By Kevin Swersky.

This is more mathematical than most of the course, but see if you can follow it.

Assume that the true class is j so the target t is zero everywhere except for a 1 in the j^{th} index. The cost function for this is going to be:

$C = -\sum_i t_i \log \frac{\exp w_i^T x}{\sum_k \exp(w_k^T x)}$

So when we actually put in the t I mentioned above, this simplifies to:

$C = -\log \frac{\exp w_j^T x}{\sum_k \exp(w_k^T x)} = -w_j^T x + \log\left(\sum_k \exp(w_k^T x)\right)$

Now suppose we look at some specific dimension of x. For example, the first dimension. Let's say that the first element of x is $x1$.

What happens if we differentiate the cross-entropy with respect to $w1j$? Remember, w is a matrix here so this is the parameter from the first dimension of the input to the j^{th} output neuron.

The derivative is:

$$\begin{aligned}
\frac{\partial C}{\partial w_{1j}} &= -x_1 + \frac{1}{\sum_k \exp(w_k^T x)}\left(\exp(w_j^T x) x_1\right) \\
&= \left(-1 + \frac{\exp(w_j^T x)}{\sum_k \exp(w_k^T x)}\right) x_1 \\
&= \left(P(\text{x belongs to class j}) - 1\right) x_1
\end{aligned}$$

Notice how a) the softmax appears in the derivative and b) the derivative is 0 when the probability under the softmax is the same as the target? So the derivative is trying to change the weights to match the neural network probabilities to the target probabilities.

Alright, we're not quite done. Let's take some other class $i \neq j$ and try its derivative as well:

$$\begin{aligned}
\frac{\partial C}{\partial w_{1i}} &= 0 + \frac{1}{\sum_k \exp(w_k^T x)}\left(\exp(w_i^T x) x_1\right) \\
&= \left(0 + \frac{\exp(w_i^T x)}{\sum_k \exp(w_k^T x)}\right) x_1 \\
&= \left(P(\text{x belongs to class i}) - 0\right) x_1
\end{aligned}$$

I left the 0 in there so that it's completely clear what this is doing. Again, the derivative is 0 when the probability under the network matches the target. So for the second case, it's trying to make the probability 0 since the target is 0.

Hopefully you can see from this that the cross-entropy does indeed seem to be doing the right thing.