Reinforcement Learning
Assignment 1

Abhishek Agarwal
2016126

Ans 1)
As evident from the plots of Average reward and %Optimal Action in the non stationary case, we can say that constant step size parameter outperforms the sample average method. Although for stationary settings sample average method is better as we know that it will converge by the law of large numbers. In the case of non stationary, since q* is changing in every step, the sample average method does not perform as good as the fixed size parameter one.

Ans 2)
Reason of mysterious spikes:

When we use the optimistic approach to maximize the reward, we initialise the action values with a high number. For instance, we have initialised it with 5 for every action. Now, the reward that we get is a function of q* and is drawn from a normal distribution with mean q*(a). The values of q*(a) have been initialised from a 0 mean gaussian. Thus, the reward that the agent gets is very less as compared to the value of that action. This makes the agent disappointed with the action choice thus leading to exploration. Since there are a total of 10 possible actions at every step and it is always exploring during the initial steps, it is guaranteed that it will choose the optimal action atleast once in the first few steps. This is the reason why we observe a spike, because in every run the optimal action is definitely chosen in the initial few steps.

Ans 3)
*Attached below

Ans 4)
We compare the three approaches namely: Epsilon Greedy, UCB and Optimistic approach to choose an action at each time step. On comparing these three we see that Optimistic approach performs the best, followed by UCB and finally by epsilon greedy when we take an average of the results over 1000 runs with 2000 steps in each run. The reason for the same can be attributed to the vast exploration caused by unsatisfactory rewards in optimistic approach. The UCB approach explores more than the epsilon greedy(epsilon = 0.1) and hence it performs better. We also notice that the UCB approach is very close to the optimistic approach after 1000 steps and

hence it might overtake it and yield greater average reward if it is ran for more runs i.e. it might perform better in the longer run.

For the non stationary setting we see that UCB approach does not perform well. This is because the optimal action is decided on the basis of q*, but this q* keeps changing because of the non stationary setting and hence the reward that we get is not always the maximum. The other two techniques explore the other actions much more and hence perform better.

Q.3

To avoid initial bias, we use a step size of β_n.
where $\beta_n = \alpha/\bar{o}_n$

$$\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}) \quad \forall n \geq 0$$
$$\bar{o}_0 = 0.$$

Claim: On using a step size of β_n, the action values are unbiased by the initial estimate.

Proof:

We know that:

$$Q_n(a) = Q_{n-1}(a) + \alpha[R_{n-1}(a) - Q_{n-1}(a)]$$

with a step size of $\beta_n = \alpha/\bar{o}_n$

Update rule is:

$$Q_n(a) = Q_{n-1}(a) + \frac{\alpha}{\bar{o}_{n-1}}[R_{n-1}(a) - Q_{n-1}(a)]$$

$$\bar{o}_0 = 0 \quad ; \quad \bar{o}_1 = 0 + \alpha(1-0) = \alpha$$
$$\bar{o}_2 = \bar{o}_1 + \alpha(1 - \bar{o}_1) = \alpha + \alpha(1-\alpha)$$
$$\bar{o}_2 = 2\alpha - \alpha^2 = \alpha(2-\alpha)$$

$$\bar{o}_3 = \bar{o}_2 + \alpha(1-\bar{o}_2) = \alpha(2-\alpha) + \alpha(1 - 2\alpha + \alpha^2)$$
$$= 2\alpha - \alpha^2 + \alpha - 2\alpha^2 + \alpha^3$$
$$= \alpha^3 - 3\alpha^2 + 3\alpha$$
$$= \alpha(\alpha^2 - 3\alpha + 3)$$

$$\beta_1 = \alpha/\alpha = 1 \quad ; \quad \beta_2 = \alpha/\alpha(2-\alpha) = \frac{1}{2-\alpha}.$$

$$\beta_3 = \frac{1}{\alpha^2 - 3\alpha + 3} \quad \ldots$$

$Q_2 = Q_1 + 1(R_1 - Q_1)$

$\boxed{Q_2 = R_1}$ i.e. independent of Q_1

Similarly,

$$Q_3 = Q_2 + \frac{1}{2-\alpha}(R_2 - Q_2)$$

$$Q_3 = R_1\left(1 - \frac{1}{2-\alpha}\right) + \frac{1}{2-\alpha}R_2$$

$$= R_1\left(\frac{2-\alpha-1}{2-\alpha}\right) + \frac{R_2}{2-\alpha}$$

$$\boxed{Q_3 = \frac{R_1(1-\alpha)}{2-\alpha} + \frac{R_2}{2-\alpha}}$$

$$Q_n = Q_{n-1} + \beta_{n-1}[R_{n-1} - Q_{n-1}]$$

$$Q_n = Q_{n-1}[1-\beta] + \beta_{n-1}R_{n-1}$$

$$\downarrow$$

$$(Q_{n-2} + \beta_{n-2}[R_{n-2} - Q_{n-2}])(1-\beta)$$

We

⋮

We can keep reducing the subscript till we reach Q_2.

Now, since $Q_2 = R_1$ is independent of Q_1.

The expression for Q_n is not dependent on Q_1 it is free of any initial bias.