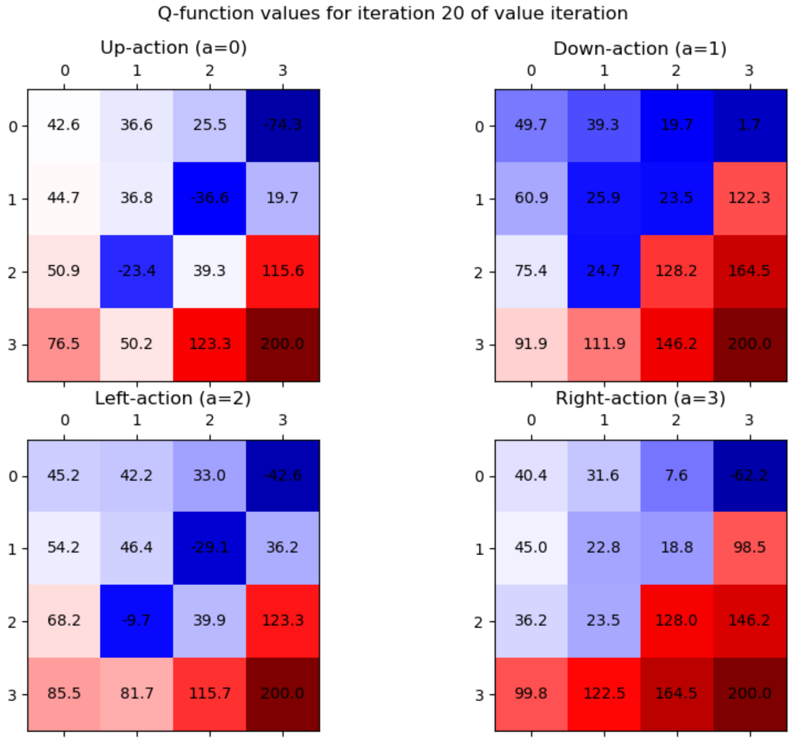
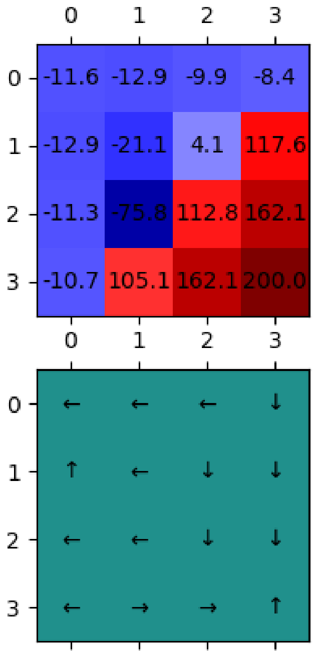
CS 5789 A1 Writeup

Name: Xingze Li NetID: xl834 15/Feb/2023

Value iteration section (20 iteration):

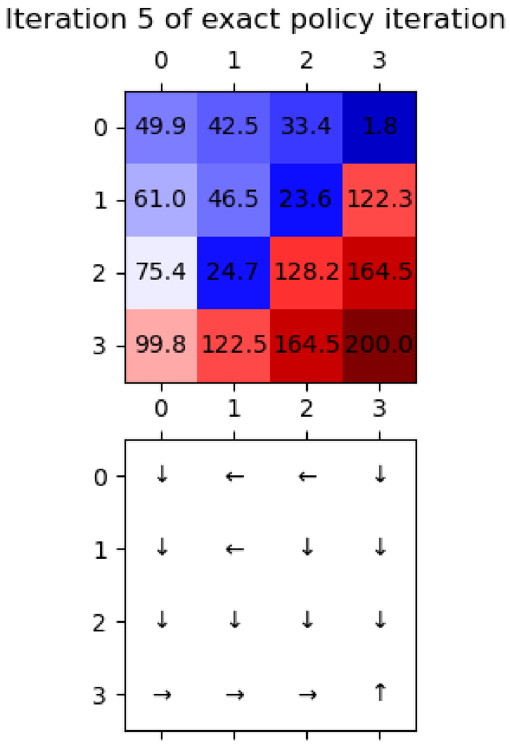


max policy and corresponding value function:



Policy iteration section:

Exact iteration:

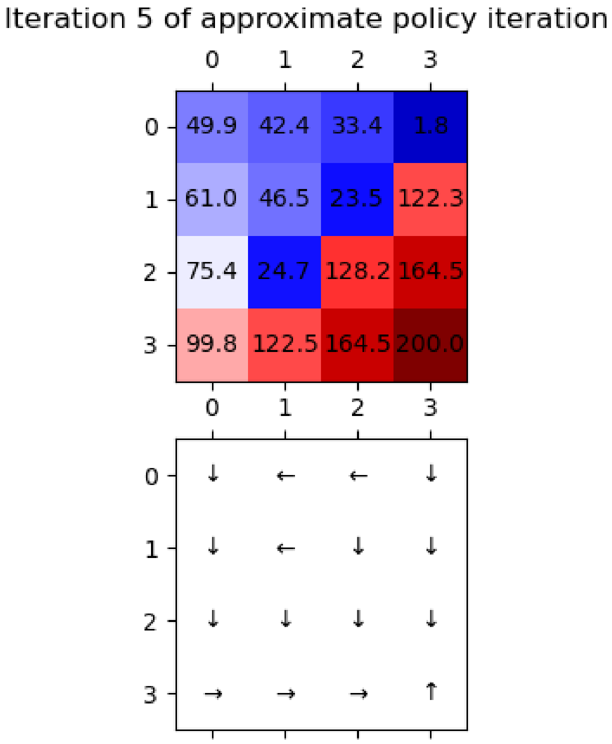


Iteration 5 of exact policy iteration

	0	1	2	3
0	49.9	42.5	33.4	1.8
1	61.0	46.5	23.6	122.3
2	75.4	24.7	128.2	164.5
3	99.8	122.5	164.5	200.0

	0	1	2	3
0	↓	←	←	↓
1	↓	←	↓	↓
2	↓	↓	↓	↓
3	→	→	→	↑

Approx iteration:



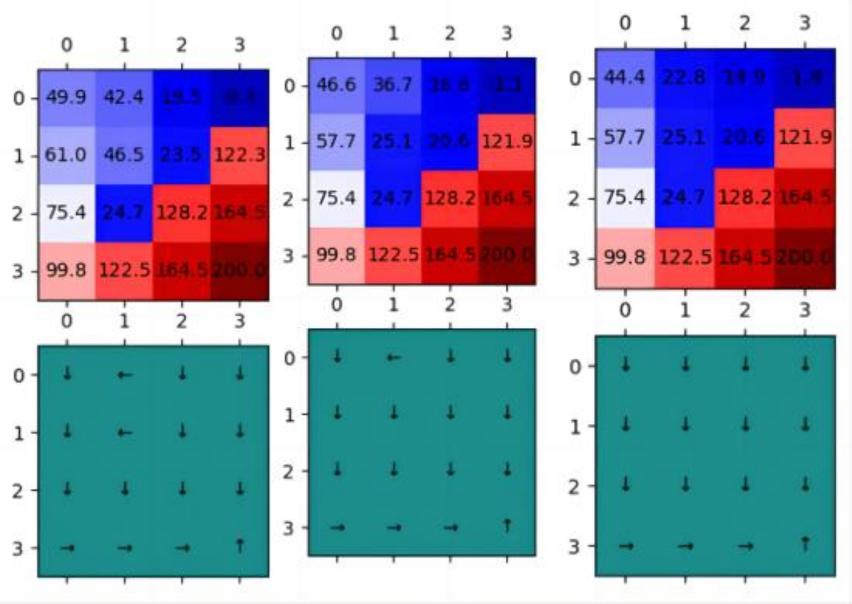
Iteration 5 of approximate policy iteration

	0	1	2	3
0	49.9	42.4	33.4	1.8
1	61.0	46.5	23.5	122.3
2	75.4	24.7	128.2	164.5
3	99.8	122.5	164.5	200.0

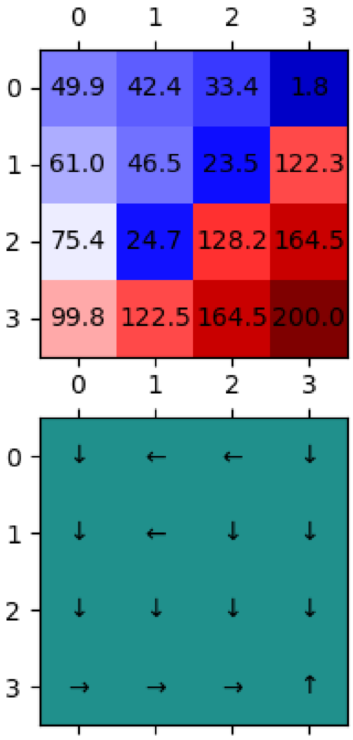
	0	1	2	3
0	↓	←	←	↓
1	↓	←	↓	↓
2	↓	↓	↓	↓
3	→	→	→	↑

IL section:

With N = 75 datapoints,
The learned policy and corresponding value:



The expert policy and corresponding value:



Comparison: The expert policy is more likely to choose go left when encountering region with reward -80. In terms of value function, if policy doesn`t choose to turn left when stays at left side of -80 reward region. Its surrounding region`s value will be lower.