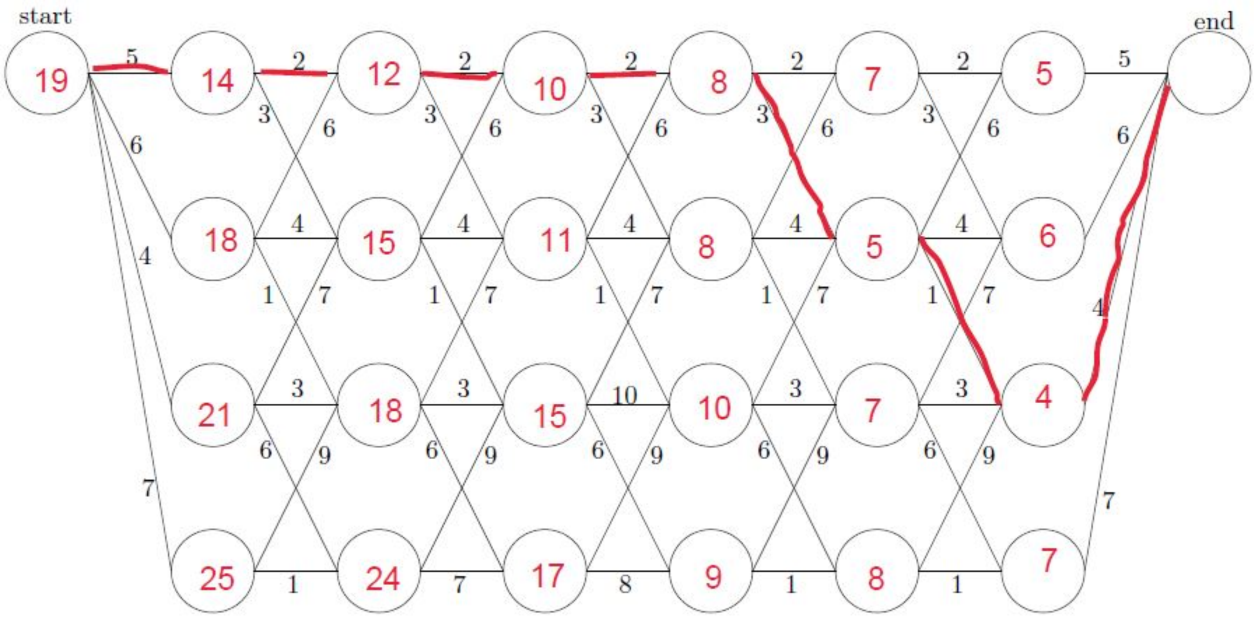
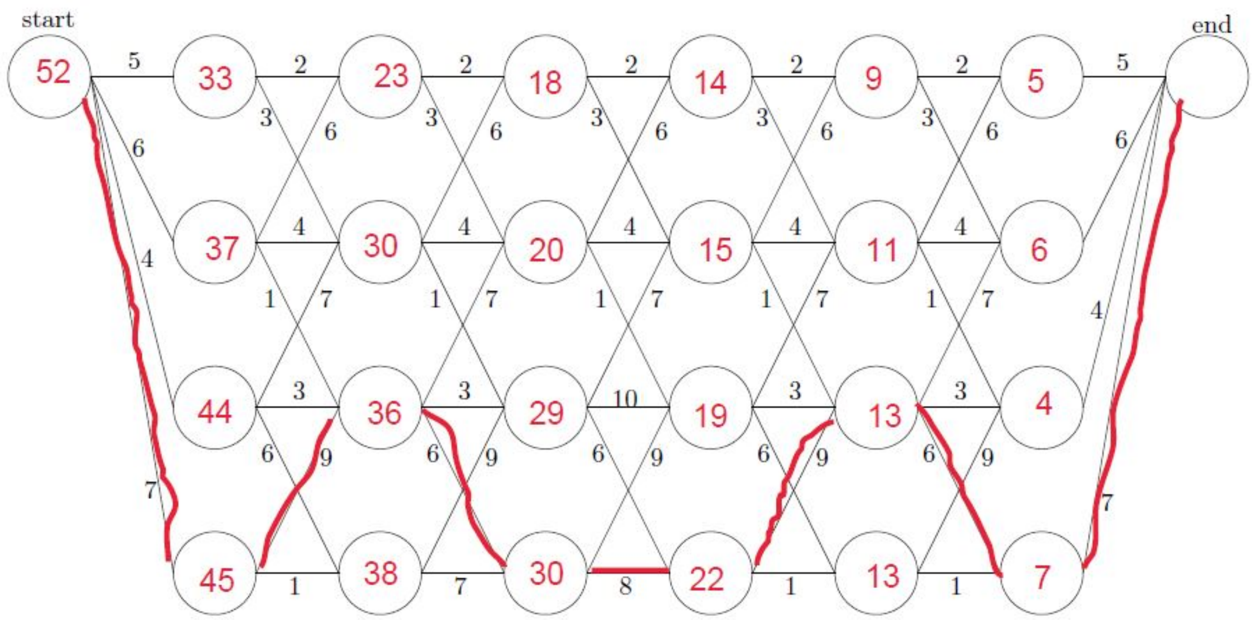
Homework 04

Problem 1





Problem 2

(a) With Bellman's expectation equation and following the policy π

$$v_\pi(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v_\pi(s')$$

For the two state $s = 0$ and $s = 1$

$$v_\pi(0) = R_0^{(1)} + \gamma(P_{00}^{(1)} v_\pi(0) + P_{01}^{(1)} v_\pi(1)) \quad (1)$$
$$v_\pi(1) = R_1^{(2)} + \gamma(P_{10}^{(2)} v_\pi(0) + P_{11}^{(2)} v_\pi(1)) \quad (2)$$

That is

$$v_\pi(0) = 1 + \frac{3}{4}(\frac{1}{3} v_\pi(0) + \frac{2}{3} v_\pi(1)) \quad (3)$$
$$v_\pi(1) = 2 + \frac{3}{4}(\frac{2}{3} v_\pi(0) + \frac{1}{3} v_\pi(1)) \quad (4)$$

$$4v_\pi(0) = 4 + (v_\pi(0) + 2v_\pi(1)) \quad (5)$$
$$4v_\pi(1) = 8 + (2v_\pi(0) + v_\pi(1)) \quad (6)$$

We can easily get

$$\begin{aligned} v_\pi(0) &= \frac{28}{5} \\ v_\pi(1) &= \frac{32}{5} \end{aligned} \quad (7)$$

(b) The initial values are set to zeros and the first 5 iteration values are

Step	$v_\pi(0)$	$v_\pi(1)$
1	1.000	2.000
2	2.250	3.000
3	3.062	3.875
4	3.703	4.500
5	4.176	4.977

With 50 iterations (code in p2b.py), the values converge to $v_\pi(0) = 5.600$ and $v_\pi(1) = 6.400$.

(c) The Bellman expectation equation for $q_\pi(s, a)$ is

$$q_\pi(s, a) = R_s^{(a)} + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')$$

$$\begin{aligned} q_\pi(0, 1) &= R_0^{(1)} + \gamma(P_{00}^{(1)} v_\pi(0) + P_{01}^{(1)} v_\pi(1)) \\ &= 1 + \frac{3}{4}(\frac{1}{3} v_\pi(0) + \frac{2}{3} v_\pi(1)) \\ &= \frac{28}{5} \end{aligned} \quad (8)$$

$$\begin{aligned} q_\pi(0, 2) &= R_0^{(2)} + \gamma(P_{00}^{(2)} v_\pi(0) + P_{01}^{(2)} v_\pi(1)) \\ &= 4 + \frac{3}{4}(\frac{1}{2} v_\pi(0) + \frac{1}{2} v_\pi(1)) \\ &= \frac{17}{2} \end{aligned} \quad (9)$$

$$\begin{aligned} q_\pi(1,1) &= R_1^{(1)} + \gamma(P_{10}^{(1)} v_\pi(0) + P_{11}^{(1)} v_\pi(1)) \\ &= 3 + \frac{3}{4}(\frac{1}{4} v_\pi(0) + \frac{3}{4} v_\pi(1)) \\ &= \frac{153}{20} \end{aligned} \tag{10}$$

$$\begin{aligned} q_\pi(1,2) &= R_1^{(2)} + \gamma(P_{10}^{(2)} v_\pi(0) + P_{11}^{(2)} v_\pi(1)) \\ &= 4 + \frac{3}{4}(\frac{1}{2} v_\pi(0) + \frac{1}{2} v_\pi(1)) \\ &= \frac{32}{5} \end{aligned} \tag{11}$$

(d) Since $q_\pi(0,1) < q_\pi(0,2)$ and $q_\pi(1,1) > q_\pi(1,2)$, a better policy $\pi^{'}$ chooses action 2 in state 0, and action 1 in state 1.

(e) From the Bellman's optimality equation

$$v_*(s) = \max_a [R_s^{(a)} + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a v_*(s^{'})]$$

$$\begin{aligned} v_*(0) &= \max_a \ [R_0^{(a)} + \gamma(P_{00}^{(a)} v_*(0) + P_{01}^{(a)} v_*(1))] \\ v_*(1) &= \max_a \ [R_1^{(a)} + \gamma(P_{10}^{(a)} v_*(0) + P_{11}^{(a)} v_*(1))] \end{aligned} \tag{12}$$

By setting the initial values to zeros, 50 iterations (code in p2e.py) give

$$\begin{aligned} v_*(0) &= 14.154 \\ v_*(1) &= 12.923 \end{aligned} \tag{13}$$

(f) With the values of $v_*(0)$ and $v_*(1)$, the optimal action-value function can be calculated with

$$q_*(s,a) = R_s^{(a)} + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a v_*(s^{'})$$

As given in p2e.py, the optimal action-values are

$$\begin{aligned} q_*(0,1) &= 11.00 \\ q_*(0,2) &= 14.15 \\ q_*(1,1) &= 12.92 \\ q_*(1,2) &= 12.31 \end{aligned} \tag{14}$$

Since $q_*(0,1) < q_*(0,2)$ and $q_*(1,1) > q_*(1,2)$, the optimal policy is: take action 2 when in state 0, and take action 1 when in state 1.

Problem 3

Model-free prediction

(a) Following the policy π, the action is determined by the current state: $S_i = 0, A_i = 1$ and $S_i = 1, A_i = 2$. The reward R_i is determined by the previous state: $S_i = 0, R_{i+1} = 1$ and $S_i = 1, R_{i+1} = 2$. I just assigned random zeros and ones to the state and make the starting state as zero. Thus the generated episode following policy π is (only the first ten listed, the entire episode in p3ab.py)

i	R_i	S_i	A_i
0	0	0	1
1	1	1	2
2	2	0	1
3	1	0	1
4	1	1	2
5	2	1	2
6	2	1	2
7	2	1	2
8	2	1	2
9	2	0	1

(b) With Monte Carlo policy evaluation, as shown in code p3ab.py, the value function is

$$\begin{aligned} v_\pi(0) &= 5.493 \\ v_\pi(1) &= 6.497 \end{aligned} \tag{15}$$

(c) The implementation of $n-$step temporal difference policy evaluation is in p3c.py. The estimated value function is

n	$v_\pi(0)$	$v_\pi(1)$
1	4.921	5.928
2	5.481	6.482
3	5.493	6.494
4	5.491	6.497
5	5.492	6.498

Model-free control

(a) With SARSA, the actions taken from current state and next state are both determined with $\epsilon-$greedy policy. The implementation is in p3_SARSA.py.

$$q_*(s,a) = \begin{bmatrix} 10.465 & 13.815 \\ 13.202 & 11.395 \end{bmatrix} \tag{16}$$

(b) With Q-learning, the action taken from the next state is with greedy policy. The implementation is in p3_QLearning.py.

$$q_*(s,a) = \begin{bmatrix} 11.608 & 14.702 \\ 13.422 & 12.301 \end{bmatrix} \tag{17}$$