JMLR: Workshop and Conference Proceedings 60:1–8, 2016 ACML 2016

The Incredible Shrinking Neural Network: Pruning to Operate in Constrained Memory Environments

Appendix A. Second Derivative Back-Propagation

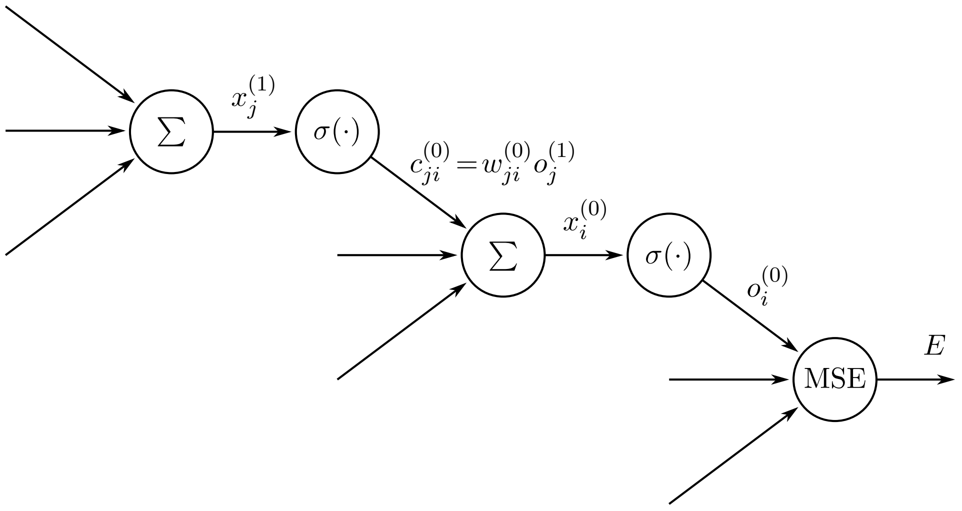

Figure 1: A computational graph of a simple feed-forward network illustrating the naming of different variables, where $\sigma(\cdot)$ is the nonlinearity, MSE is the mean-squared error cost function and E is the overall loss.

Name and network definitions:

$$E = \frac{1}{2}\sum_i (o_i^{(0)} - t_i)^2 \quad o_i^{(m)} = \sigma(x_i^{(m)}) \quad x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} \quad c_{ji}^{(m)} = w_{ji}^{(m)} o_j^{(m+1)} \tag{1}$$

Superscripts represent the index of the layer of the network in question, with 0 representing the output layer. E is the squared-error network cost function. $o_i^{(m)}$ is the ith output in layer m generated by the activation function σ, which in this paper is is the standard logistic sigmoid. $x_i^{(m)}$ is the weighted sum of inputs to the ith neuron in the mth layer, and $c_{ji}^{(m)}$ is the contribution of the jth neuron in the $m+1$ layer to the input of the ith neuron in the mth layer.

A.1. First and Second Derivatives

The first and second derivatives of the cost function with respect to the outputs:

$$\frac{\partial E}{\partial o_i^{(0)}} = o_i^{(0)} - t_i \tag{2}$$

$$\frac{\partial^2 E}{\partial {o_i^{(0)}}^2} = 1 \tag{3}$$

The first and second derivatives of the sigmoid function in forms depending only on the output:

$$\sigma'(x) = \sigma(x)\,(1 - \sigma(x)) \tag{4}$$
$$\sigma''(x) = \sigma'(x)\,(1 - 2\sigma(x)) \tag{5}$$

The second derivative of the sigmoid is easily derived from the first derivative:

$$\sigma'(x) = \sigma(x)\,(1 - \sigma(x)) \tag{6}$$
$$\sigma''(x) = \frac{\mathrm{d}}{\mathrm{d}x}\underbrace{\sigma(x)}_{f(x)}\underbrace{(1 - \sigma(x))}_{g(x)} \tag{7}$$
$$\sigma''(x) = f'(x)g(x) + f(x)g'(x) \tag{8}$$
$$\sigma''(x) = \sigma'(x)(1 - \sigma(x)) - \sigma(x)\sigma'(x) \tag{9}$$
$$\sigma''(x) = \sigma'(x) - 2\sigma(x)\sigma'(x) \tag{10}$$
$$\sigma''(x) = \sigma'(x)(1 - 2\sigma(x)) \tag{11}$$

And for future convenience:

$$\frac{\mathrm{d}o_i^{(m)}}{\mathrm{d}x_i^{(m)}} = \frac{\mathrm{d}}{\mathrm{d}x_i^{(m)}}\left(o_i^{(m)} = \sigma(x_i^{(m)})\right) \tag{12}$$
$$= \left(o_i^{(m)}\right)\left(1 - o_i^{(m)}\right) \tag{13}$$
$$= \sigma'\left(x_i^{(m)}\right) \tag{14}$$
$$\frac{\mathrm{d}^2 o_i^{(m)}}{\mathrm{d}{x_i^{(m)}}^2} = \frac{\mathrm{d}}{\mathrm{d}x_i^{(m)}}\left(\frac{\mathrm{d}o_i^{(m)}}{\mathrm{d}x_i^{(m)}} = \left(o_i^{(m)}\right)\left(1 - o_i^{(m)}\right)\right) \tag{15}$$
$$= \left(o_i^{(m)}\left(1 - o_i^{(m)}\right)\right)\left(1 - 2o_i^{(m)}\right) \tag{16}$$
$$= \sigma''\left(x_i^{(m)}\right) \tag{17}$$

Derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial x_i^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \tag{18}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (2)}} \underbrace{\sigma\left(x_i^{(0)}\right)\left(1 - \sigma\left(x_i^{(0)}\right)\right)}_{\text{from (4)}} \tag{19}$$

$$= \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right) \tag{20}$$

$$= \left(o_i^{(0)} - t_i\right)\sigma'\left(x_i^{(0)}\right) \tag{21}$$

Second derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \frac{\partial}{\partial x_i^{(0)}}\left(\frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}}\right) \tag{22}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} \partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} + \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial^2 o_i^{(0)}}{\partial {x_i^{(0)}}^2} \tag{23}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} \partial o_i^{(0)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (4)}} + \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (2)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)\left(1 - 2o_i^{(0)}\right)}_{\text{from (5)}} \tag{24}$$

$$\left(\frac{\partial^2 E}{\partial x_i^{(0)} \partial o_i^{(0)}}\right) = \frac{\partial}{\partial x_i^{(0)}} \frac{\partial E}{\partial o_i^{(0)}} = \frac{\partial}{\partial x_i^{(0)}} \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (2)}} = \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} = \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (4)}} \tag{25}$$

$$\frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)\left(1 - 2o_i^{(0)}\right) \tag{26}$$

$$= \left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right)\sigma''\left(x_i^{(0)}\right) \tag{27}$$

First derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$ from neuron j to neuron i with weight $w_{ji}^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{28}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (2)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (4)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{29}$$

$$\left(\frac{\partial x_i^{(m)}}{\partial c_{ji}^{(m)}}\right) = \frac{\partial}{\partial c_{ji}^{(m)}} \left(x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) = \frac{\partial}{\partial c_{ji}^{(m)}} \left(c_{ji}^{(m)} + k\right) = 1 \tag{30}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right) \tag{31}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right)}_{\text{from (21)}} \tag{32}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial x_i^{(0)}} \tag{33}$$

Second derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$:

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial}{\partial c_{ji}^{(0)}} \left(\frac{\partial E}{\partial c_{ji}^{(0)}} = \underbrace{\left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right)}_{\text{from (32)}}\right) \tag{34}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(x_i^{(0)}\right) - t_i\right) \sigma'\left(x_i^{(0)}\right) \tag{35}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(\sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) - t_i\right) \sigma'\left(\sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) \tag{36}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(\sum_j c_{ji}^{(0)}\right) - t_i\right) \sigma'\left(\sum_j c_{ji}^{(0)}\right) \tag{37}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \underbrace{\left(\sigma\left(c_{ji}^{(0)} + k\right) - t_i\right)}_{f\left(c_{ji}^{(0)}\right)} \underbrace{\sigma'\left(c_{ji}^{(0)} + k\right)}_{g\left(c_{ji}^{(0)}\right)} \tag{38}$$

We now make use of the abbreviations f and g:

$$= f'\left(c_{ji}^{(0)}\right) g\left(c_{ji}^{(0)}\right) + f\left(c_{ji}^{(0)}\right) g'\left(c_{ji}^{(0)}\right) \tag{39}$$

$$= \sigma'\left(c_{ji}^{(0)} + k\right) \sigma'\left(c_{ji}^{(0)} + k\right) + \left(\sigma\left(c_{ji}^{(0)} + k\right) - t_i\right) \sigma''\left(c_{ji}^{(0)} + k\right) \tag{40}$$

$$= \sigma'\left(c_{ji}^{(0)} + k\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(c_{ji}^{(0)} + k\right) \tag{41}$$

$$\left(c_{ji}^{(0)} + k = \sum_j c_{ji}^{(0)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} = x_i^{(0)}\right) \tag{42}$$

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \underbrace{\left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right)}_{\text{from } (27)} \tag{43}$$

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \tag{44}$$

A.1.1. Summary Of Output Layer Derivatives

$$\frac{\partial E}{\partial o_i^{(0)}} = o_i^{(0)} - t_i \qquad \frac{\partial^2 E}{\partial {o_i^{(0)}}^2} = 1 \tag{45}$$

$$\frac{\partial E}{\partial x_i^{(0)}} = \left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right) \qquad \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right) \tag{46}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial x_i^{(0)}} \qquad \frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \tag{47}$$

A.1.2. Hidden Layer Derivatives

The first derivative of the error with respect to a neuron with output $o_j^{(1)}$ in the first hidden layer, summing over all partial derivative contributions from the output layer:

$$\frac{\partial E}{\partial o_j^{(1)}} = \sum_i \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \frac{\partial c_{ji}^{(0)}}{\partial o_j^{(1)}} = \sum_i \underbrace{\left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right)}_{\text{from } (21)} w_{ji}^{(0)} \tag{48}$$

$$\frac{\partial c_{ji}^{(m)}}{\partial o_j^{(m+1)}} = \frac{\partial}{\partial o_j^{(m+1)}} \left(c_{ji}^{(m)} = w_{ji}^{(m)} o_j^{(m+1)}\right) = w_{ji}^{(m)} \tag{49}$$

$$\frac{\partial E}{\partial o_j^{(1)}} = \sum_i \frac{\partial E}{\partial x_i^{(0)}} w_{ji}^{(0)} \tag{50}$$

Note that this equation does not depend on the specific form of $\frac{\partial E}{\partial x_i^{(0)}}$, whether it involves a sigmoid or any other activation function. We can therefore replace the specific indexes with general ones, and use this equation in the future.

$$\frac{\partial E}{\partial o_j^{(m+1)}} = \sum_i \frac{\partial E}{\partial x_i^{(m)}} w_{ji}^{(m)} \tag{51}$$

The second derivative of the error with respect to a neuron with output $o_j^{(1)}$ in the first hidden layer:

$$\begin{align}
\frac{\partial^2 E}{\partial {o_j^{(1)}}^2} &= \frac{\partial}{\partial o_j^{(1)}} \frac{\partial E}{\partial o_j^{(1)}} \tag{52}\\
&= \frac{\partial}{\partial o_j^{(1)}} \sum_i \frac{\partial E}{\partial x_i^{(0)}} w_{ji}^{(0)} \tag{53}\\
&= \frac{\partial}{\partial o_j^{(1)}} \sum_i \left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right) w_{ji}^{(0)} \tag{54}
\end{align}$$

If we now make use of the fact, that $o_i^{(0)} = \sigma\left(x_i^{(0)}\right) = \sigma\left(\sum_j \left(w_{ji}^{(0)} o_j^{(1)}\right)\right)$, we can evaluate the expression further.

$$\begin{align}
\frac{\partial^2 E}{\partial {o_j^{(1)}}^2} &= \frac{\partial}{\partial o_j^{(1)}} \sum_i \underbrace{\left(\sigma\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) - t_i\right)}_{f\left(o_j^{(1)}\right)} \underbrace{\sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) w_{ji}^{(0)}}_{g\left(o_j^{(1)}\right)} \tag{55}\\
&= \sum_i \left(f'\left(o_j^{(1)}\right) g\left(o_j^{(1)}\right) + f\left(o_j^{(1)}\right) g'\left(o_j^{(1)}\right)\right) \tag{56}\\
&= \sum_i \sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) w_{ji}^{(0)}\ \sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) w_{ji}^{(0)} + \dots \tag{57}\\
&\sum_i \left(\sigma\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) - t_i\right) \sigma''\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) \left(w_{ji}^{(0)}\right)^2 \tag{58}\\
&= \sum_i \left(\left(\sigma'\left(x_i^{(0)}\right)\right)^2 \left(w_{ji}^{(0)}\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right) \left(w_{ji}^{(0)}\right)^2\right) \tag{59}\\
&= \sum_i \underbrace{\left(\left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right)\right)}_{\text{from } (27)} \left(w_{ji}^{(0)}\right)^2 \tag{60}
\end{align}$$

Summing up, we obtain the more general expression:

$$\frac{\partial^2 E}{\partial {o_j^{(1)}}^2} = \sum_i \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \left(w_{ji}^{(0)}\right)^2 \tag{61}$$

Note that the equation in (61) does not depend on the form of $\frac{\partial^2 E}{\partial {x_x^{(0)}}^2}$, which means we can replace the specific indexes with general ones:

$$\frac{\partial^2 E}{\partial {o_j^{(m+1)}}^2} = \sum_i \frac{\partial^2 E}{\partial {x_i^{(m)}}^2} \left(w_{ji}^{(m)}\right)^2 \tag{62}$$

At this point we are beginning to see the recursion in the form of the 2nd derivative terms which can be thought of analogously to the first derivative recursion which is central to the back-propagation algorithm. The formulation above which makes specific reference to layer indexes also works in the general case.

Consider the ith neuron in any layer m with output $o_i^{(m)}$ and input $x_i^{(m)}$. The first and second derivatives of the error E with respect to this neuron's *input* are:

$$\frac{\partial E}{\partial x_i^{(m)}} = \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \tag{63}$$

$$\begin{aligned}
\frac{\partial^2 E}{\partial {x_i^{(m)}}^2} &= \frac{\partial}{\partial x_i^{(m)}} \frac{\partial E}{\partial x_i^{(m)}} && (64)\\
&= \frac{\partial}{\partial x_i^{(m)}} \left(\frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \right) && (65)\\
&= \frac{\partial^2 E}{\partial x_i^{(m)} \partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} + \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial^2 o_i^{(m)}}{\partial {x_i^{(m)}}^2} && (66)\\
&= \frac{\partial}{\partial o_i^{(m)}} \left(\frac{\partial E}{\partial x_i^{(m)}} = \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \right) \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) && (67)\\
&= \frac{\partial^2 E}{\partial {o_i^{(m)}}^2} \left(\frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \right) + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) && (68)\\
\frac{\partial^2 E}{\partial {x_i^{(m)}}^2} &= \frac{\partial^2 E}{\partial {o_i^{(m)}}^2} \left(\sigma' \left(x_i^{(m)} \right) \right)^2 + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) && (69)
\end{aligned}$$

Note the form of this equation is the general form of what was derived for the output layer in (27). Both of the above first and second terms are easily computable and can be stored as we propagate back from the output of the network to the input. With respect to the output layer, the first and second derivative terms have already been derived above. In the case of the $m+1$ hidden layer during back propagation, there is a summation of terms calculated in the mth layer. For the first derivative, we have this from (51).

$$\frac{\partial E}{\partial o_j^{(m+1)}} = \sum_i \frac{\partial E}{\partial x_i^{(m)}} w_{ji}^{(m)} \tag{70}$$

And the second derivative for the jth neuron in the $m+1$ layer:

$$\frac{\partial^2 E}{\partial {x_j^{(m+1)}}^2} = \frac{\partial^2 E}{\partial {o_j^{(m+1)}}^2} \left(\sigma' \left(x_j^{(m+1)} \right) \right)^2 + \frac{\partial E}{\partial o_j^{(m+1)}} \sigma'' \left(x_j^{(m+1)} \right) \tag{71}$$

We can replace both derivative terms with the forms which depend on the previous layer:

$$\frac{\partial^2 E}{\partial {x_j^{(m+1)}}^2} = \underbrace{\sum_i \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \left(w_{ji}^{(0)}\right)^2}_{\text{from (62)}} \left(\sigma'\left(x_j^{(m+1)}\right)\right)^2 + \underbrace{\sum_i \frac{\partial E}{\partial x_i^{(m)}} w_{ji}^{(m)}}_{\text{from (51)}} \sigma''\left(x_j^{(m+1)}\right) \tag{72}$$

And this horrible mouthful of an equation gives you a general form for any neuron in the jth position of the $m+1$ layer. Taking very careful note of the indexes, this can be more or less translated painlessly to code. You are welcome, world.

A.1.3. Summary Of Hidden Layer Derivatives

$$\frac{\partial E}{\partial o_j^{(m+1)}} = \sum_i \frac{\partial E}{\partial x_i^{(m)}} w_{ji}^{(m)} \qquad \frac{\partial^2 E}{\partial {o_j^{(m+1)}}^2} = \sum_i \frac{\partial^2 E}{\partial {x_i^{(m)}}^2} \left(w_{ji}^{(m)}\right)^2 \tag{73}$$

$$\frac{\partial E}{\partial x_i^{(m)}} = \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \tag{74}$$

$$\frac{\partial^2 E}{\partial {x_j^{(m+1)}}^2} = \frac{\partial^2 E}{\partial {o_j^{(m+1)}}^2} \left(\sigma'\left(x_j^{(m+1)}\right)\right)^2 + \frac{\partial E}{\partial o_j^{(m+1)}} \sigma''\left(x_j^{(m+1)}\right) \tag{75}$$