Question 1. Classification with Linear Regression using Gradient Descent

a)

$$\begin{aligned}
\mathcal{L}(\mathbf{W}, \mathbf{b}, \mathbf{y}, \mathbf{x}) &= (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \\
&= (\mathbf{y} - (\mathbf{W}\mathbf{x} + \mathbf{b}))^T(\mathbf{y} - (\mathbf{W}\mathbf{x} + \mathbf{b}))
\end{aligned}$$

$$\begin{aligned}
\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}, \mathbf{b}, \mathbf{y}, \mathbf{x}) &= -\mathbf{x}^T(\mathbf{y} - (\mathbf{W}\mathbf{x} + \mathbf{b})) - \mathbf{x}(\mathbf{y} - (\mathbf{W}\mathbf{x} + \mathbf{b}))^T && \text{(by chain rule)} \\
&= -\mathbf{x}^T(\mathbf{y} - \hat{\mathbf{y}}) - x(\mathbf{y} - \hat{\mathbf{y}})^T \\
&= \mathbf{x}^T(\hat{\mathbf{y}} - \mathbf{y}) + \mathbf{x}(\hat{\mathbf{y}} - \mathbf{y})^T \\
&= 2(\hat{\mathbf{y}} - \mathbf{y})x^T = 2(\hat{\mathbf{y}} - \mathbf{y}) \otimes x
\end{aligned}$$

$$\begin{aligned}
\nabla_{\mathbf{b}}\mathcal{L}(\mathbf{W}, \mathbf{b}, \mathbf{y}, \mathbf{x}) &= (\mathbf{y} - \hat{\mathbf{y}}) + (\mathbf{y} - \hat{\mathbf{y}})^T && \text{(by chain rule)} \\
&= 2(\mathbf{y} - \hat{\mathbf{y}})
\end{aligned}$$

b) For n inputs the derivatives will change to the summation over all n inputs.

c) See the attached files.

d) The best accuracy and loss were 0.86 and 0.39, respectively, for learning rate 0.01, batch size 250 and 20 epochs.

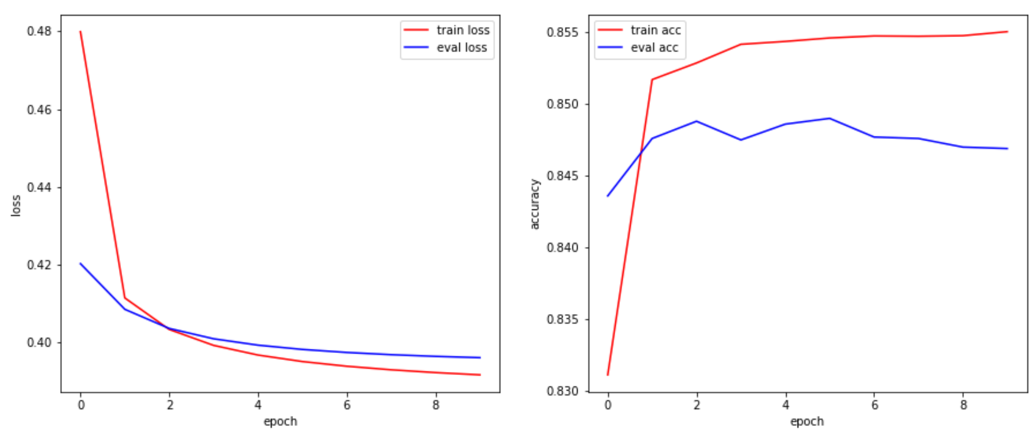

Figure 1: Train & Evaluation loss and accuracy graphs with default parameters.

e)

$$\begin{aligned}
\mathcal{L}(\mathbf{W},\mathbf{b},\mathbf{y},\mathbf{x}) &= (\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}}) \\
&= (\mathbf{y}-\sigma(\mathbf{W}\mathbf{x}+\mathbf{b}))^T(\mathbf{y}-\sigma(\mathbf{W}\mathbf{x}+\mathbf{b})) \\
&= \left(\mathbf{y}-\frac{1}{1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}\right)^T\left(\mathbf{y}-\frac{1}{1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}\right)
\end{aligned}$$

$$\begin{aligned}
\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W},\mathbf{b},\mathbf{y},\mathbf{x}) &= -\left(\frac{\mathbf{x}e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)^T(\mathbf{y}-\hat{\mathbf{y}})-\left(\frac{\mathbf{x}e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)(\mathbf{y}-\hat{\mathbf{y}})^T \quad \text{(by chain rule)} \\
&= \left(\frac{\mathbf{x}e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)^T(\hat{\mathbf{y}}-\mathbf{y})-\left(\frac{\mathbf{x}e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)(\hat{\mathbf{y}}-\mathbf{y})^T \\
&= (\hat{\mathbf{y}}-\mathbf{y})\otimes\left(\frac{\mathbf{x}e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)
\end{aligned}$$

Accordingly,

$$\begin{aligned}
\nabla_{\mathbf{b}}\mathcal{L}(\mathbf{W},\mathbf{b},\mathbf{y},\mathbf{x}) &= -\left(\frac{e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)^T(\mathbf{y}-\hat{\mathbf{y}})-\left(\frac{e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)(\mathbf{y}-\hat{\mathbf{y}})^T \quad \text{(by chain rule)} \\
&= (\hat{\mathbf{y}}-\mathbf{y})\otimes\left(\frac{e^{\mathbf{W}\mathbf{x}+\mathbf{b}}}{(1+e^{\mathbf{W}\mathbf{x}+\mathbf{b}})^2}\right)
\end{aligned}$$

Question 2. Softmax

a)

$$(\nabla_{\mathbf{a}}(\mathcal{L}(\mathbf{a},y)))_i = \frac{\partial\mathcal{L}}{\partial a_i} = \frac{\partial(\log\sum_j e^{a_j})-a_y}{\partial a_i} \overset{\text{chain rule}}{=} \frac{e^{a_i}}{\sum_j e^{a_j}} - \delta_{iy} = \mathbf{S}(\mathbf{a})_i - \delta_{iy}$$

b) See the uploaded file.

c) See the uploaded file.

d) For our preactivation values we picked the following 3 settings:

$$\mathbf{a} = [0.46608562, 0.48460241, 0.94915209]^T,\ a=2$$

$$\mathbf{a} = [0.05511136, 0.67564959, 0.29050784]^T,\ a=2$$

$$\mathbf{a} = [0.74507142, 0.16727477, 0.04435557]^T,\ a=0$$

which, with a step $e = 0.1$ yielded the following absolute errors, respectively

$$[0.01011092, 0.01022339, , 0.01239077]^T$$

$$[0.00933903, 0.01241475, 0.01076698]^T$$

$$[0.01249669, 0.01006632, 0.0093081]^T$$