4.8)

To change the output function from the sigmoid to tanh, we need only make a couple changes.

The first is to update (T4.3) we change from:

$$o_k = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}$$

$$\delta_k = -\frac{\partial E}{\partial net_n} = o_k(1 - o_k)(t_k - o_k)$$

$$\delta_k = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}} * \left(1 - \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}\right) * \left(t_k - \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x})}}\right)$$

To

$$o_k = \tanh(\vec{w} \cdot \vec{x})$$

$$\delta_k = -\frac{\partial E}{\partial net_n} = -(1 - {o_k}^2)(t_k - o_k)$$

$$\delta_k = -(1 - \tanh^2(\vec{w} \cdot \vec{x}))(t_k - \tanh(\vec{w} \cdot \vec{x}))$$

We must also update (T4.4) in a similar fashion to:

$$\delta_h = (1 - \tanh^2(\vec{w} \cdot \vec{x})) * \sum_{k \in outputs} w_{kh}\delta_k$$

4.10)

$$E(\vec{w}) = \frac{1}{2}\sum_{d\in D}\sum_{k\in outputs}(t_{kd} - o_{kd})^2 + \gamma\sum_{i,j} w_{ji}^2$$

So taking the error with respect to a single document rather than the entire network:

$$E_d(\vec{w}) = \frac{1}{2}\sum_{k\in outputs}(t_k - o_k)^2 + \gamma\sum_{i,j} w_{ji}^2$$

The gradient descent update rule is:

$$w_{ji} \mathrel{+}= \Delta w_{ji}$$

$$\Delta w_{ji} = -\eta\frac{\partial E_d}{\partial w_{ji}}$$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j}x_{ji}$$

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j}\frac{\partial o_j}{\partial net_j}$$

And now we substitute in the error equation and perform the derivative:

$$\frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j}\left(\frac{1}{2}\sum_{k\in outputs}(t_k - o_k)^2 + \gamma\sum_{i,j} w_{ji}^2\right) = \frac{\partial}{\partial o_j}\frac{1}{2}\sum_{k\in outputs}(t_k - o_k)^2 + \frac{\partial}{\partial o_j}\gamma\sum_{i,j} w_{ji}^2$$

Since the first term's sum is over only the outputs, we can simply set k=q. The second term disappears because there is no o term in it:

$$-(t_q - o_q) + 0$$

And now we need to calculate $\frac{\partial o_j}{\partial net_j}$. Note that o is the sigmoid function and the second term that disappeared in the previous derivation does not disappear here. This term is actually the same as the net for the output units:

$$\frac{\partial o_j}{\partial net_j} = \frac{\partial sigmoid(net_j)}{\partial net_j} + \frac{\partial}{\partial net_j}\gamma\sum_{i,j} w_{ji}^2$$

$$\frac{\partial o_j}{\partial net_j} = o_j(1 - o_j) + 2\gamma(1)$$

Now we combine these and obtain:

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j}\frac{\partial o_j}{\partial net_j} = -\left(t_j - o_j\right) * \left(o_j\left(1 - o_j\right) + 2\gamma\right)$$

So

$$\Delta w_{ji} = -\eta\frac{\partial E_d}{\partial w_{ji}} = \eta\left(t_j - o_j\right) * \left(o_j\left(1 - o_j\right) + 2\gamma\right)$$

And the gradient descent rule is:

$$w_{ji} \mathrel{+}= \eta\left(t_j - o_j\right) * \left(o_j\left(1 - o_j\right) + 2\gamma\right)$$