Exercise 1: Multi-armed Bandits

Prof: Robert Platt

Date: September 24th, 2021

Name: Guanang Su

Question 1.

A picture containing graphical user interface

Description automatically generated

A4 and A5 were definitely occur, others (A1, A2 and A3) could possibly have occurred.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

For step 4, , so epsilon would definitely occur.

And for step 5, , so epsilon would definitely occur.

Question 2.

Text

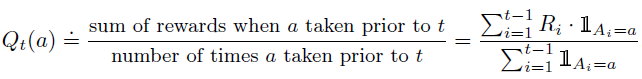
Description automatically generated with medium confidenceText, letter

Description automatically generated

Question 3.

Text

Description automatically generated



In sample-average estimate, sample mean is µ

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generatedText

Description automatically generated

It is unbiased.

Text

Description automatically generated

From the equation above, we can see that if the n is small enough like 1, there is no bias since only one sample. Also, if the data is infinite number, the biased would be small enough to regard as 0 since Q1=0, .

Text

Description automatically generated

When n is infinite, will decrease to 0,

and

So Qn is asymptotically unbiased.

Text

Description automatically generated

For a non-stationary problem, rewards close to the current state means a closer to the expected result, which help us to find better estimates for current distribution, rather than previous steps.

Question 4.

A picture containing text

Description automatically generated

Chart

Description automatically generated

Question 5.

Text

Description automatically generated with medium confidenceChart

Description automatically generated with low confidence

Text, letter

Description automatically generated

Question 6.

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

The average will not reach the asymptotic levels predicted, since it is random pick from steps, the upper bound, each greedy will explore in a range below or slightly higher than the bound value. Therefore, the average reward could not reach the asymptotic for the value we picked.

Question 7.

Chart

Description automatically generatedGraphical user interface, text

Description automatically generated

Chart, line chart

Description automatically generatedGraphical user interface, text

Description automatically generated

The sharp increase is happened to since the beginning of UCB algorithm, it tries with all arms, which is wildly optimistic to encourage the action-value to explore. After the exploration, the arm will pick the highest mean value and randomly run one of the index with highest value, the reward could have a drop.