Homework 3

Graded

Student

Jinzhi Shen

Total Points

44 / 53 pts

Question 1

1 — **20** / 20 pts

1.1 a — **5** / 5 pts

✔ **– 0 pts** Correct

1.2 b — **5** / 5 pts

✔ **– 0 pts** Correct

1.3 c i — **2** / 2 pts

✔ **– 0 pts** Correct

1.4 c ii — **3** / 3 pts

✔ **– 0 pts** Correct

1.5 c iii — **5** / 5 pts

✔ **– 0 pts** Correct

Question 2

2 — **14** / 20 pts

2.1 Deep narrow networks & shallow wide networks — **4** / 8 pts

✔ **– 1 pt** (iii) Incorrect answer for w and b in the first layer

✔ **– 2 pts** (iii) Incorrect answer for w and b in the middle layer

✔ **– 1 pt** (iii) Incorrect answer for w and b in the final layer

2.2 Network Overview — **10** / 12 pts

✔ **– 1 pt** Incorret loss

✔ **– 1 pt** Incorrect updated beta 0 (bias). Note that $\frac{\partial \ell}{\partial \beta_{1,0}}$ is not 0. Use chain rule.

Loss is positive.

Question 3

3 **7** / 8 pts

3.1 **a** **4** / 4 pts

✔ **– 0 pts** Correct

3.2 **d** **3** / 4 pts

✔ **– 1 pt** Incorrect: rbf kernel w/ sigma=5

Question 4

4 **3** / 5 pts

4.1 **a** **3** / 3 pts

✔ **– 0 pts** Correct

4.2 **d** **0** / 2 pts

✔ **– 2 pts** Incorrect

Questions assigned to the following page: 1.1 and 1.2

ECE 449 Homework 3

Question 1

(a)

Lagrangian (with Lagrangian multipliers $\alpha_i \geq 0$, $\beta_i \geq 0$):

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b) - 1 + \xi_i\right] - \sum_{i=1}^{N}\beta_i\xi_i$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)}) \qquad \Longrightarrow \mathbf{w} = \sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)})$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial b} = \sum_{i=1}^{N}\alpha_i y^{(i)} \qquad \Longrightarrow \sum_{i=1}^{N}\alpha_i y^{(i)} = 0$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \qquad \Longrightarrow C = \alpha_i + \beta_i$$

Plug in the above three derived relations, we can rewrite the Lagrangian as:

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)} < \phi\mathbf{x}^{(i)}, \phi\mathbf{x}^{(j)} >$$

From the relation $C = \alpha_i + \beta_i$, considering that $\alpha_i, \beta_i \geq 0$, we thus have $0 \leq \alpha_i \leq C$.
Therefore, the dual problem is:

$$\max_{\alpha \geq 0}\sum_{i=1}^{N}\alpha_i - \frac{1}{2}||\sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)})||^2, \qquad \text{s.t. } 0 \leq \alpha_i \leq C, \sum_{i=1}^{N}\alpha_i y^{(i)} = 0$$

with the domain described in the question.

(b)

Lagrangian (with Lagrangian multipliers $\alpha_i \geq 0$, $\beta_i \geq 0$):

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i^2 - \sum_{i=1}^{N}\alpha_i\left[y^{(i)}(\mathbf{w}^T\phi(\mathbf{x}^{(i)}) + b) - 1 + \xi_i\right] - \sum_{i=1}^{N}\beta_i\xi_i$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)}) \qquad \Longrightarrow \mathbf{w} = \sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)})$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial b} = \sum_{i=1}^{N}\alpha_i y^{(i)} \qquad \Longrightarrow \sum_{i=1}^{N}\alpha_i y^{(i)} = 0$$

$$0 = \frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \xi_i} = C\xi_i - \alpha_i - \beta_i = 0 \qquad \Longrightarrow C\xi_i = \alpha_i + \beta_i$$

Plug in the above three derived relations, we can rewrite the Lagrangian as:

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y^{(i)}y^{(j)} < \phi\mathbf{x}^{(i)}, \phi\mathbf{x}^{(j)} > -\frac{1}{2C}\sum_{i=1}^{N}(\alpha_i + \beta_i)^2$$

Therefore, the dual problem is:

$$\max_{\alpha, \beta \geq 0}\sum_{i=1}^{N}\alpha_i - \frac{1}{2}||\sum_{i=1}^{N}\alpha_i y^{(i)}\phi(\mathbf{x}^{(i)})||^2 - \frac{1}{2C}\sum_{i=1}^{N}(\alpha_i + \beta_i)^2, \qquad \text{s.t. } \alpha_i, \beta_i \geq 0, \sum_{i=1}^{N}\alpha_i y^{(i)} = 0$$

Questions assigned to the following page: 2.1, 1.3, 1.4, and 1.5

(c)

(i)

The margin is the distance between the decision boundary and the closest example, since $y^{(i)}(\mathbf{w}^{*T}\mathbf{x}^{(i)}+b)-1=0$ for $x^{(i)}$ in gutter,

$$\text{margin } \rho = \frac{1}{2}\frac{\mathbf{w}^{*T}}{||\mathbf{w}^*||}(\mathbf{x}^+ - \mathbf{x}^-) = \frac{1}{2}\cdot\frac{2}{||\mathbf{w}^*||} = \frac{1}{||\mathbf{w}^*||}$$

which can be rewritten as

$$\frac{1}{\rho^2} = ||\mathbf{w}^*||_2^2$$

(ii)

$$\mathbf{w}^{*T}\phi(\mathbf{x}_0) + b = \sum_{i=1}^{N}\alpha_i^* y^{(i)} < \phi(\mathbf{x}^{(i)}), \phi(\mathbf{x}^{(j)}) > +b = \sum_{i=1}^{N}\alpha_i^* y^{(i)} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) + b$$

Since $\mathbf{x}_0$ is far away from any $\mathbf{x}^{(i)}$, $||\mathbf{x}_0 - \mathbf{x}^{(i)}||$ is large and thus $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is very close to zero. Therefore,

$$\mathbf{w}^{*T}\phi(\mathbf{x}_0) + b \approx 0 + b = b$$

(iii)

Since $\alpha_1 = \alpha_2 = ... = \alpha_N = 1$ and $b = 0$

$$||f(\mathbf{x}^{(i)}) - y^{(i)}|| = ||\sum_{j\neq i}\alpha_i y^{(i)} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})|| \leq \sum_{j\neq i}||k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})||$$

Since $||f(\mathbf{x}^{(i)}) - y^{(i)}|| < 1$ for all i, $||k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})||$ should be smaller than $\frac{1}{N-1}$ for any (i, j). It implies $exp\left(-\frac{1}{\tau^2}||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2\right) < \frac{1}{N-1}$. Given that $||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|| \geq \epsilon$ for any $i \neq j$, we only need to satisfy $exp\left(-\frac{1}{\tau^2}\epsilon^2\right) < \frac{1}{N-1}$, which can be rewritten as $\tau < \frac{\epsilon}{\sqrt{ln(N-1)}}$

Question 2

Deep narrow networks and shallow wide networks

(a)

From the definition of the ReLU function $\vec{\sigma}_r()$, it doesn't change non-negative elements, so does $\vec{\sigma}_r^n()$, $n = 1, 2, 3....$ In addition, the ReLU function $\vec{\sigma}_r()$ changes the negative elements to non-negative values. Therefore, the elements of $\vec{\sigma}_r(\mathbf{x})$ are all non-negative.

Therefore, we have:

$$\vec{\sigma}_r^n(\mathbf{x}) = \vec{\sigma}_r^{n-1}(\vec{\sigma}_r(\mathbf{x})) = \vec{\sigma}_r(\mathbf{x})$$

(b)

$\forall \mathbf{x} \in \mathbb{R}^d$, the i-th element of $\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x})$,

$$(\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x}))_i = (\vec{\sigma}_r(\mathbf{x}))_i - (\vec{\sigma}_r(-\mathbf{x}))_i$$

If the i-th element of $\mathbf{x}$, x_i, is positive, then

$$(\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x}))_i = (\vec{\sigma}_r(\mathbf{x}))_i - (\vec{\sigma}_r(-\mathbf{x}))_i = x_i - 0 = x_i$$

If the i-th element of $\mathbf{x}$, x_i, is zero, then

$$(\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x}))_i = (\vec{\sigma}_r(\mathbf{x}))_i - (\vec{\sigma}_r(-\mathbf{x}))_i = 0 - 0 = 0 = x_i$$

If the i-th element of $\mathbf{x}$, x_i, is negative, then

$$(\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x}))_i = (\vec{\sigma}_r(\mathbf{x}))_i - (\vec{\sigma}_r(-\mathbf{x}))_i = 0 - (-x_i) = x_i$$

Therefore, for all elements of $\mathbf{x}$, $x_i = (\vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x}))_i$, so $\mathbf{x} = \vec{\sigma}_r(\mathbf{x}) - \vec{\sigma}_r(-\mathbf{x})$

Question assigned to the following page: 2.2

(c)

Network Overview

(a)

Folding the $\alpha_{j,0}$, $\beta_{k,0}$, x_0 and z_0, we have

$$\tilde{\alpha} = \begin{bmatrix} 1 & 1 & 2 & 3 & 0 & 1 & 3 \\ 1 & 3 & 1 & 2 & 1 & 0 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 0 & 2 & 1 & 2 & 2 \end{bmatrix}, \tilde{\mathbf{x}}^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \tilde{\beta} = \begin{bmatrix} 1 & 1 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 3 & 1 & 1 & 1 \end{bmatrix}$$

Then we can compute $\mathbf{a}$ as follow:

$$\mathbf{a} = \tilde{\alpha} \cdot \tilde{\mathbf{x}}^{(1)} = \begin{bmatrix} 1 & 1 & 2 & 3 & 0 & 1 & 3 \\ 1 & 3 & 1 & 2 & 1 & 0 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 0 & 2 & 1 & 2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 8 \\ 6 \end{bmatrix} \tag{1}$$

$\mathbf{z}$ can be computed as follow:

$$\mathbf{z} = \frac{1}{1 + exp(-\mathbf{a})} = \begin{bmatrix} 0.99966465 \\ 0.99908895 \\ 0.99966465 \\ 0.99752738 \end{bmatrix} \tag{2}$$

Folding the z_0, we have:

$$\tilde{\mathbf{z}} = \begin{bmatrix} 1 \\ 0.99966465 \\ 0.99908895 \\ 0.99966465 \\ 0.99752738 \end{bmatrix}$$

Then we can compute $\mathbf{b}$ as follow:

$$\mathbf{b} = \tilde{\beta} \cdot \tilde{\mathbf{z}} = \begin{bmatrix} 1 & 1 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 3 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0.99966465 \\ 0.99908895 \\ 0.99966465 \\ 0.99752738 \end{bmatrix} = \begin{bmatrix} 6.99469922 \\ 5.993473 \\ 6.99527493 \end{bmatrix}$$

Then we can compute $\hat{y}_k, k = 1, 2, 3$ as follow:

$$\hat{y}_1 = \frac{exp(b_1)}{\sum_{l=1}^{3} exp(b_l)} = 0.4222965, \quad \hat{y}_2 = \frac{exp(b_2)}{\sum_{l=1}^{3} exp(b_l)} = 0.15516382, \quad \hat{y}_3 = \frac{exp(b_3)}{\sum_{l=1}^{3} exp(b_l)} = 0.42253968$$

Then we can compute the loss:

$$loss = -\sum_{k=1}^{3} y_k log(\hat{y}_k) = -1.86327383$$

(i)

The value of a_1 is 8.

(ii)

The value of z_1 is 0.9997.

Question assigned to the following page: 2.2

(iii)

The value of a_3 is 8.

(iv)

The value of z_3 is 0.9997.

(v)

The value of b_2 is 5.9935.

(vi)

The value of $\hat{y}_2$ is 0.1552.

(vii)

I would predict a class value of 3.

(viii)

The value of total loss on this training example is -1.8633.

(b)

$$\frac{\partial loss}{\partial \hat{y}_k} = -\frac{y_k}{\hat{y}_k}, \qquad \frac{\partial loss}{\partial b_k} = \frac{\partial loss}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial b_k} = -\frac{y_k}{\hat{y}_k} \cdot \hat{y}_k(1-\hat{y}_k) = y_k(\hat{y}_k - 1)$$

(i)

$$\frac{\partial loss}{\partial \beta_{2,1}} = \frac{\partial loss}{\partial b_2} \cdot \frac{\partial b_2}{\partial \beta_{2,1}} = y_2(\hat{y}_2 - 1) \cdot z_1$$

$$\beta_{2,1}^{new} = \beta_{2,1} - \eta \cdot \frac{\partial loss}{\partial \beta_{2,1}} = 1 - 1 \cdot (0.15516382 - 1) \cdot 0.99966465 \approx 1.8446$$

(ii)

$$\frac{\partial loss}{\partial \beta_{1,0}} = \frac{\partial loss}{\partial b_1} \cdot \frac{\partial b_1}{\partial \beta_{1,0}} = \frac{\partial loss}{\partial b_1} \cdot 1 = y_1(\hat{y}_1 - 1) = 0$$

$$\beta_{1,0}^{new} = \beta_{1,0} - \eta \cdot \frac{\partial loss}{\partial \beta_{1,0}} = \beta_{1,0} = 1$$

(iii)

$$\frac{\partial loss}{\partial \alpha_{3,4}} = \frac{\partial loss}{\partial z_3} \cdot \frac{\partial z_3}{\partial a_3} \cdot \frac{\partial a_3}{\partial \alpha_{3,4}}$$

$$\frac{\partial loss}{\partial z_3} = \sum_{k=1}^{3} \frac{\partial loss}{\partial b_k} \frac{\partial b_k}{\partial z_3} = \sum_{k=1}^{3} y_k(\hat{y}_k - 1) \cdot \beta_{k,3}$$

$$\frac{\partial z_3}{\partial a_3} = z_3(1 - z_3)$$

$$\frac{\partial a_3}{\partial \alpha_{3,4}} = x_4$$

Therefore,

$$\begin{aligned} \frac{\partial loss}{\partial \alpha_{3,4}} &= \left(\sum_{k=1}^{3} y_k(\hat{y}_k - 1) \cdot \beta_{k,3} \right) \cdot z_3(1 - z_3) \cdot x_4 \\ &= (0.15516382 - 1) \cdot 0.99966465(1 - 0.99966465) \cdot 0 \\ &= 0 \end{aligned} \tag{3}$$

Questions assigned to the following page: 3.1, 3.2, and 2.2

$$\alpha_{3,4}^{new} = \alpha_{3,4} - \eta \cdot \frac{\partial loss}{\partial \alpha_{3,4}} = \alpha_{3,4} - 0 = \alpha_{3,4} = 2$$

(iv)

We would predict class 2 in that case.

Question 3

(a)

$$\Pi_{[0,\infty)^n}[\alpha] = argmin_{\alpha' \in [0,\infty)^n} ||\alpha' - \alpha||_2 = argmin_{\alpha' \in [0,\infty)^n} \sqrt{\sum_{i=1}^{n} (\alpha_i^{'} - \alpha_i)^2}$$

If $\alpha_i \geq 0$, then to minimize $||\alpha' - \alpha||_2$, $\alpha_i^{'}$ should be set to α_i, otherwise, $\alpha_i^{'}$ should be set to 0.

Therefore, we will have:

$$(\Pi_{[0,\infty)^n}[\alpha])_i = \max\{\alpha_i, 0\}$$

$$\Pi_{[0,C]^n}[\alpha] = argmin_{\alpha' \in [0,C]^n} ||\alpha' - \alpha||_2 = argmin_{\alpha' \in [0,C]^n} \sqrt{\sum_{i=1}^{n} (\alpha_i^{'} - \alpha_i)^2}$$

If $0 \leq \alpha_i \leq C$, then to minimize $||\alpha' - \alpha||_2$, $\alpha_i^{'}$ should be set to α_i, else if $\alpha_i < 0$, $\alpha_i^{'}$ should be set to 0, otherwise $\alpha_i > C$, $\alpha_i^{'}$ should be set to C.

Therefore, we will have:

$$(\Pi_{[0,C]^n}[\alpha])_i = \min\{\max\{0, \alpha_i\}, C\}$$

(d)

Polynomial kernel with degree 3

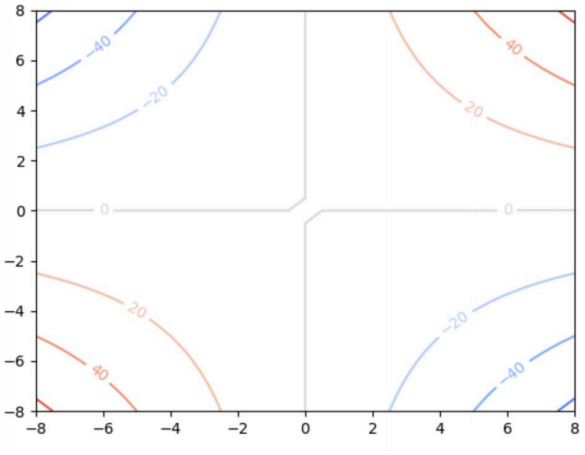

Figure 1: Polynomial kernel with degree 3.

Questions assigned to the following page: 4.1, 3.2, and 4.2

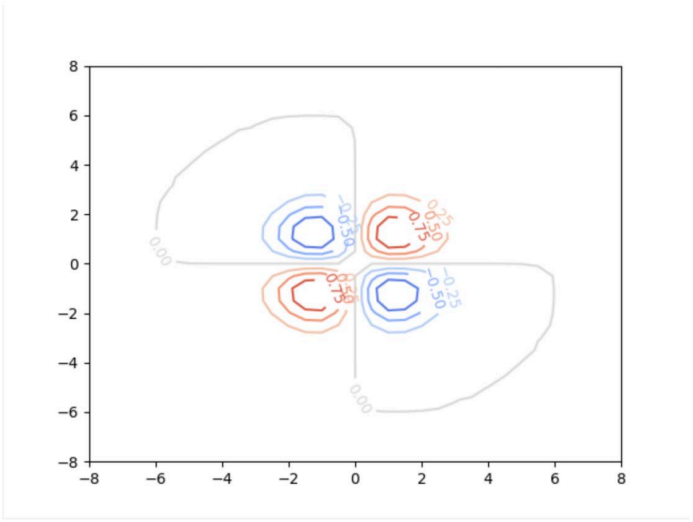

Figure 2: RBF kernel with sigma 1.

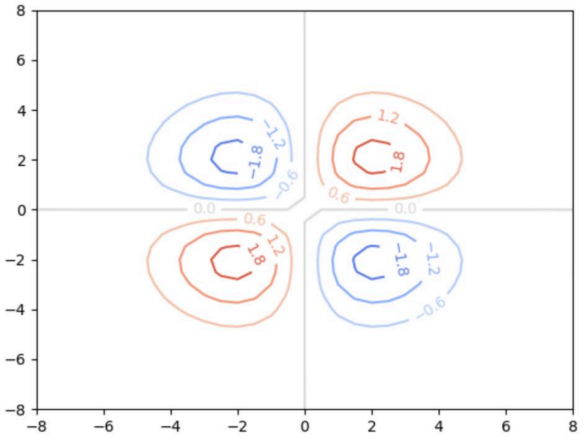

Figure 3: RBF kernel with sigma 2.

RBF kernel with sigma 1

RBF kernel with sigma 2

RBF kernel with sigma 3

Question 4

(a)

The number of input channels of the first convolutional layers is 1, and the number of input channels of the second convolutional layers is 7.

The number of input features of the last fully connected layer is 12.

(d)

Question assigned to the following page: 4.2

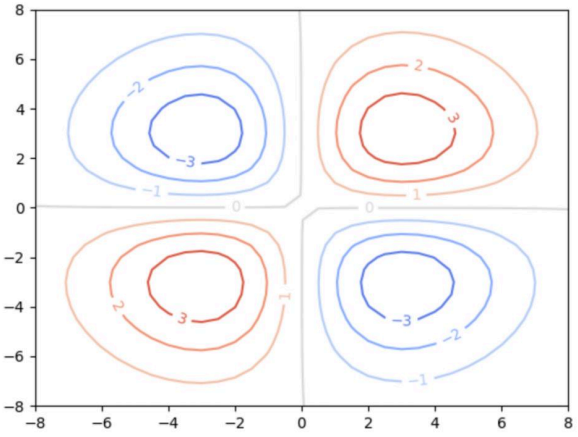

Figure 4: RBF kernel with sigma 3.

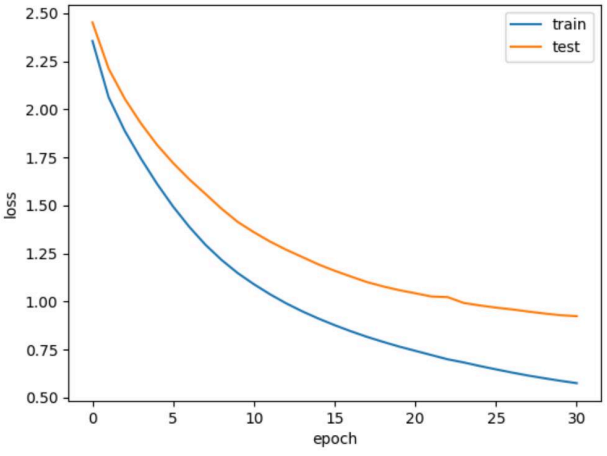

Figure 5: Epochs vs training and test loss.