Solution 1: Forward and Backward Propagation

Given the network architecture and input ($x_1^{(0)} = 2, y = 1$) the squared loss function e, and the weight matrices already provided in the example, we will first perform the forward propagation.

- $\mathbf{s}^{(1)} = \begin{bmatrix} 0.7 & 1 \end{bmatrix}^\top$
- $\mathbf{x}^{(1)} = \begin{bmatrix} 0.6044 & 0.7616 \end{bmatrix}^\top$ (in this solution, we exclude bias while expressing all layer activations)
- $s_1^{(2)} = -1.4804$
- $x_1^{(2)} = -0.9015$
- $s_1^{(3)} = -0.8031$
- $x_1^{(3)} = -0.6658$

Using the information above, we can now proceed to computing the backward propagation.

- $\boldsymbol{\delta}^{(3)} = 2\left(x_1^{(3)} - y\right) \tanh'(s_1^{(3)}) = \begin{bmatrix} -1.855 \end{bmatrix}^\top$
- $\boldsymbol{\delta}^{(2)} = \mathbf{W}^{(3)}\boldsymbol{\delta}^{(3)} \odot \tanh'(\mathbf{s}^{(2)}) = \begin{bmatrix} -0.69 \end{bmatrix}^\top$
- $\boldsymbol{\delta}^{(1)} = \mathbf{W}^{(2)}\boldsymbol{\delta}^{(2)} \odot \tanh'(\mathbf{s}^{(1)}) = \begin{bmatrix} -0.44 & 0.88 \end{bmatrix}^\top$

Thus, using the information above, we can compute the following.

- $\frac{\partial e}{\partial \mathbf{W}^{(3)}} = \begin{bmatrix} 1 & \mathbf{x}^{(2)\top} \end{bmatrix}^\top \boldsymbol{\delta}^{(3)\top} = \begin{bmatrix} -1.85 & 1.67 \end{bmatrix}^\top$
- $\frac{\partial e}{\partial \mathbf{W}^{(2)}} = \begin{bmatrix} 1 & \mathbf{x}^{(1)\top} \end{bmatrix}^\top \boldsymbol{\delta}^{(2)\top} = \begin{bmatrix} -0.69 & -0.42 & -0.53 \end{bmatrix}^\top$

Solution 2: Neural Networks and Gradients

Given the network architecture, we can derive the gradients of the sample loss with respect to the layer activations and weights. The input to the model is the feature or data vector $\mathbf{x} \in \mathbb{R}^3$ and the corresponding ground-truth response $y \in \mathbb{R}$. The model parameters are two scalar weights $\Omega = (w, v)$. We use the squared loss function, $e(\Omega) = (\hat{y} - y)^2$, where $\hat{y}$ is the model output.

a) We can express $\frac{\partial e}{\partial v} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v}$. We know that $\frac{\partial e}{\partial \hat{y}} = 2(\hat{y} - y) = 2\Delta$, and $\frac{\partial \hat{y}}{\partial v} = \frac{\partial x_1^{(3)}}{\partial v} + \frac{\partial v x_2^{(2)}}{\partial v}$. Thus, evaluting, we get $\frac{\partial e}{\partial v} = 2\Delta\left(x_2^{(2)} + v x_1^{(1)}\right)$.

b) We derive the expressions for the gradients of the loss with respect to the various layer activations below. For the same, we first express the model output in terms of first and second layer activations: $\hat{y} = x_3^{(1)} + v\left(x_2^{(1)} + v x_1^{(1)}\right)$. Then, we can write

- $\frac{\partial e}{\partial x_2^{(2)}} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial x_2^{(2)}} = 2\Delta v$
- $\frac{\partial e}{\partial x_1^{(1)}} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial x_1^{(1)}} = 2\Delta v^2$
- $\frac{\partial e}{\partial x_2^{(1)}} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial x_2^{(1)}} = 2\Delta v$
- $\frac{\partial e}{\partial x_1^{(3)}} = \frac{\partial e}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial x_1^{(3)}} = 2\Delta$

c) We know that $\frac{\partial e}{\partial w} = \frac{\partial e}{\partial x_3^{(1)}} \frac{\partial x_3^{(1)}}{\partial w} + \frac{\partial e}{\partial x_2^{(1)}} \frac{\partial x_2^{(1)}}{\partial w} + \frac{\partial e}{\partial x_1^{(1)}} \frac{\partial x_1^{(1)}}{\partial w}$. We can write $\frac{\partial x_i^{(1)}}{\partial w} = \theta'(w x_i^{(0)}) x_i^{(0)}$ for $i = 1, 2, 3$, and by using the above parts, we finally get that $\frac{\partial e}{\partial w} = 2\Delta \sum_{i=1}^{3} v^{3-i} \theta'(w x_i^{(0)}) x_i^{(0)}$.