如图所示三层神经网络：输入层 $Layer_1$、隐含层 $Layer_2$、输出层 $Layer_3$．输入层包括三个神经元，即输入为 $\boldsymbol{x}=\{x_1,x_2,x_3\}$，样本数据集为 $\mathcal{D}=\{(\boldsymbol{x}_i,y_i)\}$，目标输出为 o_1^{T} 和 o_2^{T}．$\omega_{ij}^{\mathrm{L}_k}$ 为第 $k-1$ 层的第 i 个神经元到第 k 层的第 j 个神经元之间的权重．$\boldsymbol{b}^{\mathrm{L}_k}=\{b_i^{\mathrm{L}_k}\}$ 为第 k 层神经元偏移量集合，$b_j^{\mathrm{L}_k}$ 为第 k 层第 i 个神经元的偏移量．$net_i^{\mathrm{L}_k}$ 表示第 k 层网络的第 i 个神经元的输入，$o_i^{\mathrm{L}_k}$ 表示第 k 层网络的第 i 个神经元的输出，$f_i^{\mathrm{L}_k}(\cdot)$ 表示第 k 层网络的第 i 个神经元的激活函数．

假设激活函数为 sigmoid 函数：

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

其导数为：

$$f'(x)=f(x)(1-f(x)) \tag{2}$$

对于 $Layer_2$ 层第 j 个神经元的输出：

$$net_j^{\mathrm{L}_2}=\sum_{i=1}^{3}\omega_{ij}^{\mathrm{L}_1}\times x_i+b_j^{\mathrm{L}_2} \tag{3}$$

对于 $Layer_3$ 层的第 j 个神经元的输出：

$$net_j^{\mathrm{L}_3}=\sum_{i=1}^{2}\omega_{ij}^{\mathrm{L}_2}\times o_i^{\mathrm{L}_2}+b_j^{\mathrm{L}_3} \tag{4}$$

第 L_i 层的第 j 个神经元的输出：

$$o_j^{\mathrm{L}_i}=f_j^{\mathrm{L}_i}(net_j^{\mathrm{L}_i})=\frac{1}{1+e^{-net_j^{\mathrm{L}_i}}} \tag{5}$$

输出的总误差为：

$$E_{\mathrm{Total}}=\frac{1}{2}\sum_{j=1}^{2}(o_j^{\mathrm{T}}-o_j^{\mathrm{L}_3})^2=\frac{1}{2}((o_1^{\mathrm{T}}-o_1^{\mathrm{L}_3})^2+(o_2^{\mathrm{T}}-o_2^{\mathrm{L}_3})^2) \tag{6}$$

是 $o_1^{\mathrm{L}_3}$ 和 $o_2^{\mathrm{L}_3}$ 的函数

1 $Layer_2$ 与 $Layer_3$ 间参数调整

调整 $Layer_2$ 与 $Layer_3$ 间的权重 $\omega_{ij}^{\mathrm{L}_2}$

$$\frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_2}}=\frac{\partial E_{\mathrm{Total}}}{\partial o_j^{\mathrm{L}_3}}\cdot\frac{\partial o_j^{\mathrm{L}_3}}{\partial net_j^{\mathrm{L}_3}}\cdot\frac{\partial net_j^{\mathrm{L}_3}}{\partial \omega_{ij}^{\mathrm{L}_2}} \tag{7}$$

$$=-(o_j^{\mathrm{T}}-o_j^{\mathrm{L}_3})\cdot f_j^{\mathrm{L}_3}(net_j^{\mathrm{L}_3})(1-f_j^{\mathrm{L}_3}(net_j^{\mathrm{L}_3}))\cdot o_i^{\mathrm{L}_2} \tag{8}$$

$$=-(o_j^{\mathrm{T}}-o_j^{\mathrm{L}_3})\cdot o_j^{\mathrm{L}_3}(1-o_j^{\mathrm{L}_3})\cdot o_i^{\mathrm{L}_2} \tag{9}$$

$$\tag{10}$$

令 $\delta_{ij}^{\mathrm{L}_2}$ 表示 $Layer_2$ 层第 i 个神经元与 $Layer_3$ 层第 j 个神经元间权重的梯度项，则

$$\delta_{ij}^{\mathrm{L}_2} = \frac{\partial E_{\mathrm{Total}}}{\partial o_j^{\mathrm{L}_3}} \cdot \frac{\partial o_j^{\mathrm{L}_3}}{\partial net_j^{\mathrm{L}_3}} \tag{11}$$

$$= -(o_j^{\mathrm{T}} - o_j^{\mathrm{L}_3}) \cdot o_j^{\mathrm{L}_3}(1 - o_j^{\mathrm{L}_3}) \tag{12}$$

可以看出， $Layer_2$ 层第 i 个神经元与 $Layer_3$ 层第 j 个神经元间权重的梯度项是与 i 无关的，故连接至 $Layer_3$ 层第 j 个神经元相对应的权重的梯度向均为 $\delta_{ij}^{\mathrm{L}_2}$，用 $\delta_{\cdot j}^{\mathrm{L}_2}$ 表示，

$$\delta_{\cdot j}^{\mathrm{L}_2} = \delta_{ij}^{\mathrm{L}_2} = -(o_j^{\mathrm{T}} - o_j^{\mathrm{L}_3}) \cdot o_j^{\mathrm{L}_3}(1 - o_j^{\mathrm{L}_3}) \tag{13}$$

则整体误差 E_{Total} 对 $\omega_{ij}^{\mathrm{L}_2}$ 的偏导数公式写作：

$$\frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_2}} = \delta_{\cdot j}^{\mathrm{L}_2} \cdot o_i^{\mathrm{L}_2} \tag{14}$$

则权重 $\omega_{ij}^{\mathrm{L}_2}$ 的学习公式为：

$$\widehat{\omega}_{ij}^{\mathrm{L}_2} = \omega_{ij}^{\mathrm{L}_2} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_2}} = \omega_{ij}^{\mathrm{L}_2} - \eta \cdot \delta_{\cdot j}^{\mathrm{L}_2} \cdot o_i^{\mathrm{L}_2} \tag{15}$$

η 为学习速率

$Layer_3$ 的第 j 个神经元的偏移量的梯度为

$$\frac{\partial E_{\mathrm{Total}}}{\partial b_j^{\mathrm{L}_3}} = \frac{\partial E_{\mathrm{Total}}}{\partial o_j^{\mathrm{L}_3}} \cdot \frac{\partial o_j^{\mathrm{L}_3}}{\partial net_j^{\mathrm{L}_3}} \cdot \frac{\partial net_j^{\mathrm{L}_3}}{\partial b_j^{\mathrm{L}_3}} \tag{16}$$

$$= -(o_j^{\mathrm{T}} - o_j^{\mathrm{L}_3}) \cdot f_j^{\mathrm{L}_3}(net_j^{\mathrm{L}_3})(1 - f_j^{\mathrm{L}_3}(net_j^{\mathrm{L}_3})) \cdot 1 \tag{17}$$

$$= -(o_j^{\mathrm{T}} - o_j^{\mathrm{L}_3}) \cdot o_j^{\mathrm{L}_3}(1 - o_j^{\mathrm{L}_3}) \tag{18}$$

$$= \delta_{\cdot j}^{\mathrm{L}_2} \tag{19}$$

偏移 $b_j^{\mathrm{L}_3}$ 的学习公式为：

$$\hat{b}_j^{\mathrm{L}_3} = b_j^{\mathrm{L}_3} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial b_j^{\mathrm{L}_3}} = b_j^{\mathrm{L}_3} - \eta \cdot \delta_{\cdot j}^{\mathrm{L}_2} \tag{20}$$

例如，对于权重 $\omega_{11}^{\mathrm{L}_2}$ 的更新学习公式为：

$$\widehat{\omega}_{11}^{\mathrm{L}_2} = \omega_{11}^{\mathrm{L}_2} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_2}} \tag{21}$$

$$= \omega_{11}^{\mathrm{L}_2} - \eta \cdot \delta_{11}^{\mathrm{L}_2} \cdot o_1^{\mathrm{L}_2} \tag{22}$$

$$= \omega_{11}^{\mathrm{L}_2} - \eta \cdot \big(-(o_1^{\mathrm{T}} - o_1^{\mathrm{L}_3}) \cdot o_1^{\mathrm{L}_3}(1 - o_1^{\mathrm{L}_3})\big) \cdot o_1^{\mathrm{L}_2} \tag{23}$$

$$= \omega_{11}^{\mathrm{L}_2} + \eta \cdot (o_1^{\mathrm{T}} - o_1^{\mathrm{L}_3}) \cdot o_1^{\mathrm{L}_3}(1 - o_1^{\mathrm{L}_3}) \cdot o_1^{\mathrm{L}_2} \tag{24}$$

例如，对于偏移 $b_1^{\mathrm{L}_3}$ 的更新学习公式为：

$$\hat{b}_1^{\mathrm{L}_3} = b_1^{\mathrm{L}_3} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial b_1^{\mathrm{L}_3}} \tag{25}$$

$$= b_1^{\mathrm{L}_3} + \eta \cdot (o_1^{\mathrm{T}} - o_1^{\mathrm{L}_3}) \cdot o_1^{\mathrm{L}_3}(1 - o_1^{\mathrm{L}_3}) \tag{26}$$

2 $Layer_1$ 与 $Layer_2$ 间参数调整

调整 $Layer_1$ 与 $Layer_2$ 间的权重 $\omega_{ij}^{\mathrm{L}_1}$

$$
\begin{aligned}
\frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_1}} &= (\sum_{k=1}^{2} \frac{\partial E_{\mathrm{Total}}}{\partial o_k^{\mathrm{L}_3}} \frac{\partial o_k^{\mathrm{L}_3}}{\partial net_k^{\mathrm{L}_3}} \frac{\partial net_k^{\mathrm{L}_3}}{\partial o_j^{\mathrm{L}_2}}) \cdot \frac{\partial o_j^{\mathrm{L}_2}}{\partial net_j^{\mathrm{L}_2}} \cdot \frac{\partial net_j^{\mathrm{L}_2}}{\partial \omega_{ij}^{\mathrm{L}_1}} && (27)\\
&= (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot f_j^{\mathrm{L}_2}(net_j^{\mathrm{L}_2})(1 - f_j^{\mathrm{L}_2}(net_j^{\mathrm{L}_2})) \cdot x_i && (28)\\
&= (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot o_j^{\mathrm{L}_2}(1 - o_j^{\mathrm{L}_2}) \cdot x_i && (29)\\
& && (30)
\end{aligned}
$$

令 $\delta_{ij}^{\mathrm{L}_1}$ 表示 $Layer_1$ 层第 i 个神经元与 $Layer_2$ 层第 j 个神经元间权重的梯度项，则

$$
\begin{aligned}
\delta_{ij}^{\mathrm{L}_1} &= (\sum_{k=1}^{2} \frac{\partial E_{\mathrm{Total}}}{\partial o_k^{\mathrm{L}_3}} \frac{\partial o_k^{\mathrm{L}_3}}{\partial net_k^{\mathrm{L}_3}} \frac{\partial net_k^{\mathrm{L}_3}}{\partial o_j^{\mathrm{L}_2}}) \cdot \frac{\partial o_j^{\mathrm{L}_2}}{\partial net_j^{\mathrm{L}_2}} && (31)\\
&= (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot o_j^{\mathrm{L}_2}(1 - o_j^{\mathrm{L}_2}) && (32)
\end{aligned}
$$

可以看出， $Layer_1$ 层第 i 个神经元与 $Layer_2$ 层第 j 个神经元间权重的梯度项是与 i 无关的，故连接至 $Layer_2$ 层第 j 个神经元相对应的权重的梯度向均为 $\delta_{ij}^{\mathrm{L}_1}$，用 $\delta_{\cdot j}^{\mathrm{L}_1}$ 表示，

$$
\delta_{\cdot j}^{\mathrm{L}_1} = \delta_{ij}^{\mathrm{L}_1} = (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot o_j^{\mathrm{L}_2}(1 - o_j^{\mathrm{L}_2}) \tag{33}
$$

则整体误差 E_{Total} 对 $\omega_{ij}^{\mathrm{L}_1}$ 的偏导数公式写作：

$$
\frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_1}} = \delta_{\cdot j}^{\mathrm{L}_1} \cdot x_i \tag{34}
$$

则权重 $\omega_{ij}^{\mathrm{L}_1}$ 的学习公式为：

$$
\widehat{\omega}_{ij}^{\mathrm{L}_1} = \omega_{ij}^{\mathrm{L}_1} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial \omega_{ij}^{\mathrm{L}_1}} = \omega_{ij}^{\mathrm{L}_1} - \eta \cdot \delta_{\cdot j}^{\mathrm{L}_1} \cdot x_i \tag{35}
$$

$$
\begin{aligned}
\frac{\partial E_{\mathrm{Total}}}{\partial b_j^{\mathrm{L}_2}} &= (\sum_{k=1}^{2} \frac{\partial E_{\mathrm{Total}}}{\partial o_k^{\mathrm{L}_3}} \frac{\partial o_k^{\mathrm{L}_3}}{\partial net_k^{\mathrm{L}_3}} \frac{\partial net_k^{\mathrm{L}_3}}{\partial o_j^{\mathrm{L}_2}}) \cdot \frac{\partial o_j^{\mathrm{L}_2}}{\partial net_j^{\mathrm{L}_2}} \cdot \frac{\partial net_j^{\mathrm{L}_2}}{\partial b_j^{\mathrm{L}_2}} && (36)\\
&= (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot f_j^{\mathrm{L}_2}(net_j^{\mathrm{L}_2})(1 - f_j^{\mathrm{L}_2}(net_j^{\mathrm{L}_2})) \cdot 1 && (37)\\
&= (\sum_{k=1}^{2} \delta_{\cdot k}^{\mathrm{Layer}_2} \omega_{3k}^{\mathrm{L}_2}) \cdot o_j^{\mathrm{L}_2}(1 - o_j^{\mathrm{L}_2}) && (38)\\
&= \delta_{\cdot j}^{\mathrm{L}_1} && (39)
\end{aligned}
$$

偏移 $b_j^{\mathrm{L}_2}$ 的学习公式为：

$$
\hat{b}_j^{\mathrm{L}_2} = b_j^{\mathrm{L}_2} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial b_j^{\mathrm{L}_2}} = b_j^{\mathrm{L}_2} - \eta \cdot \delta_{\cdot j}^{\mathrm{L}_1} \tag{40}
$$

例如，对于权重 $\omega_{23}^{\mathrm{L}_1}$ 的更新学习公式为：

$$
\begin{aligned}
\widehat{\omega}_{23}^{\mathrm{L}_1} &= \omega_{23}^{\mathrm{L}_2} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial \omega_{23}^{\mathrm{L}_1}} && (41)\\
&= \omega_{23}^{\mathrm{L}_2} - \eta \cdot \big((\sum_{j=1}^{2} \frac{\partial E_{\mathrm{Total}}}{\partial o_j^{\mathrm{L}_3}} \frac{\partial o_j^{\mathrm{L}_3}}{\partial net_j^{\mathrm{L}_3}} \frac{\partial net_j^{\mathrm{L}_3}}{\partial o_3^{\mathrm{L}_2}}) \cdot \frac{\partial o_3^{\mathrm{L}_2}}{\partial net_3^{\mathrm{L}_2}} \cdot \frac{\partial net_3^{\mathrm{L}_2}}{\partial \omega_{23}^{\mathrm{L}_1}} \big) && (42)\\
&= \omega_{23}^{\mathrm{L}_2} - \eta \cdot (\sum_{j=1}^{2} \delta_{\cdot j}^{\mathrm{Layer}_2} \omega_{3j}^{\mathrm{L}_2}) \cdot f_3^{\mathrm{L}_2}(net_3^{\mathrm{L}_2})(1 - f_3^{\mathrm{L}_2}(net_3^{\mathrm{L}_2})) \cdot x_2 && (43)\\
&= \omega_{23}^{\mathrm{L}_2} - \eta \cdot (\sum_{j=1}^{2} \delta_{\cdot j}^{\mathrm{Layer}_2} \omega_{3j}^{\mathrm{L}_2}) \cdot o_3^{\mathrm{L}_2}(1 - o_3^{\mathrm{L}_2}) \cdot x_2 && (44)
\end{aligned}
$$

例如，对于偏移 $b_3^{\mathrm{L}_2}$ 的更新学习公式为：

$$
\begin{aligned}
\hat{b}_3^{\mathrm{L}_2} &= b_1^{\mathrm{L}_3} - \eta \cdot \frac{\partial E_{\mathrm{Total}}}{\partial b_3^{\mathrm{L}_2}} && (45)\\
&= b_3^{\mathrm{L}_2} - \eta \cdot \big((\sum_{j=1}^{2} \frac{\partial E_{\mathrm{Total}}}{\partial o_j^{\mathrm{L}_3}} \frac{\partial o_j^{\mathrm{L}_3}}{\partial net_j^{\mathrm{L}_3}} \frac{\partial net_j^{\mathrm{L}_3}}{\partial o_3^{\mathrm{L}_2}}) \cdot \frac{\partial o_3^{\mathrm{L}_2}}{\partial net_3^{\mathrm{L}_2}} \cdot \frac{\partial net_3^{\mathrm{L}_2}}{\partial b_3^{\mathrm{L}_2}} \big) && (46)\\
&= b_3^{\mathrm{L}_2} - \eta \cdot (\sum_{j=1}^{2} \delta_{\cdot j}^{\mathrm{Layer}_2} \omega_{3j}^{\mathrm{L}_2}) \cdot f_3^{\mathrm{L}_2}(net_3^{\mathrm{L}_2})(1 - f_3^{\mathrm{L}_2}(net_3^{\mathrm{L}_2})) \cdot 1 && (47)\\
&= \omega_{23}^{\mathrm{L}_2} - \eta \cdot (\sum_{j=1}^{2} \delta_{\cdot j}^{\mathrm{Layer}_2} \omega_{3j}^{\mathrm{L}_2}) \cdot o_3^{\mathrm{L}_2}(1 - o_3^{\mathrm{L}_2}) && (48)
\end{aligned}
$$