1.1 a) $\boxed{\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}}$

$$\frac{d}{dx}\left(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right) = \frac{(e^{x}+e^{-x})(e^{x}+e^{-x}) - (e^{x}-e^{-x})(e^{x}-e^{-x})}{(e^{x}-e^{-x})^2}$$

$$= \frac{(e^{2x}+e^{0}+e^{0}+e^{-2x}) - (e^{2x}-e^{0}-e^{0}+e^{-2x})}{(e^{x}-e^{-x})^2} = \frac{4}{(e^{x}-e^{-x})^2}$$

$$= \frac{2}{(e^{x}-e^{-x})}\cdot\frac{2}{(e^{x}-e^{-x})} = \frac{1}{\cosh^2(x)} = \operatorname{sech}^2(x)$$

Therefore, $\frac{d}{dx}(\tanh(x)) = \operatorname{sech}^2(x) = 1-\tanh^2(x)$

• Case 1: j is an output unit

output from $j = o_j = \tanh(net_j)$

So, $\frac{\partial o_j}{\partial net_j} = 1-\tanh^2(net_j)$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j}\cdot\frac{\partial net_j}{\partial w_{ji}} = \frac{\partial E_d}{\partial o_j}\cdot\frac{\partial o_j}{\partial net_j}\cdot\frac{\partial net_j}{\partial w_{ji}} = -(t_j-o_j)(1-\tanh^2(net_j))x_{ji}$$

$$\Delta w_{ji} = -\eta\frac{\partial E_d}{\partial w_{ji}} = \eta(t_j-o_j)(1-o_j^2)x_{ji}$$

• Let's call $(t_j-o_j)(1-o_j^2) = \delta_j$ such that $\frac{\partial E_d}{\partial net_j} = \delta_j$

$$\boxed{\Delta w_{ji} = -\eta\delta_j x_{ji}}$$

Case 2: j is a hidden unit

$$\frac{\partial E_d}{\partial net_j} = \sum_{k\in Downstream(j)}\frac{\partial E_d}{\partial net_k}\frac{\partial net_k}{\partial net_j} = \sum_{k\in Ds(j)} -\delta_k\frac{\partial net_k}{\partial net_j} = \sum_{k\in Ds(j)} -\delta_k\frac{\partial net_k}{\partial o_j}\cdot\frac{\partial o_j}{\partial net_j} = \sum_{k\in Ds(j)} -\delta_k w_{kj}(1-\tanh^2(net_j))$$

$$\boxed{\frac{\partial E_d}{\partial net_j} = \sum_{k\in Ds(j)} -\delta_k w_{kj}(1-o_j^2)}$$

Putting it all together: $\delta_j = (1-o_j^2)\sum_{k\in Ds(j)}\delta_k w_{kj}$

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

1.1 b) $ReLu(x) = \max(0,1)$

$$\frac{\partial}{\partial x}(ReLu(x)) = \begin{cases} 1, & \text{if } x>0 \\ 0, & \text{otherwise} \end{cases}$$

Case 1: j is an output unit

output from $j = o_j = ReLu(net_j)$

So, $\frac{\partial o_j}{\partial net_j} = \begin{cases} 1, & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}} = \frac{\partial E_d}{\partial o_j} \cdot \frac{\partial o_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ji}} = \begin{cases} -(t_j - o_j)(1)(x_{ji}), & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}} = \begin{cases} \eta(t_j - o_j)x_{ji}, & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}$$

- Let's call $\left(\begin{cases} (t_j - o_j), & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}\right) = \delta_j$

such that $\frac{\partial E_d}{\partial net_j} = -\delta_j$

$$\boxed{\Delta w_{ji} = -\eta \delta_j x_{ji}}$$

Case 2: j is a hidden unit

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in Downstream(j)} \frac{\partial E_d}{\partial net_k} \frac{\partial net_k}{\partial net_j} = \sum_{k \in Ds(j)} -\delta_k \frac{\partial net_k}{\partial net_j} = \sum_{k \in Ds(j)} -\delta_k \frac{\partial net_k}{\partial o_j} \frac{\partial o_j}{\partial net_j} = \sum_{k \in Ds(j)} -\delta_k w_{kj} \left(\begin{cases} 1, & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}\right)$$

Putting it all together: $\delta_j = \begin{cases} \sum_{k \in Ds(j)} \delta_k w_{kj}, & \text{if } net_j > 0 \\ 0, & \text{otherwise} \end{cases}$

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

1.2 Gradient Descent

$$o = w_0 + w_1(x_1 + x_1^2) + \ldots + w_n(x_n + x_n^2)$$

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

$$\frac{\partial E}{\partial w_i} = \frac{1}{2}(2) \sum_{d \in D} (t_d - o_d) \frac{\partial}{\partial w_i} \left(t_d - (w_0 + w_1 x_{1d} + w_1 x_{1d}^2 + \ldots + w_n x_{nd} + w_n x_{nd}^2\right)$$

$$\frac{\partial E}{\partial w_i} = \sum_{d \in D} (t_d - o_d)(-x_{id} - x_{id}^2)$$

$$\Delta w_i = \eta \sum_{d \in D} (t_d - o_d)(x_{id} + x_{id}^2)$$

* w/ $f(x) = x$ activation function

$f'(x) = 1$

1.3 Comparing Activation Function

a)

Node	Net	Output
1	x_1	x_1
2	x_2	x_2
3	$net_3 = x_1 w_{31} + x_2 w_{32}$	$x_3 = h(net_3)$
4	$net_4 = x_1 w_{41} + x_2 w_{42}$	$x_4 = h(net_4)$
5	$net_5 = x_3 w_{53} + x_4 w_{54}$	$y_5 = h(net_5)$

b) $\left(W^{(1)}_{2\times2} \cdot X_{2\times1}\right) = \begin{bmatrix} x'_3 \\ x'_4 \end{bmatrix} \xrightarrow{\text{Use Activation Function}} h\left[\begin{bmatrix} x'_3 \\ x'_4 \end{bmatrix}\right] = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}$

- Let's use this notation as in part a)
- x' is before $h(x)$ activation function

$$\left(W^{(2)}_{1\times2} \cdot \begin{bmatrix} x_3 \\ x_4 \end{bmatrix}_{2\times1}\right) = \begin{bmatrix} w_{53} & w_{54} \end{bmatrix}\begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = w_{53}x_3 + w_{54}x_4$$

$$y_5 = h(w_{53}x_3 + w_{54}x_4)$$

So, $h\left(W^{(2)} \cdot h\left(W^{(1)} \cdot X\right)\right) = y_5$ = output

c) Sigmoid: $h_s(x) = \frac{1}{1+e^{-x}} = \sigma(x)$ | Tanh: $h_t(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

- $\sigma(-x) = 1 - \sigma(x)$
- $h_s(-x) = 1 - h_s(x)$

- Let's adjust Tanh(x) such that $h_t(x) = \frac{e^x + e^{-x} - e^{-x} - e^{-x}}{e^x + e^{-x}}$

$$h_t(x) = \frac{e^x + e^{-x} - 2e^{-x}}{e^x + e^{-x}} = 1 - \frac{2e^{-x}}{e^x + e^{-x}}$$

$$h_t(x) = 1 - \frac{2e^{-x}}{e^x + e^{-x}} \frac{\left(\frac{1}{e^{-x}}\right)}{\left(\frac{1}{e^{-x}}\right)} = 1 - \frac{2}{e^{2x}+1} = \underbrace{1 - \frac{2}{1+e^{2x}}}$$

Using this relationship

$$h_t(x) = 1 - 2h_s(-2x) = 1 - 2(1 - h_s(2x))$$

$$h_t(x) = 1 - 2 + 2h_s(2x)$$

$$h_t(x) = 2h_s(2x) - 1$$

- This shows that $h_t(x)$ is a rescaled $h_s(x)$.

1.4 Gradient Descent with a Weight Penalty

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

- Output Layer Node: $\frac{\partial E}{\partial w_{ji}} = (t_j - o_j) o_j (1 - o_j) x_{ji} - 2\eta\gamma w_{ji}$

* Assuming Sigmoid Activation Function *

Thus, $\Delta w_{ji} = \eta (t_j - o_j) o_j (1 - o_j) x_{ji} - 2\eta\gamma w_{ji}$

Weight Update: $\overset{new}{w_{ji}} = w_{ji} + \Delta w_{ji}$

new $w_{ji} = w_{ji}(1 - 2\eta\gamma) + \eta (t_j - o_j) o_j (1 - o_j) x_{ji}$

- Hidden Neuron Node: $\frac{\partial E}{\partial w_{ji}} = \eta o_j (1 - o_j) x_{ji} \sum_{k \in downstream(j)} \delta_k w_{kj} - 2\eta\gamma w_{ji}$

* Assume Sigmoid Activation Function * | * $\left(\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}\right)$ *

Weight Update: $\overset{new}{w_{ji}} = w_{ji} + \Delta w_{ji}$

new $w_{ji} = w_{ji}(1 - 2\eta\gamma) + \eta o_j (1 - o_j) x_{ji} \sum_{k \in downstream(j)} \delta_k w_{kj}$