Assignment 2 Report

Course Code: AI3000
Course Name: Reinforcement Learning

Ahmik Virani
Roll Number: ES22BTECH11001
IIT Hyderabad

9 October, 2025

Contents

Chapter 1

Problem 1: Importance Sampling

1.1 Part(a)

From class notes, we know that:

$$V(s_t) \leftarrow V(s_t) + \alpha_t \left[\frac{\pi(s_t|a_t)}{\mu(s_t|a_t)} (r_{t+1} + \gamma V(s_{t+1})) - V(s_t) \right]$$

Where, π is the target policy and μ is the behavior policy.

- In the question we are also given that the horizon length is 1. Thus we can eliminate V_{t+1} and also take $\alpha_t = 1$
- We are also given that the action taken is a and the reward is r.
- In the question, behavior policy is given as π_b

Using the above facts, the estimate becomes:

$$V^\pi = \frac{\pi(a)}{\pi_b(a)} r$$

The above is an unbiased estimate because this is essentially the Monte Carlo estimate, since the horizon length is just 1.

1.2 Part(b)

$$\begin{aligned} \mathbb{E}_{\pi_b} \left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} \right] &= \sum_a \pi_b(a) \frac{\pi(a)}{\pi_b(a)} \\ &= \sum_a \pi(a) \\ &= 1 \end{aligned}$$

Thus

$$\mathbb{E}_{\pi_b} \left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)} \right] = 1$$

1.3 Part(c)

We need

$$\frac{\pi(a)}{\pi_b(a)}$$

- Given π_b is a uniformly random policy, thus, $\pi_b(a) = \frac{1}{K}$
- Given π is deterministic, thus it is 1 if we pick some action a' else 0.

$$\frac{\pi(a)}{\pi_b(a)} = \begin{cases} K & \text{if } a = a' \\ 0 & \text{otherwise.} \end{cases}$$

1.4 Part(d)

$$P(\tau) = \mu(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t) P(s_{t+1}|a_t, s_t)$$

$$Q(\tau) = \mu(s_0) \prod_{t=0}^{\infty} \pi_b(a_t|s_t) P(s_{t+1}|a_t, s_t)$$

Thus,

$$\frac{P(\tau)}{Q(\tau)} = \prod_{t=0}^{\infty} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$$

Chapter 2

Problem 2: Q Learning

2.1 Part(a)

Since there is no stochasticity, the deterministic optimal policy π^* would be just to take the path which would lead to the maximum reward, which is 100 (state 2 in the diagram).

Thus, the value of any non-terminal state for this policy would just be 100 (where I have taken $\gamma = 1$ as it is finite horizon).

$$V^*(s) = 100, \text{for s} = \{0,3,4,5\}$$

2.2 Part(b)

We start with the initial state action pair table for Q and initialize all values to 0.

States vs. Actions	Left	Right	Up	Down
0	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0

Table 2.1: Initial State Action pair table for Q

Episode 1

1. (0,Down,3,0), Q(0, Down) remains unaffected since all values are 0. Q(0, Down) = 0
2. (3,Right,4,0), Q(3, Right) remains unaffected since all values are 0. Q(3, Right) = 0
3. (4, Down, 7, -100):

$$\begin{aligned} Q(4,\ \text{Down}) &\leftarrow Q(4,\ \text{Down}) + 0.5\big(r - Q(4,\ \text{Down})\big) \\ &\leftarrow 0 + 0.5(-100 - 0) \end{aligned}$$

Thus $Q(4,\ \text{Down}) = -50$.

States vs. Actions	Left	Right	Up	Down
0	0	0	0	0
3	0	0	0	0
4	0	0	0	-50
5	0	0	0	0

Table 2.2: State Action pair table for Q after episode 1

Episode 2

1. (0, Down, 3, 0):

$$\begin{aligned} Q(0,\ \text{Down}) &\leftarrow Q(0,\ \text{Down}) + 0.5\big(r + 0.5 * \max_{a'} Q(3, a') - Q(0,\ \text{Down})\big) \\ &\leftarrow 0 + 0.5(0 + 0 - 0) \end{aligned}$$

   Thus $Q(0,\ \text{Down}) = 0$.

2. (3, Right, 4, 0):

$$\begin{aligned} Q(3,\ \text{Right}) &\leftarrow Q(3,\ \text{Right}) + 0.5\big(r + 0.5 * \max_{a'} Q(4, a') - Q(3,\ \text{Right})\big) \\ &\leftarrow 0 + 0.5(0 + 0 - 0) \end{aligned}$$

   Thus $Q(3,\ \text{Right}) = 0$.

3. (4, Right, 5, 0):

$$\begin{aligned} Q(4,\ \text{Right}) &\leftarrow Q(4,\ \text{Right}) + 0.5\big(r + 0.5 * \max_{a'} Q(5, a') - Q(4,\ \text{Right})\big) \\ &\leftarrow 0 + 0.5(0 + 0 - 0) \end{aligned}$$

   Thus $Q(4,\ \text{Right}) = 0$.

4. (5, Up, 2, 100):

$$\begin{aligned} Q(5,\ \text{Up}) &\leftarrow Q(5,\ \text{Up}) + 0.5\big(r - Q(5,\ \text{Up})\big) \\ &\leftarrow 0 + 0.5(100 - 0) \end{aligned}$$

   Thus $Q(5,\ \text{Up}) = 50$.

States vs. Actions	Left	Right	Up	Down
0	0	0	0	0
3	0	0	0	0
4	0	0	0	-50
5	0	0	50	0

Table 2.3: State Action pair table for Q after episode 2

Episode 3

1. (0, Down, 3, 0):

$$\begin{aligned} Q(0,\ \text{Down}) &\leftarrow Q(0,\ \text{Down}) + 0.5\big(r + 0.5 * \max_{a'} Q(3, a') - Q(0,\ \text{Down})\big) \\ &\leftarrow 0 + 0.5(0 + 0 - 0) \end{aligned}$$

   Thus $Q(0,\ \text{Down}) = 0$.

2. (3, Right, 4, 0):

$$\begin{aligned} Q(3,\ \text{Right}) &\leftarrow Q(3,\ \text{Right}) + 0.5\big(r + 0.5 * \max_{a'} Q(4, a') - Q(3,\ \text{Right})\big) \\ &\leftarrow 0 + 0.5(0 + 0 - 0) \end{aligned}$$

   Thus $Q(3,\ \text{Right}) = 0$.

3. (4, Right, 5, 0):

$$\begin{aligned} Q(4,\ \text{Right}) &\leftarrow Q(4,\ \text{Right}) + 0.5\big(r + 0.5 * \max_{a'} Q(5, a') - Q(4,\ \text{Right})\big) \\ &\leftarrow 0 + 0.5(0 + 25 - 0) \end{aligned}$$

   Thus $Q(4,\ \text{Right}) = 12.5$.

4. (5, Down, 8, 80):

$$\begin{aligned} Q(5,\ \text{Down}) &\leftarrow Q(5,\ \text{Down}) + 0.5\big(r - Q(5,\ \text{Down})\big) \\ &\leftarrow 0 + 0.5(80 - 0) \end{aligned}$$

   Thus $Q(5, \text{Down}) = 40$.

States vs. Actions	Left	Right	Up	Down
0	0	0	0	0
3	0	0	0	0
4	0	12.5	0	-50
5	0	0	50	40

Table 2.4: State Action pair table for Q after episode 2

Thus, the question asked us:

- $Q(5,\ \text{Up}) = 50$
- $Q(3,\ \text{Down}) = 0$
- $Q(4,\ \text{Right}) = 12.5$

2.3 Part(c)

In class we saw that the Robbins-Monroe condition is:

- $\sum_t \alpha_t = \infty$
- $\sum_t \alpha_t^2 < \infty$

2.3.1 Sub-part(i)

Given, $\alpha_t = \frac{1}{t}$, thus by integral test we have,

- $\int_1^\infty \frac{1}{t} dt = [ln(t)]_1^\infty = \infty$. Thus, $\sum_t \frac{1}{t}$ diverges.
- $\int_1^\infty \frac{1}{t^2} dt = [-\frac{1}{t}]_1^\infty = 1$. Thus, $\sum_t \frac{1}{t^2}$ converges.

Hence, $\alpha_t = \frac{1}{t}$ **obeys** Robbins-Monroe condition.

2.3.2 Sub-part(ii)

Given, $\alpha_t = \frac{1}{t^2}$. We saw in the above part (i) itself that $\sum_t \frac{1}{t^2}$ converges. Hence, $\alpha_t = \frac{1}{t^2}$ **does not obey** Robbins-Monroe condition.

Chapter 3

Problem 3: Game of Tic-Tac-Toe

This question is entirely answered in the python notebook attached.