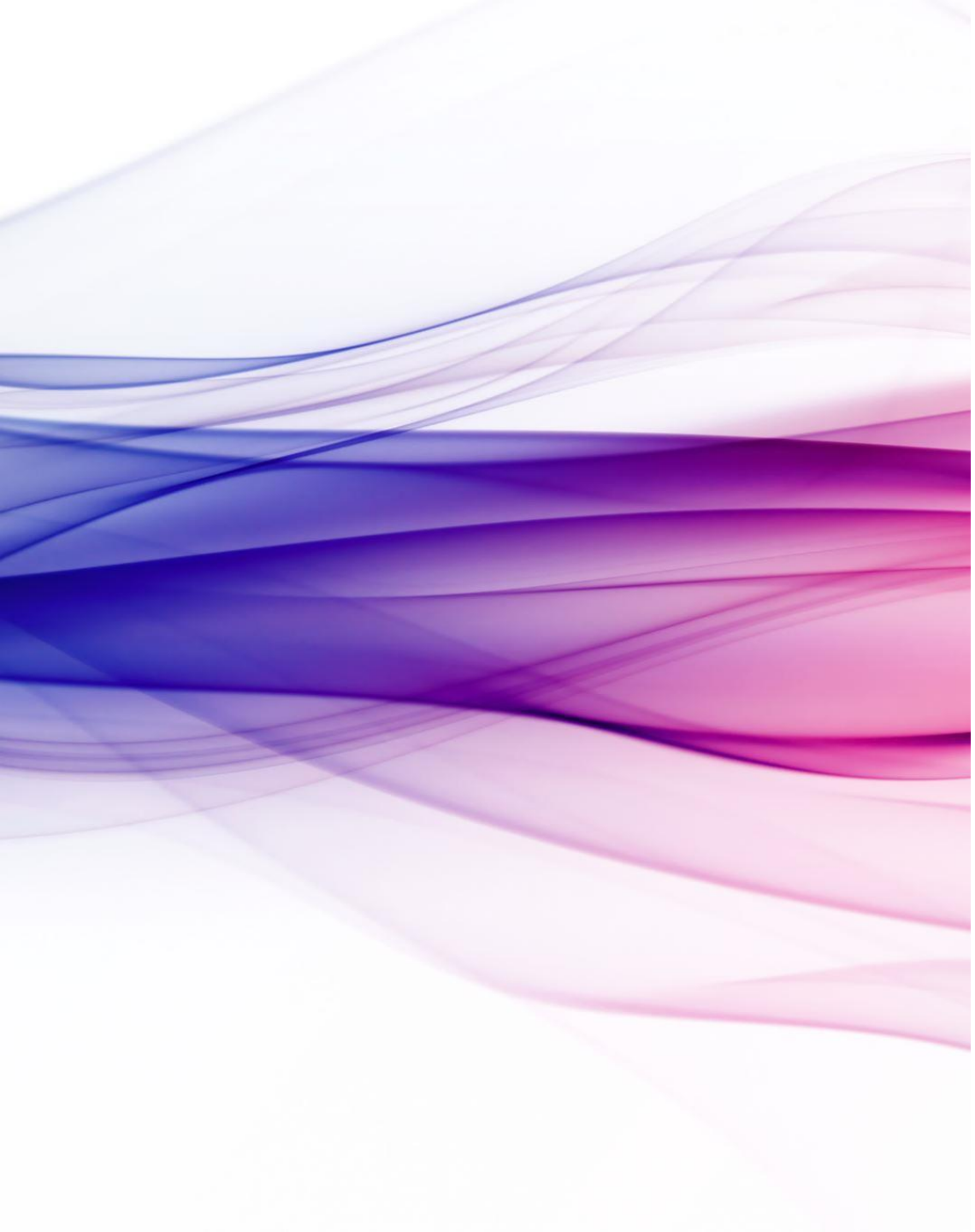

Q_8 – HW_3

Alireza Abbasi - 403616204

First 6 episode Trajectories

```
Trajectory:
['RU_8p', 'R', 0, 'RU_10p', 'P', 2, 'RU_10a', 'any', 0, 'Terminal']
Return of episode 1:
1.6
Trajectory:
['RU_8p', 'P', 2, 'TU_10p', 'R', 0, 'RU_8a', 'P', 0, 'TU_10a', 'any', -1, 'Terminal']
Return of episode 2:
1.488
Trajectory:
['RU_8p', 'P', 2, 'TU_10p', 'R', 0, 'RU_8a', 'R', 0, 'RU_10a', 'any', 0, 'Terminal']
Return of episode 3:
2.0
Trajectory:
['RU_8p', 'P', 2, 'TU_10p', 'R', 0, 'RU_8a', 'S', 0, 'RD_10a', 'any', 4, 'Terminal']
Return of episode 4:
4.048
Trajectory:
['RU_8p', 'R', 0, 'RU_10p', 'R', 0, 'RU_8a', 'S', 0, 'RD_10a', 'any', 4, 'Terminal']
Return of episode 5:
2.0480000000000005
Trajectory:
['RU_8p', 'P', 2, 'TU_10p', 'P', 2, 'RU_10a', 'any', 0, 'Terminal']
Return of episode 6:
3.6
```

Mean return after 50 episodes

```
Mean return of all episodes:
1.7342400000000004
```

Manually calculating values:

RU 8P => s_1
TU 10p => s_2
RU 10p => s_3
RD 10p => s_4
RU 8a => s_5
RD 8a => s_6
TU 10a => s_7
RU 10a => s_8
RD 10a => s_9
TD 10a => s_{10}
Terminal state => s_{11}
$V(s_{11}) = 0$

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)[r + \gamma v_\pi(s')], \text{ for all } s \in \mathcal{S},$$

$$V(s_7) = \pi(P \mid s_7)p(s', r \mid s_7, P)[r + \gamma v_\pi(s_{11})] + \pi(R \mid s_7)p(s_{11}, r \mid s_7, R)[r + \gamma v_\pi(s_{11})] + \pi(S \mid s_7)p(s_{11}, r \mid s_7, S)[r + \gamma v_\pi(s_{11})]$$

$$V(s_7) = [-1 + 0.8v_\pi(s_{11})] + [-1 + 0.8v_\pi(s_{11})] + [-1 + 0.8v_\pi(s_{11})] = -1 - 1 - 1 = -3$$

$$V(s_8) = [0 + 0.8v_\pi(s_{11})] + [0 + 0.8v_\pi(s_{11})] + [0 + 0.8v_\pi(s_{11})] = 0 + 0 + 0 = 0$$

$$V(s_9) = [4 + 0.8v_\pi(s_{11})] + [4 + 0.8v_\pi(s_{11})] + [4 + 0.8v_\pi(s_{11})] = 4 + 4 + 4 = 12$$

$$V(s_{10}) = [3 + 0.8v_\pi(s_{11})] + [3 + 0.8v_\pi(s_{11})] + [3 + 0.8v_\pi(s_{11})] = 3 + 3 + 3 = 9$$

$$V(s_5) = [2 + 0.8V(s_7)] + [0 + 0.8V(s_8)] + [-1 + 0.8V(s_9)] = -0.4 + 0 + 8.6 = 8.2$$

$$V(s_6) = [0 + 0.8V(s_9)] + [2 + 0.8V(s_{10})] = 9.6 + 9.2 = 18.8$$

$$V(s_4) = [0 + 0.8V(s_6)] + 0.5[2 + 0.8V(s_6)] + 0.5[2 + 0.8V(s_9)] = 15.04 + 8.52 + 5.8 = 29.36$$

$$V(s_3) = [0 + 0.8V(s_5)] + 0.5[2 + 0.8V(s_5)] + 0.5[2 + 0.8V(s_8)] + [-1 + 0.8v(s_6)] = 6.56 + 4.28 + 1 = 11.84$$

$$V(s_2) = [0 + 0.8V(s_5)] + [2 + 0.8V(s_8)] = 6.56 + 2 = 8.56$$

$$V(s_1) = [2 + 0.8V(s_2)] + [0 + 0.8V(s_3)] + [-1 + 0.8V(s_4)] = 8.848 + 9.472 + 22.488 = 40.8080$$