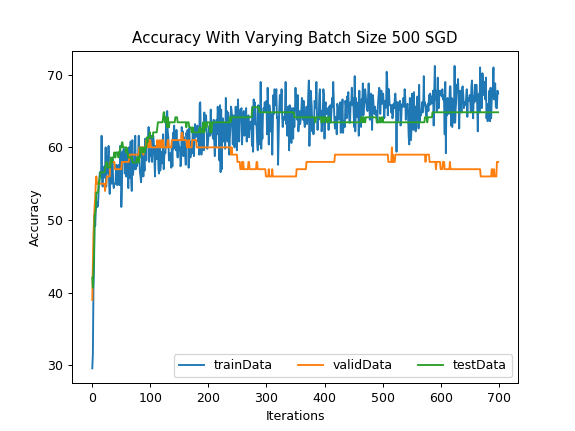
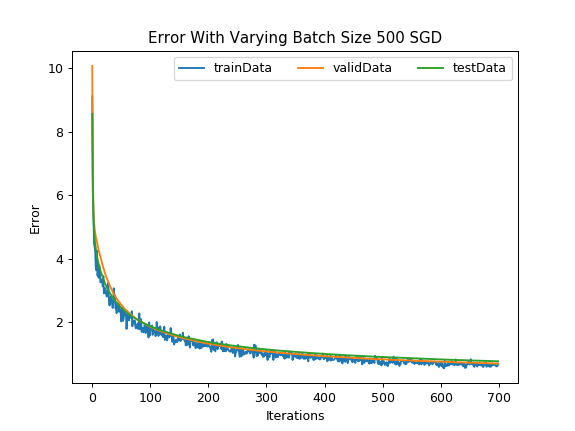
**Part 3 Batch Gradient Descent vs. SGD and Adam**

**1. MSE**

1. Loss and Accuracy graph of SGD(Gradient Descent optimizer) algorithm for a minibatch size of 500 optimizing over 700 epochs , minimizing the MSE. Other parameters: Lambda = 0, Learning rate = 0.001



Final training accuracy: 67.4

Final Training error: 0.671037

Final Validation accuracy: 57.99999999999999

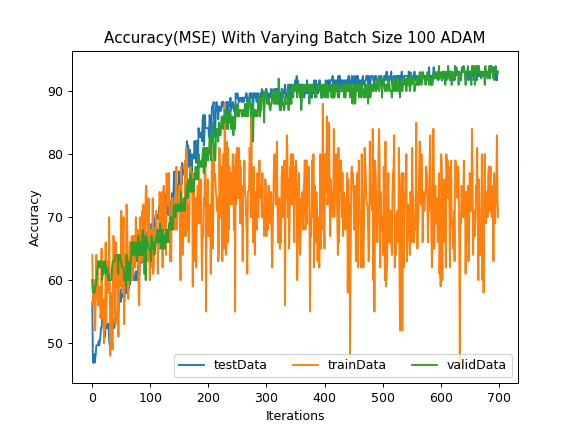
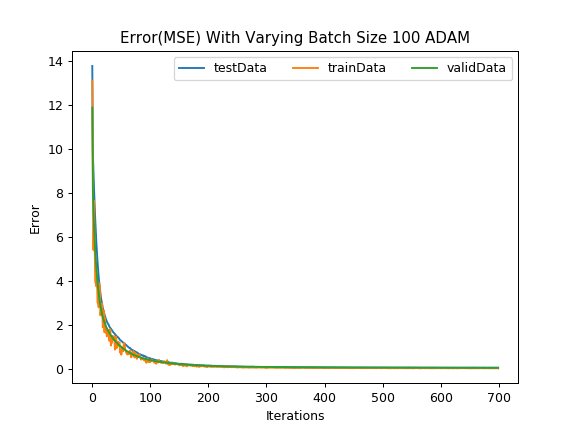
Final Validation error: 0.706172985749

Final Test accuracy: 64.82758620689654

Final Test error: 0.777026486581

2. SGD algorithm optimized using Adam by optimizing the model for minimizing the MSE. Used parameters: epochs =700, Lambda = 0, Learning rate = 0.001 with variable batch sizes:

1) Batch size = 100



Final training accuracy: 70.0

Final Training error: 0.00981315

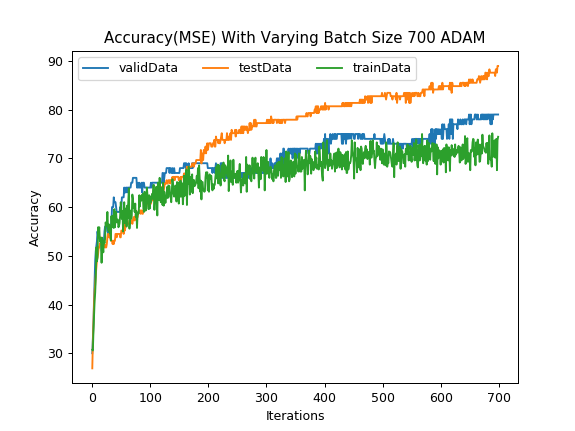
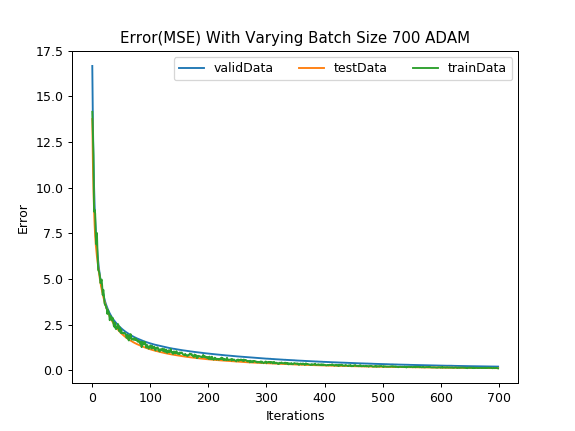
Final Validation accuracy: 93.0

Final Validation error: 0.0335003148801

Final Test accuracy: 93.10344827586206

Final Test error: 0.0389473910031

2) Batch size = 700



Final training accuracy: 67.0

Final Training error: 0.745089

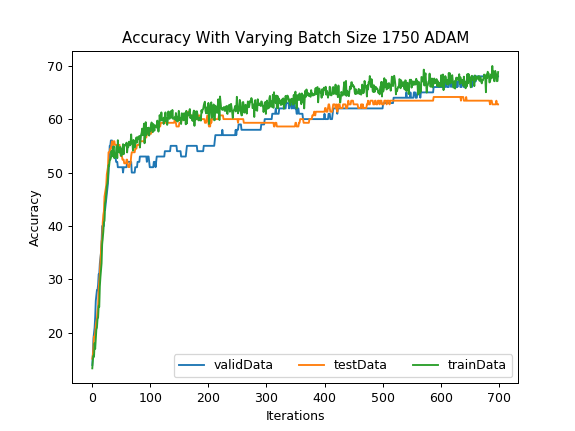
Final Validation accuracy: 56.99999999999999

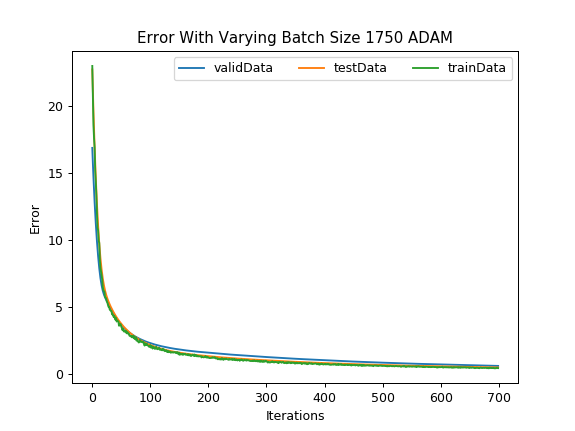
Final Validation error: 0.953214459308

Final Test accuracy: 61.37931034482759

Final Test error: 0.700888780314

3) Batch size = 1750





Final Training accuracy: 66.685715

Final Training error: 0.5300796

Final Validation accuracy: 67.0

Final Validation error: 0.7199412678757632

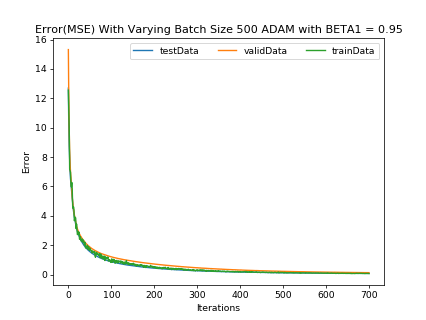
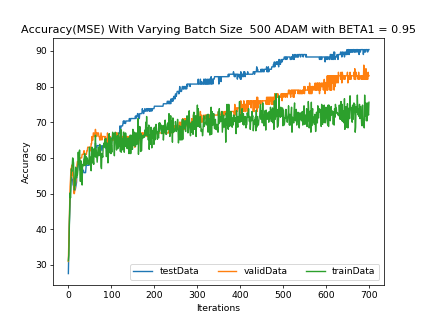
Final Test accuracy: 61.37931034482759

Final Test error: 0.5585539424018097

3.

Final training, validation and test accuracies with following Adam hyper-Parameters. Used parameters: epochs =700, Lambda = 0, Learning rate = 0.001, batch size = 500 :

1. Beta1 = 0.95 (Other parameters: Adam’s default values)



Final Training accuracy: 75.6

Final Training error: 0.0621561

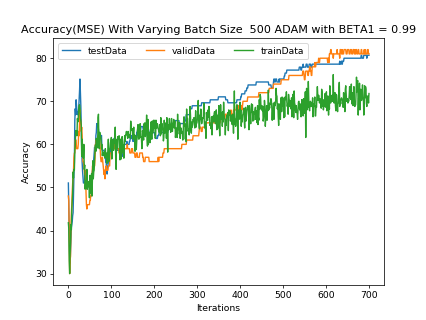
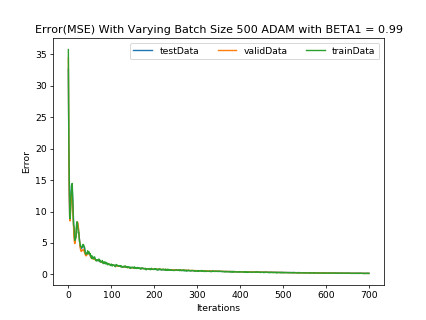
Final Validation accuracy: 83.0

Final Validation error: 0.126025

Final Test accuracy: 90.3448275862069

Final Test error: 0.0766271

1. Beta1 = 0.99 (Other parameters: Adam’s default values)



Final Training accuracy: 71.8

Final Training error: 0.157338

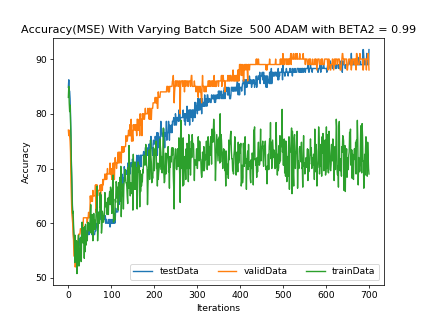
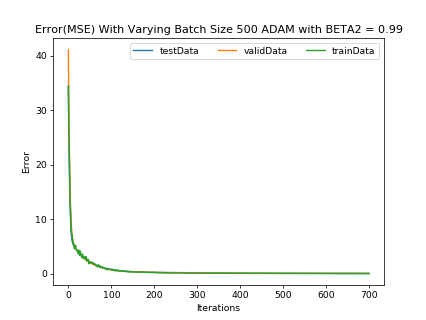
Final Validation accuracy: 81.0

Final Validation error: 0.166965

Final Test accuracy: 80.6896551724138

Final Test error: 0.176577

1. Beta2 = 0.99 (Other parameters: Adam’s default values)



Final Training accuracy: 69.0

Final Training error: 0.0314697

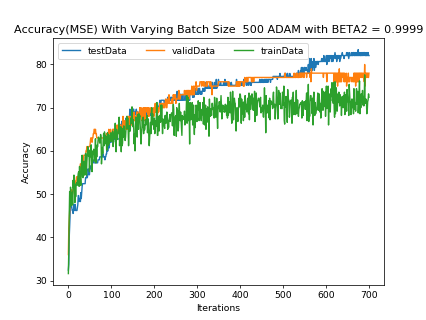
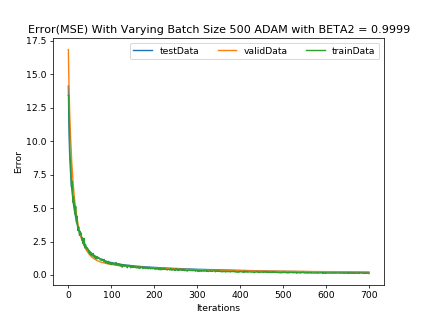
Final Validation accuracy: 88.0

Final Validation error: 0.0677183

Final Test accuracy: 91.72413793103448

Final Test error: 0.0577196

1. Beta2 = 0.9999 (Other parameters: Adam’s default values)



Final Training accuracy: 72.4

Final Training error: 0.101777

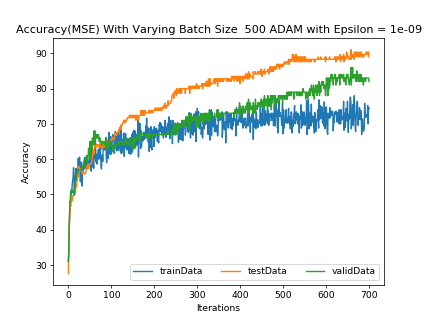
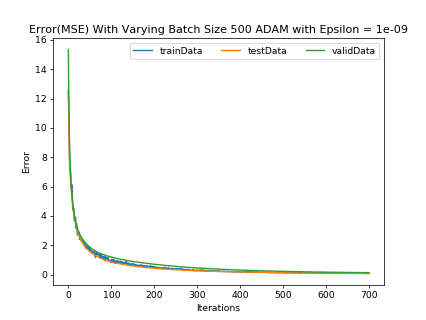
Final Validation accuracy: 78.0

Final Validation error: 0.227036

Final Test accuracy: 82.06896551724138

Final Test error: 0.182887

1. Epsilon = 1e – 09 (Other parameters: Adam’s default values)



Final Training accuracy: 74.4

Final Training error: 0.0624906

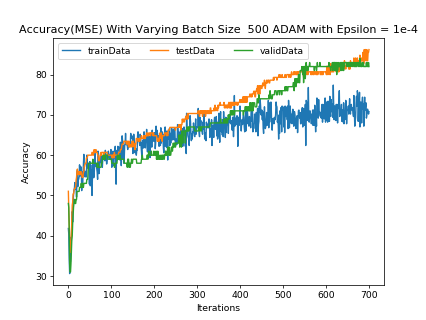
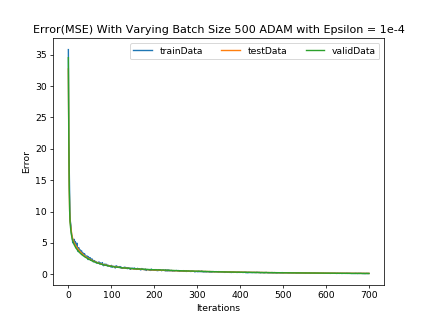
Final Validation accuracy: 82.0

Final Validation error: 0.126558

Final Test accuracy: 88.96551724137932

Final Test error: 0.0772777

1. Epsilon = 1e – 4 (Other parameters: Adam’s default values)



Final Training accuracy: 70.8

Final Training error: 0.124076

Final Validation accuracy: 82.0

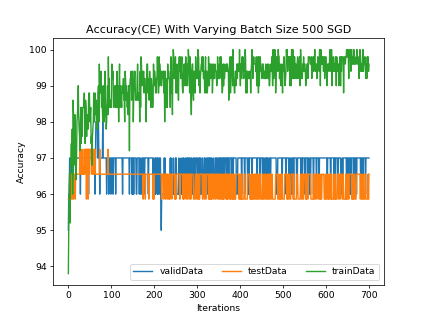
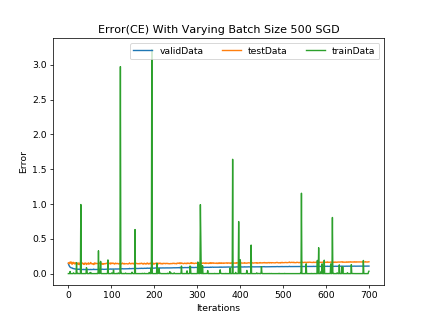
Final Validation error: 0.12162

Final Test accuracy: 86.20689655172413

Final Test error: 0.139391

**2.CE**

1. SGD(Gradient Descent optimizer) algorithm for a minibatch size of 500 optimizing over 700 epochs . Other parameters: Lambda = 0, Learning rate = 0.001



Final training accuracy: 99.6

Final Training error: 0.034631

Final Validation accuracy: 97.0

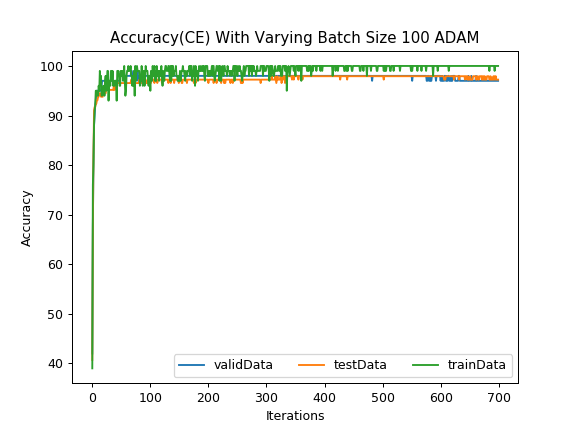
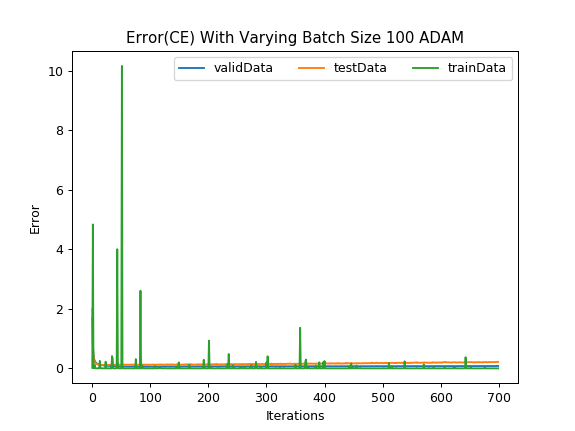
Final Validation error: 0.108054090004

Final Test accuracy: 96.55172413793103

Final Test error: 0.169762087977

1. SGD algorithm optimized using Adam by optimizing the model for minimizing the CE. Used parameters: epochs =700, Lambda = 0, Learning rate = 0.001 with variable batch sizes:

1) Batch size = 100



Final training accuracy: 100.0

Final Training error: 2.02908e-10

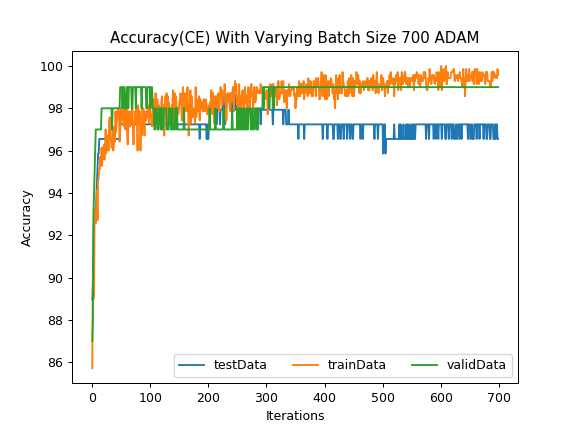
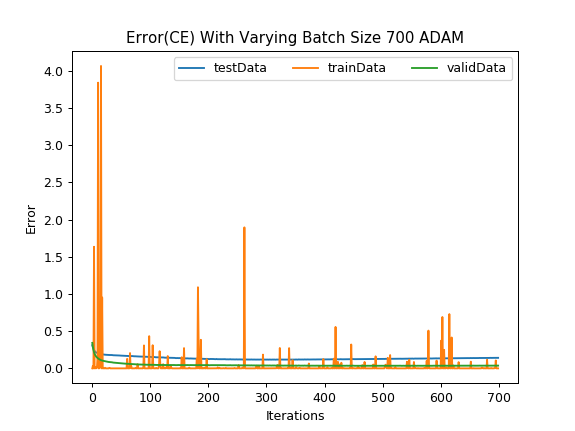
Final Validation accuracy: 97.0

Final Validation error: 0.0790904108438

Final Test accuracy: 97.24137931034483

Final Test error: 0.217853601998

2) Batch size = 700



Final training accuracy: 99.5714

Final Training error: 4.07625e-07

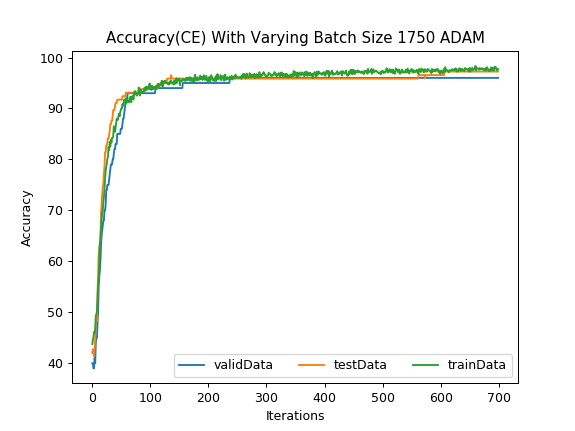
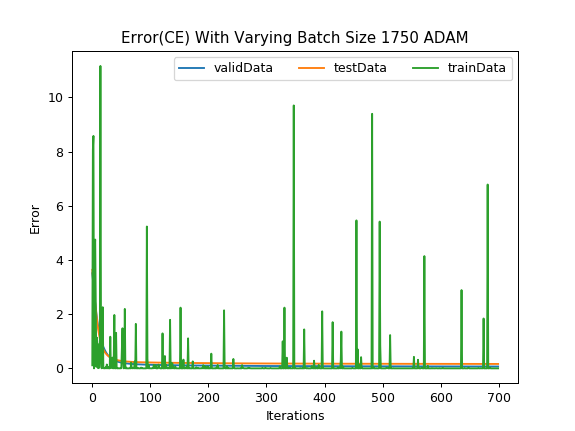
Final Validation accuracy: 99.0

Final Validation error: 0.0366850147505

Final Test accuracy: 96.55172413793103

Final Test error: 0.141456545833

3) Batch size = 1750



Final training accuracy: 97.7143

Final Training error: 1.36819e-05

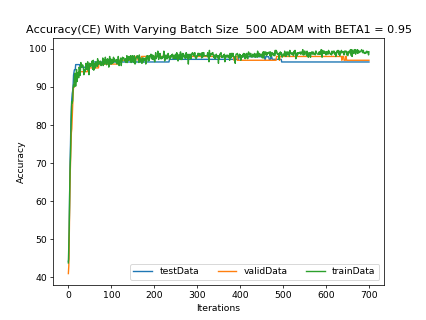
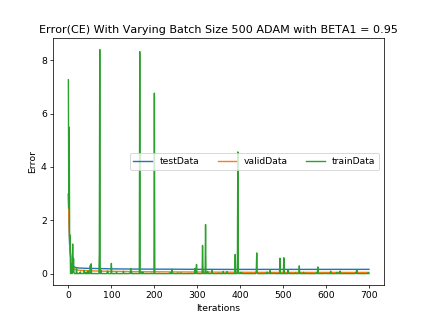
Final Validation accuracy: 96.0

Final Validation error: 0.0681662998078

Final Test accuracy: 97.24137931034483

Final Test error: 0.162773706021

1. Final training, validation and test accuracies with following Adam hyper-Parameters. Used parameters: epochs =700, Lambda = 0, Learning rate = 0.001, batch size = 500:
2. Beta1 = 0.95 (Other parameters: Adam’s default values)



Final Training accuracy: 99.2

Final Training error: 2.29499e-05

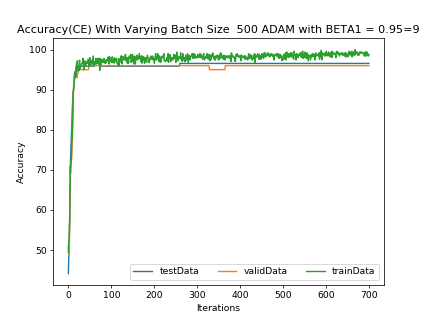
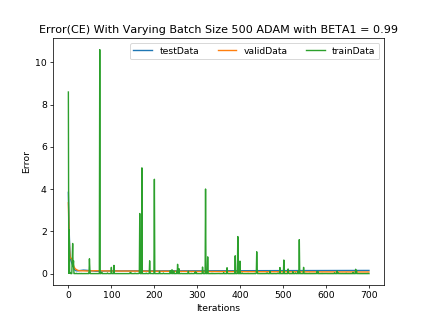
Final Validation accuracy: 97.0

Final Validation error: 0.0441717

Final Test accuracy: 96.55172413793103

Final Test error: 0.159736

1. Beta1 = 0.99 (Other parameters: Adam’s default values)



Final Training accuracy: 98.8

Final Training error: 5.64003e-05

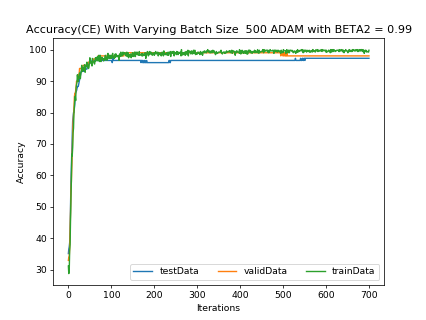
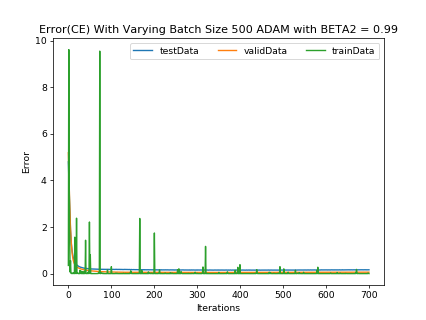
Final Validation accuracy: 96.0

Final Validation error: 0.0767184

Final Test accuracy: 96.55172413793103

Final Test error: 0.149259

1. Beta2 = 0.99 (Other parameters: Adam’s default values)



Final Training accuracy: 99.8

Final Training error: 1.87791e-05

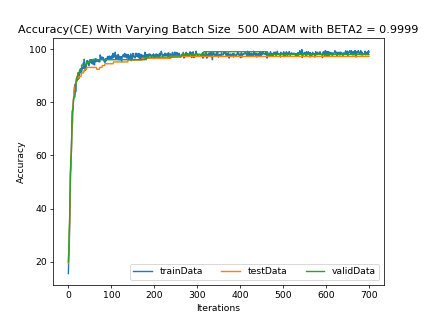
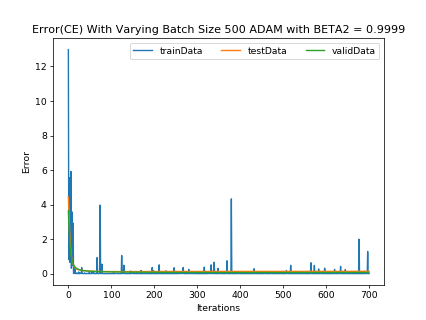
Final Validation accuracy: 98.0

Final Validation error: 0.0613664

Final Test accuracy: 97.24137931034483

Final Test error: 0.163428

1. Beta2 = 0.9999 (Other parameters: Adam’s default values)



Final Training accuracy: 98.8

Final Training error: 5.39271e-05

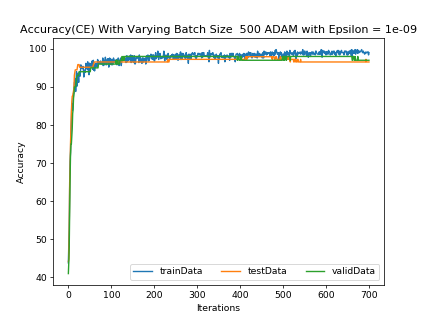
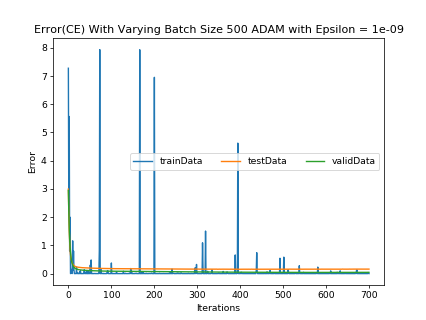
Final Validation accuracy: 98.0

Final Validation error: 0.0677422

Final Test accuracy: 97.24137931034483

Final Test error: 0.132327

5) Epsilon = 1e – 09 (Other parameters: Adam’s default values)



Final Training accuracy: 99.0

Final Training error: 2.2373e-05

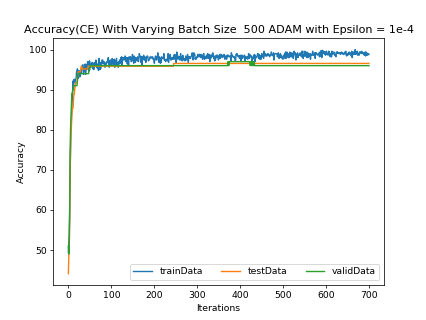
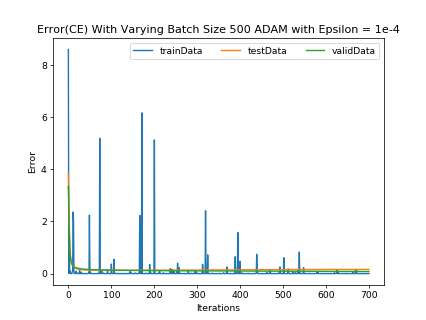
Final Validation accuracy: 97.0

Final Validation error: 0.0441423

Final Test accuracy: 96.55172413793103

Final Test error: 0.158898

6) Epsilon = 1e – 4 (Other parameters: Adam’s default values)



Final Training accuracy: 98.8

Final Training error: 5.47557e-05

Final Validation accuracy: 96.0

Final Validation error: 0.079528

Final Test accuracy: 96.55172413793103

Final Test error: 0.159296

**Hyperparameter Impact Analysis**

For the value of beta1 and beta2, they control the decay rates of moving average. From the graph, higher beta1 value and lower beta2 value will increase the error and decrease the accuracy slightly. Epsilon specifies the second of two hyperparameters for the adaptive learning rate, it is a small number to prevent any division by zero in the implementation. It does not significantly change the graphs as long as it close to zero.

**Comparison Between MSE and CE:**

There are more fluctuations in error and accuracy graphs for MSE than CE. However, the error curves with CE sometimes increases to high and decrease immediately after, this may caused by the data chosen in the batches cause the value of log function becomes really high.

For the final classification accuracy, MSE always gets smaller final accuracy than CE. Besides, MSE gets higher final error than CE. This is because MSE is doing regression estimation and CE is preferred in classification. Also, trainingData always has higher error and lower accuracy value than validData and testData.

**Analysis**

For the overall performance, SGD algorithm converges faster with less iteration numbers than batch gradient descent algorithm. But the final error loss and accuracy are similar. The SGD algorithm has fluctuations on error and accuracy since the data in every batch are randomly selected from training data. Also, Smaller batch size provides smaller error and higher accuracy. At the same time, smaller batch size increases the magnitude of fluctuation.