1. **Hard-Coding a Network** In this problem, you need to find a set of weights and biases for a multilayer perceptron which determines if a list of length 4 is in sorted order. More specifically, you receive four inputs $x_1, \cdots, x_4$, where $x_i \in \mathbb{R}$, and the network must output 1 if $x_1 < x_2 < x_3 < x_4$, and 0 otherwise. All hidden units and the output unit use a hard threshold activation function

$$\phi(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

Solution. □

We will set weight and biases such that h_j activates if $x_j < x_{j+1}$. For example, we want $h_1 = 1$ if $x_1 < x_2$, that is for

$$w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + w_{14}x_4 + b_1 > 0$$

weights $(1, -1, 0, 0)$ and bias of -0.1 satisfies the constraint. Similarly, we can hard code weights and biases for h_2 and h_3.

$$\mathbf{W}^{(-1)} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \qquad \mathbf{b}^{(1)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Now we set weights and bias for computing y such that y activates if and only if all h_1, h_2, h_3 activates

$$\mathbf{w}^{(2)} = \begin{pmatrix} -10 & -10 & -10 \end{pmatrix} \qquad b^{(2)} = 0.5$$

2. **Backprop** Consider a neural network with N input units, N output units, and K hidden units. The activations are computed as follows:

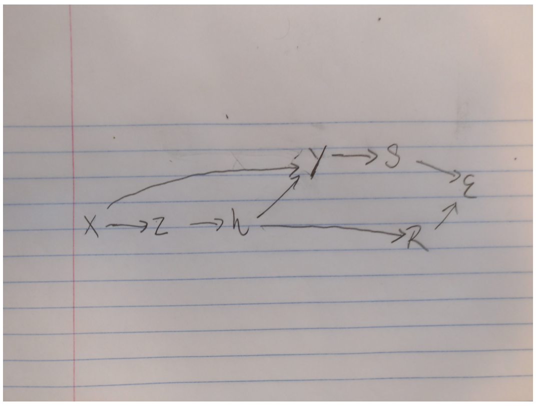
$$\begin{aligned} \mathbf{z} &= \mathbf{W^{(1)}x} + \mathbf{b^{(1)}} \\ \mathbf{h} &= \sigma(\mathbf{z}) \\ \mathbf{y} &= \mathbf{x} + \mathbf{W^{(2)}h} + \mathbf{b^{(2)}} \end{aligned}$$

The cost involves both $\mathbf{h}$ and $\mathbf{y}$

$$\begin{aligned} \mathcal{E} &= \mathcal{R} + \mathcal{S} \\ \mathcal{R} &= \mathbf{r^T h} \\ \mathcal{S} &= \frac{1}{2}||\mathbf{y} - \mathbf{s}||^2 \end{aligned}$$

for given vector $\mathbf{r}$ and $\mathbf{s}$

(a) Draw computation graph



(b) Derive backprop equations for computing $\overline{x} = {}^{\partial \mathcal{E}}/_{\partial \mathbf{x}}$

Solution. □

We first derive scalar form for the forward pass

$$
\begin{aligned}
\mathcal{E} &= \mathcal{R} + \mathcal{S} \\
\mathcal{S} &= \frac{1}{2}\sum_{i=1}^{N}(y_i - s_i)^2 \\
\mathcal{R} &= \sum_{j=1}^{K} r_j h_j \\
y_i &= x_i + \sum_{j}^{K} w_{ij}^{(2)} h_j + b_i^{(2)} \\
h_j &= \sigma(z_j) \\
z_j &= \sum_{i=1}^{N} w_{ji}^{(1)} x_i + b_j^{(1)}
\end{aligned}
$$

for $i = 1, 2, \ldots, N$ and $j = 1, 2, \cdots, K$. Then we derive the scalar form for the reverse pass,

$$\begin{aligned}
\bar{\mathcal{E}} &= 1 \\
\bar{\mathcal{R}} &= \bar{\mathcal{E}}\frac{\partial \mathcal{E}}{\partial \mathcal{R}} = 1 \\
\bar{\mathcal{S}} &= \bar{\mathcal{E}}\frac{\partial \mathcal{R}}{\partial \mathcal{S}} = 1 \\
\bar{y}_i &= \bar{\mathcal{S}}\frac{\partial \mathcal{S}}{\partial y_i} = y_i - s_i \\
\bar{h}_j &= \bar{\mathcal{R}}\frac{\partial \mathcal{R}}{\partial h_j} + \sum_{i=1}^{N} \bar{y}_i \frac{\partial y_i}{\partial h_j} = r_j + \sum_{i=1}^{N} \bar{y}_i w_{ij}^{(2)} \\
\bar{z}_j &= \bar{h}_j \frac{\partial h_j}{\partial z_j} = \bar{h}_j \sigma'(z_j) \\
\bar{x}_i &= \sum_{j=1}^{K} \bar{z}_j \frac{\partial z_j}{\partial x_i} + \bar{y}_i \frac{\partial y_i}{\partial x_i} = \sum_{j=1}^{K} \bar{z}_j w_{ji}^{(1)} + \bar{y}_i
\end{aligned}$$

Then vectorize the result

$$\begin{aligned}
\bar{\mathcal{E}} &= 1 \\
\bar{\mathcal{R}} &= 1 \\
\bar{\mathcal{E}} &= 1 \\
\bar{\mathbf{y}} &= \mathbf{y} - \mathbf{s} \\
\bar{\mathbf{h}} &= \mathbf{r} + \mathbf{W^{(2)T}}\bar{\mathbf{y}} \\
\bar{\mathbf{z}} &= \bar{\mathbf{h}} \circ \sigma'(\mathbf{z}) \\
\bar{\mathbf{x}} &= \mathbf{W^{(1)T}}\bar{\mathbf{z}} + \bar{\mathbf{y}}
\end{aligned}$$

3. **Sparsifying Activation Function**

Solution. □

$$\begin{array}{cc}
\dfrac{\partial \mathcal{L}}{\partial w_1} & \text{YES} \\
\dfrac{\partial \mathcal{L}}{\partial w_2} & \text{YES} \\
\dfrac{\partial \mathcal{L}}{\partial w_3} & \text{NO}
\end{array}$$

If h_1 is 0, then y do not depend on $w_1 h_1 = 0$ and so the a measure in change of $\mathcal{L}$ with respect to w_1 while holding other variables constant is zero, i.e. $\dfrac{\partial \mathcal{L}}{\partial w_1} = 0$. A infinitesimal change in w_2 yields negative value as input to h_1, by similar argument shown previously, we have $\dfrac{\partial \mathcal{L}}{\partial w_2} = 0$. However, a small change in w_3 might result in changes in y if both h_2, h_3 activates, and so $\dfrac{\partial \mathcal{L}}{\partial w_3}$ might not be zero.