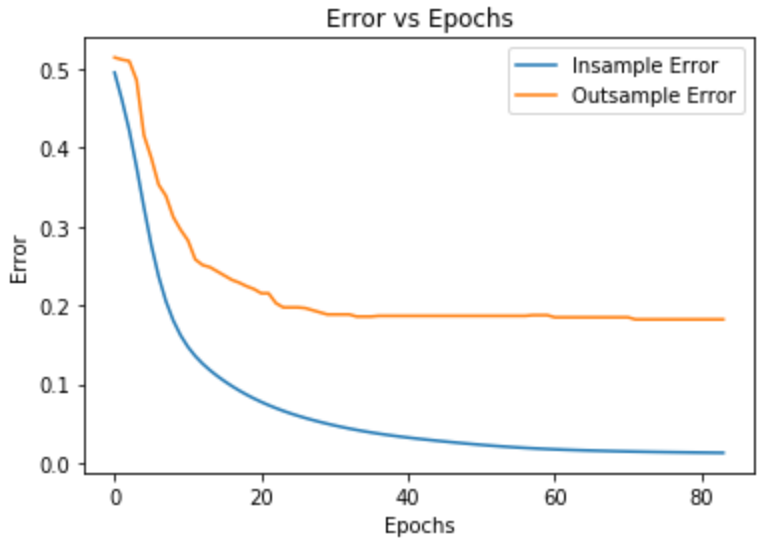
For faster training the batch size is set to 1000. You could set it to 1 for single training example for each gradient update as in Stochastic Gradient Descent.

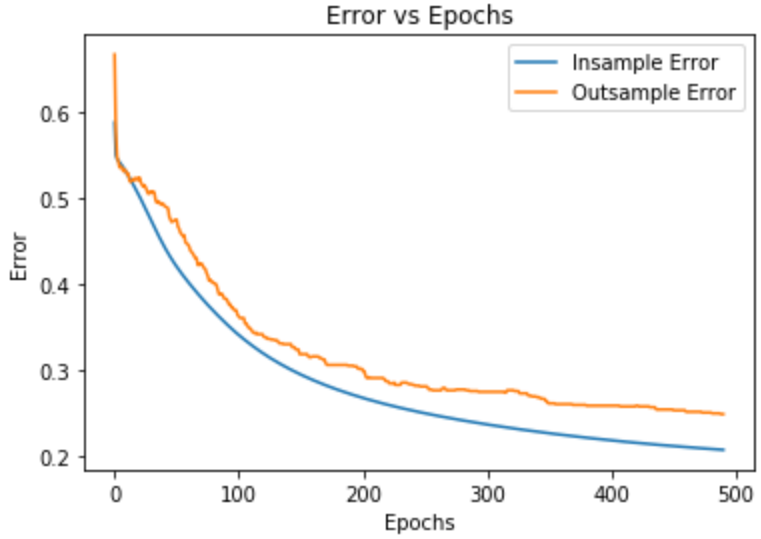
Part A

(A1) Tanh
Plot:



Optimal Number Of Iterations: 84

(A2) Sigmoid
Plot:

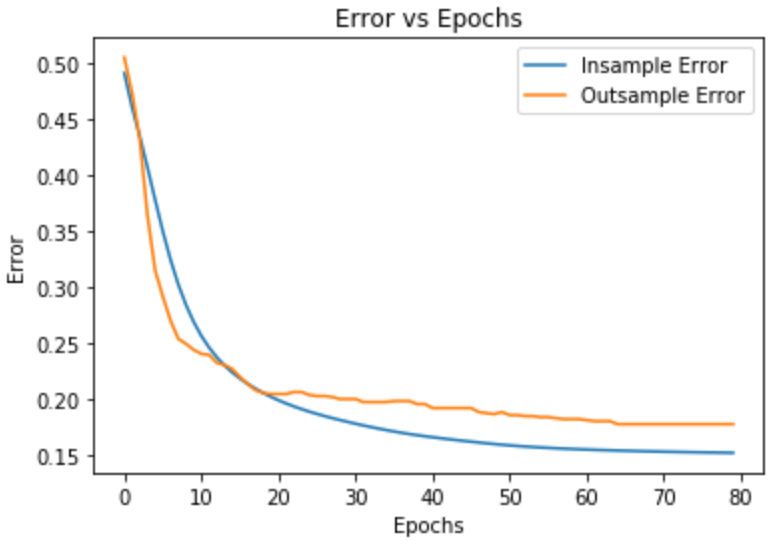


Optimal Number Of Iterations: 491

Result: Activation function “tanh” performs the best

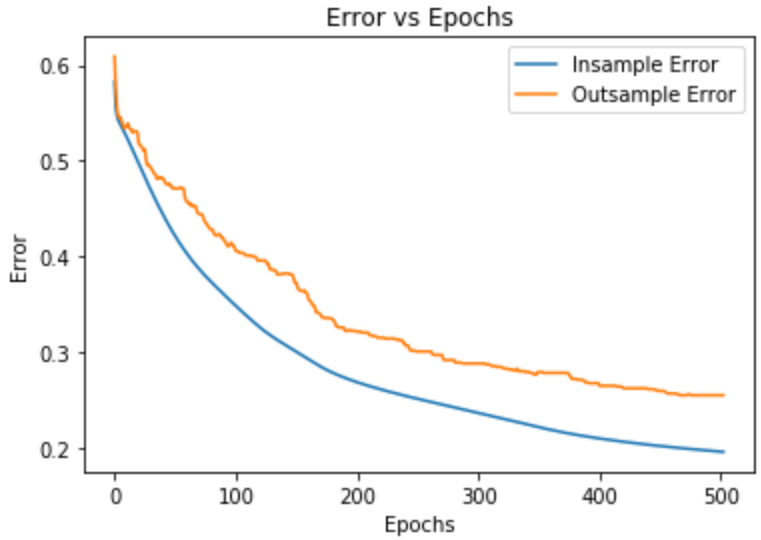
Part B

(B1) Tanh



Plot:
Optimal Number Of Iterations: 80

(B2) Sigmoid
Plot:



Optimal Number Of Iterations: 921

Result: Activation function “tanh” performs the best

Report :

As it is evident from the graphs that tanh activation performs better than the sigmoid activation and the reason behind this is the zeros centered nature of tanh compared to that of the sigmoid which is always positive. This always positive nature of sigmoid puts a restriction on the gradient direction of the weights that follow the activation function and causes a jitter in the training which is evident from the graphs.