Reinforcement Learning: An Introduction
Attempted Solutions
Chapter 2

Scott Brownlie & Rafael Rui

1 Exercise 2.1

In ϵ-greedy action selection, for the case of two actions and $\epsilon = 0.5$, what is the probability that the greedy action is selected?

The greedy action is initially selected with probability 0.5, and if not, then one of the two actions is selected randomly (each with probability 0.5). Therefore, the overall probability that the greedy action is selected is

$$0.5 + 0.5 \cdot 0.5 = 0.75.$$

2 Exercise 2.2: Bandit example

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 the initial sequence of actions and rewards is $A_1 = 1, R_1 = 1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = 2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

The action value estimates on each time step are as follows:

1. $Q_1(1) = 0, Q_1(2) = 0, Q_1(3) = 0, Q_1(4) = 0$
2. $Q_1(1) = 1, Q_1(2) = 0, Q_1(3) = 0, Q_1(4) = 0$
3. $Q_1(1) = 1, Q_1(2) = 1, Q_1(3) = 0, Q_1(4) = 0$
4. $Q_1(1) = 1, Q_1(2) = 3/2, Q_1(3) = 0, Q_1(4) = 0$
5. $Q_1(1) = 1, Q_1(2) = 5/3, Q_1(3) = 0, Q_1(4) = 0$
6. $Q_1(1) = 1, Q_1(2) = 5/3, Q_1(3) = 0, Q_1(4) = 0$

The highest actions values on each time step were $\{1, 2, 3, 4\}, \{1\}, \{1, 2\}, \{2\}, \{2\}$ and the chosen actions were $1, 2, 2, 2, 3$ respectively. Therefore, the ϵ case definitely occurred on time steps 2 and 5. As it is possible that the greedy action is chosen randomly when the ϵ case occurs, the ϵ case could possibly have occurred on any of the remaining time steps.

3 Exercise 2.3

In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively.

On each time step the probability of selecting the best action is $1 - \epsilon$ times the probability that the greedy action is the best action, plus ϵ times the probability that the best action is chosen randomly. Clearly the probability that the best action is chosen randomly is $1/k$, thus we just need to work out the probability that the greedy action is the best action.

On the first time step all actions have value 0 and we select an action A_1 at random and receive a reward R_1. Suppose that $R_1 < 0$. What is the probability that A_1 is the best action?