1.

$$s^{(1)} = w^{(1)} x + b^{(1)}$$

$$W^{(1)} = \begin{bmatrix} 1 & -2 & 1 \\ 3 & 4 & -2 \end{bmatrix} \quad b^{(1)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \quad x = \begin{bmatrix} +1 \\ -1 \\ +1 \end{bmatrix}$$

$$s^{(1)} = \begin{bmatrix} 1 & -2 & 1 \\ 3 & 4 & -2 \end{bmatrix} \begin{bmatrix} +1 \\ -1 \\ +1 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} 1+2+1+1 \\ 3-4-2-2 \end{bmatrix} = \boxed{\begin{bmatrix} 5 \\ -5 \end{bmatrix}}$$

$$a^{(1)} = h(s^{(1)}) = \max(0, s^{(1)})$$

$$= \begin{bmatrix} \max(0, 5) \\ \max(0, -5 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 5 \\ 0 \end{bmatrix}}$$

$$a^{(1)} = \boxed{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}$$

$$s^{(2)} = W^{(2)} a^{(1)} + b^{(2)}$$

$$W^{(2)} = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \quad b^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad a^{(1)} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

$$s^{(2)} = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 5+1 \\ 15+0 \end{bmatrix} = \boxed{\begin{bmatrix} 6 \\ 15 \end{bmatrix}}$$

$$a^{(2)} = h(s^{(2)}) = \max(0, s^{(2)})$$

$$= \begin{bmatrix} \max(0, 6) \\ \max(0, 15) \end{bmatrix} = \boxed{\begin{bmatrix} 6 \\ 15 \end{bmatrix}}$$

$$a^{(2)} = \boxed{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}$$

$$s^{(3)} = W^{(3)} \cdot a^{(2)} + b^{(3)}$$

$$W^{(3)} = \begin{bmatrix} 2 & 2 \\ 3 & -3 \\ 2 & 1 \end{bmatrix} \qquad b^{3} = \begin{bmatrix} 0 \\ -4 \\ -2 \end{bmatrix}$$

$$s^{(3)} = \begin{bmatrix} 2 & 2 \\ 3 & -3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 15 \end{bmatrix} + \begin{bmatrix} 0 \\ -4 \\ -2 \end{bmatrix}$$

$$= \begin{bmatrix} 12 + 30 + 0 \\ 18 - 45 - 4 \\ 12 + 15 - 2 \end{bmatrix} = \boxed{\begin{bmatrix} 42 \\ -31 \\ 25 \end{bmatrix}}$$

$$a_i^{(3)} = \frac{e^{s_i^{(3)}}}{\sum_{j=1}^{3} e^{s_j^{(3)}}}$$

$$e^{42} = 1.739 \times 10^{18}$$

$$e^{-31} = 3.442 \times 10^{-14}$$

$$e^{25} = 7.2 \times 10^{10}$$

$$\sum_{j=1}^{3} e^{s_j^{(3)}} = e^{42} + e^{-31} + e^{25} = 1.739 \times 10^{18}$$

$$a_1^{(3)} = \frac{e^{42}}{1.739 \times 10^{18}} = 1$$

$$a_2^{(3)} = \frac{e^{-31}}{1.739 \times 10^{18}} = 1.979 \times 10^{-32} \doteqdot 0$$

$$a_3^{(3)} = \frac{e^{25}}{1.739 \times 10^{18}} = 4.14 \times 10^{-8}$$

$$a^{(3)} = \boxed{\begin{bmatrix} 1 \\ 0 \\ 4.14 \times 10^{-8} \end{bmatrix}}$$

(a) **Feedforward Computation:** Perform the feedforward calculation for the input vector $\mathbf{x} = [\,+1 \;\; -1 \;\; +1\,]^{\mathrm{T}}$. Fill in the following table. Follow the notation used in the slides, *i.e.*, $\mathbf{s}^{(l)}$ is the linear activation, $\mathbf{a}^{(l)} = \underline{h}(\mathbf{s}^{(l)})$, and $\dot{\mathbf{a}}^{(l)} = \dot{\underline{h}}(\mathbf{s}^{(l)})$.

l:	1	2	3
$\mathbf{s}^{(l)}$:	$\begin{bmatrix} 5 \\ -5 \end{bmatrix}$	$\begin{bmatrix} 6 \\ 15 \end{bmatrix}$	$\begin{bmatrix} 42 \\ -31 \\ 25 \end{bmatrix}$
$\mathbf{a}^{(l)}$:	$\begin{bmatrix} 5 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 6 \\ 15 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 4.14\times10^{-8} \end{bmatrix}$
$\dot{\mathbf{a}}^{(l)}$:	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	(not needed)

(b)

$$y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\delta^{(3)} = \frac{\partial C}{\partial s^{(3)}} = a^{(3)} - y$$

$$= \begin{bmatrix} 1 - 0 \\ 0 - 0 \\ 4.14\times10^{-8} - 1 \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}}$$

$$\delta^{(2)} = \left[(W^{(3)})^{T} \cdot \delta^{(3)} \right] \odot \dot{a}^{(2)}$$

$$= \begin{bmatrix} 2 & 3 & 2 \\ 2 & -3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \odot \dot{a}^{(2)}$$

$$= \begin{bmatrix} 2 + 0 - 2 \\ 2 + 0 - 1 \end{bmatrix} \odot \dot{a}^{(2)}$$

$$= \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}$$

$$\delta^{(1)} = \left[(W^{(2)})^T \cdot \delta^{(2)} \right] \odot \dot{a}(1)$$

$$= \begin{bmatrix} 1 & 3 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \odot \dot{a}(1)$$

$$= \begin{bmatrix} 3 \\ 4 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 3 \\ 0 \end{bmatrix}}$$

$W^{(3)}$ (update) $= W^{(3)} - 0.5 \cdot \delta^{(3)} [a^{(2)}]^T$ $\quad 3\times1 \cdot 1\times2$

$$= \begin{bmatrix} 2 & 2 \\ 3 & -3 \\ 2 & 1 \end{bmatrix} - 0.5 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 6 & 15 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 2 \\ 3 & -3 \\ 2 & 1 \end{bmatrix} - 0.5 \begin{bmatrix} 6 & 15 \\ 0 & 0 \\ -6 & -15 \end{bmatrix}$$

$$= \begin{bmatrix} 2-3 & 2-7.5 \\ 3 & -3 \\ 2+3 & 1+7.5 \end{bmatrix} = \boxed{\begin{bmatrix} -1 & -5.5 \\ 3 & -3 \\ 5 & 8.5 \end{bmatrix}}$$

$b^{(3)}$ update $= b^{(3)} - 0.5\, \delta^{(3)}$

$$= \begin{bmatrix} 0 \\ -4 \\ -1 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0 \\ -0.5 \end{bmatrix} = \boxed{\begin{bmatrix} -0.5 \\ -4 \\ -1.5 \end{bmatrix}}$$

$W^{(2)}$ update $= W^{(2)} - 0.5\, \delta^{(2)} [a^{(1)}]^T$

$$= \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} - 0.5 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 5 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} - 0.5 \begin{bmatrix} 0 & 0 \\ 5 & 0 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 1 & -2 \\ 0.5 & 4 \end{bmatrix}}$$

$b^{(2)}$ update $= b^{(2)} - 0.5\,\delta^{(2)}$

$$= \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0.5 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ -0.5 \end{bmatrix}}$$

$w^{(1)}$ update $= w^{(1)} - 0.5\,\delta^{(1)} \cdot [x]^T$

$$= \begin{bmatrix} 1 & -2 & 1 \\ 3 & 4 & -2 \end{bmatrix} - 0.5 \cdot \begin{bmatrix} 3 \\ 0 \end{bmatrix} \begin{bmatrix} +1 & -1 & +1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -2 & 1 \\ 3 & 4 & -2 \end{bmatrix} - 0.5 \begin{bmatrix} 3 & -3 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} -0.5 & -0.5 & -0.5 \\ 3 & 4 & -2 \end{bmatrix}}$$

$b^{(1)}$ update $= b^{(1)} - 0.5\,\delta^{(1)}$

$$= \begin{bmatrix} 1 \\ -2 \end{bmatrix} - 0.5 \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \boxed{\begin{bmatrix} -0.5 \\ -2 \end{bmatrix}}$$

(b) **Backpropagation Computation:** Apply standard SGD backpropagation for the input assuming a multi-category cross-entropy loss function and one-hot labeled target: $\mathbf{y} = [\,0\ \ 0\ \ 1\,]^{\mathrm{T}}$. Follow the notation used in the slides, *i.e.*, $\delta^{(l)} = \nabla_{\mathbf{s}^{(l)}} C$. Enter the delta values in the table below and provide the updated weights and biases assuming a learning rate $\eta = 0.5$.

l:	1	2	3
$\delta^{(l)}$:	$\begin{bmatrix} 3 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$
$\mathbf{W}^{(l)}$:	$\begin{bmatrix} -0.5 & -0.5 & -0.5 \\ 3 & 4 & -2 \end{bmatrix}$	$\begin{bmatrix} 1 & -2 \\ 0.5 & 4 \end{bmatrix}$	$\begin{bmatrix} -1 & -5.5 \\ 3 & -3 \\ 5 & 8.5 \end{bmatrix}$
$\mathbf{b}^{(l)}$:	$\begin{bmatrix} -0.5 \\ -2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} -0.5 \\ -4 \\ -1.5 \end{bmatrix}$

2.

i. How did you determine a learning rate? What values did you try? What was your final value?

ii. Describe the method you used to establish model convergence.

iii. What regularizers did you try? Specifically, how did each impact your model or improve its performance?

iv. Plot log-loss (*i.e.,* learning curve) of the training set and test set on the same figure. On a separate figure plot the accuracy against iteration number of your model on the training set and test set. Plot each as a function of the iteration number.

v. Clasify each input to the binary output "`digit is a 2`" using a 0.5 threshold. Compute the final loss and final accuracy for both your training set and test set.

(i)

I experimented with different learning rates from 0.001 and increasing up to 0.5. I found that a learning rate of 0.5 was too high, causing the model to struggle to converge. The test loss became very volatile and fluctuated significantly, indicating poor generalization to the test set. I determined that a learning rate of 0.001 for my final value. It allowed the model to converge smoothly and reduced the training and test loss consistently.

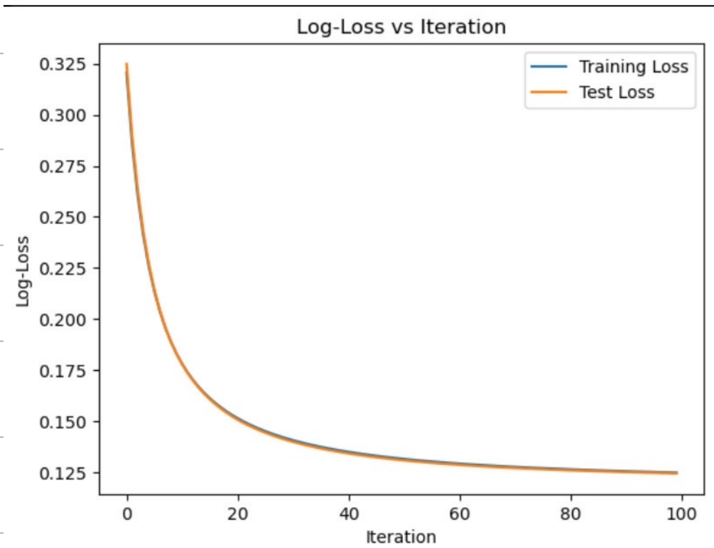
(ii)

I established model convergence by tracking both the binary log-loss provided in the question and accuracy for the training and test sets after each epoch. Additionally, I applied a 0.5 threshold to the predicted probabilities to classify each input. The model is considered converged when the log-loss and accuracy stabilize.

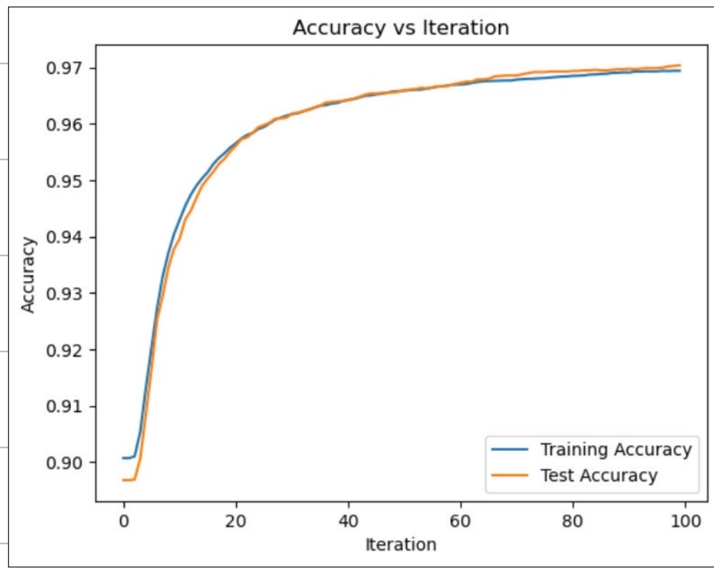
(iii)

I used L2 Regularization in my model. I experimented with different lambda value from 0.00001 to 1. Using too strong regularization lead to underfitting and cannot generalize well. Also using too weak value do not help much to constraint, so I decided 0.01 for my regularization value.

(iv)



Log-Loss for my model using L2 Regularization, lambda = 0.01 and learning rate = 0.001.



Accuracy of the training and testing set.

(v)

Final Train Loss: 0.12479557917330604
Final Test Loss: 0.12444388318822013
Final Train Accuracy: 0.9694333333333334
Final Test Accuracy: 0.9704