EE526 Homework 5

Xingche Guo

December 8, 2019

Problem 1

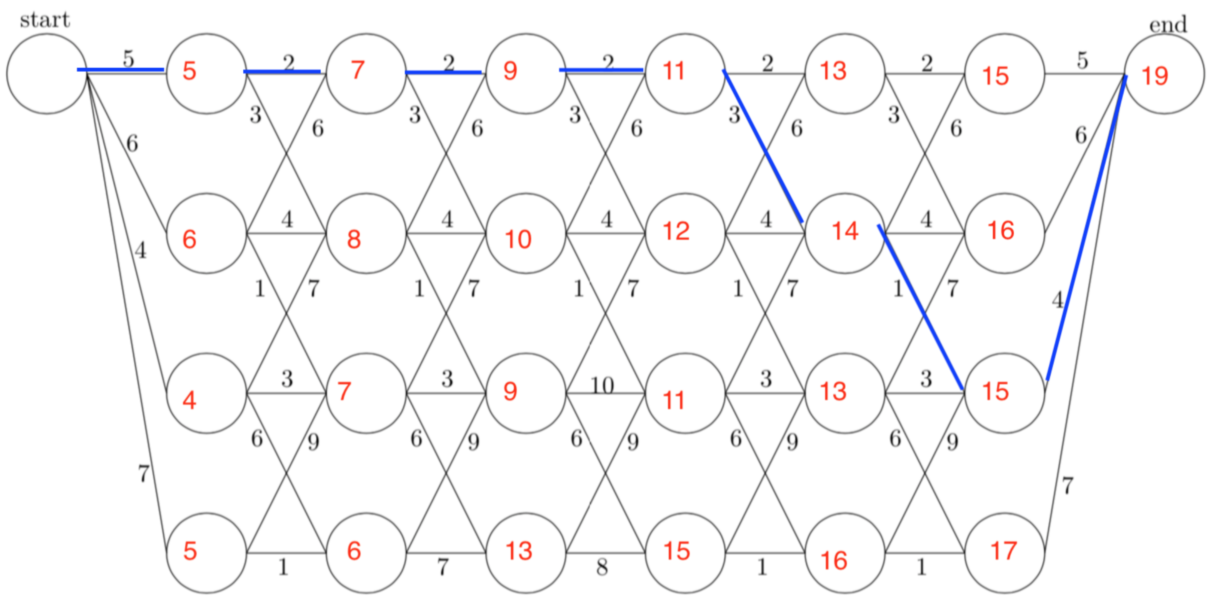

Figure 1: Shortest Path

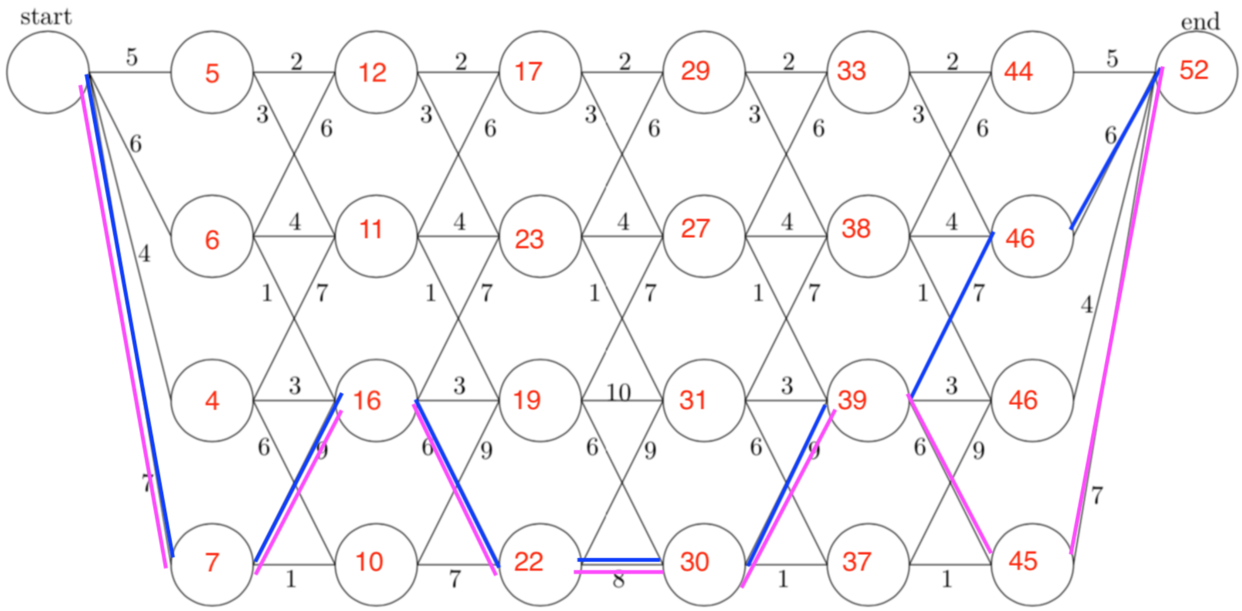

Figure 2: Longest Path

Problem 2

(a)

By Bellman's Expectation equation:

$$V_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a);$$
$$q_\pi(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^{(a)} V_\pi(s').$$

Therefore:

$$V_\pi(0) = q_\pi(0,1) = R_0^1 + \gamma(P_{00}^{(1)} V_\pi(0) + P_{01}^{(1)} V_\pi(1));$$
$$V_\pi(1) = q_\pi(1,2) = R_1^2 + \gamma(P_{10}^{(2)} V_\pi(0) + P_{11}^{(2)} V_\pi(1)).$$

Thus $(V_\pi(0), V_\pi(1)) = (5.6, 6.4)$.

(b)

$$V_\pi^{(t+1)}(0) \leftarrow R_0^1 + \gamma(P_{00}^{(1)} V_\pi^{(t)}(0) + P_{01}^{(1)} V_\pi^{(t)}(1));$$
$$V_\pi^{(t+1)}(1) \leftarrow R_1^2 + \gamma(P_{10}^{(2)} V_\pi^{(t)}(0) + P_{11}^{(2)} V_\pi^{(t)}(1)).$$

Let $(V_\pi^{(0)}(0), V_\pi^{(0)}(1)) = (0, 0)$, then:

$$
\begin{aligned}
(V_\pi^{(1)}(0), V_\pi^{(1)}(1)) &= (2.25, 3);\\
(V_\pi^{(2)}(0), V_\pi^{(2)}(1)) &= (3.0625, 3.875);\\
(V_\pi^{(3)}(0), V_\pi^{(3)}(1)) &= (3.703125, 4.5);\\
(V_\pi^{(4)}(0), V_\pi^{(4)}(1)) &= (4.175781, 4.976562);\\
(V_\pi^{(5)}(0), V_\pi^{(5)}(1)) &= (4.532227, 5.332031);\\
(V_\pi^{(100)}(0), V_\pi^{(100)}(1)) &= (5.6, 6.4).
\end{aligned}
$$

(c)

$$
\begin{pmatrix} q_\pi(0,1) \\ q_\pi(1,1) \\ q_\pi(0,2) \\ q_\pi(1,2) \end{pmatrix} = \begin{pmatrix} R_0^1 \\ R_q^1 \\ R_0^2 \\ R_1^2 \end{pmatrix} + \gamma \begin{pmatrix} P_{00}^{(1)} & 0 & 0 & P_{01}^{(1)} \\ P_{10}^{(1)} & 0 & 0 & P_{11}^{(1)} \\ P_{00}^{(2)} & 0 & 0 & P_{01}^{(2)} \\ P_{10}^{(2)} & 0 & 0 & P_{11}^{(2)} \end{pmatrix} \begin{pmatrix} R_0^1 \\ R_q^1 \\ R_0^2 \\ R_1^2 \end{pmatrix} \tag{1}
$$

Then $(q_\pi(0,1), q_\pi(1,1), q_\pi(0,2), q_\pi(1,2)) = (5.6, 7.65, 8.5, 6.4)$.

(d)

$$
\begin{aligned}
\because\ \pi^{'}(s) &= \arg\max_a q_\pi(s,a)\\
\therefore\ \pi^{'}(0) &= 2;\ \pi^{'}(1) = 1.
\end{aligned}
$$

(e)

By Bellman?s optimality equation:

$$
V_*(s) = \max_a \left(R_s^a + \gamma \sum_{s^{'}} P_{ss^{'}}^{(a)} V_*(s^{'}) \right),
$$

therefore let:

$$
\begin{aligned}
V_*^{(t+1)}(0) &\leftarrow \max_a \left(R_0^a + \gamma P_{00}^{(a)} V_*^{(t)}(0) + \gamma P_{01}^{(a)} V_*^{(t)}(1) \right);\\
V_*^{(t+1)}(1) &\leftarrow \max_a \left(R_1^a + \gamma P_{10}^{(a)} V_*^{(t)}(0) + \gamma P_{11}^{(a)} V_*^{(t)}(1) \right).
\end{aligned}
$$

therefore:

$$
(V_*^{(100)}(0), V_*^{(100)}(1)) = (14.154, 12.923).
$$

(f)

$$\pi_*(0) = \arg\max_a \left(R_0^a + \gamma P_{00}^{(a)} V_*^{(t)}(0) + \gamma P_{01}^{(a)} V_*^{(t)}(1) \right) = 2;$$
$$\pi_*(1) = \arg\max_a \left(R_1^a + \gamma P_{10}^{(a)} V_*^{(t)}(0) + \gamma P_{11}^{(a)} V_*^{(t)}(1) \right) = 1.$$

Problem 3

(a), (b), (c)

The estimated value function $v_\pi(s)$ based on Monte Carlo policy evaluation method is:

$$(v_\pi(0), v_\pi(1)) = (5.59, 6.40).$$

The estimated value function $v_\pi(s)$ based on 5-step temporal difference policy evaluation method is:

$$(v_\pi(0), v_\pi(1)) = (5.59, 6.39).$$

(d), (e)

For learning rate $\alpha = 0.1$, exploration probability $\epsilon_t = t^{-1/2}$, number of time $N = 50000$:
The optimal action-value function $q_*(s, a)$ estimated based on SARSA algorithm is:

$$(q_*(0,1), q_*(0,2), q_*(1,1), q_*(1,2)) = (10.73, 14.23, 12.88, 12.29).$$

The optimal action-value function $q_*(s, a)$ estimated based on Q-learning algorithm is:

$$(q_*(0,1), q_*(0,2), q_*(1,1), q_*(1,2)) = (11.06, 14.28, 12.65, 12.20).$$

Both of the methods indicate that:

$$\begin{cases} \pi_*(0) = 2; \\ \pi_*(1) = 1. \end{cases}$$