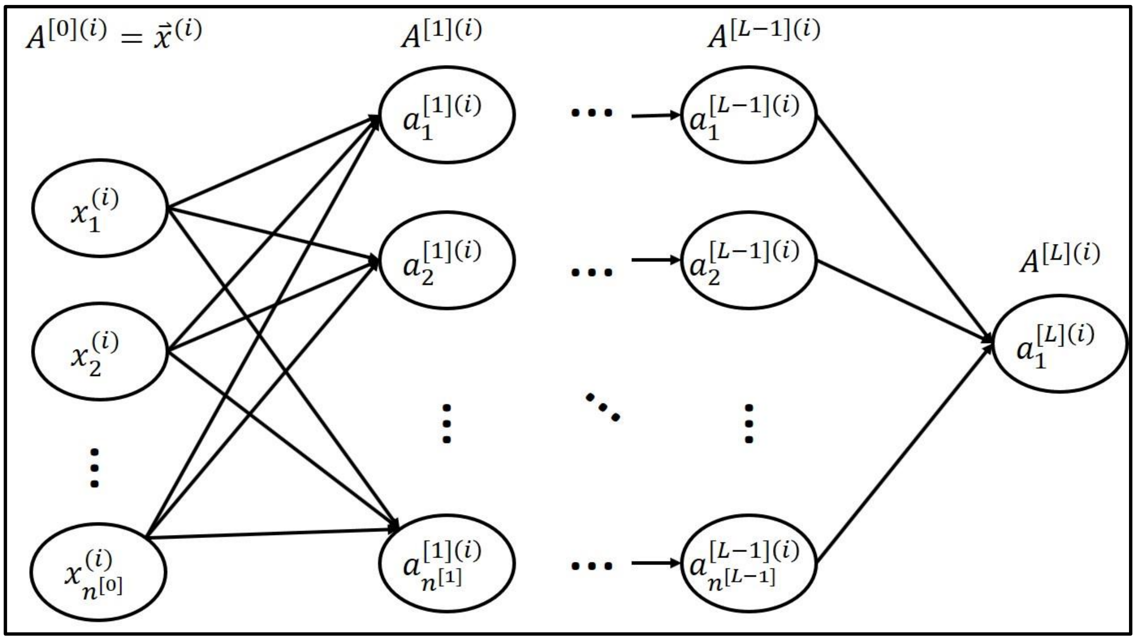
Gradient Descent Derivative Derivations for a single-output L-Layer Neural Network

1. Neural Network Description

The following diagram depicts the structure of a one output L-layer neural network for training sample i, where $i \in \{1,2,\ldots,m\}$:



Variables:

- $x_j^{(i)} \rightarrow$ jth input feature for training sample i, where $j \in \{1,2,\ldots,n^{[0]}\}$
- $\vec{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_{n^{[0]}}^{(i)} \end{bmatrix}$
- $X = [\vec{x}^{(1)} \quad \vec{x}^{(2)} \quad \ldots \quad \vec{x}^{(m)}]$
- $a_j^{[\ell](i)} \rightarrow$ jth activation in layer ℓ for training sample i, where $j \in \{1,2,\ldots,n^{[\ell]}\}$ and $\ell \in \{0,2,\ldots,L\}$, with $a_j^{[0](i)} = x_j^{(i)}$. In this model, we assume that $n^{[L]} = 1$.
- $\vec{a}^{[\ell](i)} = \begin{bmatrix} a_1^{[\ell](i)} \\ a_2^{[\ell](i)} \\ \vdots \\ a_{n^{[\ell]}}^{[\ell](i)} \end{bmatrix}$
- $A^{[\ell]} = [\vec{a}^{[\ell](1)} \quad \vec{a}^{[\ell](2)} \quad \ldots \quad \vec{a}^{[\ell](m)}]$

- $y^{(i)} \rightarrow$ True output for training sample i
- $\vec{y} = \left[y^{(1)}, y^{(2)}, \ldots, y^{(m)}\right]$
- $w_{j,k}^{[\ell]} \rightarrow$ Weight coefficient of activation $a_k^{[\ell-1](i)}$ for $i \in \{1,2,\ldots,m\}$, where $\ell \in \{1,2,\ldots,\ell\}$
- $\vec{w}_j^{[\ell]} = \begin{bmatrix} w_{j,1}^{[\ell]} \\ w_{j,2}^{[\ell]} \\ \vdots \\ w_{j,n^{[\ell-1]}}^{[\ell]} \end{bmatrix}$
- $W^{[\ell]} = \begin{bmatrix} \vec{w}_1^{[\ell]^T} \\ \vec{w}_2^{[\ell]^T} \\ \vdots \\ \vec{w}_{n^{[\ell]}}^{[\ell]^T} \end{bmatrix}$
- $b_j^{[\ell]} \rightarrow$ Bias for activation $a_j^{[\ell](i)}$ for $i \in \{1,2,\ldots,m\}$
- $\vec{b}^{[\ell]} = \begin{bmatrix} b_1^{[\ell]} \\ b_2^{[\ell]} \\ \vdots \\ b_{n^{[\ell]}}^{[\ell]} \end{bmatrix}$
- $z_j^{[\ell](i)} = \vec{w}_j^{[\ell]^T} \vec{a}^{[\ell-1](i)} + b_j^{[\ell]}$
- $\vec{z}^{[\ell](i)} = W^{[\ell]} \vec{a}^{[\ell-1](i)} + \vec{b}^{[\ell]}$
- $\mathbf{1}_{p \times q} \rightarrow$ Matrix of ones with p rows and q columns
- $Z^{[\ell]} = W^{[\ell]} A^{[\ell-1]} + \vec{b}^{[\ell]} \mathbf{1}_{1 \times m}$
- $g^{[\ell]}\left(z_j^{[\ell](i)}\right) = a_j^{[\ell](i)} \rightarrow$ Activation function for members in layer ℓ, where $\ell \in \{1,2,\ldots,L\}$
- $\vec{g}^{[\ell]}\left(\vec{z}^{[\ell](i)}\right) = \begin{bmatrix} g^{[\ell]}\left(z_1^{[\ell](i)}\right) \\ g^{[\ell]}\left(z_2^{[\ell](i)}\right) \\ \vdots \\ g^{[\ell]}\left(z_{n^{[\ell]}}^{[\ell](i)}\right) \end{bmatrix} = \vec{a}^{[\ell](i)}$
- $G^{[\ell]}\left(Z^{[\ell]}\right) = \left[\vec{g}^{[\ell]}\left(\vec{z}^{[\ell](i)}\right) \quad \vec{g}^{[\ell]}\left(\vec{z}^{[\ell](2)}\right) \quad \ldots \quad \vec{g}^{[\ell]}\left(\vec{z}^{[\ell](m)}\right)\right] = A^{[\ell]}$
- $\mathcal{L}\left(a_1^{[L](i)}, y^{(i)}\right) \rightarrow$ Loss function for training sample i
- $\mathcal{J}\left(A^{[L]}, \vec{y}\right) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(a_1^{[L](i)}, y^{(i)}\right)$

The goal is to find the values in $W^{[\ell]}$ and $\vec{b}^{[\ell]}$, $\ell \in \{1,2,\ldots,m\}$, that minimize the value of $\mathcal{J}\left(A^{[L]}, \vec{y}\right)$

2. Methodology

2.1 Assumptions

We assume that the functions $\mathcal{L}(.,.)$ and $g^{[\ell]}(.)$, for $\ell \in \{1,2,\dots,L\}$ are designed/chosen such that:

- $\frac{\partial^2 \mathcal{J}(A^{[L]},\vec{y})}{\partial {w_{j,k}^{[\ell]}}^2} > 0$ for $j \in \{1,2,\dots,n^{[\ell]}\}$, $k \in \{1,2,\dots,n^{[\ell-1]}\}$, and $\ell \in \{1,2,\dots,L\}$
- $\frac{\partial \mathcal{J}^2(A^{[L]},\vec{y})}{\partial {b_j^{[\ell]}}^2} > 0$ for $j \in \{1,2,\dots,n^{[\ell]}\}$ and $\ell \in \{1,2,\dots,L\}$

In other words, we assume that $\mathcal{L}(.,.)$ and $g^{[\ell]}(.)$ are at least twice differentiable and result in making $\mathcal{J}(.,.)$ strictly convex with respect to the parameters we wish to estimate.

2.2 Procedure High Level Description

Given the assumptions above, we aim to use gradient descent to recalibrate parameter values iteratively in a manner that they result in having the general loss function $\mathcal{J}(.,.)$ converge to a minimum value. For iteration $n \in \{0,1,\dots,N\}$ and a set learning rate α (this is a hyper-parameter which can be calibrating via a development/validation dataset) the procedure can be described as follows (we add the superscript $\{n\}$ to the parameter of interest to denote the iteration number it is being calculated at):

- Standardize input features (subtract mean and divide by standard deviation)
- For $n = 0; n \leq N; n++$:
  - For $\ell = 1$; $\ell \leq L$; $\ell++$://Forward propagation
    - If $n == 0$:
      - $W^{[\ell]\{0\}} = random\ normal(0,1)$
      - $\vec{b}^{[\ell]\{0\}} = zero\ vector\left(\left(n^{[\ell]},0\right)\right)$
    - Calculate activations and store key variables
  - Calculate $\frac{\partial \mathcal{J}(A^{[L]\{n\}},\vec{y})}{\partial A^{[L]\{n\}}}$
  - For $\ell = L$; $\ell \geq 1$; $\ell--$://Backward propagation
    - Calculate $\frac{\partial \mathcal{J}(A^{[L]\{n\}},\vec{y})}{\partial W^{[\ell]\{n\}}}$
    - Calculate $\frac{\partial \mathcal{J}(A^{[L]\{n\}},\vec{y})}{\partial \vec{b}^{[\ell]\{n\}}}$
    - $W^{[\ell]\{n+1\}} = W^{[\ell]\{n\}} - \alpha \frac{\partial \mathcal{J}(A^{[L]\{n\}},\vec{y})}{\partial W^{[\ell]\{n\}}}$
    - $\vec{b}^{[\ell]\{n+1\}} = \vec{b}^{[\ell]\{n\}} - \alpha \frac{\partial \mathcal{J}(A^{[L]\{n\}},\vec{y})}{\partial \vec{b}^{[\ell]\{n\}}}$

2.3 Derivations

If we wish to minimize $\mathcal{J}(.,.)$, then we wish to minimize $\mathcal{L}(.,.)$ for every training sample $i \in \{1,2,\dots,m\}$. Thus, our focus for now will be on finding

$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{j,k}^{[\ell]}}$ and $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b_j^{[\ell]}}$ where $j \in \left\{1,2,\dots,n^{[L]}\right\}$, $k \in \left\{1,2,\dots,n^{[L-1]}\right\}$, and $\ell \in \{1,2,\dots,L\}$ (note: we intentionally dropped the superscripts (i) and $\{n-1\}$ for notational convenience; they will be reinserted once we vectorize the results).

2.3.1 Scalar Derivations

The derivations are made based on breaking up the derivatives into smaller pieces via the chain rule and then finding recursive relationships between them. Let $g^{[\ell]\prime}\left(z_j^{[\ell]}\right) = \frac{\partial g\left(z_j^{[\ell]}\right)}{\partial z_j^{[\ell]}} = \frac{\partial a_j^{[\ell]}}{\partial z_j^{[\ell]}}$, we have the following expressions:

$$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{j,k}^{[\ell]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}} \times \frac{\partial a_j^{[\ell]}}{\partial z_j^{[\ell]}} \times \frac{\partial z_j^{[\ell]}}{\partial w_{j,k}^{[\ell]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}} \times g^{[\ell]\prime}\left(z_j^{[\ell]}\right) \times a_k^{[\ell-1]}$$

and

$$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b_j^{[\ell]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}} \times \frac{\partial a_j^{[\ell]}}{\partial z_j^{[\ell]}} \times \frac{\partial z_j^{[\ell]}}{\partial b_j^{[\ell]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}} \times g^{[\ell]\prime}\left(z_j^{[\ell]}\right) \times 1$$

For $\ell = L$, we compute $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[L]}}$ directly

For $\ell < L$, we can recursively calculate $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}}$ as follows:

$$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_j^{[\ell]}} = \sum_{p=1}^{n^{[\ell]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_j^{[\ell+1]}} \times \frac{\partial z_j^{[\ell+1]}}{\partial a_j^{[\ell]}} = \sum_{p=1}^{n^{[\ell]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} \times \frac{\partial z_p^{[\ell+1]}}{\partial a_j^{[\ell]}} = \sum_{p=1}^{n^{[\ell+1]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} \times w_{p,j}^{[\ell+1]},$$

where $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_p^{[\ell+1]}} \times \frac{\partial a_p^{[\ell+1]}}{\partial z_p^{[\ell+1]}} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_p^{[\ell+1]}} \times g^{[\ell+1]\prime}\left(z_p^{[\ell+1]}\right)$

2.3.2 Matrix Derivations of a Single Training Sample

Here, we combine all units in a layer.

We write the derivatives as $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial W^{[L]}}$ and $\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b^{[\ell]}}$ in the following forms and express the results in the subsequent formulas:

- $$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial W^{[\ell]}} = \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{1,1}^{[\ell]}} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{1,2}^{[\ell]}} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{1,n^{[\ell-1]}}^{[\ell]}} \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{2,1}^{[\ell]}} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{2,2}^{[\ell]}} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{2,,n^{[\ell-1]}}^{[\ell]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{n^{[\ell]},1}^{[\ell]}} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{j,k}^{[\ell]}} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial w_{n^{[\ell]},n^{[\ell-1]}}^{[\ell]}} \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_1^{[\ell]}} \times g^{[\ell]\prime}\left(z_1^{[\ell]}\right) \times a_1^{[\ell-1]} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_1^{[\ell]}} \times g^{[\ell]\prime}\left(z_1^{[\ell]}\right) \times a_2^{[\ell-1]} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_1^{[\ell]}} \times g^{[\ell]\prime}\left(z_1^{[\ell]}\right) \times a_{n^{[\ell-1]}}^{[\ell-1]} \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_2^{[\ell]}} \times g^{[\ell]\prime}\left(z_2^{[\ell]}\right) \times a_1^{[\ell-1]} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_2^{[\ell]}} \times g^{[\ell]\prime}\left(z_2^{[\ell]}\right) \times a_2^{[\ell-1]} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_2^{[\ell]}} \times g^{[\ell]\prime}\left(z_2^{[\ell]}\right) \times a_{n^{[\ell-1]}}^{[\ell-1]} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_{n^{[\ell]}}^{[\ell]}} \times g^{[\ell]\prime}\left(z_{n^{[\ell]}}^{[\ell]}\right) \times a_1^{[\ell-1]} & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_{n^{[\ell]}}^{[\ell]}} \times g^{[\ell]\prime}\left(z_{n^{[\ell]}}^{[\ell]}\right) \times a_2^{[\ell-1]} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_{n^{[\ell]}}^{[\ell]}} \times g^{[\ell]\prime}\left(z_{n^{[\ell]}}^{[\ell]}\right) \times a_{n^{[\ell-1]}}^{[\ell-1]} \end{bmatrix}$$

$$= \left(\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial \vec{a}^{[\ell]}} \circ \vec{g}^{[\ell]'}\left(\vec{z}^{[\ell]}\right) \right) \times \vec{a}^{[\ell-1]^T}$$

- $$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b^{[\ell]}} = \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b_1^{[\ell]}} \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b_2^{[\ell]}} \\ \vdots \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial b_{n^{[\ell]}}^{[\ell]}} \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_1^{[\ell]}} \times g^{[\ell]'}\left(z_1^{[\ell]}\right) \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_2^{[\ell]}} \times g^{[\ell]'}\left(z_2^{[\ell]}\right) \\ \vdots \\ \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial a_{n^{[\ell]}}^{[\ell]}} \times g^{[\ell]'}\left(z_{n^{[\ell]}}^{[\ell]}\right) \end{bmatrix} = \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial \vec{a}^{[\ell]}} \circ \vec{g}^{[\ell]'}\left(\vec{z}^{[\ell]}\right)$$

$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial \vec{a}^{[\ell]}}$ can be derived as follows:

$$\frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial \vec{a}^{[\ell]}} = \begin{bmatrix} \sum_{p=1}^{n^{[\ell+1]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} \times w_{p,1}^{[\ell+1]} \\ \sum_{p=1}^{n^{[\ell+1]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} \times w_{p,2}^{[\ell+1]} \\ \vdots \\ \sum_{p=1}^{n^{[\ell+1]}} \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial z_p^{[\ell+1]}} \times w_{p,n^{[\ell]}}^{[\ell+1]} \end{bmatrix} = W^{[\ell+1]^T} \times \frac{\partial \mathcal{L}\left(a_1^{[L]},y\right)}{\partial \vec{z}^{[\ell+1]}}, \text{ where}$$

$$\frac{\partial \mathcal{L}\left(a_1^{[L]}, y\right)}{\partial \vec{z}^{[\ell+1]}} = \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L]}, y\right)}{\partial a_1^{[\ell+1]}} \times g^{[\ell+1]'}\left(z_1^{[\ell+1]}\right) \\ \frac{\partial \mathcal{L}\left(a_1^{[L]}, y\right)}{\partial a_2^{[\ell+1]}} \times g^{[\ell+1]'}\left(z_2^{[\ell+1]}\right) \\ \vdots \\ \frac{\partial \mathcal{L}\left(a_1^{[L]}, y\right)}{\partial a_{n^{[\ell+1]}}^{[\ell+1]}} \times g^{[\ell+1]'}\left(z_{n^{[\ell+1]}}^{[\ell+1]}\right) \end{bmatrix} = \frac{\partial \mathcal{L}\left(a_1^{[L]}, y\right)}{\partial \vec{a}^{[\ell+1]}} \circ \vec{g}^{[\ell+1]'}\left(\vec{z}^{[\ell+1]}\right)$$

2.3.3 Matrix Derivations of Entire Sample

Here, we derive the final form of the formulas:

$$\frac{\partial \mathcal{J}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial W^{[\ell]\{n-1\}}}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \frac{\partial \mathcal{L}\left(a_1^{[L](i)\{n-1\}}, y^{(i)}\right)}{\partial W^{[\ell]}}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial \mathcal{L}\left(a_1^{[L](i)\{n-1\}}, y\right)}{\partial \vec{a}^{[\ell](i)\{n-1\}}} \circ \vec{g}^{[\ell]'}\left(\vec{z}^{[\ell](i)\{n-1\}}\right)\right) \times \vec{a}^{[\ell-1](i)\{n-1\}^T}$$

$$= \frac{1}{m} \times \left(\frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial A^{[\ell]\{n-1\}}} \circ G^{[\ell]'}\left(Z^{[\ell]\{n-1\}}\right)\right) \times A^{[\ell-1]\{n-1\}^T}$$

$$\frac{\partial \mathcal{J}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial \vec{b}^{[\ell]\{n-1\}}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial \mathcal{L}\left(a_1^{[L](i)\{n-1\}}, y^{(i)}\right)}{\partial \vec{b}^{[\ell]}} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\partial \mathcal{L}\left(a_1^{[L](i)\{n-1\}}, y\right)}{\partial \vec{a}^{[\ell](i)\{n-1\}}} \circ \vec{g}^{[\ell]'}\left(\vec{z}^{[\ell](i)\{n-1\}}\right)\right)$$

$$= \frac{1}{m} \times \left(\frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial A^{[\ell]\{n-1\}}} \circ G^{[\ell]'}\left(Z^{[\ell]\{n-1\}}\right)\right) \times \mathbf{1}_{m\times 1}$$

The expression $\frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial A^{[\ell]\{n-1\}}}$ can be calculated as follows:

$$\frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial A^{[\ell]\{n-1\}}} = \begin{bmatrix} \frac{\partial \mathcal{L}\left(a_1^{[L](1)\{n-1\}}, y^{(1)}\right)}{\partial \vec{a}^{[\ell](1)\{n-1\}}} & \frac{\partial \mathcal{L}\left(a_1^{[L](2)\{n-1\}}, y^{(2)}\right)}{\partial \vec{a}^{[\ell](2)\{n-1\}}} & \cdots & \frac{\partial \mathcal{L}\left(a_1^{[L](m)\{n-1\}}, y^{(m)}\right)}{\partial \vec{a}^{[\ell](m)\{n-1\}}} \end{bmatrix}$$

$$= \left[W^{[\ell+1]\{n-1\}^T} \times \frac{\partial \mathcal{L}\left(a_1^{[L](1)\{n-1\}}, y^{(1)}\right)}{\partial \vec{z}^{(1)[\ell+1]\{n-1\}}} \quad W^{[\ell+1]\{n-1\}^T} \times \frac{\partial \mathcal{L}\left(a_1^{[L](2)\{n-1\}}, y^{(1)}\right)}{\partial \vec{z}^{(2)[\ell+1]\{n-1\}}} \quad \dots \quad W^{[\ell+1]\{n-1\}^T} \times \frac{\partial \mathcal{L}\left(a_1^{[L](m)\{n-1\}}, y^{(1)}\right)}{\partial \vec{z}^{(m)[\ell+1]\{n-1\}}} \right]$$

$$= W^{[\ell+1]\{n-1\}^T} \times \frac{\partial \mathcal{L}(A^{[L]\{n-1\}}, \vec{y})}{\partial Z^{[\ell+1]\{n-1\}}}, \text{ where}$$

$$\frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial Z^{[\ell+1]}}$$

$$= \left[\frac{\partial \mathcal{L}\left(a_1^{[L](1)\{n-1\}}, y^{(1)}\right)}{\partial \vec{a}^{[\ell+1](1)\{n-1\}}} \circ \vec{g}^{[\ell+1]'}\left(\vec{z}^{[\ell+1](1)\{n-1\}}\right) \quad \frac{\partial \mathcal{L}\left(a_1^{[L](1)\{n-1\}}, y^{(2)}\right)}{\partial \vec{a}^{[\ell+1](2)\{n-1\}}} \circ \vec{g}^{[\ell+1]'}\left(\vec{z}^{[\ell+1](2)\{n-1\}}\right) \quad \dots \quad \frac{\partial \mathcal{L}\left(a_1^{[L](1)\{n-1\}}, y^{(m)}\right)}{\partial \vec{a}^{[\ell+1](m)\{n-1\}}} \circ \vec{g}^{[\ell+1]'}\left(\vec{z}^{[\ell+1](m)\{n-1\}}\right) \right]$$

$$= \frac{\partial \mathcal{L}\left(A^{[L]\{n-1\}}, \vec{y}\right)}{\partial A^{[\ell+1]\{n-1\}}} \circ G^{[\ell+1]'}\left(Z^{[\ell+1]\{n-1\}}\right)$$

2.4 Procedure Detailed Description

Using the structure and derivations above, the parameters can be calculated as follows. For iteration $n \in \{0,1,\dots,N\}$ and a set learning rate α, we have the following steps

- Standardize input features (subtract mean and divide by standard deviation)

- For $n = 0; n \leq N; n++$:
  - For $\ell = 1$; $\ell \leq L$; $\ell++$://Forward propagation
    - If $n == 0$:
      - $W^{[\ell]\{0\}} = random\ normal(0,1)$
      - $\vec{b}^{[\ell]\{0\}} = zero\ vector\left(\left(n^{[\ell]}, 0\right)\right)$
    - Calculate $A^{[\ell]\{n\}}$, $Z^{[\ell]\{n\}}$, $G^{[\ell]}\left(Z^{[\ell]\{n\}}\right)$, $G^{[\ell]'}\left(Z^{[\ell]\{n\}}\right)$for $\ell \in \{1,2,\dots,L\}$ (note: we assume that all activation functions $G^{[\ell]}(.)$ have calculable derivative values) and store $G^{[\ell]'}\left(Z^{[\ell]\{n\}}\right)$ to be used in the backward propagation step
  - Calculate $\frac{\partial \mathcal{L}\left(A^{[L]\{n\}}, \vec{y}\right)}{\partial A^{[L]\{n\}}}$ (note: we assume that the loss function $\mathcal{L}(.,.)$ has a calculable derivative value)

  - For $\ell = L$; $\ell \geq 1$; $\ell--$://Backward propagation
    - $\frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial Z^{[\ell]\{n\}}} = \frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial A^{[\ell]\{n\}}} \circ G^{[\ell]'}\left(Z^{[\ell]\{n\}}\right)$
    - $\frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial A^{[\ell-1]\{n\}}} = W^{[\ell]\{n\}^T} \times \frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial Z^{[\ell]\{n\}}}$
    - $\frac{\partial \mathcal{J}(A^{[L]\{n\}}, \vec{y})}{\partial W^{[\ell]\{n\}}} = \frac{1}{m} \times \frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial Z^{[\ell]\{n\}}} \times A^{[\ell-1]\{n\}^T}$
    - $\frac{\partial \mathcal{J}(A^{[L]\{n\}}, \vec{y})}{\partial \vec{b}^{[\ell]\{n\}}} = \frac{1}{m} \times \frac{\partial \mathcal{L}(A^{[L]\{n\}}, \vec{y})}{\partial Z^{[\ell]\{n\}}} \times \mathbf{1}_{m \times 1}$
    - $W^{[\ell]\{n+1\}} = W^{[\ell]\{n\}} - \alpha \frac{\partial \mathcal{J}(A^{[L]\{n\}}, \vec{y})}{\partial W^{[\ell]\{n\}}}$
    - $\vec{b}^{[\ell]\{n+1\}} = \vec{b}^{[\ell]\{n\}} - \alpha \frac{\partial \mathcal{J}(A^{[L]\{n\}}, \vec{y})}{\partial \vec{b}^{[\ell]\{n\}}}$