1 Multiarmed Bandits

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?

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?

3. Exercise 2.3 In the comparison shown in Figure 2.2 (below), 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.

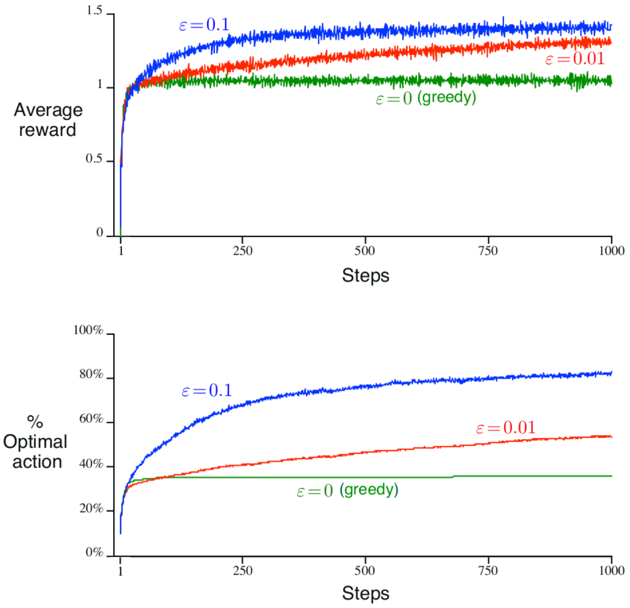


Figure 2.2: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 runs with different bandit problems. All methods used sample averages as their action-value estimates.

4. Implementation Task: 10 armed-testbed

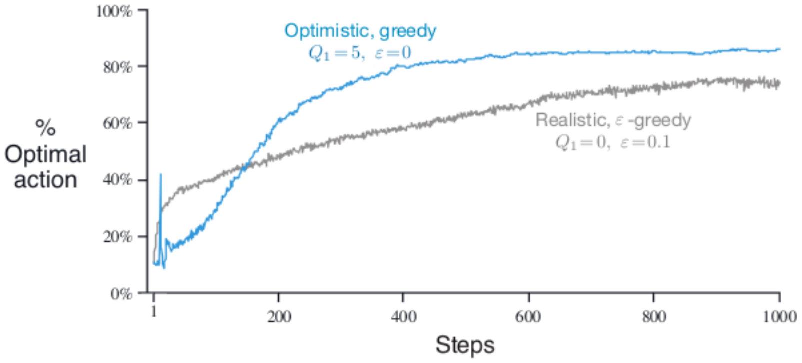
   Implement a simple bandit algorithm for the "10-armed-testbed". (Ten independent bandits whose action value a_i is taken from a normal distribution with mean 0 and variance 1 for $i \in \{1, 2..., 10\}$; the distribution of each reward should be defined from a normal distribution with mean a_i and variance 1)

   Repeat the experiment 1000 time steps and average over 1000 independent runs for at least 3 different epsilon values and plot the results.

5. 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 a 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?

6. Exercise 2.6 Mysterious Spikes

   The results shown in Figure 2.3 (below) should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps?


Optimistic, greedy
$Q_1 = 5,\ \varepsilon = 0$
Realistic, ε-greedy
$Q_1 = 0,\ \varepsilon = 0.1$
Optimal
action
Steps