Reinforcement Learning - Assignment 1

Deepak Srivatsav

2016030

1 Question 1

1.0.1 Graphs

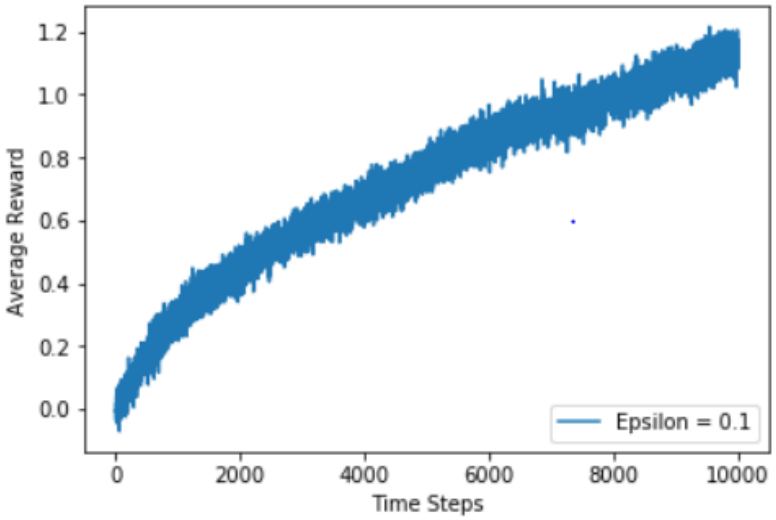

Figure 1: Sample Mean - Average Reward vs Time steps

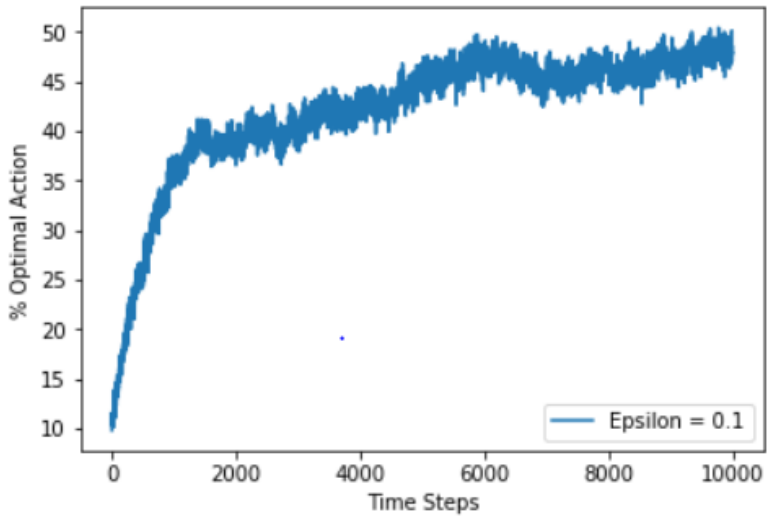

Figure 2: Sample Mean - % Optimal Solution vs Time steps

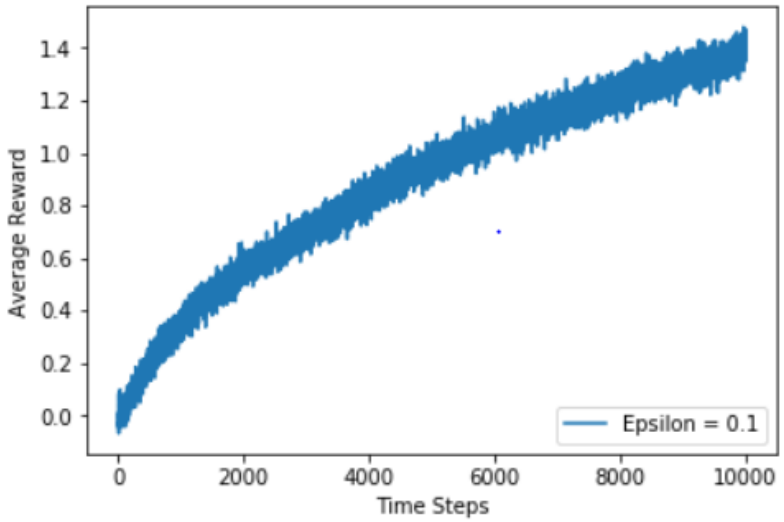

Figure 3: Sample Mean - Average Reward vs Time steps

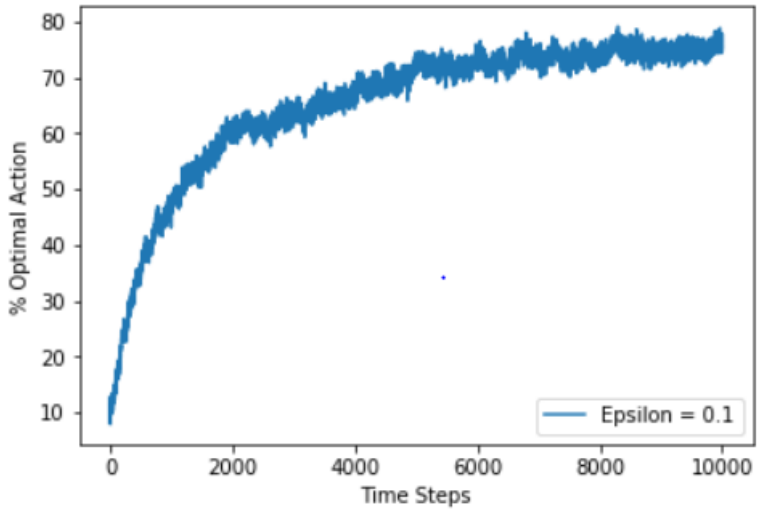

Figure 4: Constant Alpha - % Optimal Solution vs Time steps

2 Question 2

2.1 Graphs

:

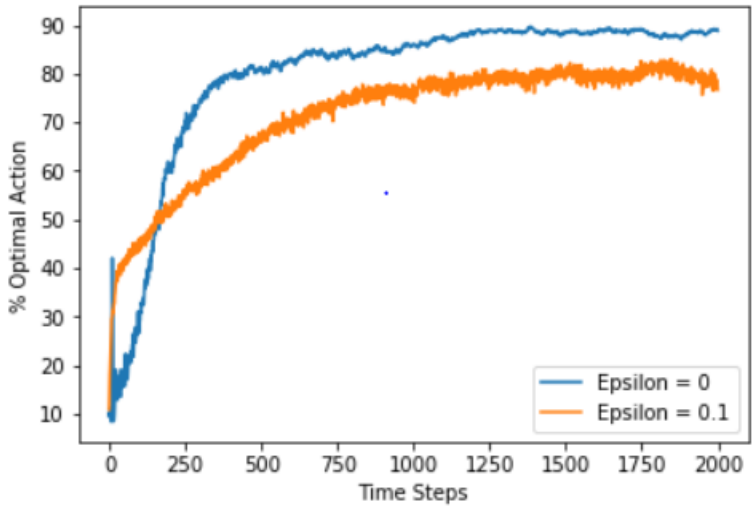

Figure 5: Stationary environment - % Optimal Action vs Time steps

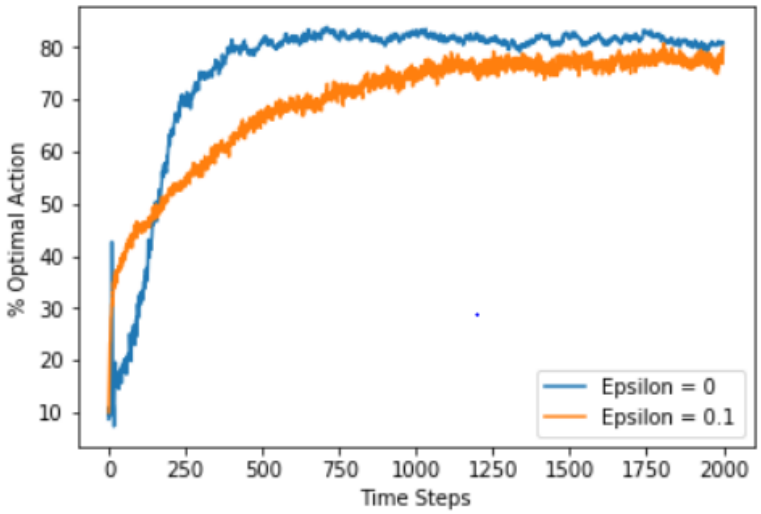

Figure 6: Non-Stationary environment - % Optimal Action vs Time steps

2.2 Analysis and Reasoning

We would see spikes during the initial steps when we use optimistic initialization since this would encourage exploration during the early steps. Exploring may or may not result in an optimal value. When we play an optimal value initially, we would continue playing the same value resulting in an increase in the number of times we play an optimal action. On the other hand, if a non-optimal action is played multiple times at early steps, this would result in a lesser percentage of actions being optimal actions at a time step. This is why we observe spikes in initial stages - and it settles down once each action value estimate has been updated to a value closer to the true value.

3 Question 3

3.1 Analysis and Reasoning

We have -

$$o_n = o_{n-1} + \alpha(1 - o_{n-1})$$

$$o_0 = 0$$

Let $\beta_n = \frac{\alpha}{o_n}$

$$o_1 = o_0 + \alpha(1 - o_0) = 0 + \alpha(1 - 0) = \alpha \tag{1}$$

Using (1), we get

$$\beta_1 = \frac{\alpha}{o_1} = \frac{\alpha}{\alpha} = 1 \tag{2}$$

We have $Q_{n+1} = Q_n + \beta[R_n - Q_n]$, from which we get, $Q_{n+1} = \beta_n R_n + (1-\beta_n)Q_n$
We can expand Q_n and rewrite the above equation as

$$Q_{n+1} = \beta_n R_n + (1-\beta_n)[\beta_{n-1}R_{n-1} + (1-\beta_{n-1}Q_{n-1}]$$

On repeatedly expanding Q_i, we will end up with this equation -

$$Q_{n+1} = \beta_n R_n + \sum_{i=1}^{n-1}(\prod_{j=n-i+1}^{n}(1-\beta_j))\beta_{n-i}R_{n-i} + Q_1\prod_{i=1}^{n}(1-\beta_i)$$

From (2), we can replace the value of $beta_1$ in the above equation, so that $Q_1\prod_{i=1}^{n}(1-\beta_i)$ reduces to 0 as $(1-\beta_1)$ would be 0. This removes the bias on Q1.

4 Question 4

5 Graphs

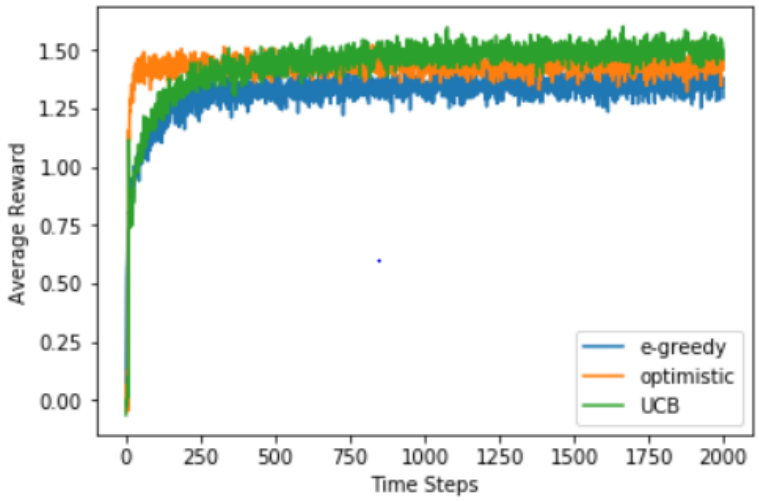

Figure 7: Stationary environment - Average Reward vs Time steps

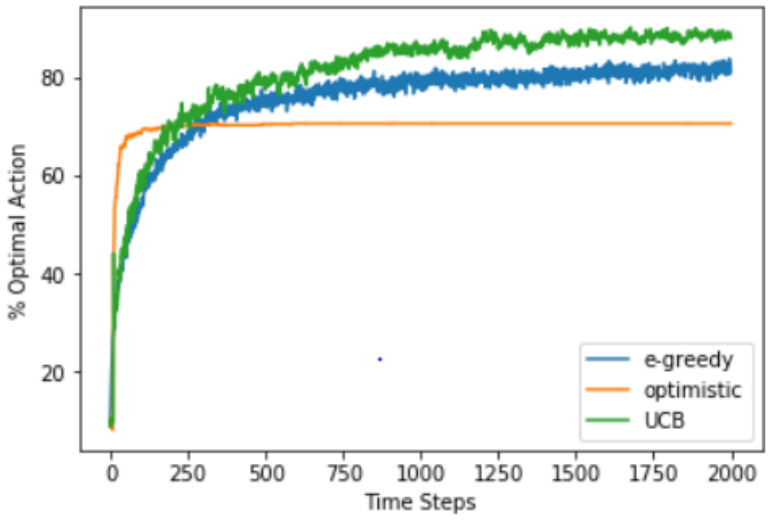

Figure 8: Stationary environment - % Optimal Action vs Time steps

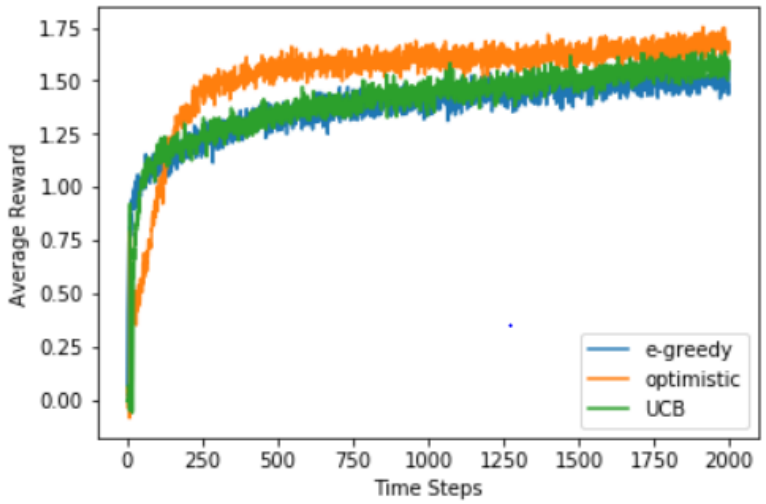

Figure 9: Non-Stationary environment - Average Reward vs Time steps

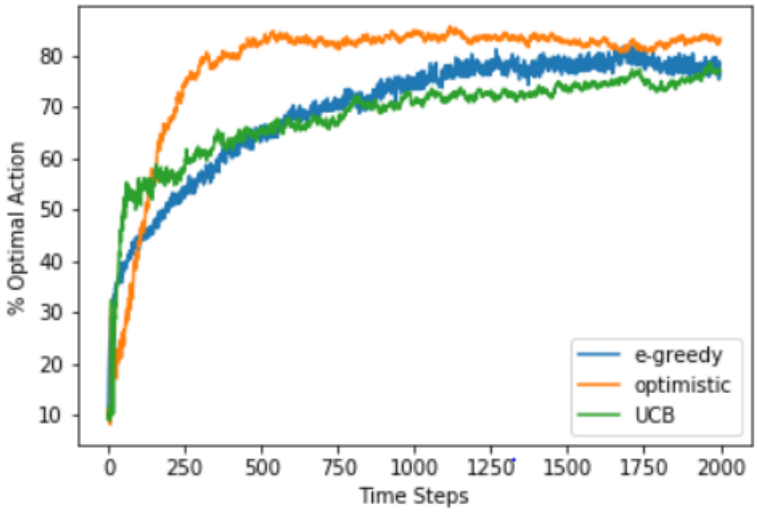

Figure 10: Non-Stationary environment - % Optimal Action vs Time steps