1.

(a) Recall the softmax function $y_k = \dfrac{\exp(a_k)}{\sum_j \exp(a_j)}$ Find $\dfrac{\partial y_k}{\partial a_j}$?

If $k = j$,

$$\frac{\partial y_k}{\partial a_k} = \frac{\exp(a_k) \cdot \sum_j \exp(a_j) - \exp(a_k) \cdot \exp(a_k)}{\left(\sum_j \exp(a_j)\right)^2}$$

$$= \frac{\exp(a_k)}{\sum_j \exp(a_j)} - \left[\frac{\exp(a_k)}{\sum_j \exp(a_j)}\right]^2 = y_k - y_k^2$$

If $k \neq j$,

$$\frac{\partial y_k}{\partial a_j} = \frac{-\exp(a_k) \cdot \exp(a_j)}{\left(\sum_j \exp(a_j)\right)^2} = -\left(\frac{\exp(a_k)}{\sum_j \exp(a_j)}\right)\left(\frac{\exp(a_j)}{\sum_j \exp(a_j)}\right) = -y_k y_j$$

Observe that

$$y_k(\delta_{kj} - y_j) = \begin{cases} y_k(1 - y_k) & \text{if } k = j \\ -y_k y_j & \text{if } k \neq j \end{cases}$$

Hence, $\dfrac{\partial y_k}{\partial a_k} = y_k(\delta_{kj} - y_j)$

12

(b). Recall the cross-entroy error function is

$$E W = - \sum_{i=1}^{n} \sum_{k=1}^{c} t_{ik} \ln y_{ik}$$

Since $y_k(\vec{x}_i) = \frac{\exp(a_k)}{\sum_j \exp(a_{ij})}$, from prob. (a), $\frac{\partial y_{ik}}{\partial a_{ij}} = y_{ik}(\delta_{kj} - y_{ij})$

Differentiating EW w.r.t a_{ij} using chain rule,

$$\frac{\partial E}{\partial a_{ij}} = \sum_k \frac{\partial y_{ik}}{\partial a_{ij}} \frac{\partial E}{\partial y_{ik}} = -\sum_k \cancel{y_{ik}}(\delta_{kj} - y_{ij}) \times \frac{t_{ik}}{\cancel{y_{ik}}}$$

$$= - \sum_k \delta_{kj} t_{ik} + \sum_k y_{ij} t_{ik}$$

$$= - t_{ij} + y_{ij} \sum_k t_{ik} = y_{ij} - t_{ij} \quad \left(\because \vec{t}_i = (0,0,\cdots 0,1,0,0,\cdots 0) \Rightarrow \sum_k t_{ik} = 1 \right)$$

Differentiating w.r.t. w_j using chain rule again,

$$\frac{\partial E}{\partial w_j} = \sum_i \frac{\partial a_{ij}}{\partial w_j} \frac{\partial E}{\partial a_{ij}} = \sum_{i=1}^{n} \vec{x}_i (y_{ij} - t_{ij}) \quad \left(\because a_{ij} = w_j^T \cdot \vec{x}_i \quad \frac{\partial a_{ij}}{\partial w_j} = \vec{x}_i \right)$$

So the <u>batch learning rule</u> is

$$\boxed{\vec{w}_j \leftarrow \vec{w}_j + \eta \sum_{i=1}^{n} (y_{ij} - t_{ij}) \vec{x}_i}$$

And the <u>single-sample rule</u> is

$$\boxed{\vec{w}_j \leftarrow \vec{w}_j + \eta (y_{ij} - t_{ij}) \vec{x}_i}$$

where η is small enough.