Exercise 1

3.

(a) training set is used for machine learning model training, where it is used to adjust the parameters and weights of model and minimize training loss.

Validation set is for checking if model generalize well during training. By using it seperately from training set, it can evaluate the models performance during training.

Test set is for after training, which provides an evaluation on how the model might performs on the expected data.

(b)

By avoiding to use test set during training, it would prevent model from knowing the information ahead and adapt it into the model. which could lead to a biased evaluation because the model will most likely perform well since it has used the set to train itself. It could also likely to increase variance of estimate since it overfits if test set is used during training.

(c)

First define some possible values for each hyperparameters. which is a grid. Then among these assumptions, we use different combination to train and evaluate the model using training and validation set.

Compare the results and get the best result grid which then should be the final hyperparameters we choose.

4.

(a) RMSProp is a optimizer which solves the problem of learning rate too large or small. It automatically adjust learning rate for each parameter which helps the model to converge faster and more smoothly

Momentum optimizer is instead of update parameter base on instant gradient result; it saves the results from previous updates too and take them into account.

(b)

(i) $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ → updated first moment

$= 0.9 \cdot 0.5 + (1-0.9) \cdot 2.0$

$= 0.65$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ → updated second moment

$= 0.99 \cdot 0.2 + (1-0.99) \cdot (2.0)^2$

$= 0.238$

(ii)

$\Delta w_t = -\alpha \frac{m_t}{\sqrt{v_t} + \varepsilon}$

$= -0.01 \cdot \frac{0.65}{\sqrt{0.238} + \varepsilon}$

$= -0.01 \cdot 1.33$

$= 0.0133$

(iii)

v_t' would be a lot larger than v_t because $20 \gg 0.2(v_{t-1})$

$|\Delta w_t'|$ would be smaller than $|\Delta w_t|$ because again v_t' is large and then $\sqrt{v_t'}$ on denominator would make the result small

It implies that Adam automatically adjust the learning rate for different parameters.