

(a) Derive the gradient with respect to the input of a softmax function when cross entropy loss is used for evaluation, i.e., find the gradients with respect to the softmax input vector θ , when the prediction is made by $\hat{y} = \text{softmax}(\theta)$. Cross entropy and softmax are defined as:

$$\text{CE}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i y_i \cdot \log(\hat{y}_i)$$

$$\text{softmax}(\theta)_i = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$$

The gold vector \mathbf{y} is a one-hot vector, and the predicted vector $\hat{\mathbf{y}}$ is a probability distribution over the output space.