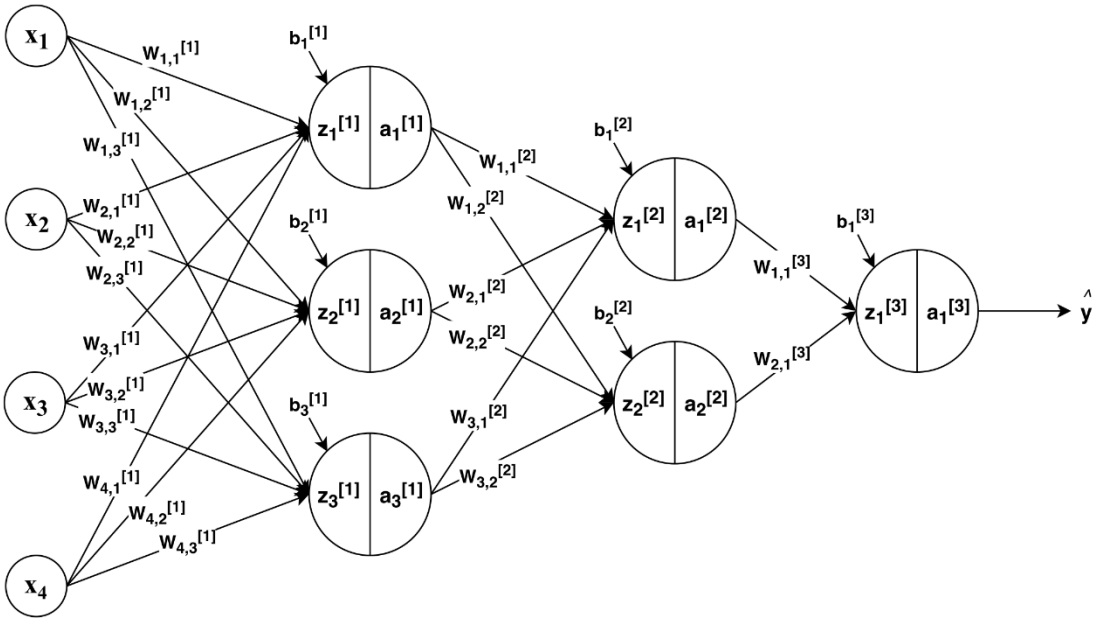
Back Propagation



Forward propagation

Note: It is common that the vector is presented as a column, for example, $a^{[0]} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$. However, we usually present it as a row vector when implementing algorithms. Therefore here we choose the latter way so that the mathematic result is consistent with coding result.

The first layer:

$$a^{[0]} = [x_1 \quad x_2 \quad x_3 \quad x_4]$$

$$z_1^{[1]} = W_{1,1}^{[1]} a_1^{[0]} + W_{2,1}^{[1]} a_2^{[0]} + W_{3,1}^{[1]} a_3^{[0]} + W_{4,1}^{[1]} a_4^{[0]} + b_1^{[1]} = a^{[0]} W_1^{[1]} + b_1^{[1]}$$

$$z_2^{[1]} = W_{1,2}^{[1]} a_1^{[0]} + W_{2,2}^{[1]} a_2^{[0]} + W_{3,2}^{[1]} a_3^{[0]} + W_{4,2}^{[1]} a_4^{[0]} + b_2^{[1]} = a^{[0]} W_2^{[1]} + b_2^{[1]}$$

$$z_3^{[1]} = W_{1,3}^{[1]} a_1^{[0]} + W_{2,3}^{[1]} a_2^{[0]} + W_{3,3}^{[1]} a_3^{[0]} + W_{4,3}^{[1]} a_4^{[0]} + b_3^{[1]} = a^{[0]} W_3^{[1]} + b_3^{[0]}$$

$$z^{[1]} = \begin{bmatrix} z_1^{[1]} & z_2^{[1]} & z_3^{[1]} \end{bmatrix} = a^{[0]} \begin{bmatrix} W_1^{[1]} & W_2^{[1]} & W_3^{[1]} \end{bmatrix} + \begin{bmatrix} b_1^{[1]} & b_2^{[1]} & b_3^{[1]} \end{bmatrix} = a^{[0]} W^{[1]} + b^{[1]}$$

Where $W^{[1]} \in \mathbb{R}^{4\times3}$,4 rows correspond to 4 inputs; 3 columns correspond to 3 outputs. $a^{[0]}$ has the shape 1×4, $b^{[1]}$ has the shape 1×3. Therefore, $z^{[1]}$ has the shape 1×3.

$$W^{[1]} = \begin{bmatrix} W_1^{[1]} & W_2^{[1]} & W_3^{[1]} \end{bmatrix} = \begin{bmatrix} W_{1,1}^{[1]} & W_{1,2}^{[1]} & W_{1,3}^{[1]} \\ W_{2,1}^{[1]} & W_{2,2}^{[1]} & W_{2,3}^{[1]} \\ W_{3,1}^{[1]} & W_{3,2}^{[1]} & W_{3,3}^{[1]} \\ W_{4,1}^{[1]} & W_{4,2}^{[1]} & W_{4,3}^{[1]} \end{bmatrix}$$

Since:

$$a_1^{[1]} = g_h\left(z_1^{[1]}\right) = \frac{1}{1 - e^{-z_1^{[1]}}}$$

$$a_2^{[1]} = g_h\left(z_2^{[1]}\right) = \frac{1}{1 - e^{-z_2^{[1]}}}$$
$$a_3^{[1]} = g_h\left(z_3^{[1]}\right) = \frac{1}{1 - e^{-z_3^{[1]}}}$$

$a^{[1]}$ has the same shape as $z^{[1]}$: 1×3: $a^{[1]} = \begin{bmatrix} a_1^{[1]} & a_2^{[1]} & a_3^{[1]} \end{bmatrix}$

The second layer:

$$z_1^{[2]} = W_{1,1}^{[2]} a_1^{[1]} + W_{2,1}^{[2]} a_2^{[1]} + W_{3,1}^{[2]} a_3^{[1]} + b_1^{[2]} = a^{[1]} W_1^{[2]} + b_1^{[2]}$$

$$z_2^{[2]} = W_{1,2}^{[2]} a_1^{[1]} + W_{2,2}^{[2]} a_2^{[1]} + W_{3,2}^{[2]} a_3^{[1]} + b_2^{[2]} = a^{[1]} W_2^{[2]} + b_2^{[2]}$$

$$z^{[2]} = \begin{bmatrix} z_1^{[2]} & z_2^{[2]} \end{bmatrix} = a^{[1]} \begin{bmatrix} W_1^{[2]} & W_2^{[2]} \end{bmatrix} + \begin{bmatrix} b_1^{[2]} & b_2^{[2]} \end{bmatrix} = a^{[1]} W^{[2]} + b^{[2]}$$

Where $W^{[2]} \in \mathbb{R}^{3\times 2}$, 3 rows correspond to 3 inputs; 2 columns correspond to 2 outputs. Since $a^{[1]}$ has the shape 1×3, $b^{[1]}$ has the shape 1×2. Therefore, $z^{[2]}$ has the shape 1×2.

$$W^{[2]} = \begin{bmatrix} W_1^{[2]} & W_2^{[2]} \end{bmatrix} = \begin{bmatrix} W_{1,1}^{[2]} & W_{1,2}^{[2]} \\ W_{2,1}^{[2]} & W_{2,2}^{[2]} \\ W_{3,1}^{[2]} & W_{3,2}^{[2]} \end{bmatrix}$$

Since:

$$a_1^{[2]} = g_h\left(z_1^{[2]}\right) = \frac{1}{1 - e^{-z_1^{[2]}}}$$
$$a_2^{[2]} = g_h\left(z_2^{[2]}\right) = \frac{1}{1 - e^{-z_2^{[2]}}}$$

$a^{[2]}$ has the same shape as $z^{[2]}$: 1×2: $a^{[2]} = \begin{bmatrix} a_1^{[2]} & a_2^{[2]} \end{bmatrix}$

The third layer:

$$z_1^{[3]} = W_{1,1}^{[3]} a_1^{[2]} + W_{2,1}^{[3]} a_2^{[2]} + b_1^{[3]} = a^{[2]} W_1^{[3]} + b_1^{[2]}$$

$$z^{[3]} = \left[z_1^{[3]}\right] = {W_1^{[3]}}^T a^{[2]} + b_1^{[2]} = a^{[2]} W^{[3]} + b^{[2]}$$

Where $W^{[3]}$ has the shape 2×1, $a^{[2]}$ has the shape 1×2, $b^{[3]}$ has the shape 1×1. Therefore, $z^{[3]}$ has the shape 1×1.

$$W^{[3]} = W_1^{[3]} = \begin{bmatrix} W_{1,1}^{[3]} \\ W_{2,1}^{[3]} \end{bmatrix}$$

$W^{[3]} \in \mathbb{R}^{2\times 1}$,2 rows correspond to 2 inputs; 1 column correspond to 1 output.

Since:

$$a_1^{[3]} = g_o\left(z_1^{[3]}\right) = \frac{1}{1 - e^{-z_1^{[3]}}}$$
$$\hat{y} = a_1^{[3]}$$

$a^{[3]}$ has the same shape as $z^{[3]}$: 1×1: $a^{[3]} = \begin{bmatrix} a_1^{[3]} \end{bmatrix}$

- Performance measure
  1. We need something that is a continuous function of $\hat{y}$ so that we can optimize it to find the solution for $W^{[k]}, b^{[k]}, k = 1,2,3$.
  2. We can use $\mathcal{L}_v(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2$ as before, we can also use:

$$\mathcal{L}(\hat{y}, y) = \mathcal{L}\big(y, a^{[0]}, W\big) = -[(1 - y)\log(1 - \hat{y}) + y\log(\hat{y})]$$

  3. It turns out that the latter one gives a nice form of expression after taking derivatives and it can also be seen as the log maximum likelihood function or cross-entropy.

Back propagation

First, we need to initialize the values of $W_0^{[k]}, b_0^{[k]}, k = 1,2,3$, by giving them random values.

Second, we need to perform minimization of $\mathcal{L}(\hat{y}, y)$ using gradient descent method to update values of $W^{[k]}, b^{[k]}, k = 1,2,3$. Let α be a parameter controlling the learning rate. Then

$$W_{s+1}^{[k]} = W_s^{[k]} - \alpha\frac{\partial\mathcal{L}(\hat{y}, y)}{\partial W_s^{[k]}}$$

Where $W_s^{[k]}$ is the value of $W^{[k]}$ in step s. When it converges, we have obtained the values of $W^{[k]}, b^{[k]}, k = 1,2,3$.

Now, let's see how to calculate $\frac{\partial\mathcal{L}(\hat{y},y)}{\partial \mathrm{W}}$ using back propagation. Here we use a variable denoted as $g^{[nx]}$, where n is the index of layer and x is the gradient variable (a, z, W, b). For example, $g^{[3a]}$ is the gradient of loss function w.r.t to activation a in the third layer.

Back propagation begins with the output layer $z^{[3]}|a^{[3]}$. Read the following content with the network architecture in the very beginning part, which gives you the clue of back propagation.

There are only one neuron in the last layer. Let's look at it in backward flow:

Note that from here we present the result by only taking the result in the last step (only use its notation rather than full expansion of it) and multiplying it by the local gradient.

$$g^{[3a]} = \frac{d\mathcal{L}(\hat{y}, y)}{da_1^{[3]}} = -\frac{d}{da_1^{[3]}}\left[(1 - y)\log\left(1 - a_1^{[3]}\right) + y\log\left(a_1^{[3]}\right)\right] = \frac{1 - y}{1 - a_1^{[3]}} - \frac{y}{a_1^{[3]}}$$

$$= \frac{a_1^{[3]}(1 - y) - y\left(1 - a_1^{[3]}\right)}{a_1^{[3]}\left(1 - a_1^{[3]}\right)} = \frac{a_1^{[3]} - y}{a_1^{[3]}\left(1 - a_1^{[3]}\right)}$$

$$g_o'\left(z_1^{[3]}\right) = \frac{da_1^{[3]}}{dz_1^{[3]}} = \frac{d}{dz_1^{[3]}}\left(\frac{1}{1 - e^{-z_1^{[3]}}}\right) = \frac{d}{dz_1^{[3]}}\left(1 - e^{-z_1^{[3]}}\right)^{-1} = -1 \times \left(1 - e^{-z_1^{[3]}}\right)^{-2} \cdot \left(-e^{-z_1^{[3]}}\right)$$

$$= \frac{e^{-z_1^{[3]}}}{\left(1 - e^{-z_1^{[3]}}\right)^2} = \frac{1}{1 - e^{-z_1^{[3]}}} \cdot \frac{e^{-z_1^{[3]}}}{1 - e^{-z_1^{[3]}}} = a_1^{[3]}\left(1 - a_1^{[3]}\right)$$

Where $g_o'(\dots)$ is the derivate of activation function in output layer. And we will see $g_h'(\dots)$ in the following derivation, denoting this derivate in hidden layers.

By chaining the local gradient with the upstream gradient, we have:

$$g^{[3z]} = \frac{d\mathcal{L}(\hat{y}, y)}{dz_1^{[3]}} = \frac{da_1^{[3]}}{dz_1^{[3]}} \cdot \frac{d\mathcal{L}(\hat{y}, y)}{da_1^{[3]}} = g_o'\left(z_1^{[3]}\right) \cdot g^{[3a]}$$

Since

$$\frac{\partial z_1^{[3]}}{\partial W_{1,1}^{[3]}} = \frac{\partial}{\partial W_{1,1}^{[3]}}\left(W_{1,1}^{[3]}a_1^{[2]} + W_{2,1}^{[3]}a_2^{[2]} + b_1^{[3]}\right) = a_1^{[2]}$$

By applying chain rule again, we have $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[3]}} = \frac{\partial z_1^{[3]}}{\partial W_{1,1}^{[3]}} \cdot \frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[3]}} = a_1^{[2]} \cdot g^{[3z]}$

Similarly, by repeating above steps, we have

$$\frac{\partial z_1^{[3]}}{\partial W_{2,1}^{[3]}} = \frac{\partial}{\partial W_{1,1}^{[3]}}\left(W_{1,1}^{[3]}a_1^{[2]} + W_{2,1}^{[3]}a_2^{[2]} + b_1^{[3]}\right) = a_2^{[2]}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[3]}} = a_2^{[2]} \cdot g^{[3z]}$$

And the computations of this layer stop here. To summarize the result obtained in the layer:

$$g^{[3W]} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W^{[3]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[3]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[3]}} \end{bmatrix}_{2\times 1} = \begin{bmatrix} a_1^{[2]} \cdot g^{[3z]} \\ a_2^{[2]} \cdot g^{[3z]} \end{bmatrix}_{2\times 1} = \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \end{bmatrix}_{2\times 1} \cdot g^{[3z]}{}_{1\times 1} = \left(a^{[2]}\right)^T{}_{2\times 1} \cdot g^{[3z]}{}_{1\times 1}$$

By seeing $b^{[3]}$ as the coefficient of $z^{[3]}$,we have:

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial b^{[3]}} = g^{[3z]}{}_{1\times 1}$$

Now, let's consider the hidden layer $z^{[2]}|a^{[2]}$. There are two neurons in the layer. Let's consider the first neuron and once the gradient of its corresponding parameters are derived, that of the second neuron is similar to get.

Let's first consider $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[2]}}$. We have:

$$\frac{\partial z_1^{[3]}}{\partial a_1^{[2]}} = \frac{\partial}{\partial a_1^{[2]}}\left(W_{1,1}^{[3]}a_1^{[2]} + W_{2,1}^{[3]}a_2^{[2]} + b_1^{[3]}\right) = W_{1,1}^{[3]}$$

By chaining it with $\frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[3]}} = g^{[3z]}$, which is obtained in last layer, we can get:

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}} = \frac{\partial z_1^{[3]}}{\partial a_1^{[2]}} \cdot \frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[3]}} = W_{1,1}^{[3]} \cdot \frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[3]}} = W_{1,1}^{[3]} \cdot g^{[3z]}$$

By directly using the result of logistic function derivative and chaining it with upstream gradient:

$$g_{\mathrm{h}}'\left(z_1^{[2]}\right) = \frac{da_1^{[2]}}{dz_1^{[2]}} = a_1^{[2]}\left(1 - a_1^{[2]}\right)$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[2]}} = \frac{da_1^{[2]}}{dz_1^{[2]}} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}} = g_h'\left(z_1^{[2]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}}$$

Compute the local gradient and apply chain rule, we have:

$$\frac{\partial z_1^{[2]}}{\partial W_{1,1}^{[2]}} = \frac{\partial}{\partial W_{1,1}^{[2]}}\left(W_{1,1}^{[2]}a_1^{[1]} + W_{2,1}^{[2]}a_2^{[1]} + W_{3,1}^{[2]}a_3^{[1]} + b_1^{[2]}\right) = a_1^{[1]}$$

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{1,1}^{[2]}} = \frac{\partial z_1^{[2]}}{\partial W_{1,1}^{[2]}} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} = a_1^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}$$

Similarly, $\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{2,1}^{[2]}}$ and $\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{3,1}^{[2]}}$ can be obtained by using the result of $\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}$ since they are also related to the first neuron:

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{2,1}^{[2]}} = a_2^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} \qquad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{3,1}^{[2]}} = a_3^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}$$

By seeing $b_1^{[2]}$ as the coefficient of $z_1^{[2]}$, we have:

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b_1^{[2]}} = \frac{\partial z_1^{[2]}}{\partial b_1^{[2]}} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}$$

Then,

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_1^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{1,1}^{[2]}} \\ \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{2,1}^{[2]}} \\ \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{3,1}^{[2]}} \end{bmatrix}_{3\times 1} = \begin{bmatrix} a_1^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} \\ a_2^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} \\ a_3^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}} \end{bmatrix}_{3\times 1} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \end{bmatrix}_{3\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}_{1\times 1} = \left(a^{[1]}\right)^T{}_{3\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[2]}}_{1\times 1}$$

Next, we consider parameters related to the second neuron. Using the same method above, we have:

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_2^{[2]}} = \frac{\partial z_1^{[3]}}{\partial a_2^{[2]}} \cdot \frac{d\mathcal{L}(\hat{y}, y)}{dz_1^{[3]}} = W_{2,1}^{[3]} \cdot \frac{d\mathcal{L}(\hat{y}, y)}{dz_1^{[3]}} = W_{2,1}^{[3]} \cdot g^{[3z]}$$

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}} = \frac{da_2^{[2]}}{dz_2^{[2]}} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_2^{[2]}} = g_h'\left(z_2^{[2]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_2^{[2]}}$$

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{1,2}^{[2]}} = \frac{\partial z_2^{[2]}}{\partial W_{1,2}^{[2]}} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}} = a_1^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}}$$

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{2,2}^{[2]}} = a_2^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}} \qquad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_{3,2}^{[2]}} = a_3^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}}$$

$$\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b_2^{[2]}} = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[2]}}$$

Then,

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_2^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,2}^{[2]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,2}^{[2]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,2}^{[2]}} \end{bmatrix}_{3\times 1} = \begin{bmatrix} a_1^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \\ a_2^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \\ a_3^{[1]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \end{bmatrix}_{3\times 1} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \end{bmatrix}_{3\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}}_{1\times 1} = \left(a^{[1]}\right)^T{}_{3\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}}_{1\times 1}$$

To conclude all the stuff in the hidden layer $z^{[2]}|a^{[2]}$:

$$g^{[2a]} = \frac{d\mathcal{L}(\hat{y},y)}{\partial a^{[2]}}_{1\times 2} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[2]}} \end{bmatrix}_{1\times 2} = \begin{bmatrix} W_{1,1}^{[3]} \cdot g^{[3z]} & W_{2,1}^{[3]} \cdot g^{[3z]} \end{bmatrix}_{1\times 2}$$

$$= g^{[3z]}{}_{1\times 1} \cdot \begin{bmatrix} W_{1,1}^{[3]} & W_{2,1}^{[3]} \end{bmatrix}_{1\times 2} = g^{[3z]}{}_{1\times 1} \cdot \left(W^{[3]}\right)^T{}_{1\times 2}$$

$$g^{[2z]} = \frac{d\mathcal{L}(\hat{y},y)}{\partial z^{[2]}}_{1\times 2} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \end{bmatrix}_{1\times 2}$$

$$= \begin{bmatrix} g_h'\left(z_1^{[2]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}} & g_h'\left(z_2^{[2]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[2]}} \end{bmatrix}_{1\times 2}$$

$$= \begin{bmatrix} g_h'\left(z_1^{[2]}\right) & g_h'\left(z_2^{[2]}\right) \end{bmatrix}_{1\times 2} \circ \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[2]}} \end{bmatrix}_{1\times 2}$$

$$= g_h'\left(z^{[2]}\right)_{1\times 2} \circ g^{[2a]}{}_{1\times 2}$$

Where "$\circ$" is called the Hadamard product, which is the elementwise product of the matrices.

$$g^{[2W]} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_2^{[2]}} \end{bmatrix}_{3\times 2} = \begin{bmatrix} \left(a^{[1]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[2]}} & \left(a^{[1]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \end{bmatrix}_{3\times 2}$$

$$= \left(a^{[1]}\right)^T{}_{3\times 1} \cdot \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \end{bmatrix}_{1\times 2} = \left(a^{[1]}\right)^T{}_{3\times 1} \cdot g^{[2z]}{}_{1\times 2}$$

$$g^{[2b]} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial b^{[2]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial b_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial b_2^{[2]}} \end{bmatrix}_{1\times 2} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[2]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[2]}} \end{bmatrix}_{1\times 2} = g^{[2z]}{}_{1\times 2}$$

Now, let's consider the input layer $z^{[1]}|a^{[1]}$. Again, there are three neurons in this layer and we only show you the derivation of gradients related to the first one and that of the rest neurons are similar.

First, we consider $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[1]}}$, $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[1]}}$, $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,1}^{[1]}}$, $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,1}^{[1]}}$ since they are related to the first neuron.

Note that there are two flows back to $a_1^{[1]}$ since $a_1^{[1]}$ contributes to both $z_1^{[2]}$ and $z_2^{[2]}$ in the forward propagation. So we need to consider two parts separately and chain each with its own

upstream gradient:

$$\frac{\partial z_1^{[2]}}{\partial a_1^{[1]}} = \frac{\partial}{\partial a_1^{[1]}}\left(W_{1,1}^{[2]}a_1^{[1]} + W_{2,1}^{[2]}a_2^{[1]} + W_{3,1}^{[2]}a_3^{[1]} + b_1^{[2]}\right) = W_{1,1}^{[2]}$$

$$\frac{\partial z_2^{[2]}}{\partial a_1^{[1]}} = \frac{\partial}{\partial a_1^{[1]}}\left(W_{1,2}^{[2]}a_1^{[1]} + W_{2,2}^{[2]}a_2^{[1]} + W_{3,2}^{[2]}a_3^{[1]} + b_2^{[2]}\right) = W_{1,2}^{[2]}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[1]}} = \left(\frac{\partial z_1^{[2]}}{\partial a_1^{[1]}}\cdot\frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[2]}} + \frac{\partial z_2^{[2]}}{\partial a_1^{[1]}}\cdot\frac{d\mathcal{L}(\hat{y},y)}{dz_2^{[2]}}\right) = \left(W_{1,1}^{[2]}\cdot\frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[2]}} + W_{1,2}^{[2]}\cdot\frac{d\mathcal{L}(\hat{y},y)}{dz_2^{[2]}}\right)$$

$$= \begin{bmatrix}\frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[2]}} & \frac{d\mathcal{L}(\hat{y},y)}{dz_2^{[2]}}\end{bmatrix}\cdot\begin{bmatrix}W_{1,1}^{[2]}\\ W_{1,2}^{[2]}\end{bmatrix} = g^{[2z]}\cdot\begin{bmatrix}W_{1,1}^{[2]}\\ W_{1,2}^{[2]}\end{bmatrix}$$

$$\frac{da_1^{[1]}}{dz_1^{[1]}} = g_h'\left(z_1^{[1]}\right) = a_1^{[1]}\left(1 - a_1^{[1]}\right)$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} = \frac{da_1^{[1]}}{dz_1^{[1]}}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[1]}} = g_h'\left(z_1^{[1]}\right)\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[1]}}$$

$$\frac{\partial z_1^{[1]}}{\partial W_{1,1}^{[1]}} = \frac{\partial}{\partial W_{1,1}^{[1]}}\left(W_{1,1}^{[1]}a_1^{[0]} + W_{2,1}^{[1]}a_2^{[0]} + W_{3,1}^{[1]}a_3^{[0]} + W_{4,1}^{[1]}a_4^{[0]} + b_1^{[1]}\right) = a_1^{[0]}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[1]}} = \frac{\partial z_1^{[1]}}{\partial W_{1,1}^{[1]}}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} = a_1^{[0]}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}$$

Similarly, $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[1]}}$, $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,1}^{[1]}}$ and $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,1}^{[1]}}$ can use the result of $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}$ since they are only related to the first neuron:

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[1]}} = a_2^{[0]}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \qquad \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,1}^{[1]}} = a_3^{[0]}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \qquad \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,1}^{[1]}} = a_4^{[0]}\cdot\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}$$

By the way, we can get $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial}$ by seeing $b_1^{[1]}$as a coefficient of $z_1^{[1]}$:

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial b_1^{[1]}} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}\cdot\frac{\partial z_1^{[1]}}{\partial b_1^{[1]}} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}$$

Then,

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_1^{[1]}} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,1}^{[1]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,1}^{[1]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,1}^{[1]}} \\ \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,1}^{[1]}} \end{bmatrix}_{4\times 1} = \begin{bmatrix} a_1^{[0]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \\ a_2^{[0]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \\ a_3^{[0]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \\ a_4^{[0]} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}} \end{bmatrix}_{4\times 1} = \begin{bmatrix} a_1^{[0]} \\ a_2^{[0]} \\ a_3^{[0]} \\ a_4^{[0]} \end{bmatrix}_{4\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}_{1\times 1} = \left(a^{[0]}\right)^T{}_{4\times 1} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_1^{[1]}}_{1\times 1}$$

Second, we consider $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,2}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,2}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,2}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,2}^{[1]}}$ since they are related to the second neuron.

By using the same method above, we have:

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[1]}} = \left(\frac{\partial z_1^{[2]}}{\partial a_2^{[1]}} \cdot \frac{d\mathcal{L}(\hat{y},y)}{dz_1^{[2]}} + \frac{\partial z_2^{[2]}}{\partial a_2^{[1]}} \cdot \frac{d\mathcal{L}(\hat{y},y)}{dz_2^{[2]}}\right) = g^{[2z]} \cdot \begin{bmatrix} W_{3,1}^{[2]} \\ W_{3,2}^{[2]} \end{bmatrix}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[1]}} = \frac{da_2^{[1]}}{dz_2^{[1]}} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[1]}} = g_h'\left(z_2^{[1]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[1]}}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_2^{[1]}} = \left(a^{[0]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[1]}}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial b_2^{[1]}} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_2^{[1]}}$$

Lastly, we consider $\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{1,3}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{2,3}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{3,3}^{[1]}}, \frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_{4,3}^{[1]}}$ since they are related to the third neuron.

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_3^{[1]}} = g^{[2z]} \cdot \begin{bmatrix} W_{3,1}^{[2]} \\ W_{3,2}^{[2]} \end{bmatrix}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_3^{[1]}} = \frac{da_3^{[1]}}{dz_3^{[1]}} \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_3^{[1]}} = g_h'\left(z_3^{[1]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_3^{[1]}}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial W_3^{[1]}} = \left(a^{[0]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_3^{[1]}}$$

$$\frac{\partial \mathcal{L}(\hat{y},y)}{\partial b_3^{[1]}} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial z_3^{[1]}}$$

To conclude all the stuff in the hidden layer $z^{[1]}|a^{[1]}$:

$$g^{[1a]} = \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a^{[1]}}_{1\times 3} = \begin{bmatrix} \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_1^{[1]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_2^{[1]}} & \frac{\partial \mathcal{L}(\hat{y},y)}{\partial a_3^{[1]}} \end{bmatrix}_{1\times 3}$$

$$= \begin{bmatrix} g^{[2z]} \cdot \begin{bmatrix} W_{1,1}^{[2]} \\ W_{1,2}^{[2]} \end{bmatrix} & g^{[2z]} \cdot \begin{bmatrix} W_{2,1}^{[2]} \\ W_{2,2}^{[2]} \end{bmatrix} & g^{[2z]} \cdot \begin{bmatrix} W_{3,1}^{[2]} \\ W_{3,2}^{[2]} \end{bmatrix} \end{bmatrix}_{1\times 3}$$

$$= g^{[2z]}{}_{1\times 2} \cdot \begin{bmatrix} W_{1,1}^{[2]} & W_{2,1}^{[2]} & W_{3,1}^{[2]} \\ W_{1,2}^{[2]} & W_{2,2}^{[2]} & W_{3,2}^{[2]} \end{bmatrix}_{2\times 3} = g^{[2z]}{}_{1\times 2} \cdot \left(W^{[2]}\right)^T$$

$$g^{[1z]} = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z^{[1]}}_{1\times 3} = \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_3^{[1]}}\right]_{1\times 3}$$

$$= \left[g_h'\left(z_1^{[1]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_1^{[1]}} \quad g_h'\left(z_2^{[1]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_2^{[1]}} \quad g_h'\left(z_3^{[1]}\right) \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_3^{[1]}}\right]_{1\times 3}$$

$$= \left[g_h'\left(z_1^{[1]}\right) \quad g_h'\left(z_2^{[1]}\right) \quad g_h'\left(z_3^{[1]}\right)\right]_{1\times 3} \circ \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial a_3^{[1]}}\right]_{1\times 3}$$

$$= g_h'\left(z^{[1]}\right)_{1\times 3} \circ g^{[1a]}$$

$$g^{[1w]} = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W^{[1]}} = \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial W_3^{[1]}}\right]_{4\times 3}$$

$$= \left[\left(a^{[0]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[1]}} \quad \left(a^{[0]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[1]}} \quad \left(a^{[0]}\right)^T \cdot \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_3^{[1]}}\right]_{4\times 3}$$

$$= \left(a^{[0]}\right)^T{}_{4\times 1} \cdot \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_3^{[1]}}\right]_{1\times 3} = \left(a^{[0]}\right)^T{}_{4\times 1} \cdot g^{[1z]}{}_{1\times 3}$$

$$= \left(a^{[0]}\right)^T{}_{4\times 1} \cdot g^{[1z]}{}_{1\times 3}$$

$$g^{[1b]} = \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b^{[1]}} = \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial b_3^{[1]}}\right]_{1\times 3} = \left[\frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_1^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_2^{[1]}} \quad \frac{\partial \mathcal{L}(\hat{y}, y)}{\partial z_3^{[1]}}\right]_{1\times 3}$$

$$= g^{[1z]}{}_{1\times 3}$$

So far, the derivation of back propagation is completed.