

STATS232C Cognitive Artificial Intelligence

Assignment 4: Signaling Policy

Jingyuan Hu (#605435853)

1 Model

We first obtain the state values $V(s)$ using value iteration with original reward:

$$V(s) = \max_a \sum_{s'} P(s', R_{original}|s, a) [R_{original} + \gamma V(s')]$$

then obtain the state-action Q -function

$$Q(s, a) = \sum_{s'} P(s'|s, a) [R_{original}(s, a, s') + \gamma V(s')]$$

The corresponding Boltzmann Policy where actions are drawn from a ‘softmax’ is given by

$$\pi(a|s, g) \propto e^{\beta Q_g^{\pi}(s, a)}$$

where β is a measure of how much noise an agent is predicted to have from its optimal path. Recall that

$$P(s_{t+1}|s_t, g) = \sum_{a_t \in A_{\pi_t}} P(s_{t+1}|s_t, a_t) \pi(a_t|s_t, g)$$

Now consider how much ambiguity we are reducing by taking a particular action by

$$r_{info} = \frac{P(s_{t+1}|s_t, g_{true})}{\sum_{g \in G} P(s_{t+1}|s_t, g)}$$

This is the likelihood ratio of the next state given the actual goal compared to all goals. It should be more rewarding to try to get to next states that are likely for the true goal and unlikely for the other goals. Thus, we have

$$r_{new} = r_{original} + \alpha r_{info}$$

or more specifically in this context

$$R_{new}^g(s, a, s') = R(s, a, s') + \alpha \frac{P(s_{t+1}|s_t, g_{true})}{\sum_{g \in G} P(s_{t+1}|s_t, g)}$$

where α affects how much preference is given to the information added. Then we can derive the new state values as well as the state-action Q -function

$$\begin{aligned} V_{new}^g(s) &= \max_a \sum_{s'} P(s', R_{new}^g|s, a) [R_{new}^g + \gamma V(s')] \\ Q_{new}^g(s, a) &= \sum_{s'} P(s'|s, a) [R_{new}^g(s, a, s') + \gamma V_{new}^g(s')] \end{aligned}$$

and the Boltzmann Policy with intention to signal is given by

$$\pi_{new}(a|s, g) \propto e^{\beta Q_{new}^g(s, a)}$$

This would give us the policy for a cooperative agent with a particular goal.











































2 Results

The resulting graphs of the value table and policy under the two cases with three possible goals are given below.

2.1 Original Reward











































2.1.1 Value and Boltzmann Policy Table (Goal A)

5 -	28.0	32.226	36.92	42.134	47.927	54.363	61.515
4 -	32.226	36.92	42.134	47.927	54.363	61.515	69.461
3 -	36.92	42.134	47.927	-44.637	61.515	69.461	78.29
2 -	42.133	47.927	54.363	61.515	69.461	78.29	88.1
1 -	36.92	42.134	47.927	-29.539	78.29	88.1	98.872
0 -	32.226	36.92	42.133	-37.485	69.461	78.29	88.1
	0	1	2	3	4	5	6

5 -							
4 -							
3 -							
2 -							
1 -							
0 -							
	0	1	2	3	4	5	6

2.1.2 Value and Boltzmann Policy Table (Goal B)

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	98.872
3 -	42.134	47.927	54.363	-37.485	69.461	78.29	88.1
2 -	36.92	42.134	47.927	54.363	61.515	69.461	78.29
1 -	32.226	36.92	42.134	-51.073	54.363	61.515	69.461
0 -	28.0	32.226	36.919	-56.866	47.927	54.363	61.515
	0	1	2	3	4	5	6

5 -							
4 -							
3 -							
2 -							
1 -							
0 -							
	0	1	2	3	4	5	6

2.1.3 Value and Boltzmann Policy Table (Goal C)











































5 -	88.1	98.872	88.1	78.29	69.461	61.515	54.363
4 -	78.29	88.1	78.29	69.461	61.515	54.363	47.927
3 -	69.461	78.29	69.461	-37.485	54.363	47.927	42.134
2 -	61.515	69.461	61.515	54.363	47.927	42.134	36.92
1 -	54.363	61.515	54.363	-51.073	42.134	36.92	32.226
0 -	47.927	54.363	47.927	-56.866	36.919	32.226	28.0
	0	1	2	3	4	5	6

5 -	→	↑	←	←	←	←	←
4 -	↖	↑	↗	↗	↖	↖	↖
3 -	↖	↑	↗	↗	↑	↖	↖
2 -	↖	↑	↗	←	↖	↖	↖
1 -	↖	↑	↗	↗	↑	↖	↖
0 -	↖	↑	↗	←	↑	↖	↖
	0	1	2	3	4	5	6

2.2 Reward with Signal











































2.2.1 Value and Boltzmann Policy Table (Goal A)

5 -	68.809	74.298	79.782	85.495	93.225	100.643	108.99
4 -	74.828	80.977	87.053	93.329	102.002	110.245	119.433
3 -	78.654	85.454	92.304	0.688	108.82	117.979	128.259
2 -	83.074	90.638	98.116	106.425	116.464	126.64	138.066
1 -	75.322	81.85	89.304	15.917	125.093	136.399	148.834
0 -	68.417	74.603	81.374	4.05	112.739	123.308	135.067
	0	1	2	3	4	5	6

5 -							
4 -							
3 -							
2 -							
1 -							
0 -							
	0	1	2	3	4	5	6

2.2.2 Value and Boltzmann Policy Table (Goal B)

5 -	81.217	88.54	95.893	104.199	113.973	124.152	135.598
4 -	88.677	96.863	105.032	114.11	125.122	136.432	148.871
3 -	81.177	88.544	96.862	5.06	113.322	124.289	136.432
2 -	74.641	80.475	88.652	95.279	104.06	113.973	125.122
1 -	67.748	74.159	80.808	-11.813	94.446	104.161	114.944
0 -	61.736	67.715	73.725	-18.79	87.428	95.392	104.451
	0	1	2	3	4	5	6

5 -							
4 -							
3 -							
2 -							
1 -							
0 -							
	0	1	2	3	4	5	6

2.2.3 Value and Boltzmann Policy Table (Goal C)

5 -	135.566	148.833	138.064	128.258	119.432	111.489	104.34
4 -	125.938	138.064	128.258	119.432	111.489	104.34	97.906
3 -	114.8	126.591	117.887	11.034	102.673	96.287	90.727
2 -	104.904	116.266	108.52	101.775	95.542	89.987	85.018
1 -	96.083	106.139	99.474	-5.56	88.37	83.412	79.144
0 -	87.855	97.026	91.272	-12.809	81.029	76.809	73.174
	0	1	2	3	4	5	6

5 -	→	↑	←	←	←	←	←
4 -	↑	↑	↗	↗	↗	↗	↗
3 -	↖	↑	←	←	↑	←	←
2 -	↖	↑	←	←	←	←	←
1 -	←	↑	←	←	↑	←	←
0 -	←	↑	←	←	↑	←	←
	0	1	2	3	4	5	6

3 Conclusion

Take goal A as an example (the observations and conclusions are similar for the other two cases): in the original goal policies, the optimal policies are indifferent between several actions (e.g. equal probability between moving towards two directions) in most of the states, as they are approaching the given goal with same rewards and costs but only different paths. However, under the new signaling policies, we are assigning more rewards to certain directions to distinguish between goals. Specifically, for state (0, 5), our new signaling

policy would prefer moving downward instead of being indifferent between moving downward and moving right. The latter action could be confuse as it could distinguish goal A $(6, 1)$ and goal C $(1, 5)$.