1. 前向传播公式

输入到隐藏层的表示（如果有多层网络）：

$$z = X_{\text{aug}} \cdot W^T$$

- X_{aug}：包含偏置项的输入矩阵，形状 $(nd, nh+1)$。
- W：权重矩阵，形状 $(no, nh+1)$，每列表示输出层与隐藏层的连接权重。

---

2. Softmax 函数

Softmax 是输出层的激活函数，用来计算分类任务的概率分布：

$$\hat{y}_j = s(z_j) = \frac{\exp(z_j)}{\sum_{k=1}^{no} \exp(z_k)}$$

其中：

- z_j：第 j 类的线性组合（logits）。
- $\hat{y}_j$：第 j 类的概率，满足：

$$\sum_{j=1}^{no} \hat{y}_j = 1$$

---

3. 损失函数：交叉熵

交叉熵损失用于多分类问题，衡量预测 $\hat{y}$ 和真实标签 y 之间的差距：

$$J = -\sum_{d=1}^{nd} \sum_{j=1}^{no} y_{ij}^{(d)} \log(\hat{y}_j^{(d)})$$

- y_{ij}：真实标签，采用 one-hot 编码，1 表示样本属于对应类别。
- $\hat{y}_j$：预测的概率。

---

4. Softmax 的导数（雅可比矩阵）

Softmax 的导数是一个 $no \times no$ 的矩阵（即雅可比矩阵）：

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1 - \hat{y}_i) & \text{if } i = j \\ -\hat{y}_i \cdot \hat{y}_j & \text{if } i \neq j \end{cases}$$

矩阵形式表示为：

$$\frac{\partial \hat{y}}{\partial z} = \text{diag}(\hat{y}) - \hat{y} \cdot \hat{y}^T$$

---

5. 损失函数对 softmax 输出的导数

通过交叉熵损失函数，计算损失对 softmax 输出 $\hat{y}$ 的梯度：

$$\frac{\partial J}{\partial \hat{y}_j} = -\frac{y_j}{\hat{y}_j}$$

其中：

- y_j：真实标签（1 或 0）。
- $\hat{y}_j$：softmax 的输出概率。

---

6. 损失函数对 z 的梯度

通过链式法则，计算损失 J 对 z 的梯度：

$$\frac{\partial J}{\partial z_j} = \sum_{i=1}^{no} \frac{\partial J}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_j}$$

将公式展开：

$$\frac{\partial J}{\partial z_j} = \sum_{i=1}^{no} \left(-\frac{y_i}{\hat{y}_i} \right) \cdot (\hat{y}_i \cdot (\delta_{ij} - \hat{y}_j))$$

最终简化为：

$$\frac{\partial J}{\partial z_j} = \hat{y}_j - y_j$$

---

7. 损失函数对权重 w 的梯度

权重 w 是从隐藏层到输出层的连接权重，损失函数对权重 w 的梯度公式为：

$$\frac{\partial J}{\partial w_{ij}} = \sum_{d=1}^{nd} \frac{\partial J}{\partial z_j^{(d)}} \cdot \frac{\partial z_j^{(d)}}{\partial w_{ij}}$$

对于单个样本，展开为：

$$\frac{\partial J}{\partial w_{ij}} = \frac{\partial J}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}} = \frac{\partial J}{\partial z_j} \cdot x_i$$

其中：

- $\frac{\partial J}{\partial z_j} = \hat{y}_j - y_j$。
- x_i：隐藏层的输入。

---

8. 矩阵形式的总梯度

将权重的梯度累积到矩阵形式，使用矢量化表示：

$$\frac{\partial J}{\partial W} = X_{\text{aug}}^T \cdot \left(\frac{\partial J}{\partial Z} \right)$$

- X_{aug}：输入矩阵（包含偏置项）。
- $\frac{\partial J}{\partial Z} = \hat{Y} - Y$，即预测值与真实值的差异。

完整公式总结

1. **前向传播**：

$$z = X_{\text{aug}} \cdot W^T, \quad \hat{y} = \text{softmax}(z)$$

2. **损失函数**:

$$J = -\sum_{d=1}^{nd}\sum_{j=1}^{no} y_j^{(d)} \log(\hat{y}_j^{(d)})$$

3. **梯度计算**:

- $\frac{\partial J}{\partial \hat{y}} = -\frac{Y}{\hat{Y}}$
- $\frac{\partial J}{\partial Z} = \hat{Y} - Y$
- $\frac{\partial J}{\partial W} = X_{\text{aug}}^T \cdot \left(\frac{\partial J}{\partial Z}\right)$