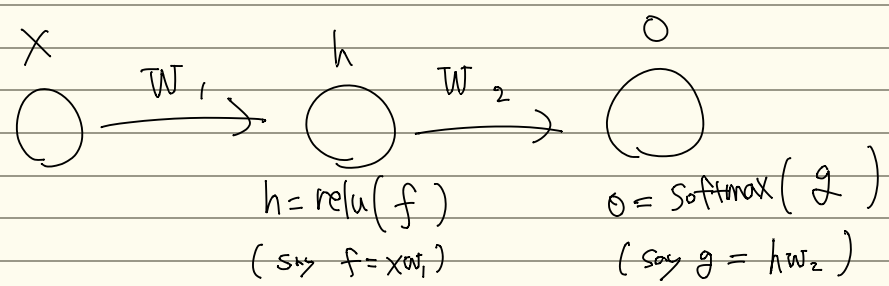
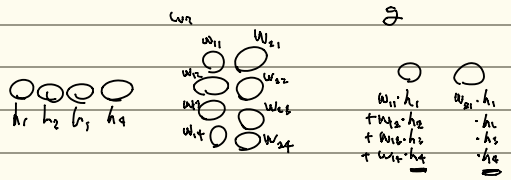
(for simplicity)





$$E = -\sum_{j \in D} t_j \cdot \log(o_j)$$

Now we want to compute $\frac{\partial E}{\partial W_1}$ (12 weights), $\frac{\partial E}{\partial W_2}$ (8 weights) — total 20 weights' gradient.

$$\frac{\partial E}{\partial W_2} = \frac{\partial E}{\partial o}\frac{\partial o}{\partial g}\frac{\partial g}{\partial W_2}$$

→ (Note W_2 here is not a matrix, just simple notation for any weight in W_2, such as W_{kj}, $k=1,2$, $j=1,2,3,4$)

$$\frac{\partial E}{\partial W_1} = \frac{\partial E}{\partial o}\frac{\partial o}{\partial g}\frac{\partial g}{\partial h}\frac{\partial h}{\partial f}\frac{\partial f}{\partial W_1}$$

→ (same for W_1, W_1 here means any weight in W_1, such as W_{ji}) $i=1,2,3$

Now we can easily see $\frac{\partial E_d}{\partial o_n} = -\frac{t}{o_n}$, $\frac{\partial o_n}{\partial g_k} = \begin{cases} o_n(1-o_n) & : k=n \\ -o_n \cdot o_k & : k \neq n \end{cases}$

↳ for each data. ↳ property of softmax. ↳ product of these two may be little bit confussing so Let's just skip.

$$\therefore \frac{\partial E_d}{\partial o_n}\cdot\frac{\partial o_n}{\partial g_k} = o_k - t_k$$

and for $\frac{\partial g_k}{\partial W_{kj}} = h_j$ $\therefore$ $\boxed{\frac{\partial E_d}{\partial W_{kj}} = (o_k - t_k)h_j}$

with $\sum_D$, ∇E for W_{kj}

Now for W_1, $\frac{\partial g_k}{\partial h_j} = W_{kj}$, $\frac{\partial h}{\partial f} = \begin{cases} 0 & : f \le 0 \\ 1 & : f > 0 \end{cases}$, $\frac{\partial f_j}{\partial W_{ji}} = X_i$

$$\therefore \boxed{\frac{\partial E_d}{\partial W_{ji}} = \sum_{k=1}^{2}(o_k - t_k)\cdot W_{kj}\cdot(\text{if}(f>0))\cdot X_i}$$

(if(f>0)): 1 = true, 0 = false

$\sum_{d \in D} \Rightarrow \nabla E$ for W_{ji}

for weight in W_1, we must consider all output nodes ($k=1\sim2$)