Reinforcement Learning Exercise 1

Alexander Jäggle, Johannes Haberstock, and Daniel Arnold

M.Sc. Autonomous Systems, University of Stuttgart

May 3, 2021

1 Multi-armed Bandits

a) The probability of the greedy action being selected is $p = .75$ since $p = (1 - \epsilon) + \frac{\epsilon}{k}$ where $k = 2$.

b)
1. random = [2,5]
2. greedy = [1,3,4]

Explanation to b) Before step three is executed, only the rewards for action 1 and action 2 are known, which both are 1. Every other action, which was not explored yet, is assumed with a reward of 0. Thus, at timestep $t = 3$, action 2 with a reward of 1 was selected at greedy.

A similar scenario happened at timestep $t = 4$, when action-value estimates Q are known. For timestep $t = 3$, the action-value estimate is 1.5, while the other two estimates are 1 respectively 0. Since the next action, which was selected was action 2, and thus it was greedy.

2 Action Selection Strategies

The solution of our group will be submitted as *ex01-bandits.py* and is also available on our GitHub.

a) Changes were made between line 24 - 54

b) Changes were made between line 58 - 99

c) E-Greedy performs better with a total amount of 805.76 compared to a score of 797.63 of the Greedy as can be seen in fig. 1. After roughly 300 executions, e-greedy surpasses the greedy in amount of rewards.

d) Possible ways to improve the methods is to change ϵ or increase the number of executions.

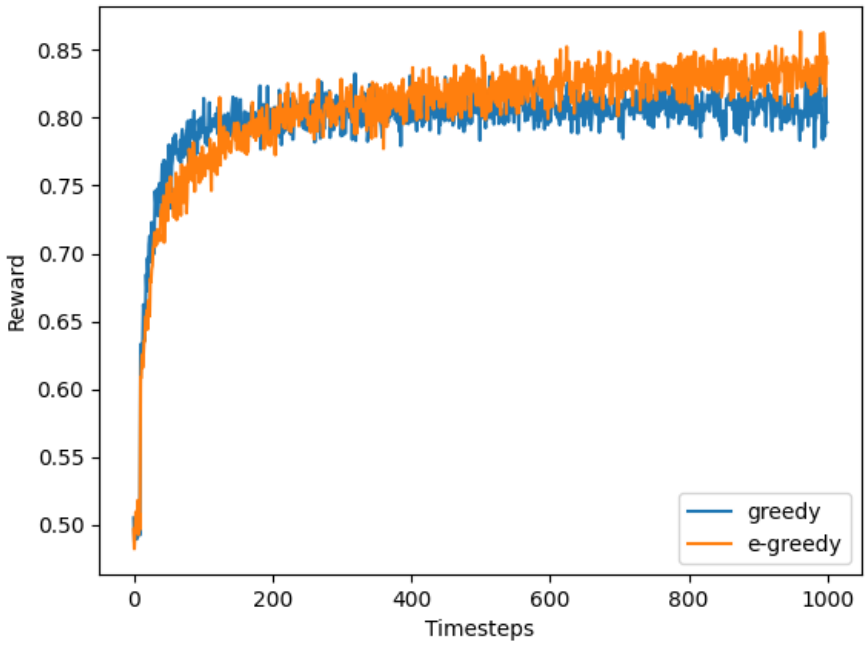

Figure 1: ϵ-Greedy vs. Greedy