We implemented the network for three different learning rates and following are the results for each of them for two architectures with parameters as below along with different weight initializations heuristics.

**Architecture 1**

Nodes in first hidden layer =500

Nodes in second hidden layer =100

Iterations = 2000

Training Samples 10000

Validation Samples 1000

Test Samples 4000

**Architecture 2**

Nodes in first hidden layer =100

Nodes in second hidden layer =10

Iterations = 2000

Training Samples 20000

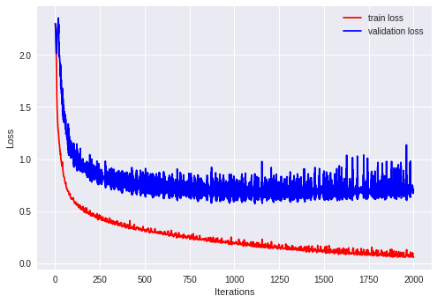
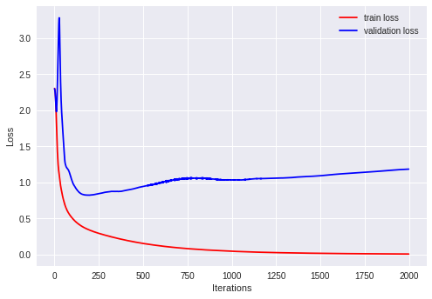
Validation Samples 1000

Test Samples 5000

At first we began with small dataset namely 10000 samples and after observing the trends in various momentum techniques, we increased the dataset size and trained the network on parameters that showed relatively good performance.

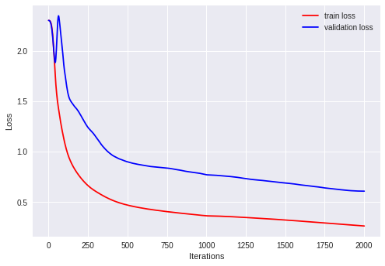
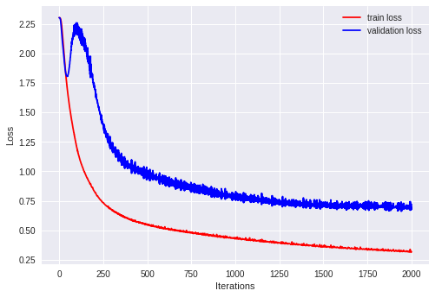
Learning Rate = 0.001:

For lower learning rates, RMS and ADAM perform better as portrayed by our results too. Other techniques took larger time to converge. Although RMS shows good accuracy, it is not as stable as ADAM. Same is evident in the plots below



Architecture 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Test |
| NoMomentum | 0.453 | 0.23 | 0.212 |
| Poly | 80 | 57 | 56 |
| RMS | 0.989 | 0.773 | 0.7335 |
| ADAM | 1 | 0.664 | 0.6575 |
| NAG | 0.7249 | 0.389 | 0.4245 |



Architecture 2

For this architecture, no momentum technique fails to converge because of low learning rate.

|  |  |  |  |
| --- | --- | --- | --- |
| NoMomentum | 0.1 | 0.1 | 0.1 |
| Poly | 0.55 | 0.155 | 0.165 |
| RMS | 0.883 | 0.766 | 0.73025 |
| ADAM | 90 | 78 | 76 |
| NAG | 0.4816 | 0.1099 | 0.1086 |

Learning 0.05

As compared to learning rate 0.001, this learning rate is favorable to the other three techniques i.e. NoMomentum, Momentum(Ploy) and NAG.

Architecture 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Test |
| NoMomentum | 0.69 | 0.341 | 0.309 |
| Poly | 0.8729 | 0.733 | 0.71525 |
| RMS | 0.7051 | 0.442 | 0.393 |
| ADAM | 93 | 66 | 66 |
| NAG | 0.8635 | 0.712 | 0.686 |

For architecture 2, the performance of RMS and ADAM deteriorates because the nodes in the hidden layer are significantly less than architecture 1 and so they get stuck and fail to converge partly because of high learning rate too. And when we say high learning rate, we are talking in context for RMS and ADAM.

Architecture 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Test |
| NoMomentum | 0.842 | 0.604 | 0.61 |
| Poly | 0.984 | 0.7 | 0.69 |
| RMS | 11 | 10 | 10 |
| ADAM | 0.374 | 0.143 | 0.218 |
| NAG | 84 | 64 | 63 |

Learning 0.01

Architecture 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Test |
| NoMomentum | 0.6516 | 0.221 | 0.2255 |
| Poly | 0.8586 | 0.633 | 0.6355 |
| RMS | 0.9845 | 0.711 | 0.6965 |
| ADAM | 1 | 0.639 | 0.63075 |
| NAG | 88 | 67 | 65 |

Architecture 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Training | Validation | Test |
| NoMomentum | 0.5529 | 0.157 | 0.169 |
| Poly | 0.81295 | 0.665 | 0.6396 |
| RMS | 0.9913 | 0.761 | 0.7682 |
| ADAM | 0.99985 | 0.651 | 0.6264 |
| NAG | 0.885 | 0.765 | 0.75 |