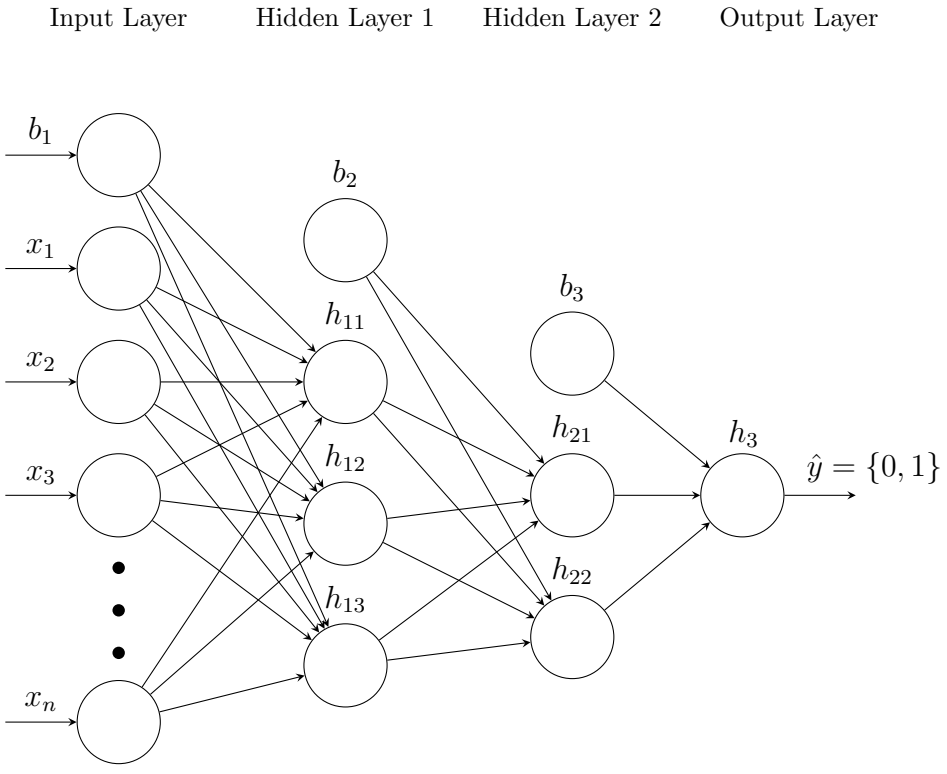
Two Layer Neural Network Derivation

Background

Consider the two-class classification problem using a single image as input. Since an image can be represented as a 3-dimensional matrix (or third-order tensor), we can consider the input as a flattened image vector with size $n = a \times b \times c$, where a and b correspond to the dimensions of the rows and columns of the matrix, respectively, and $c = 3$ corresponds to the rgb channel.

For this derivation, suppose we have a two layer neural network. Specifically, there are 3 neurons in the first hidden layer, 2 neurons in the second hidden layer, 1 neuron in the output layer, and a bias neuron in every layer except the output layer. An illustration of the model's architecture is shown below.



Forward Propagation

First, let's consider the equations of the neural network model from a *forward propagation* approach. This allows us to think about the equations moving forward through the network from input to output.

Consider a flattened image vector which can be expressed as $x \in \mathbb{R}^{n\times 1}$. The data points x_i are individually fed through the neural network and each data point x_i is mapped to all possible neurons in the first hidden layer. For each of the three neurons, the vector x initially encounters the linear part of the first hidden layer which can be expressed as the following set of functions:

$$\underbrace{z_1^{[1]}}_{1\times 1} = \underbrace{w_1^{[1]}}_{1\times n}\underbrace{x}_{n\times 1} + \underbrace{b_1^{[1]}}_{1\times 1} \tag{1}$$

$$z_2^{[1]} = w_2^{[1]}x + b_2^{[1]} \tag{2}$$

$$z_3^{[1]} = w_3^{[1]}x + b_3^{[1]} \tag{3}$$

where the superscript identifies the layer and the subscript identifies the neuron in the given layer. Note that the bias term $b^{[1]}$ is located in the input layer. To avoid any ambiguity, the previous illustration should make it clear that $b^{[1]}$ is not located in the first hidden layer.

We can express these three linear functions in the following simplified form:

$$\underbrace{z^{[1]}}_{3\times 1} = \underbrace{w^{[1]}}_{3\times n}\underbrace{x}_{n\times 1} + \underbrace{b^{[1]}}_{3\times 1} \tag{4}$$

Next, the linear functions $z_k^{[1]}$, where $k = \{1, 2, 3\}$, from each of the three neurons is mapped through independent activation functions. For this derivation, we will use the sigmoid function for all activation functions. Other common activation functions are the hyperbolic tangent function and the Rectified Linear Unit (ReLU) function. The sigmoid activation functions can expressed as follows:

$$\underbrace{a_1^{[1]}}_{1\times 1} = \underbrace{\sigma\left(z_1^{[1]}\right)}_{1\times 1}$$

$$a_2^{[1]} = \sigma\left(z_2^{[1]}\right)$$

$$a_3^{[1]} = \sigma\left(z_3^{[1]}\right)$$

where σ is the sigmoid function which is a scalar map $\sigma : \mathbb{R} \to \mathbb{R}$ expressed as:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \tag{5}$$

In a similar fashion, we can express these three activation functions in the following simplified form:

$$a^{[1]} = \sigma\left(z^{[1]}\right) \tag{6}$$

where $a^{[1]} \in \mathbb{R}^{3\times 1}$.

Next, the data encounter the second hidden layer of the network which consists of a second set of linear functions. This set of functions can be expressed as follows:

$$\underbrace{z_1^{[2]}}_{1\times 1} = \underbrace{w_1^{[2]}}_{1\times 3}\underbrace{a^{[1]}}_{3\times 1} + \underbrace{b_1^{[2]}}_{1\times 1} \tag{7}$$

$$z_2^{[2]} = w_2^{[2]}a^{[1]} + b_2^{[2]} \tag{8}$$

We can express these two functions in the following simplified form:

$$\underbrace{z^{[2]}}_{2\times 1} = \underbrace{w^{[2]}}_{2\times 3}\underbrace{a^{[1]}}_{3\times 1} + \underbrace{b^{[2]}}_{2\times 1} \tag{9}$$

Then we encounter the second set of sigmoid activation functions on $z^{[2]}$ which can be expressed as follows:

$$\underbrace{a_1^{[2]}}_{1\times 1} = \underbrace{\sigma\left(z_1^{[2]}\right)}_{1\times 1}$$

$$a_2^{[2]} = \sigma\left(z_2^{[2]}\right)$$

In a similar fashion, we can express these two functions in the following simplified form:

$$a^{[2]} = \sigma\left(z^{[2]}\right) \tag{10}$$

where $a^{[2]} \in \mathbb{R}^{2\times 1}$.

Then the data encounter the last linear part of the model in the output layer with the following function:

$$\underbrace{z^{[3]}}_{1\times1} = \underbrace{w^{[3]}}_{1\times2}\underbrace{a^{[2]}}_{2\times1} + \underbrace{b^{[3]}}_{1\times1} \tag{11}$$

Lastly, we encounter the final activation function on $z^{[3]}$ expressed as follows:

$$a^{[3]} = \sigma\left(z^{[3]}\right) \tag{12}$$

where $a^{[3]} \in \mathbb{R}^{1\times1}$.

Backpropagation

With the previous equations identified through forward propagation, we can now consider moving backwards through the neural network from the loss function to the input using a technique known as *backpropagation* (BP).

In this derivation, we can use the cross-entropy loss function which is suitable for binary classification problems. If regression was desired, we may consider using the squared error loss instead.

In order to define an objective function, let's first consider the log-likelihood loss function for the single data point under consideration:

$$\mathcal{L} = ylog\left(\hat{y}\right) + (1-y)log\left(1-\hat{y}\right) \tag{13}$$

To define an objective (or cost) function, consider obtaining more data points and summing over all the individual losses. If we normalize the summed loss for n observations, the objective can be defined as:

$$\mathcal{J}^*(\hat{y}, y) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)}$$

To optimize the parameters in the neural network, we want to *minimize* the *negative* loss function such that we obtain the cross entropy loss which we can then use in our gradient descent algorithm. In other words, our objective function we wish to minimize becomes the following:

$$\mathcal{J}(\hat{y}, y) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}^{(i)} \tag{14}$$

Therefore, to perform gradient descent we need to derive the following equations $\forall \ell = 1, 2, 3$:

$$w^{[\ell]}_{(new)} = w^{[\ell]} - \eta \frac{\partial \mathcal{L}}{\partial w^{[\ell]}} \tag{15}$$

$$b^{[\ell]}_{(new)} = b^{[\ell]} - \eta \frac{\partial \mathcal{L}}{\partial b^{[\ell]}} \tag{16}$$

Note that we can use the previously identified forward propagation equations to identify the correct path to take when performing the chain rule calculations when solving the derivatives of interest.

First, let's evaluate the derivatives of the loss with respect to $w^{[3]}$ and $b^{[3]}$.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w^{[3]}} &= -\frac{\partial}{\partial w^{[3]}} \left[y^{(i)} log\left(\hat{y}^{(i)}\right) + (1 - y^{(i)}) log\left(1 - \hat{y}^{(i)}\right) \right] \\
&= -\left[y^{(i)} \frac{\partial}{\partial w^{[3]}} \left(log\left(\sigma(w^{[3]} a^{[2]} + b^{[3]})\right) + (1 - y^{(i)}) \frac{\partial}{\partial w^{[3]}} log\left(1 - \sigma(w^{[3]} a^{[2]} + b^{[3]})\right) \right] \right] \\
&= -[y^{(i)} \frac{1}{\sigma(w^{[3]} a^{[2]} + b^{[3]})} \frac{\partial}{\partial w^{[3]}} \left(\sigma(w^{[3]} a^{[2]} + b^{[3]})\right) \frac{\partial}{\partial w^{[3]}} (w^{[3]} a^{[2]} + b^{[3]}) \\
&\quad + (1 - y^{(i)}) \frac{1}{1 - \sigma(w^{[3]} a^{[2]} + b^{[3]})} \frac{\partial}{\partial w^{[3]}} (1 - \sigma(w^{[3]} a^{[2]} + b^{[3]}) \frac{\partial}{\partial w^{[3]}} (\sigma(w^{[3]} a^{[2]} + b^{[3]}))] \\
&= -[y^{(i)} \frac{1}{a^{[3]}} a^{[3]} (1 - a^{[3]}) a^{[2]T} + (1 - y^{(i)}) \frac{1}{1 - a^{[t]}} (-1) a^{[3]} (1 - a^{[3]}) a^{[2]T}] \\
&= -[y^{(i)} (1 - a^{(3)}) a^{[2]T} - (1 - y^{(i)}) a^{[3]} a^{[2]T}] \\
&= -[y^{(i)} a^{[2]T} - y^{(i)} a^{[3]} a^{[2]T} - a^{[3]} a^{[2]T} + y^{(i)} a^{[3]} a^{[2]T}] \\
&= -[y^{(i)} a^{[2]T} - a^{[3]} a^{[2]T}] \\
&= (a^{[3]} - y^{(i)}) \cdot \frac{\partial z^{[3]}}{\partial w^{[3]}} \\
&= \underbrace{\underbrace{(a^{[3]} - y^{(i)})}_{1 \times 1} \underbrace{a^{[2]T}}_{1 \times 2}}_{1 \times 2}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b^{[3]}} &= -\frac{\partial}{\partial b^{[3]}}\left[y^{(i)} log\left(\hat{y}^{(i)}\right) + (1-y^{(i)}) log\left(1-\hat{y}^{(i)}\right)\right] \\
&= -\left[y^{(i)}\frac{\partial}{\partial b^{[3]}}\left(log\left(\sigma(w^{[3]}a^{[2]}+b^{[3]}\right)\right) + (1-y^{(i)})\frac{\partial}{\partial b^{[3]}} log\left(1-\sigma(w^{[3]}a^{[2]}+b^{[3]})\right)\right] \\
&= -[y^{(i)}\frac{1}{\sigma(w^{[3]}a^{[2]}+b^{[3]})}\frac{\partial}{\partial b^{[3]}}\left(\sigma(w^{[3]}a^{[2]}+b^{[3]})\right)\frac{\partial}{\partial b^{[3]}}(w^{[3]}a^{[2]}+b^{[3]}) \\
&\quad + (1-y^{(i)})\frac{1}{1-\sigma(w^{[3]}a^{[2]}+b^{[3]})}\frac{\partial}{\partial b^{[3]}}(1-\sigma(w^{[3]}a^{[2]}+b^{[3]})\frac{\partial}{\partial b^{[3]}}(\sigma(w^{[3]}a^{[2]}+b^{[3]}))] \\
&= -[y^{(i)}\frac{1}{a^{[3]}}a^{[3]}(1-a^{[3]})(1) + (1-y^{(i)})\frac{1}{1-a^{[3]}}(-1)a^{[3]}(1-a^{[3]})(1)] \\
&= -[y^{(i)}(1-a^{[3]}) - (1-y^{(i)})a^{[3]}] \\
&= -[y^{(i)} - y^{(i)}a^{[3]} - a^{[3]} + y^{(i)}a^{[3]}] \\
&= \underbrace{a^{[3]} - y^{(i)}}_{1\times 1}
\end{aligned}
$$

Before solving the next derivative, let's consider the following. If we want to compute the gradient of a function (e.g., the sigmoid function) $\phi : \mathbb{R}^n \to \mathbb{R}^n$ that takes the form of an element-wise operator such that

$$[\phi(\mathbf{x})]_i = f(\mathbf{x}_i), \forall i \in \{1, 2, ..., n\}$$

it can be helpful to compute the derivative of the scalar component function, f'. This can then be used in combination with differentials to compute the gradient of ϕ as follows:

$$d\phi(\mathbf{x}) = diag(f'[\mathbf{x}])d\mathbf{x}$$

where the notation $f'[\mathbf{x}]$ is used to denote that f' is applied element-wise to each element of $\mathbf{x}$ and the dimensions are $d\phi(\mathbf{x}) \in \mathbb{R}^{n\times 1}, diag(f'[\mathbf{x}]) \in \mathbb{R}^{n\times n}$, and $d\mathbf{x} \in \mathbb{R}^{n\times 1}$.

Note that for $i = \{1, 2\}$ and $j = \{1, 2, 3\}$, we have $\frac{\partial \mathcal{L}}{\partial w_{ij}^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_{11}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{12}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{13}^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial w_{21}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{22}^{[2]}} & \frac{\partial \mathcal{L}}{\partial w_{23}^{[2]}} \end{bmatrix}$.

Next, let's evaluate the derivatives of the loss with respect to $w^{[2]}$ and $b^{[2]}$.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w^{[2]}} &= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial w_{ij}^{[2]}} \\
&= (a^{[3]} - y^{(i)}) w^{[3]} \frac{\partial \sigma(z^{[2]})}{\partial z^{[2]}} \frac{\partial (w^{[2]} a^{[1]} + b^{[2]})}{\partial w_{ij}^{[2]}} \\
&= (a^{[3]} - y^{(i)}) w^{[3]} \frac{\partial \sigma(z^{[2]})}{\partial z^{[2]}} \frac{\partial (w^{[2]} a^{[1]})}{\partial w_{ij}^{[2]}} \\
&= (a^{[3]} - y^{(i)}) w^{[3]} d\sigma(\mathbf{z}^{[2]}) a_j^{[1]} \mathbf{e}_i \\
&= [\underbrace{(a^{[3]} - y^{(i)})}_{1\times 1} \underbrace{w^{[3]}}_{1\times 2} \underbrace{diag(\sigma'[\mathbf{z}^{[2]}]) d\mathbf{z}^{[2]}}_{2\times 2}] \underbrace{a_j^{[1]} \mathbf{e}_i}_{2\times 1} \\
&= \underbrace{[(a^{[3]} - y^{(i)}) w^{[3]} \circ (a^{[3]}(1 - a^{[3]}))]}_{1\times 2} \underbrace{a_j^{[1]} \mathbf{e}_i}_{2\times 1} \\
&= \underbrace{[(a^{[3]} - y^{(i)}) w^{[3]} \circ (a^{[3]}(1 - a^{[3]}))]_i a_j^{[1]}}_{1\times 1} \\
&= \underbrace{\left[(a^{[3]} - y^{(i)}) w^{[3]} \circ (a^{[3]}(1 - a^{[3]}))\right] a^{[1]T}}_{2\times 3}
\end{aligned}$$

where $a^{[1]} \in \mathbb{R}^{3\times 1}$, $\mathbf{e}_i \in \mathbb{R}^{2\times 1}$ is the i^{th} basis vector, and $\circ$ denotes the Hadamard product.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b^{[2]}} &= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}} \\
&= (a^{[3]} - y^{(i)}) w^{[3]} \frac{\partial \sigma(z^{[2]})}{\partial z^{[2]}} \cdot \frac{\partial (w^{[2]} a^{[1]} + b^{[2]})}{\partial b^{[2]}} \\
&= (a^{[3]} - y^{(i)}) w^{[3]} d\sigma(\mathbf{z}^{[2]}) \cdot \frac{\partial (b^{[2]})}{\partial b^{[2]}} \\
&= [(a^{[3]} - y^{(i)}) w^{[3]} d\sigma(\mathbf{z}^{[2]})] \mathbf{e}_i \\
&= [(a^{[3]} - y^{(i)}) w^{[3]} diag(\sigma'[\mathbf{z}^{[2]}]) d\mathbf{z}^{[2]}] \mathbf{e}_i \\
&= \underbrace{[(a^{[3]} - y^{(i)}) w^{[3]} \circ \sigma(z^{[2]})(1 - \sigma(z^{[2]}))]}_{1\times 2} \underbrace{\mathbf{e}_i}_{2\times 1} \\
&= \underbrace{[(a^{[3]} - y^{(i)}) w^{[3]} \circ \sigma(z^{[2]})(1 - \sigma(z^{[2]}))]_i}_{1\times 1}
\end{aligned}$$

$$= \underbrace{[(a^{[3]} - y^{(i)})w^{[3]} \circ (a^{[3]}(1 - a^{[3]}))]}_{2 \times 1}$$

where $b^{[2]} \in \mathbb{R}^{2\times 1}$ and $\mathbf{e}_i \in \mathbb{R}^{2\times 1}$ is the i^{th} basis vector. Finally, we can evaluate the derivatives of the loss with respect to $w^{[1]}$ and $b^{[1]}$.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w^{[1]}} &= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}} \cdot \frac{\partial a^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial w_{ij}^{[1]}} \\
&= (a^{[3]} - y^{(i)})w^{[3]} d\sigma(\mathbf{z}^{[2]}) w^{[2]} d\sigma(\mathbf{z}^{[1]}) \frac{\partial (w^{[1]}x + b^{[1]})}{\partial w_{ij}^{[1]}} \\
&= (a^{[3]} - y^{(i)})w^{[3]} d\sigma(\mathbf{z}^{[2]}) w^{[2]} d\sigma(\mathbf{z}^{[1]}) \frac{\partial (w^{[1]}x)}{\partial w_{ij}^{[1]}} \\
&= \underbrace{(a^{[3]} - y^{(i)})}_{1\times 1} \underbrace{w^{[3]}}_{1\times 2} \underbrace{diag(\sigma'[\mathbf{z}^{[2]}])d\mathbf{z}^{[2]}}_{2\times 2} \underbrace{w^{[2]}}_{2\times 3} \underbrace{diag(\sigma'[\mathbf{z}^{[1]}])d\mathbf{z}^{[1]}}_{3\times 3} \underbrace{x_j \mathbf{e}_i}_{3\times 1} \\
&= \underbrace{(a^{[3]} - y^{(i)})w^{[3]} \circ \sigma(z^{[2]})(1 - \sigma(z^{[2]}))w^{[2]} \circ \sigma(z^{[1]})(1 - \sigma(z^{[1]}))}_{1\times 3} \underbrace{x_j \mathbf{e}_i}_{3\times 1} \\
&= \underbrace{\left[(a^{[3]} - y^{(i)})w^{[3]} \circ \sigma(z^{[2]})(1 - \sigma(z^{[2]}))w^{[2]} \circ \sigma(z^{[1]})(1 - \sigma(z^{[1]}))\right]_i x_j}_{1\times 1} \\
&= \underbrace{\left[(a^{[3]} - y^{(i)})w^{[3]} \circ \sigma(z^{[2]})(1 - \sigma(z^{[2]}))w^{[2]} \circ \sigma(z^{[1]})(1 - \sigma(z^{[1]}))\right] x^T}_{3\times n}
\end{aligned}$$

where $x \in \mathbb{R}^{n\times 1}$ and $\mathbf{e}_i \in \mathbb{R}^{3\times 1}$ is the i^{th} basis vector.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial b^{[1]}} &= \frac{\partial \mathcal{L}}{\partial a^{[3]}} \cdot \frac{\partial a^{[3]}}{\partial z^{[3]}} \cdot \frac{\partial z^{[3]}}{\partial a^{[2]}} \cdot \frac{\partial a^{[2]}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial a^{[1]}} \cdot \frac{\partial a^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial b^{[1]}} \\
&= (a^{[3]} - y^{(i)})w^{[3]} d\sigma(\mathbf{z}^{[2]}) w^{[2]} d\sigma(\mathbf{z}^{[1]}) \frac{\partial (w^{[1]}x + b^{[1]})}{\partial b^{[1]}} \\
&= (a^{[3]} - y^{(i)})w^{[3]} d\sigma(\mathbf{z}^{[2]}) w^{[2]} d\sigma(\mathbf{z}^{[1]}) \frac{\partial (b^{[1]})}{\partial b^{[1]}} \\
&= \underbrace{(a^{[3]} - y^{(i)})}_{1\times 1} \underbrace{w^{[3]}}_{1\times 2} \underbrace{diag(\sigma'[\mathbf{z}^{[2]}])d\mathbf{z}^{[2]}}_{2\times 2} \underbrace{w^{[2]}}_{2\times 3} \underbrace{diag(\sigma'[\mathbf{z}^{[1]}])d\mathbf{z}^{[1]}}_{3\times 3} \underbrace{\mathbf{e}_i}_{3\times 1} \\
&= \underbrace{\left[(a^{[3]} - y^{(i)})w^{[3]} \circ \left(\sigma(z^{[2]})(1 - \sigma(z^{[2]})\right) w^{[2]} \circ \left(\sigma(z^{[1]})(1 - \sigma(z^{[1]}))\right)\right]}_{1\times 3} \underbrace{\mathbf{e}_i}_{3\times 1}
\end{aligned}$$

$$= \underbrace{\left[(a^{[3]} - y^{(i)})w^{[3]} \circ \left(\sigma(z^{[2]})(1-\sigma(z^{[2]})\right) w^{[2]} \circ \left(\sigma(z^{[1]})(1-\sigma(z^{[1]})\right)\right]_i}_{1\times 1}$$
$$= \underbrace{\left[(a^{[3]} - y^{(i)})w^{[3]} \circ \left(\sigma(z^{[2]})(1-\sigma(z^{[2]})\right) w^{[2]} \circ \left(\sigma(z^{[1]})(1-\sigma(z^{[1]})\right)\right]}_{3\times 1}$$

where $b^{[1]} \in \mathbb{R}^{3\times 1}$ and $\mathbf{e}_i \in \mathbb{R}^3$ is the i^{th} basis vector. With the necessary derivatives evaluated, let's summarize our results.

First, let's consider the derivatives of the loss with respect to the weight parameters.

$$\frac{\partial \mathcal{L}}{\partial w^{[3]}} = (a^{[3]} - y^{(i)})a^{[2]T} \in \mathbb{R}^{1\times 2}$$
$$\frac{\partial \mathcal{L}}{\partial w^{[2]}} = \left[(a^{[3]} - y^{(i)})w^{[3]} \circ (a^{[3]}(1-a^{[3]}))\right] a^{[1]T} \in \mathbb{R}^{2\times 3}$$
$$\frac{\partial \mathcal{L}}{\partial w^{[1]}} = \left[(a^{[3]} - y^{(i)})w^{[3]} \circ \sigma(z^{[2]})(1-\sigma(z^{[2]}))w^{[2]} \circ \sigma(z^{[1]})(1-\sigma(z^{[1]}))\right] x^T \in \mathbb{R}^{3\times n}$$

Second, lets consider the derivatives of the loss with respect to the bias parameters.

$$\frac{\partial \mathcal{L}}{\partial b^{[3]}} = (a^{[3]} - y^{(i)}) \in \mathbb{R}^{1\times 1}$$
$$\frac{\partial \mathcal{L}}{\partial b^{[2]}} = (a^{[3]} - y^{(i)})w^{[3]} \circ (a^{[3]}(1-a^{[3]})) \in \mathbb{R}^{2\times 1}$$
$$\frac{\partial \mathcal{L}}{\partial b^{[1]}} = (a^{[3]} - y^{(i)})w^{[3]} \circ \left(\sigma(z^{[2]})(1-\sigma(z^{[2]})\right) w^{[2]} \circ \left(\sigma(z^{[1]})(1-\sigma(z^{[1]})\right) \in \mathbb{R}^{3\times 1}$$

With these derivative equations, we can perform gradient descent with updates (15) and (16).