HW1

- HW1-1
- Simulate a Function

1.Describe the models you use, including the number of parameters (at least two models) and the function you use. (0.5%)

Ans:

a.Shallow

為一層hidden layer dnn，size為280，共有843個parameters。optimizer為Adam，loss function為mse，batch size設定為一，epoch為10000。

b.Mid model

為四層hidden layer dnn，size為15、20、15、10，共有838個parameters。optimizer為Adam，loss function為mse，batch size設定為一，epoch為10000。

c.Deep model

為七層hidden layer dnn，size為10、10、10、15、15、10、5，共有868個parameters。optimizer為Adam，loss function為mse，batch size設定為一，epoch為10000。

使用函式：

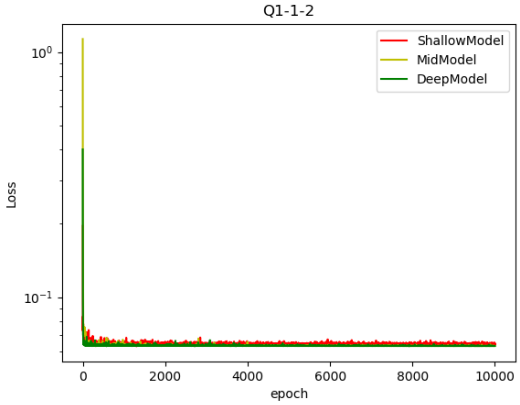
a. $y = sin\left(\frac{x\pi}{180}\right)$

b. $y = \frac{sin\,(10x\pi)}{10x\pi}$

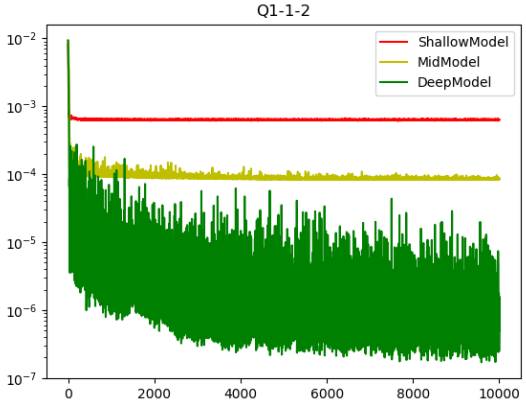
2.In one chart, plot the training loss of all models. (0.5%)

Ans:

a. $y = sin\left(\frac{x\pi}{180}\right)$



b. $y = \frac{sin\,(10x\pi)}{10x\pi}$

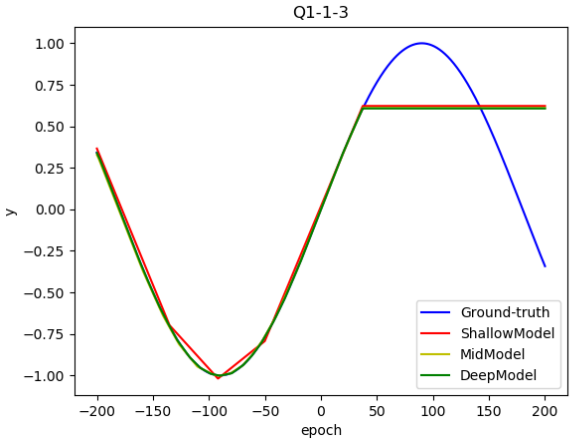


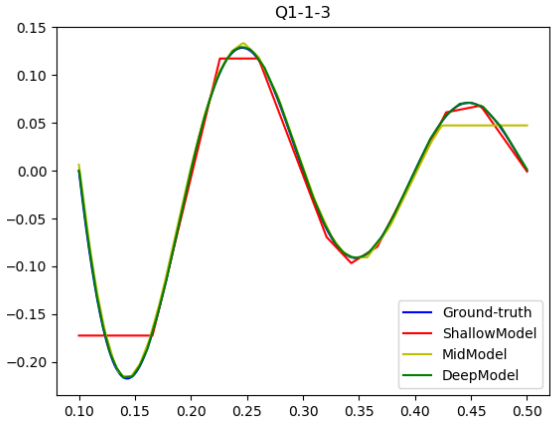
3.In one graph, plot the predicted function curve of all models and the ground-truth function curve.(0.5%)

Ans:

a. $y = sin\left(\frac{x\pi}{180}\right)$

b. $y = \frac{sin(10x\pi)}{10x\pi}$





4.Comment on your results. (1%)

Ans:

經比較可以發現，在參數接近的情況之下，較深的model可以較為擬合要近似的曲線。而shallow的模型較不擬合，推測某些地方在經過平均點後，便達到了local minima，而不再繼續減少loss。

- Train on Actual Tasks

1.Describe the models you use and the task you chose. (0.5%)

Ans:

a.1-hidden layer with 1583866 params

一層3*3，size為612的conv layer，再進入2*2的pooling layer，最後攤平後送至10維的softmax output layer。共有1583866個parameters。

其中，optimizer為Adadelta，lr=0.1, rho=0.95, epsilon=1e-08，batch size為256。loss function為cross entropy，epoch為10000。

b.4-hidden layer with 1554026 params

一層3*3，size為64的conv layer，再進入2*2的pooling layer，再進入一層3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為256的conv layer，再進入2*2的pooling

layer，接著進入3*3，size為480的conv layer，再進入2*2的pooling layer，最後攤平後送至10維的softmax output layer。共有1554026個parameters。

其中，optimizer為Adadelta，lr=0.1, rho=0.95, epsilon=1e-08，batch size為256。loss function為cross entropy，epoch為10000。

c.8-hidden layer with 1554026 params

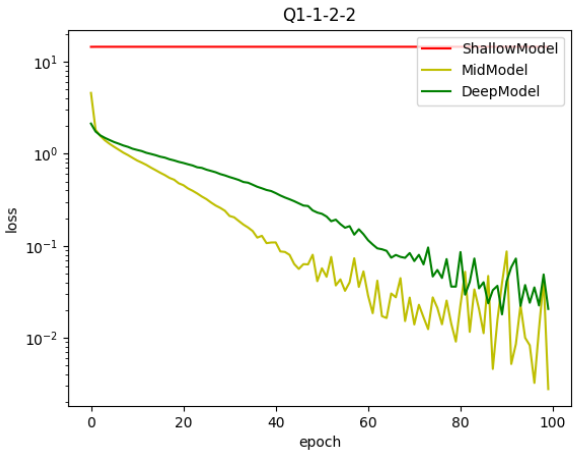
一層3*3，size為64的conv layer，再進入2*2的pooling layer，再進入一層3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為128的conv layer，再進入2*2的pooling layer，接著進入3*3，size為256的conv layer，再進入2*2的pooling layer，接著進入3*3，size為256的conv layer，再進入2*2的pooling layer，最後攤平後送至10維的softmax output layer。共有1554026個parameters。

其中，optimizer為Adadelta，lr=0.1, rho=0.95, epsilon=1e-08，batch size為256。loss function為cross entropy，epoch為10000。

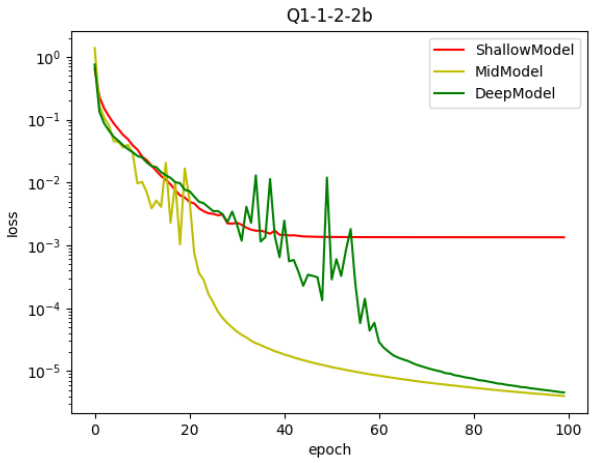
2.In one chart, plot the training loss of all models. (0.5%)

Ans:

a.cifar10



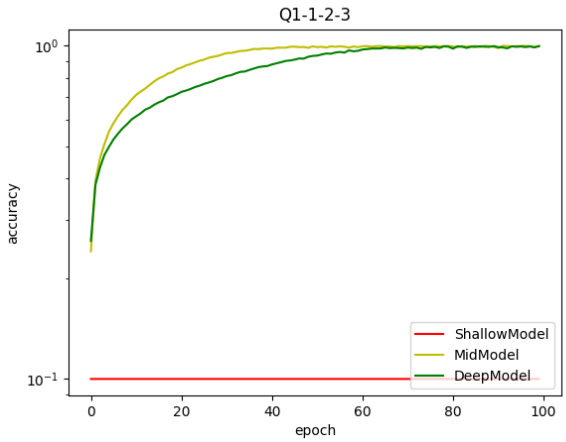
b.Mnist



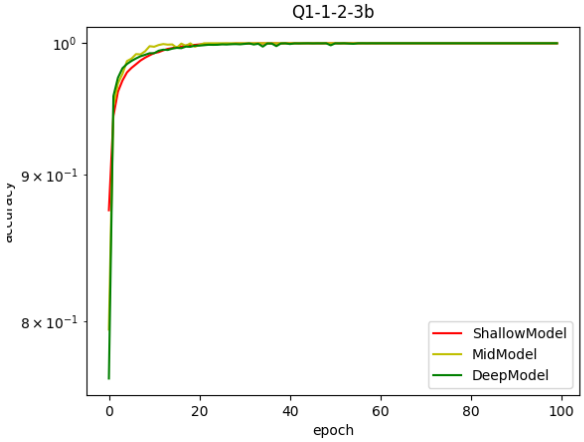
3.In one chart, plot the training accuracy. (0.5%)

Ans:

a.cifar10



b.Mnist



4.Comment on your results. (1%)

由實驗結果可以明顯看出，只有一層hidden-layer的CNN，在兩種不同的task下都明顯表現較其他兩種model差，雖然參數幾乎相同。

而單純比較4層與8層hidden-layer的狀況下，會發現在最後兩者皆達到了接近1的準確度，不分上下。

但單就過程而言，反倒是較深的8層hidden-layer model，準確率提升的較4層的慢。

- HW1-2
- How to reach the point where the gradient norm is zero?

1.Describe your experiment settings. (The cycle you record the model parameters, optimizer, dimension reduction method, etc) (1%)

Ans:

在mnist上訓練，用fully connected neural network model，input 28*28的影像，第一層128個neural，第二層32個neural，最後output一個10維的向量。

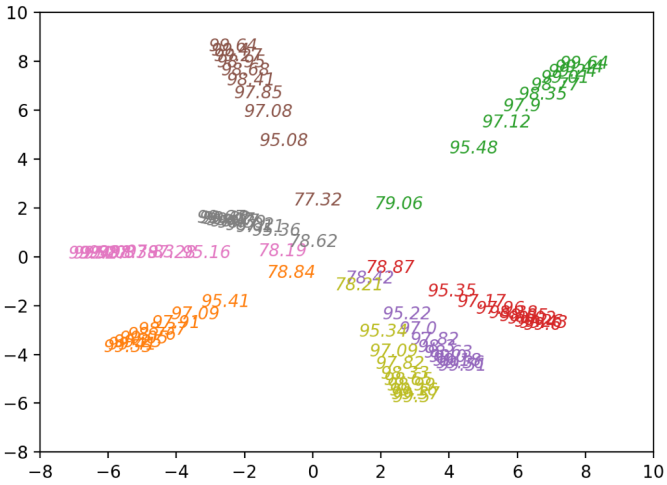
optimizer用SGD(lr=0.01, momentum=0.9)

loss function用cross entropy

Dimension reduction用sklearn.decomposition的PCA

2.Train the model for 8 times, selecting the parameters of any one layer and whole model and plot them on the figures separately. (1%)

Ans:



3.Comment on your result. (1%)

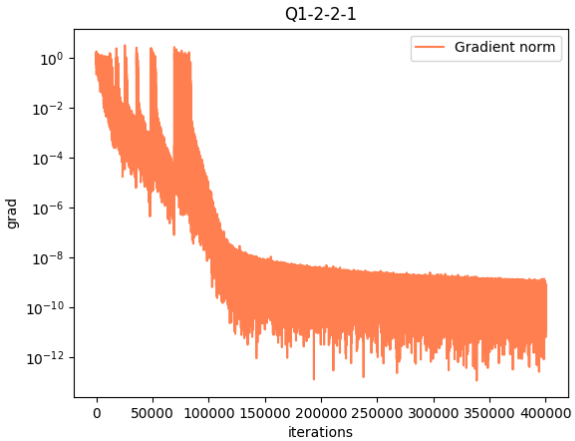
Ans:

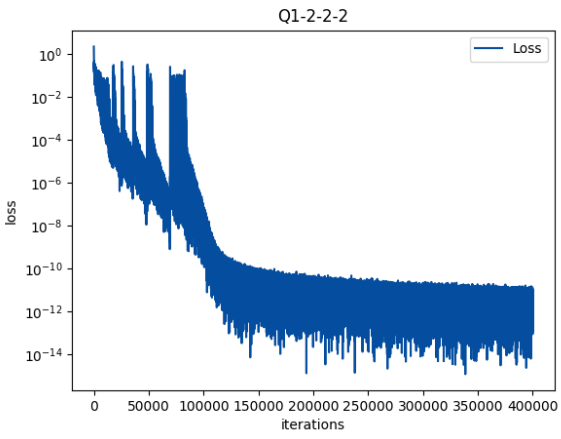
視覺化的結果發現accuracy較低的地方越靠近圖的中間，accuracy較高的地方較靠近圖的外側，推測係數的optimum應該在圖的外側。

- Observe gradient norm during training.

1.Plot one figure which contain gradient norm to iterations and the loss to iterations. (1%)

Ans:





2.Comment your result. (1%)

Ans:

由第一題中的兩張圖可以發現，其實gradient norm與loss的關係非常密切，若norm下降，則loss也在下降，兩個曲線相似。在70000次iterations後，可以發現loss與gradient norm都大幅下降，接著便維持在近乎0的地方，也代表到達了local或global minimum。

- What happens when gradient is almost zero?

1.State how you get the weight which gradient norm is zero and how you define the minimal ratio. (2%)

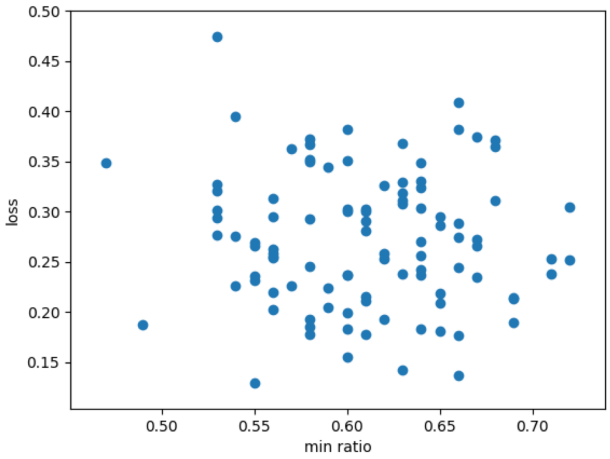
Ans:

將model用原本的loss train到收斂之後，將loss改成gradient norm繼續train，即可將gradient norm變小(但很難接近零)。

minimal ratio的計算採用：在要計算的點的附近sample100個點，loss比該點高的比率即為minimal ratio。

2.Train the model for 100 times. Plot the figure of minimal ratio to the loss. (2%)

Ans:



3.Comment your result. (1%)

Ans:

從上圖看minimal ratio大約分佈在[0.5, 0.75]，點很分散，看不出跟loss有什麼相關性。但是minimal ratio幾乎都大於0.5，推測gradient descent找到的點是有一定參考性的。

- HW1-3
- Can network fit random variables?

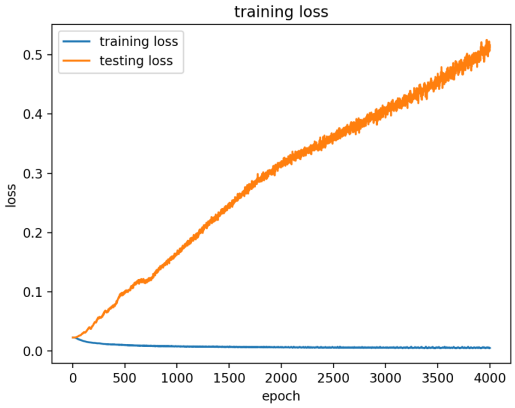
1.Describe your settings of the experiments. (e.g. which task, learning rate, optimizer) (1%)

Ans:

Dataset用mnist，用fully connected neural network model，input 28*28的影像，第一層256個neural，第二層128個neural，最後output一個10維的向量。optimizer用SGD(lr=0.01, momentum=0.9)，loss function用cross entropy。

2.Plot the figure of the relationship between training and testing, loss and epochs. (1%)

Ans:



- Number of parameters v.s. Generalization

1.Describe your settings of the experiments. (e.g. which task, the 10 or more structures you choose) (1%)

Ans:

Dataset用mnist，模型用fully connected neural network model，optimizer用SGD(lr=0.01, momentum=0.9)，loss function用cross entropy，epoch=10，不然accuracy會普遍太高

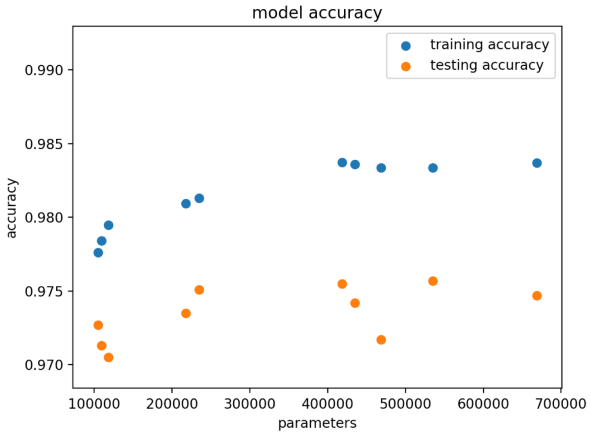
model每層的neural 數，扣除第一層28*28的input和10的output

Model name	Neurals of 1st layer	Neurals of 2nd layer
model1	256	128
model2	256	64
model3	128	64
model4	128	128

model5	512	256
model6	512	512
model7	512	128
model8	512	64
model9	512	32
model10	128	32

2.Plot the figures of both training and testing, loss and accuracy to the number of parameters. (1%)

Ans:



3.Comment your result. (1%)

Ans:

參數越多，training的accuracy有升高、loss有降低的趨勢，testing的結果比較震盪，但參數多的model準確率還是普遍較參數少的model準確率高，只是上升程度有限、loss下降程度也有限，表示參數越多、模型越複雜，對準確率/loss的影響還是有限。

- Flatness v.s. Generalization

Part1

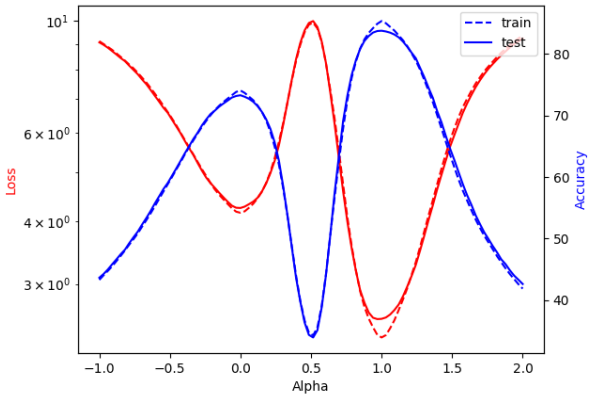
1.Describe the settings of the experiments (e.g. which task, what training approaches) (0.5%)

Ans:

本題實驗在mnist上，model為兩層hidden layer的架構，size分別為256、128，而最後一層為十維的softmax，Optimizer採用Adadelta，learning rate為0.1，rho為0.95, epsilon為1e-08，loss採用cross entropy實驗變數為batch size，使用128與1024兩種來進行比較。

2.Plot the figures of both training and testing, loss and accuracy to the number of interpolation ratio. (1%)

Ans:



3.Comment your result. (1%)

Ans:

由part1的圖可以看出，loss與accuracy呈鏡像的對稱，在alpha為0與1處為使用單一模型來進行測試的數據，可以看出個別的模型都能得到超過70%的準確率，但這也是內差後的模型能獲得最高準確度的兩個maxima處，往兩側則準確度會再次下降。由圖還可以發現，在兩個模型參數取平均的地方，也是一個local minima。

Part2

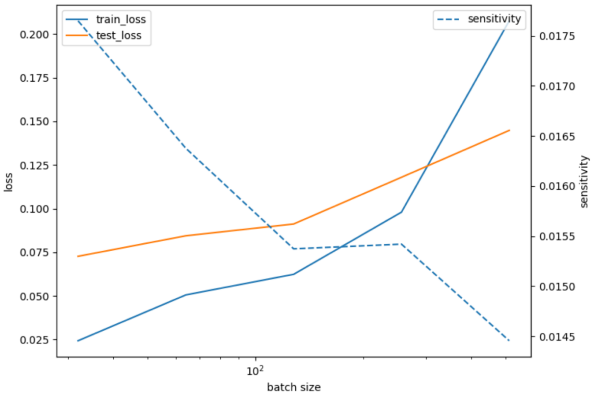
1.Describe the settings of the experiments (e.g. which task, what training approaches) (0.5%)

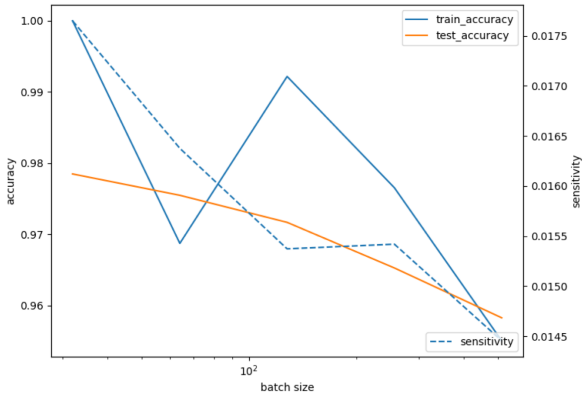
Ans:

本題用1 hidden layer dnn實驗在mnist上(hidden size=128)，Optimizer採用Adam，learning rate=1e-3，實驗變數為batch size={2^5, 2^6, 2^7, 2^8, 2^9}。Sensitivity的定義採用：「Frobenius norm of gradients of loss to input」

2.Plot the figures of both training and testing, loss and accuracy, sensitivity to your chosen variable. (1%)

Ans:





3.Comment your result. (1%)

Ans:

很明顯得可以看到，batch size越大，sensitivity越小，似乎跟上課所述相反。

但有可能是因為對sensitivity定義不同所導致，原本為「Frobenius norm of Jacobian matrix of model output (class probability) to input」，為簡化計算而改成「Frobenius norm of gradients of loss to input」。

- 分工表

楊碩碉：hw1-2-3、hw1-3-3-2

鄭雅文：hw1-2-1、hw1-3-1、hw1-3-2

陳品君：hw1-1、hw1-2-2、hw1-3-3-1