Lab Assignment 3A

Artificial Intelligence

Name: Prakhar Gupta and Vibesh Kumar
Roll No: 2103126 and 2103140
Instructor: Dr. Divya Padmanabhan
Date: September 5, 2023

Q1.)
Ans1. Our policy evaluation works on the principal of convergence. We could have create n linear equation with n unknowns with linear algebra but due to high time complexity we have used convergence method. So we are starting with 0 as utility of all states and calculating new utility based on previous utility using $U^{pi}_{i+1}(s)$ = Expectation[gamma*$U^{pi}_{i}(s)$]. We will do till norm converges.

Q2.)
Ans2.Value Iteration works principal of starting with assumed utility vector (utility of each state). Then we try to calculate new utility based on values of old utility (assumed utility for 1^{st} iteration). New utility is calculated by considering all the possible actions for each state and then considering the max expected utility for those actions. So what we get is which action should be taken to get best utility.
We do this over and over updating old and new utility vector till convergence happens.

Q3.)
Ans3. Policy Iterations starts with using any policy. We calculates its utility and then we start with any state and check if we have a better option than current policy for this state. We decide better by calculating expected utility of each action. If we found any better action than we replace our current policy for this state with better option and move to next state.
We do this over and over till our new policy(at end of loop) is same as old policy (at start of loop) meaning that over 1 iteration our policy has not changed.

Output of Q1, Q2, and Q3:

```
State Policy  Utility

Policy 1
0     -1      0
1     1      0.53764
2     2      1.49344
3     3      2.88233
4     4      4.14846
5     5      7.2
6     4      8.00646
7     3      9.44017
8     2      11.5235
9     1      13.4227
10     -1      20

Policy 2
0      -1      0
1      1      0.0205948
2      1      0.0572078
3      1      0.128018
4      1      0.269795
5      1      0.557403
6      1      1.14365
7      1      2.3407
8      1      4.78647
9      1      9.78469
10      -1      20

Optimal Policy (Value Iteration)
0      0     0
1      1     0.53764
2      2     1.49344
3      2     2.88233
4      1     4.14846
5      5     7.2
6      4     8.00646
7      3     9.44017
8      2     11.5235
9      1     13.4227
10      0     20

Optimal Policy (Policy Iteration)
0     0      0
1     1      0.53764
2      2      1.49344
3      2      2.88233
4      1      4.14846
5      5      7.2
6      4      8.00646
7      3      9.44017
8      2      11.5235
9      1      13.4227
10      0      20
```

Experiment:
We modified our original code to incorporate -1 reward for each transition except to N.
We can compare the result here: Left is without -1 reward (Original Code) ; Right is with -1 reward.

```
State Policy  Utility

Policy 1
0     -1      0
1     1       0.53764
2     2       1.49344
3     3       2.88233
4     4       4.14846
5     5       7.2
6     4       8.00646
7     3       9.44017
8     2       11.5235
9     1       13.4227
10     -1      20

Policy 2
0      -1      0
1      1       0.0205948
2      1       0.0572078
3      1       0.128018
4      1       0.269795
5      1       0.557403
6      1       1.14365
7      1       2.3407
8      1       4.78647
9      1       9.78469
10      -1      20

Optimal Policy (Value Iteration)
0      0     0
1      1     0.53764
2      2     1.49344
3      2     2.88233
4      1     4.14846
5      5     7.2
6      4     8.00646
7      3     9.44017
8      2     11.5235
9      1     13.4227
10      0     20
```

```
State Policy  Utility

Policy 1
0     -1      0
1     1       -1.04586
2     2       -0.127394
3     3       1.20723
4     4       2.42391
5     5       6.2
6     4       6.13121
7     3       7.50891
8     2       9.51085
9     1       11.3359
10     -1      20

Policy 2
0      -1      0
1      1       -2.63059
2      1       -4.52941
3      1       -5.85804
4      1       -6.70045
5      1       -7.04751
6      1       -6.74796
7      1       -5.3953
8      1       -2.08721
9      1       5.0729
10      -1      20

Optimal Policy (Value Iteration)
0      0     0
1      1     -1.04586
2      2     -0.127394
3      3     1.20723
4      4     2.42391
5      5     6.2
6      4     6.13121
7      3     7.50891
8      2     9.51085
9      1     11.3359
10      0     20
```

Here we can observe various trend that compared to no reward, -1 reward reduced utility of each state for every policy.
Slower we approach N, more negative utility we receive. Policy 2 gives a glimpse of this trend. Policy 2 also gives more negative utility as we reach closer to N from 0. Like utility(5) < utility(2). This is because the our agent gets -1 for each transistion and till cant quit the game. It state1 can easily quit game by going reaching 0 whereas state5 would have to get multiple times negative reward to 0 and quit game. Although it can also reach N state but due to Probability of getting head is low therefore utility are as above.

Experiment:
N = 100
We have attached an output txt file for N = 100 for different policy including optimal policy;