This note does the necessary computations to show that the softmax loss gradient is "$p-1$ at the correct class and $p$ at all other classes".

Suppose that $x \in \mathbb{R}^d$ is a training example that receives score $s \in \mathbb{R}^c$ from the network, while the correct answer is some $y \in \{1, \dots, c\}$. The *softmax loss* of this training example is defined as

$$\ell(s) = -\log p_y,$$

where we write $p_i := \dfrac{e^{s_i}}{\Sigma}$ for each $1 \leq i \leq c$ with $\Sigma := \sum_{i=1}^{c} e^{s_i}$.

We want to compute $\dfrac{\partial \ell}{\partial s}$. By the chain rule and the fact that the derivative of $\log z$ is $\dfrac{1}{z}$, we have

$$\frac{\partial \ell}{\partial s} = -\frac{1}{p_y} \cdot \frac{\partial p_y}{\partial s}. \tag{1}$$

Now $\frac{\partial p_y}{\partial s}$ is a vector in $\mathbb{R}^c$, which we compute with the quotient rule for derivatives, distinguishing the coordinate $y$ from the other coordinates with $i \neq y$:

$$\begin{aligned}
\left(\frac{\partial p_y}{\partial s}\right)_y &= \frac{\Sigma \cdot e^{s_y} - e^{s_y} \cdot e^{s_y}}{\Sigma^2} = p_y \frac{\Sigma - e^{s_y}}{\Sigma} = p_y(1 - p_y), \\
\left(\frac{\partial p_y}{\partial s}\right)_i &= \frac{-e^{s_y} e^{s_i}}{\Sigma^2} = -p_y p_i \text{ for every } i \neq y.
\end{aligned}$$

If we plug this into (1), we get

$$\begin{aligned}
\left(\frac{\partial \ell}{\partial s}\right)_y &= -\frac{1}{p_y} \cdot p_y(1 - p_y) = p_y - 1, \\
\left(\frac{\partial \ell}{\partial s}\right)_i &= -\frac{1}{p_y} \cdot -p_y p_i = p_i \text{ for every } i \neq y.
\end{aligned}$$