Assignment 2

Team number: 6
Rishabh Singhal: 20171213
Pratyush Kumar: 20171168

PART B

1. b) 0.99 , -X/10

The value iteration algorithm is run for this input state, and the output after each iteration is shown below.

Iteration #0

0.6	0.0	4.152	6.0
-0.125	-0.6	2.629	4.412
-0.6	-0.659	0.0	2.895
-0.6	-0.778	-1.2	1.574

Iteration #1

0.6	0.0	4.823	6.0
-0.197	1.357	3.791	4.964
-0.88	0.323	0.0	3.905
-1.212	-0.583	-1.2	2.53

Iteration #2

0.6	0.0	5.005	6.0
0.447	2.569	4.202	5.059
-0.301	1.437	0.0	4.18
-1.016	0.319	-1.2	2.842

Iteration #3

0.6	0.0	5.064	6.0
1.464	3.125	4.324	5.081
0.672	2.084	0.0	4.252
-0.137	0.918	-1.2	2.93

Iteration #4

0.6	0.0	5.081	6.0
2.001	3.341	4.355	5.086
1.258	2.377	0.0	4.27
0.473	1.21	-1.2	2.953

Iteration #5

0.6	0.0	5.086	6.0
2.23	3.415	4.363	5.087

1.55 2.494 0.0 4.275
0.794 1.335 -1.2 2.959

Iteration #6
0.6 0.0 5.087 6.0
2.318 3.44 4.365 5.088
1.683 2.538 0.0 4.276
0.944 1.385 -1.2 2.961

Iteration #7
0.6 0.0 5.088 6.0
2.351 3.449 4.365 5.088
1.737 2.555 0.0 4.276
1.006 1.404 -1.2 2.961

Iteration #8
0.6 0.0 5.088 6.0
2.363 3.452 4.365 5.088
1.757 2.561 0.0 4.276
1.03 1.411 -1.2 2.961

Iteration #9
0.6 0.0 5.088 6.0
2.367 3.453 4.365 5.088
1.764 2.563 0.0 4.276
1.039 1.414 -1.2 2.961

Policy :::

- - E -
E E N/E N
E N - N
N N - N

Since the discount factor ≈ 1 , there is not much decay in the rewards of the farther states. As such, the policy always tends to direct us to the goal state at (0,3), as it has a very high value.

1. a) 0.1 , -X/10

Iteration #0

0.6 0.0 -0.12 6.0
-0.552 -0.6 -0.601 -0.126
-0.6 -0.606 0.0 -0.6
-0.6 -0.618 -1.2 -0.606

Iteration #1
0.6 0.0 -0.127 6.0
-0.564 -0.657 -0.617 -0.127

-0.657 -0.661 0.0 -0.622
-0.66 -0.666 -1.2 -0.661

_____

Iteration #2

0.6 0.0 -0.127 6.0
-0.564 -0.658 -0.618 -0.127
-0.658 -0.666 0.0 -0.623
-0.666 -0.667 -1.2 -0.666
Policy ::
- - E -
N W N/E N
N N/W - N
N N/W E N

The algorithm converges rapidly, and the final utilities are all mostly negative. This is because our discount factor is much less, and as such, the farther rewards are given much less favour. So, the immediate negative step cost dominates the utility values.

2. a) 0.99 , X

The value iteration algorithm is run for this input state, and the output after each iteration is shown below.

Only 60, 61 state is shown as the number of iterations is large

Iteration #60

_____

0.6 0.0 398.94 6.0
416.597 418.327 417.817 415.189
418.563 420.322 0.0 416.6
420.15 421.778 -1.2 401.693

Iteration #61

0.6 0.0 402.82 6.0
420.159 421.855 421.352 418.765
422.086 423.812 0.0 420.148
423.643 425.239 -1.2 405.503

Policy ::

- - S -
S S E E
E S - E
E N W N

Since the step cost is very large, larger than the reward of the best goal state, the utilities are skewed and the policy sends us in circles. The cell at (1,3) has the highest utility, so it assumes that we cannot move from there.

2. b) 0.99 , -X/5

The value iteration algorithm is run for this input state, and the output after each iteration is shown below.

Iteration #8

```
 0.6     0.0     4.259    6.0
-0.413   1.141   2.88     4.259
-1.753   -0.522   0.0     2.709
-3.099   -2.039  -1.2     0.918
```

Iteration #9

```
 0.6     0.0     4.259    6.0
-0.41    1.142   2.88     4.259
-1.75    -0.521   0.0     2.709
-3.095   -2.037  -1.2     0.918
```

Policy ::

```
-  -  E  -
E  E  N/E N
N  N  -   N
B  N  E   N
```

Here, the negative factor of the step cost initially outweighs all positive benefits, as they occur much farther away. However, as we approach the goal state at (0,3), the utilities become larger as it has a very high reward. So, the policy directs us towards that state.

2. c) 0.99 , -X/4

The value iteration algorithm is run for this input state, and the output after each iteration is shown below.

Iteration #10

```
 0.6     0.0     3.844    6.0
-1.138   -0.005  2.137    3.844
-2.883   -1.986   0.0     1.926
-4.522   -2.938  -1.2     -0.104
```

Iteration #11

```
 0.6     0.0     3.844    6.0
-1.138   -0.005  2.137    3.844
-2.883   -1.986   0.0     1.926
-4.522   -2.938  -1.2     -0.104
```

Policy ::
- - E -
N E N/E N
N N - N
N E E N

This case is also almost exact same as 2 b.) expect for the fact that, since the step cost is higher, the negative values are present till farther along. As such, for example at (0,1), the policy directs it to the north as opposed to the east as in the previous case.

2. d) 0.99 , -X

The value iteration algorithm is run for this input state, and the output after each iteration is shown below.

Iteration #4

0.6 0.0 -2.368 6.0
-7.775 -15.281 -8.997 -2.373
-15.281 -16.68 0.0 -9.816
-16.679 -9.543 -1.2 -8.792

Iteration #5

0.6 0.0 -2.373 6.0
-7.807 -15.348 -9.005 -2.374
-15.348 -16.729 0.0 -9.824
-16.729 -9.551 -1.2 -8.793

Policy ::
- - E -
N W N N
N S - N
E E E W

Here, the unit step cost is extremely high, and as such, even the negative termination state is a valid place for the agent. So, the agent directs itself towards it's closest termination state, as the step cost is high for it to even consider the farther states.

Part C.) The linear program for the given MDP was constructed and solved using Excel, and it was found that the answer exactly matched the answer obtained using value iteration with gamma = 1.