Homework 1 – Deep Learning (CS/DS541, Whitehill, Fall 2024)

Ashwin Disa, Manoj Velmurugan

Q2.

Parameter
Number of Epochs - 1000
Learning rate - 1e-2
Input size - 48*48
Hidden size - 64
Output size - 1
Train:val - 9:1

Epoch: 0, Training Loss: 1775.2651274594298, Validation Loss: 1511.7177229525662
Epoch: 100, Training Loss: 541.4972022592882, Validation Loss: 570.094799970118
Epoch: 200, Training Loss: 268.2796873638299, Validation Loss: 286.2836231216435
Epoch: 300, Training Loss: 231.9070857303789, Validation Loss: 245.78145440267872
Epoch: 400, Training Loss: 218.73087460581664, Validation Loss: 229.21043785871467
Epoch: 500, Training Loss: 219.8483208730668, Validation Loss: 226.16514486587027
Epoch: 600, Training Loss: 217.3758356566733, Validation Loss: 226.25482954355718
Epoch: 700, Training Loss: 240.04346621864337, Validation Loss: 227.7359612049683
Epoch: 800, Training Loss: 242.44212719610647, Validation Loss: 231.25296712317217
Epoch: 900, Training Loss: 223.11028421024147, Validation Loss: 239.20073689069707

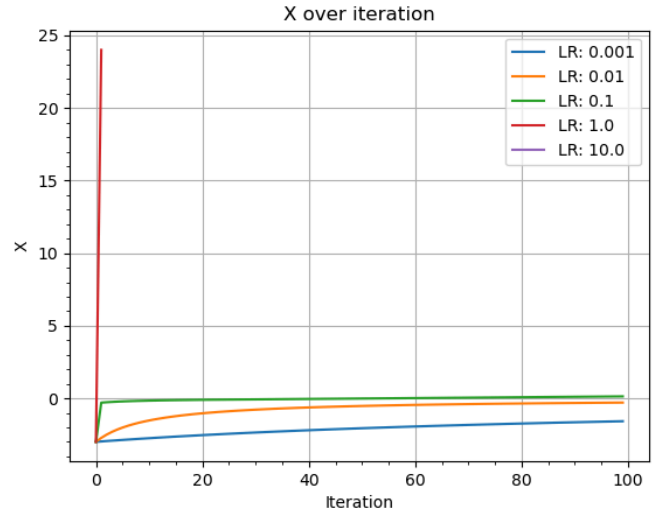
Last 10 iterations
Training Loss: 229.32640215564854
Training Loss: 266.4009998561433
Training Loss: 224.1200747776349
Training Loss: 224.09266021238977
Training Loss: 226.7081924983344
Training Loss: 259.0717499918084
Training Loss: 224.11250827835474
Training Loss: 227.766804080435
Training Loss: 265.9897288197635

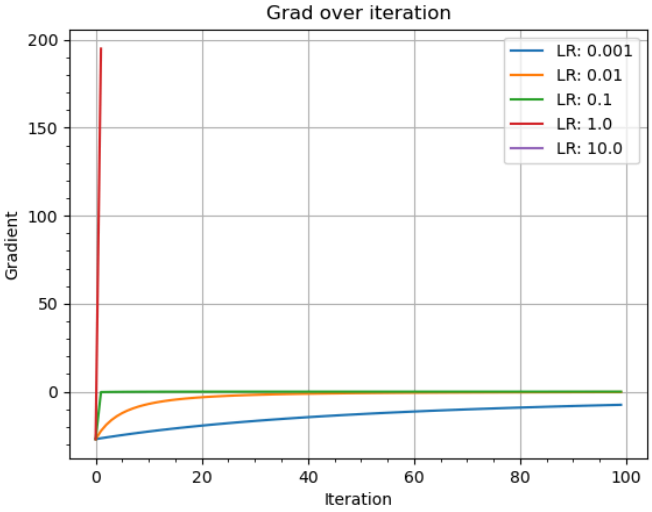
Final losses
Training Loss: 265.9897288197635, Validation Loss: 239.20073689069707

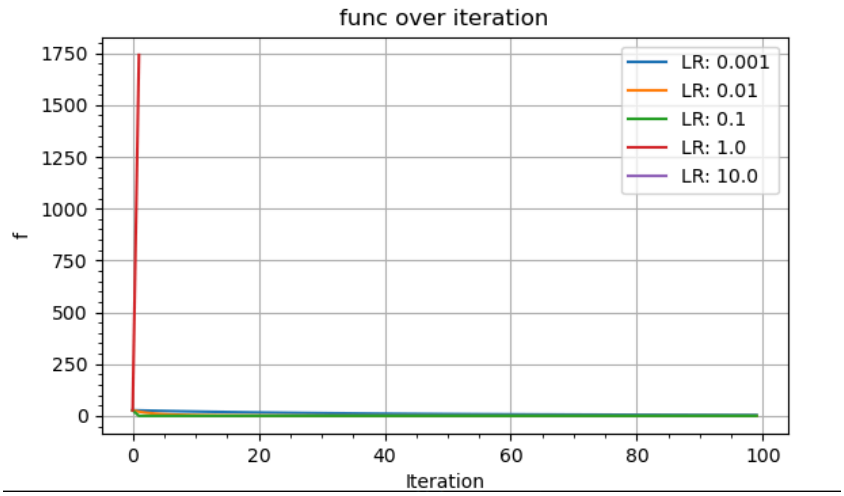
Test Loss: 230.07504908595635

Question 3:

3.ii

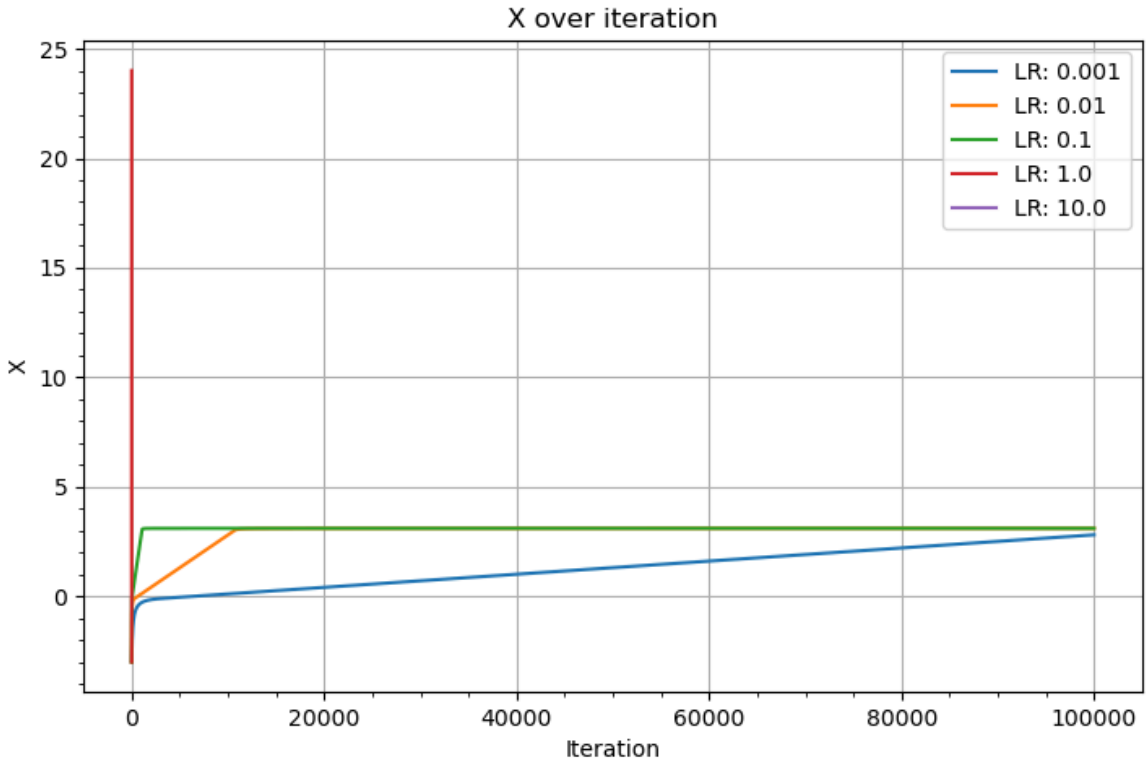




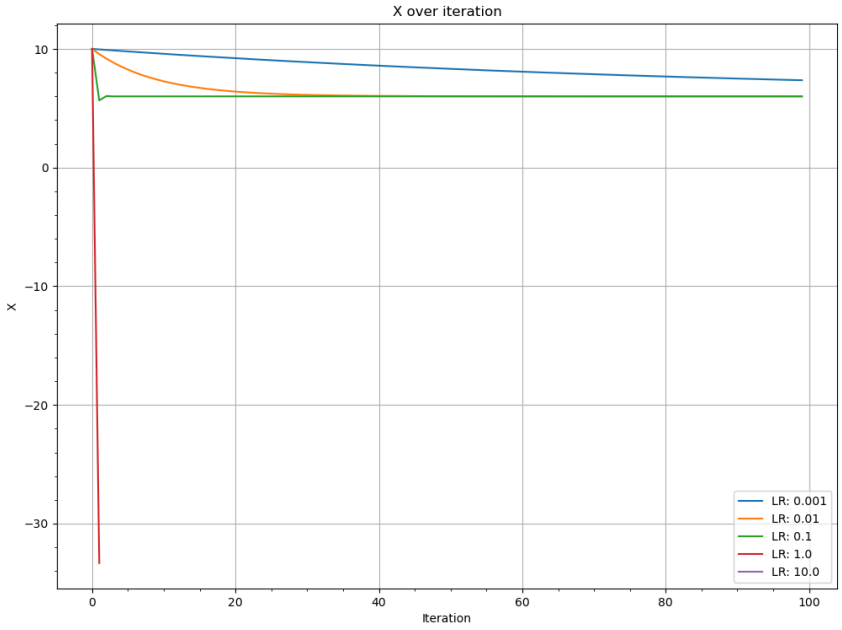


3.iii. The gradients are going to zero and the progress towards global minima stops completely.

3.iv - Looks like we have a local minima at x = 3.1 where the gradient goes to zero. Hence when we start from -3, it always gets stuck at 3.1 as evident in the plot below,



Starting from 10 settles correctly at 6 for 1e-2 and 1e-1 LR.



3.b:

I. The pattern/trajectory will not change. It will still take too long to converge.

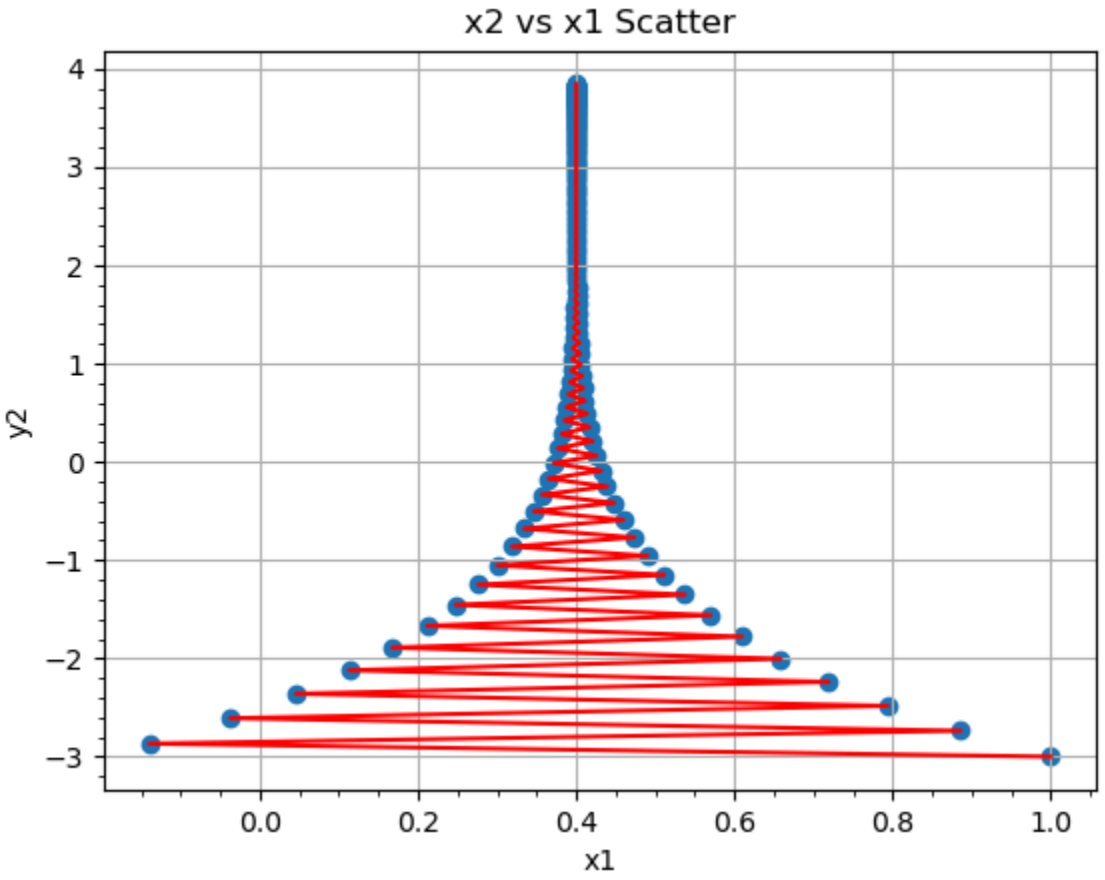
Ii. Just by looking at the sample trajectory one can say that c1 is close to 0.4 and c2 is close to 4.

The jump along x1 direction is higher indicating that x1 gradient is really high probably because of a high a1.

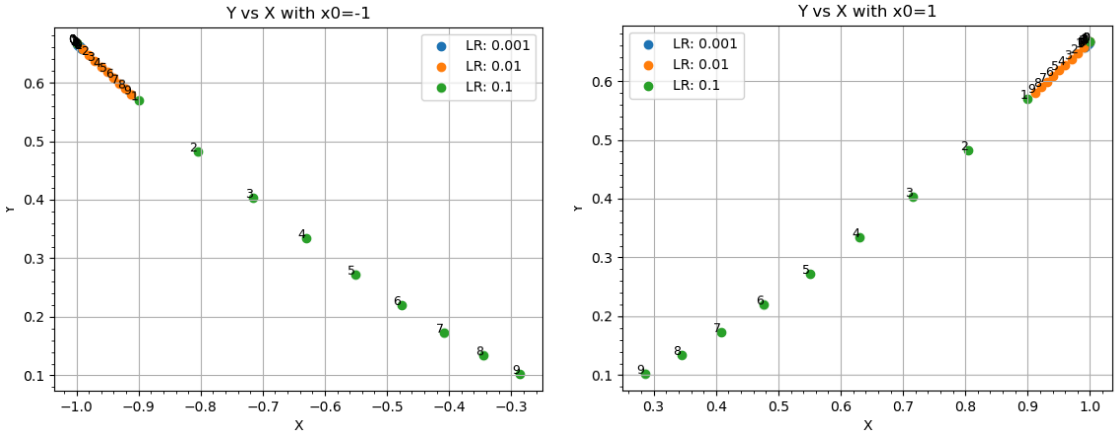
using , [a1, a2, c1, a2] as [100, 1, 0.4, 4] and using a lr of 0.0095 yields the following scatter plot for 200 iterations.

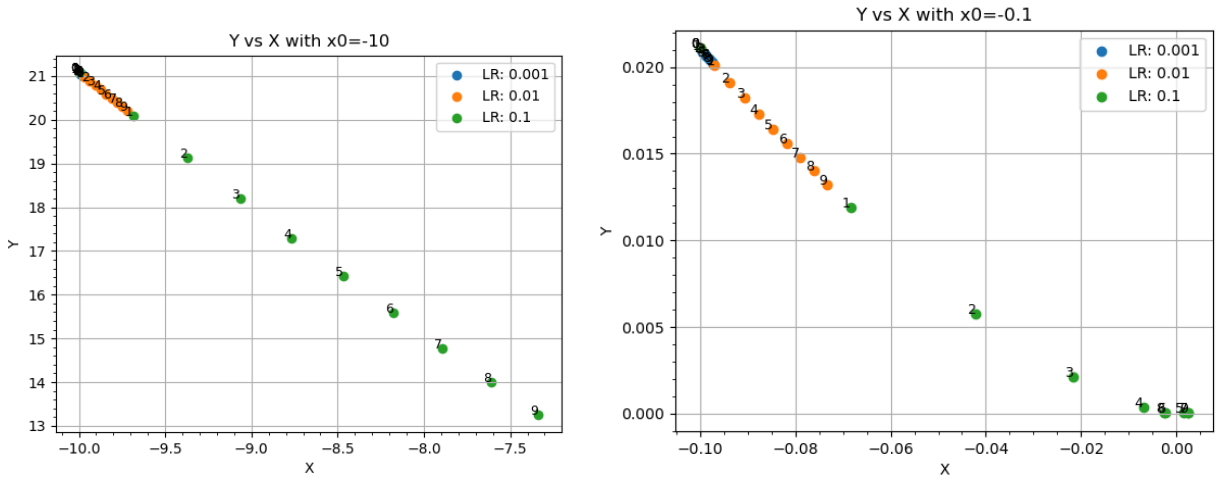
This problem is called an illconditioned problem. Because of the behavior of x2 and x1 are very different, using a single high learning rate for both x1 and x2 will lead to one of them converging stably and the other one violently oscillating. This problem forces one to lower LR but that can decrease the settling time greatly (just because of one rogue variable). It can be

avoided by scaling the first part of the loss function or the variable x1 with a linear transformation.



3.C:





For low LR, it converges so slowly that it may never reach 0. The behavior is seen for different starting points.

3. D: