Q2:

1. Assume you have a bandit problem with 4 actions, where the agent can see rewards from the set R = {−3.0, −0.1, 0, 4.2}. Assume you have the probabilities for rewards for each action: p(r|a) for a ∈ {1, 2, 3, 4} and r ∈ {−3.0, −0.1, 0, 4.2}. How can you write this problem as an MDP? Remember that an MDP consists of (S, A, R, P, γ).

Answer:

S={s0} (one state),
A={1,2,3,4} (four possible actions),
R={−3.0,−0.1,0,4.2} (set of possible rewards),
P(s′|s,a)=1 (the agent remains in state s0 after any action),
γ=0 (since it's a bandit problem with no future time horizon).

2. Compute Advantage of action for the following example.

π

→	→	H
→	→	→
G	→	→

$Q_\pi(s, a)$

0.536	0.351	H
0.725	0.539	0.492 ←
G	0.682	0.521

$V_\pi(s)$

0.335	0.152	H
0.412	0.321	0.192
G	0.398	0.212

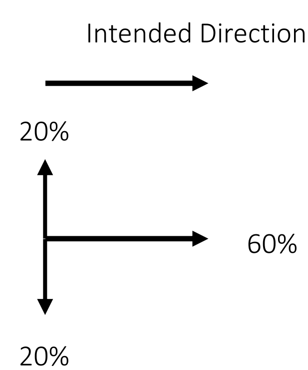
Action Advantage function

0.201	0.199	H
0313	0.218	0.3
G	0.284	0.309

3. Fill out the table for the following: states **3, 4** and actions **up and down** only. **[3 marks]**
   Given:

0	1	2 **H**
3 **G**	4	5

State transition:



$Actions\ set\ = \{up, down, left, right\}$, , **Rewards** G= +2, H= -1, 0 other states

State	Action	Next State	Transition Probability	Reward
3	up	3	1	0
3	up	3	1	0
3	up	3	1	0
3	down	3	1	0
3	down	3	1	0

State	Action	Next State	Transition Probability	Reward
3	down	3	1	0
4	up	1	0.6	0
4	up	5	0.2	0
4	up	3	0.2	2
4	down	4	0.6	0
4	down	5	0.2	0
4	down	3	0.2	2

4.

	Policy 1	Policy 2	Policy 3
Policy table	(see below)	(see below)	(see below)
State values and action values	(see below)	(see below)	(see below)

Policy 1

Policy table

	a1	a2
S1	0.4	0.6
S2	0.2	0.8
S3	0.7	0.3
S4	0.6	0.4
S5	0.5	0.5

State values and action values

	V
S1	1.6
S2	1.4
S3	1.7
S4	2.3
S5	0.9

	a1	a2
S1	1.8	1.2
S2	1.5	1.3
S3	1.2	1.6
S4	2.1	1.8
S5	0.8	1.0

Policy 2

Policy table

	a1	a2
S1	0.0	1.0
S2	0.0	1.0
S3	1.0	0.0
S4	1.0	0.0
S5	0.0	1.0

State values and action values

	V
S1	3.1
S2	3.3
S3	3.5
S4	2.9
S5	2.8

	a1	a2
S1	3.0	3.3
S2	3.2	3.4
S3	3.6	3.1
S4	3.0	2.8
S5	2.8	3.2

Policy 3

Policy table

	a1	a2
S1	0.6	0.4
S2	0.3	0.7
S3	0.7	0.3
S4	0.5	0.5
S5	0.2	0.8

State values and action values

	V
S1	2.4
S2	2.1
S3	2.5
S4	2.9
S5	1.9

	a1	a2
S1	2.8	2.2
S2	2.5	1.9
S3	2.3	1.8
S4	2.7	3.0
S5	1.8	2.1

a. Is Policy 1 deterministic or stochastic? Explain why?

Stochastic, because the policies are derived from a distribution e.g form s1: $\pi(a1|s1) = 0.4$ and $\pi(a2|s1) = 0.6$

b. Which of the policies given the figure is an optimal policy?

Policy 2

c. Choose optimal state values and optimal action values:

Policy 2

Q3:

Part I: 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?

(b) Assume the policy is $\pi(a = 1|s_i)$ = 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.

MDP Diagram:

1. States: s0s_0s0, s1s_1s1, s2s_2s2
2. Actions: Action 1, Action 2
3. Transitions: Defined by the action taken and the resulting state.
4. Rewards: Given for each transition.
5. Terminal States: Both s1s_1s1 and s2s_2s2 are terminal after taking action 1 or 2.

```
              (10)
     s_0 -----------------> s_1
      |                      |
(5)   |                      | (5)
      |                      |
      v                      v
      s_2 <--------------- s_1
              (10)
```

Discount factor:

γ=0.5.

Policy $\boldsymbol{\pi(a = 1 \mid si) = 0.3}$ for all si:

- Given $\pi(a = 1 \mid si) = 0.3$, the probability of taking action 1 in any state is 0.3.
- The probability of taking action 2 is the complement of this, i.e., $\pi(a = 2 \mid si) = 1 - \pi(a = 1 \mid si) = 0.7$
- So, for all $si \in \{s0, s1, s2\}$:
    - $\pi(a = 1 \mid si) = 0.3$
    - $\pi(a = 2 \mid si) = 0.7$

Q4: For the following environment given in figure and episode (trajectory), compute discounted return. The discount factor is $\gamma = 0.8$. The reward at the goal state is $+3$. The reward for being Hole is -3. For each transition (other than terminal states) maintenance reward is -0.07.

Episode: 5, 8, 7, 4, 1, 0, 3, 4, 7, 4, 3, 6

π

0 →	1 →	2 H
3 →	4 →	5 →
6 G	7 →	8 →

= 1 * (-0.07) + 0.8 *(-0.07) + 0.64 *(-0.07) + 0.512 *(-0.07) + 0.4096* (-0.07) +

0.327*(-0.07) + 0.26*(-0.07) +0.2*(-0.07) + 0.16 * (-0.07) + 0.13*(-0.07) + 0.1*(-0.07)

+0.08*3

= -0.07 - 0.056 - 0.0448 - 0.03584 - 0.028 - 0.022 - 0.0182 - 0.014 - 0.0112 - 0.0091 - 0.007

+**0.24**

= -0.31614 +0.24

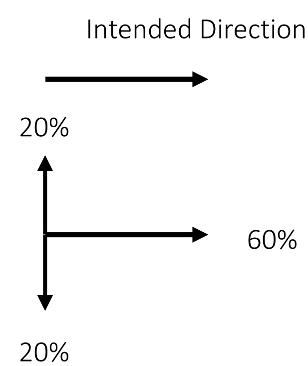
= - 0.07614

Q5:

Write python code for the following environment:

0	1	2 **H**
3 **G**	4	5

State transition:



0 left, 1 up, 2 right, 3 down

```
P = {
    0: {
        0: [ (0.6, 0, 0.0, FALSE) , (0.2, 0, 0.0, FALSE) , (0.2, 3, 1.0, TRUE) ] ,
        1: [ (0.6, 0, 0.0, FALSE) , (0.2, 0, 0.0, FALSE) , (0.2, 1, 0.0, FALSE) ],
        2: [ (0.6, 1, 0.0, FALSE) , (0.2, 0, 0.0, FALSE) , (0.2, 3, 1.0, TRUE) ],
        3: [ (0.6, 0, 0.0, FALSE) , (0.2, 0, 0.0, FALSE) , (0.2, 3, 1.0, TRUE) ],
    },
    1:{
        0: [ (0.6, 0, 0.0, FALSE) , (0.2, 4, 0.0, FALSE) , (0.2, 1, 0.0, FALSE) ] ,
```

```
        1: [ (0.6, 1, 0.0, FALSE) ,  (0.2, 0, 0.0, FALSE) ,   (0.2, 2, -1.0, TRUE)  ],
        2: [ (0.6, 2, -1.0, TRUE) ,  (0.2, 1, 0.0, FALSE) ,   (0.2, 4, 0.0,  FALSE)  ],
        3: [ (0.6, 4, 0.0, FALSE) ,  (0.2, 2, -1.0, TRUE) ,   (0.2, 1, 0.0, FALSE)  ],
    },
    2:{
        0: [ (1.0, 2, 0.0, TRUE) ,  (1.0, 2, 0.0, TRUE) ,   (1.0, 2, 0.0, TRUE)  ] ,
        1: [ (1.0, 2, 0.0, TRUE) ,  (1.0, 2, 0.0, TRUE) ,   (1.0, 2, 0.0, TRUE)  ],
        2: [ (1.0, 2, 0.0, TRUE) ,  (1.0, 2, 0.0, TRUE) ,   (1.0, 2, 0.0, TRUE)  ],
        3: [ (1.0, 2, 0.0, TRUE) ,  (1.0, 2, 0.0, TRUE) ,   (1.0, 2, 0.0, TRUE)  ],
    },
    3:{
        0: [ (1.0, 3, 0.0, TRUE) ,  (1.0, 3, 0.0, TRUE) ,   (1.0, 3, 0.0, TRUE)  ] ,
        1: [ (1.0, 3, 0.0, TRUE) ,  (1.0, 3, 0.0, TRUE) ,   (1.0, 3, 0.0, TRUE)  ],
        2: [ (1.0, 3, 0.0, TRUE) ,  (1.0, 3, 0.0, TRUE) ,   (1.0, 3, 0.0, TRUE)  ],
        3: [ (1.0, 3, 0.0, TRUE) ,  (1.0, 3, 0.0, TRUE) ,   (1.0, 3, 0.0, TRUE)  ],
    },

    4:{
        0: [ (0.6, 3, 1.0, TRUE) ,  (0.2, 1, 0.0, FALSE) ,   (0.2, 4, 1.0, FALSE)  ],
        1: [ (0.6, 1, 0.0, FALSE) ,  (0.2, 3, 1.0, TRUE) ,   (0.2, 5, 0.0, FALSE)  ],
        2: [ (0.6, 5, 0.0, FALSE) ,  (0.2, 1, 0.0, FALSE) ,   (0.2, 4, 0.0, FALSE)  ],
        3: [ (0.6, 4, 0.0, FALSE) ,  (0.2, 5, 0.0, FALSE) ,   (0.2, 3, 1.0, TRUE)  ],
    },
    5:{
        0: [ (0.6, 4, 0.0, FALSE) ,  (0.2, 5, 0.0, FALSE) ,   (0.2, 2, -1.0, TRUE)  ] ,
        1: [ (0.6, 2, -1.0, TRUE) ,  (0.2, 4, 0.0, FALSE) ,   (0.2, 5, 0.0, FALSE)  ],
        2: [ (0.6, 5, 0.0, FALSE) ,  (0.2, 2, -1.0, TRUE) ,   (0.2, 5, 0.0, FALSE)  ],
        3: [ (0.6, 5, 0.0, FALSE) ,  (0.2, 5, 0.0, FALSE) ,   (0.2, 4, 1.0, FALSE)  ],
    },
}
```

***