Name: Alfred
Student No.: 2020-12788

Exercise 1

Data Science & Reinforcement Learning

Spring 2023

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

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

   2 actions: Coin A or Coin B
   (first) (second)

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

   Stochastic, the actions are conditioned

   (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 textbook) to compute the estimates of the value of each action.

   Coin A: $= \frac{-1}{3}$ Coin B: $= \frac{1}{3}$

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

   we flip Coin B to exploit more. (second Coin)
   because the probability of good reward is higher.

epsilon greedy
timing for better
perfor.
UCB variations

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?

K. Constraint.
1 no wait → q_1
2 wait 1min → q_2
3 wait 2min → q_3
K wait K min → q_K

it would be better to do it by unit of time, or discretizing time step. K would be the max time the agent can hypothetically wait. At each step decide to go or not.

(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 on the wait time in this bus-waiting problem reasonable? Justify your answer using properties of the Gaussian distribution.

Since gaussian distribution needs to be symmetric. and there's a need for a peak, from agent pov the waiting time there couldn be random, we don't know actual starting point.

3. 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?

A_n	1	2	2	4	5
R_n	-1	1	-2	2	0

	1	2	3	4
1	-1	0	0	0
2	-1	1	0	0
3	-1	r_2	0	0
4	-1	r_2	0	0

$Q_1 = 0$ $Q_2 = 0$ $Q_3 = 0$ $Q_4 = 0$

it's definitely random at step 4 at 5 since the previous one have defined action values

4. Express the action-value function q_π in terms of v_π. The formula will also include p and π.

$$q_*(s,a) \doteq \mathbb{E}[R_{t+1} + \gamma_t(S_{t+1}) | S_t = s, A_t = a] = \sum_{s' \in S} P(s'|s,a)\left[R(s,a,s') + \gamma V(s')\right]$$

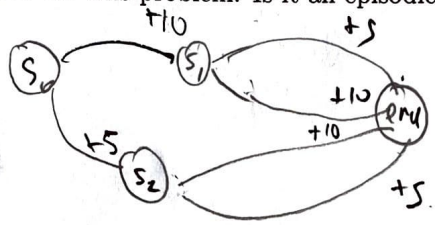
where $s' = S_{t+1}$, $s = S_t$

$$V_\pi = \sum_{a \in A} \pi(a|s) q_\pi(s|a)$$

5. In this question, you will take a word specification of an MDP, and write the formal terms and determine the optimal policy. Suppose you have a problem with two actions. The agent always starts in the same state, s_0. From this state, if it takes action 1 it transitions to a new state s_1 and receives reward 10; if it takes action 2 it transitions to a new state s_2 and receives reward 5. From s_1 if it takes action 1 it receives a reward of 5 and terminates; if it takes action 2 it receives a reward of 10 and terminates. From s_2 if it takes action 1 it receives a reward of 10 and terminates; if it takes action 2 it receives a reward of 5 and terminates. Assume the agent cares equally about long-term reward as about immediate reward.

(a) Draw the MDP for this problem. Is it an episodic or continuing problem? What is γ?

upper: action 1 → lower action 2



because needs long term $\gamma=1$ & episodic.

(b) Assume $\pi(a=1|s_i)=0.3$ for all $s_i \in \{s_0, s_1, s_2\}$. What is $\pi(a=2|s_i)$? And what is the value function for this policy? In other words, find $v_\pi(s)$ for all three states.

& $\pi(a=2|s_i) = 1-\pi(a=1|s_i) = 0.7$

then:

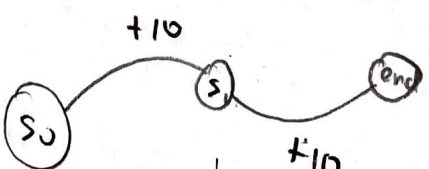
$V_\pi(s_2) = 0.3\times 10 + 5\times 0.7 = 6.5$

$V_\pi(s_1) = 0.3\times 5 + 10\times 0.7 = 8.5$

$V_\pi(s_0) = 0.3(10+V_\pi(s_1)) + 0.7(5+V_\pi(s_2)) = 13.16$

(c) What is the optimal policy in this environment?

the optimal policy is by principal of optimality. for the second part we pick upper with action 2 and lower with action 1 then at the first step we see the max reward happens for when taking action 1.



Max reward.

$\pi(a=1|s_0)=1$ | $\pi(a=2|s_1)=1$ | $\pi(a=1|s_2)=1$