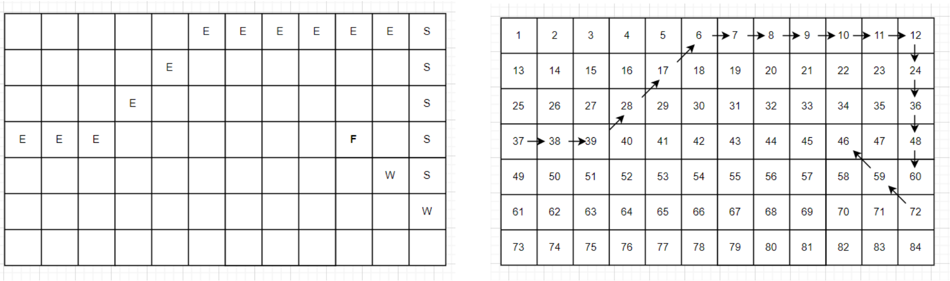
ELL729 Coding Assignment 3

Suraj Joshi

January 16, 2021

1 Troubled Waters Problem I

The optimal path from A = 37 to F = 46 is given by [37, 38, 39, 28, 17, 6, 7, 8, 9, 10, 11, 12, 24, 36, 48, 60, 72, 59, 46]. This path along with the actions taken from source to destination is shown in the following figure:



					E	E	E	E	E	E	S
				E							S
			E								S
E	E	E							F		S
										W	S
											W

(a) Policy followed along the path from A to F (the final state is just indicated by F)

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81	82	83	84

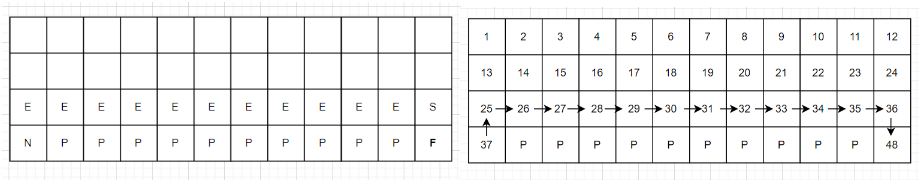
(b) The actual path from A = 37 to F = 46

Figure 1: The path from A = 37 to F = 46 for Problem 1

The difference between different epsilon values: All epsilon values do not result in optimal convergence equally regularly. Some balance between exploration and exploitation is required for good results. Generally for epsilon too large or too small the convergence takes some time. However in such cases increasing the number of episodes improves the probability of converging to the optimal path. Usually with 1000 or more episodes, the algorithm always converges to the optimal path.

2 Troubled Waters Problem II

The optimal path from A = 37 to F = 48 is given by: [37, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 48]. This path along with the actions taken from source to destination is shown in the following figure:



E	E	E	E	E	E	E	E	E	E	E	S
N	P	P	P	P	P	P	P	P	P	P	F

(a) Policy from the path from A to F (the final state is just indicated by F)

1	2	3	4	5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20	21	22	23	24
25 →	26 →	27 →	28 →	29 →	30 →	31 →	32 →	33 →	34 →	35 →	36 ↓
↑ 37	P	P	P	P	P	P	P	P	P	P	48

(b) Moves from A to F

Figure 2: The path from A = 37 to F = 48 for Problem 2 (P indicates pirates in a location)

This task is technically episodic because there is an end state (F). However, there is a possibility of going in circles by repeatedly moving into the locations with pirates and circling from A to the pirates back to A repeatedly. The epsilon greedy strategy poses a problem because it sometimes leads to the agent learning to go to the pirates again and again in a cycle rather than learning the optimal path. Even after the optimal path is obtained and the algorithm is run for near optimal Q values, the expected average reward for every episode should be -1 which is the cost of steps along the shortest path without visiting the pirates, however we can see on the plot of average rewards that in several episodes rewards lower than these are also obtained indicating that even in a near optimal policy there is a tendency to go off the optimal path and move into the location with the pirates, in other words no learning has happened. Hence epsilon greedy Q learning is not a good strategy for this problem. The plot of average rewards with near optimal policy is as shown below:

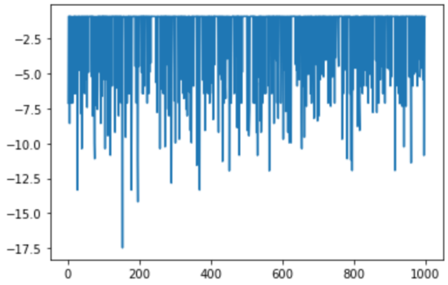


Figure 3: The average reward per episode using (near) optimal policy