RL Assignment2

shagun16088

September 2019

1 Question 2

In order to solve the grid problem, we solve the linear system of equations:

$$Ax = b \tag{1}$$

where the grid size is 5x5, thereby giving a total of 25 states and an action space of 4 possible states. We model this problem as a linear system shown above where A (25x25 matrix) depicts the transition from each state to every other possible state and b is the 25x1 vector depicting the reward. The solution x is the required value function i.e. :

```
Value [ 3.30899634  8.78929186  4.42761918  5.32236759  1.49217876  1.52158807
  2.99231786  2.25013995  1.9075717   0.54740271  0.05082249  0.73817059
  0.67311326  0.35818621 -0.40314114 -0.9735923  -0.43549543 -0.35488227
 -0.58560509 -1.18307508 -1.85770055 -1.34523126 -1.22926726 -1.42291815
 -1.97517905]
```

Figure 1: Value Function

2 Question 4

To obtain the state-action value function, we solve the linear inequality:

$$Ax >= b \tag{2}$$

Here, we have the action space of 4 possible actions and a total of 25 states. Therefore, A as a 100x25 matrix since for every state we have a transition to every other state for each possible action. b is the 100x1 vector and x is 25x1 for the 25 states in the grid.

By solving the above non-linear system, we obtain the state-action value function and the policy as follows:

```
Value [[21.98 24.42 21.98 19.42 17.48]
 [19.78 21.98 19.78 17.8  16.02]
 [17.8  19.78 17.8  16.02 14.42]
 [16.02 17.8  16.02 14.42 12.98]
 [14.42 16.02 14.42 12.98 11.68]]
Policy
['right '] ['up ', 'down ', 'left ', 'right '] ['left '] ['up ', 'down ', 'left ', 'right '] ['left ']
['up ', 'right '] ['up '] ['up ', 'left '] ['left '] ['left ']
['up ', 'right '] ['up '] ['up ', 'left '] ['up ', 'left '] ['up ', 'left ']
['up ', 'right '] ['up '] ['up ', 'left '] ['up ', 'left '] ['up ', 'left ']
['up ', 'right '] ['up '] ['up ', 'left '] ['up ', 'left '] ['up ', 'left ']
```

Figure 2: Value Function and Policy

3 Question 6

3.1 Policy Iteration

The policy and the state value function obtained after policy iteration method is shown below:

```
Value [[ 0. -1. -2. -3.]
 [-1. -2. -3. -2.]
 [-2. -3. -2. -1.]
 [-3. -2. -1.  0.]]
Policy
['---- ', '---- '] ['left '] ['left '] ['down ', 'left ']
['up '] ['up ', 'left '] ['up ', 'down ', 'left ', 'right '] ['down ']
['up '] ['up ', 'down ', 'left ', 'right '] ['down ', 'right '] ['down ']
['up ', 'right '] ['right '] ['right '] ['---- ', '---- ']
```

The per iteration difference in value function is as follows:

```
Iteration:  0
Difference between previous and current value function is :  5.766122522572036

Iteration:  10
Difference between previous and current value function is :  2.446230466082198

Iteration:  20
Difference between previous and current value function is :  1.0196738914812349

Iteration:  30
Difference between previous and current value function is :  0.424935441320615

Iteration:  40
Difference between previous and current value function is :  0.17708605754237206

Iteration:  50
Difference between previous and current value function is :  0.0737982024700002

Iteration:  60
Difference between previous and current value function is :  0.030754395706637752

Iteration:  70
Difference between previous and current value function is :  0.012816475518696007

Iteration:  80
Difference between previous and current value function is :  0.005341091604862473

Iteration:  90
Difference between previous and current value function is :  0.0022258271776756725

Iteration:  100
Difference between previous and current value function is :  0.0009275831592852963

Iteration:  110
Difference between previous and current value function is :  0.000386557647431135
```

3.2 Value Iteration

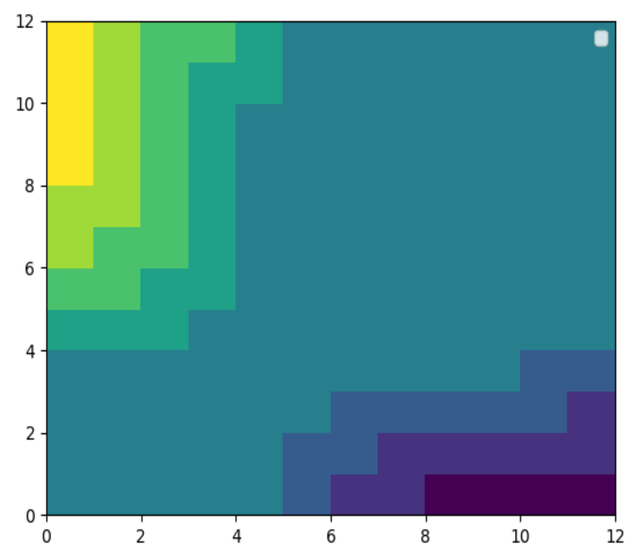
The policy and the state value function obtained after value iteration method and the per iteration difference in value function is shown below:

```
Iteration:  0
Difference between previous and current value function is :  3.7416573867739413
Iteration:  1
Difference between previous and current value function is :  3.1622776601683795
Iteration:  2
Difference between previous and current value function is :  2.0
Iteration:  3
Difference between previous and current value function is :  0.0
Value [[ 0. -1. -2. -3.]
 [-1. -2. -3. -2.]
 [-2. -3. -2. -1.]
 [-3. -2. -1.  0.]]
Policy

['----', '----'] ['left '] ['left '] ['down ', 'left ']
['up '] ['up ', 'left '] ['up ', 'down ', 'left ', 'right '] ['down ']
['up '] ['up ', 'down ', 'left ', 'right '] ['down ', 'right '] ['down ']
['up ', 'right '] ['right '] ['right '] ['----', '----']
```

4 Question 7

The optimal policy plot for original question as highlighted in the textbook is as follows:



For the modified problem as highlighted in Exercise 4.7, the optimal policy plot and the 3D vlaue plot are as follows: