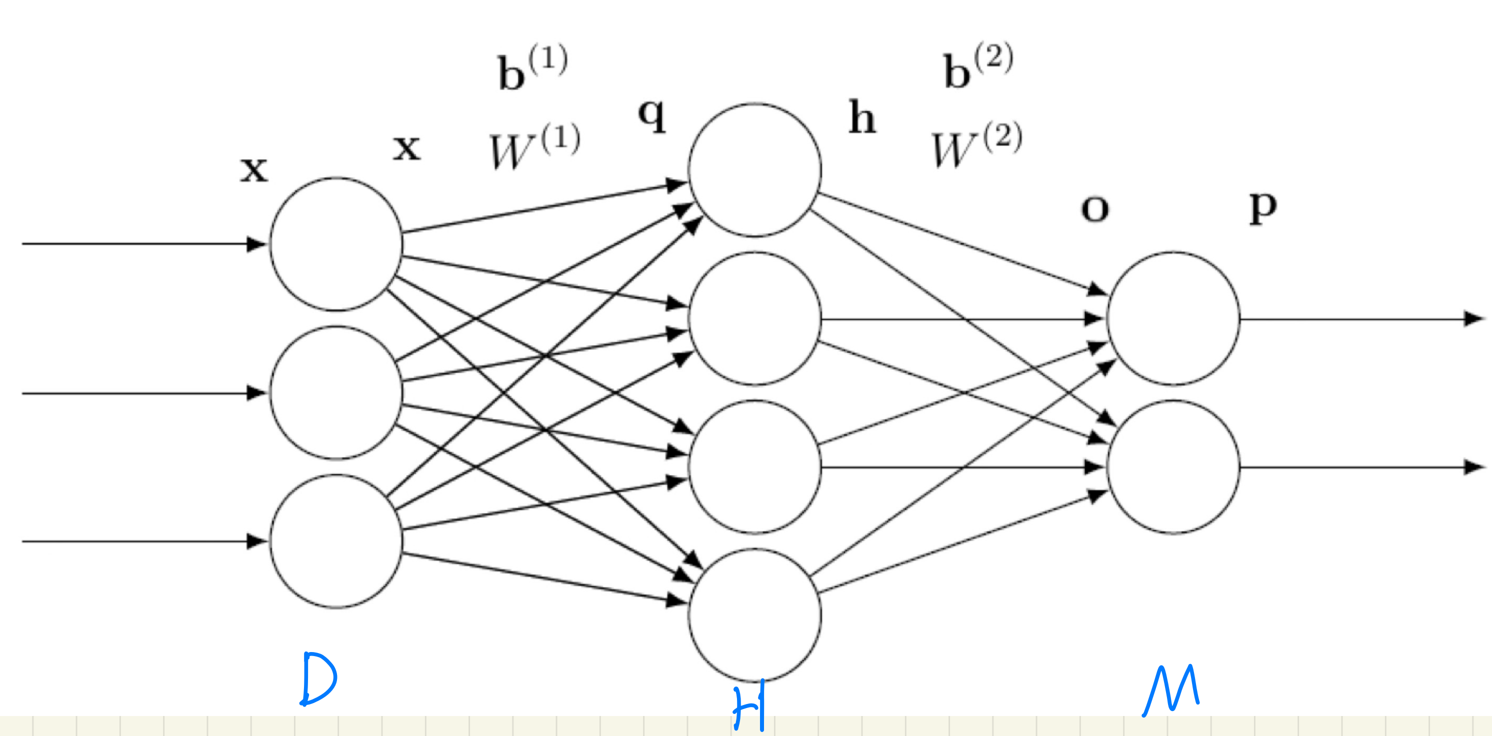

Forward Process

$$q = W^1 X + b^1$$

$$h = \mathrm{ReLU}(q)$$

$$o = W^2 h + b^2$$

$$p = \mathrm{softmax}(o), \text{ where } p_i = \frac{e^{o_i}}{\sum_{k=1}^{M} e^{o_k}}$$

$$L(p, y) = -\sum_{i=1}^{M} y_i \log p_i$$

Backward Process

$$\frac{\partial P_i}{\partial O_j} = \frac{\partial\left(\frac{e^{O_i}}{\sum_{k=1}^{M} e^{O_k}}\right)}{\partial \quad O_j} = \delta_{ij} \frac{e^{O_i}}{\sum_{k=1}^{M} e^{O_k}} - \frac{e^{O_i}}{\left(\sum_{k=1}^{M} e^{O_k}\right)^2} \cdot e^{O_j}$$

$$\delta_{ij} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

$$= P_i(\delta_{ij} - P_j)$$

$$\frac{\partial L}{\partial O_j} = \sum_{i=1}^{M} \frac{\partial L}{\partial P_i} \frac{\partial P_i}{\partial O_j} = \sum_{i=1}^{M} \frac{\partial\left(-\sum_{k=1}^{M} y_k \log P_k\right)}{\partial \quad P_i} \cdot \frac{\partial P_i}{\partial O_j}$$

$$= \sum_{i=1}^{M} -y_i \frac{1}{\cancel{P_i}} \cdot \cancel{P_i}(\delta_{ij} - P_j)$$

$$= \sum_{i=1}^{M} y_i P_j - \sum_{i=1}^{M} y_i \delta_{ij}$$

$$= P_j - y_j$$

$$\frac{\partial O_i}{\partial b_j^{(2)}} = \delta_{ij}$$

$$\frac{\partial L}{\partial b_j^{(2)}} = \frac{\partial L}{\partial O_j} \frac{\partial O_j}{\partial b_j} = P_j - y_j$$

$$\frac{\partial O_i}{\partial h_j} = \frac{\partial \left(\sum_{k=1}^{H} W_{ik} h_k + b_i \right)}{\partial h_j} = W_{ij}$$

$$\frac{\partial L}{\partial h_j} = \sum_{i=1}^{M} \frac{\partial L}{\partial O_i} \frac{\partial O_i}{\partial h_j} = \sum_{i=1}^{M} (P_i - y_i) \cdot W_{ij}$$

$$\frac{\partial O_k}{\partial W_{ij}^{(2)}} = \frac{\partial \left(\sum_{n=1}^{H} W_{kn} \cdot h_n + b_k \right)}{\partial W_{ij}^{(2)}} = \sum_{n=1}^{H} \delta_{ik} \delta_{nj} h_n = \delta_{ki} h_j$$

$$\frac{\partial L}{\partial W_{ij}} = \sum_{k=1}^{M} \frac{\partial L}{\partial O_k} \frac{\partial O_k}{\partial W_{ij}^{(2)}} = \sum_{k=1}^{M} (P_k - y_k) \cdot \delta_{ki} h_j = (P_i - y_i) h_j$$

$$\frac{\partial h_i}{\partial q_j} = \delta_{ij} \cdot a \qquad a = \begin{cases} 1 & \text{if } q_j \geq 0 \\ 0 & \text{if } q_j < 0 \end{cases}$$

$$\frac{\partial L}{\partial q_i} = \frac{\partial L}{\partial h_i} \frac{\partial h_i}{\partial q_j} = \frac{\partial L}{\partial h_j} \cdot a$$