Problem 1

Question1.

In this question use vector-form Policy Iteration technique to find the optimal policy and the optimal state values. Choose action Left in all states as your initial policy for the Policy Iteration method and all-zero initial state values for the Value Iteration method.

In the Base Scenario factor, because the probability of selecting the optimal action is high , so the algorithm is likely to converge more rapidly to the optimal policy. This is because the agent is more likely to select the action with the highest expected reward, which leads to a more efficient search for the optimal policy. In the Large Stochasticity Scenario factor, because the optima policy would be selected by a small probability, which will result in a suboptimal policies. So compare to the Large Stochasticity Scenario, it is easy for Base Scenario factor to get the optimal policy.

In the Small Discount Factor Scenario , because it has the lower gamma compare to the other two factors, so the agent will give higher priority to immediate rewards over long-term rewards, in this situation the agent would rather take risky or uncertain actions, so in the give gamma of the question, the agent was trapped in the loop.

The plots of Policy Iteration values,actions and path are as follow:

a. P =0.02, γ =0.95， θ =0.01

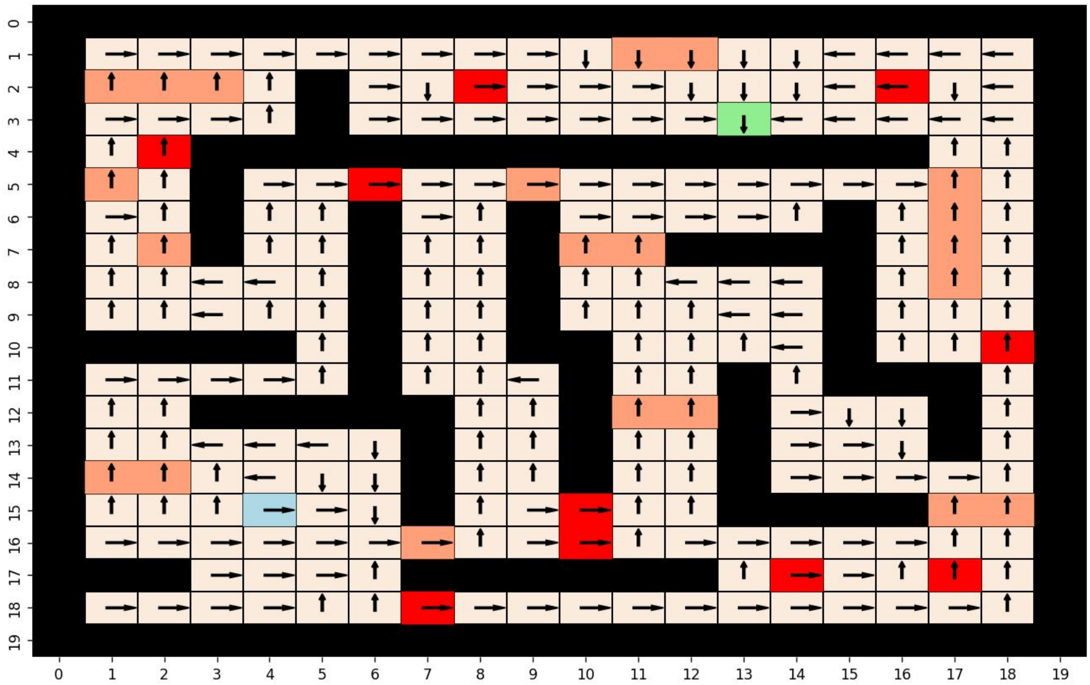


Figure 1.1 The optimal policy for each state for the base scenario under the Policy Iteration method.

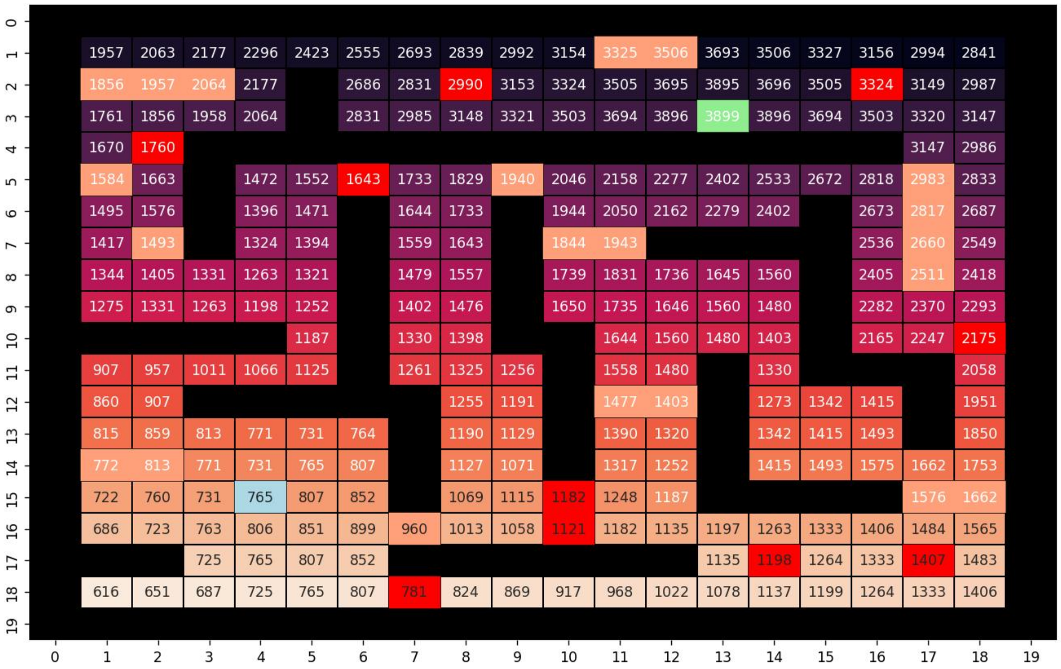


Figure 1.2 The optimal value function values for each state for the base scenario under the Policy Iteration method

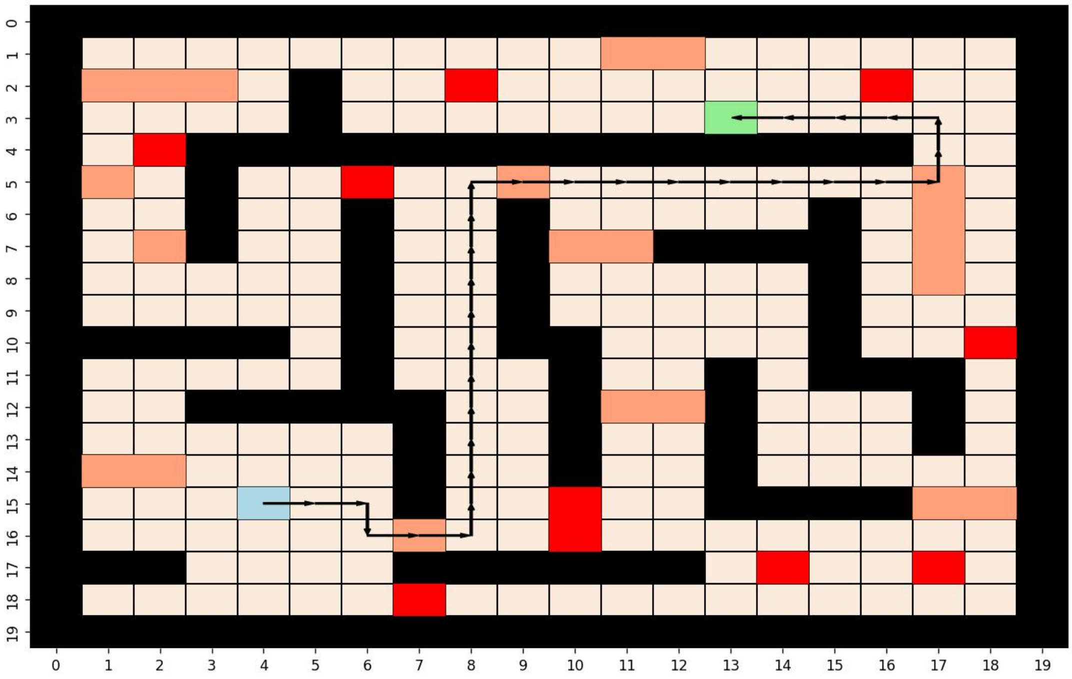


Figure 1.3 The optimal path from the start state to the end state for the base scenario under the Policy Iteration method.

b. P =0.5, γ =0.95， θ =0.01

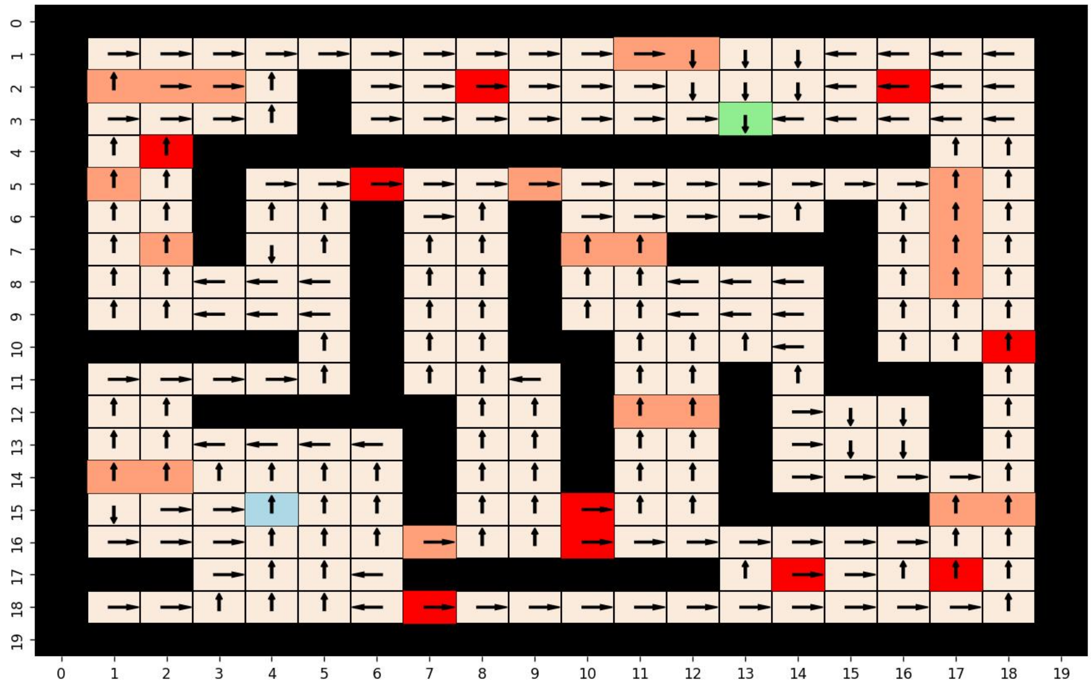


Figure 1.4The optimal policy for each state for the base scenario under the Policy Iteration method.

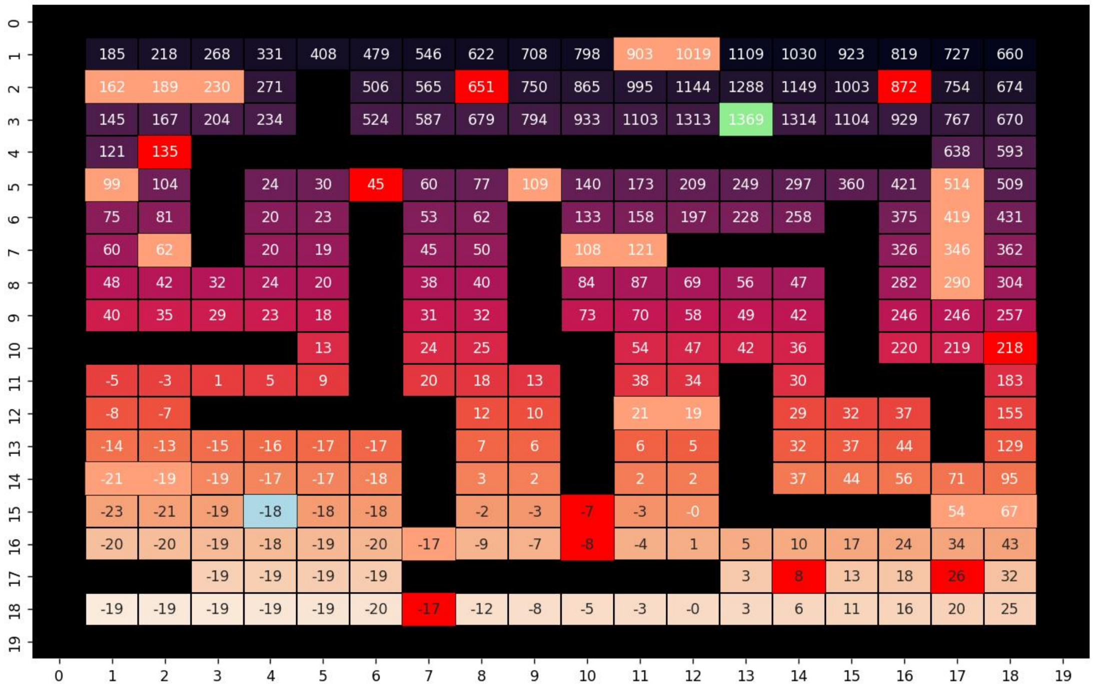


Figure 1.5The optimal value function values for each state for the base scenario under the Policy Iteration method

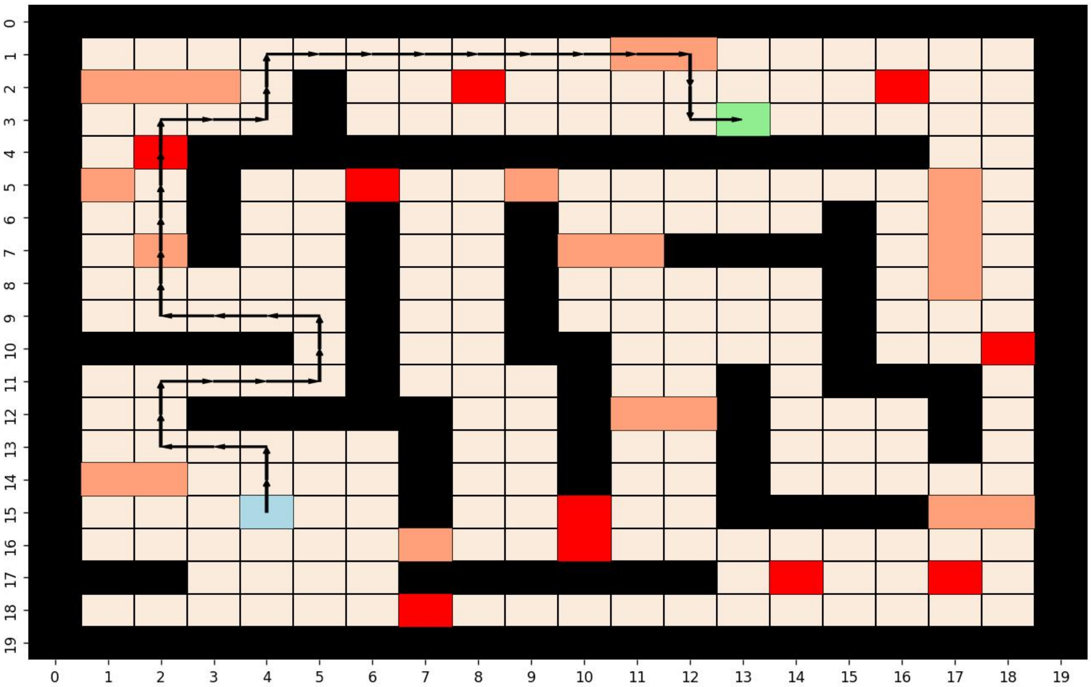


Figure 1.6The optimal path from the start state to the end state for the base scenario under the Policy Iteration method.

c. P =0.02, γ =0.55， θ =0.01

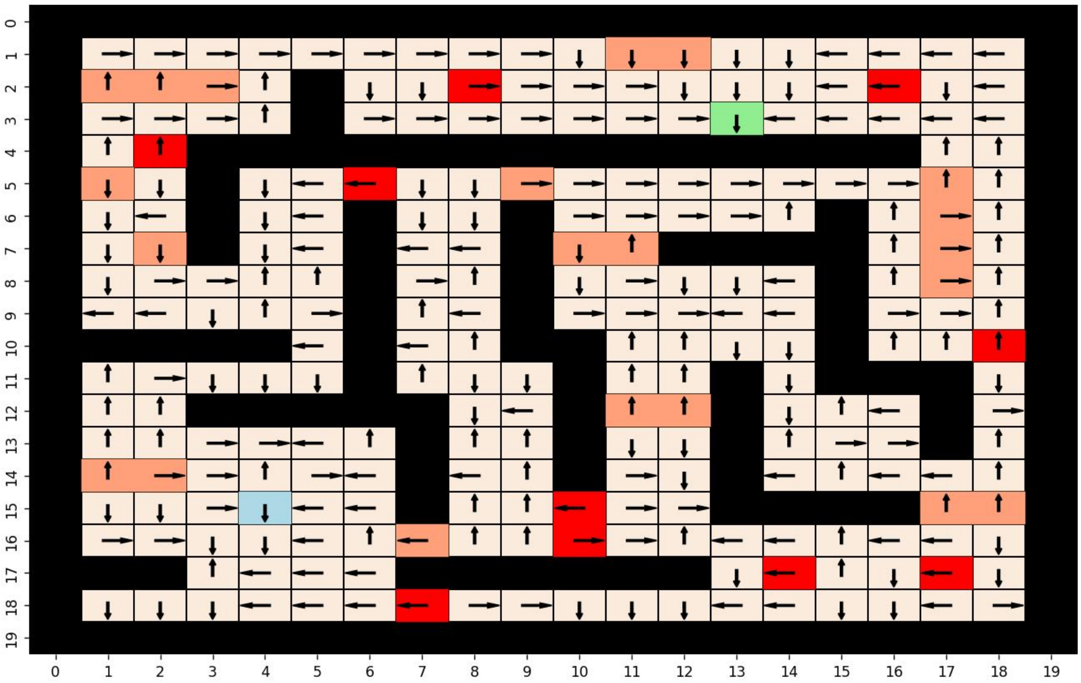
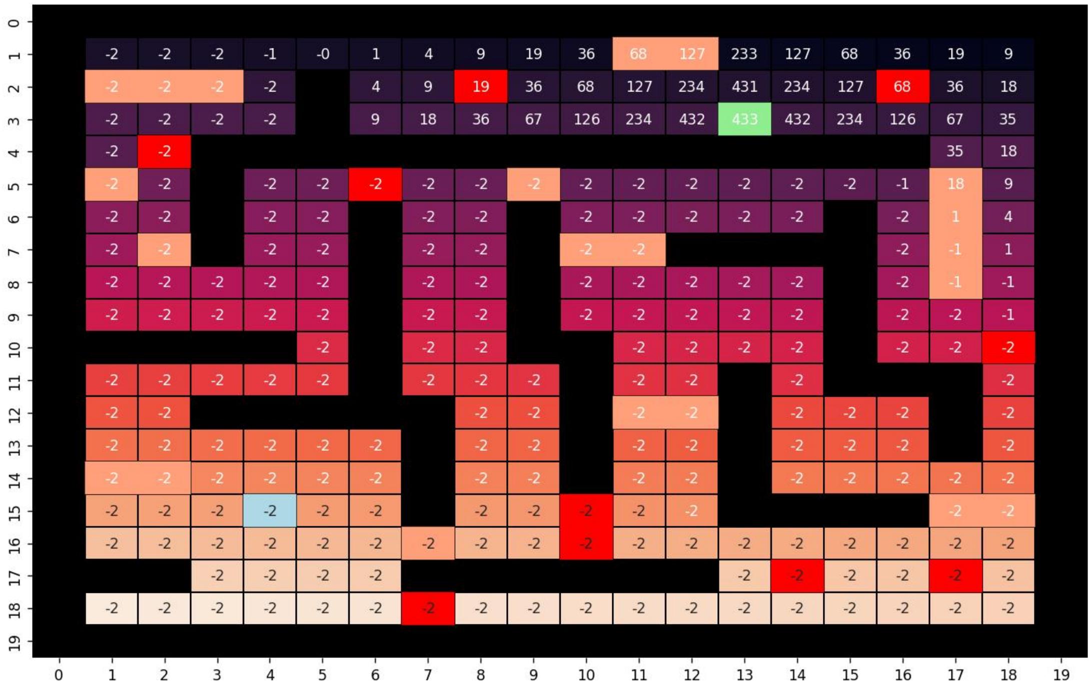
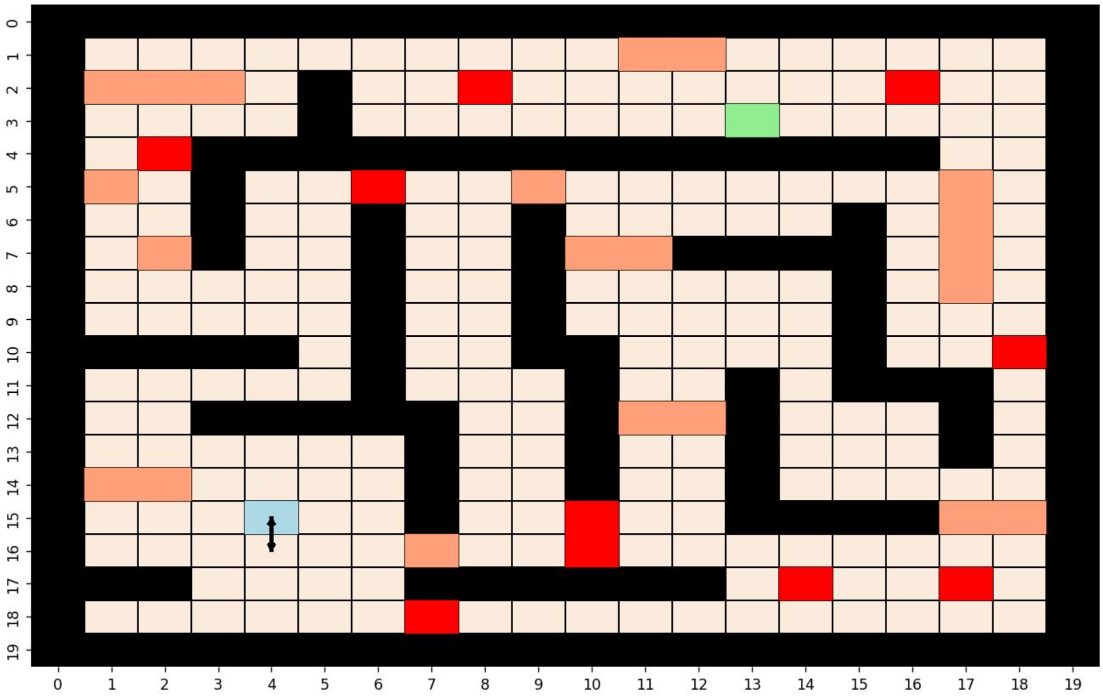


Figure 1.7 The optimal policy for each state for the base scenario under the Policy Iteration method.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1		-2	-2	-2	-1	-0	1	4	9	19	36	68	127	233	127	68	36	19	9	
2		-2	-2	-2	-2		4	9	19	36	68	127	234	431	234	127	68	36	18	
3		-2	-2	-2	-2		9	18	36	67	126	234	432	433	432	234	126	67	35	
4		-2	-2															35	18	
5		-2	-2		-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-1	18	9	
6		-2	-2		-2	-2		-2	-2		-2	-2	-2	-2	-2		-2	1	4	
7		-2	-2		-2	-2		-2	-2		-2	-2					-2	-1	1	
8		-2	-2	-2	-2	-2		-2	-2		-2	-2	-2	-2	-2		-2	-1	-1	
9		-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																				

Figure 1.8The optimal value function values for each state for the base scenario under the Policy Iteration method



.Figure 1.9The optimal path from the start state to the end state for the base scenario under the Policy Iteration method.

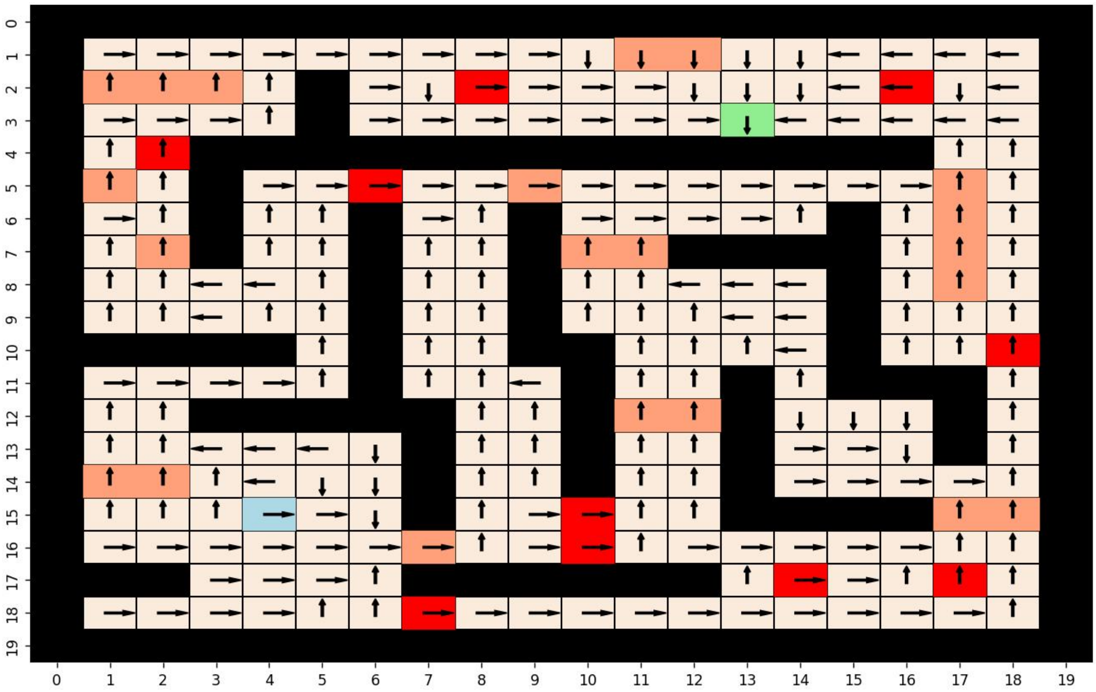
Question2.

Repeat the same process as part 1 with the vector-form Value Iteration method. Compare Value Iteration and Policy Iteration techniques for this problem

According to the different factors outcome, using Base Scenario factor and Large Stochasiticity Scenario factor in Value Iteration would take more iterations than Policy Iteration. Which because that Policy Iteration is better suited for situations where the optimal policy is needed quickly. So contract to the Policy Iteration, if the optimal policy is not required immediately, and MDP also has the large state space, value iteration would be better to use.

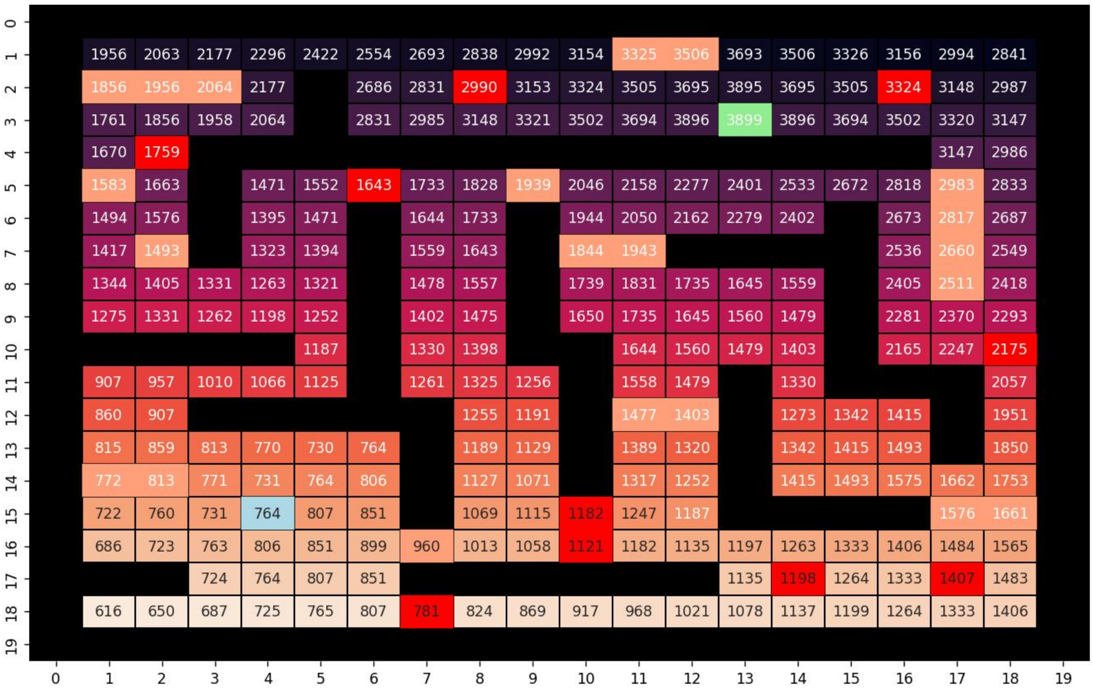
The plots of Policy Iteration values,actions and path are as follow:

a. P =0.02, γ =0.95， θ =0.01



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1		→	→	→	→	→	→	→	→	→	↓	↓	↓	↓	↓	←	←	←	←	
2		↑	↑	↑	↑		→	↓	→	→	→	→	↓	↓	↓	←	←	↓	←	
3		→	→	→	↑		→	→	→	→	→	→	→	↓	←	←	←	←	←	
4		↑	↑															↑	↑	
5		↑	↑		→	→	→	→	→	→	→	→	→	→	→	→	→	↑	↑	
6		→	↑		↑	↑		→	↑		→	→	→	→	↑		↑	↑	↑	
7		↑	↑		↑	↑		↑	↑		↑	↑					↑	↑	↑	
8		↑	↑	←	←	↑		↑	↑		↑	↑	←	←	←		↑	↑	↑	
9		↑	↑	←	↑	↑		↑	↑		↑	↑	↑	←	←		↑	↑	↑	
10						↑		↑	↑			↑	↑	↑	←		↑	↑	↑	
11		→	→	→	→	↑		↑	↑	←		↑	↑		↑				↑	
12		↑	↑						↑	↑		↑	↑		↓	↓	↓		↑	
13		↑	↑	←	←	←	↓		↑	↑		↑	↑		→	→	↓		↑	
14		↑	↑	↑	←	↓	↓		↑	↑		↑	↑		→	→	→	→	↑	
15		↑	↑	↑		→	↓		↑	→	→	↑	↑					↑	↑	
16		→	→	→	→	→	→	→	↑	→	→	↑	→	→	→	→	→	↑	↑	
17				→	→	→	↑							↑	→	→	↑	↑	↑	
18		→	→	→	→	↑	↑	→	→	→	→	→	→	→	→	→	→	→	↑	
19																				

Figure 2.1 The optimal policy for each state for the base scenario under the Value Iteration method



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1		1956	2063	2177	2296	2422	2554	2693	2838	2992	3154	3325	3506	3693	3506	3326	3156	2994	2841	
2		1856	1956	2064	2177		2686	2831	2990	3153	3324	3505	3695	3895	3695	3505	3324	3148	2987	
3		1761	1856	1958	2064		2831	2985	3148	3321	3502	3694	3896	3899	3896	3694	3502	3320	3147	
4		1670	1759															3147	2986	
5		1583	1663		1471	1552	1643	1733	1828	1939	2046	2158	2277	2401	2533	2672	2818	2983	2833	
6		1494	1576		1395	1471		1644	1733		1944	2050	2162	2279	2402		2673	2817	2687	
7		1417	1493		1323	1394		1559	1643		1844	1943					2536	2660	2549	
8		1344	1405	1331	1263	1321		1478	1557		1739	1831	1735	1645	1559		2405	2511	2418	
9		1275	1331	1262	1198	1252		1402	1475		1650	1735	1645	1560	1479		2281	2370	2293	
10						1187		1330	1398			1644	1560	1479	1403		2165	2247	2175	
11		907	957	1010	1066	1125		1261	1325	1256		1558	1479		1330				2057	
12		860	907						1255	1191		1477	1403		1273	1342	1415		1951	
13		815	859	813	770	730	764		1189	1129		1389	1320		1342	1415	1493		1850	
14		772	813	771	731	764	806		1127	1071		1317	1252		1415	1493	1575	1662	1753	
15		722	760	731	764	807	851		1069	1115	1182	1247	1187					1576	1661	
16		686	723	763	806	851	899	960	1013	1058	1121	1182	1135	1197	1263	1333	1406	1484	1565	
17				724	764	807	851							1135	1198	1264	1333	1407	1483	
18		616	650	687	725	765	807	781	824	869	917	968	1021	1078	1137	1199	1264	1333	1406	
19																				

Figure 2.2 The optimal value function values for each state for the base scenario under the Value Iteration method

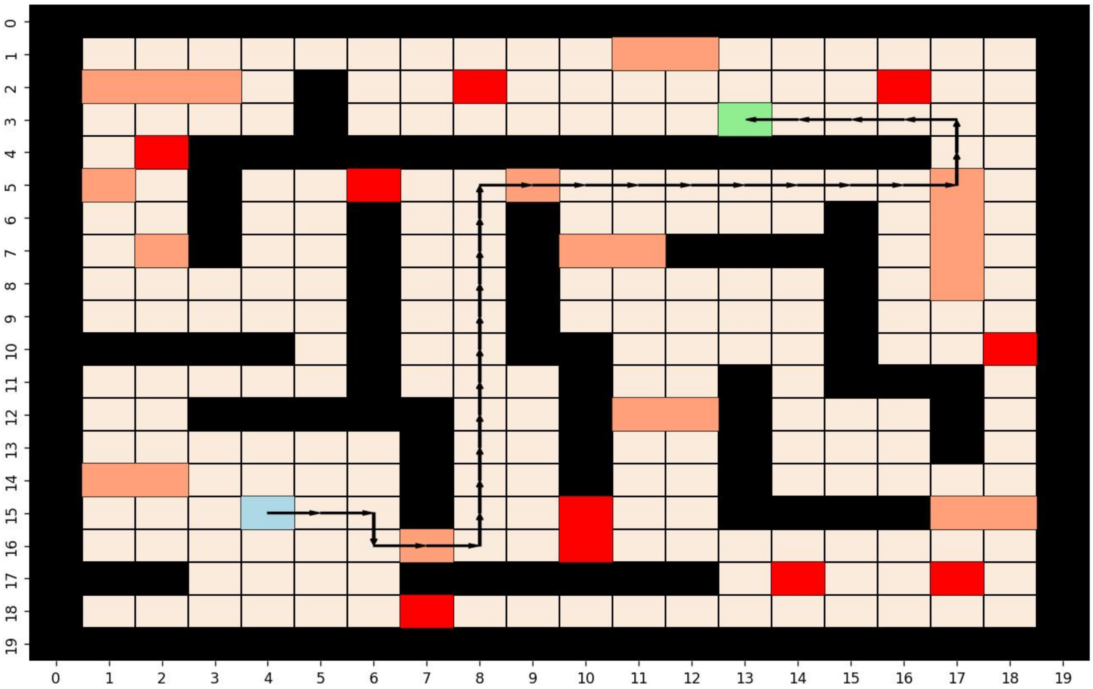


Figure 2.3 The optimal path from the start state to the end state for the base scenario under the Value Iteration method.

b. P =0.5, γ =0.95， θ =0.01

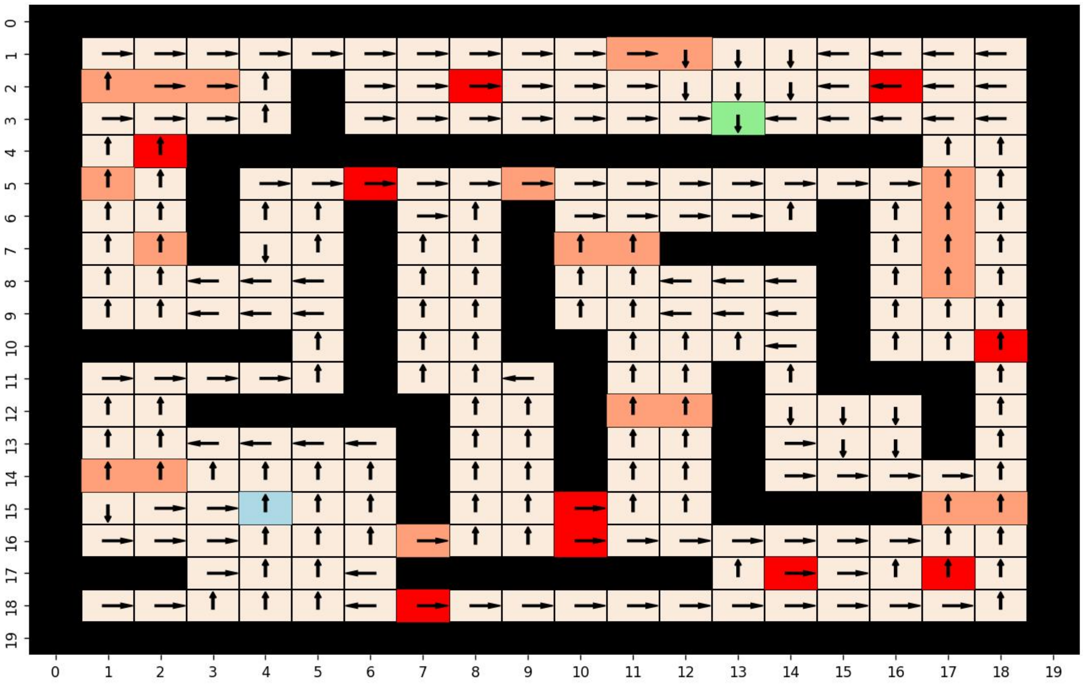
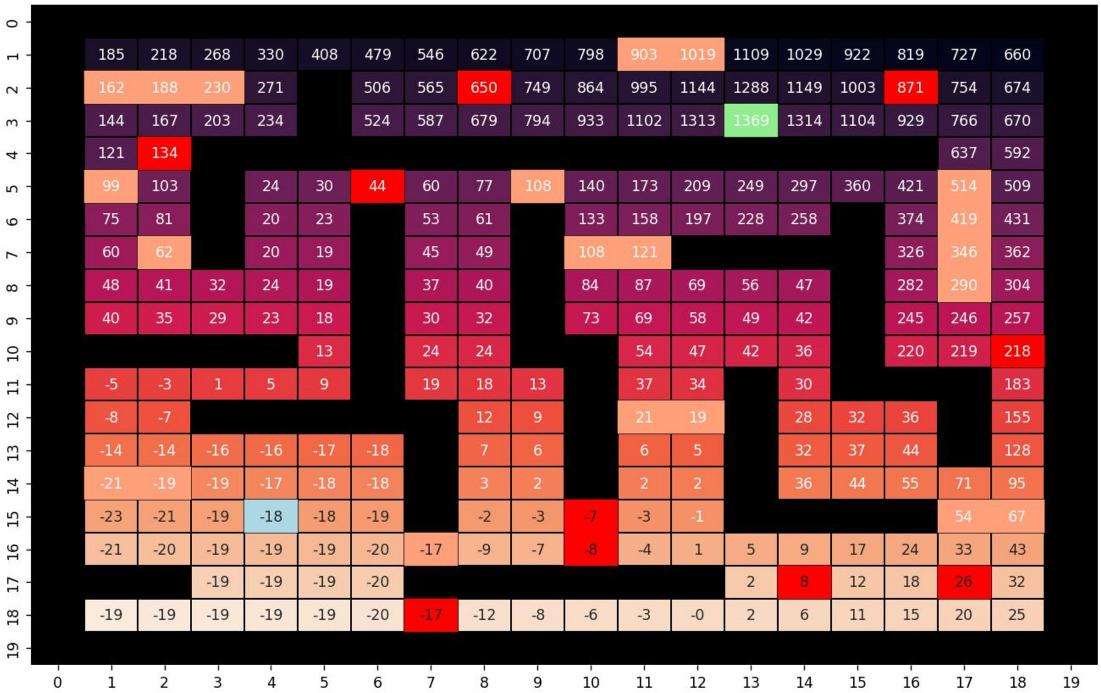


Figure 2.4 The optimal policy for each state for the base scenario under the Value Iteration method



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1		185	218	268	330	408	479	546	622	707	798	903	1019	1109	1029	922	819	727	660	
2		162	188	230	271		506	565	650	749	864	995	1144	1288	1149	1003	871	754	674	
3		144	167	203	234		524	587	679	794	933	1102	1313	1369	1314	1104	929	766	670	
4		121	134															637	592	
5		99	103		24	30	44	60	77	108	140	173	209	249	297	360	421	514	509	
6		75	81		20	23		53	61		133	158	197	228	258		374	419	431	
7		60	62		20	19		45	49		108	121					326	346	362	
8		48	41	32	24	19		37	40		84	87	69	56	47		282	290	304	
9		40	35	29	23	18		30	32		73	69	58	49	42		245	246	257	
10						13		24	24			54	47	42	36		220	219	218	
11		-5	-3	1	5	9		19	18	13		37	34		30				183	
12		-8	-7						12	9		21	19		28	32	36		155	
13		-14	-14	-16	-16	-17	-18		7	6		6	5		32	37	44		128	
14		-21	-19	-19	-17	-18	-18		3	2		2	2		36	44	55	71	95	
15		-23	-21	-19	-18	-18	-19		-2	-3	-7	-3	-1					54	67	
16		-21	-20	-19	-19	-19	-20	-17	-9	-7	-8	-4	1	5	9	17	24	33	43	
17				-19	-19	-19	-20							2	8	12	18	26	32	
18		-19	-19	-19	-19	-19	-20	-17	-12	-8	-6	-3	-0	2	6	11	15	20	25	
19																				

Figure 2.5 The optimal value function values for each state for the base scenario under the Value Iteration method

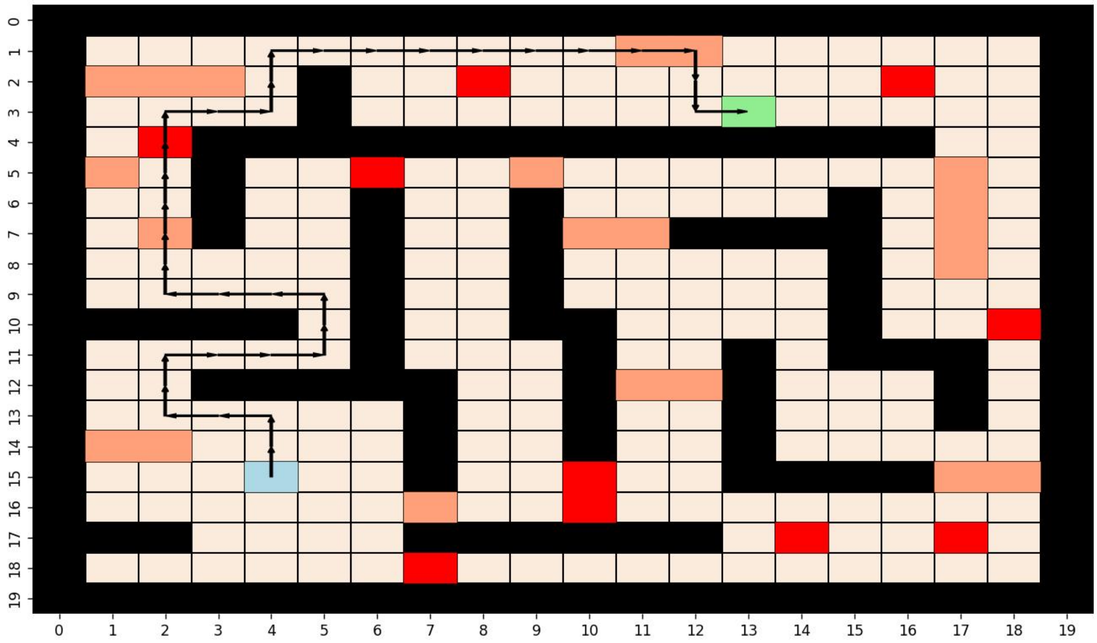


Figure 2.6 The optimal path from the start state to the end state for the base scenario under the Value Iteration method.

c. P =0.02, γ =0.55， θ =0.01

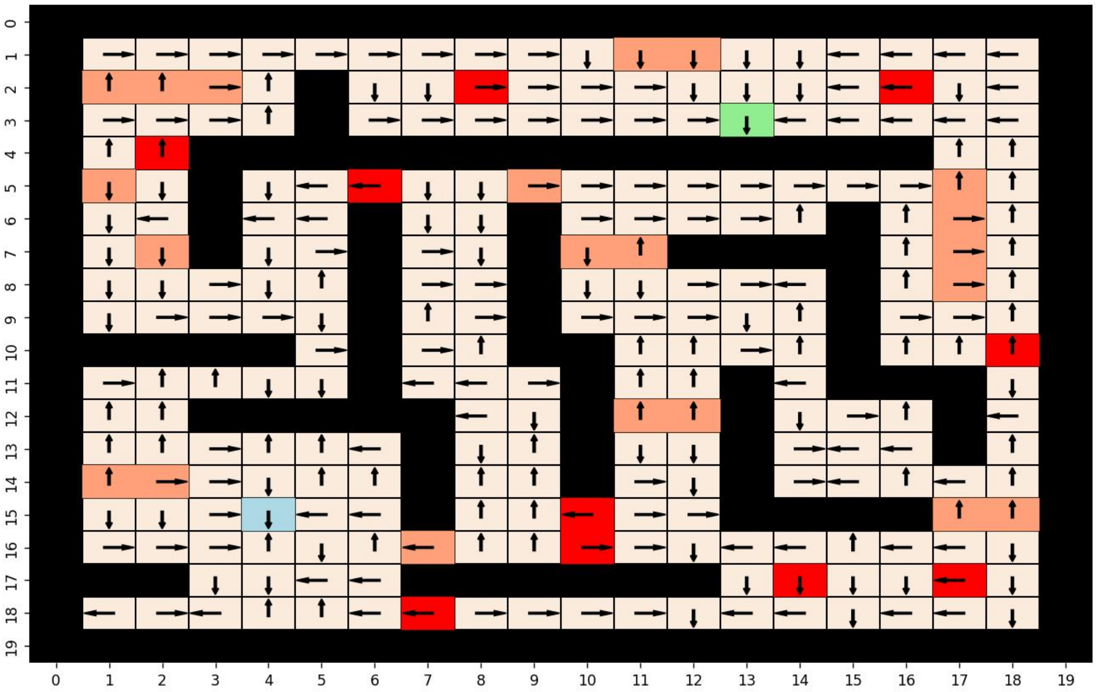


Figure 2.7 The optimal policy for each state for the base scenario under the Value Iteration method

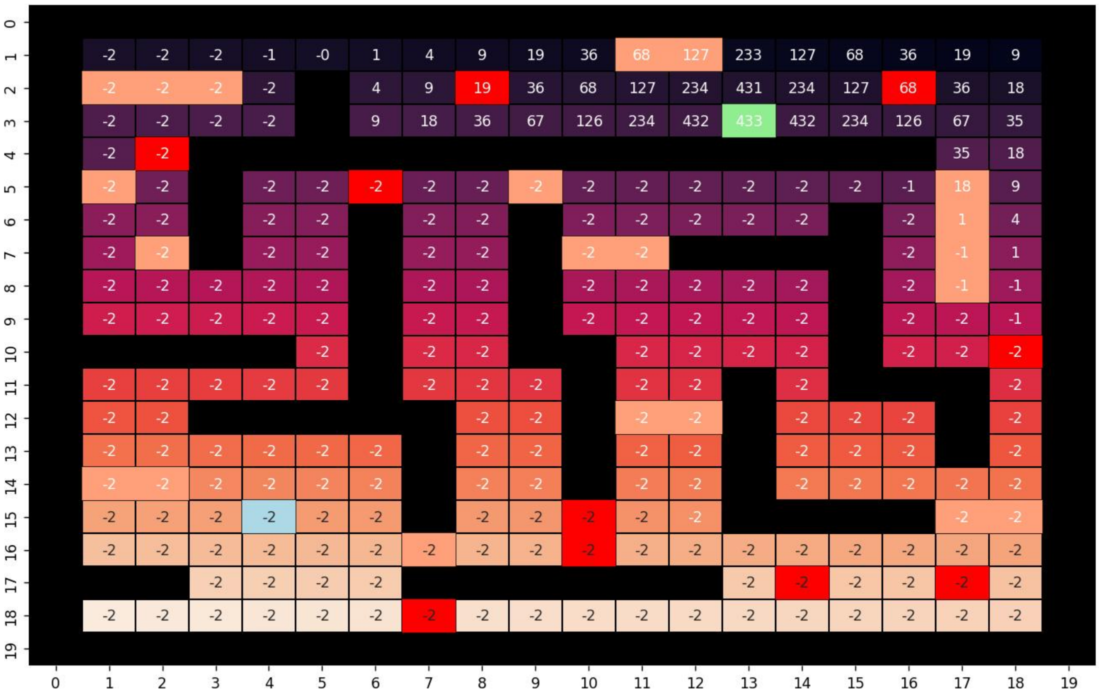


Figure 2.8 The optimal value function values for each state for the base scenario under the Value Iteration method

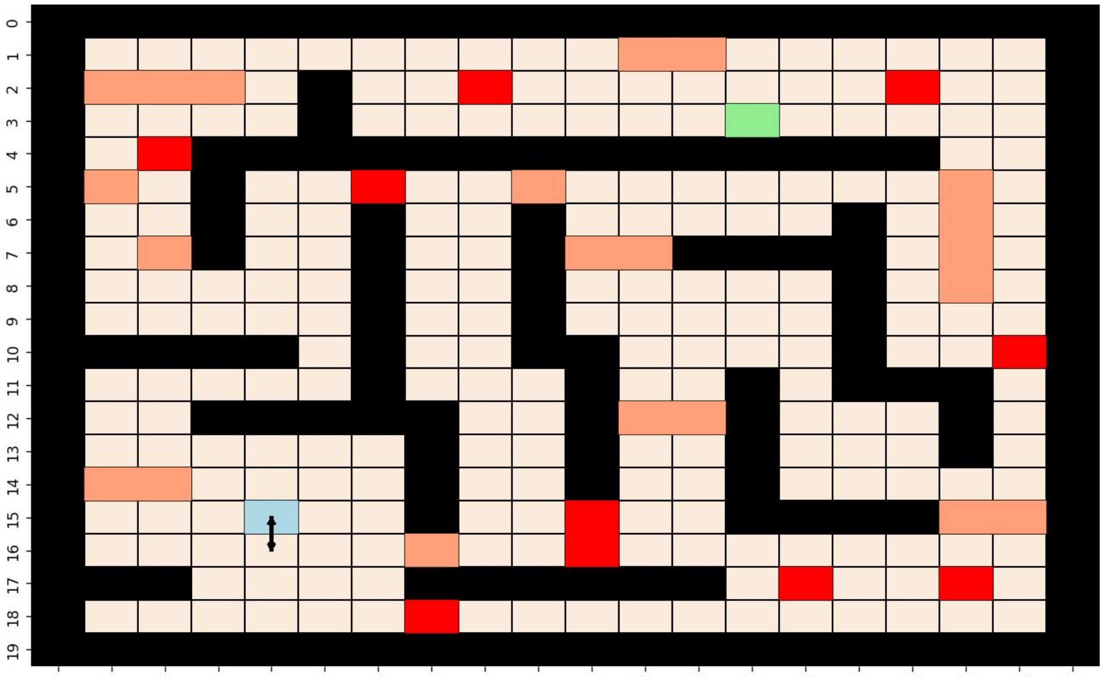


Figure 2.9 The optimal path from the start state to the end state for the base scenario under the Value Iteration method.

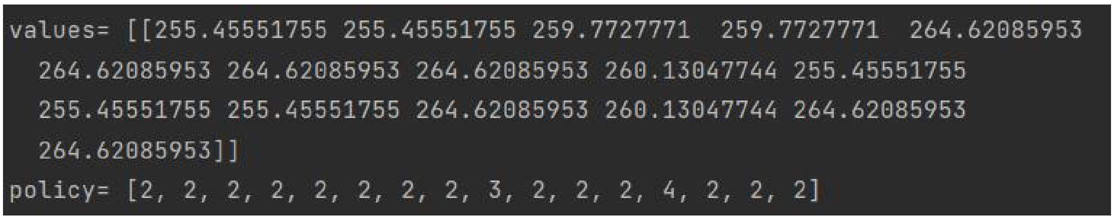
Problem 2

Part A

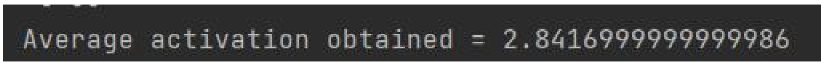
For p=0.05 use matrix-form Value Iteration method for computing the optimal control policy π*,which is a vector of size 16, specifying the best action at any given state. Use all-zero initial state values and θ =0.01.Compare the AvgA under the obtained optimal control policy with AvgA under no control policy(taking a-in all states)

.

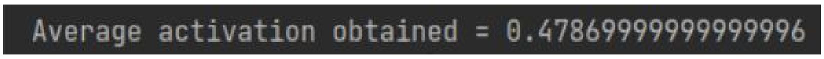
p=0.05



```
values= [[255.45551755 255.45551755 259.7727771  259.7727771  264.62085953
  264.62085953 264.62085953 264.62085953 260.13047744 255.45551755
  255.45551755 255.45551755 264.62085953 260.13047744 264.62085953
  264.62085953]]
policy= [2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 4, 2, 2, 2]
```



```
 Average activation obtained = 2.8416999999999986
```



```
 Average activation obtained = 0.47869999999999996
```

Under the obtained optimal control policy : AvgA = 2.84

Under no control policy control policy : AvgA =0.48

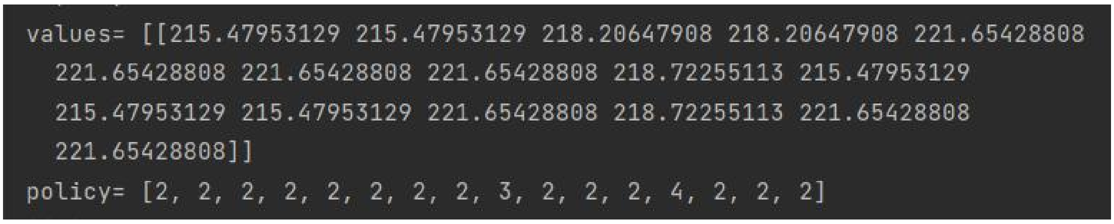
When p=0.05, the optimal control policy $\pi^*=[a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^4,a^3,a^3,a^3,a^5,a^3,a^3,a^3]^T$

Compare to the AvgA under no control policy, the AvgA optimal control policy is greater, which is because In matrix value iteration, the optimal control policy is obtained by finding the policy that maximizes the value function at each state. Since the optimal control policy maximizes the expected total reward at each state, it follows that the AvgA under the optimal control policy should be greater.

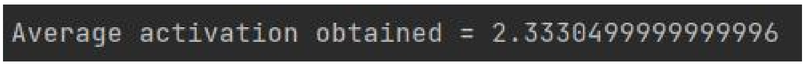
Part B

Repeat part(a) for p = 0.2 and p = 0.45.What is your observation? Compare the obtained optimal policy and AvgA for two noise parameters with the results of part(a).

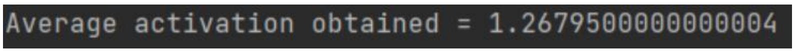
p = 0.2



```
values= [[215.47953129 215.47953129 218.20647908 218.20647908 221.65428808
  221.65428808 221.65428808 221.65428808 218.72255113 215.47953129
  215.47953129 215.47953129 221.65428808 218.72255113 221.65428808
  221.65428808]]
policy= [2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 4, 2, 2, 2]
```



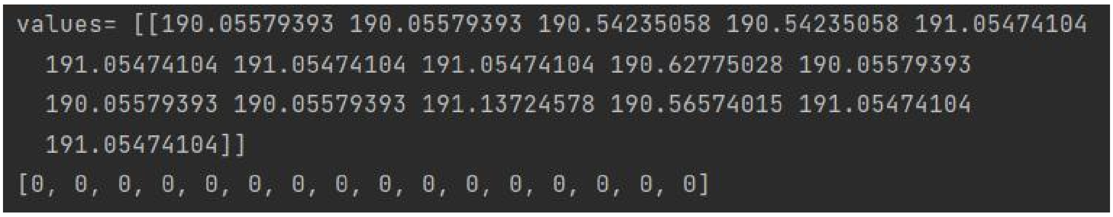
```
Average activation obtained = 2.3330499999999996
```



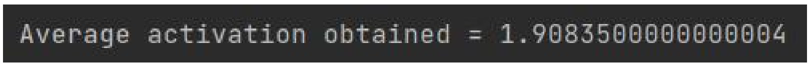
```
Average activation obtained = 1.2679500000000004
```

When p = 0.2, the optimal control policy $\pi^*=[a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^4,a^3,a^3,a^3,a^5,a^3,a^3,a^3]^T$
Under the obtained optimal control policy : AvgA =2.33
Under no control policy control policy : AvgA =1.26

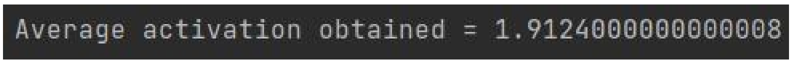
p = 0.45



```
values= [[190.05579393 190.05579393 190.54235058 190.54235058 191.05474104
  191.05474104 191.05474104 191.05474104 190.62775028 190.05579393
  190.05579393 190.05579393 191.13724578 190.56574015 191.05474104
  191.05474104]]
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```



```
Average activation obtained = 1.9083500000000004
```



```
Average activation obtained = 1.9124000000000008
```

When p = 0.45, the optimal control policy $\pi^*=[a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1,a^1]^T$
'Under the obtained optimal control policy : AvgA = 1.91
Under no control policy control policy : AvgA = 1.91

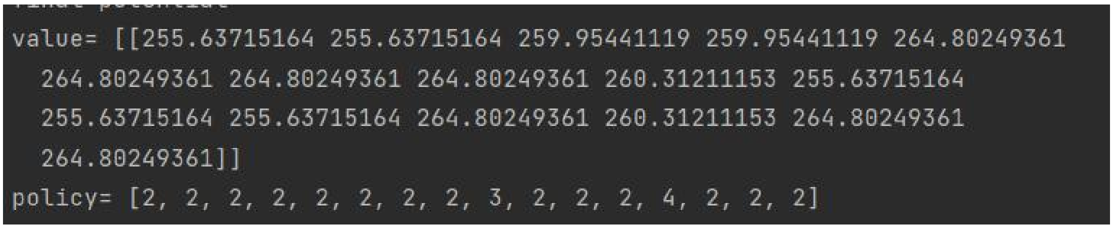
When p=0.2 or p= 0.42, the Average Activation is lower than part (a), since with the increase of probability, the noise is also increase,it might result in the selection of random action increase, so in this situation, the agent can hardly make choice to gain the greater reward.

Part C

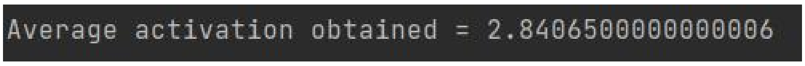
Similar to part(a) use p=0.05,an action a^1 in all states as the initial policy. Perform matrix- form

Policy Iteration method for computing the optimal control policy π^*. Compare the results with part (a). Are the obtained policies the same?

When p = 0.05, initial policy = $a^{1,}$,using policy iteration:



```
value= [[255.63715164 255.63715164 259.95441119 259.95441119 264.80249361
  264.80249361 264.80249361 264.80249361 260.31211153 255.63715164
  255.63715164 255.63715164 264.80249361 260.31211153 264.80249361
  264.80249361]]
policy= [2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 4, 2, 2, 2]
```

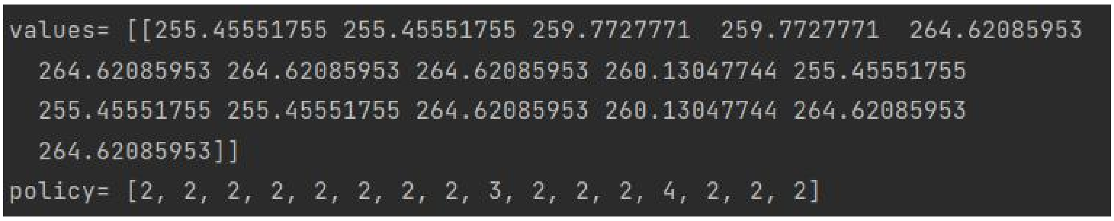


```
Average activation obtained = 2.8406500000000006
```

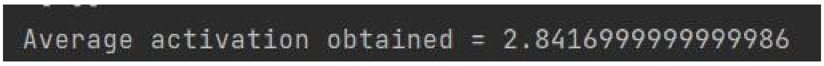
The optimal control policy $\pi^*=[a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^4,a^3,a^3,a^3,a^5,a^3,a^3,a^3]^T$

'Under the obtained optimal control policy : AvgA = 2.84

When p=0.05,using value iteration:



```
values= [[255.45551755 255.45551755 259.7727771  259.7727771  264.62085953
  264.62085953 264.62085953 264.62085953 260.13047744 255.45551755
  255.45551755 255.45551755 264.62085953 260.13047744 264.62085953
  264.62085953]]
policy= [2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 4, 2, 2, 2]
```



```
 Average activation obtained = 2.8416999999999986
```

The optimal control policy $\pi^*=[a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^3,a^4,a^3,a^3,a^3,a^5,a^3,a^3,a^3]^T$

'Under the obtained optimal control policy : AvgA = 2.84

The Policy iteration has the same policies with the Value Iteration when p=0.05. Compare the iterations with Policy iteration and Value iteration, the Policy Iteration has the less steps.