Answer: To find $\frac{\partial}{\partial z_j}(-\sum y_k \log \hat{y}_k) = (*)$ we note that $\hat{y}_k = \text{softmax}(\hat{y})_k = \frac{e^{z_k}}{\sum_i e^{z_i}}$, and so $\frac{\partial}{\partial z_j}\hat{y}_k = \frac{e^{z_k}(\sum_i e^{z_i})\delta_{jk} - e^{z_j}e^{z_k}}{(\sum_i e^{z_i})^2} = \hat{y}_k\delta_{jk} - \hat{y}_j\hat{y}_k$. So then $(*) = -\sum_k (y_k/\hat{y}_k)(\hat{y}_k\delta_{jk} - \hat{y}_j\hat{y}_k) = \hat{y}_j - y_j$. I guess this sort of makes sense. In gradient descent we'll try to decrease cross entropy loss, i.e. stepping along $-\nabla_z = y - \hat{y}$, i.e. seeking to make the final pre-softmax layer increase/decrease s.t. $\hat{y}$ is more like y