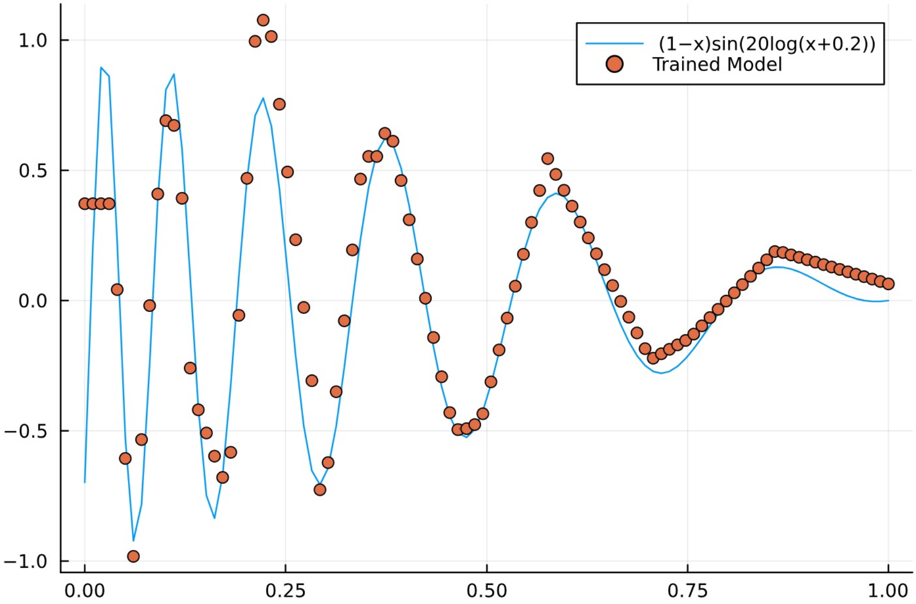
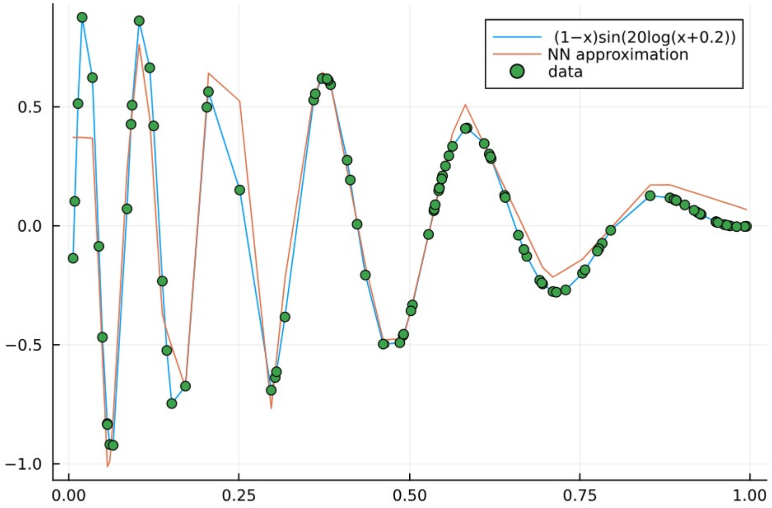
Amanda Marlow HW5 DMU

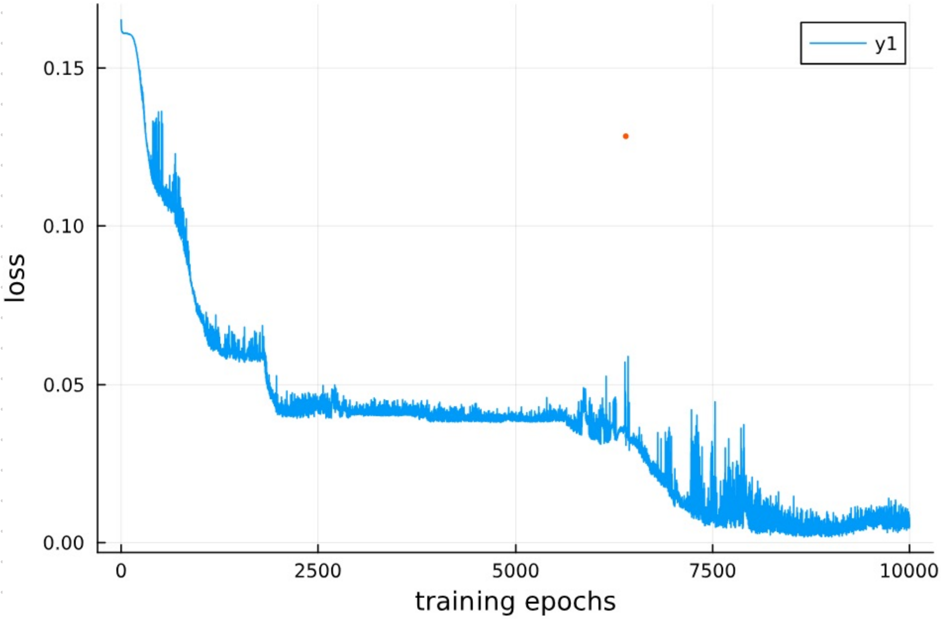
1] mean reward: 38.329

2) 100 points in trained model:



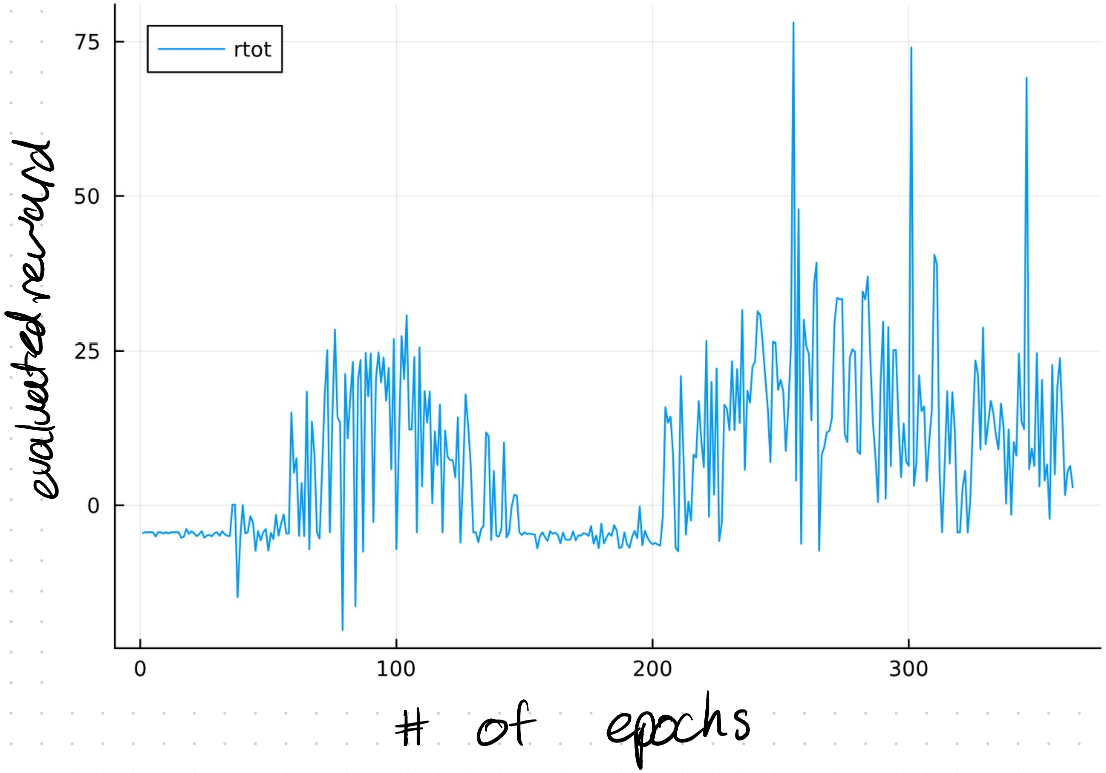
Plot from training:




loss
0.00
0.05
0.10
0.15
0
2500
5000
7500
10000
training epochs
y1

3)

a) I used DQN learning. I use an epsilon greedy policy initialized with $\epsilon = 0.5$ and then decays. I use the policy to take 100 steps in the environment that add data to the buffer. Then, I train with 1000 random samples from the buffer, I continue to interact with the environment in this way, and every 1000 steps I update Q_θ' to be Q_θ. Once, it reaches a terminal state, or $\gamma < 0.01$, the epoch ends and I evaluate the current Q to see if it should be saved as the best Q. I continue looping through epochs until I reach 100,000 steps in the environment. Then I return the max Q.


rtot
75
50
25
0
0
100
200
300
evaluated reward
# of epochs