Assignment 1 - Bandits

Jonathan Stumber ?? email@stud.uni-stuttgart.de
Christian Reiser 3131934 st141151@stud.uni-stuttgart.de

1. Multi-armed Bandits

1.1 Greedy action probability

With the given ϵ of 0.5, the greedy action is selected with a probability of

$$1 - \epsilon = 0.5. \tag{1}$$

1.2 On which steps did the bandit explore?

ϵ definitely occurred at actions:

1. $A_2 = 2$ - as it has an unknown reward, whereas greedy $A_2 = 1$ has an expected reward of 1.
2. $A_5 = 3$ - as it has an unknown reward, whereas $A_5 = 2$ would be the greedy action.

ϵ might have occurred on action $A_1 = 1$, because all action-values are unknown and it's maximum cannot be chosen.

2. Action Selection Strategies

c)
The greedy method always chooses the arm with the highest Q value. Consequently, the method does not explore if there are better actions.

The ϵ-greedy method explores other actions randomly with a chance of ϵ. This has the downside of taking non-promising actions and might explain inferior results for the first 200 timesteps, with $\epsilon = 0.1$. With the chance of 1-ϵ greedy action is selected, with the benefit of better Q values due to exploration.

With more realistic Q values, the greedy method works better, which might explain the superior e=performance of the ϵ-greedy method for timesteps > 200.

d)

1. It might be beneficial to explore more in the beginning and be greedy later on. This could be implemented such that ϵ is not constant but decays with increasing timesteps.
2. It might be beneficial not to explore randomly, but explore the arm which was selected fewest.

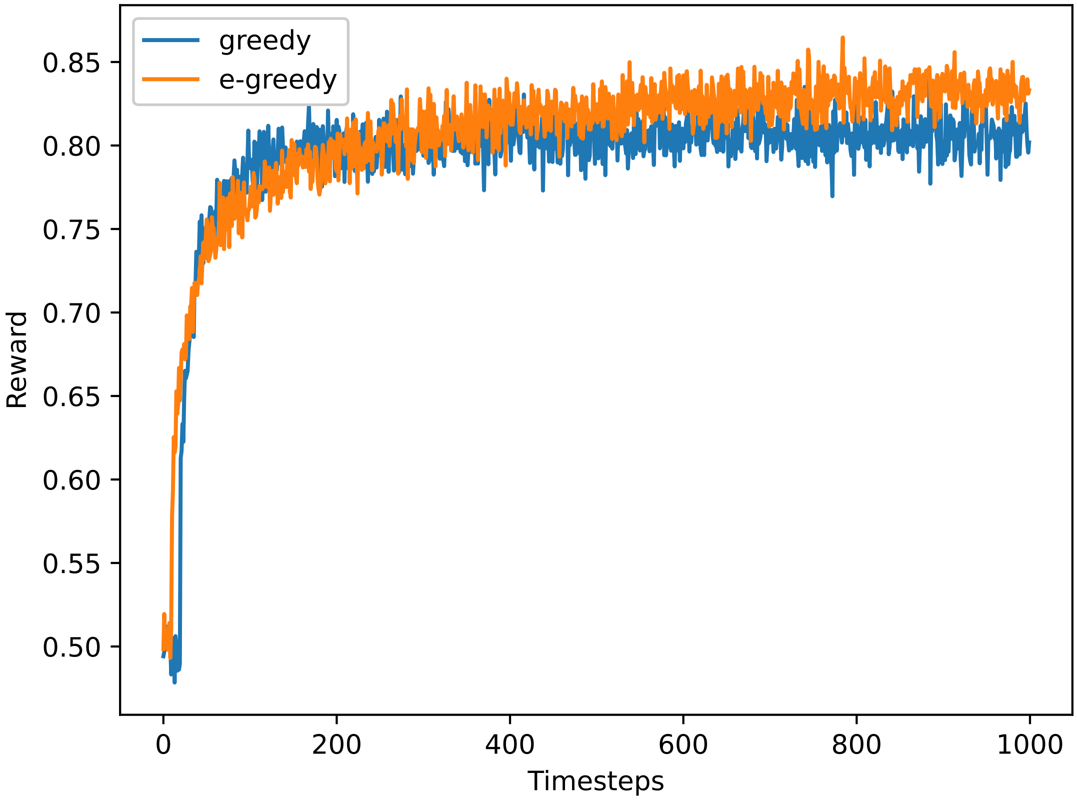

Figure 1: Bandit Strategies