Chapter 2: Multi-armed Bandits - Exercises

Jacob Taylor Cassady

November 26, 2022

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

50 percent.

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.

See Figure 1 at the end of the document for the output of the code I wrote for this.

2.1 On which time steps did this definitely occur?

Note. I am indexing the time steps by 0. Answer: [3, 4]

2.2 On which time steps could this possibly have occurred?

Note. I am indexing the time steps by 0. Answer: [0, 1, 2]

3 In the comparison shown in Figure 2.2,

Express your answer quantitatively.

3.1 which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action?

ε=0.01

3.2 How much better will it be?

ε=0.1 would perform at best 91% optimal. ε=0.01 would perform at best 99.1% optimal. After convergence, ε=0.01 would perform 8.1% better.

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 a weighting different from that given by (2.6).

4.1 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?

5 programming

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 $\varepsilon = 0.1$ and longer runs, say of 10,000 steps.

6 Mysterious Spikes

7 Unbiased Constant-Step-Size Trick

8 UCB Spikes

9 Show that in the case of two actions, the softmax distribution is the same as that given by the logistic, or sigmoid, function often used in statistics and artificial neural networks.

10 Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 0.1 and 0.2 with probability 0.5 (case A), and 0.9 and 0.8 with probability 0.5 (case B).

10.1 If you are not able to tell which case you face at any step, what is the best expectation of success you can achieve and how should you behave to achieve it?

10.2 Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expectation of success you can achieve in this task, and how should you behave to achieve it?

11 programming

Make a figure analogous to Figure 2.6 for the nonstationary case outlined in Exercise 2.5. Include the constant-step-size ε-greedy algorithm with α=0.1. Use runs of 200,000 steps and, as a performance measure for each algorithm and parameter setting, use the average reward over the last 100,000 steps.

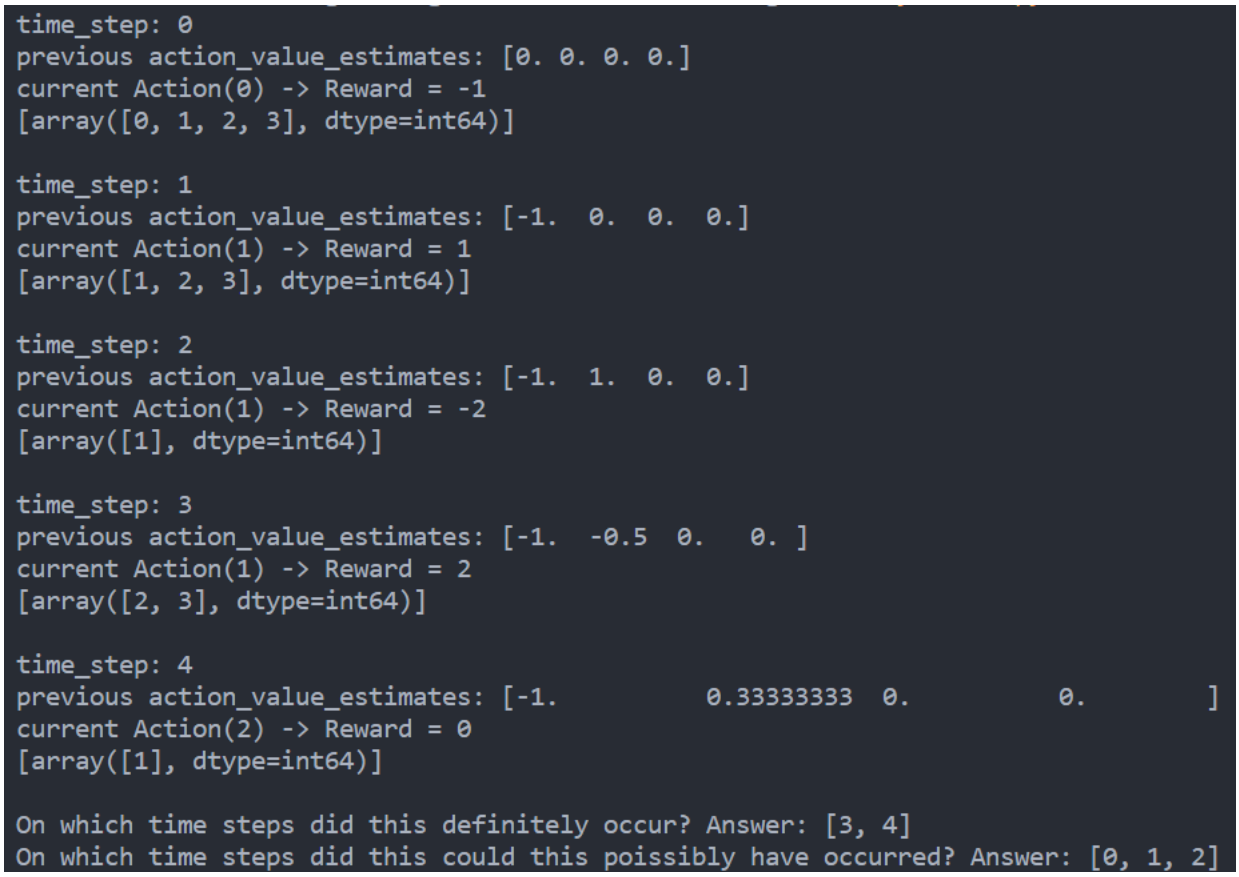

Figure 1: Output of bandit_example.py