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?

4 Exercise 2.4

If the step-size parameters, α_n, are not constant, then the estimate Q_n is a weighted average of previously received rewards with weighting different from that given by (2.6). What is the weighting on each prior reward for the general case, analogous to (2.6), in terms of the sequence of step-size parameters?

We have

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= \alpha_n R_n + (1-\alpha_n)Q_n \\
&= \alpha_n R_n + (1-\alpha_n)[\alpha_{n-1}R_{n-1} + (1-\alpha_{n-1})Q_{n-1}] \\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})Q_{n-1} \\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + \quad \cdots + (1-\alpha_n)(1-\alpha_{n-1})\ldots(1-\alpha_2)\alpha_1 R_1 + \\
&\quad (1-\alpha_n)(1-\alpha_{n-1})\ldots(1-\alpha_2)(1-\alpha_1)Q_1 \\
&= Q_1\prod_{j=1}^{n}(1-\alpha_j) + \alpha_n R_n + \sum_{i=1}^{n-1}\alpha_i R_i \prod_{j=i+1}^{n}(1-\alpha_j).
\end{aligned}$$

Therefore, the weighting on R_n is α_n and the weighting on R_i for $i = 1, \ldots, n-1$ is $\alpha_i \prod_{j=i+1}^{n}(1-\alpha_j)$. Note that when $\alpha_i = \alpha$ for all i the expression above reduces to

$$(1-\alpha)^n Q_1 + \alpha R_n + \sum_{i=1}^{n-1}\alpha(1-\alpha)^{n-i}R_i = (1-\alpha)^n Q_1 + \sum_{i=1}^{n}\alpha(1-\alpha)^{n-i}R_i,$$

which is exactly (2.6).

5 Exercise 2.5

Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for nonstationary problems. Use a modified version of the 10-armed testbed in which all the $q_*(a)$ start out equal and then take independent random walks (say by adding a normally distributed increment with mean zero and standard deviation 0.01 to all the $q_*(a)$ on each step). Prepare plots like Figure 2.2 for an action-value method using sample averages, incrementally computed, and another action-value method using a constant step-size parameter, $\alpha = 0.1$. Use $\epsilon = 0.1$ and longer runs, say of 10,000 steps.

As we can see from Figure 1, the sample averaging method chooses the optimal action less than 50% of the time, even after 10,000 time steps. Using a constant step-size of 0.1 works better, as shown in Figure 2 where the optimal action is chosen around 75% of the time after 10,000 time steps. The sample averaging method does not work as well because after a significant number of time steps the step-size of $1/n$ becomes very small and the action-value estimates cannot adapt to the non-stationary environment.

Figure 1: Average performance of non-stationary K-armed bandit with sample averaging.

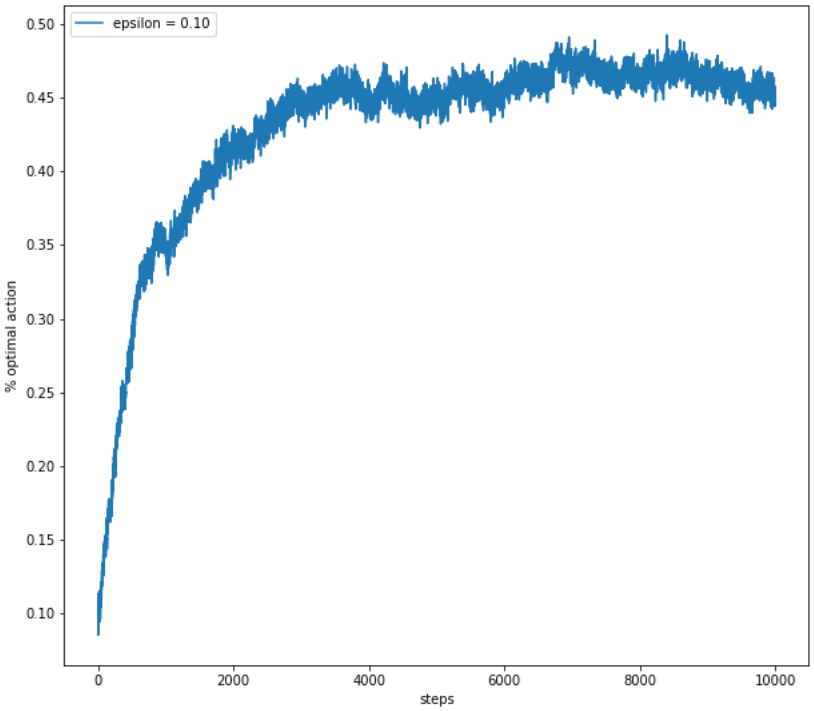

Figure 2: Average performance of non-stationary K-armed bandit with constant step-size $\alpha = 0.1$.