Newton Method Computation

Training Sample:
$x = 2$
$y = 200$

Parameters:
$w_1 = 3$
$w_2 = 5$
$w_3 = 7$

All activation functions are linear:
$\phi_k(x) = x$

Squared error is used as a loss function:
$L(y, \hat{y}) = (y - \hat{y})^2$

Forward Propagation

First, forward values are computed, storing intermediate values for use during backpropagation.

$a_1 = xw_1 = 6$
$h_1 = \phi_1(a_1) = 3$
$a_2 = h_1 w_2 = 30$
$h_2 = \phi_2(a_2) = 30$
$a_3 = h_2 w_3 = 210$
$h_3 = \phi_3(a_3) = 210$
$L = (y - h_3)^2 = 100$

Backpropagation

Next, first- and second-order derivatives are computed from the loss function backwards, starting with the seed the value of 1. All first and second order derivatives are stored for use in future computations.

Layer 3

$$\frac{\partial L}{\partial L} = 1, \frac{\partial^2 L}{\partial L^2} = 1$$

$$\frac{\partial L}{\partial h_3} = 2(h_3 - y) = 20$$

- Look-up the derivative of the loss function and evaluate at the point $h_3 = 210$.

$$\frac{\partial^2 L}{\partial h_3^2} = 2$$

- Look-up second-order derivative of the loss function and evaluate at the point $h_3 = 210$. Here, the second order derivative is constant, because the loss function is quadratic.

$$\frac{\partial L}{\partial a_3} = \frac{\partial h_3}{\partial a_3} \frac{\partial L}{\partial h_3} = 1 \cdot 20 = 20$$

- Compute the local derivative $\frac{\partial h_3}{\partial a_3}$ by looking up the derivative of ϕ_3 and evaluating it at the point a_3.
- Multiply by the previously computed value $\frac{\partial L}{\partial h_3}$.

$$\frac{\partial^2 L}{\partial a_3^2} = \left(\frac{\partial h_3}{\partial a_3}\right)^2 \frac{\partial^2 L}{\partial h_3^2} + \frac{\partial^2 h_3}{\partial a_3^2} \frac{\partial L}{\partial h_3} = 1^2 \cdot 2 = 2$$

- We just computed the local derivative $\frac{\partial h_3}{\partial a_3}$, multiply the same value by the previously computed value $\frac{\partial^2 L}{\partial h_3^2}$.

- The term $\frac{\partial^2 h_3}{\partial a_3^2}$ is the second-order local derivative of the activation function ϕ_k. In this example, ϕ_k is linear, so the second-order derivative is always 0. Future expressions will omit this term.

$$\frac{\partial L}{\partial w_3} = \frac{\partial a_3}{\partial w_3}\frac{\partial L}{\partial a_3} = 30 \cdot 20 = 600$$

- Compute the local derivative $\frac{\partial a_3}{\partial w_3}$ by looking up the forward propagated value h_2. $a_3 = w_3 h_2$ so the local first order derivative is just h_2.
- Multiply by the previously computed value $\frac{\partial L}{\partial a_3}$.

$$\frac{\partial^2 L}{\partial w_3^2} = \left(\frac{\partial a_3}{\partial w_3}\right)^2 \frac{\partial^2 L}{\partial a_3^2} + \frac{\partial^2 a_3}{\partial w_3^2}\frac{\partial L}{\partial a_3} = h_2^2 \frac{\partial^2 L}{\partial a_3^2} = 30^2 \cdot 2 = 1800$$

- $\frac{\partial^2 a_3}{\partial w_3^2}$ is the local second order derivative of the linear function $w_3 h_2$, which is always 0.
- We just computed $\frac{\partial a_3}{\partial w_3}$, multiply this by the previously computed $\frac{\partial^2 L}{\partial a_3^2}$.

Layer 2

$$\frac{\partial L}{\partial h_2} = \frac{\partial a_3}{\partial h_2}\frac{\partial L}{\partial a_3} = 7 \cdot 20 = 140$$

- The local derivative $\frac{\partial a_3}{\partial h_2}$ is the weight of the edge, w_3.
- As above, compute the local derivative, multiply by the accumulated derivative.

$$\frac{\partial^2 L}{\partial h_2^2} = \left(\frac{\partial a_3}{\partial h_2}\right)^2 \frac{\partial^2 L}{\partial a_3^2} + \frac{\partial^2 a_3}{\partial h_2^2}\frac{\partial L}{\partial a_3} = 7^2 \cdot 2 = 98$$

- The local derivative $\frac{\partial a_3}{\partial h_2}$ was just computed, multiply it by the accumulated second-order derivative.
- The second order local derivative $\frac{\partial^2 a_3}{\partial h_2^2}$ is always zero, as we are crossing a linear function.

$$\frac{\partial L}{\partial a_2} = \frac{\partial h_2}{\partial a_2}\frac{\partial L}{\partial h_2} = 1 \cdot 140 = 140$$

- Zero term is omitted as explained above.

$$\frac{\partial^2 L}{\partial a_2^2} = \left(\frac{\partial h_2}{\partial a_2}\right)^2 \frac{\partial^2 L}{\partial h_2^2} = 1^2 \cdot 98 = 98$$

- Zero term also omitted as explained above.

$$\frac{\partial L}{\partial w_2} = \frac{\partial a_2}{\partial w_2}\frac{\partial L}{\partial a_2} = 6 \cdot 140 = 840$$

- Same pattern as above, multiply the local derivative by the previously computed value.

$$\frac{\partial^2 L}{\partial w_2^2} = \left(\frac{\partial a_2}{\partial w_2}\right)^2 \frac{\partial^2 L}{\partial a_2^2} = 6^2 \cdot 98 = 3528$$

- Zero term also omitted as explained above.
- Reusing already computed local derivative and previously computed second-order derivative.

$$\frac{\partial^2 L}{\partial w_3 \partial w_2} = \frac{\partial h_2}{\partial w_2}\frac{\partial L}{\partial a_3} + \frac{\partial a_3}{\partial w_3}\frac{\partial a_3}{\partial w_2}\frac{\partial^2 L}{\partial a_3^2} = 6 \cdot 20 + 30 \cdot 42 \cdot 2 = 2640$$

- The first "bridge" term $\frac{\partial h_2}{\partial w_2} = \frac{\partial a_2}{\partial w_2}\frac{\partial h_2}{\partial a_2}$ is a product of local derivatives we previously computed.
- $\frac{\partial a_3}{\partial w_3} = h_2$ is looked up from previously computed forward values.
- $\frac{\partial a_3}{\partial w_2} = \frac{\partial a_2}{\partial w_2}\frac{\partial h_2}{\partial a_2}\frac{\partial a_3}{\partial h_2}$ is product of previously computed values, a "bridge" term between the positions of w_1 and w_2.

- $\frac{\partial^2 L}{\partial a_3^2}$ is a previously computed second order derivative.

Layer 1

All computations follow the same pattern as above, computing additional terms and multiplying them be previously computed values.

$$\frac{\partial L}{\partial h_1} = \frac{\partial a_2}{\partial h_1}\frac{\partial L}{\partial a_2} = 5 \cdot 140 = 700$$

$$\frac{\partial^2 L}{\partial h_1^2} = \left(\frac{\partial a_2}{\partial h_1}\right)^2 \frac{\partial L}{\partial a_2^2} = 5^2 \cdot 98 = 2450$$

$$\frac{\partial L}{\partial a_1} = \frac{\partial h_1}{\partial a_1}\frac{\partial L}{\partial h_1} = 1 \cdot 700 = 700$$

$$\frac{\partial^2 L}{\partial a_1^2} = \left(\frac{\partial h_1}{\partial a_1}\right)^2 \frac{\partial^2 L}{\partial h_1^2} = 1^2 \cdot 2450 = 2450$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial a_1}{\partial w_1}\frac{\partial L}{\partial a_1} = 2 \cdot 700 = 1400$$

$$\frac{\partial^2 L}{\partial w_1^2} = \left(\frac{\partial a_1}{\partial w_1}\right)^2 \frac{\partial^2 L}{\partial a_1^2} = 2^2 \cdot 2450 = 9800$$

$$\frac{\partial^2 L}{\partial w_2 \partial w_1} = \frac{\partial h_1}{\partial w_1}\frac{\partial L}{\partial a_2} + \frac{\partial a_2}{\partial w_2}\frac{\partial a_2}{\partial w_1}\frac{\partial^2 L}{\partial a_2^2} = 2 \cdot 140 + 6 \cdot 10 \cdot 98 = 6160$$

$$\frac{\partial^2 L}{\partial w_3 \partial w_1} = \frac{\partial h_2}{\partial w_1}\frac{\partial L}{\partial a_3} + \frac{\partial a_3}{\partial w_3}\frac{\partial a_3}{\partial w_1}\frac{\partial^2 L}{\partial a_3^2} = 10 \cdot 20 + 30 \cdot 70 \cdot 2 = 4400$$

- Here the bridge terms $\frac{\partial h_2}{\partial w_1}$ and $\frac{\partial a_3}{\partial w_1}$ are longer because they span from the first layer to the third, but still consist of previously computed values.

Output

We have now computed the lower triangle of the Hessian matrix and may fill in the remaining entries by symmetry.

$$H_L = \begin{bmatrix} \frac{\partial L^2}{\partial w_1^2} & \frac{\partial L^2}{\partial w_1 \partial w_2} & \frac{\partial L^2}{\partial w_1 \partial w_3} \\ \frac{\partial L^2}{\partial w_2 \partial w_1} & \frac{\partial L^2}{\partial w_2^2} & \frac{\partial L^2}{\partial w_2 \partial w_3} \\ \frac{\partial L^2}{\partial w_3 \partial w_1} & \frac{\partial L^2}{\partial w_3 \partial w_2} & \frac{\partial L^2}{\partial w_3^2} \end{bmatrix} = \begin{bmatrix} 9800 & 6160 & 4400 \\ 6160 & 3528 & 2640 \\ 4400 & 2640 & 1800 \end{bmatrix}$$

The algorithm returns the first-order derivatives and the Hessian matrix of second-order derivatives for use in gradient descent.