Softmax/Categorical CE Top Gradient

Softmax $\quad y_k = P(l=k, \vec{x})$

$$= \frac{e^{z_k}}{\sum_j e^{z_j}}$$

Note $\quad y_k = P(l=k, z_1, \ldots, z_K)$

Thus, y_k depends on $z_1, \ldots, z_K$.

$$\frac{\partial y_k}{\partial z_j} = \frac{\partial}{\partial z_j}\left(\frac{e^{z_k}}{e^{z_1} + \cdots + e^{z_K}}\right)$$

If $\boxed{j=k}$...

Two indices, so two cases ✓

$$\frac{\partial y_k}{\partial z_k} = \frac{e^{z_k} Z - (e^{z_k}) e^{z_k}}{(Z)^2}$$

$$= \frac{e^{z_k}}{Z} - \left(\frac{e^{z_k}}{Z}\right)^2 = y_k(1-y_k)$$

$Z = \sum_j e^{z_j}$

If $\boxed{j \neq k}$,

$$\frac{\partial y_k}{\partial z_j} = \frac{0 - (e^{z_k})(e^{z_j})}{(Z)^2} = -y_k y_j$$

Putting those together

✓

$$\boxed{\frac{\partial y_k}{\partial z_j} = y_k(\delta_{jk} - y_j)}$$

δ_{jk} is the Kronecker delta

$$\delta_{jk} = \begin{cases} 1 & \text{if } j=k \\ 0 & \text{if } j \neq k \end{cases}$$

For cross-entropy, $E(\vec{y},\vec{t}) = -\sum_{k=1}^{K} t_k \ln y_k$

$$\frac{\partial E}{\partial z_j} = \sum_{k=1}^{K} \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial z_j}$$

✓ $$\frac{\partial E}{\partial y_k} = \frac{-t_k}{y_k}$$

$$\therefore \frac{\partial E}{\partial z_j} = \sum_{k=1}^{K} \frac{-t_k}{\cancel{y_k}} \overbrace{\cancel{y_k}(\delta_{jk} - y_j)}$$

$$= -\sum_{k=1}^{K} t_k(\delta_{jk} - y_j)$$

Simplify ✓

$$= -(t_\gamma - y_j)$$

$$= y_j - t_\gamma$$

since $t_k = 1$ only for the correct class, which we call γ here.

Or, stated as a vector

$$\frac{\partial E}{\partial \vec{z}} = \vec{y} - \vec{t}$$

Alternative Solution

Softmax Gradient

We substitute the softmax activation function for the output node

$$y_k = \frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}}$$

inside the loss function

$$E(\vec{y},\vec{t}) = -\sum_{k}^{K} t_k ln(y_k)$$

to get

$$E(\vec{y},\vec{t}) = -\sum_{k}^{K} t_k ln\left(\frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}}\right)$$

and evaluate the gradient with respect to the input current to the output layer

$$\begin{aligned}
\frac{\partial E}{\partial z_j} &= \frac{\partial}{\partial z_j}\left[-\sum_{k}^{K} t_k ln\left(\frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}}\right)\right] \\
&= \frac{\partial}{\partial z_j}\left[-\sum_{k}^{K} t_k\left(z_k - ln\left(\sum_{i=1}^{K} e^{z_i}\right)\right)\right] \\
&= \frac{\partial}{\partial z_j}\left[\sum_{k}^{K} t_k\left(ln\left(\sum_{i=1}^{K} e^{z_i}\right) - z_k\right)\right] \\
&= \frac{\partial}{\partial z_j}\left[\sum_{k}^{K} t_k ln\left(\sum_{i=1}^{K} e^{z_i}\right) - t_k z_k\right] \\
&= \frac{\partial}{\partial z_j}\left[\sum_{k}^{K} t_k ln\left(\sum_{i=1}^{K} e^{z_i}\right)\right] - t_j \\
&= \sum_{k}^{K}\left(\frac{t_k e^{z_j}}{\sum_{i=1}^{K} e^{z_i}}\right) - t_j \\
&= \sum_{k} t_k y_j - t_j \\
&= y_j\left(\sum_{k} t_k\right) - t_j \\
&= y_j - t_j \\
&= \begin{cases} y_j & \text{if } j \neq \gamma \\ y_j - 1 & \text{if } j = \gamma \end{cases}
\end{aligned}$$

As a vector: $\frac{\partial E}{\partial \vec{z}} = \vec{y} - \vec{t}$.