Task 1.1

Over a single data point we get:

$$-\mathcal{E}(w) = t^n \ln(y^n) + (1-t^n)\ln(1-y^n)$$

using $y^n = g_w^n$ and $\frac{\partial g_w^n}{\partial w_j} = x_j^n g_w^n (1-g_w^n)$

allows the usage of the chain rule to find $\frac{\partial -\mathcal{E}^n(w)}{\partial w_j}$

$$\frac{\partial(t^n \ln(g_w^n))}{\partial w_j} = \frac{t^n x^n \cancel{g_w^n}(1-g_w^n)}{\cancel{g_w^n}} \rightarrow \text{example chain rule}$$

$$\frac{\partial((1-t^n)\ln(1-g_w^n))}{\partial w_j} = -\frac{(1-t^n)}{\cancel{1-g_w^n}}(x^n g_w^n)\cancel{(1-g_w^n)}$$

$\Rightarrow$

$$\begin{aligned}
\frac{-\partial \mathcal{E}^n(w)}{\partial w_j} &= t^n x_j^n (1-g_w^n) - (1-t^n)(x_j^n g_w^n) \\
&= t^n x_j^n - t^n x_j^n g_w^n - (x_j^n g_w^n - t^n x_j^n g_w^n) \\
&= t^n x_j^n \cancel{- t^n x_j^n g_w^n} - x_j^n g_w^n \cancel{+ t^n x_j^n g_w^n} \\
&= t^n x_j^n - x_j^n g_w^n = (t^n - g_w^n) x_j^n \\
&= \underline{(t^n - y^n) x_j^n} \quad \blacksquare
\end{aligned}$$

Scanned by CamScanner

Task 1.2

$$y_k^n = \frac{e^{w_k^T x^n}}{\sum_{k'} e^{w_{k'}^T x^n}}$$

Using the quotient rule:

$$\frac{\partial y_k^n}{\partial w_j} = \frac{\frac{\partial}{\partial w_j} e^{w_k^T x^n} \cancel{\sum_{k'} e^{w_{k'}^T x^n}}}{(\sum_{k'} e^{w_{k'}^T x^n})^{\cancel{2}}} - \frac{e^{w_k^T x^n}}{(\sum_{k'} e^{w_k^T x^n})^2} \frac{\partial}{\partial w_j} \sum e^{w_k^T x^n}$$

Now: we split this into parts of $k=j$ and $k \neq j$, such that we differ between whether the weights are inputs to the wanted output or not:

mind this j!

$$\boxed{k=j:} \Rightarrow \frac{x^n e^{w_k^T x^n}}{\sum_{k'} e^{w_{k'}^T x^n}} - \frac{e^{w_k^T x^n}}{(\sum_{k'} e^{w_{k'}^T x^n})^2} \cdot x^n e^{w_j^T x^n}$$

$$\Rightarrow \frac{\partial y_k^n}{\partial w_j} = x^n y_k^n - x^n y_k^n y_j = \underline{x^n y_k^n (1 - y_j^n)}$$

$$\boxed{k \neq j}: \quad 0 - x^n y_k y_j = \frac{\partial y_k^n}{\partial w_j}$$ (by the same method as above for $k=j$)

With the derivative of softmax, we may obtain the gradient of the loss function:

$$\frac{\partial E}{\partial w_{kj}} = -\sum_k^C t_k \ln(y_k^n) = -\sum_{k=1}^{C} \frac{t_k}{y_k^n} \frac{\partial y_k^n}{\partial w_{kj}}$$ (move the minus sign)

Kronecker delta.

$$\frac{\partial y_k^n}{\partial w_{kj}} = \begin{cases} x^n y_k^n (1 - y_j^n), & k=j \\ -x^n y_k^n y_j^n, & k \neq j \end{cases} \Rightarrow y_k^n x^n (\delta_{kj} - y_j^n)$$

(we are using the weight w_{kj}, making $j=k$ here, not to be confused with the $j=k$ from the derivation)

$$\text{Thus: } -\frac{\partial E}{\partial w_{kj}} = \sum_{k=1}^{C} \frac{t_k}{\cancel{y_k^n}} \cancel{y_k^n} (\delta_{kj} - y_k^n)$$

$$= \sum_{k=1}^{C} t_k x^n \delta_{kj} - \sum_{k=1}^{C} t_k x^n y_j = x^n (t_k - y_k^n)$$

Task 2.2 a)

$$J(w) = \mathcal{E}(w) + \lambda C(w)$$

$$\frac{J(w)}{\partial w} = \frac{\mathcal{E}(w)}{\partial (w)} + \frac{\lambda C(w)}{\partial w}$$

↳ given from previous task

$$\frac{\lambda C(w)}{\partial w} = \frac{\lambda \left(\lambda \sum_{i,j} w_{i,j}^2\right)}{\partial w} = \underline{\lambda 2 w}$$

(The summation disappears, as we are differentiating over one specific weight, whilst the others remain constant).

I've also seen $\frac{\lambda}{2n} C(w)$ used, which gives:

$$\frac{\partial \frac{\lambda}{2n} \sum w^2}{\partial w} = \underline{\frac{\lambda}{n} w}.$$