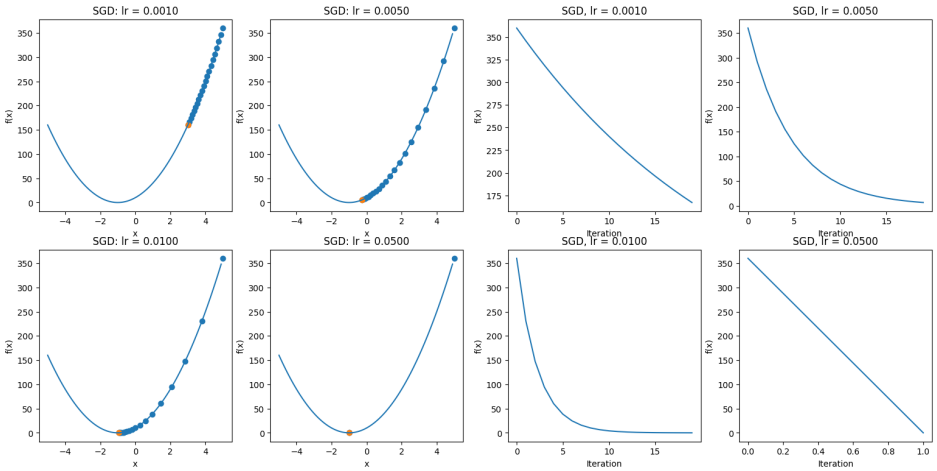
1. Gradient Descent
   1.1. Optimizer.sgd
      1.1.a.Test function q1()
         1.1.a.i. Describe the termination criteria used in the test_sgd function in the tests_A2.py file. (1 marks)

The termination criteria is defined by *update_thres*. If the update vector is smaller than *update_thres*, then the algorithm will terminate.

         1.1.a.ii. Include the figures generated by q1() in your PA2_qa.pdf file. (1 marks)

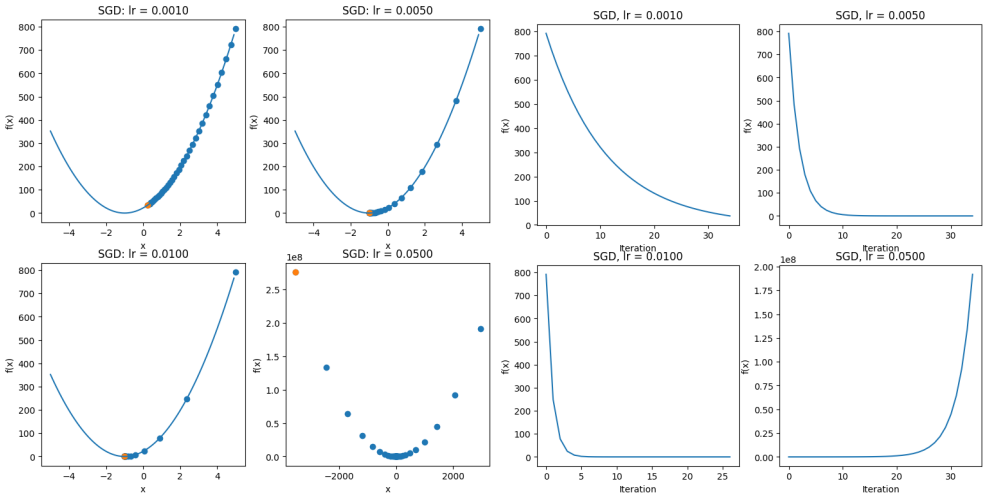



         1.1.a.iii. With learning rate η = 0.05, what would be the value of w1, i.e., after one iteration of SGD update? Show your mathematical process. If you implemented SGD correctly, the figures generated by q1() should verify your w1. (1 marks)

$w_0 = 5$
$w_1 = w_0 + (-0.05)*(20w_0+20) = -1$

1.1.b.Test function q2()

1.1.b.i. Include the figures generated by q2() in your PA2_qa.pdf file. (1 marks)

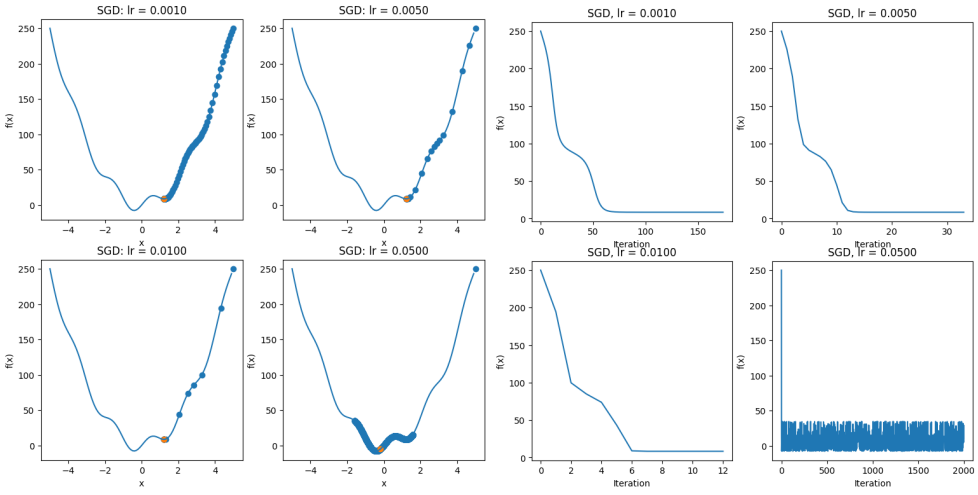



1.1.b.ii. When η = 0.05, SGD would fail to converge to the optimal solution. What causes such behavior? (1 marks)

The learning rate is set too high, causing SGD to overshoot.

1.1.c.Test function q3()

1.1.c.i. Include the figures generated by q3() in your PA2_qa.pdf file. (1 marks)



1.1.c.ii. In 1-2 sentences describe the behavior of SGD in q3() when η = 0.001, 0.005, and 0.01. Explain why SGD fails to find the global optimum point? (1 marks)

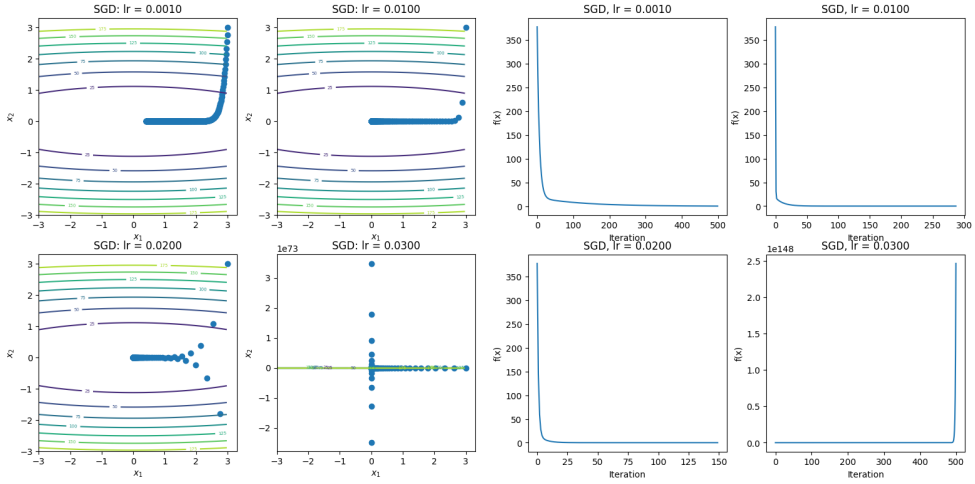
When η = 0.001, 0.05, and 0.01, SGD converges to a local minimum but fails to find the global minimum. This is because the function is not convex, causing SGD to be trapped in local minima.

1.1.c.iii. In 1-2 sentences describe the behavior of SGD in q3() when η = 0.05. (1 marks)

When η = 0.05, the learning rate is too high so SGD overshoots and bounces back and forth, which fails to converge.

1.1.d.Test function q4()

1.1.d.i. Include the figures generated by q4() in your PA2_qa.pdf file. (1 marks)




1.1.d.ii. In 1-2 sentences describe the behavior of SGD in q4() when η = 0.001 and 0.01. How is this behavior related to the stretched nature of the function f (w)? (1 marks)

When η=0.001 and 0.01, SGD converges. The function is more influenced by changes in $w_1$, so $w_0$ starts at (3, 3) in the top right corner, and quickly drops as small change in $w_1$ quickly updates $w_0$.

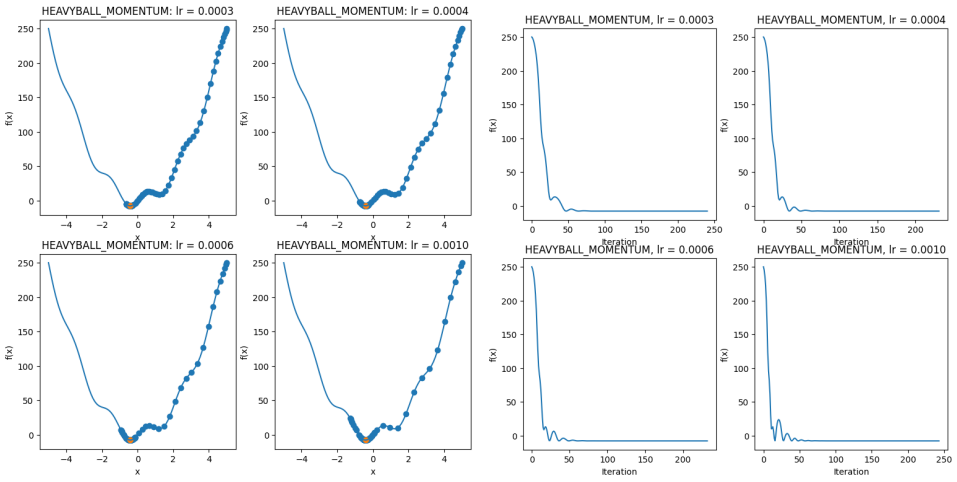
1.1.d.iii. In 1-2 sentences describe the behavior of SGD in q4() when η = 0.03. (1 marks)

When η=0.03, the learning rate is too high, which causes SGD to diverge and overshoot the minimum.

1.2. Optimizer.heavyball_momentum and Optimizer.nestrov_momentum

1.2.a.test function q5()

1.2.a.i. Include the figures generated by q5() in your PA2_qa.pdf file. (1 marks)

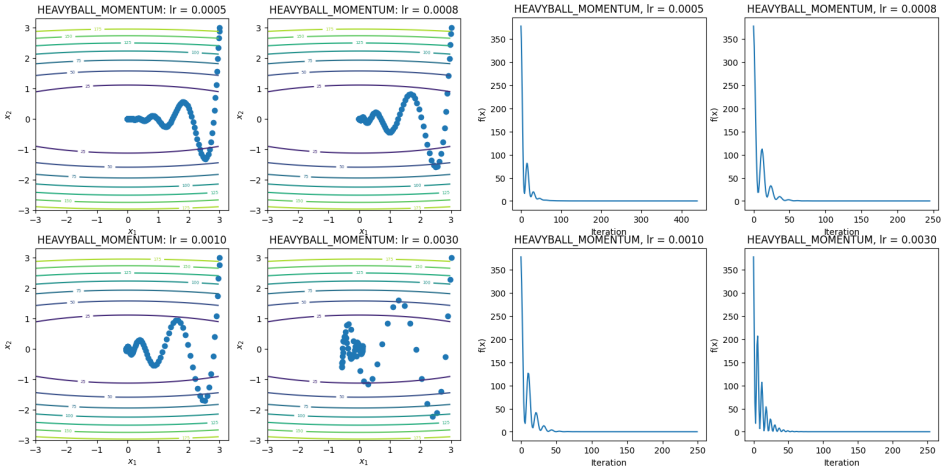



1.2.a.ii. In 1-2 sentences, compare the performance of SGD with and without heavy-ball momentum by comparing the outcome of tests q3() and q5() (2 marks)

Without heavy-ball momentum, SGD stops at local minimum. With heavy-ball momentum, SGD passes over local minimum and finds optimal solution

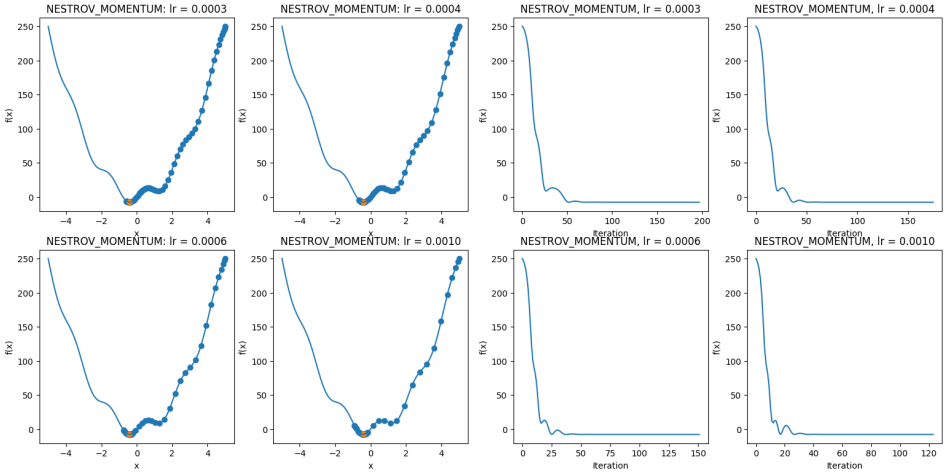
1.2.b.test function q6()

1.2.b.i. Include the figures generated by q6() in your PA2_qa.pdf file. (1 marks)



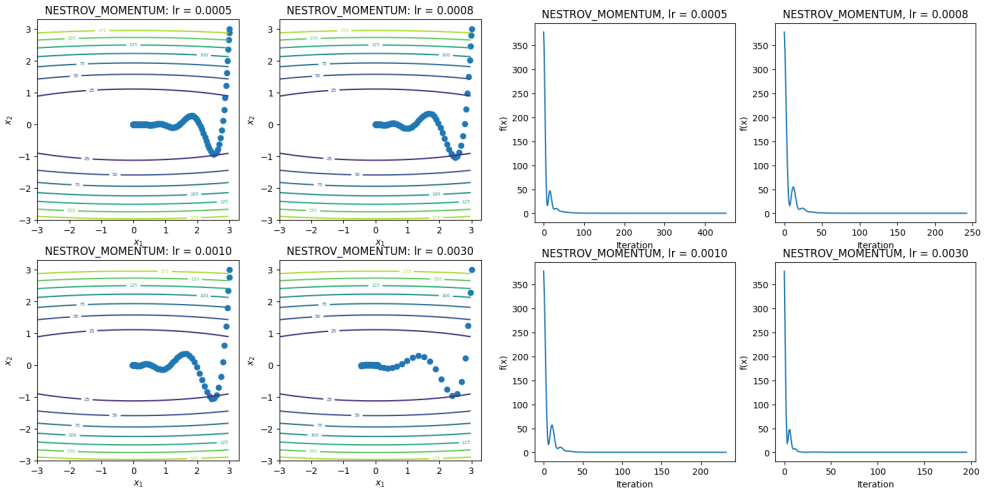

1.2.c.test function q7()

1.2.c.i. Include the figures generated by q7() in your PA2_qa.pdf file. (1 marks)




1.2.d.test function q8()

1.2.d.i. Include the figures generated by q8() in your PA2_qa.pdf file. (1 marks)



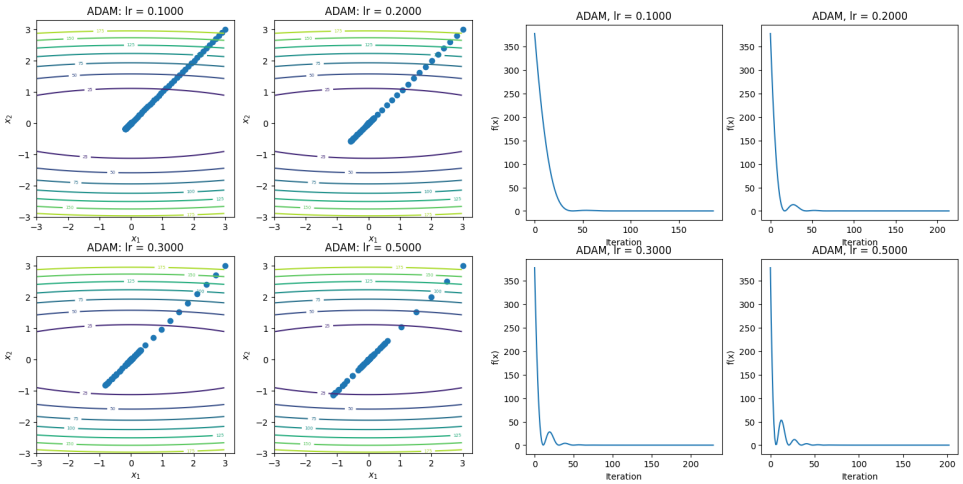

1.2.d.ii. In 1-2 sentences, compare the performance of Nestrov Momentum with the heavy-ball momentum by comparing the outcome of tests q5() and q6() with that of q7() and q8(). (1 marks)

Nestrov momentum performs more stable and smooth convergence, especially at higher learning rates (i.e., lr = 0.0030). The convergence happens more smoothly. It adds a fraction of the previous update vector to the current gradient.

1.3. Optimizer.adam

1.3.a.test function q9()

1.3.a.i. Include the figures generated by q9() in your PA2_qa.pdf file. (1 marks)

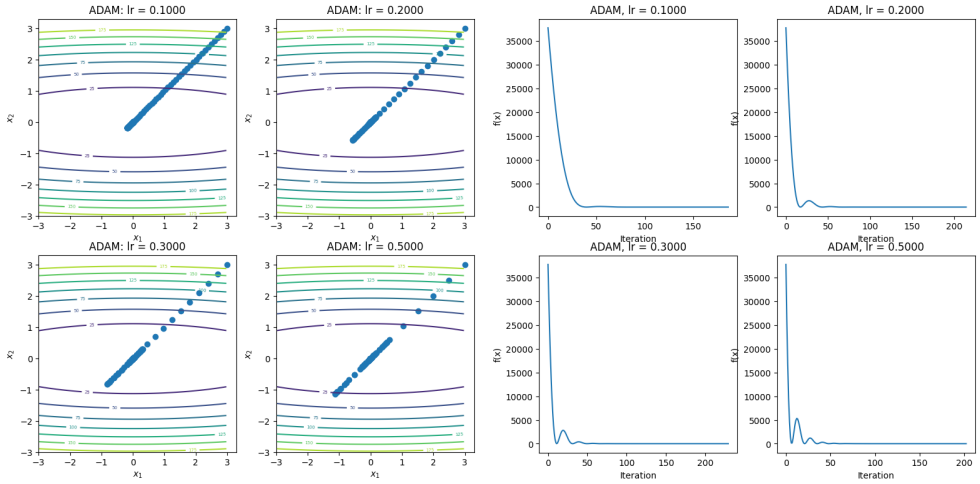


1.3.a.ii. In 1-2 sentences, compare the performance of adam with momentum method (heavy-ball or Nestrov) (2 marks)

Adam converges faster with fewer oscillations than the momentum method.

1.3.b.test function q10()

1.3.b.i. Include the figures generated by q10() in your PA2_qa.pdf file. (1 marks)




1.3.b.ii. Based on the outcome of q9() and q10(), describe the advantage of Adam in 1-2 sentence. (2 marks)

Adam is more robust against scaling of functions or gradients. The second moment helps Adam to scale the learning rate for each parameter.

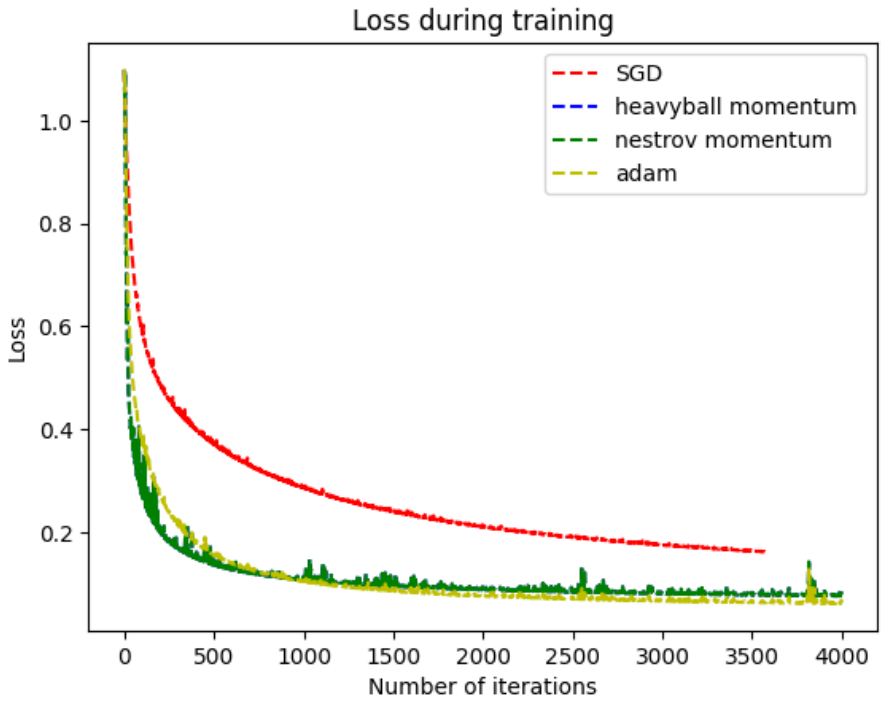
2. Multiclass Logistic Regression

2.1. Implementing the Learning Model

2.2. Implementing the Learning Algorithm

2.2.a.test function q22()

2.2.a.i. Include the figures generated by q22() in your PA2_qa.pdf file. (2 marks)




2.2.a.ii. In 1-2 sentences, compare the performance of the four variants of gradient descent on this dataset (2 marks)

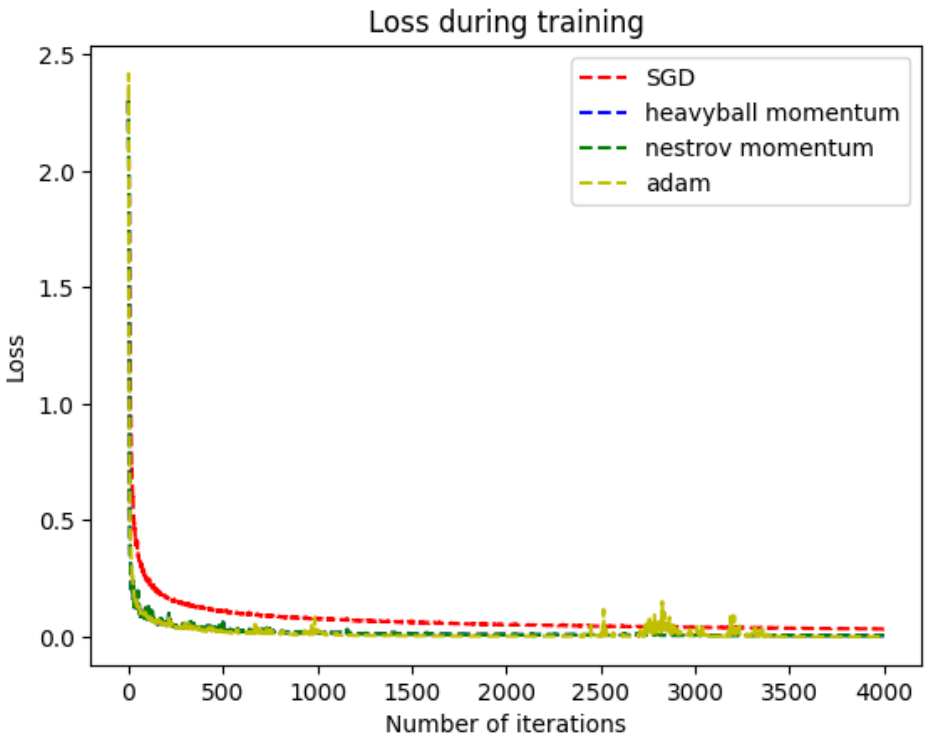
SGD converges the slowest with the highest loss among all four. The other three performs significantly better. Adam seems to have the least amount of oscillation among the four.

2.2.a.iii. In 1-2 sentences, explain how is it possible that the loss derived by the Adam optimizer is smaller than that of Heavy-ball Momentum, but the evaluation score of Adam is equal to the evaluation score of the heavy-ball momentum. (2 marks)

Loss is specific to how well the model fits the training data, but the evaluation score shows how well the algorithm performs with new data.

2.2.b.test function q23()

2.2.b.i. Include the figures generated by q23() in your PA2_qa.pdf file. (2 marks)




3. K-Means Clustering