Homework 4
Stats 232-C

Parth Shettiwar `parthshettiwar@g.ucla.edu`

March 1, 2022

1 Regarding code

The code generates all of the following 12 plots on the run. The names with which they are saved is intuitive. All the plots are saved in the same folder as the code.

2 Value Table Visualisations

Value table for Goal A before signaling

	0	1	2	3	4	5	6
5	28.006	32.229	36.921	42.134	47.927	54.363	61.515
4	32.229	36.921	42.134	47.927	54.363	61.515	69.461
3	36.921	42.134	47.927	-44.637	61.515	69.461	78.29
2	42.134	47.927	54.363	61.515	69.461	78.29	88.1
1	36.921	42.134	47.927	-29.539	78.29	88.1	99.0
0	32.229	36.921	42.134	-37.485	69.461	78.29	88.1

Value table for Goal A after signaling

	0	1	2	3	4	5	6
5	68.898	74.414	79.866	85.496	93.329	100.748	108.99
4	74.89	81.017	87.074	93.329	102.033	110.275	119.433
3	78.766	85.574	92.304	0.783	108.925	118.083	128.259
2	83.074	90.638	98.116	106.425	116.583	126.759	138.066
1	75.433	81.874	89.304	15.917	125.093	136.399	148.962
0	68.557	74.737	81.374	4.05	112.834	123.426	135.067

Value table for Goal B before signaling

	0	1	2	3	4	5	6
5	42.134	47.927	54.363	61.515	69.461	78.29	88.1
4	47.927	54.363	61.515	69.461	78.29	88.1	99.0
3	42.134	47.927	54.363	-37.485	69.461	78.29	88.1
2	36.921	42.134	47.927	54.363	61.515	69.461	78.29
1	32.229	36.921	42.134	-51.073	54.363	61.515	69.461
0	28.006	32.229	36.921	-56.866	47.927	54.363	61.515

Value table for Goal B after signaling

	0	1	2	3	4	5	6
5	81.311	88.677	96.029	104.199	114.11	124.289	135.598
4	88.677	96.863	105.032	114.11	125.122	136.432	148.998
3	81.309	88.676	96.862	5.199	113.36	124.289	136.432
2	74.678	80.475	88.676	95.279	104.199	114.11	125.122
1	67.876	74.227	80.808	-11.749	94.446	104.199	114.944
0	61.753	67.854	73.727	-18.79	87.455	95.506	104.451

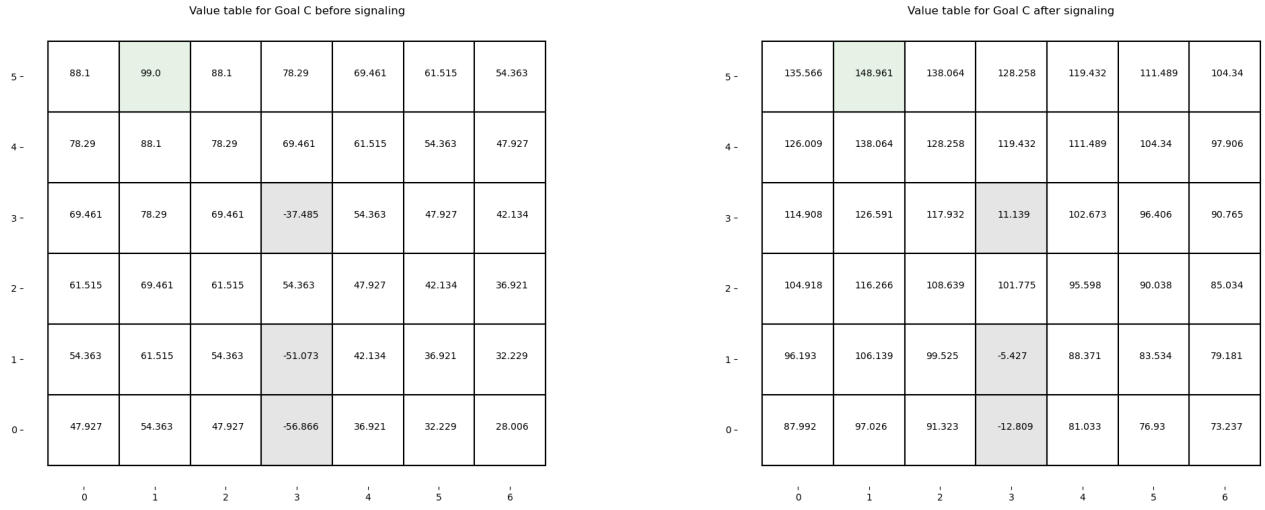
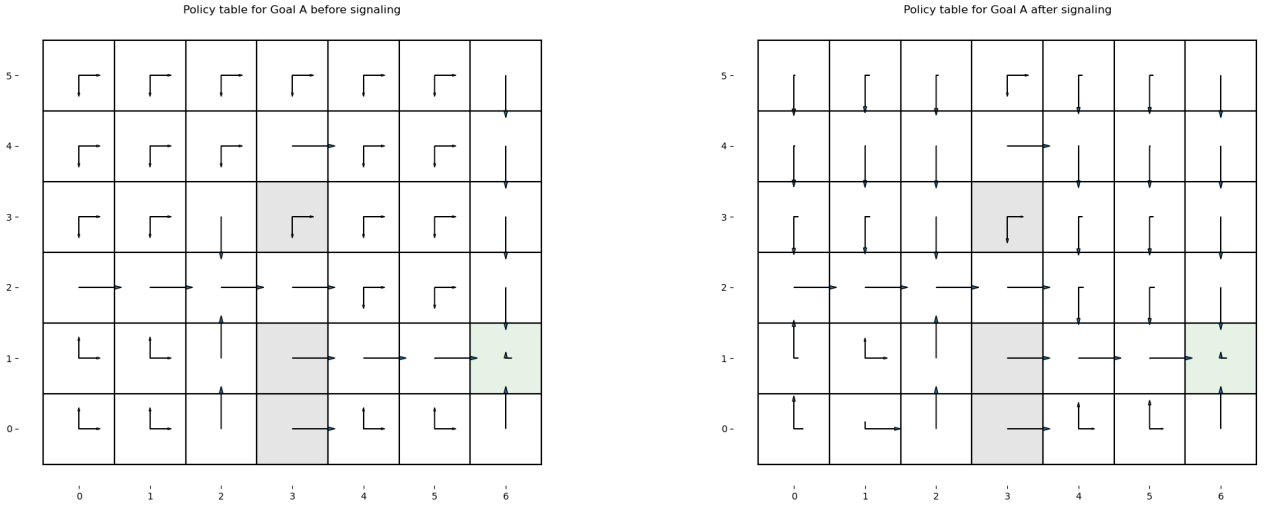


Figure 1: Value Tables for all Goals before and after adding signal as part of reward

2.1 Observations

- In all three cases, its observed that for all states, the values increases after signalling, which was expected as we increase the reward by some positive value always as described by equation for all states
- Secondly we observe that, the difference of values have become amplified around a cell. For example, previously for Goal C, (1,3) cell led to three nearby cells with values 69.461 each (for cells (0,3),(2,3),(1,2)). The cell (1,4) has highest value relatively. Hence, when we calculate probability we take action to cell (1,4) with highest probability as compared to other actions as this action leads to desired goal and highest value. But now after changing the reward, the probability of taking action to cell (1,4) becomes much higher as the value is much much higher than other neighbouring cell values. This helps in signalling the observer as the agent now would take the route to goal C with higher probability and observer can distinguish unambiguously that agent is going to goal C and not other goals
- Another observation is that, as mentioned in previous example, that neighbouring sub-optimal cells take same values previously, as all of them are equally far from desired goal. But now after incorporating signal, the values have become asymmetric. For example,in same example as before, we see that all three cells (0,3),(2,3),(1,2) have same values. Hence before signalling, any state of these is equally possible from state (1,3). However now after signalling, we see that state (0,3) has value 114.908, state (2,3) has 117.932 and state (1,2) has value 116.266. Such assymetricity actually helps the observer, in unambiguously deciding the goal the agent is aiming for (in case agent takes sub-optimal actions too)

3 Policy Table Visualisations



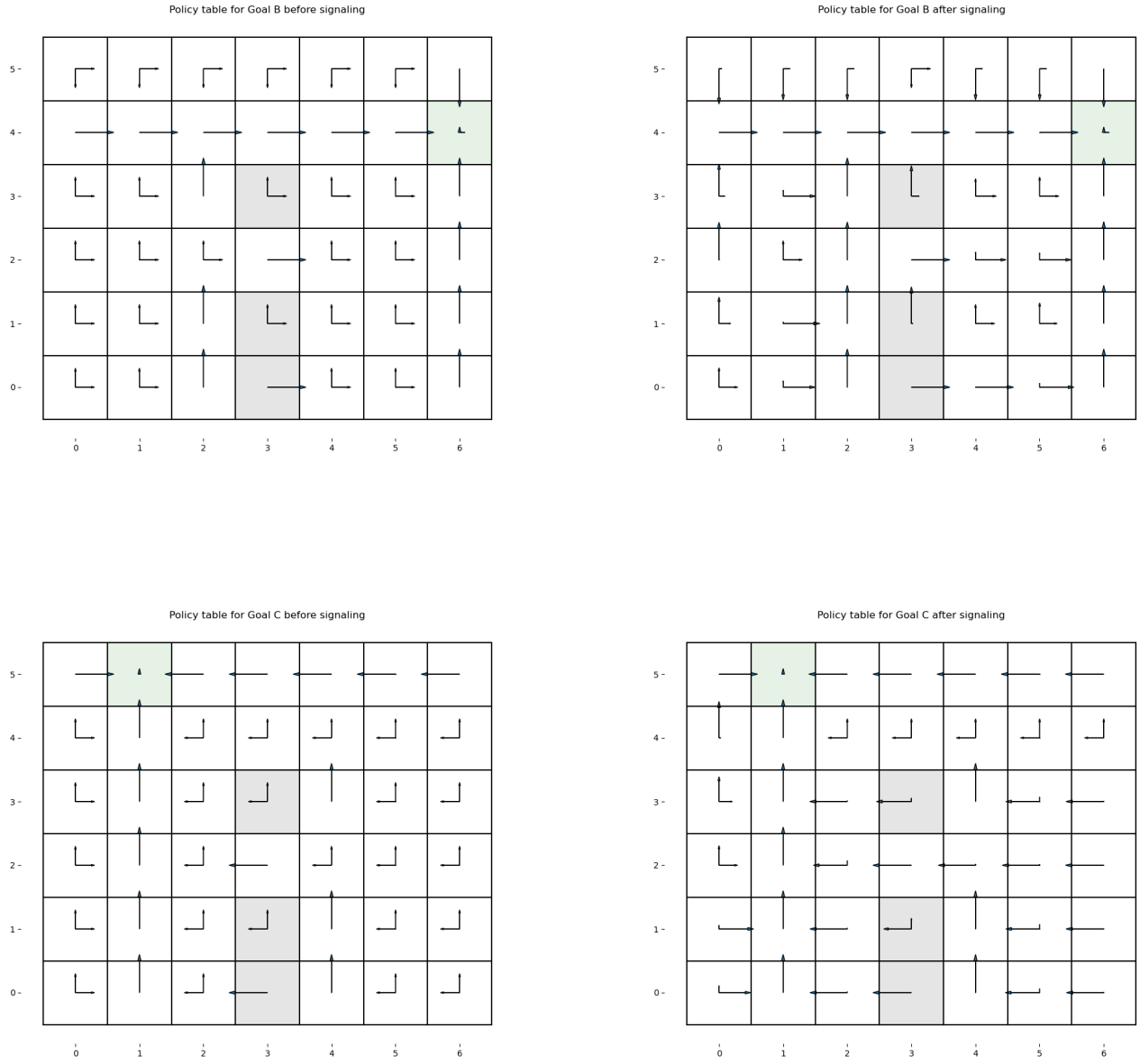


Figure 2: Policy Tables for all Goals before and after adding signal as part of reward

3.1 Observations

- As we see, for all three cases, as compared to previous goal policies, the arrows have become elongated in a particular direction almost all of times
- This is direct implication from signalling since now, as described in previous section, the action, which leads to highest closing the distance between current and goal state, will be taken with highest probability from the state. To be more formal, we know the following:

$$\pi\left(a_t \mid s_t, g\right) \propto e^{\beta Q_g^{\pi}(s_t, a_t)}$$

where Q is defined as:

$$Q_{g,w}^{\pi}\left(s_t, a_t\right) = \sum P\left(s_{t+1} \mid s_t, a_t, w\right) V_{g,w}^{\pi}\left(s_{t+1}\right) + C_{g,w}\left(a_t, s_t\right)$$

As we see, the values have increased, as well as addition of r_info reward while calculating Q function, we get a larger increase in Q value for actions from a state which lead to closing the distance between current and final goal state. This in turn leads to higher probability allocated for action from that state.

- Now we try to understand which arrows elongated and in which direction. For this we see what our signal was. As we see, the signal was devised in way that an observer should be able to unambiguously distinguish to which goal our agent is moving. Previously before signalling, all sub optimal directions had same length arrows. But now arrows size have become different. For example, for Goal C, going from state (0,4), we see that actions leading to states (0,5) and (1,4) have equal sized arrows implying that we can take any action. However after signalling, we see that from state (0,4), the arrow has become bigger and pointing towards (0,5) and arrow towards (1,4) has become smaller. Why? If an observer is watching this and if he sees agent to take action to state (1,4) then he might get confused and think for moment that final goal might be goal B too. To avoid this ambiguity, agent only takes action to state (0,5), even though both (0,5) and (1,4) take agent one step closer to desired goal.