- What is the gradient computed with respect to?
  - Weights - $m$ at hidden nodes and $k$ at output nodes
  - $\mathbf{w}_j$ $(j = 1 \ldots m)$
  - $\mathbf{w}_l$ $(l = 1 \ldots k)$
- $\mathbf{w}_j \leftarrow \mathbf{w}_j - \eta \frac{\partial J}{\partial \mathbf{w}_j} = \mathbf{w}_j - \eta \sum_{i=1}^{N} \frac{\partial J_i}{\partial \mathbf{w}_j}$
- $\mathbf{w}_l \leftarrow \mathbf{w}_l - \eta \frac{\partial J}{\partial \mathbf{w}_l} = \mathbf{w}_l - \eta \sum_{i=1}^{N} \frac{\partial J}{\partial \mathbf{w}_l}$

$$\nabla J_i = \begin{bmatrix} \frac{\partial J_i}{\partial \mathbf{w}_1} \\ \frac{\partial J_i}{\partial \mathbf{w}_2} \\ \vdots \\ \frac{\partial J_i}{\partial \mathbf{w}_{m+k}} \end{bmatrix}$$

$$\frac{\partial J_i}{\partial \mathbf{w}_r} = \begin{bmatrix} \frac{\partial J_i}{\partial w_{r1}} \\ \frac{\partial J_i}{\partial w_{r2}} \\ \vdots \end{bmatrix}$$

- Need to compute $\frac{\partial J_i}{\partial w_{rq}}$
- Update rule for the $q^{th}$ entry in the $r^{th}$ weight vector:

$$w_{rq} \leftarrow w_{rq} - \eta \frac{\partial J}{\partial w_{rq}} = w_{rq} - \eta \sum_{i=1}^{N} \frac{\partial J_i}{\partial w_{rq}}$$

## 4.1 Derivation of the Backpropagation Rules

Assume that we only one training example, i.e., $i = 1$, $J = J_i$. Dropping the subscript $i$ from here onwards.

- Consider any weight $w_{rq}$
- Let $u_{rq}$ be the $q^{th}$ element of the input vector coming in to the $r^{th}$ unit.

**Observation 1**

Weight $w_{rq}$ is connected to $J$ through $net_r = \sum_i w_{rq} u_{rq}$.

$$\frac{\partial J}{\partial w_{rq}} = \frac{\partial J}{\partial net_r}\frac{\partial net_r}{\partial w_{rq}} = \frac{\partial J}{\partial net_r} u_{rq}$$

**Observation 2**

$net_l$ for an **output node** is connected to $J$ only through the output value of the node (or $o_l$)

$$\frac{\partial J}{\partial net_l} = \frac{\partial J}{\partial o_l}\frac{\partial o_l}{\partial net_l}$$

The first term above can be computed as:

$$\frac{\partial J}{\partial o_l} = \frac{\partial}{\partial o_l}\frac{1}{2}\sum_{l=1}^{k}(y_l - o_l)^2$$

The entries in the summation in the right hand side will be non zero only for $l$. This results in:

$$\begin{aligned}\frac{\partial J}{\partial o_l} &= \frac{\partial}{\partial o_l}\frac{1}{2}(y_l - o_l)^2 \\ &= -(y_l - o_l)\end{aligned}$$

Moreover, the second term in the chain rule above can be computed as:

$$\begin{aligned}\frac{\partial o_l}{\partial net_l} &= \frac{\partial \sigma(net_l)}{\partial net_l} \\ &= o_l(1 - o_l)\end{aligned}$$

The last result arises from the fact $o_l$ is a sigmoid function. Using the above results, one can compute the following.

$$\frac{\partial J}{\partial net_l} = -(y_l - o_l)o_l(1 - o_l)$$

Let

$$\delta_l = (y_l - o_l)o_l(1 - o_l)$$

Therefore,

$$\frac{\partial J}{\partial net_l} = -\delta_l$$