Vishaal Udandarao
20161119

CLASSTIME | Page No.
Date / /

THEORY ANSWERS

Q.1) We see from the plots that sample average action-value method is sub-optimal to the constant step size parameter action value method. Thus, for the non stationary setting, Constant step size parameter method outperforms sample average action-value method.

Q.2) The generated figure 2.3 is shown in the code.

Ex. 2.6

The oscillations and spikes in the early part of the curve for optimistic method are due to the high exploratory nature of the method. In the initial steps, the methods will choose all actions and since the reward received from all those actions is low compared to the estimates of those actions, the method will be "disappointed" in the rewards it receives.
Now, even though it is "disappointed" by all the actions, the method will choose the optimal action after trying out all the actions at least

once since the maximum reward (albeit "disappointing") is from the optimal action. Therefore, all the agents will almost simultaneously choose the optimal action and hence this causes the spike in the graph.
However, even after choosing the optimal actions, the reward it receives is again bound to be "disappointing" and hence the method does more exploration.
Thus, the oscillation in the plot can be explained by this.

Therefore, for the initial steps in the method, the optimistic greedy initialization method is not a good methodology since it encourages exploration only in the beginning and is useful only for stationary problems.

Q.3) Ex. 2.7

As suggested in the question, when we use sample averages to estimate action values, the action values are unbiased by the initial estimate i.e they are independent of Q_1.

But when we use constant step size parameter method to estimate

action values, these actions values are biased by the initial estimate Q_1.

Therefore, we need to propose a method which can be used to modify the constant step size parameter so that the action values estimated are independent of Q_1.

For this, we can use the following as our step size in the action value update step ⇒

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

Here, β_n is the step size we will use and α is the constant step size.

o_n is defined as ⇒

$$o_n = o_{n-1} + \alpha \cdot (1 - o_{n-1}) \quad \text{for } n \geq 0$$
$$o_0 = 0.$$

Now, by this formula, we see ⇒

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

$$\therefore \beta_1 = \frac{\alpha}{o_1} = \frac{\alpha}{\alpha} = 1 \Rightarrow \boxed{\beta_1 = 1}$$

Therefore, our action value estimate for step 2 is ⇒

$$Q_2 = Q_1 + \beta_1 (R_1 - Q_1)$$
$$\Rightarrow Q_2 = Q_1 + (R_1 - Q_1)$$
$$\Rightarrow Q_2 = Q_1 + R_1 - Q_1 = R_1$$
$$\Rightarrow \boxed{Q_2 = R_1}$$

Therefore Q_2 is independent of the initial estimate Q_1.

Similarly, we can extend this ⇒
For step 3 ⇒
$$Q_3 = Q_2 + \beta_2 (R_2 - Q_2)$$

We see ⇒
$$Q_2 = Q_1 + \alpha (1 - Q_1)$$
$$= \alpha + \alpha (1 - \alpha)$$
$$= \alpha (2 - \alpha) \Rightarrow \boxed{Q_2 = \alpha (2 - \alpha)}$$

$$\therefore \beta_2 = \frac{\alpha}{Q_2} = \frac{\alpha}{\alpha (2-\alpha)} = \frac{1}{2-\alpha} \Rightarrow \boxed{\beta_2 = \frac{1}{2-\alpha}}$$

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

Also, $Q_2 = R_1$ from above ⇒
$$Q_3 = R_1 + \frac{1}{2-\alpha} (R_2 - R_1)$$

$$\Rightarrow \boxed{Q_3 = \frac{R_1(1-\alpha) + R_3}{2-\alpha}}$$

This is also independent of Q_1.

∴ In a similar fashion, we can show for all action values $Q_4, Q_5, \ldots, Q_n$ that they are independent of Q_1.

∴ The step size that we should use for removing the dependence on initial estimate Q_1 is β.

Q.2 contd)

Analysis for stationary

The stationary plot shows that optimistic estimate performs better than epsilon greedy over the long run since it chooses optimal value actions greedily over the long run.

Analysis for non stationary

The non stationary plot shows that epsilon greedy should perform better over the long run. This is because in ~~greedy~~ non stationary environment, the optimal value changes at every step. Hence optimistic

greedy performs poorer.

Q. 4) Analysis for stationary

In stationary setting, we see that the optimistic estimate method performs best followed by UCB and then epsilon greedy. This is because in stationary setting, the optimal action remains same and hence optimistic estimate method is the best since in the long run it exploits the most. UCB is a slightly more explor-atory method and hence performs worse than optimistic estimate. Epsilon-greedy is not an entirely exploitation method and hence it performs worst.

Analysis for non stationary

In non stationary setting, we see that UCB performs the worst among all three. This is because in non stationary setting, the optimal action keeps changing and hence the exploratory aspect of a method should be high. But we see that UCB is highly exploitation favouring as time passes by. Hence, UCB performs the worst in non stationary setting.