Worksheet C1M1

1. Suppose a game where you choose to flip one of two (possibly unfair) coins. You win $1 if your chosen coin shows heads and lose $1 if it shows tails.

(a) Model this as a K-armed bandit problem: define the action set.

(b) Is the reward a deterministic or stochastic function of your action?

(c) You do not know the coin flip probabilities. Instead, you are able to view 6 sample flips for each coin respectively: (T,H,H,T,T,T) and (H,T,H,H,H,T). Use the sample average formula (equation 2.1 in the book) to compute the estimates of the value of each action.

(d) Decide on which coin to flip next! Assume it's an exploit step.

2. Consider a problem where an agent is trying to get to school and must choose how long to wait at the bus stop. The agent can walk to school, but wants to catch the bus if possible. At the same time, the agent doesn't want to wait too long because of delays. Unfortunately, the time it takes for a bus to arrive is effectively random.

(a) This is not a K-armed bandit problem because your action set, how long to wait, is not a positive integer. How could you reformulate the bus-waiting problem as a K-armed bandit?

(b) In problems with continuous random variables, we rarely know the distribution of a variable. Instead, we often make assumptions on its distribution. One commonly assumed distribution for continuous random variables is the Gaussian (or Normal) distribution. Is the Gaussian assumption in this bus-waiting problem reasonable? Justify your answer using properties of the Gaussian distribution and other assumptions about the distribution of time spent waiting at the bus stop.

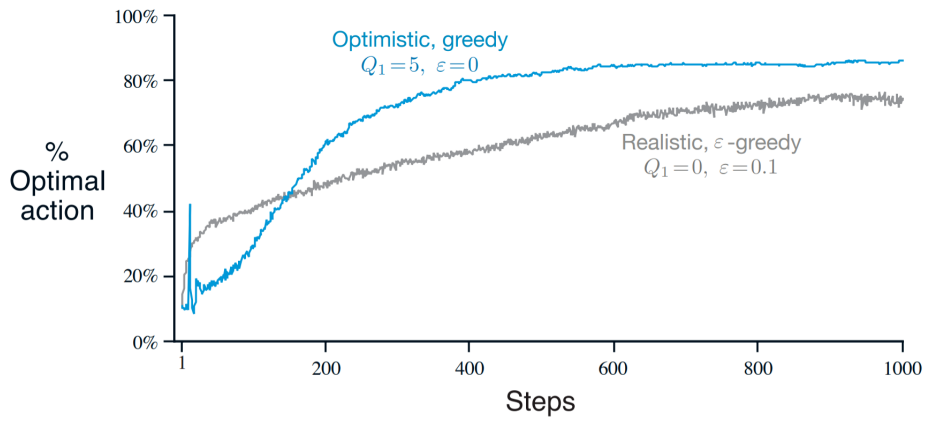

Figure 2.3: The effect of optimistic initial action-value estimates on the 10-armed testbed. Both methods used a constant step-size parameter, $\alpha = 0.1$.

3. (Exercise 2.2 from S&B 2nd edition) 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?

4. **Challenge Problem:** (Exercise 2.6 from S&B 2nd edition) The results shown in Figure 2.3 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?