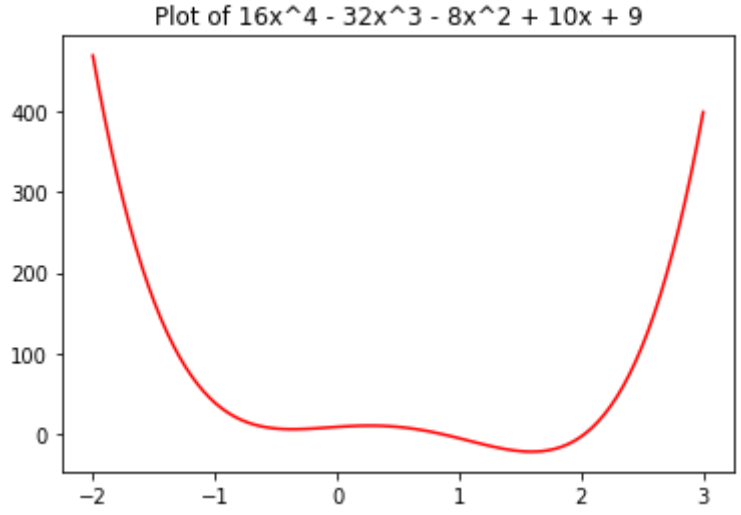
RAKSHIT SAREEN (rs5606)

PART II Solutions (**Source Code is present in accompanying Jupyter notebook**)

Q1
The plot of the function is



As we can see the function has two minima, one local and one global.

a) The value x for local minimum : [-0.36425]
   The value x for global minimum : [1.5953125]
b) eta = 0.001 , x = -1 and 5 iterations
   The values for x, f(x)
   [(-0.866, 24.122076702976),
   (-0.776294662656, 17.196933470898333),
   (-0.7109220204970491, 13.432358919601501),
   (-0.6607817416581312, 11.182090980505741),
   (-0.6209723259625907, 9.746936730571225)]

   eta = 0.001 , x = -1 and 1000 iterations
   The values for x, f(x)
   [(-0.36422374257061546, 6.124226461427285),
   (-0.36422374257061546, 6.124226461427285),
   (-0.36422374257061546, 6.124226461427285),
   (-0.36422374257061546, 6.124226461427285),
   (-0.36422374257061546, 6.124226461427285)]

Yes, according to the values above, we can notice that the value of x has converged. Gradient Descent has found a minimum. The value of x at which gradient descent has found a minimum is the local minimum. We can say this because the value of x is negative (-0.36) where the local minima occurs as we can verify from the graph plotted above

c) eta = 0.001 , x = 2 and 5 iterations
The values for x, f(x) are
[(1.894, -12.280893572864002),
(1.823848257024, -16.471988590606028),
(1.7740854376290425, -18.621275496013634),
(1.7372611541368062, -19.813420352480247),
(1.7092287072971393, -20.51063658566802)]

eta = 0.001 , x = 2 1000 iterations
[(1.5953147000510854, -21.69623302490436),
(1.5953147000510852, -21.69623302490436),
(1.5953147000510854, -21.69623302490436),
(1.5953147000510852, -21.69623302490436),
(1.5953147000510854, -21.69623302490436)]

Yes, according to the values above, we can notice that the value of x has converged. Gradient Descent has found a minimum. The value of x at which gradient descent has found a minimum is the global minimum. We can say this because the value of x is positive (1.595) where the global minimum occurs as we can verify from the graph plotted above

d) eta = 0.01 , x = -1 and 1000 iterations
First Five values
[(0.3400000000000001, 10.43128576),
(0.38022144000000013, 10.221092268269524),
(0.44466297051775, 9.676878465941176),
(0.5493557211358288, 8.231163257954496),
(0.7208664355233224, 4.384921376586129)]

Next Five Values
[(0.9953254006215324, -4.822543299773962),
(1.3745565522920227, -18.359057075000234),
(1.646174283580392, -21.471761448124713),
(1.5560478035180898, -21.572196325826845),
(1.6181625286423753, -21.651974555361676)]

Last Five Values
[(1.5953147000510854, -21.69623302490436),
(1.5953147000510852, -21.69623302490436),
(1.5953147000510854, -21.69623302490436),
(1.5953147000510852, -21.69623302490436),
(1.5953147000510854, -21.69623302490436)]

With the new value of eta we were able to reach the global minimum even with the same starting point as -1. Using too small eta made gradient descent stay within the local minimum and could not et out of the trap. Using a larger eta changed the problem.

e)

x = −1 and η = 0.05
Overflow Detected in f(x)
Overflow Detected in f(x)

Explanation

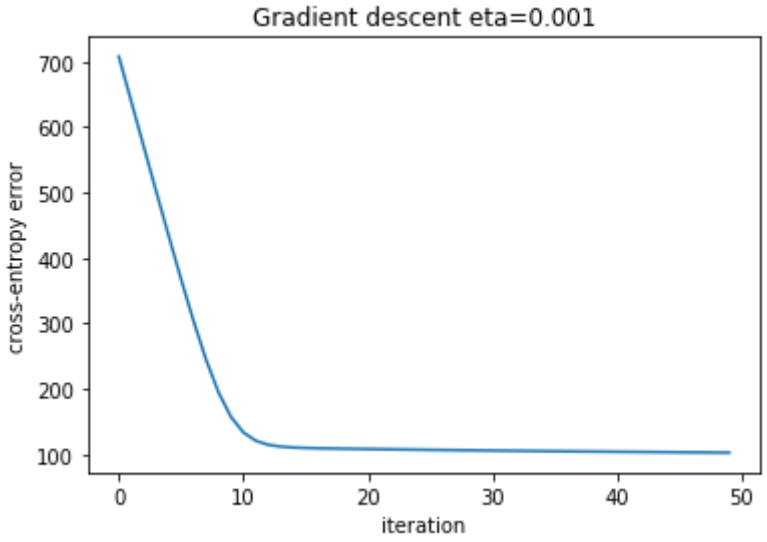
Our function has only one local minima and one global maxima. When we set eta to 0.05 (particularly >.03) we get Overflow error. This is happening because at the `x` which is being produced by gradient descent, our function is almost straight. Therefore both the derivative and the function are Overflowing (tending to infinity)

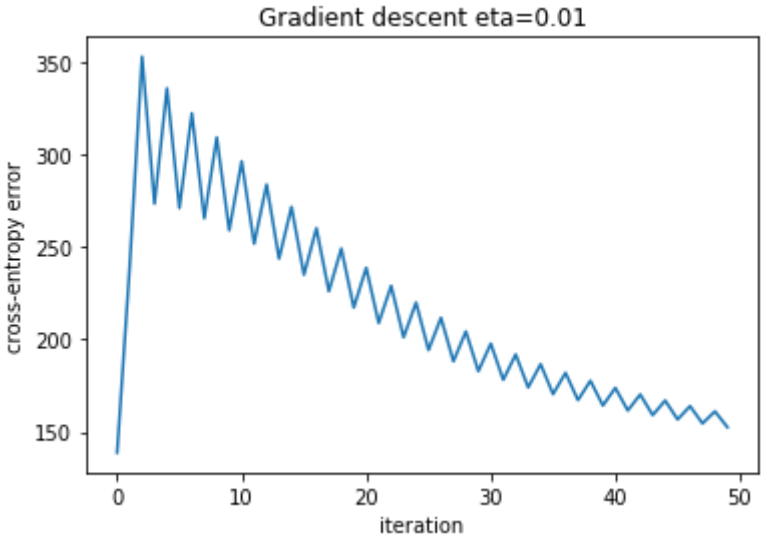
Q2

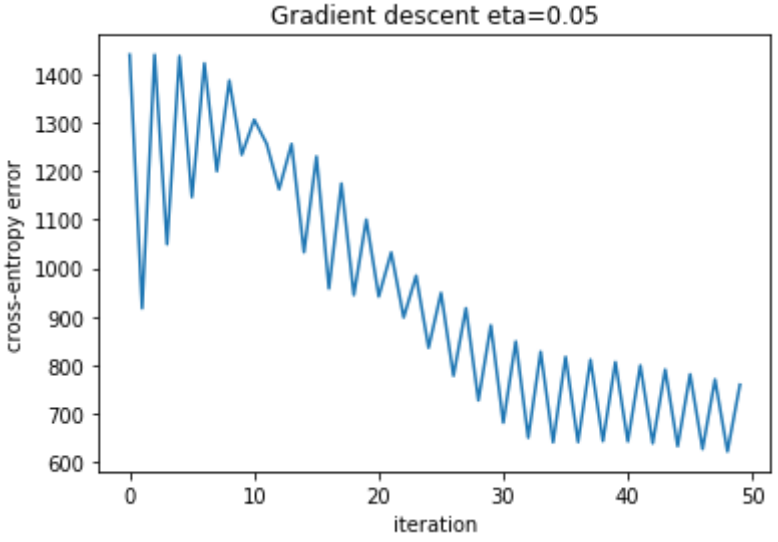
- a)
  - i)

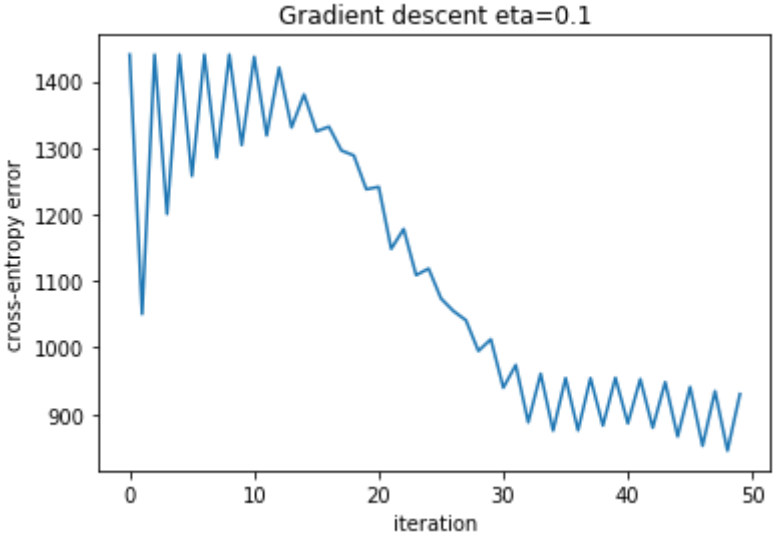
Rate 0.001 cross error :: 102.695001295 classification error :: 28.3333333333% and L2Norm :: 2.61565476966
Rate 0.01 cross error :: 152.48539954 classification error :: 36.1111111111% and L2Norm :: 6.59687582841
Rate 0.05 cross error :: 759.86320188 classification error :: 36.1111111111% and L2Norm :: 32.7597411458
Rate 0.1 cross error :: 930.220966831 classification error :: 36.6666666667% and L2Norm :: 65.0075032233
Rate 0.5 cross error :: 1024.83557566 classification error :: 36.6666666667% and L2Norm :: 324.2618561
Rate 1.0 cross error :: 1047.44078536 classification error :: 36.6666666667% and L2Norm :: 649.271427939
Rate 1.5 cross error :: 1056.34449418 classification error :: 36.6666666667% and L2Norm :: 973.892726438

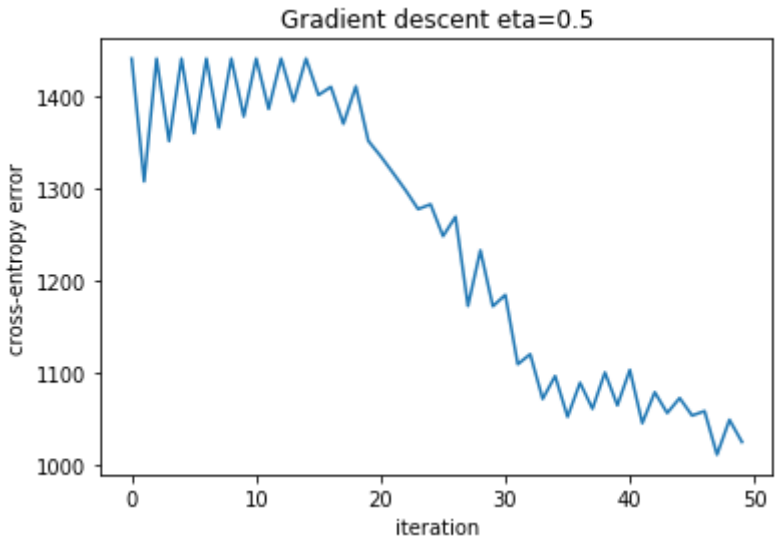
Graphs are as follows:

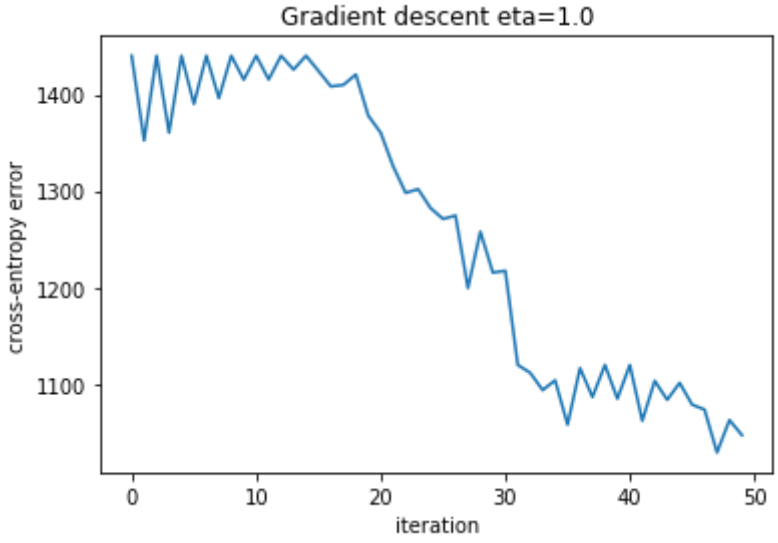


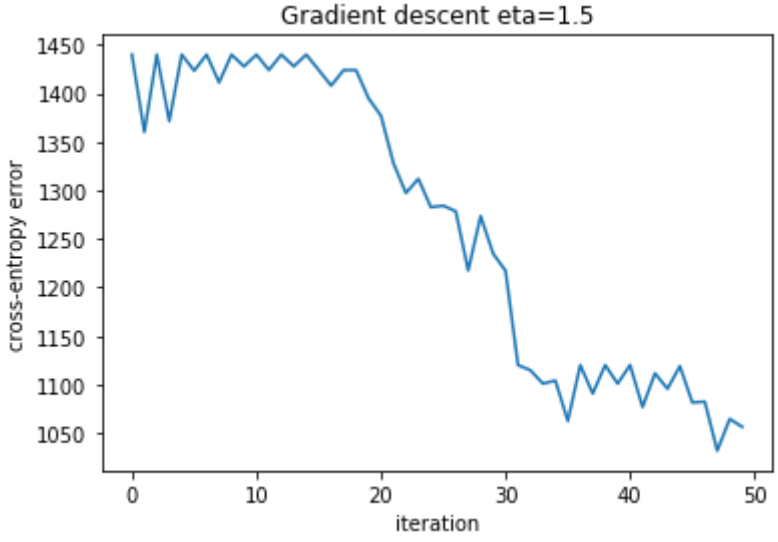

Gradient descent eta=0.01
cross-entropy error
iteration
350
300
250
200
150
0
10
20
30
40
50


Gradient descent eta=0.05
cross-entropy error
iteration
1400
1300
1200
1100
1000
900
800
700
600
0
10
20
30
40
50


Gradient descent eta=0.1
cross-entropy error
iteration
1400
1300
1200
1100
1000
900
0
10
20
30
40
50


Gradient descent eta=0.5
cross-entropy error
iteration
1400
1300
1200
1100
1000
0
10
20
30
40
50


Gradient descent eta=1.0
cross-entropy error
iteration
1400
1300
1200
1100
0
10
20
30
40
50


Gradient descent eta=1.5
cross-entropy error
iteration
1450
1400
1350
1300
1250
1200
1150
1100
1050
0
10
20
30
40
50

b)

Lambda 0 cross entropy error :: 102.695001 classification error :: 28.333% and L2Norm :: 2.61565476966
Lambda 0.05 cross entropy error :: 102.884414 classification error :: 28.333% and L2Norm :: 2.60974858562
Lambda 0.1 cross entropy error :: 103.072319 classification error :: 28.333% and L2Norm :: 2.60385691692
Lambda 0.2 cross entropy error :: 103.443651 classification error :: 28.333% and L2Norm :: 2.59211698406
Lambda 0.3 cross entropy error :: 103.809078 classification error :: 27.777% and L2Norm :: 2.58043468893
Lambda 0.4 cross entropy error :: 104.168684 classification error :: 27.777% and L2Norm :: 2.56880975075
Lambda 0.5 cross entropy error :: 104.522549 classification error :: 27.777% and L2Norm :: 2.55724189014

Lambda := 0 CV Error := 30.0
Lambda := 0.005 CV Error := 30.0
Lambda := 0.1 CV Error := 30.0
Lambda := 0.2 CV Error := 29.7777777778
Lambda := 0.3 CV Error := 29.6666666667
Lambda := 0.4 CV Error := 29.6666666667
Lambda := 0.5 CV Error := 29.6666666667

c)

When we increased the value of λ, then the weights were not allowed to have much higher values as higher lambda acts as larger weight decay in the weights which does not let the weights to increase.
Due to this, the effect is that the cross entropy error increases, and the model become less flexible and thus error on the training set increases.
Since the weights are not allowed to have higher values due to regularization and thus the L2 norm will decrease overall.
The cross validation classification error decreases.

Q3

The problem with choosing hyperparameters with cross validation is that it is very computationally expensive. Cross validation is used to figure out which degree model to use or the model parameters. This is possible because the weights we learn using regularization are generally good to use in other folds as well, but lambda is not.