$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \quad \text{or equivalently} \quad L_i = -f_{y_i} + \log\sum_j e^{f_j}$$

where:

$f(x_i, W) = Wx_i$, f_j is score of jth class,

f_{y_i} is score of the correct class.

let $P_k = \frac{e^{f_k}}{\sum_j e^{f_j}}$, normalized prob for class k. $\therefore L_i = -\log(P_{y_i})$

$$\frac{\partial L_i}{\partial f_k} = \frac{\partial L_i}{\partial P_{y_i}} \cdot \frac{\partial P_{y_i}}{\partial f_k} \cdots ① \qquad \frac{\partial L_i}{\partial P_{y_i}} = [-\log(P_{y_i})]' = -\frac{1}{P_{y_i}} \cdots ②$$

$$\frac{\partial P_{y_i}}{\partial f_k} = \left[\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right]'$$

if $k = y_i$, $\frac{\partial P_{y_i}}{\partial f_k} = \left[\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right]' = \left(\frac{\sum_j e^{f_j} - \sum_{j=y_i} e^{f_j}}{\sum_j e^{f_j}}\right)' = \left(1 - \frac{\sum_{j=y_i} e^{f_j}}{\sum_j e^{f_j}}\right)'$

$$= -\sum_{j \neq y_i} e^{f_j} \cdot \left[-\frac{1}{(\sum_j e^{f_j})^2} \cdot e^{f_{y_i}}\right] = \frac{(\sum_j e^{f_j} - e^{f_{y_i}}) \cdot e^{f_{y_i}}}{(\sum_j e^{f_j})^2}$$

$$= \frac{\sum_j e^{f_j} - e^{f_{y_i}}}{\sum_j e^{f_j}} \cdot \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} = \frac{\sum_j e^{f_j} - e^{f_{y_i}}}{\sum_j e^{f_j}} \cdot P_{y_i} \cdots ③$$

put ② and ③ back in ①

$$\therefore \frac{\partial L_i}{\partial f_k} = -\frac{1}{P_{y_i}} \cdot \left(\frac{\sum_j e^{f_j} - e^{f_{y_i}}}{\sum_j e^{f_j}} \cdot P_{y_i}\right) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} - 1 = P_{y_i} - 1$$

if $k \neq y_i$, $\frac{\partial P_{y_i}}{\partial f_k} = e^{f_{y_i}} \cdot \left[-\frac{1}{(\sum_j e^{f_j})^2} \cdot e^{f_k}\right] = -P_{y_i} \cdot P_k \cdots ④$

put ② and ④ back in ①

$$\therefore \frac{\partial L_i}{\partial f_k} = -\frac{1}{P_{y_i}} \cdot (-P_{y_i} \cdot P_k) = P_k$$

$\therefore$ OVERALL: $\frac{\partial L_i}{\partial f_k} = P_k - 1(y_i = k)$

Also refer to section:

"put it together: Minimal Neural Network Case Study"