\frac{\partial z^{[k](i)}}{\partial w_j^{[k]}} = \frac{\partial\left(\vec{w}^{[k]^T}\vec{x}^{(i)} + b\right)}{\partial w_j^{[k]}} = \frac{\partial\left(\left(\sum_{j=1}^{n_x} w_j^{[k]} x_j^{(i)}\right) + b\right)}{\partial w_j^{[k]}} = x_j^{(i)}$$

$$\frac{\partial z^{[k](i)}}{\partial b^{[k]}} = \frac{\partial\left(\vec{w}^{[k]^T}\vec{x}^{(i)} + b\right)}{\partial b^{[k]}} = \frac{\partial\left(\left(\sum_{j=1}^{n_x} w_j^{[k]} x_j^{(i)}\right) + b\right)}{\partial b^{[k]}} = 1$$

Thus, we have the following results:

$$\frac{\partial \vec{z}^{[k]^T}}{\partial w_j^{[k]}} = \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}, \frac{\partial \vec{z}^{[k]^T}}{\partial b^{[k]}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbf{1}_{m\times 1}$$

5.1.4 Putting It All Together

Using the derivations above, we have the following results:

5.1.4.1 $\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}}$

$$\begin{aligned}
\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}} = \frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}} \times \frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} \times \frac{\partial \vec{z}^{[k]^T}}{\partial w_j^{[k]}} \\
&= \frac{1}{m}\left[\frac{a^{(1)[k]} - y^{(1)}}{a^{(1)[k]} \times (1 - a^{(1)[k]})}, \frac{a^{(2)[k]} - y^{(2)}}{a^{(2)[k]} \times (1 - a^{(2)[k]})}, \dots, \frac{a^{(m)[k]} - y^{(m)}}{a^{(m)[k]} \times (1 - a^{(m)[k]})}\right] \\
&\times \begin{bmatrix} a^{(1)[k]} \times \left(1 - a^{(1)[k]}\right) & 0 & \dots & 0 \\ 0 & a^{(2)[k]} \times \left(1 - a^{(2)[k]}\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a^{(m)[k]} \times \left(1 - a^{(m)[k]}\right) \end{bmatrix} \\
&\times \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}
\end{aligned}$$

$$= \frac{1}{m}\left[a^{(1)[k]} - y^{(1)}, a^{(2)[k]} - y^{(2)}, \dots, a^{(m)[k]} - y^{(m)}\right] \times \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}$$

$$= \frac{1}{m}\left(\left[a^{(1)[k]}, a^{(2)[k]}, \dots, a^{(m)[k]}\right] - \left[y^{(1)}, y^{(2)}, \dots, y^{(m)}\right]\right) \times \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}$$

$$= \frac{1}{m}\left(A^{[k]} - Y\right) \times \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}$$

$$= \frac{1}{m} \times \left[x_j^{(1)}, x_j^{(2)}, \dots, x_j^{(m)}\right] \times \left(A^{[k]} - Y\right)^T$$

In vector form, we have the following result:

$$\frac{\partial \mathcal{J}^{[k]}}{\partial \vec{w}^{[k]^T}} = \begin{bmatrix} \frac{\partial \mathcal{J}^{[k]}}{\partial w_1^{[k]}} \\ \frac{\partial \mathcal{J}^{[k]}}{\partial w_2^{[k]}} \\ \vdots \\ \frac{\partial \mathcal{J}^{[k]}}{\partial w_{n_x}^{[k]}} \end{bmatrix} = \begin{bmatrix} \frac{1}{m} \times \left[x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(m)}\right] \times \left(A^{[k]} - Y\right)^T \\ \frac{1}{m} \times \left[x_2^{(1)}, x_2^{(2)}, \dots, x_2^{(m)}\right] \times \left(A^{[k]} - Y\right)^T \\ \vdots \\ \frac{1}{m} \times \left[x_{n_x}^{(1)}, x_{n_x}^{(2)}, \dots, x_{n_x}^{(m)}\right] \times \left(A^{[k]} - Y\right)^T \end{bmatrix}$$

$$= \frac{1}{m} \times \begin{bmatrix} \left[x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(m)}\right] \\ \left[x_2^{(1)}, x_2^{(2)}, \dots, x_2^{(m)}\right] \\ \vdots \\ \left[x_{n_x}^{(1)}, x_{n_x}^{(2)}, \dots, x_{n_x}^{(m)}\right] \end{bmatrix} \times \left(A^{[k]} - Y\right)^T$$

$$= \frac{1}{m} \times \begin{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_{n_x}^{(1)} \end{bmatrix} & \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ \vdots \\ x_{n_x}^{(2)} \end{bmatrix} & \dots & \begin{bmatrix} x_1^{(m)} \\ x_2^{(m)} \\ \vdots \\ x_{n_x}^{(m)} \end{bmatrix} \end{bmatrix} \times \left(A^{[k]} - Y\right)^T$$

$$= \frac{1}{m} \times [\vec{x}^{(1)} \quad \vec{x}^{(2)} \quad \dots \quad \vec{x}^{(m)}] \times \left(A^{[k]} - Y\right)^T$$

$$= \frac{1}{m} \times X \times \left(A^{[k]} - Y\right)^T$$

5.1.4.2 $\frac{\partial \mathcal{J}^{[k]}}{\partial b^{[k]}}$

Leveraging previous derivations, we have the following result:

$$\frac{\partial \mathcal{J}^{[k]}}{\partial b^{[k]}} = \frac{\partial \mathcal{J}^{[k]}}{\partial \vec{a}^{[k]}} \times \frac{\partial \vec{a}^{[k]}}{\partial \vec{z}^{[k]^T}} \times \frac{\partial \vec{z}^{[k]^T}}{\partial b^{[k]}} = \frac{1}{m}\left[a^{(1)[k]} - y^{(1)}, a^{(2)[k]} - y^{(2)}, \dots, a^{(m)[k]} - y^{(m)}\right] \times \mathbf{1}_{m\times 1}$$
$$= \frac{1}{m}\sum_{i=1}^{m} a^{(i)[k]} - y^{(i)}$$

5.2 Checking Direction of Gradient Descent

We conduct a second derivative test on each coefficient:

$$\frac{\partial \mathcal{J}^{2[k]}}{\partial w_j^{[k]2}} = \frac{\partial\left(\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}}\right)}{\partial w_j^{[k]}} = \frac{\partial\left(\frac{1}{m}\left[a^{(1)[k]} - y^{(1)}, a^{(2)[k]} - y^{(2)}, \dots, a^{(m)[k]} - y^{(m)}\right] \times \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(m)} \end{bmatrix}\right)}{\partial w_j^{[k]}}$$
$$= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \times \frac{\partial a^{(i)[k]}}{\partial w_j^{[k]}}$$
$$= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \times \frac{\partial a^{(i)[k]}}{\partial z^{(i)[k]}} \times \frac{\partial z^{(i)[k]}}{\partial w_j^{[k]}}$$
$$= \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \times \left(a^{(i)[k]} \times \left(1 - a^{(i)[k]}\right)\right) \times x_j^{(i)}$$
$$= \frac{1}{m}\sum_{i=1}^{m} \left(x_j^{(i)}\right)^2 \times \left(a^{(i)[k]} \times \left(1 - a^{(i)[k]}\right)\right)$$

Note that since $a^{(i)[k]} = g\left(z^{(i)[k]}\right) = \frac{1}{1+e^{-z^{(i)[k]}}}$, therefore $a^{(i)[k]} \in (0,1)$; thus $1 - a^{(i)[k]} \in (0,1)$. Finally, since $m > 0$ and $\left(x_j^{(i)}\right)^2 \geq 0$, if we assume that there is at least one $x_j^{(i)} > 0$, we thus have $\frac{\partial \mathcal{J}^{2[k]}}{\partial w_j^{[k]2}} > 0$ ensuring that $\mathcal{J}^{[\mathrm{k}]}$ is strictly convex with respect $w_j^{[k]}$ for $j = 1,2, \dots, n_x$.

$$\frac{\partial \mathcal{J}^{2[k]}}{\partial b^{[k]2}} = \frac{\partial\left(\frac{\partial \mathcal{J}^{[k]}}{\partial w_j^{[k]}}\right)}{\partial b^{[k]}} = \frac{\partial\left(\frac{1}{m}\sum_{i=1}^{m} a^{(i)[k]} - y^{(i)}\right)}{\partial b^{[k]}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial a^{(i)[k]}}{\partial b^{[k]}}$$
$$= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial a^{(i)[k]}}{\partial z^{(i)[k]}} \times \frac{\partial z^{(i)[k]}}{\partial b^{[k]}}$$
$$= \frac{1}{m}\sum_{i=1}^{m} \left(a^{(i)[k]} \times \left(1 - a^{(i)[k]}\right)\right) \times 1 > 0$$

Thus, $\frac{\partial \mathcal{J}^{2^{[k]}}}{\partial b^{[k]^2}} > 0$, ensuring that $\mathcal{J}^{[\mathrm{k}]}$ is strictly convex with respect $b^{[k]}$ regardless of the values of $x_j^{(i)}$.

5.3 Final Derivations

- $\vec{w}$ (more efficient to express in vector form):
  - $\vec{w}^{[0]} = vector\ of\ random\ numbers\ (typically\ chosen\ to\ be\ close\ to\ zero)$
  - $\vec{w}^{[k+1]} = \vec{w}^{[k]} - \alpha \times \left(\frac{1}{m} \times X \times \left(A^{[k]} - Y\right)^T\right)$, where α is a set learning rate (e.g., 0.005), and $k = 0,1,\dots,N$, where N is the total number of iterations
- b:
  - $b^{[0]} = random\ number\ (typically\ chosen\ to\ be\ zero)$
  - $b^{[k+1]} = b^{[k]} - \alpha \times \left(\frac{1}{m}\sum_{i=1}^{m} a^{(i)[k]} - y^{(i)}\right)$, where α is a set learning rate (e.g., 0.005), and $k = 0,1,\dots,N$, where N is the total number of iterations