Reinforcement Learning Exercises
Exercise 1

Daniel Arnold

April 26, 2021

1 Multiarmed Bandits

a) What is the probability that the greedy action is selected? (2P)

Answer: 1- $\epsilon = 0.5$

b) Consider a k-armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using -greedy action selection, sample-average action-value estimates, and initial estimates of Q_1(a) = 0, for all a. Suppose, you observe the following sequence of actions and rewards: $A_1 = 1$, $R_1 = 1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = 2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the case may have occurred, causing an action to be selected at random. (2P)

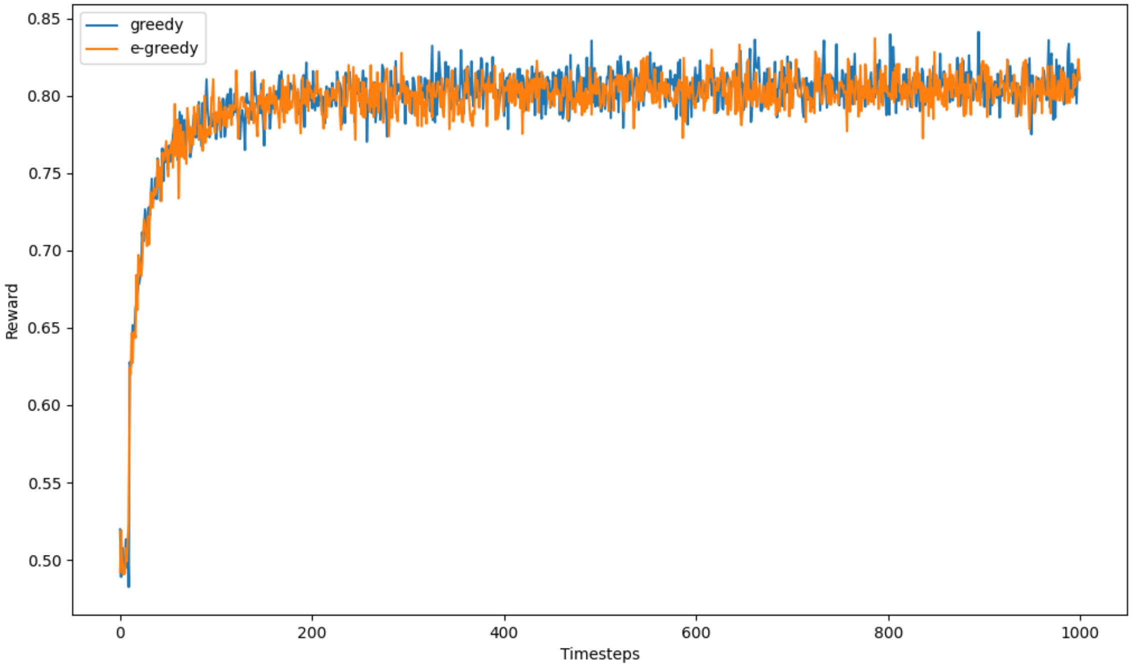
1) On which time steps did this definitely occur?
2) On which time steps could this possibly have occurred?

1. Answer: $T_1 \in \{1, 2, 3, 5\}$
2. Answer: $T_2 \in \{4\}$

2 Action Selection Strategies

c) In the main function set n episodes=10000 to create a plot with less noise (this might take some time). Add the plot into your submission pdf (The code template already stores it as an eps file). Which of the 2 methods performs better, why? (1P)

Answer: Both methods perform equally well, because ϵ
is near 0.
So most of the time (90%) the greedy acation is selected.


greedy
e-greedy
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Reward
0
200
400
600
800
1000
Timesteps

d) Think about possible ways to improve the implemented methods. What changes could you make to the strategies in order to improve them? (1P)

Answer: Try out different values for ϵ
Incrase Number of Timesteps.