Iowa State University

Department of Electrical and Computer Engineering

Deep Machine Learning: Theory and Practice

EE 526X

Homework 5

Author:
Vishal Deep

Instructor:
Dr. Zhengdao Wang

December 7, 2019

Iowa State University

1 Problem 1

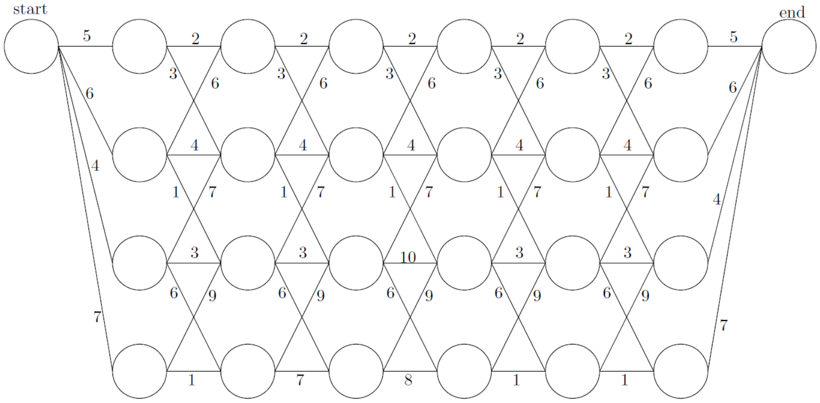

Figure 1: Network Graph

2 Problem 2

States: S = 0, 1
Actions: A = 1, 2
Rewards:

$$R_S^{(a)} = \begin{cases} 1 & (s,a) = (0,1) \\ 4 & (s,a) = (0,2) \\ 3 & (s,a) = (1,1) \\ 2 & (s,a) = (1,2) \end{cases}$$

Transition Probabilities:

$$\begin{bmatrix} P_{00}^{(1)} & P_{00}^{(2)} \\ P_{10}^{(1)} & P_{10}^{(2)} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{2} \\ \frac{1}{4} & \frac{2}{3} \end{bmatrix}$$

Discount factor: $\gamma = \frac{3}{4}$

The Bellman's expectation equation is given by

$$V_\pi(S) = R_S + \gamma \sum_{s' \epsilon S} P_{SS'} V_\pi(S')$$

2(a):

choosing action 1 in state 0, and action 2 in state 1,

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

Substituting the values,

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

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

Solving these 2 equations, we get

$$V_\pi(0) = \frac{28}{5}$$

$$V_\pi(1) = \frac{32}{5}$$

2(b):

```
vpi_0 =0
vpi_1 =0
gamma =3/4

for i in range(0, 5):
    vpi_0_temp =1 +3/4 *( 1/3 *vpi_0 +2/3 *vpi_1)
    vpi_1 =2 +3/4 *( 2/3 *vpi_0 +1/3 *vpi_1)
    vpi_0 =vpi_0_temp
    print(f"Iteration = {i}, vpi_0 = {vpi_0}, vpi_1 = {vpi_1}")
```

```
Iteration = 0, vpi__0 = 1.0, vpi__1 = 2.0
Iteration = 1, vpi__0 = 2.25, vpi__1 = 3.0
Iteration = 2, vpi__0 = 3.0625, vpi__1 = 3.875
Iteration = 3, vpi__0 = 3.7031, vpi__1 = 4.5
Iteration = 4, vpi__0 = 4.17578125, vpi__1 = 4.9765625
```

2(c):

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

$$q_\pi(s, a) = R_S^{(a)} + \gamma \sum_{s' \epsilon S} P_{SS'}^{(a)} V_\pi(S')$$

$$q_\pi(0, 1) = R_0^{(1)} + \gamma(P_{00}^{(1)} V_\pi(0) + P_{01}^{(1)} V_\pi(1))$$

$$q_\pi(0, 1) = 1 + \frac{3}{4}(\frac{1}{3}V_\pi(0) + \frac{2}{3}V_\pi(1)) = \frac{28}{5}$$

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

$$q_\pi(0, 2) = 4 + \frac{3}{4}(\frac{1}{2}V_\pi(0) + \frac{1}{2}V_\pi(1)) = \frac{17}{2}$$

$$q_\pi(1, 1) = R_1^{(1)} + \gamma(P_{10}^{(1)} V_\pi(0) + P_{11}^{(1)} V_\pi(1))$$

$$q_\pi(1, 1) = 3 + \frac{3}{4}(\frac{1}{4}V_\pi(0) + \frac{3}{4}V_\pi(1)) = \frac{153}{20}$$

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

$$q_\pi(1, 2) = 4 + \frac{3}{4}(\frac{1}{2}V_\pi(0) + \frac{1}{2}V_\pi(1)) = \frac{32}{5}$$

2(d)

From the part c,

$$State0 : q_\pi(0, 1) < q_\pi(0, 2)$$

$$State1 : q_\pi(1, 1) > q_\pi(1, 2)$$

Improved policy would be to choose action 2 in state 0 and action 1 in state 1.

2(e)

```
import numpy as np

def q_pi(s, a, V0, V1):
    gamma =3/4

    if s==0 and a==1:
        R =1
        P0 =1/3
        P1 =1 -P0
    elif s==0 and a==2:
        R =4
        P0 =1/2
        P1 =1- P0
    elif s==1 and a==1:
        R =3
        P0 =1/4
        P1 =1 -P0
    elif s==1 and a==2:
        R =2
        P0 =2/3
        P1 =1- P0

    v_star =R +gamma *(P0*V0 +P1*V1)
    return v_star

# Intialize to zero
V0 =0
V1 =0

for i in range(0, 10):
    V0_temp =np.maximum(q_pi(0, 1, V0, V1), q_pi(0, 2, V0, V1))
    V1 =np.maximum(q_pi(1, 1, V0, V1), q_pi(1, 2, V0, V1))
    V0 =V0_temp
    print(f"Iteration: {i}, V0: {V0:.4f}, V1: {V1:.4f}")

q01 =q_pi(0, 1, V0, V1)
q02 =q_pi(0, 2, V0, V1)
q11 =q_pi(1, 1, V0, V1)
q12 =q_pi(1, 2, V0, V1)

print(f"q(0,1) = {q01:.4f}")
print(f"q(0,2) = {q02:.4f}")
print(f"q(1,1) = {q11:.4f}")
print(f"q(1,2) = {q12:.4f}")
```

```
Iteration: 0, V0: 4.0000, V1: 3.0000
Iteration: 1, V0: 6.6250, V1: 5.4375
Iteration: 2, V0: 8.5234, V1: 7.3008
Iteration: 3, V0: 9.9341, V1: 8.7048
Iteration: 4, V0: 10.9896, V1: 9.7591
Iteration: 5, V0: 11.7808, V1: 10.5500
Iteration: 6, V0: 12.3741, V1: 11.1433
Iteration: 7, V0: 12.8190, V1: 11.5882
Iteration: 8, V0: 13.1527, V1: 11.9219
Iteration: 9, V0: 13.4030, V1: 12.1722
q(0,1) = 10.4369
q(0,2) = 13.5907
q(1,1) = 12.3599
q(1,2) = 11.7446
```

2(f)

As calculated from the code above, Optimal Policy is

$$q(0,1) = 10.4369$$

$$q(0,2) = 13.5907$$

$$q(1,1) = 12.3599$$

$$q(1,2) = 11.7446$$

Problem 3