Question 2: Softmax/Categorical CE

[5 marks]

Consider a classification problem in which you have K classes. Suppose you have a labelled dataset containing pairs of inputs and class labels, $(\mathbf{x}, \ell)$, where $\mathbf{x} \in \mathbb{R}^X$ and $\ell \in \{1, 2, \ldots, K\}$.

Your neural network's output is a classification vector based on the softmax activation function, so that if z_k is the input current for output node k, then the activation of output node y_k is

$$y_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \quad , \quad k = 1, \ldots, K \, .$$

Thus, $\mathbf{y} \in [0, 1]^K$, and $y_k = P\left(k = \ell \,|\, \mathbf{x}\right)$.

Suppose that your loss function is categorical cross-entropy,

$$L(\mathbf{y}, \mathbf{t}\,) = -\sum_{k=1}^{K} t_k \ln y_k \, ,$$

where $\mathbf{t}$ is the one-hot indicator vector for class ℓ, so that $t_k = \delta_{k\ell}$ (Kronecker delta). Derive the simplest expression you can for $\nabla_{\mathbf{z}} L$, the gradient of the loss function with respect to the input currents to the output layer.

Make sure your derivation is organized, and explain your steps.

Submitting Mathematical Derivations

You will submit your solution as a PDF to Kritik. In preparing your solution, you may:

- typeset your answer in a word-processing application, like Word or LaTeX,
- handwrite your answer on a tablet computer, or
- handwrite your answer on paper, and take a photo or scan.

No matter which option you choose, it is your responsibility to ensure the PDF you submit is of sufficient quality that others can reasonably understand your work.