Q 8 Deep Reinforcement Learning

5.1

I have implemented the neural network which has three layers in build\_model() function.

"select\_action" function returns argmax(qvals) of a particular observation which is used in train() function to get the next observation.

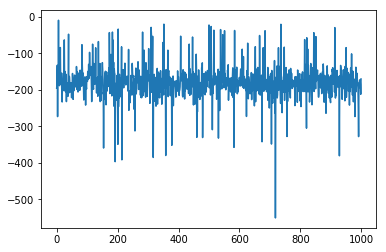
In the train() function, for each episode I am storing the values in the replay memory. Appending the qvalues and target values.

I have used q-learning algorithm in this function to find the best q value. For every batch value, update function is called which takes the replay memory of this batch size and find the q value. I am using the Adam Optimizer for optimizing the gradient.

5.2

Visualization :

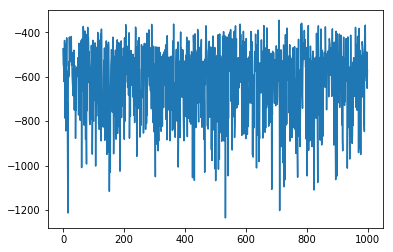
Below graph is for 0.05



Episodes (X-axis)

Rewards (Y-axis)

Below graph is for 0.95



Episodes (X-axis)

Rewards (Y-axis)

5.3 Q-Values

For State 1 [[-0.12048396 -0.03056725 -0.23096715 -0.07373706]]

State 2 [[-0.11854298 -0.03516214 -0.21915746 -0.06428962]]

State 3 [[-0.11552157 -0.04469136 -0.19858706 -0.05804461]]

These values are not accurate because they are exploring more in the environment and not exploiting based on the the alpha value. Below is the way to calculate the Q values which makes the value inaccurate.

Q(s,a) = (1-alpha) \* Q(s,a) + alpha\*(r + gamma \* maxQ(s',a'))

MaxQ(s',a') is only the approximation which we used in the training which turns out to be wrong initially.