在 Softmax 回归中，我们利用交叉熵损失函数来计算训练集上的经验风险，即：

$$\mathcal{R}(W) \ = -\frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}^{(n)})^{\mathrm{T}}\log\hat{\mathbf{y}}^{(n)}$$

在训练过程中，需要计算风险函数对 W 的梯度，计算公式如下：

$$\frac{\partial\mathcal{R}(W)}{\partial W} = -\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(n)}\left(\mathbf{y}^{(n)}-\hat{\mathbf{y}}^{(n)}\right)^{\mathrm{T}}$$

过程见下页↓

证：设有K个类别，则 $\hat{y}^{(n)}$ 和 $y^{(n)}$ 的形状为 $K\times 1$；设 $x^{(n)}$ 存有D个特征，则 $x^{(n)}$ 的形状为 $D\times 1$，则 W 为形状为 $D\times K$ 的矩阵.

令 $z^{(n)}=W^{T}\cdot x^{(n)} \Rightarrow z^{(n)}$ 为 $K\times 1$ 的列向量.

则 $\hat{y}_i^{(n)}=\dfrac{exp(z_i^{(n)})}{\sum_{j=1}^{K} exp(z_j^{(n)})}$

(1) 首先求 softmax 函数中 $\hat{y}_i^{(n)}$ 对 $z_m^{(n)}$ 的偏导.

①当 $m=i$ 时，有

$$\frac{\partial \hat{y}_i^{(n)}}{\partial z_m^{(n)}}=\frac{\partial \hat{y}_i^{(n)}}{\partial z_i^{(n)}}=\frac{exp(z_i^{(n)})\cdot\sum_{j=1}^{K} exp(z_j^{(n)})-exp(z_i^{(n)})\cdot exp(z_i^{(n)})}{\left(\sum_{j=1}^{K} exp(z_j^{(n)})\right)^2}$$

$$=\hat{y}_i^{(n)}-(\hat{y}_i^{(n)})^2$$

$$=\hat{y}_i^{(n)}(1-\hat{y}_i^{(n)})$$

②当 $m\neq i$ 时，有

$$\frac{\partial \hat{y}_i^{(n)}}{\partial z_m^{(n)}}=\frac{-exp(z_m^{(n)})\cdot exp(z_i^{(n)})}{\left(\sum_{j=1}^{K} exp(z_j^{(n)})\right)^2}=-\hat{y}_m^{(n)}\hat{y}_i^{(n)}$$

(2) 接着求交叉熵损失来对 $z_m^{(n)}$ 的偏导.

$$\cancel{\frac{\partial (y^{(n)})^T\cdot \log \hat{y}^{(n)}}{\partial z_m^{(n)}}}=\frac{\partial -(y^{(n)})^T\cdot \log \hat{y}^{(n)}}{\partial z_m^{(n)}}=\frac{\partial -\sum_{i=1}^{K} y_i^{(n)}\cdot \log \hat{y}_i^{(n)}}{\partial z_m^{(n)}}$$

$$=-\sum_{i=1}^{K} y_i^{(n)}\cdot\frac{1}{\hat{y}_i^{(n)}}\cdot\frac{\partial \hat{y}_i^{(n)}}{\partial z_m^{(n)}} \Rightarrow$$

$$\Rightarrow =-\left(y_m^{(n)}\cdot\frac{1}{\hat{y}_m^{(n)}}\cdot\frac{\partial \hat{y}_m^{(n)}}{\partial z_m^{(n)}}+\sum_{i\neq m}^{K} y_i^{(n)}\cdot\frac{1}{\hat{y}_i^{(n)}}\cdot\frac{\partial \hat{y}_i^{(n)}}{\partial z_m^{(n)}}\right)$$

$$=-\left[y_m^{(n)}\cdot(1-\hat{y}_m^{(n)})+\sum_{i\neq m}^{K} y_i^{(n)}\cdot(-\hat{y}_m^{(n)})\right]$$

$$=-\left[y_m^{(n)}-y_m^{(n)}\cdot\hat{y}_m^{(n)}-\hat{y}_m^{(n)}\cdot\sum_{i\neq m}^{K} y_i^{(n)}\right]=\hat{y}_m^{(n)}-y_m^{(n)}$$

(3) 最后求交叉熵损失对 w_{om} 的偏导，其中 $o=1,2\cdots D$, $m=1,2\cdots K$

$$\frac{\partial -(y^{(n)})^T\cdot \log \hat{y}^{(n)}}{\partial w_{om}} = \frac{\partial -(y^{(n)})^T\cdot \log \hat{y}^{(n)}}{\partial z_m} \cdot \frac{\partial z_m}{\partial w_{om}}$$

$$= (\hat{y}_m^{(n)} - y_m^{(n)}) \cdot \frac{\partial \left(\sum_{p=1}^{D} (W^T)_{mp}\, x_p^{(n)}\right)}{\partial w_{om}}$$

$$= (\hat{y}_m^{(n)} - y_m^{(n)}) \cdot x_o^{(n)}$$

$$\cancel{\Rightarrow \frac{\partial R(W)}{\partial W} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}}$$

⇒ 对 W 的梯度如下所示：

$$\begin{bmatrix} (\hat{y}_1^{(n)} - y_1^{(n)})\cdot \cancel{x_o}\, x_1^{(n)} & \cdots & (\hat{y}_K^{(n)} - y_K^{(n)})\cdot x_1^{(n)} \\ \vdots & & \vdots \\ (\hat{y}_1^{(n)} - y_1^{(n)})\cdot x_D^{(n)} & \cdots & (\hat{y}_K^{(n)} - y_K^{(n)})\cdot x_D^{(n)} \end{bmatrix}_{D\times K}$$

可写为 $x^{(n)}\cdot(\hat{y}^{(n)} - y^{(n)})^T$

~~对其中一个样本的~~ 第n个样本的风险函数对W的梯度为 $x^{(n)}\cdot(\hat{y}^{(n)} - y^{(n)})^T$

$$\Rightarrow \frac{\partial R(W)}{\partial W} = -\frac{1}{N}\sum_{n=1}^{N} x^{(n)}\cdot(\hat{y}^{(n)} - y^{(n)})^T$$