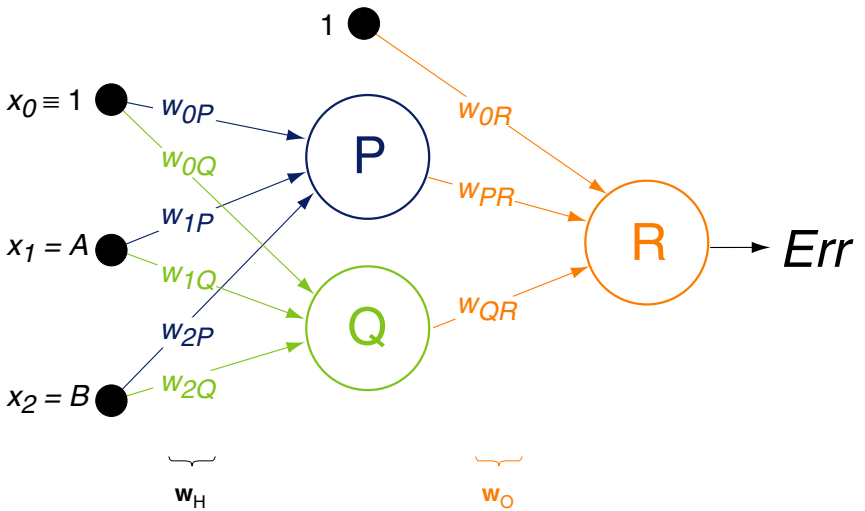
Chapter ML-LAB:VI

Some hints to help with the Multilayer Perceptron implementation

XOR Network



$$
\begin{aligned}
y_R(\mathbf{w}, \mathbf{x}) &= y_R(\mathbf{w}_O, \mathbf{y}_H(\mathbf{w}_H, \mathbf{x})) = \sigma\left(w_{0R} + w_{PR} \cdot y_P(\mathbf{w}_H, \mathbf{x}) + w_{QR} \cdot y_Q(\mathbf{w}_H, \mathbf{x})\right) \\
&= \sigma\left(w_{0R} + w_{PR} \cdot \sigma\left(w_{0P} + w_{1P}x_1 + w_{2P}x_2\right) + w_{QR} \cdot \sigma\left(w_{0Q} + w_{1Q}x_1 + w_{2Q}x_2\right)\right)
\end{aligned}
$$

$$
Err(\mathbf{w}) = \frac{1}{2} \sum_{(\mathbf{x}, \mathbf{c}(\mathbf{x})) \in D} (c(\mathbf{x}) - y_R(\mathbf{x}))^2
$$

Backpropagation

Idea

- Find out how each parameter in $\mathbf{w}$ contributes to the error *Err*
- Adjust $\mathbf{w}$ in the direction where *Err* decreases the most: $-\frac{\partial \mathit{Err}}{\partial \mathbf{w}}$
- Weight update by layer: $\mathbf{w}_O^{(t+1)} := \mathbf{w}_O^{(t)} - \eta \cdot \frac{\partial \mathit{Err}}{\partial \mathbf{w}_O}$ (η is the learning rate)

  and $\mathbf{w}_H^{(t+1)} := \mathbf{w}_H^{(t)} - \eta \cdot \frac{\partial \mathit{Err}}{\partial \mathbf{w}_H}$

Backpropagation

Idea

- Find out how each parameter in $\mathbf{w}$ contributes to the error *Err*
- Adjust $\mathbf{w}$ in the direction where *Err* decreases the most: $-\frac{\partial Err}{\partial \mathbf{w}}$
- Weight update by layer: $\mathbf{w}_O^{(t+1)} := \mathbf{w}_O^{(t)} - \eta \cdot \frac{\partial Err}{\partial \mathbf{w}_O}$ (η is the learning rate)

  and $\mathbf{w}_H^{(t+1)} := \mathbf{w}_H^{(t)} - \eta \cdot \frac{\partial Err}{\partial \mathbf{w}_H}$

Write the error as a nested function:

$$Err(\mathbf{w}, \mathbf{x}) = Err(y_R(\mathbf{w}, \mathbf{x})) = Err(y_R(\mathbf{w}_O, \mathbf{y}_H(\mathbf{w}_H, \mathbf{x})))$$

where $\mathbf{y}_H := \begin{bmatrix} y_P(\mathbf{x}, \mathbf{w}_H) \\ y_Q(\mathbf{x}, \mathbf{w}_H) \end{bmatrix}$

Backpropagation

Idea

- Find out how each parameter in $\mathbf{w}$ contributes to the error *Err*
- Adjust $\mathbf{w}$ in the direction where *Err* decreases the most: $-\frac{\partial Err}{\partial \mathbf{w}}$
- Weight update by layer: $\mathbf{w}_O^{(t+1)} := \mathbf{w}_O^{(t)} - \eta \cdot \frac{\partial Err}{\partial \mathbf{w}_O}$ (η is the learning rate)

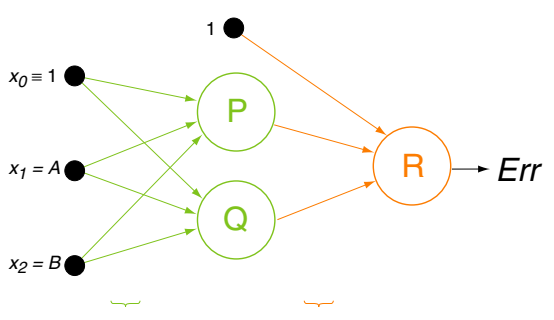
  and $\mathbf{w}_H^{(t+1)} := \mathbf{w}_H^{(t)} - \eta \cdot \frac{\partial Err}{\partial \mathbf{w}_H}$

Write the error as a nested function:

$$Err(\mathbf{w}, \mathbf{x}) = Err(y_R(\mathbf{w}, \mathbf{x})) = Err(y_R(\mathbf{w}_O, \mathbf{y}_H(\mathbf{w}_H, \mathbf{x})))$$

where $\mathbf{y}_H := \begin{bmatrix} y_P(\mathbf{x}, \mathbf{w}_H) \\ y_Q(\mathbf{x}, \mathbf{w}_H) \end{bmatrix}$

Decompose the gradient with respect to each layer of weights (using the chain rule):

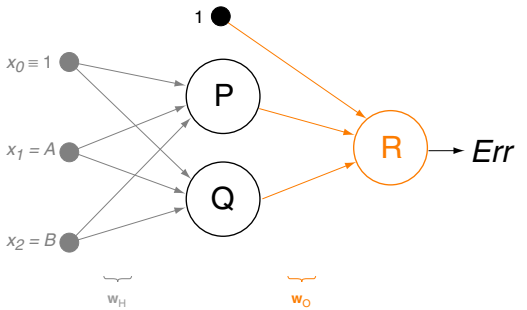


$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

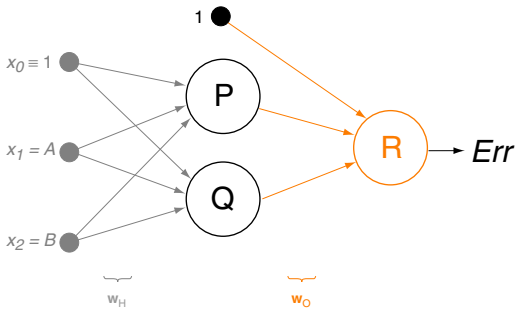
and

$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

Gradient for Output Layer Weights

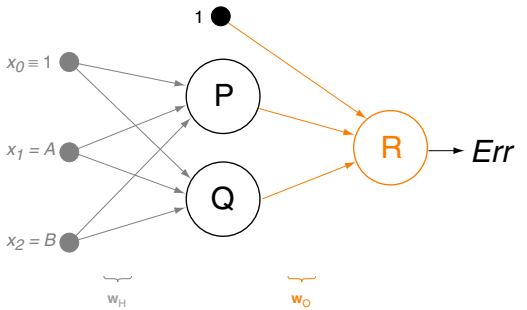


Gradient for Output Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

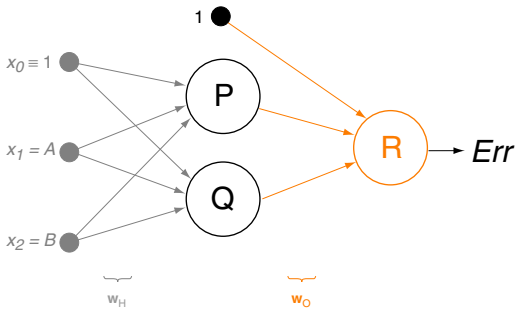
Gradient for Output Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

$$\frac{\partial Err}{\partial y_R} = \frac{\partial}{\partial y_R} \frac{1}{2} \left(c(x) - y_R\right)^2 = -1 \cdot \left(c(x) - y_R\right)$$

Gradient for Output Layer Weights



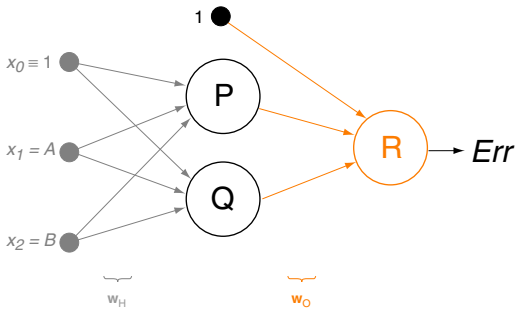
$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

$$\frac{\partial Err}{\partial y_R} = \frac{\partial}{\partial y_R} \frac{1}{2} \left(c(x) - y_R\right)^2 = -1 \cdot \left(c(x) - y_R\right)$$

$$\frac{\partial y_R}{\partial \mathbf{w}_O} = \frac{\partial}{\partial \mathbf{w}_O} \sigma \left(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H\right)$$

where $\widetilde{\mathbf{y}}_H := \begin{bmatrix} 1 \\ \mathbf{y}_H \end{bmatrix}$

Gradient for Output Layer Weights



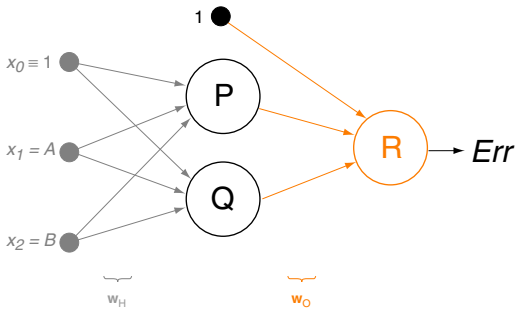
$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

$$\frac{\partial Err}{\partial y_R} = \frac{\partial}{\partial y_R} \frac{1}{2} \left(c(x) - y_R\right)^2 = -1 \cdot \left(c(x) - y_R\right)$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{w}_O} &= \frac{\partial}{\partial \mathbf{w}_O} \sigma\left(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H\right) \\ &= \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \quad \cdot \quad \frac{\partial}{\partial \mathbf{w}_O} \mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H \end{aligned}$$

where $\widetilde{\mathbf{y}}_H := \begin{bmatrix} 1 \\ \mathbf{y}_H \end{bmatrix}$

Gradient for Output Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

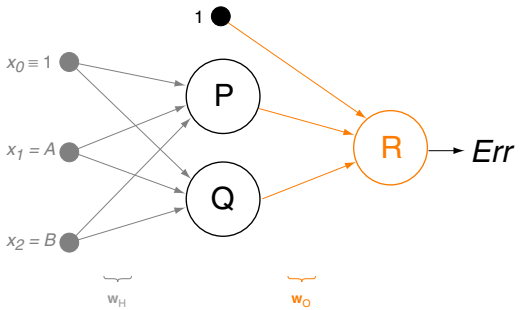
$$\frac{\partial Err}{\partial y_R} = \frac{\partial}{\partial y_R} \frac{1}{2} \left(c(x) - y_R\right)^2 = -1 \cdot \left(c(x) - y_R\right)$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{w}_O} &= \frac{\partial}{\partial \mathbf{w}_O} \sigma\left(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H\right) \\ &= \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \quad \cdot \quad \frac{\partial}{\partial \mathbf{w}_O} \mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H \\ &= \left(\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \odot (1 - \sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H))\right) \quad \cdot \quad \widetilde{\mathbf{y}}_H \end{aligned}$$

where $\widetilde{\mathbf{y}}_H := \begin{bmatrix} 1 \\ \mathbf{y}_H \end{bmatrix}$

$\odot :=$ element-wise product

Gradient for Output Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O}$$

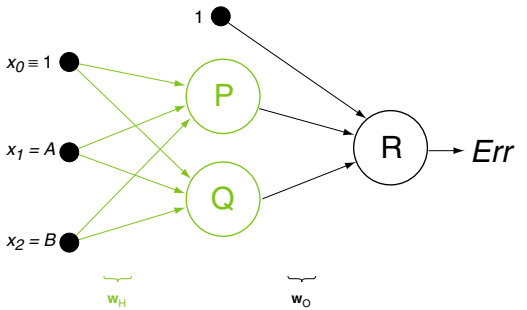
$$\frac{\partial Err}{\partial y_R} = \frac{\partial}{\partial y_R} \frac{1}{2} \left(c(x) - y_R\right)^2 = -1 \cdot \left(c(x) - y_R\right)$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{w}_O} &= \frac{\partial}{\partial \mathbf{w}_O} \sigma\left(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H\right) \\ &= \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \quad \cdot \quad \frac{\partial}{\partial \mathbf{w}_O} \mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H \\ &= \left(\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \odot (1 - \sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H))\right) \quad \cdot \quad \widetilde{\mathbf{y}}_H \\ &= \left(y_R \odot (1 - y_R)\right) \cdot \widetilde{\mathbf{y}}_H \end{aligned}$$

where $\widetilde{\mathbf{y}}_H := \begin{bmatrix} 1 \\ \mathbf{y}_H \end{bmatrix}$

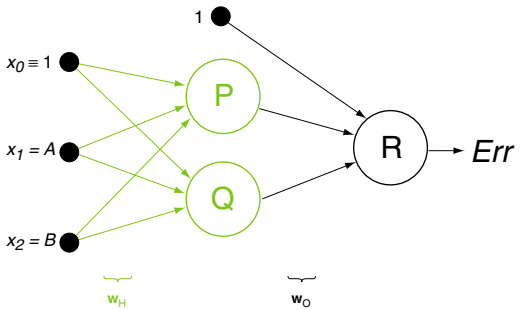
$\odot :=$ element-wise product

Gradient for Hidden Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

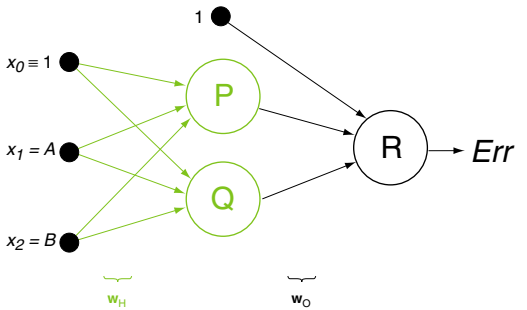
Gradient for Hidden Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\frac{\partial y_R}{\partial \mathbf{y}_H} = \frac{\partial}{\partial y_H} \sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H)$$

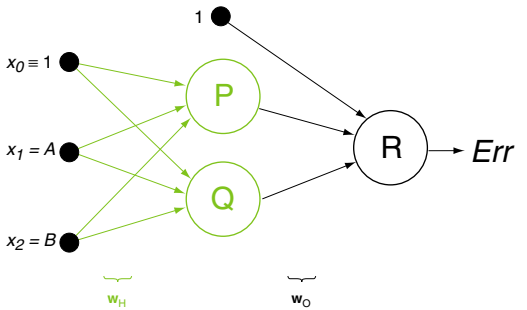
Gradient for Hidden Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\frac{\partial y_R}{\partial \mathbf{y}_H} = \frac{\partial}{\partial y_H} \sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) = \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \cdot \mathbf{w}_O$$

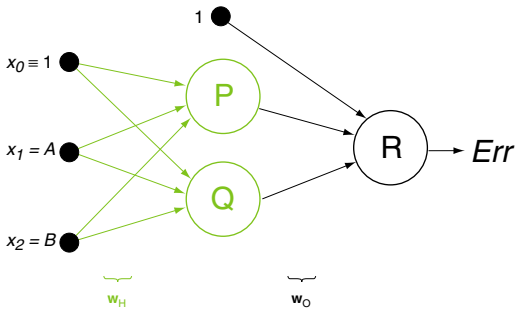
Gradient for Hidden Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{y}_H} &= \frac{\partial}{\partial y_H}\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) = \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \cdot \mathbf{w}_O \\ &= (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \end{aligned}$$

Gradient for Hidden Layer Weights

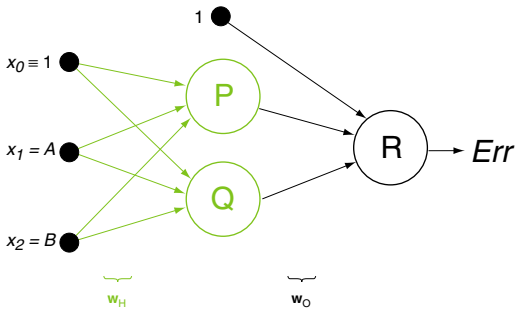


$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{y}_H} &= \frac{\partial}{\partial y_H}\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) = \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \cdot \mathbf{w}_O \\ &= (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \end{aligned}$$

$$\frac{\partial \mathbf{y}_H}{\mathbf{w}_H} = \frac{\partial}{\partial \mathbf{w}_H}\sigma(\mathbf{x} \cdot \mathbf{w}_H)$$

Gradient for Hidden Layer Weights

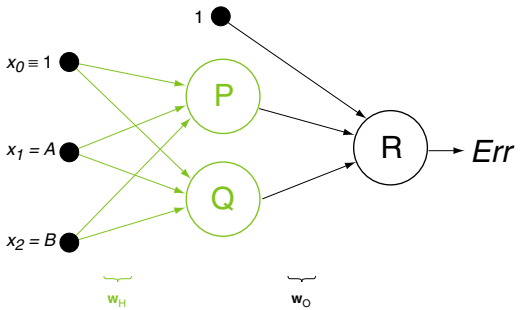


$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{y}_H} &= \frac{\partial}{\partial y_H}\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) = \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \cdot \mathbf{w}_O \\ &= (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \end{aligned}$$

$$\frac{\partial \mathbf{y}_H}{\mathbf{w}_H} = \frac{\partial}{\partial \mathbf{w}_H}\sigma(\mathbf{x} \cdot \mathbf{w}_H) = \sigma'(\mathbf{x} \cdot \mathbf{w}_H) \cdot \mathbf{x}$$

Gradient for Hidden Layer Weights



$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H}$$

$$\begin{aligned} \frac{\partial y_R}{\partial \mathbf{y}_H} &= \frac{\partial}{\partial y_H}\sigma(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) = \sigma'(\mathbf{w}_O \cdot \widetilde{\mathbf{y}}_H) \cdot \mathbf{w}_O \\ &= (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathbf{y}_H}{\mathbf{w}_H} &= \frac{\partial}{\partial \mathbf{w}_H}\sigma(\mathbf{x} \cdot \mathbf{w}_H) = \sigma'(\mathbf{x} \cdot \mathbf{w}_H) \cdot \mathbf{x} \\ &= (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} \end{aligned}$$

Deriving the Weight Update Rules

Recap:

$$\frac{\partial Err(\mathbf{w}_O)}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \widetilde{\mathbf{y}}_H$$

$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \cdot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x}$$

Deriving the Weight Update Rules

Recap:

$$\frac{\partial Err(\mathbf{w}_O)}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \widetilde{\mathbf{y}}_H = -\delta_O \cdot \widetilde{\mathbf{y}}_H$$

$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \cdot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} = -(\delta_O \cdot \mathbf{w}_O) \odot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x}$$

Deriving the Weight Update Rules

Recap:

$$\frac{\partial Err(\mathbf{w}_O)}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \widetilde{\mathbf{y}}_H = -\delta_O \cdot \widetilde{\mathbf{y}}_H$$

$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \cdot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} = -(\delta_O \cdot \mathbf{w}_O) \odot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} = -\delta_H \cdot \mathbf{x}$$

Deriving the Weight Update Rules

Recap:

$$\frac{\partial Err(\mathbf{w}_O)}{\partial \mathbf{w}_O} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{w}_O} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \widetilde{\mathbf{y}}_H = -\delta_O \cdot \widetilde{\mathbf{y}}_H$$

$$\frac{\partial Err(\mathbf{w})}{\partial \mathbf{w}_H} = \frac{\partial Err}{\partial y_R} \cdot \frac{\partial y_R}{\partial \mathbf{y}_H} \cdot \frac{\partial \mathbf{y}_H}{\partial \mathbf{w}_H} = -1 \cdot (c(x) - y_R) \odot (y_R \odot (1 - y_R)) \cdot \mathbf{w}_O \cdot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} = -(\delta_O \cdot \mathbf{w}_O) \odot (\mathbf{y}_H \odot (1 - \mathbf{y}_H)) \cdot \mathbf{x} = -\delta_H \cdot \mathbf{x}$$

In the code:

$$y_R := \text{(forward pass)}$$
$$\delta_O := (c(x) - y_R) \odot (y_R \odot (1 - y_R))$$
$$\delta_H := (\delta_O \cdot \mathbf{w}_O) \odot (\mathbf{y}_H \odot (1 - \mathbf{y}_H))$$

$$\mathbf{w}_O^{(t+1)} := \mathbf{w}_O^{(t)} + \eta \cdot \delta_O \cdot \widetilde{\mathbf{y}}_H$$
$$\mathbf{w}_H^{(t+1)} := \mathbf{w}_H^{(t)} + \eta \cdot \delta_H \cdot \mathbf{x}$$