The University of Vermont

CS253A QR: Reinforcement Learning: Assignment №2

Ayat Ospanov

September 9, 2018

Contents

1 2.1

If $\varepsilon = 0.5$, each second step is the greedy step. This means the probability of choosing the greedy action is at least 0.5. Further, as we do random step with the probability of 0.5 and select greedy action in this step with the probability of $\frac{1}{2}$ (because we have two actions and one of them is the greedy action), the probability of randomly selecting the greedy action is 0.25. Therefore, the overall probability of selecting the greedy action is 0.75.

We can expand this task to the case of n options/actions and any ε. The answer is $(1-\varepsilon)+\frac{\varepsilon}{n}$.

2 2.2

Table 1: Work of a bandit algorithm

t	1	2	3	4	5
A_i	1	2	2	2	3
R_i	-1	1	-2	2	0
$Q_i(1)$	0	-1	-1	-1	-1
$Q_i(2)$	0	0	1	-0.5	0.(3)
$Q_i(3)$	0	0	0	0	0
$Q_i(4)$	0	0	0	0	0
Choice	Random	Random	Greedy	ε case	Random

On the table 1 the work of a bandit algorithm is provided by time step. Each arrow shows the Q-value of a chosen action. As on greedy step an algorithm choose the $\arg\max_a Q_t(a)$, the only case of choosing the argmax is step 3. On the step 4 the algorithm chose the value of -0.5 which is not the argmax. It means that at this step the ε case has occured. On the other steps (1, 2, 5) as we have more that one maximum value of Q, the alogrithm chose random argmax. On these steps ε case could possibly have occurred.

3 2.3

In the long run, $\varepsilon = 0.01$ will act better as 99.1% of the time (see 1) it choose the correct actions, while in the case of $\varepsilon = 0.1$ the rate of correct actions is 0.91 or 91%.