Reinforcement learning

Neethu.R.R

PhD19108

Q ①

$\boxed{\text{See figure 2.2.png}}$

The plot for average reward & % optimal action for steps from 1 to 1000 is plotted.

Three different ε values are taken.

(i) $\varepsilon = 0 \Rightarrow$ Greedy policy.

Here we can see that average reward and % optimal action is very less. Because here only exploitation is taking place. Exploration of non-optimal arms which can give higher rewards are not considered.

(ii) $\varepsilon = 0.1 \Rightarrow$ 90% exploit & 10% explore.

Here we can see higher average reward and % optimal action is obtained. This is because of the high rewards obtained from the exploitation.

(iii) $\varepsilon = 0.01 \Rightarrow$ 99% exploit & 1% explore.

The average reward & % optimal action is much more than greedy action.

ε changing with time $\Rightarrow$ $\boxed{\text{See figure 2.2\_1.png}}$

Here we can see that the average reward and % optimal action is more than the greedy action.

It is behaving like $\varepsilon = 0.01$ which will be the best in the long run.

Q(2) $\boxed{\text{See figure Q2.png}}$

Variance of the distribution = 4.

Here we can see a decrease in the % optimal action because of the uncertainity in picking the arm. And the average reward has also large variation during the steps.

For $\varepsilon = 0.1$, the average reward obtained during the steps are very less compared to variance of 1.

Q3 $\boxed{\text{See figure Q3.png}}$

From Fig 2.2, we can see that the graph of $\varepsilon = 0.01$ is increasing in an optimal way compared to other values.

So in the long run, $\varepsilon = 0.01$ will perform better than other values in terms of cumulative reward & probability of selecting optimal action.

Comparing with other values,

$\varepsilon = 0.01 \Rightarrow$ 99% exploit & 1% explore.

$\therefore$ The chance of finding the optimal action is more.

$\varepsilon = 0.1 \Rightarrow$ 90% exploit & 10% explore.

(greedy) $\varepsilon = 0 \Rightarrow$ 0% explore which is not the ideal scenario.

In the plot, $\varepsilon = 0.01$ curve, overtakes the $\varepsilon = 0.1$ curve for large value of time steps.

$\therefore$ If $\varepsilon = 0.1$,

Suppose the average estimate of non-optimal values is q_n & the optimal value is q_{opt}

$$\therefore E[R_t] = 0.90 \times q_{opt} + 0.1 \times q_n.$$

If $\varepsilon = 0.01$.

$$E[R_t] = 0.99\, q_{opt} + 0.01\, q_n.$$

∴ The chance of ~~op~~ exploiting is more & finding the optimal arm.

Q4)

Sample mean.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}}$$

It is the summation of rewards for selecting a particula arm a.

eg: $Q_1(1) = Q_1(2) = Q_1(3) = Q_1(4) = 1$

$t = 1$ $A_1 = 1$ $R_1 = 1$

$t = 2$ $Q_2(1) = R_1 = 1$

What ever be the initial value of the estimate, it will not effect the total estimate of an arm.

But, in exponentially weighted recency average, where the step size is constant,

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$$

where Q_1 is the initial estimate.

But as time increases ie 'n' increases, the dependence on the initial estimate decreases.

A method to have constant stepsize α but no dependence of $Q_t(a)$ & $Q_1(a)$ is using a new stepsize.

Let $\beta_n = \alpha / \bar{O}_n$ be the new stepsize parameter where α is constant.

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

(i) When $n = 1$

$$\bar{O}_1 = \bar{O}_0 + \alpha(1 - \bar{O}_0) = \underline{\underline{\alpha}}.$$

$$\therefore \beta_1 = 1$$

Exponential recency weighted average :-

$$Q_{n} = (1-\beta_1)^n Q_1 + \sum_{i=1}^{n} \beta_1 (1-\beta_1)^{n-i} R_i$$

$$= \underline{\underline{0}} + \ldots$$

Which have no. dependency with the initial estimate Q_1

(iii) When $n=2$

$$\bar{O}_2 = \bar{O}_1 + \alpha(1-\bar{O}_1)$$

$$= \bar{O}_0 + \alpha(1-\bar{O}_0) + \alpha(1-(\bar{O}_0+\alpha(1-\bar{O}_0)))$$

$$\bar{O}_2 = \frac{1}{\alpha+\alpha(1-\alpha)}$$

$$\beta_2 = \frac{1}{2+\alpha}$$

$$\therefore \bar{Q}_3 = (1-\beta_2)^2 Q_1 + \sum_{i=1}^{2} \beta_2 (1-\beta_2)^{2-i} R_i$$

$$= \left(\frac{1+\alpha}{2+\alpha}\right)^2 Q_1 + \sum_{i=1}^{2} \frac{1}{2+\alpha}\left(\frac{1+\alpha}{2+\alpha}\right)^{2-i} R_i$$

Here $\left(\frac{1+\alpha}{2+\alpha}\right)^2$ is a small amount which will decrease the effect of Q_1. The research will be weighted by $\frac{1}{2+\alpha}$ so that dependency on the research increases.

Eventually, as n increases, the effect of Q_1 or the dependency of Q_t over Q_1 will be negligible for a constant stepsize (α).

Q(6) Generated figure 2.4

In $\boxed{\text{figure 2.4.png}}$, we can see that there is a sudden increase in the value of average reward of UCB at the 11^{th} step and it is maintained. As there are 10 arms, in the first round ie after completing 10 steps, the arm with highest reward is chosen at the 11^{th} step from this 10 arms. So there will be increment in the average reward because of the addition of reward at the 11^{th} step.

Also in $\boxed{\text{figure 2.4\_1.png}}$, there is a decrease in the average reward of UCB for the initial steps as shown in figure 2.4 in the text. This is for $c = 2$. For $c > 2$, ~~UCB~~ ~~with~~ the average reward for UCB will be less for the initial steps for a long time. But in $\boxed{\text{figure 2.4\_2.png}}$, where $c = 1$, the average reward for UCB is less than ε-greedy for the initial steps for a short duration of time. And for $C=4$, the average reward for UCB is always greater than ε-greedy.

The average reward for UCB $C=4$ > UCB $C=2$ > UCB $C=1$.

At 11th step.
- UCB $C=1$ ⇒ Average reward peak = 0.9
- UCB $C=2$ ⇒ Average reward peak = 1
- UCB $C=4$ ⇒ Average reward peak = 1.2.

Q7 figure 2.5 is generated (without baseline)

$\boxed{\text{See figure 2.5.png.}}$

Here stepsize $\alpha = 0.1$ gives more optimality in picking the actions compared to $\alpha = 0.4$.

A gibbs distribution is generated & compared with the best arm distribution (Gaussian)

The average reward is taken as zero. ∴ the % of optimal action is less.