Reinforcement Learning - Project 2

Fridolin Paiki

March 4, 2024

1 Problem 1

Given:

Inital policy $\pi^0(S) = Left$ and $V_0(S) = 0$ for all $S \in S$. The map for state ID can be seen in Figure 1



Figure 1: State IDs for all states

1.1 Part 1 - Policy Iteration

1.1.1 Base Scenario

$$p = 0.02, \gamma = 0.95, \theta = 0.01$$

	Maze Problem																			
0 -	-14																			
- 1		1445	1524	1608	1696	1790	1888	1990	2098	2212	2332	2442	0	2716	2594	2461	2334	2214	2101	
5 -		1370	1445	1524	1608		1986	2093	2210	2331	2458	2592	2718	2882	2734	2593	2459	2329	2209	
m -		1300	1370	1446	1524		2093	2207	2329	2456	2591	2733	2883	2826	2883	2733	2591	2456	2328	
4 -		1232	1299															2328	2208	
2 -		1168	1226		1081	1140	1209	1275	1346	1431	1509	1593	1680	1773	1870	1973	2081	2206	2095	
9 -		1101	1161		1025	1080		1209	1275		1434	1512	1595	1682	1773		1974	2080	1987	
7 -		1044	1100		972	1024		1147	1208		1360	1433					1872	1961	1884	
ω -		990	1032	978	927	970		1087	1145		1280	1348	1277	1210	1147		1776	1848	1787	
ი -		939	978	927	879	919		1031	1085		1214	1277	1211	1147	1088		1684	1742	1695	
10						871		977	1028			1210	1147	1088	1031		1597	1651	1607	
11		664	701	741	782	825		927	974	923		1146	1088		978				1519	
12		629	664						922	875		1086	1031		938	989	1043		1440	
13		596	629	595	563	534	557		874	829		1019	968		989	1043	1100		1365	
14		565	595	564	534	557	588		828	786		965	917		1043	1100	1161	1226	1293	
15		525	553	534	557	588	621		784	815	866	914	870					1162	1226	
16		499	527	556	588	621	656	704	743	773	821	866	834	879	928	979	1034	1091	1151	
17				528	557	588	621							833	880	929	980	1034	1091	
18		448	473	500	528	557	588	572	603	637	673	710	750	791	835	881	929	980	1034	
19																				
•	0	í	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 2: The Optimal Value Function values at all states for the Base Scenario under the Policy Iteration method

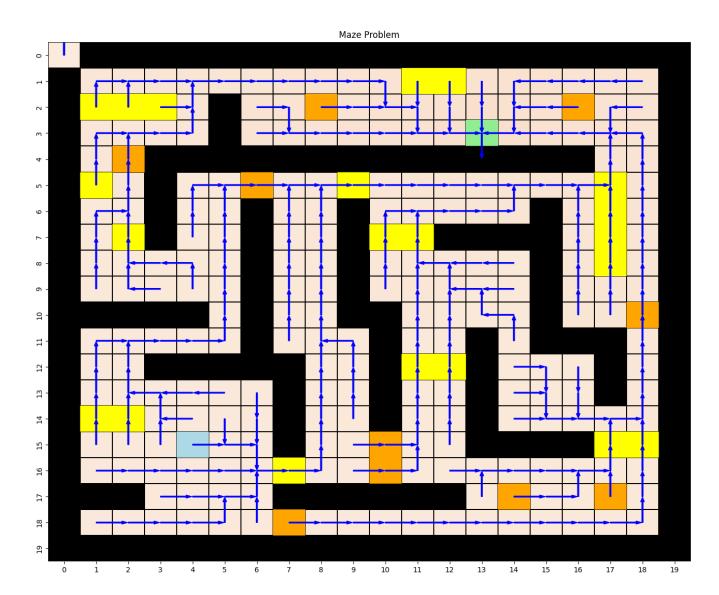


Figure 3: The Optimal Policy values at all states for the Base Scenario under the Policy Iteration method

1.1.2 Large Stochasticity Scenario

$$p = 0.5, \gamma = 0.95, \theta = 0.01$$

-	Maze Problem																			
0 -	-1																			
п-		-1	-1	-1	-1	-1	-1	-5	4	19	63	98	0	301	254	145	77	21	2	
2 -		-2	-3	-2	-1		-1	-1	11	56	112	225	348	548	382	244	117	51	8	
m -		-1	-2	-1	-1		-3	4	31	83	190	343	580	472	585	346	190	72	28	
4 -		-1	-2															27	5	
- 22		-2	-1		-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	3	-2	
9 -		-1	-1		-1	-1		-1	-1		-1	-2	-1	-1	-1		-1	-2	-1	
7		-1	-2		-1	-1		-1	-1		-3	-3					-1	-2	-1	
o -		-1	-2	-1	-2	-1		-1	-1		-1	-2	-1	-1	-1		-1	-2	-1	
6 -		-1	-1	-1	-1	-1		-1	-1		-1	-2	-2	-2	-1		-1	-2	-1	
10						-1		-1	-1			-1	-2	-1	-1		-1	-1	-2	
11 -		-1	-1	-1	-1	-1		-1	-2	-1		-1	-1		-1				-1	
12		-1	-1						-1	-1		-2	-2		-1	-1	-1		-1	
13		-1	-2	-1	-1	-1	-1		-1	-1		-1	-1		-1	-2	-1		-1	
14		-3	-3	-3	-3	-3	-1		-1	-1		-1	-1		-1	-1	-1	-1	-1	
15		-1	-3	-4	-4	-3	-1		-1	-2	-3	-2	-1					-3	-2	
16		-1	-1	-3	-4	-4	-3	-3	-1	-1	-3	-1	-1	-1	-1	-1	-1	-2	-1	
17				-1	-3	-3	-1							-1	-2	-2	-2	-2	-1	
18		-1	-1	-1	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
19																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 4: The Optimal Value Function values at all states for the Large Stochastic Scenario under the Policy Iteration method

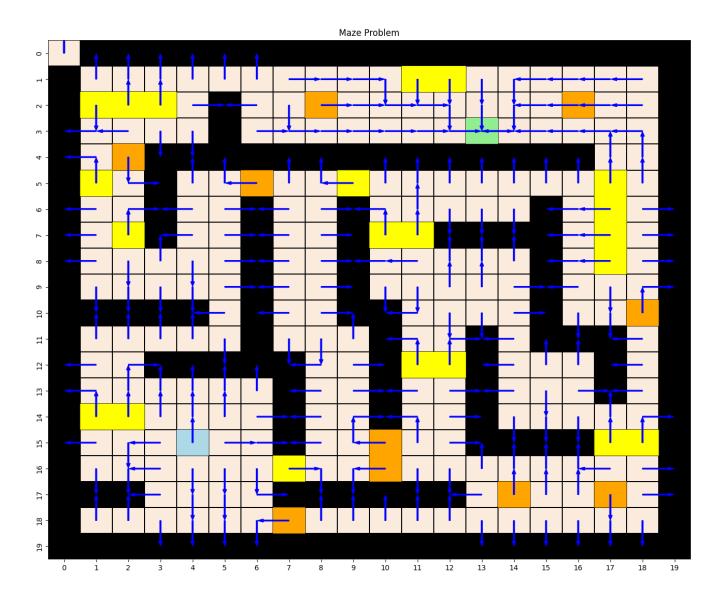


Figure 5: The Optimal Policy values at all states for the Large Stochastic Scenario under the Policy Iteration method

1.1.3 Small Discount Factor Scenario

$$p = 0.02, \gamma = 0.55, \theta = 0.01$$

-	Maze Problem																			
0 -	-1																			
п-		-1	-1	-1	-1	-1	-1	-5	4	19	63	98	0	301	254	145	77	21	2	
2 -		-2	-3	-2	-1		-1	-1	11	56	112	225	348	548	382	244	117	51	8	
m -		-1	-2	-1	-1		-3	4	31	83	190	343	580	472	585	346	190	72	28	
4 -		-1	-2															27	5	
- 22		-2	-1		-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	3	-2	
9 -		-1	-1		-1	-1		-1	-1		-1	-2	-1	-1	-1		-1	-2	-1	
7		-1	-2		-1	-1		-1	-1		-3	-3					-1	-2	-1	
o -		-1	-2	-1	-2	-1		-1	-1		-1	-2	-1	-1	-1		-1	-2	-1	
6 -		-1	-1	-1	-1	-1		-1	-1		-1	-2	-2	-2	-1		-1	-2	-1	
10						-1		-1	-1			-1	-2	-1	-1		-1	-1	-2	
11 -		-1	-1	-1	-1	-1		-1	-2	-1		-1	-1		-1				-1	
12		-1	-1						-1	-1		-2	-2		-1	-1	-1		-1	
13		-1	-2	-1	-1	-1	-1		-1	-1		-1	-1		-1	-2	-1		-1	
14		-3	-3	-3	-3	-3	-1		-1	-1		-1	-1		-1	-1	-1	-1	-1	
15		-1	-3	-4	-4	-3	-1		-1	-2	-3	-2	-1					-3	-2	
16		-1	-1	-3	-4	-4	-3	-3	-1	-1	-3	-1	-1	-1	-1	-1	-1	-2	-1	
17				-1	-3	-3	-1							-1	-2	-2	-2	-2	-1	
18		-1	-1	-1	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
19																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 6: The Optimal Value Function values at all states for the Small Discount Factor Scenario under the Policy Iteration method

Figure 7: The Optimal Policy values at all states for the Small Discount Factor Scenario under the Policy Iteration method

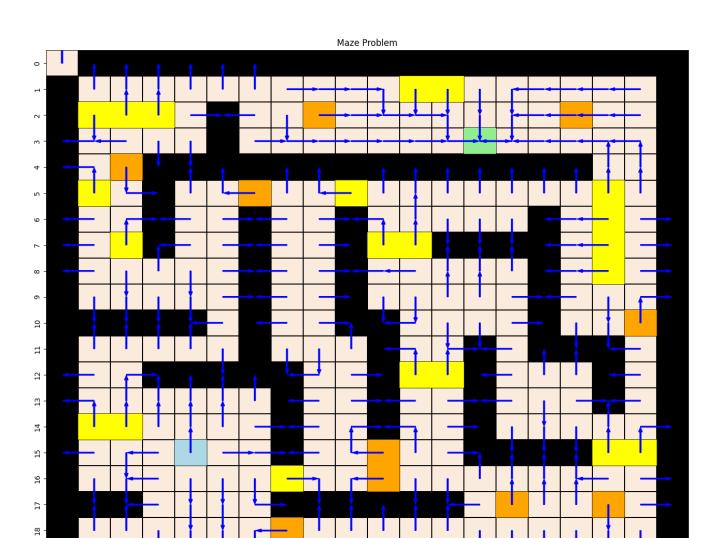


Figure 8: Enter Caption

1.2 Part 2 - Value Iteration

Base Scenario

$$p = 0.02, \gamma = 0.95, \theta = 0.01$$

Figure 9: The Optimal Value Function values at all states for the Base Scenario under the Value Iteration method

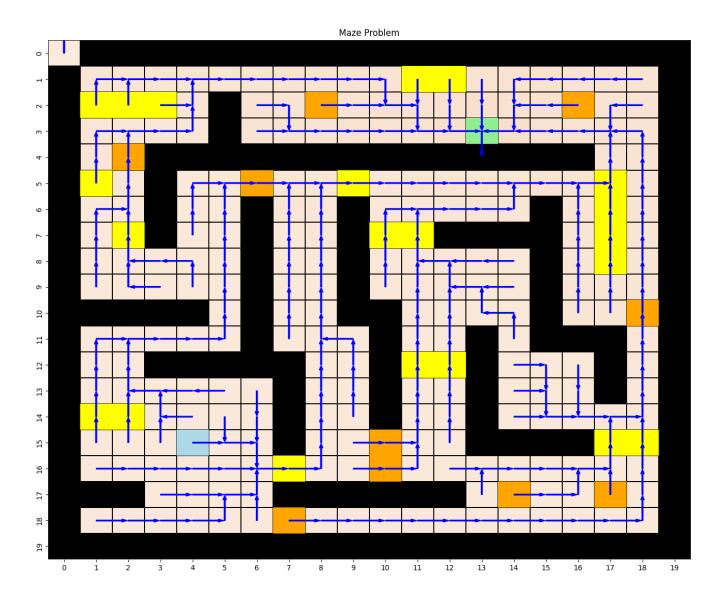


Figure 10: The Optimal Policy values at all states for the Base Scenario under the Value Iteration method

1.2.1 Large Stochasticity Scenario

$$p = 0.5, \gamma = 0.95, \theta = 0.01$$

	Maze Problem																			
0 -	-1																			
1		-1	-1	-1	-1	-1	-1	2	23	72	146	206	0	493	505	421	341	268	216	
7 -		-2	-3	-2	-1		-1	10	58	137	249	389	548	805	694	559	440	334	267	
m -		-1	-2	-1	-1		5	31	94	201	357	566	840	781	918	701	531	381	294	
4 -		-1	-2															272	237	
2 -		-2	-1		-1	-1	-2	-1	-1	-2	-1	-1	-1	-0	9	34	80	177	179	
9 -		-1	-1		-1	-1		-1	-1		-1	-2	-1	-1	3		61	106	128	
7		-1	-2		-1	-1		-1	-1		-3	-3					41	60	86	
œ -		-1	-2	-1	-2	-1		-1	-1		-1	-3	-1	-1	-1		25	33	55	
ი -		-1	-1	-1	-1	-1		-1	-1		-1	-2	-3	-2	-1		14	18	34	
10						-1		-1	-1			-1	-2	-1	-1		7	12	19	
11		-1	-1	-1	-1	-1		-1	-2	-1		-1	-1		-1				3	
12		-1	-1						-1	-1		-2	-2		-1	-1	-1		-0	
13		-1	-2	-1	-1	-1	-1		-1	-1		-1	-1		-1	-2	-1		-1	
14		-3	-3	-3	-3	-3	-1		-1	-1		-1	-1		-1	-1	-1	-1	-1	
15		-1	-3	-4	-4	-3	-1		-1	-2	-3	-2	-1					-3	-2	
16		-1	-1	-3	-4	-4	-3	-3	-1	-1	-3	-1	-1	-1	-1	-1	-1	-2	-1	
17				-1	-3	-3	-1							-1	-2	-2	-2	-2	-1	
18		-1	-1	-1	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	
19																				
	0	í	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 11: The Optimal Value Function values at all states for the Large Stochastic Scenario under the Value Iteration method

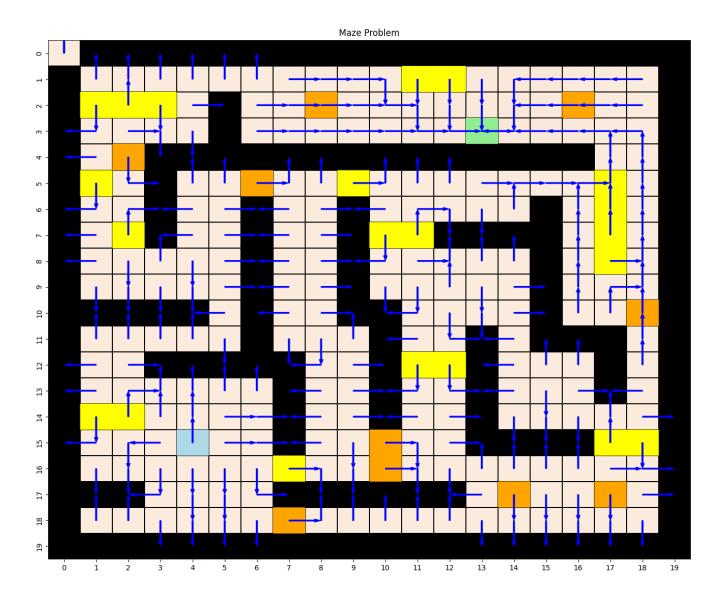


Figure 12: The Optimal Policy values at all states for the Large Stochastic Scenario under the Value Iteration method

1.2.2 Small Discount Factor Scenario

$$p = 0.02, \gamma = 0.55, \theta = 0.01$$

	Maze Problem																			
0 -	-2																			
1		-2	-2	-2	-2	-2	-2	1	6	15	33	64	0	229	125	67	36	18	9	
2 -		-2	-2	-2	-2		0	6	15	33	64	123	230	427	232	125	67	35	18	
m -		-2	-2	-2	-2		6	15	32	64	122	230	427	420	429	232	125	67	35	
4 -		-2	-2															35	18	
2 -		-2	-2		-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	18	9	
9 -		-2	-2		-2	-2		-2	-2		-2	-2	-2	-2	-2		-2	1	4	
7		-2	-2		-2	-2		-2	-2		-2	-2					-2	-0	1	
ω -		-2	-2	-2	-2	-2		-2	-2		-2	-2	-2	-2	-2		-2	-1	-0	
ი -		-2	-2	-2	-2	-2		-2	-2		-2	-2	-2	-2	-2		-2	-2	-1	
10						-2		-2	-2			-2	-2	-2	-2		-2	-2	-2	
11 -		-2	-2	-2	-2	-2		-2	-2	-2		-2	-2		-2				-2	
12		-2	-2						-2	-2		-2	-2		-2	-2	-2		-2	
13		-2	-2	-2	-2	-2	-2		-2	-2		-2	-2		-2	-2	-2		-2	
14		-2	-2	-2	-2	-2	-2		-2	-2		-2	-2		-2	-2	-2	-2	-2	
15		-2	-2	-2	-2	-2	-2		-2	-2	-2	-2	-2					-2	-2	
16		-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	
17				-2	-2	-2	-2							-2	-2	-2	-2	-2	-2	
18		-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	
19																				
	0	í	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Figure 13: The Optimal Value Function values at all states for the Small Discount Factor Scenario under the Value Iteration method



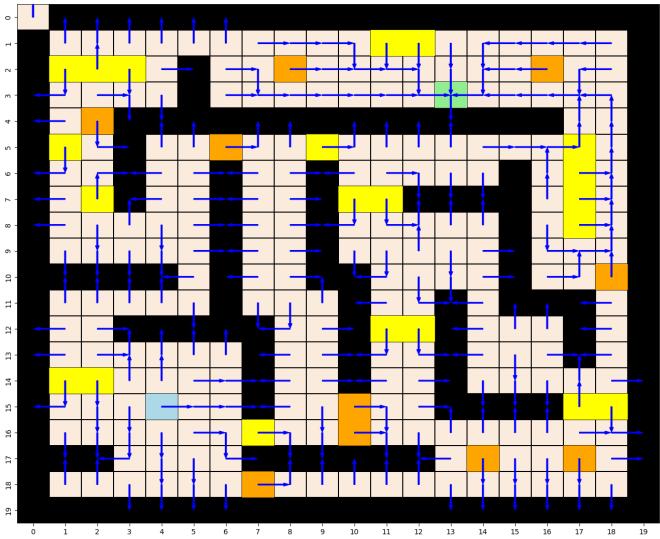


Figure 14: The Optimal Policy values at all states for the Small Discount Factor Scenario under the Value Iteration method

Just like in the policy iteration case, the agent in the ideal scenario finds the fastest route to the goal with a high overall value. However, when randomness (stochasticity) is introduced, the agent might deviate from the optimal policy, leading to a longer path. In the case of a very small discount factor, the agent prioritizes immediate rewards so much that it might never reach the goal at all, getting stuck chasing short-term gains.

2 Problem 2

Given $2^4 = 16$ states and 5 possible actions in action space $a \in A = \{a^1 = [0, 0, 0, 0]^T, a^2 = [1, 0, 0, 0]^T, a^3 = [0, 1, 0, 0]^T, a^4 = [0, 0, 1, 0]^T, a^5 = [0, 0, 0, 1]^T\}$

2.1 Part a - Matrix-Form Value Iteration

$$p = 0.05, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:

2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2

The average activation genes AvgA are 0.48 for no-control policy and 2.83 for the optimal policy.

2.2 Part b - Matrix-Form Value Iteration

$$p = 0.2, \gamma = 0.95, \theta = 0.01$$

The average activation genes AvgA are 1.25 for no-control policy and 2.41 for the optimal policy.

$$p = 0.45, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:

The average activation genes AvgA are 1.92 for no-control policy and 1.90 for the optimal policy.

Basically, as the value iteration algorithm gets more random (the higher p value), it struggles to find the best reward. When the randomness parameter (p) is set to 0.45, the algorithm mostly takes random actions, which is no different from having no control strategy at all. In this case, it always picks action zero because it offers the most immediate reward. However, with less randomness, the algorithm can better balance exploring new options and exploiting what it already knows, ultimately resulting in a better strategy than simply doing nothing.

2.3 Part c - Matrix-Form Policy Iteration

$$p = 0.05, \gamma = 0.95, \theta = 0.01$$

Based on the calculation of matrix-form value iteration, I found that the optimal policy is:

2 2 2 2 2 2 2 2 3 2 2 2 4 2 2 2

The average activation genes AvgA are 0.48 for no-control policy and 2.85 for the optimal policy.

Both policy iteration and value iteration converge to the same optimal policy. Likewise, the average activation rate they achieve is very similar, with slight variations due to the inherent randomness of the problem. The key difference lies in the number of iterations required. Value iteration takes 137 iterations, while policy iteration takes 3 iterations. This efficiency boost in policy iteration comes from performing the policy evaluation step in a single iteration, essentially doing one big matrix calculation.