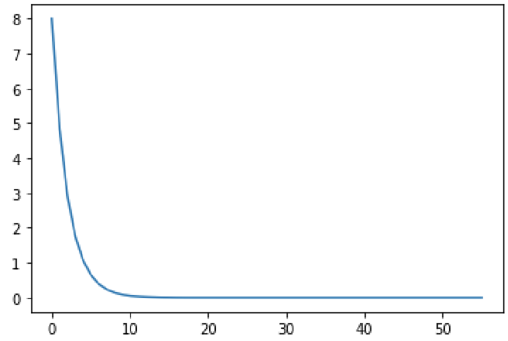
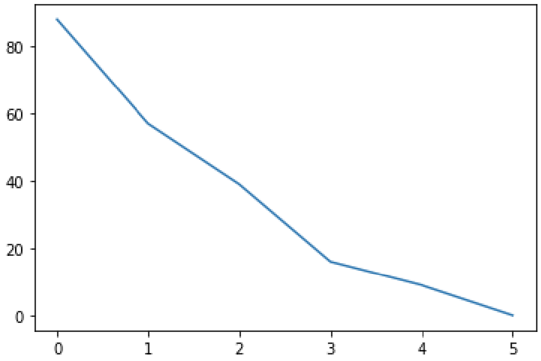
Question 5

1) I have implemented the code in q5_goal1.ipynb and q5_goal2.ipynb (they are the same just with different Goals and some minor changes).

2) a) Plot $max_s |J_{i+1}(s) - J_i(s)|$ $vs\ iterations$

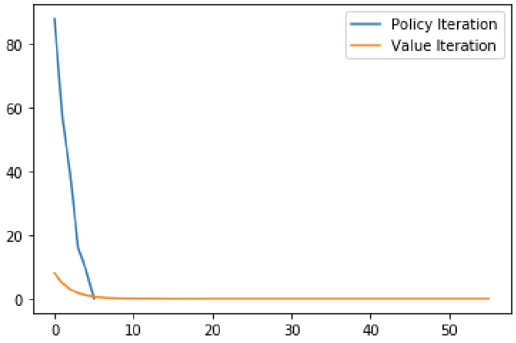


Plot for $\sum_s \pi_{i+1}(s)\ != \pi_i(s)$ vs iterations

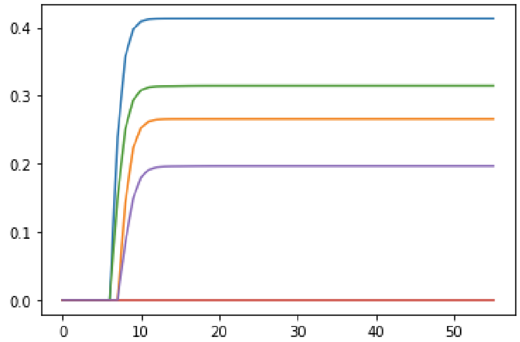


Here in the policy iteration, there are very few transitions thus we are unable to see the exponential true exponential nature of J in policy iterations.

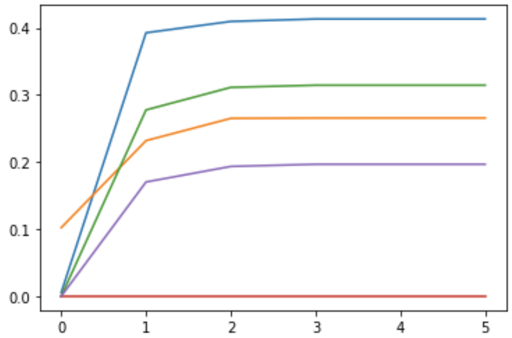
Both on same plot -



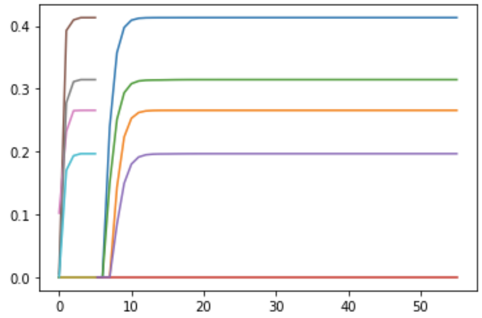
b) For value iteration -



For policy iteration -



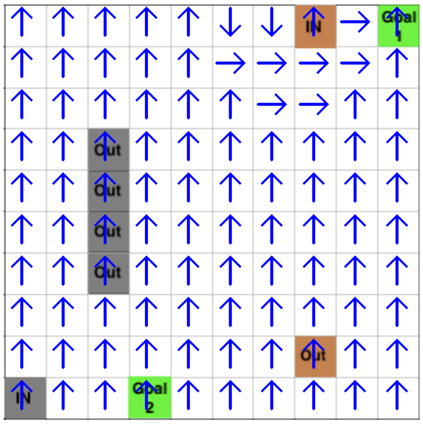
Both on same plot -



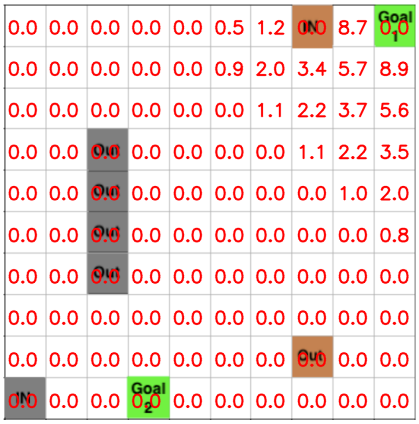
Notice that policy iteration stabilizes in ~5 steps whereas value iteration takes lot more steps to converge.

c) Policy and J for Value Iteration (Greedy Policy) -

After 5 iterations-

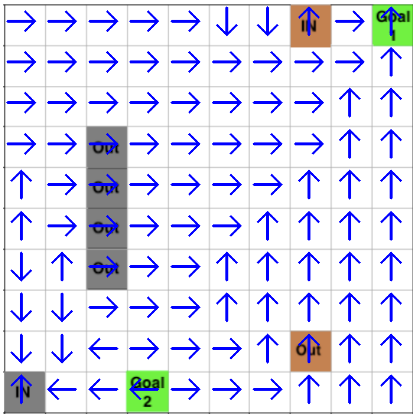


↑	↑	↑	↑	↑	↓	↓	IN ↑	→	Goal 1 ↑
↑	↑	↑	↑	↑	→	→	→	→	↑
↑	↑	↑	↑	↑	↑	→	→	↑	↑
↑	↑	Out ↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	Out ↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	Out ↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	Out ↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	↑	↑	↑	↑	↑	Out ↑	↑	↑
IN ↑	↑	↑	Goal 2 ↑	↑	↑	↑	↑	↑	↑

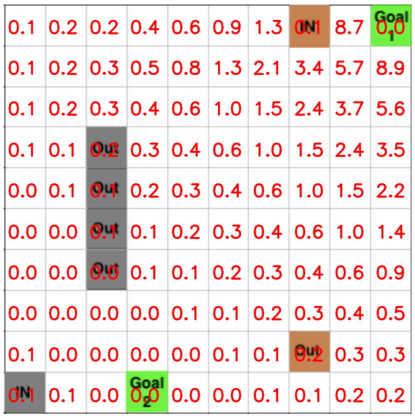


0.0	0.0	0.0	0.0	0.0	0.5	1.2	IN 0.0	8.7	Goal 1 0.0
0.0	0.0	0.0	0.0	0.0	0.9	2.0	3.4	5.7	8.9
0.0	0.0	0.0	0.0	0.0	0.0	1.1	2.2	3.7	5.6
0.0	0.0	Out 0.0	0.0	0.0	0.0	0.0	1.1	2.2	3.5
0.0	0.0	Out 0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0
0.0	0.0	Out 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8
0.0	0.0	Out 0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	Out 0.0	0.0	0.0
IN 0.0	0.0	0.0	Goal 2 0.0	0.0	0.0	0.0	0.0	0.0	0.0

After Stopping -



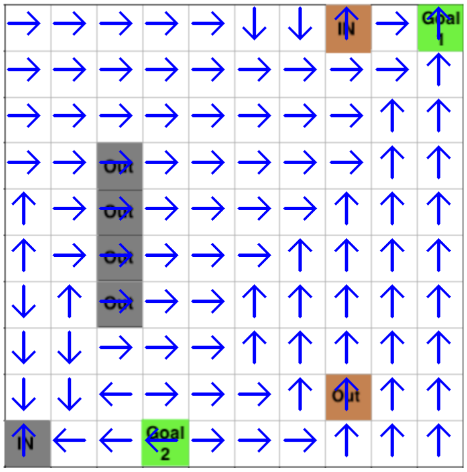
→	→	→	→	→	↓	↓	IN ↑	→	Goal 1 ↑
→	→	→	→	→	→	→	→	→	↑
→	→	→	→	→	→	→	→	↑	↑
→	→	Out →	→	→	→	→	→	↑	↑
↑	→	Out →	→	→	→	→	↑	↑	↑
↑	→	Out →	→	→	→	↑	↑	↑	↑
↓	↑	Out →	→	→	↑	↑	↑	↑	↑
↓	↓	→	→	→	↑	↑	↑	↑	↑
↓	↓	←	→	→	→	↑	Out ↑	↑	↑
IN ↑	←	←	Goal 2 ←	→	→	→	↑	↑	↑



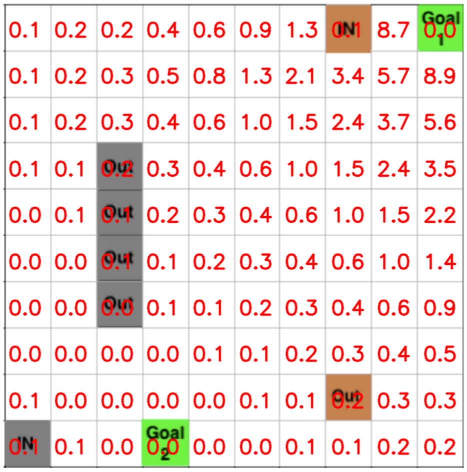
0.1	0.2	0.2	0.4	0.6	0.9	1.3	IN 0.1	8.7	Goal 1 0.0
0.1	0.2	0.3	0.5	0.8	1.3	2.1	3.4	5.7	8.9
0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.7	5.6
0.1	0.1	Out 0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.5
0.0	0.1	Out 0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.2
0.0	0.0	Out 0.1	0.1	0.2	0.3	0.4	0.6	1.0	1.4
0.0	0.0	Out 0.0	0.1	0.1	0.2	0.3	0.4	0.6	0.9
0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5
0.1	0.0	0.0	0.0	0.0	0.1	0.1	Out 0.2	0.3	0.3
IN 0.1	0.1	0.0	Goal 2 0.0	0.0	0.0	0.1	0.1	0.2	0.2

d) Policy and J for policy iteration

After 5 iterations-

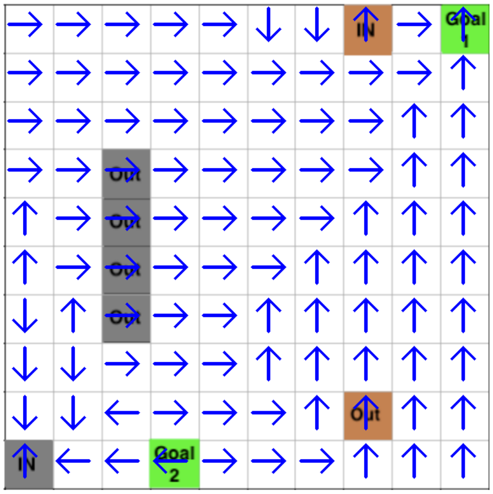


→	→	→	→	→	↓	↓	IN ↑	→	Goal 1 ↑
→	→	→	→	→	→	→	→	→	↑
→	→	→	→	→	→	→	→	↑	↑
→	→	Out →	→	→	→	→	→	↑	↑
↑	→	Out →	→	→	→	→	↑	↑	↑
↑	→	Out →	→	→	→	↑	↑	↑	↑
↓	↑	Out →	→	→	↑	↑	↑	↑	↑
↓	↓	→	→	→	↑	↑	↑	↑	↑
↓	↓	←	→	→	→	↑	Out ↑	↑	↑
IN ↑	←	←	Goal 2 ←	→	→	→	↑	↑	↑

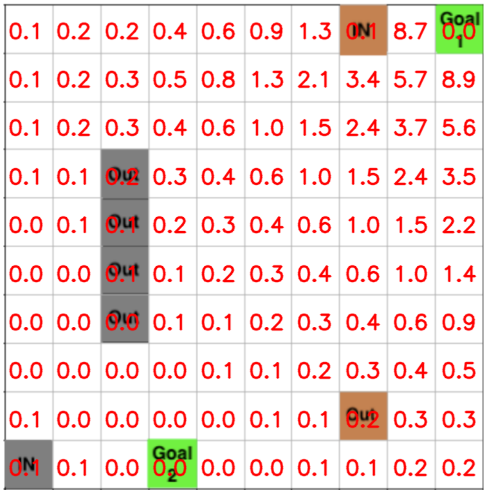


0.1	0.2	0.2	0.4	0.6	0.9	1.3	IN 0.1	8.7	Goal 1 0.0
0.1	0.2	0.3	0.5	0.8	1.3	2.1	3.4	5.7	8.9
0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.7	5.6
0.1	0.1	Out 0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.5
0.0	0.1	Out 0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.2
0.0	0.0	Out 0.1	0.1	0.2	0.3	0.4	0.6	1.0	1.4
0.0	0.0	Out 0.0	0.1	0.1	0.2	0.3	0.4	0.6	0.9
0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5
0.1	0.0	0.0	0.0	0.0	0.1	0.1	Out 0.2	0.3	0.3
IN 0.1	0.1	0.0	Goal 2 0.0	0.0	0.0	0.1	0.1	0.2	0.2

After Stopping-



→	→	→	→	→	↓	↓	↑ IN	→	↑ Goal 1
→	→	→	→	→	→	→	→	→	↑
→	→	→	→	→	→	→	→	↑	↑
→	→	→ Out	→	→	→	→	→	↑	↑
↑	→	→ Out	→	→	→	→	↑	↑	↑
↑	→	→ Out	→	→	→	↑	↑	↑	↑
↓	↑	→ Out	→	→	↑	↑	↑	↑	↑
↓	↓	→	→	→	↑	↑	↑	↑	↑
↓	↓	←	→	→	→	↑	↑ Out	↑	↑
↑ IN	←	←	← Goal 2	→	→	→	↑	↑	↑



0.1	0.2	0.2	0.4	0.6	0.9	1.3	IN 0.1	8.7	Goal 1 0.0
0.1	0.2	0.3	0.5	0.8	1.3	2.1	3.4	5.7	8.9
0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.7	5.6
0.1	0.1	Out 0.2	0.3	0.4	0.6	1.0	1.5	2.4	3.5
0.0	0.1	Out 0.1	0.2	0.3	0.4	0.6	1.0	1.5	2.2
0.0	0.0	Out 0.1	0.1	0.2	0.3	0.4	0.6	1.0	1.4
0.0	0.0	Out 0.0	0.1	0.1	0.2	0.3	0.4	0.6	0.9
0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.3	0.4	0.5
0.1	0.0	0.0	0.0	0.0	0.1	0.1	Out 0.2	0.3	0.3
IN 0.1	0.1	0.0	Goal 2 0.0	0.0	0.0	0.1	0.1	0.2	0.2

e) The greedy policy obtained at 5th iteration and one obtained at termination is very different that is because value iteration takes a long time to converge whereas as we can see policy iteration has already converged in 5 iterations (as the policies and J is the same).

We can also notice that both the value iteration and policy iteration give the exact same policy at termination. (J differed but the difference was of the order of 1e-10, for each term thus we can see that here as J has been rounded off).

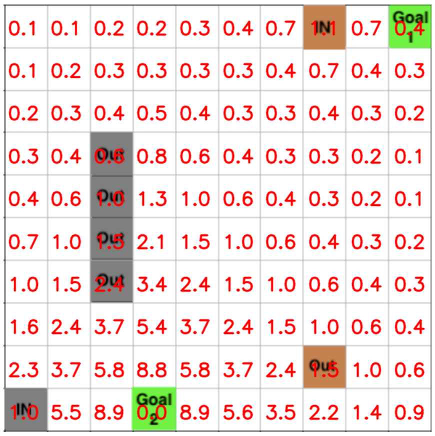
One other key observation we can make is that in the case of 5 steps value iteration, the J and policy behave as non-zero (different than 'UP') only for the states which can reach the goal in less than or equal to 5 steps. This is true because the operator T, corresponds to DP-like approach and hence only the cells which can reach the goal in 5 steps of DP are changed.

In policy iteration, we don't observe any such phenomenon and even one step of policy iteration can possibly change the policy for the whole grid. This is because while policy evaluation we take many steps (but for the same policy).

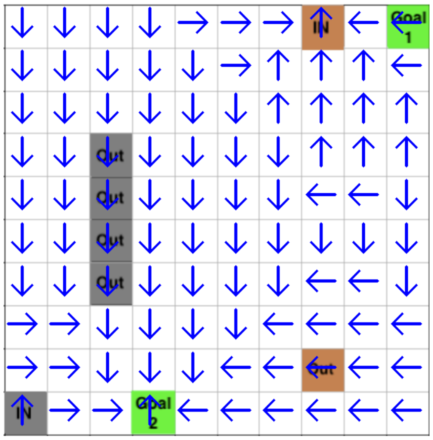
Since the policies obtained after stopping value iteration are the same, let's analyze them together -

- Arrows near Goal1 are trying to reach the Goal1 and also trying to avoid the 'Brown IN' since that takes it away from Goal1.
- All the arrows in the middle are again trying to reach Goal1.
- Arrows near the 'Grey IN' are trying to reach this IN cell because reaching there would teleport it closer to the Goal1 cell.
- Value of J is larger near the Goal (zero **at** the goal because it is a terminal state) and smaller farther away. This is due to the discount factor. Also, the 'IN' states have anomalous J values because they have special transitions closer/farther to the Goal.

3) Value Iteration -

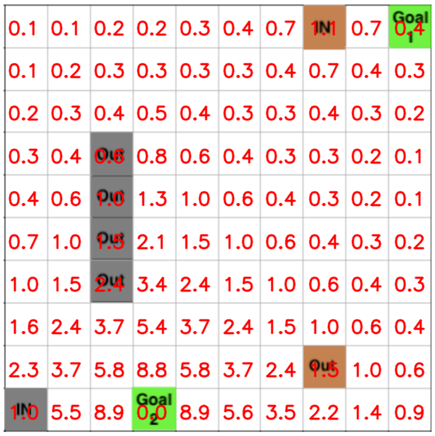


0.1	0.1	0.2	0.2	0.3	0.4	0.7	IN 1.1	0.7	Goal 1 0.4
0.1	0.2	0.3	0.3	0.3	0.3	0.4	0.7	0.4	0.3
0.2	0.3	0.4	0.5	0.4	0.3	0.3	0.4	0.3	0.2
0.3	0.4	Out 0.6	0.8	0.6	0.4	0.3	0.3	0.2	0.1
0.4	0.6	Out 1.0	1.3	1.0	0.6	0.4	0.3	0.2	0.1
0.7	1.0	Out 1.5	2.1	1.5	1.0	0.6	0.4	0.3	0.2
1.0	1.5	Out 2.4	3.4	2.4	1.5	1.0	0.6	0.4	0.3
1.6	2.4	3.7	5.4	3.7	2.4	1.5	1.0	0.6	0.4
2.3	3.7	5.8	8.8	5.8	3.7	2.4	Out 1.5	1.0	0.6
IN 1.0	5.5	8.9	Goal 2 0.0	8.9	5.6	3.5	2.2	1.4	0.9

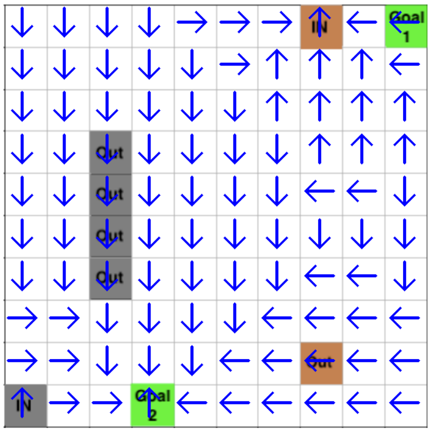


↓	↓	↓	↓	→	→	→	IN ↑	←	Goal 1 ←
↓	↓	↓	↓	↓	→	↑	↑	↑	←
↓	↓	↓	↓	↓	↓	↑	↑	↑	↑
↓	↓	Out ↓	↓	↓	↓	↓	↑	↑	↑
↓	↓	Out ↓	↓	↓	↓	↓	←	←	↓
↓	↓	Out ↓	↓	↓	↓	↓	↓	↓	↓
↓	↓	Out ↓	↓	↓	↓	↓	←	←	↓
→	→	↓	↓	↓	↓	←	←	←	←
→	→	↓	↓	↓	←	←	Out ←	←	←
IN ↑	→	→	Goal 2 ↑	←	←	←	←	←	←

Policy Iteration-



0.1	0.1	0.2	0.2	0.3	0.4	0.7	IN 1.1	0.7	Goal 1 0.4
0.1	0.2	0.3	0.3	0.3	0.3	0.4	0.7	0.4	0.3
0.2	0.3	0.4	0.5	0.4	0.3	0.3	0.4	0.3	0.2
0.3	0.4	Out 0.6	0.8	0.6	0.4	0.3	0.3	0.2	0.1
0.4	0.6	Out 1.0	1.3	1.0	0.6	0.4	0.3	0.2	0.1
0.7	1.0	Out 1.5	2.1	1.5	1.0	0.6	0.4	0.3	0.2
1.0	1.5	Out 2.4	3.4	2.4	1.5	1.0	0.6	0.4	0.3
1.6	2.4	3.7	5.4	3.7	2.4	1.5	1.0	0.6	0.4
2.3	3.7	5.8	8.8	5.8	3.7	2.4	Out 1.5	1.0	0.6
IN 1.0	5.5	8.9	Goal 2 0.0	8.9	5.6	3.5	2.2	1.4	0.9



↓	↓	↓	↓	→	→	→	IN ↑	←	Goal 1 ←
↓	↓	↓	↓	↓	→	↑	↑	↑	←
↓	↓	↓	↓	↓	↓	↑	↑	↑	↑
↓	↓	Out ↓	↓	↓	↓	↓	↑	↑	↑
↓	↓	Out ↓	↓	↓	↓	↓	←	←	↓
↓	↓	Out ↓	↓	↓	↓	↓	↓	↓	↓
↓	↓	Out ↓	↓	↓	↓	↓	←	←	↓
→	→	↓	↓	↓	↓	←	←	←	←
→	→	↓	↓	↓	←	←	Out ←	←	←
IN ↑	→	→	Goal 2 ↑	←	←	←	←	←	←

Again the terminal policy obtained by both the algorithm is the same.
We can observe patterns similar to the previous case in the terminal policies.

Some things we notice are-

- Arrows in general point towards the Goal, except near the Brown IN, where the arrows point towards this IN, since the Brown OUT is close to the goal. (Due to alpha)
- The value of J is larger near the Goal2 and smaller farther away. This is due to the discount factor. Gray IN has an exceptionally low value (compared to neighbors) because the corresponding “OUT”s are farther to Goal2 than IN.
- Brown In has an exceptionally high value because the corresponding OUT is comparatively closer to Goal2.