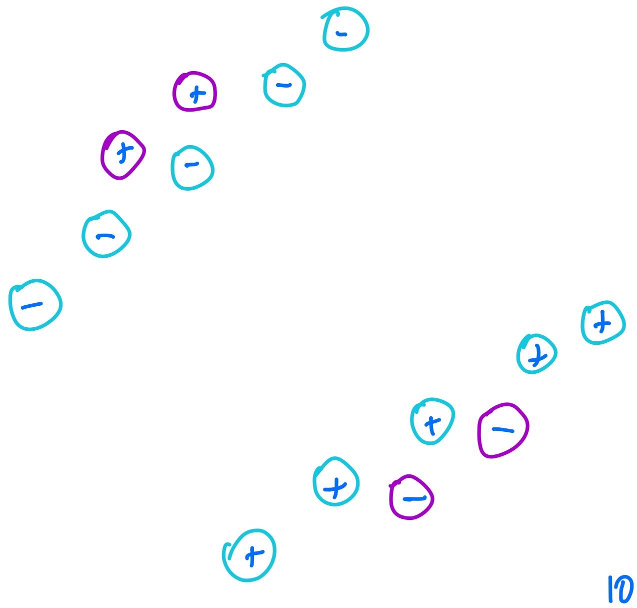
1) a)



$10/14$

ACCURACY = 0.7142857 //

ERROR = $4/14$ = 0.285714 //

b. False. Need not be the case. The least training accuracy occurs when we choose k is 1. In which case it is the closest point and hence 100% accuracy. But this will not give good test accuracy. This is because of overfitting.

ii. Method 2 is better. If we choose hyperparameters that minimize the training error, we are overfitting the hyperparameters. Therefore, this may not generalize well to new test data. In case of method 2, we are setting the hyperparamter on a new set of data and hence can avoid overfitting.

c. A and B are True. As we increase k, we are considering more points and hence outliers are removed, smoothening the boundary.

C is false. Increasing k decreases overfitting as it is more generalized now.

D is true. We can use cross validation as it is a hyperparameter.

E is false. We should set k based on validation set.

d. i. We have overfit the data in this instance. The classifier overfits to the training data and hence has high tesr error and low training errror.

ii. A will improve the classifier. As more data will help reduce overfitting.
B will not help as lesser data will be more easily overfit.
C will not help as it will become more easy to overfit.
D will help as this disables overfiting.
E will improve accuracy but the classifier is by no means better. it is just overfitiing to the test data.
F is incorrect.

A, D

2. a) i)

$$L = s \times (-y)$$

$$\frac{\partial L}{\partial s} = -y$$

$$s = \cos(z, \hat{z}) = \frac{z^T \hat{z}}{\|z\|_2 \, \|\hat{z}\|_2}$$

$$\frac{\partial s}{\partial z} = \frac{\left(\|z\|_2 \|\hat{z}\|_2 \frac{\partial}{\partial z}(z^T \hat{z}) - z^T \hat{z} \frac{\partial}{\partial z}(\|z\|_2 \|\hat{z}\|_2)\right)}{(\|z\|_2 \|\hat{z}\|_2)^2}$$

$$= \frac{(\|z\|_2 \|\hat{z}\|_2 \hat{z} - z^T \hat{z} \left(\frac{\|\hat{z}\|_2}{\|z\|_2}\right) z}{\|z\|_2^2 \, \|\hat{z}\|_2^2}$$

$$\frac{\partial s}{\partial \hat{z}} = \frac{(\|z\|_2 \|\hat{z}\|_2) z - z^T \hat{z} \left(\frac{\|z\|}{\|\hat{z}\|}\right) \hat{z}}{\|z\|_2^2 \, \|\hat{z}\|_2^2}$$

$$\frac{\partial L}{\partial z} = \frac{\partial s}{\partial z} \frac{\partial L}{\partial s} = \frac{(\|z\|_2 \|\hat{z}\|_2 \hat{z} - z^T \hat{z} \left(\frac{\|\hat{z}\|_2}{\|z\|_2} \right) z}{\|z\|_2^2 \|\hat{z}\|_2^2} \times (-y)$$

$$\frac{\partial L}{\partial \hat{z}} = \frac{\partial s}{\partial \hat{z}} \cdot \frac{\partial L}{\partial s} = \frac{(\|z\|_2 \|\hat{z}\|_2) z - z^T \hat{z} \left(\frac{\|z\|_2}{\|\hat{z}\|_2} \right) \hat{z}}{\|z\|_2^2 \|\hat{z}\|_2^2} \times (-y)$$

ii) $\nabla_{W_2} L$

$$\frac{\partial L}{\partial W_2}$$

$$z = W_2 h_1$$

$$\hat{z} = W_2 \hat{h}_1$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z} \cdot h_1^T$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{z}} \cdot \hat{h}_1^T$$

$$\frac{\partial L}{\partial W_2} = \delta_z \cdot h_1^T + \delta_{\hat{z}} \hat{h}_1^T$$

iii) $$\frac{\partial z}{\partial h_1} = W_2^T \frac{\partial L}{\partial z}$$

$$= W_2^T \delta_z$$

$$\frac{\partial L}{\partial \hat{h}_1} = W_2^T \delta_{\hat{z}}$$

$$z = W_2 h_1$$

$$\hat{z} = W_2 \hat{h}_1$$

iv) $\frac{\partial L}{\partial m}$ $\frac{\partial L}{\partial n}$

$$h_1 = \text{ReLU}(m)$$

$$\frac{\partial h_1}{\partial m} = \begin{cases} 1 , & m > 0 \\ 0 , & m \leq 0 \end{cases}$$

$$= I(m > 0)$$

$$\frac{\partial L}{\partial m} = \frac{\partial h_1}{\partial m} \frac{\partial L}{\partial h_1} = I(m > 0) \odot \delta_{h_1}$$

$$\frac{\partial L}{\partial n} = \frac{\partial \hat{h}_1}{\partial n} \frac{\partial L}{\partial \hat{h}_1} = I(n > 0) \odot \delta_{\hat{h}_1}$$

v) $\frac{\partial L}{\partial W_1}$

$$m = W_1 x$$

$$n = W_1 \hat{x}$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial m} \cdot x^T + \frac{\partial L}{\partial n} \cdot \hat{x}^T$$

$$= \delta_m \cdot x^T + \delta_n \cdot \hat{x}^T$$

b) i) $\{x^{(g)}, \hat{x}^{(g)}, +1\}$

$$z^{(g)} = \hat{z}^{(g)}$$

$$L = -y\,s$$

$$= -s$$

$$= -\cos(z^{(g)}, \hat{z}^{(g)})$$

$$= -\left(\frac{z^{(g)T}\hat{z}^{(g)}}{\|z^{(g)}\|_2\,\|\hat{z}^{(g)}\|_2}\right)$$

if $z^{(g)} = \hat{z}^{(g)}$

$$z^{(g)} \cdot \hat{z}^{(g)} = \|z^{(g)}\|^2$$

$$= -\left(\frac{\|z^{(g)}\|^2}{\|z^{(g)}\|^2}\right)$$

$$= -1$$

ii) $z^{(g)}$, $\hat{z}^{(g)}$ are orthogonal.

$$z^{(g)^T} \hat{z}^{(g)} = 0$$

$$L = -y\, s$$

$$= -s$$

$$= -\cos\left(z^{(g)}, \hat{z}^{(g)}\right)$$

$$= -\left(\frac{z^{(g)^T} \hat{z}^{(g)}}{\|z^{(g)}\|_2 \, \|\hat{z}^{(g)}\|_2}\right)$$

$$= 0$$

iii) $z^{(g)} = -\hat{z}^{(g)}$

$$L = -y\, s$$

$$= -s$$

$$= -\cos(z^{(g)}, \hat{z}^{(g)})$$

$$= -\left(\frac{z^{(g)T}\hat{z}^{(g)}}{\|z^{(g)}\|_2 \, \|\hat{z}^{(g)}\|_2}\right)$$

if $z^{(g)} = -\hat{z}^{(g)}$

$$z^{(g)} \cdot \hat{z}^{(g)} = -\|z^{(g)}\|^2$$

$$= +\left(\frac{\|z^{(g)}\|^2}{\|z^{(g)}\|^2}\right)$$

$$= +1$$

c) Yes. Loss is more when the embeddings are not aligned. When aligned loss is 0, when at 90°, they are 0 and when totally misaligned, they are maximum. So by decreasing the loss, we can make the embeddings align.

3. a. ReLU solves the vanishing gradient problem.
Tanh causes vanishing gradient problem as it saturates on higher and lower values of x.
Sigmoid causes vanishing gradient problem as it saturates on higher and lower values of x.
Leaky ReLU doesnt cause vanishing gradient problem.
Identity doesnt cause vanishing gradient problem.

tanh , sigmoid

b. A. False. Batch normalization is just normalizing the data and doesnt cause any increase in iterations.
B. The mean and standard deviations can be learned. So false.
C. Yes it is non linear and ensures output has mean 1 and std 0.
D. True. It normalizes the data and hence helps in regularization. Regularization can be viewed as introducing noise.
E. False. It is applicable at test time.

It doesn't explicitly add noise. But the normalization regularization effect could be viewed as noise addition

C, D

c. A. True.
B. False. It will avoid overfitting. So it might increase training loss.
C. True.It is an ensemble method like bagging.
D. True. It can make the weights go to 0.

A, C, D

d. A. True. Only the last layers are re trained.
B. True. Data augmentation can help add data.
C. Same data is run over multiple task. So False. If the particular models are simple and can be trained with lesser data, we can proceed.
D. True. As we are combining resutls from multiple model, it can help improve performance.

A, B, D

e. 1 is the Loss/error
2 is number of epochs ↗ loss
3 is the least validation error that can be achieved
4 is the number of epoch that achieves least error. So stop at that point.
5 is the validation error /loss
6 is training error/ loss

4. a. i. Gradient Descent with momentum has highest chance to get out of the trap. We all know that vanilla gradient descent will have diminishing gardients and might stop at plateau. Momentum carries the historic trend and historically we have been moving to the right in the last 5 steps. So momentum will have a component towards the right and will push the gradient over the plateau and after that gradient are in the right direction and hence it can reach a local minima. We might consider adagrad or adam as well. Adagrad punishes the gradients for moving in a particular direction. It carries the running sum of squares of gradient norms and hence as we moved toward down in the last few steps, the gradient along y axis will be very less and it may not help overcome the trap. Adam is a combination of rmsprop and momentum. So it has the same disadvantages as adagrad because of the rmsprop component and the advantages of momentum. So it might help overcome the trap but momentum alone is a better bet.

ii. Gradient Descent with momentum > Adam > Adagrad

Momentum carries the historic trend and historically we have been moving to the right in the last 5 steps. So momentum will have a component towards the right with a very large magnitude. Adagrad punishes the gradients for moving in a particular direction. It carries the running sum of squares of gradient norms and hence as we moved toward down in the last few steps, the gradient along y axis will be very less decreasing the magnitude. Adam is a combination of rmsprop and momentum. So it has the same disadvantages as adagrad because of the rmsprop component and the advantages of momentum. So it might lie in between.

b) $\gamma_1 E[v_t] = \gamma_1(1-\beta_1) E\left(\sum_{i=1}^{t} \beta_1^{t-i} g_i\right)$

$$= \gamma_1(1-\beta_1) E\left(\beta_1^{t-1} g_1 + \beta_2^{t-2} g_2 + \beta_3^{t-3} g_3 \ldots + \beta^0 g_t\right)$$

$$= \gamma_1(1-\beta_1)\left[\beta_1^{t-1} E(g_1) + \beta_2^{t-2} E(g_2) + \cdots + E(g_t)\right]$$

$$E|g_t] = \mu$$

$$= \gamma_1((1-\beta_1)\left[\beta_1^{t-1}\mu + \beta_2^{t-2}\mu + \ldots + \mu\right]$$

$$= \gamma_1(1-\beta_1)(\beta_1^0 + \beta_1 + \ldots \beta_1^{n-1})\mu$$

$$= \gamma_1(1-\beta_1)\frac{(1-\beta_1^t)}{1-\beta_1}\mu$$

$$= \frac{1}{(1-\beta_1^t)} \underline{(1-\beta_1^t)} \mu$$

$$= \mu$$

$$\gamma_2 \mathbb{E}[a_t] = \gamma_2 (1-\beta_2) \mathbb{E}\left(\sum_{i=1}^{t} \beta_2^{t-i} g_i\right)$$

Similar to previous we get

$$= \gamma_2 (1-\beta_2) \frac{(1-\beta_2^t)}{1-\beta_2} s$$

$$= \frac{1}{1-\beta_2 t} \cdot (1-\beta_2 t) \cdot s$$

$$= s //$$

5) $v \leftarrow \alpha v - \epsilon \nabla_{\theta} L(\theta + \alpha v)$

$\theta \leftarrow \theta + v$

Lets start with θ_1 , v_1

$v_2 \leftarrow \alpha v_1 - \epsilon \nabla_{\theta} L(\theta_1 + \alpha v_1)$

$\theta_2 \leftarrow \theta_1 + v_2$

We get θ_2, v_2

Start with θ_1, v_1

Second gives $\tilde{\theta}_{old} = \theta_1 + \alpha v_1$

$v_2 = \alpha v_1 - \epsilon \nabla_{\theta} L(\theta_1 + \alpha v_1)$ — $\boxed{\text{SAME}}$

$\tilde{\theta}_2 = (\theta_1 + \alpha v_1) + v_2 + \alpha(v_2 - v_1)$

$= \theta_1 + \alpha v_1 + v_2 + \alpha v_2 - \alpha v_1$

$= \theta_1 + v_2 + \alpha v_2$

$\theta_2 = \tilde{\theta}_2 - \alpha v_2$

$= \theta_1 + v_2 + \alpha v_2 - \alpha v_2$

$= \theta_1 + v_2$

We return θ_2, v_2

So both θ_2 and v_2 are same for both //