Question 1

(a)

Since $\sigma_i(\boldsymbol{x}) > 0$ for all x, i, and its value is dependent on all other $j \neq i$, every entry of the Jacobian matrix is non-zero.

In order to get rid of the exponential functions, we can try to use the logarithm of the softmax function:

$$\begin{aligned}\frac{\partial \log \sigma_i(\boldsymbol{x})}{\partial x_j} &= \frac{1}{\sigma_i(\boldsymbol{x})}\frac{\partial \sigma_i(\boldsymbol{x})}{\partial x_j} \\ \frac{\partial \sigma_i(\boldsymbol{x})}{\partial x_j} &= \sigma_i(\boldsymbol{x})\frac{\partial \log \sigma_i(\boldsymbol{x})}{\partial x_j}\end{aligned}$$

$$\begin{aligned}\log \sigma_i(\boldsymbol{x}) &= \log\left(\frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)}\right) \\ &= x_i - \log\left(\sum_{j=1}^n \exp(x_j)\right)\end{aligned}$$

In the following we use:

$$\frac{\partial x_i}{\partial x_z} = \begin{cases} 1 & \text{if } i = z \\ 0 & \text{if } i \neq z \end{cases}$$

$$\begin{aligned}\frac{\partial \log \sigma_i(\boldsymbol{x})}{\partial x_j} &= \frac{\partial x_i}{\partial x_j} - \frac{\partial \log\left(\sum_{j=1}^n \exp(x_j)\right)}{\partial x_j} \\ &= \mathbb{1}_{i=j} - \frac{\partial \log\left(\sum_{j=1}^n \exp(x_j)\right)}{\partial x_j} \\ &= \mathbb{1}_{i=j} - \frac{1}{\sum_{j=1}^n \exp(x_j)}\left(\frac{\partial}{\partial x_j}\sum_{j=1}^n \exp(x_j)\right) \\ &= \mathbb{1}_{i=j} - \frac{\exp(x_j)}{\sum_{j=1}^n \exp(x_j)} \\ &= \mathbb{1}_{i=j} - \sigma_j(\boldsymbol{x})\end{aligned}$$

Finally, convert back to the original derivative:

$$\begin{aligned}\frac{\partial \sigma_i(\boldsymbol{x})}{\partial x_j} &= \sigma_i(\boldsymbol{x})\frac{\partial \log \sigma_i(\boldsymbol{x})}{\partial x_j} \\ &= \sigma_i(\boldsymbol{x}) \cdot (\mathbb{1}_{i=j} - \sigma_j(\boldsymbol{x}))\end{aligned}$$

$$\begin{aligned} D_{\boldsymbol{x}}\sigma(\boldsymbol{x}) &= \begin{pmatrix} \frac{\partial\sigma_1(\boldsymbol{x})}{\partial x_1} & \dots & \frac{\partial\sigma_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial\sigma_n(\boldsymbol{x})}{\partial x_1} & \dots & \frac{\partial\sigma_n(\boldsymbol{x})}{\partial x_n} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1(\boldsymbol{x})\cdot(1-\sigma_1(\boldsymbol{x})) & \dots & -\sigma_1(\boldsymbol{x})\cdot\sigma_n(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ -\sigma_n(\boldsymbol{x})\cdot\sigma_1(\boldsymbol{x}) & \dots & \sigma_n(\boldsymbol{x})\cdot(1-\sigma_n(\boldsymbol{x})) \end{pmatrix} \end{aligned}$$

So especially, the diagonal entries are:

$$\frac{\partial\sigma_i(\boldsymbol{x})}{\partial x_i} = \sigma_i(\boldsymbol{x})\cdot(1-\sigma_i(\boldsymbol{x}))$$

And the off-diagonal entries are:

$$\frac{\partial\sigma_i}{\partial x_j} = -\sigma_i(\boldsymbol{x})\cdot\sigma_j(\boldsymbol{x})$$

And the matrix is symmetric. Thus:

$$\begin{aligned} \frac{\partial\sigma_i}{\partial x_j} &= -\sigma_i(\boldsymbol{x})\cdot\sigma_j(\boldsymbol{x}) \\ &= -\sigma_j(\boldsymbol{x})\cdot\sigma_i(\boldsymbol{x}) \\ &= \frac{\partial\sigma_j}{\partial x_i} \end{aligned}$$

(b)

$$\begin{aligned} \boldsymbol{z} &= \boldsymbol{v}\cdot D_{\boldsymbol{x}}\sigma(\boldsymbol{x}) \\ &= (v_1\dots v_n)\cdot \begin{pmatrix} \sigma_1(\boldsymbol{x})\cdot(1-\sigma_1(\boldsymbol{x})) & \dots & -\sigma_1(\boldsymbol{x})\cdot\sigma_n(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ -\sigma_n(\boldsymbol{x})\cdot\sigma_1(\boldsymbol{x}) & \dots & \sigma_n(\boldsymbol{x})\cdot(1-\sigma_n(\boldsymbol{x})) \end{pmatrix} \\ &= \begin{pmatrix} v_1\cdot\sigma_1(\boldsymbol{x})\cdot(1-\sigma_1(\boldsymbol{x})) + \dots + v_n\cdot -\sigma_n(\boldsymbol{x})\cdot\sigma_1(\boldsymbol{x}) \\ \vdots \\ v_1\cdot -\sigma_1(\boldsymbol{x})\cdot\sigma_n(\boldsymbol{x}) + \dots + v_n\cdot\sigma_n(\boldsymbol{x})\cdot(1-\sigma_n(\boldsymbol{x})) \end{pmatrix}^\top \\ &= \begin{pmatrix} \sigma_1(\boldsymbol{x})\cdot(v_1\cdot(1-\sigma_1(\boldsymbol{x})) - v_2\cdot\sigma_2(\boldsymbol{x}) - \dots - v_n\cdot\sigma_n(\boldsymbol{x})) \\ \vdots \\ \sigma_n(\boldsymbol{x})\cdot(v_1\cdot\sigma_1(\boldsymbol{x}) - v_{n-1}\cdot\sigma_{n-1}(\boldsymbol{x}) - \dots + v_n\cdot(1-\sigma_n(\boldsymbol{x}))) \end{pmatrix}^\top \\ &= \begin{pmatrix} \sigma_1(\boldsymbol{x})\cdot(v_1 - v_1\cdot\sigma_1(\boldsymbol{x}) - v_2\cdot\sigma_2(\boldsymbol{x}) - \dots - v_n\cdot\sigma_n(\boldsymbol{x})) \\ \vdots \\ \sigma_n(\boldsymbol{x})\cdot(v_1\cdot\sigma_1(\boldsymbol{x}) - v_{n-1}\cdot\sigma_{n-1}(\boldsymbol{x}) - \dots + v_n - v_n\cdot\sigma_n(\boldsymbol{x})) \end{pmatrix}^\top \\ &= \begin{pmatrix} \sigma_1(\boldsymbol{x})\cdot(v_1 - \boldsymbol{v}\cdot\sigma(\boldsymbol{x})^\top) \\ \vdots \\ \sigma_n(\boldsymbol{x})\cdot(v_n - \boldsymbol{v}\cdot\sigma(\boldsymbol{x})^\top) \end{pmatrix}^\top \end{aligned}$$

(c)

$$\begin{aligned}
\frac{\partial l(\boldsymbol{z}, \boldsymbol{t})}{z_j} &= -\frac{\partial}{\partial z_j} \sum_{i=1}^{n} t_i \cdot \log(z_i) \\
&= -\sum_{i=1}^{n} t_i \cdot \frac{\partial}{\partial z_j} \log(z_i) \\
&= -\sum_{i=1}^{n} \frac{t_i}{z_i} \cdot \frac{\partial z_i}{\partial z_j} \\
&= -\frac{t_i}{z_i} \cdot \frac{\partial}{\partial z_j} \sum_{i=1}^{n} z_i \\
&= -\frac{t_i}{z_i}
\end{aligned}$$

$$D_{\boldsymbol{z}} l(\boldsymbol{z}, \boldsymbol{t}) = \begin{pmatrix} -\frac{t_1}{z_1} \\ \vdots \\ -\frac{t_n}{z_n} \end{pmatrix}^\top$$

(d)

If one of the terms $z_i = 0$, then $D_{\boldsymbol{z}} l(\boldsymbol{z}, \boldsymbol{t})$ is not computable.

We have already observed that $\sigma_i(\boldsymbol{x}) > 0$. Thus, $z_i = 0$ can only occur when the following is satisfied:

$$v_j = \boldsymbol{v} \cdot \sigma(\boldsymbol{x})^\top$$

Question 3

(a)

We are sure that we should prove $\frac{\partial}{\partial x}\tanh(x) = 1 - \tanh(x)^2 = 1 - \tanh(x) \cdot \tanh(x)$ *and not* $\frac{\partial}{\partial x}\tanh(x) = 1 - \tanh^2(x) = 1 - (\tanh \circ \tanh)(x) = 1 - \tanh(\tanh(x))$.

$$\begin{aligned}
\frac{\partial}{\partial x}\tanh(x) &= \frac{\partial}{\partial x}\frac{e^x - e^{-x}}{e^x + e^{-x}} \\
&= \frac{e^x + e^{-x}}{(e^x + e^{-x})^2} \cdot \frac{\partial}{\partial x}(e^x - e^{-x}) - \frac{e^x - e^{-x}}{(e^x + e^{-x})^2} \cdot \frac{\partial}{\partial x}(e^x + e^{-x}) \\
&= \frac{e^x + e^{-x}}{(e^x + e^{-x})^2} \cdot (e^x + e^{-x}) - \frac{e^x - e^{-x}}{(e^x + e^{-x})^2} \cdot (e^x - e^{-x}) \\
&= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - \tanh(x)^2
\end{aligned}$$