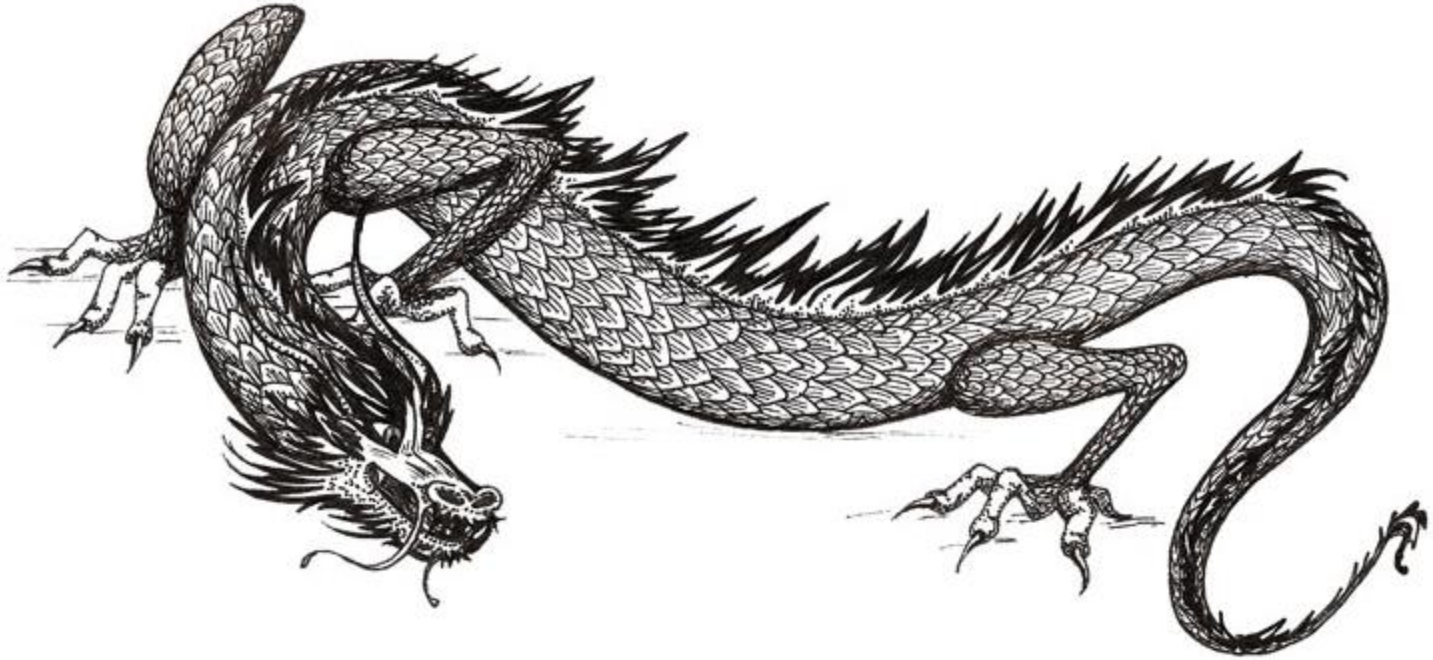
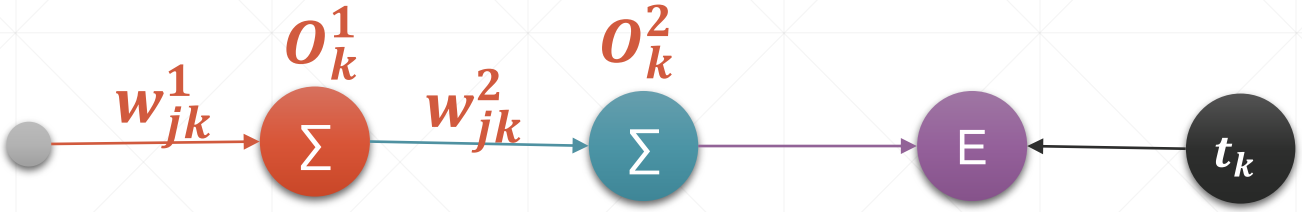


多层感知机梯度

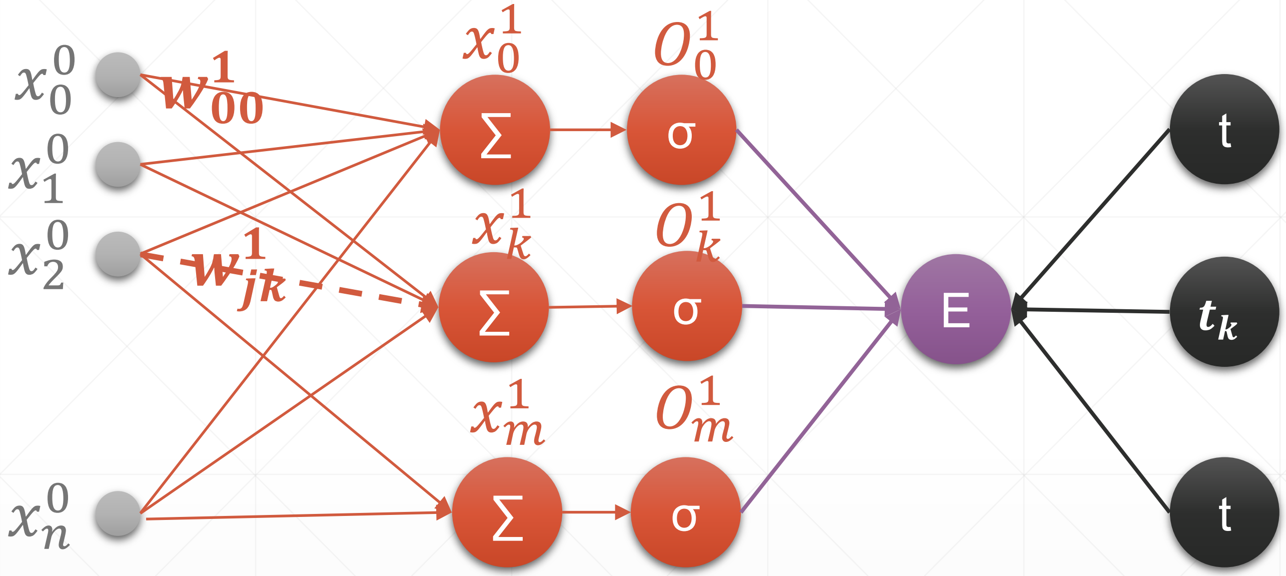
主讲：龙良曲

Chain rule



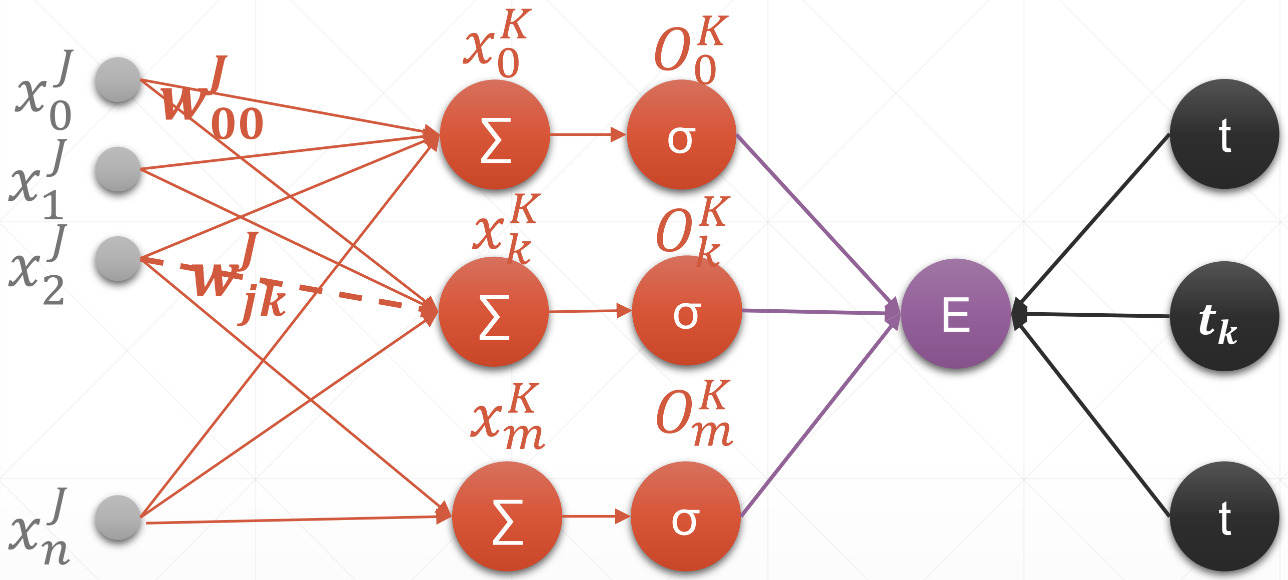
- $$\frac{\partial E}{\partial w_{jk}^1} = \frac{\partial E}{\partial o_k^1}\frac{\partial o_k^1}{\partial x} = \frac{\partial E}{\partial o_k^2}\frac{\partial o_k^2}{\partial o_k^1}\frac{\partial o_k^1}{\partial x}$$

Multi-output Perceptron

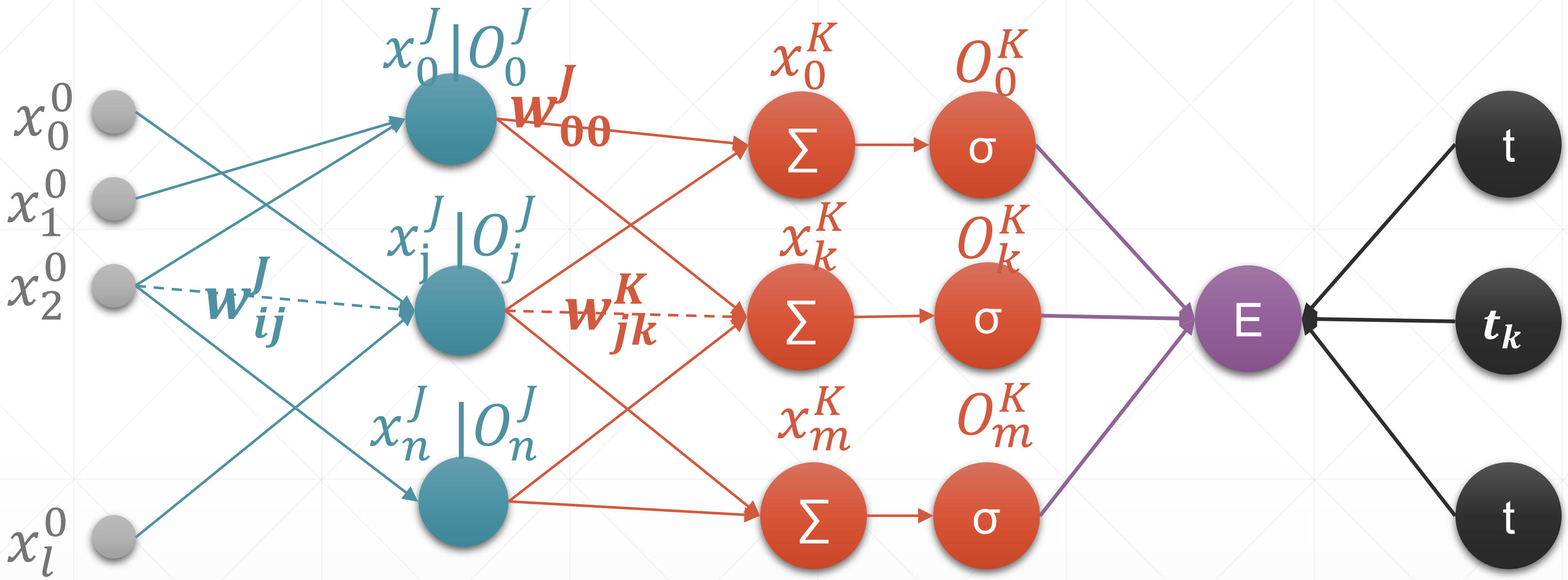


$$\frac{\partial E}{\partial w_{jk}} = (O_k - t_k)\mathrm{O_k}\,(1 - O_k)\, x_j^0$$

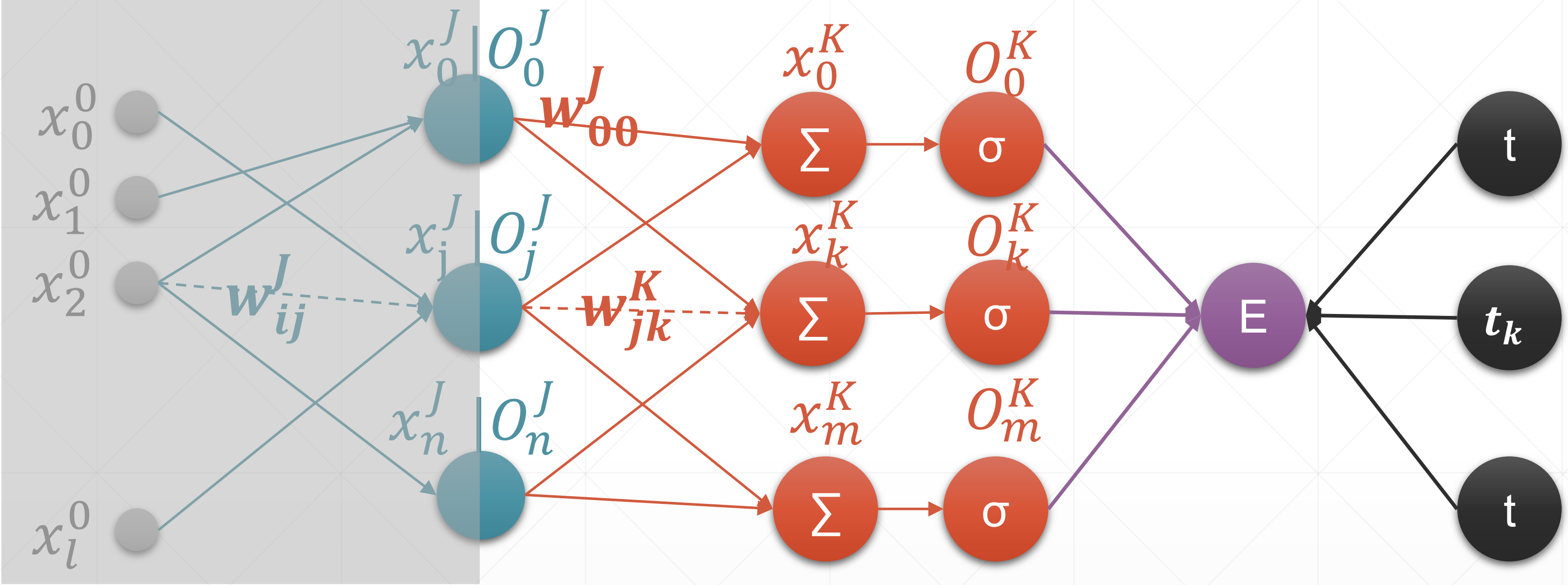
Multi-Layer Perceptron



Multi-Layer Perceptron



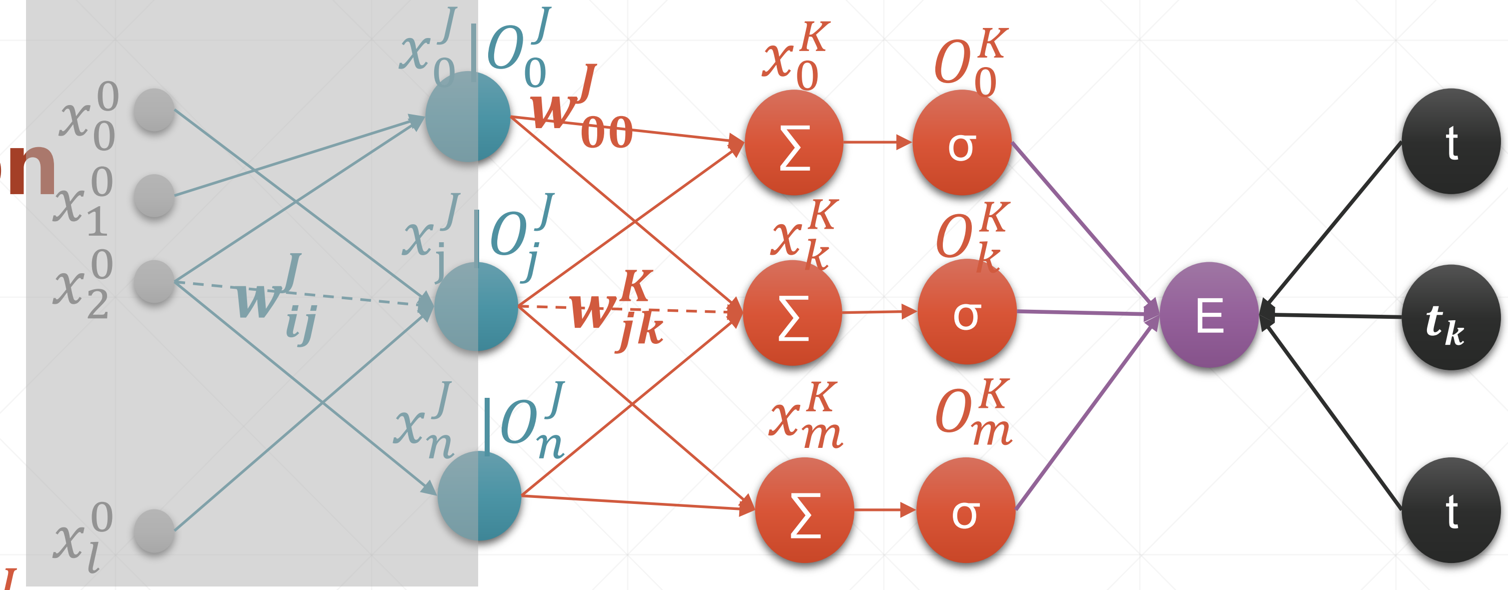
Multi-Layer Perceptron



$$\frac{\partial E}{\partial w_{jk}} = (O_k - t_k)\mathrm{O_k}\,(1 - O_k)\; x_j^0$$

$$\frac{\partial E}{\partial w_{jk}} = (O_k - t_k)\mathrm{O_k}\,(1 - O_k)\; \mathrm{O}_j^J$$

Multi-Layer Perceptron



$$\frac{\partial E}{\partial w_{jk}} = \underbrace{(O_k - t_k)O_k(1 - O_k)}_{} O_j^J$$

$$\frac{\partial E}{\partial w_{jk}} = \delta_k^K \quad O_j^J$$

$$\frac{\partial E}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} \frac{1}{2} \sum_{k \in K} (\mathcal{O}_k - t_k)^2$$

$$\frac{\partial E}{\partial W_{ij}} = \sum_{k \in K} (\mathcal{O}_k - t_k) \frac{\partial}{\partial W_{ij}} \mathcal{O}_k$$

$$\frac{\partial E}{\partial W_{ij}} = \sum_{k \in K} (\mathcal{O}_k - t_k) \frac{\partial}{\partial W_{ij}} \sigma(x_k)$$

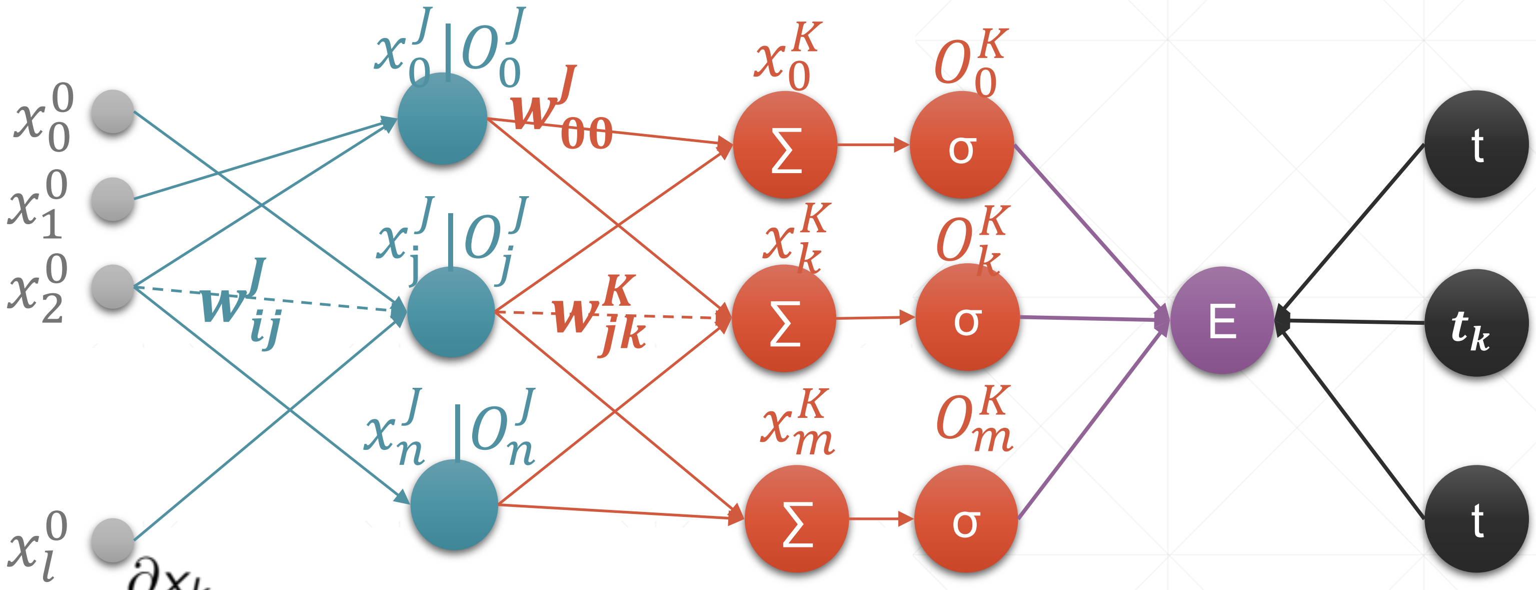
$$\frac{\partial E}{\partial W_{ij}} = \sum_{k \in K} (\mathcal{O}_k - t_k) \sigma(x_k)(1 - \sigma(x_k)) \frac{\partial x_k}{\partial W_{ij}}$$

$$\frac{\partial E}{\partial W_{ij}} = \sum_{k \in K} (\mathcal{O}_k - t_k) \mathcal{O}_k (1 - \mathcal{O}_k) \frac{\partial x_k}{\partial \mathcal{O}_j} \cdot \frac{\partial \mathcal{O}_j}{\partial W_{ij}}$$

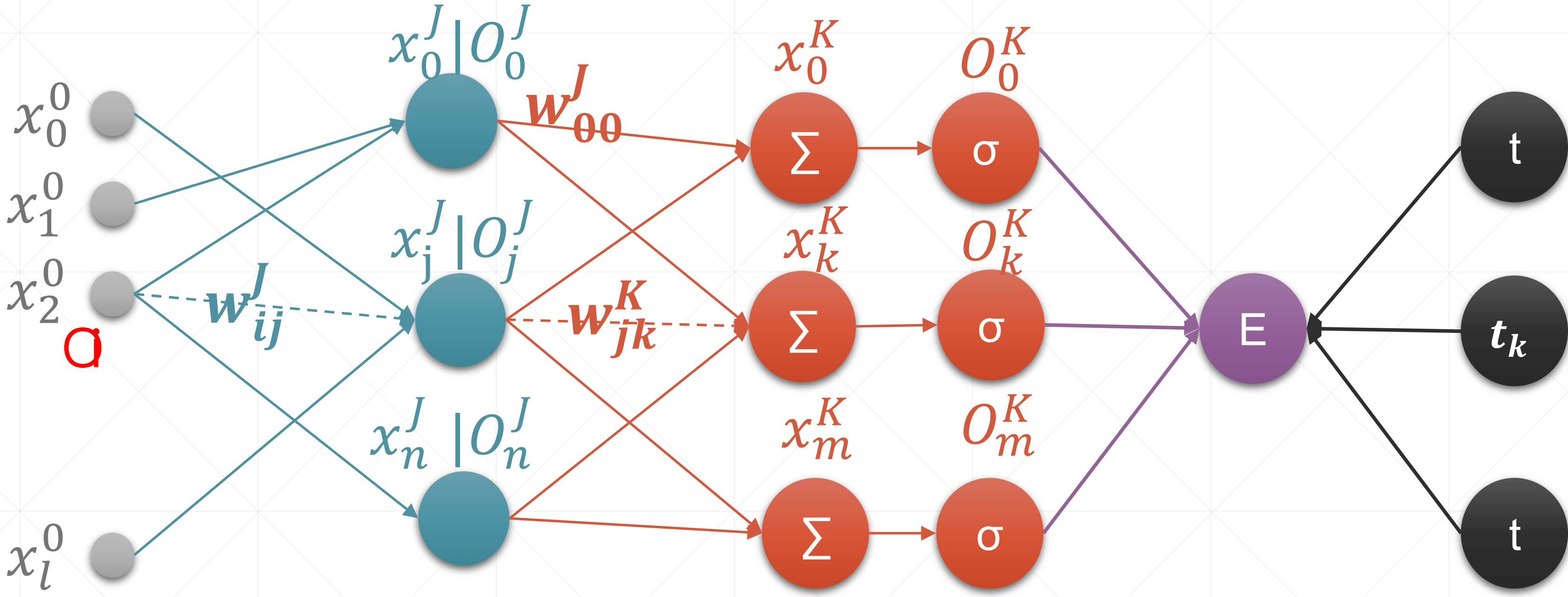
$$\frac{\partial E}{\partial W_{ij}} = \sum_{k \in K} (\mathcal{O}_k - t_k) \mathcal{O}_k (1 - \mathcal{O}_k) W_{jk} \frac{\partial \mathcal{O}_j}{\partial W_{ij}}$$

$$\frac{\partial E}{\partial W_{ij}} = \frac{\partial \mathcal{O}_j}{\partial W_{ij}} \sum_{k \in K} (\mathcal{O}_k - t_k) \mathcal{O}_k (1 - \mathcal{O}_k) W_{jk}$$

$$\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_j (1 - \mathcal{O}_j) \frac{\partial x_j}{\partial W_{ij}} \sum_{k \in K} (\mathcal{O}_k - t_k) \mathcal{O}_k (1 - \mathcal{O}_k) W_{jk}$$



$$\boxed{\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_j (1 - \mathcal{O}_j) \mathcal{O}_i \sum_{k \in K} (\mathcal{O}_k - t_k) \mathcal{O}_k (1 - \mathcal{O}_k) W_{jk}}$$

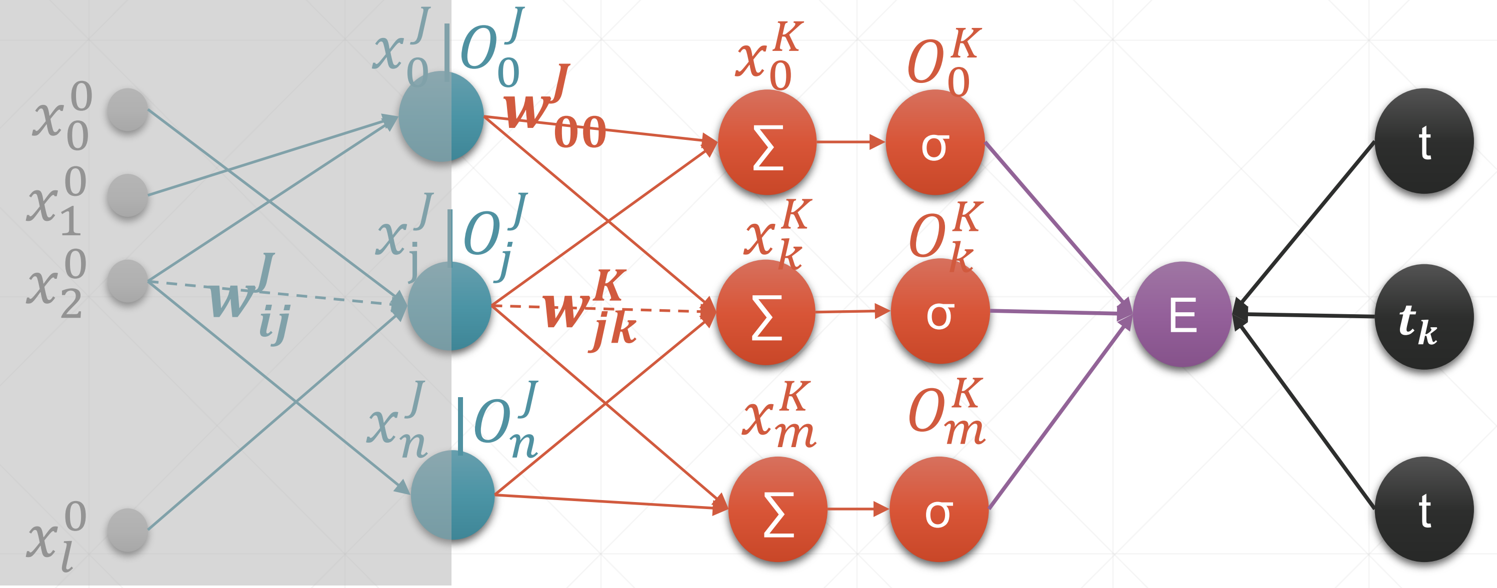


$$\boxed{\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_j(1-\mathcal{O}_j)\mathcal{O}_i \sum_{k\in K}(\mathcal{O}_k - t_k)\mathcal{O}_k(1-\mathcal{O}_k)W_{jk}}$$

$$\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_i\mathcal{O}_j(1-\mathcal{O}_j)\sum_{k\in K}\delta_k W_{jk}$$

$$\frac{\partial E}{\partial w_{jk}} = (O_k - t_k)O_k(1 - O_k)\, O_j^J$$

$$\frac{\partial E}{\partial w_{jk}} = \delta_k^K \, O_j^J$$



$$\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_j(1 - \mathcal{O}_j)\mathcal{O}_i \sum_{k \in K}(\mathcal{O}_k - t_k)\mathcal{O}_k(1 - \mathcal{O}_k)W_{jk}$$

$$\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_i\mathcal{O}_j(1 - \mathcal{O}_j) \sum_{k \in K} \delta_k W_{jk}$$

$$\frac{\partial E}{\partial W_{ij}^{(l)}} = O_i^{(l-1)} \delta_j^{(l)}$$

For an output layer node $k \in K$

输出层: $\delta^{(l)} = O^{(l)} .* (1 - O^{(l)}) .* (O^{(l)} - t)$

隐藏层: $\delta^{(l)} = O^{(l)} .* (1 - O^{(l)}) .* \delta^{(l+1)} (W^{(l+1)})^T$

$$\frac{\partial E}{\partial W_{jk}} = \mathcal{O}_j \delta_k$$

where

$$\delta_k = \underline{\mathcal{O}_k(1 - \mathcal{O}_k)}(\mathcal{O}_k - t_k)$$

若最后一层无激活函数，则没有这一部分

For a hidden layer node $j \in J$

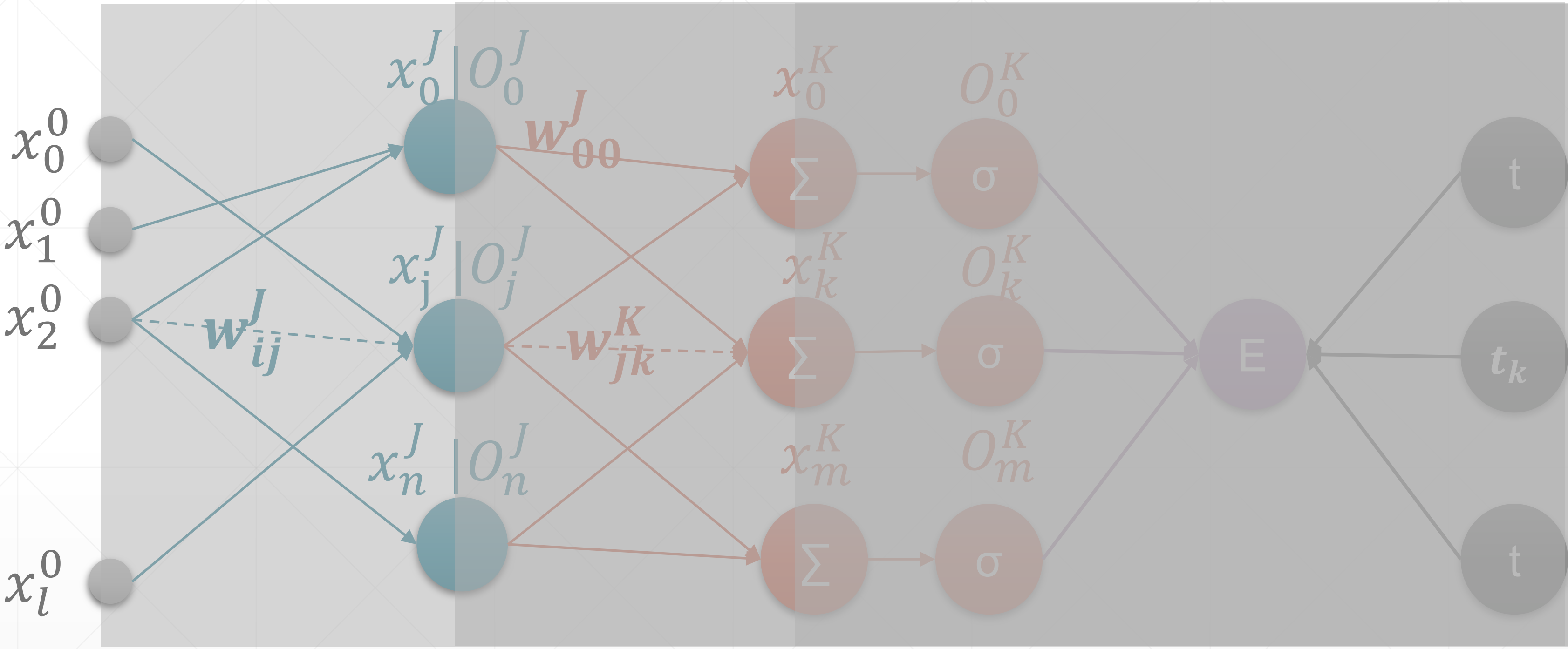
$$\frac{\partial E}{\partial W_{ij}} = \mathcal{O}_i \delta_j$$

此处的参数矩阵与Ng课程中的互为转置，因为Ng中每层为列向量，而此处为行向量

where

$$\delta_j = \mathcal{O}_j(1 - \mathcal{O}_j) \sum_{k \in K} \delta_k W_{jk}$$

- δ_k^K
- $\frac{\partial E}{\partial w_{jk}}$
- δ_j^J
- $\frac{\partial E}{\partial w_{ij}}$
- δ_i^I
- $\frac{\partial E}{\partial w_{ni}}$



Congratulations!

下一课时

优化与训练

Thank You.