4a) $\frac{\partial E}{\partial z_k^n} = y_k^{(n)} - t_k^{(n)}$

$$\frac{\partial E}{\partial z_k^n} = \sum_j \overset{(\mathrm{I})}{\frac{\partial E}{\partial y_j^n}} \cdot \overset{(\mathrm{II})}{\frac{\partial y_j^n}{\partial z_k^n}}$$ by Chain Rule

(I) $\frac{\partial E}{\partial y_j^n} = \frac{\partial}{\partial y_j^n}\left(-\sum_k t_k^n \log(y_k^n)\right)$ by def of E

$= -\sum_k \frac{\partial t_k^n \log(y_k^n)}{y_j^n}$

$= -\sum_k t_k^n \cdot \frac{1}{y_j^n} \delta(j,k)$

$= -\frac{1}{y_j^n} \sum_k t_k^n \delta(j,k)$ by property of delta

$= -\frac{t_j^n}{y_j^n}$

(II) $\frac{\partial y_j^n}{\partial z_k^n} = \delta(k,j) \cdot y_j^n - y_j^n y_k^n$ by the third last equation

$\frac{\partial E}{\partial z_k^n} = \sum_j -\frac{t_j^n}{y_j^n} \cdot \left(\delta(k,j) y_j^n - y_j^n y_k^n\right)$ Implement (I) and (II) in equation

$= \sum_j -t_j^n \delta(k,j) + t_j^n \cdot y_k^n$ by simplifing y_j^n

$= -t_k^n + \sum_j t_j^n \cdot y_k^n$ by property of delta

$= -t_k^n + y_k^n \cancel{\sum_j t_j^n} \to 1$ by 1-of-k

$$\boxed{\frac{\partial E}{\partial z_k^n} = y_k^n - t_k^n}$$

4b) $\frac{\partial E}{\partial w} = x^T(Y-T)$

$$\frac{\partial E}{\partial w} = \sum_k \frac{\partial E}{\partial w_k}$$

$$= \sum_k \sum_i \frac{\partial E}{\partial w_{ki}}$$

$$= \sum_k \sum_i \sum_n [y_k^n - t_k^n]\, x_i^n \qquad \text{by def of } \frac{\partial E}{\partial w_{ki}}$$

$$= \sum_k \sum_i [Y_k - T_k]\, X_i$$

$$= \sum_k \sum_i X_i^T [Y_k - T_k] \qquad \text{by law of transpose}$$

$$= \sum_k X^T [Y_k - T_k]$$

$$\boxed{\frac{\partial E}{\partial w} = X^T [Y-T]}$$

4c) $\frac{\partial E}{\partial w_{k0}} = \sum_k [y_k^n - t_k^n]$

This is true because for an X_0, it wouldn't affect the bias of our function. Since the bias W_0 is independent to whatever X is. Therefore at $i=0$, the X wouldn't affect W_0.

5b)

(i) Overflow

$$\text{softmax}(z') = \frac{e^{z'_k}}{\sum_j e^{z'_j}}$$

$$= \frac{e^{z_k - \max(z)}}{\sum e^{z_j - \max(z)}} = \frac{e^{z_k} \cdot e^{-\max(z)}}{\sum e^{z_j} \cdot e^{-\max(z)}}$$

$$= \frac{e^{z_k} \cdot \cancel{e^{-\max(z)}}}{\cancel{e^{-\max(z)}} \cdot \sum_j e^{z_j}} = \frac{e^{z_k}}{\sum_j e^{z_j}} = \text{softmax}(z)$$

A function can overflow in three scenarios

- Numerator overflows: in this case it is impossible since we substract the $\max(z)$ from z_k the highest possible value is $z_k - \max(z) \leq 0$.

- Denominator overflows: it is also impossible since $z_k - \max(z) \leq 0$, so it won't have numbers that overflow.

- The division overflows: can't happend since the denominator will always be greater then the numerator since the sum of all z is greater then your z_k so it will always be a fraction.

(ii) Underflow

$$\log(y_k) = \log\left(\frac{e^{z_k}}{\sum_j e^{z_j}}\right) = \log(e^{z_k}) - \log\left(\sum_j e^{z_j}\right)$$ by log division property

$$= z_k - \log\left(\sum_j e^{z_j}\right) = z_k - \log\left(\sum_j e^{z_j} \cdot \frac{e^{\max(z)}}{e^{\max(z)}}\right)$$

$$= z_k - \max(z) - \log\left(\sum_j e^{z_j - \max(z)}\right)$$

$\sum_j e^{z_j - \max(z)}$ has a max value of e^0, so the log of these values will be large negative number

2f)

$$P(C=1|x) = \frac{e^{-(w^T x + w_0)}}{1 + e^{-(w^T x + w_0)}}$$

Since $C=1$, $w^T x + w_0 = 1$

$$\frac{e^{-1}}{1 + e^{-1}} = \boxed{0.268}$$