Exercise 1 - Bandits

Tim Schäfer, Timo Mahringer

Exercise 1: Multi-armed bandits

a) The probability that the bandit choses a random action is

$$P(a_{\text{random}}) = 1 - \epsilon = 1 - \frac{1}{2} = \frac{1}{2}$$

As it is possible that the greedy action may be chosen randomly or chosen by maximising the reward the probability sums up to

$$P(a_{\text{greedy}}) = \frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{3}{4}$$

Where the first summand comes from the probability that the greedy action is chosen by the policy and the second one if the greedy action is chosen between the two possible actions when the policy decides to make a random choice.

b)

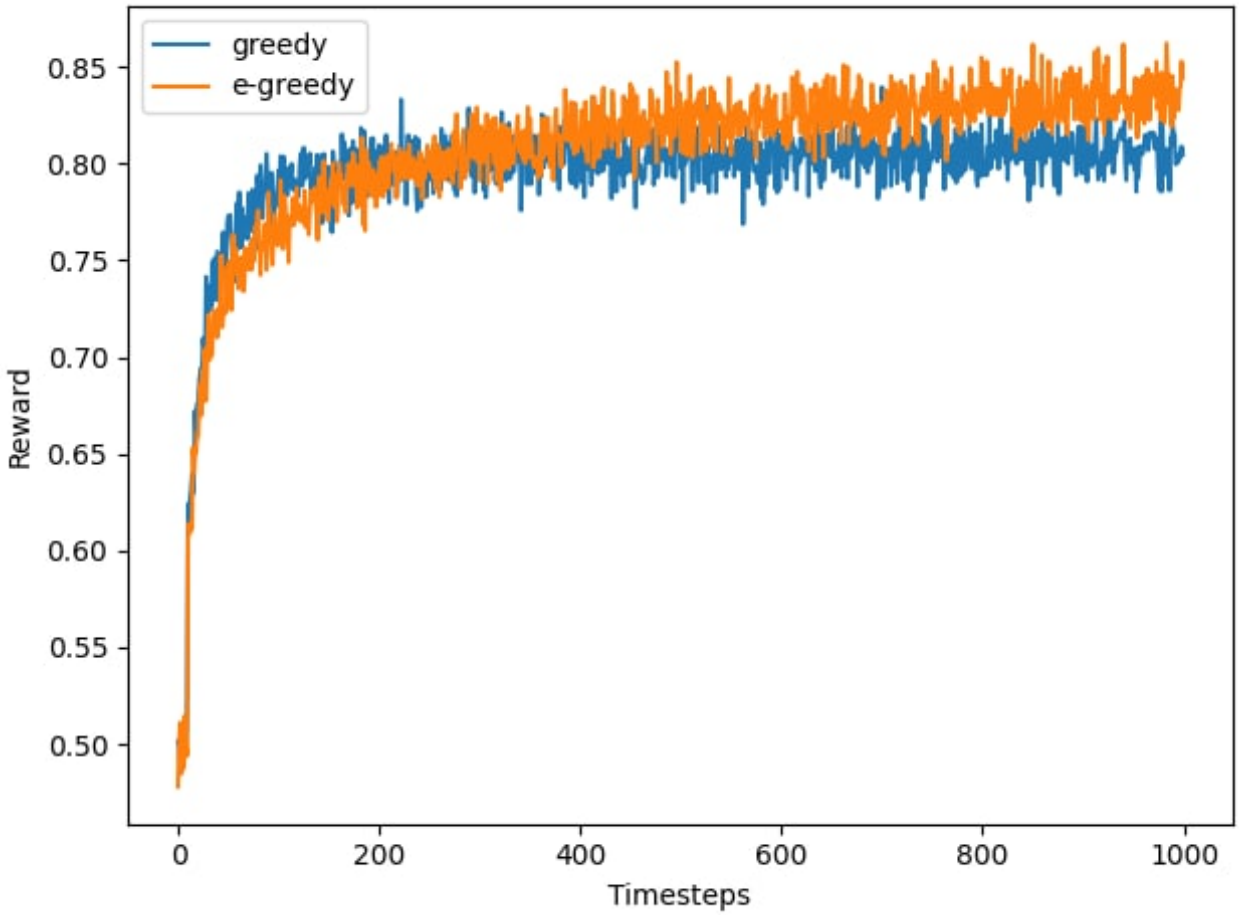
	Action 1	Action 2	Action 3	Action 4		Chosen next action	May be random choice	Must have been random choice
Q_1	0	0	0	0	As given	1	Yes	Maybe
Q_2	1	0	0	0	Calculated	2	Yes	Yes
Q3	1	1	0	0	Calculated	2	Yes	Maybe
Q_4	1	1.5	0	0	Calculated	2	Yes	
Q_5	1	5/3	0	0	Calculated	2	Yes	
Q_6	1	5/3	0	0	Calculated	3	Yes	Yes

Since the probability of a random choice is never zero, every choice could have been a random choice.
The choices 2 and 5 must have been random choices since they are not maximising the expected reward. As we don't no, how the algorithm decides if two actions have the same expected reward, the choice 1 could have been a random choice (all have zero as expected reward) and also choice 3 (action 1 and 2 have the same expected reward).

Exercise 2: Action selection strategies

a) See code
b) See code
c) The epsilon-greedy strategy performs better since only maximising could lead into a local maximum while the global maximum could be another action but which has not been played (enough) before the greedy strategy gets stuck in the local maximum.

The screenshot of our run:



d) There are too much ideas we have so we start with a simple one:

We would start with an very exploring strategy to get a good overview about maybe several maxima. Then starting to decrease epsilon so that the focus would more lie on the found maxima. One could improve this behaviour by having multiply different probabilities for choosing a maximum by decreasing with each lower found maximum. For example:

For the highest expected reward take p = 0.5
For the second one p = 0.2
For the third one p = 0.1
For the fourth one p = 0.05
Random choice p = 0.15

And same here: with better understanding of the different expected rewards the probability for a random choice or the lower maxima can be reduced to really take advantage of the highest reward per round (exploitation).