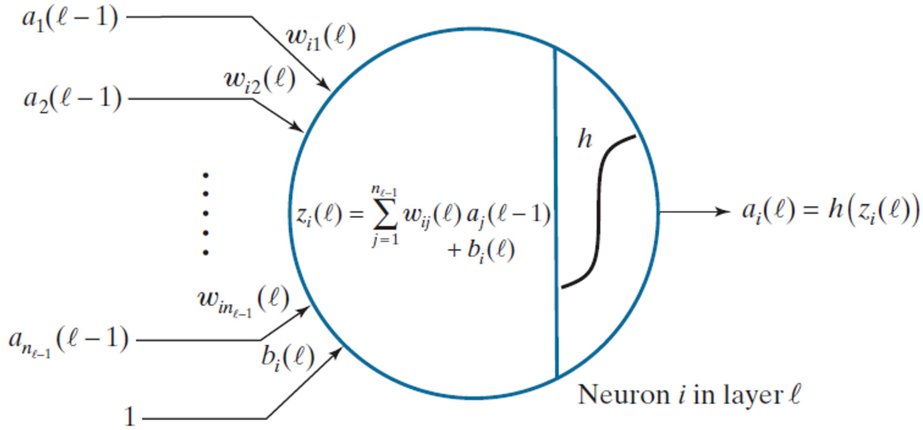
Lecture 26



- Previously, the activation function is a linear (identify) function, where $\frac{\partial E}{\partial a_i(l)} = \frac{\partial E}{\partial z_i(l)}$, since $a_i(l)=h(z_i(l)) = z_i(l)$.
- In general, $\frac{\partial E}{\partial z_i(l)} = \frac{\partial E}{\partial a_i(l)} \frac{\partial a_i(l)}{\partial z_i(l)} = \frac{\partial E}{\partial a_i} \frac{d(z_i(l))}{dz_i(l)} = \frac{\partial E}{\partial a_i} h'(z_i(l))$.
- Therefore, we need to integrate $h'(z_i(l))$ in the backpropagation formulation derived earlier.

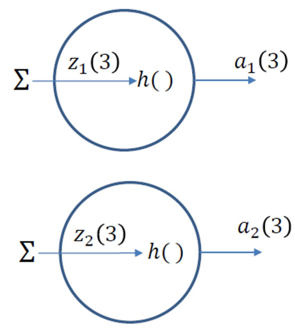
The newly defined $\boldsymbol{D}(3)$

The gradient of the output error with respect to the input of the final layer $\boldsymbol{Z}(3)$ is:

$$\boldsymbol{D}(3) = \frac{\partial E}{\partial \boldsymbol{Z}(3)} = \begin{bmatrix} \frac{\partial E}{\partial z_1(3)} \\ \frac{\partial E}{\partial z_2(3)} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial a_1(3)} \frac{\partial a_1(3)}{\partial z_1(3)} \\ \frac{\partial E}{\partial a_1(3)} \frac{\partial a_1(3)}{\partial z_1(3)} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial E}{\partial a_1(3)} h'(z_1(3)) \\ \frac{\partial E}{\partial a_1(3)} h'(z_2(3)) \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial a_1(3)} \\ \frac{\partial E}{\partial a_2(3)} \end{bmatrix} \odot \begin{bmatrix} h'(z_1(3)) \\ h'(z_2(3)) \end{bmatrix}$$

Elementwise Multiplication



Since $\frac{\partial E}{\partial \boldsymbol{A}(3)} = \begin{bmatrix} \frac{\partial E}{\partial a_1(3)} \\ \frac{\partial E}{\partial a_2(3)} \end{bmatrix} = \boldsymbol{A}(3) - \mathbf{r}$

$$\boldsymbol{D}(3) = [\boldsymbol{A}(3) - \mathbf{r}] \odot \begin{bmatrix} h'(z_1(3)) \\ h'(z_2(3)) \end{bmatrix}$$

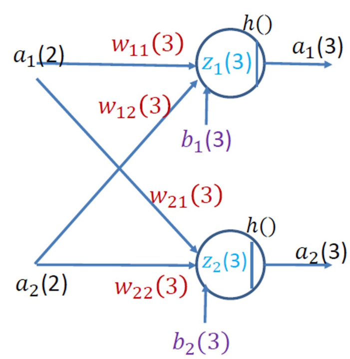
Modified Gradient of the Error wrt. Weights

$$\frac{\partial E}{\partial w_{11}(3)} = \frac{\partial E}{\partial z_1(3)} \frac{\partial z_1(3)}{\partial w_{11}(3)} = \frac{\partial E}{\partial z_1(3)} a_1(2)$$

$$\frac{\partial E}{\partial w_{12}(3)} = \frac{\partial E}{\partial z_1(3)} \frac{\partial z_1(3)}{\partial w_{12}(3)} = \frac{\partial E}{\partial z_1(3)} a_2(2)$$

$$\frac{\partial E}{\partial w_{21}(3)} = \frac{\partial E}{\partial z_2(3)} \frac{\partial z_2(3)}{\partial w_{21}(3)} = \frac{\partial E}{\partial z_2(3)} a_1(2)$$

$$\frac{\partial E}{\partial w_{22}(3)} = \frac{\partial E}{\partial z_2(3)} \frac{\partial z_2(3)}{\partial w_{22}(3)} = \frac{\partial E}{\partial z_2(3)} a_2(2)$$



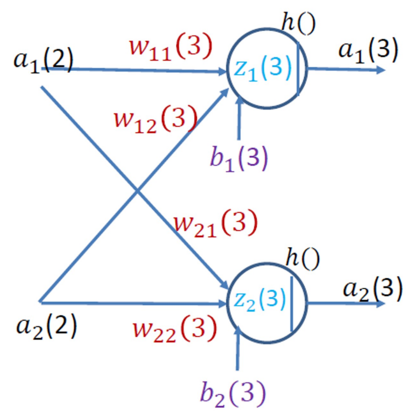
$$\frac{\partial E}{\partial \boldsymbol{W}(3)} = \begin{bmatrix} \frac{\partial E}{\partial w_{11}(3)} & \frac{\partial E}{\partial w_{12}(3)} \\ \frac{\partial E}{\partial w_{21}(3)} & \frac{\partial E}{\partial w_{22}(3)} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(3)} \\ \frac{\partial E}{\partial z_2(3)} \end{bmatrix} [a_1(2) \quad a_2(2)] = \frac{\partial E}{\partial \boldsymbol{Z}(3)} \boldsymbol{A}(2)^{\mathrm{T}} = \boldsymbol{D}(3)\boldsymbol{A}(2)^{\mathrm{T}}$$

where $\boldsymbol{Z}(3) = \begin{bmatrix} z_1(3) \\ z_2(3) \end{bmatrix} = \boldsymbol{W}(3)\boldsymbol{A}(2) + \boldsymbol{b}(3)$

Modified Gradient of the Error wrt. Biases

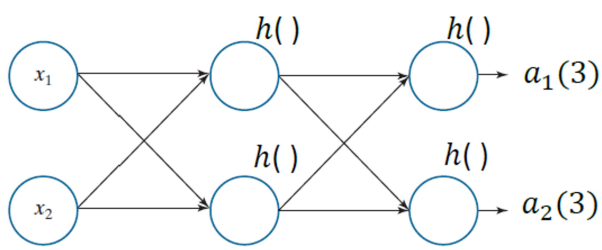
$$\frac{\partial E}{\partial b_1(3)} = \frac{\partial E}{\partial z_1(3)} \frac{\partial z_1(3)}{\partial b_1(3)} = \frac{\partial E}{\partial z_1(3)}$$

$$\frac{\partial E}{\partial b_2(3)} = \frac{\partial E}{\partial z_2(3)} \frac{\partial z_2(3)}{\partial b_2(3)} = \frac{\partial E}{\partial z_2(3)}$$



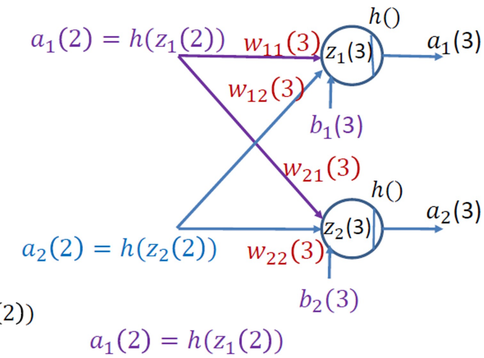
$$\frac{\partial E}{\partial \boldsymbol{b}(3)} = \begin{bmatrix} \frac{\partial E}{\partial b_1(3)} \\ \frac{\partial E}{\partial b_2(3)} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(3)} \\ \frac{\partial E}{\partial z_2(3)} \end{bmatrix} = \frac{\partial E}{\partial \boldsymbol{Z}(3)} = \boldsymbol{D}(3)$$

Forward Pass



Modified Relation between $\boldsymbol{D}(2)$ and $\boldsymbol{D}(3)$

$$\boldsymbol{D}(2) = \frac{\partial E}{\partial \boldsymbol{Z}(2)} = \begin{bmatrix} \frac{\partial E}{\partial z_1(2)} \\ \frac{\partial E}{\partial z_2(2)} \end{bmatrix}$$



$$\frac{\partial E}{\partial z_1(2)} = \frac{\partial E}{\partial z_1(3)}\frac{\partial z_1(3)}{\partial z_1(2)} + \frac{\partial E}{\partial z_2(3)}\frac{\partial z_2(3)}{\partial z_1(2)}$$

where

$$\frac{\partial z_1(3)}{\partial z_1(2)} = \frac{\partial z_1(3)}{\partial a_1(2)}\frac{\partial a_1(2)}{\partial z_1(2)} = w_{11}(3)h'(z_1(2))$$

$$\frac{\partial z_2(3)}{\partial z_1(2)} = \frac{\partial z_2(3)}{\partial a_1(2)}\frac{\partial a_1(2)}{\partial z_1(2)} = w_{21}(3)h'(z_1(2))$$

Thus

$$\frac{\partial E}{\partial z_1(2)} = \frac{\partial E}{\partial z_1(3)} w_{11}(3)h'(z_1(2)) + \frac{\partial E}{\partial z_2(3)} w_{21}(3)h'(z_1(2))$$

Modified Backpropagation Rule

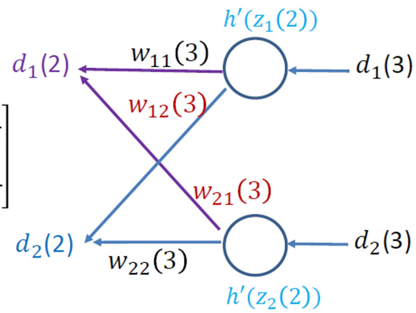
$$\frac{\partial E}{\partial z_1(2)} = \frac{\partial E}{\partial z_1(3)} w_{11}(3)h'(z_1(2)) + \frac{\partial E}{\partial z_2(3)} w_{21}(3)h'(z_1(2))$$

Similarly,

$$\frac{\partial E}{\partial z_2(2)} = \frac{\partial E}{\partial z_1(3)} w_{12}(3)h'(z_2(2)) + \frac{\partial E}{\partial z_2(3)} w_{22}(3)h'(z_2(2))$$

Thus $\boldsymbol{D}(2) = \begin{bmatrix} \frac{\partial E}{\partial z_1(2)} \\ \frac{\partial E}{\partial z_2(2)} \end{bmatrix} = \left\{ \begin{bmatrix} w_{11}(3) & w_{12}(3) \\ w_{21}(3) & w_{22}(3) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \frac{\partial E}{\partial z_1(3)} \\ \frac{\partial E}{\partial z_1(3)} \end{bmatrix} \right\} \odot \begin{bmatrix} h'(z_1(2)) \\ h'(z_2(2)) \end{bmatrix}$

$$\boldsymbol{D}(2) = [\boldsymbol{W}(3)^{\mathrm{T}}\boldsymbol{D}(3)] \odot h'(\boldsymbol{Z}(2))$$



$$\boldsymbol{D}(2) = \begin{bmatrix} d_1(2) \\ d_2(2) \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(2)} \\ \frac{\partial E}{\partial z_2(2)} \end{bmatrix} \qquad \boldsymbol{D}(3) = \begin{bmatrix} d_1(3) \\ d_2(3) \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(3)} \\ \frac{\partial E}{\partial z_2(3)} \end{bmatrix}$$

Modified Gradient of the Error wrt. Weights (Level Two)

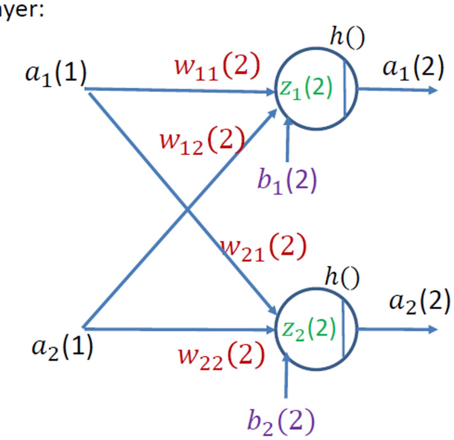
Similar to the previous derivations, for the 2nd layer:

$$\frac{\partial E}{\partial w_{11}(2)} = \frac{\partial E}{\partial z_1(2)} \frac{\partial z_1(2)}{\partial w_{11}(2)} = \frac{\partial E}{\partial z_1(2)} a_1(1)$$

$$\frac{\partial E}{\partial w_{12}(2)} = \frac{\partial E}{\partial z_1(2)} \frac{\partial z_1(2)}{\partial w_{12}(2)} = \frac{\partial E}{\partial z_1(2)} a_2(1)$$

$$\frac{\partial E}{\partial w_{21}(2)} = \frac{\partial E}{\partial z_2(2)} \frac{\partial z_2(2)}{\partial w_{21}(2)} = \frac{\partial E}{\partial z_2(2)} a_1(1)$$

$$\frac{\partial E}{\partial w_{22}(2)} = \frac{\partial E}{\partial z_2(2)} \frac{\partial z_2(2)}{\partial w_{22}(2)} = \frac{\partial E}{\partial z_2(2)} a_2(1)$$



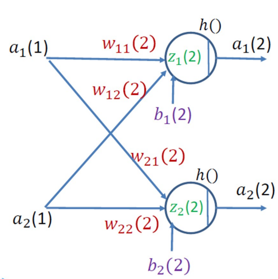
$$\frac{\partial E}{\partial \boldsymbol{W}(2)} = \begin{bmatrix} \frac{\partial E}{\partial w_{11}(2)} & \frac{\partial E}{\partial w_{12}(2)} \\ \frac{\partial E}{\partial w_{21}(2)} & \frac{\partial E}{\partial w_{22}(2)} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(2)} \\ \frac{\partial E}{\partial z_2(2)} \end{bmatrix} [a_1(1) \quad a_2(1)] = \frac{\partial E}{\partial \boldsymbol{Z}(2)} \boldsymbol{A}(1)^{\mathrm{T}} = \boldsymbol{D}(2) \boldsymbol{A}(1)^{\mathrm{T}}$$

Where $\boldsymbol{D}(2)$ is obtained by back propagating $\boldsymbol{D}(3)$, and $\boldsymbol{A}(1) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is the input vector.

Modified Gradient wrt. Biases (2nd Layer)

$$\frac{\partial E}{\partial b_1(2)} = \frac{\partial E}{\partial z_1(2)} \frac{\partial z_1(2)}{\partial b_1(2)} = \frac{\partial E}{\partial z_1(2)}$$

$$\frac{\partial E}{\partial b_2(2)} = \frac{\partial E}{\partial z_2(2)} \frac{\partial z_2(2)}{\partial b_2(2)} = \frac{\partial E}{\partial z_2(2)}$$

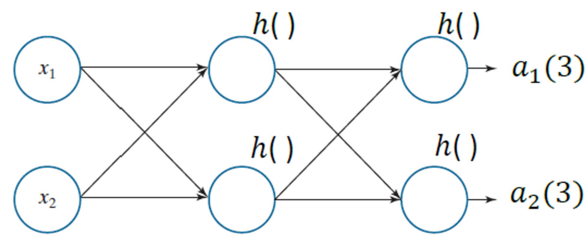


$$\frac{\partial E}{\partial \boldsymbol{b}(2)} = \begin{bmatrix} \frac{\partial E}{\partial b_1(2)} \\ \frac{\partial E}{\partial b_2(2)} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial z_1(2)} \\ \frac{\partial E}{\partial z_2(2)} \end{bmatrix} = \frac{\partial E}{\partial \boldsymbol{Z}(2)} = \boldsymbol{D}(2)$$

Summary of the Results

$$\boldsymbol{A}(1) = \begin{bmatrix} a_1(1) \\ a_2(1) \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Forward Pass



$$\boldsymbol{Z}(2) = \begin{bmatrix} z_1(2) \\ z_2(2) \end{bmatrix} = \boldsymbol{W}(2)\boldsymbol{A}(1) + \boldsymbol{b}(2)$$

$$\boldsymbol{A}(2) = h(\boldsymbol{Z}(2)) = \begin{bmatrix} h(z_1(2)) \\ h(z_2(2)) \end{bmatrix}$$

$$\boldsymbol{Z}(3) = \begin{bmatrix} z_1(3) \\ z_2(3) \end{bmatrix} = \boldsymbol{W}(3)\boldsymbol{A}(2) + \boldsymbol{b}(3)$$

$$\boldsymbol{A}(3) = h(\boldsymbol{Z}(3)) = \begin{bmatrix} h(z_1(3)) \\ h(z_2(3)) \end{bmatrix}$$

Error: $E = \frac{1}{2}\|\mathbf{r} - \boldsymbol{A}(3)\|^2$

$$\boldsymbol{D}(3) = [\boldsymbol{A}(3) - \mathbf{r}] \odot \begin{bmatrix} h'(z_1(3)) \\ h'(z_2(3)) \end{bmatrix}$$

Backpropagation:

$$\boldsymbol{D}(2) = \begin{bmatrix} d_1(3) \\ d_2(2) \end{bmatrix} = \boldsymbol{W}(3)^{\mathrm{T}}\boldsymbol{D}(3) \odot h'(\boldsymbol{Z}(2))$$

$$\frac{\partial E}{\partial \boldsymbol{W}(3)} = \boldsymbol{D}(3)\boldsymbol{A}(2)^{\mathrm{T}}, \quad \frac{\partial E}{\partial \boldsymbol{b}(3)} = \boldsymbol{D}(3)$$

$$\frac{\partial E}{\partial \boldsymbol{W}(2)} = \boldsymbol{D}(2)\boldsymbol{A}(1)^{\mathrm{T}}, \quad \frac{\partial E}{\partial \boldsymbol{b}(2)} = \boldsymbol{D}(2)$$

Backpropagation of error gradient from output to hidden layer:

$$\boldsymbol{D}(\boldsymbol{2}) = \boldsymbol{W}(3)^{\mathrm{T}}\boldsymbol{D}(3)$$