Reinforcement Learning
Exercise 1

Mathias Niepert, Vinh Tong

April 8, 2024

1 Multi-armed Bandits (4 Points)

a) Consider ϵ-greedy action selection for a bandit with two actions ($k = 2$) and $\epsilon = 0.5$. What is the probability that the greedy action is selected? (2P)

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 ϵ-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: $(A_1 = 1, R_2 = 1), (A_2 = 2, R_3 = 1), (A_3 = 2, R_4 = 2), (A_4 = 2, R_5 = 2), (A_5 = 3, R_6 = 0)$. On some of these time steps the ϵ 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?

2 Action Selection Strategies (6 points)

The source code for programming exercises will be published on Ilias. The first exercise can be found as python script in *ex01-bandits/ex01-bandits.py*. The code implements a 10-armed Gaussian bandit.

a) Implement the greedy action selection strategy in the function *greedy*. Initialize the values by playing each arm once. (3P)

b) Implement the ϵ-greedy strategy in the function *epsilon_greedy*. Use $\epsilon = 0.1$. (1P)

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

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