Intelligent Agents CS 533: Homework-3 Report

Vivswan Shitole and Rogen George

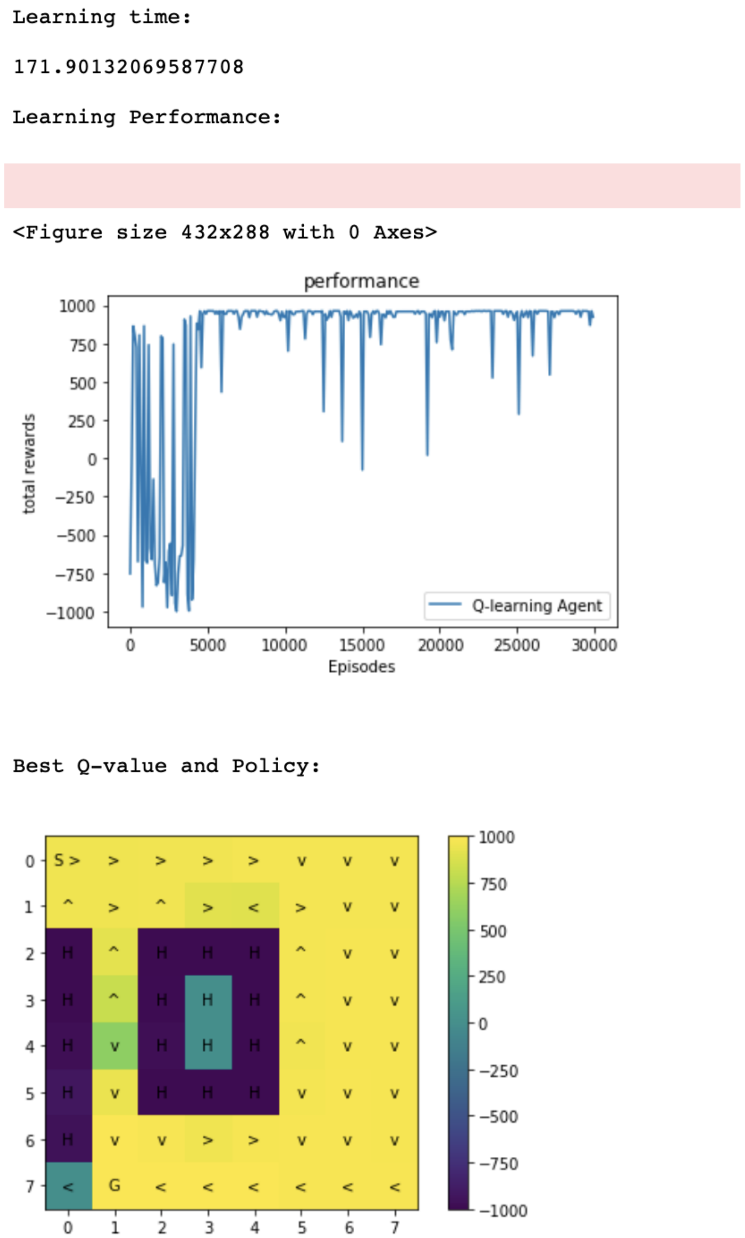
Oregon State University

1 Learning curves and Final Q-tables

1.1 Dangerous Halway: Non-Distributed

Experiment 1 Parameters used:

- epsilon = 0.3
- learning rate = 0.1
- learning episodes = 30000
- test interval = 100
- do test = True

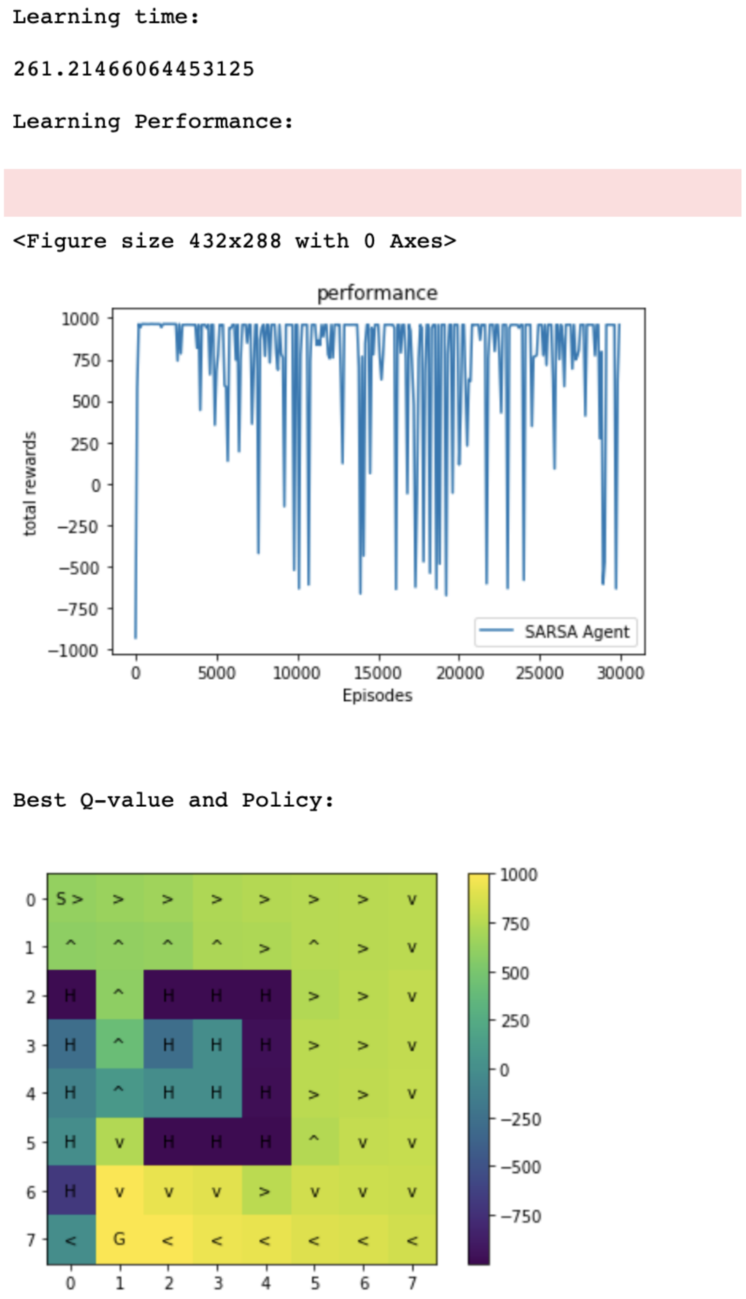


```
Learning time:

171.90132069587708

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

	0	1	2	3	4	5	6	7
0	S >	>	>	>	>	v	v	v
1	^	>	^	>	<	>	v	v
2	H	^	H	H	H	^	v	v
3	H	^	H	H	H	^	v	v
4	H	v	H	H	H	^	v	v
5	H	v	H	H	H	v	v	v
6	H	v	v	>	>	v	v	v
7	<	G	<	<	<	<	<	<

Fig. 1: Q learning.



```
Learning time:

261.21466064453125

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

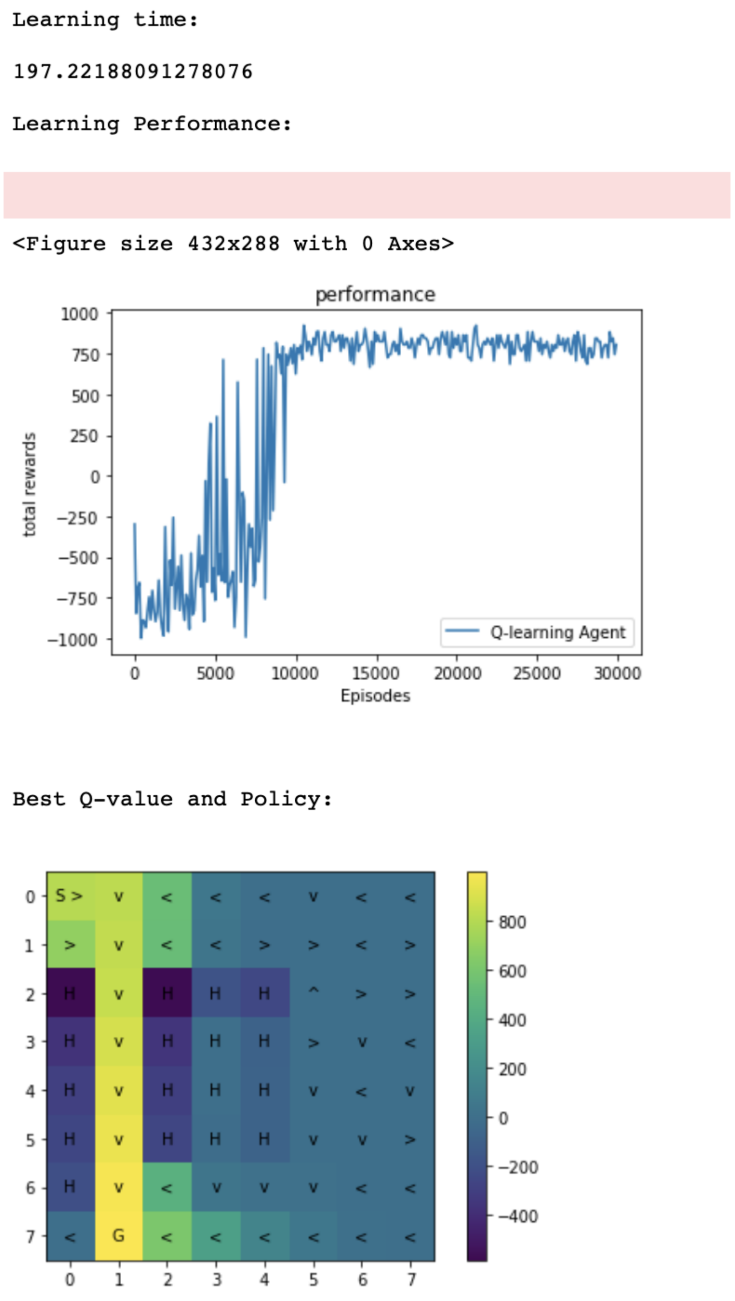
```
Best Q-value and Policy:
```

	0	1	2	3	4	5	6	7
0	S >	>	>	>	>	>	>	v
1	^	^	^	^	>	^	>	v
2	H	^	H	H	H	>	>	v
3	H	^	H	H	H	>	>	v
4	H	^	H	H	H	>	>	v
5	H	v	H	H	H	^	v	v
6	H	v	v	v	>	v	v	v
7	<	G	<	<	<	<	<	<

Fig. 2: SARSA.

Experiment 2 Parameters used:

- epsilon = 0.3
- learning rate = 0.001
- learning episodes = 30000
- test interval = 100
- do test = True

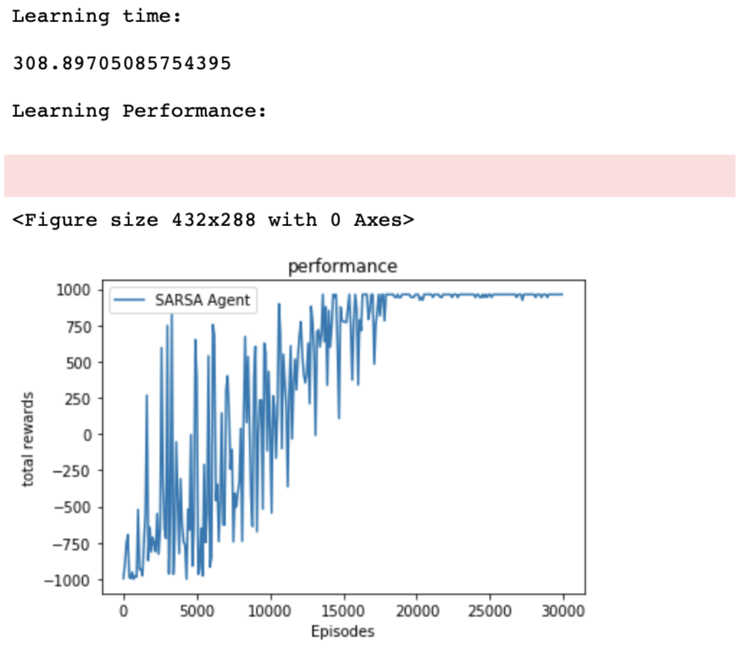


```
Learning time:

197.22188091278076

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

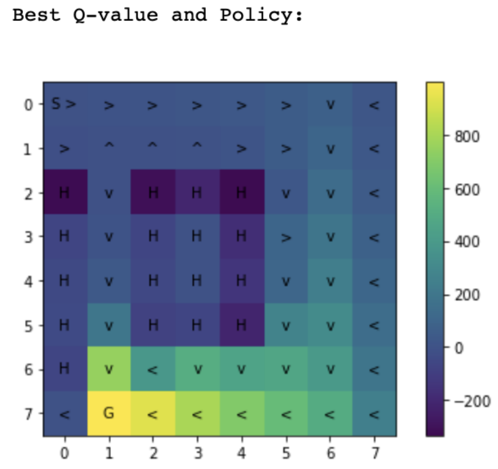
```
Best Q-value and Policy:
```

	0	1	2	3	4	5	6	7
0	S >	v	<	<	<	v	<	<
1	>	v	<	<	>	>	<	>
2	H	v	H	H	H	^	>	>
3	H	v	H	H	H	>	v	<
4	H	v	H	H	H	v	<	v
5	H	v	H	H	H	v	v	>
6	H	v	<	v	v	v	<	<
7	<	G	<	<	<	<	<	<

Fig. 3: Q learning.



```
Learning time:

308.89705085754395

Learning Performance:

<Figure size 432x288 with 0 Axes>
```



```
Best Q-value and Policy:
```

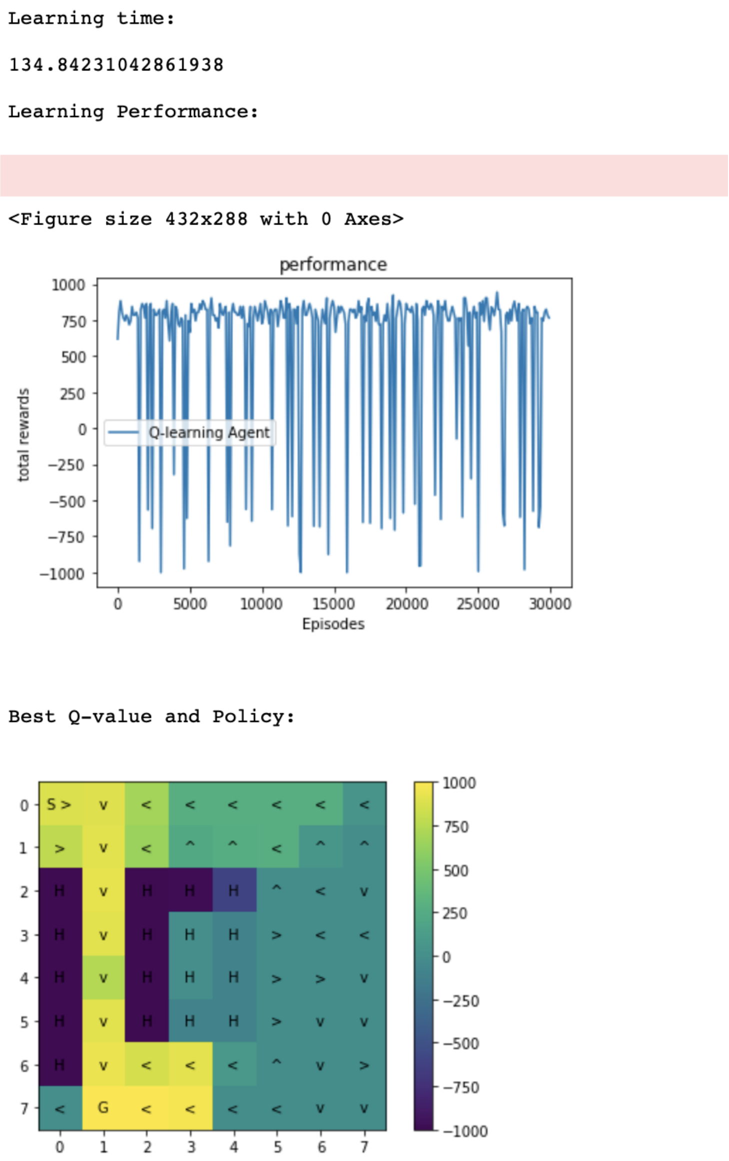
	0	1	2	3	4	5	6	7
0	S >	>	>	>	>	>	v	<
1	>	^	^	^	>	>	v	<
2	H	v	H	H	H	v	v	<
3	H	v	H	H	H	>	v	<
4	H	v	H	H	H	v	v	<
5	H	v	H	H	H	v	v	<
6	H	v	<	v	v	v	v	<
7	<	G	<	<	<	<	<	<

Fig. 4: SARSA.

Experiment 3 Parameters used:

- epsilon = 0.05
- learning rate = 0.1
- learning episodes = 30000

- test interval = 100
- do test = True

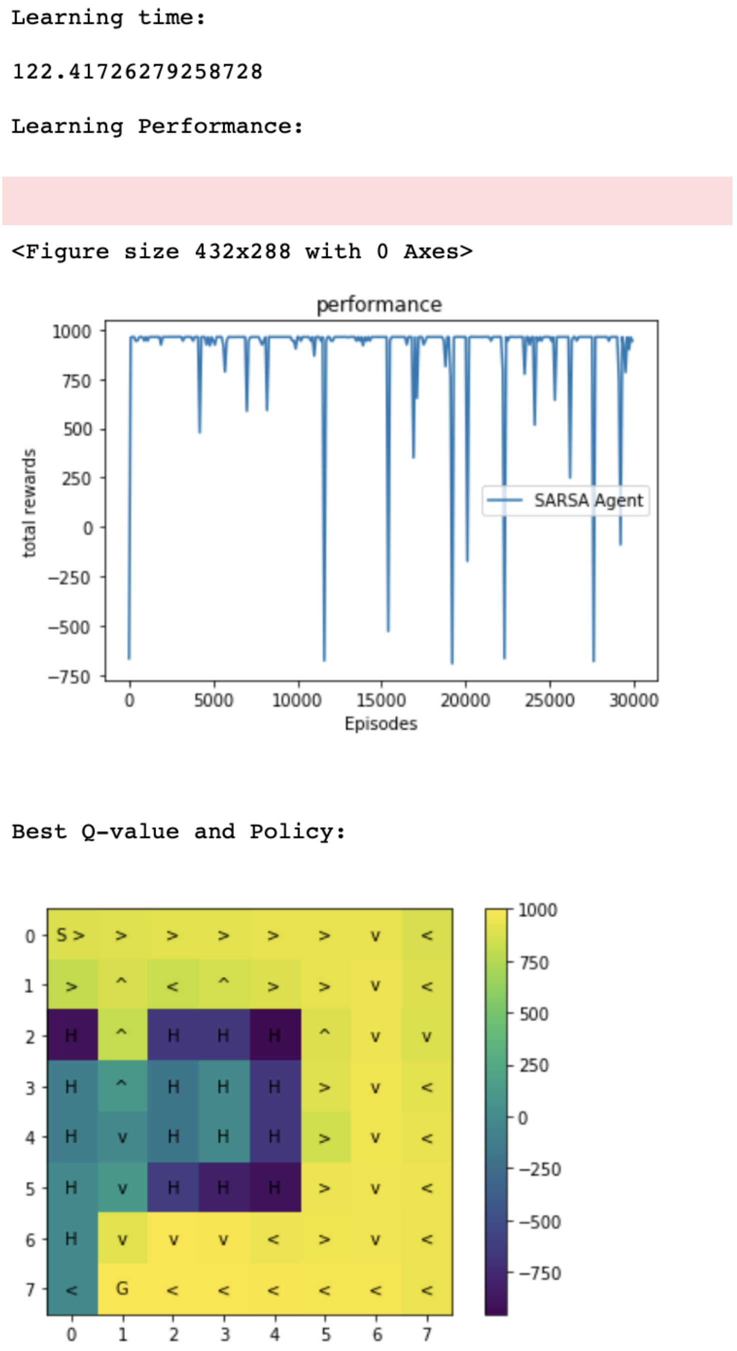


```
Learning time:

134.84231042861938

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

	0	1	2	3	4	5	6	7
0	S >	v	<	<	<	<	<	<
1	>	v	<	^	^	<	^	^
2	H	v	H	H	H	^	<	v
3	H	v	H	H	H	>	<	<
4	H	v	H	H	H	>	>	v
5	H	v	H	H	H	>	v	v
6	H	v	<	<	<	^	v	>
7	<	G	<	<	<	<	v	v

Fig. 5: Q learning.



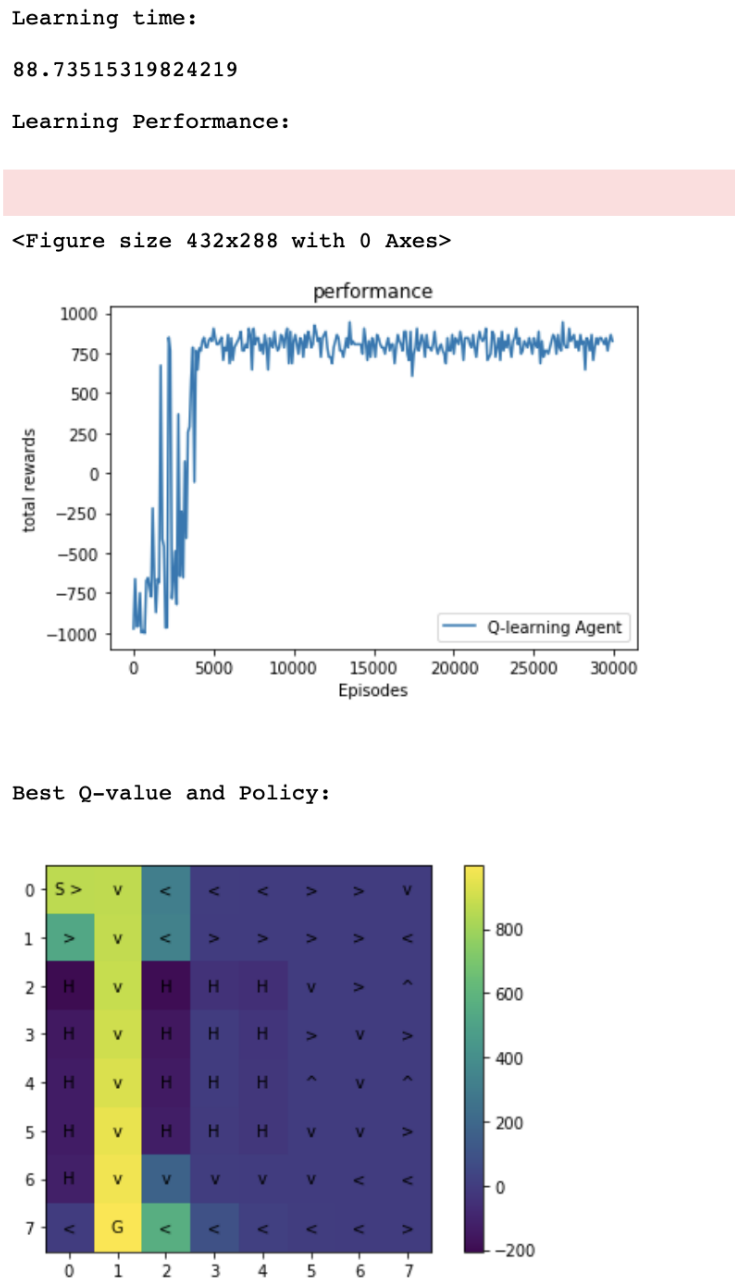
```
Learning time:

122.41726279258728

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

Fig. 6: SARSA.

Experiment 4 Parameters used:

- epsilon = 0.05
- learning rate = 0.001
- learning episodes = 30000
- test interval = 100
- do test = True



```
Learning time:

88.73515319824219

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

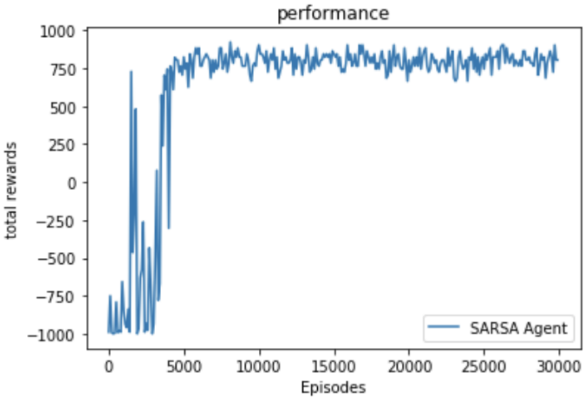
```
Best Q-value and Policy:
```

	0	1	2	3	4	5	6	7
0	S >	v	<	<	<	>	>	v
1	>	v	<	>	>	>	>	<
2	H	v	H	H	H	v	>	^
3	H	v	H	H	H	>	v	>
4	H	v	H	H	H	^	v	^
5	H	v	H	H	H	v	v	>
6	H	v	v	v	v	v	<	<
7	<	G	<	<	<	<	<	>

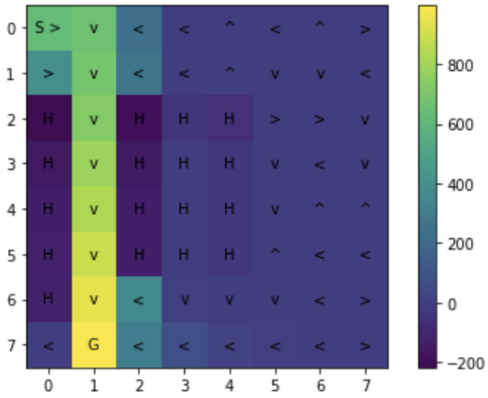
Fig. 7: Q learning.

```
Learning time:

93.19903635978699

Learning Performance:
```

```
<Figure size 432x288 with 0 Axes>
```



```
Best Q-value and Policy:
```



	0	1	2	3	4	5	6	7
0	S >	v	<	<	^	<	^	>
1	>	v	<	<	^	v	v	<
2	H	v	H	H	H	>	>	v
3	H	v	H	H	H	v	<	v
4	H	v	H	H	H	v	^	^
5	H	v	H	H	H	^	<	<
6	H	v	<	v	v	v	<	>
7	<	G	<	<	<	<	<	>

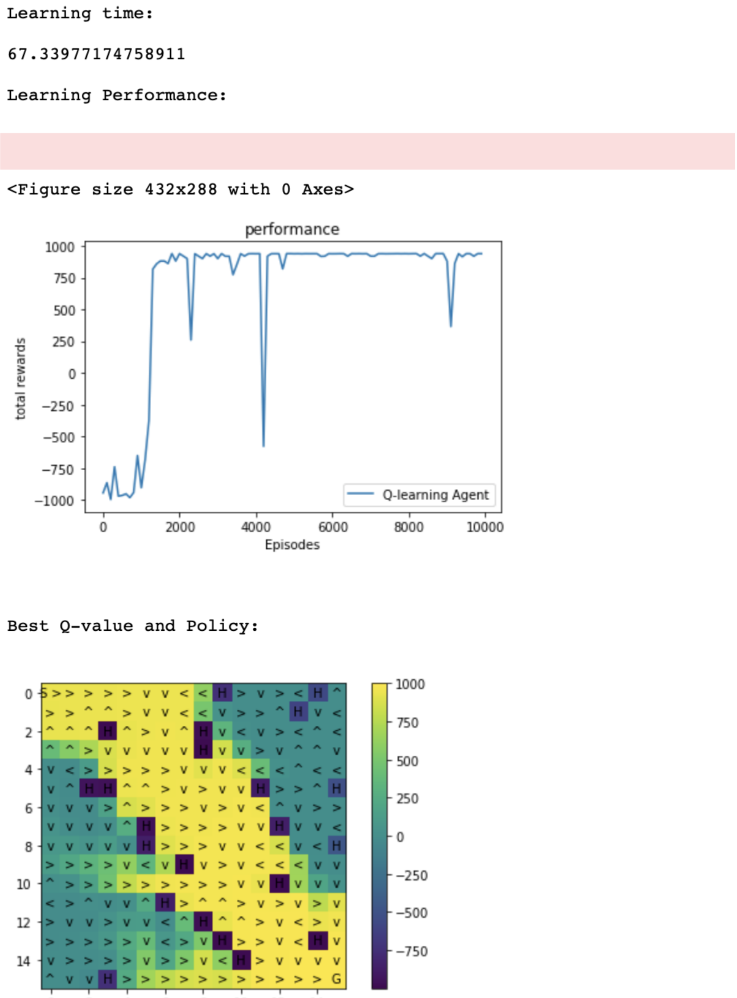
Fig. 8: SARSA.

1.2 Map 16: Non-Distributed

Experiment 1 Parameters used:

- epsilon = 0.3

- learning rate = 0.1
- learning episodes = 100000
- test interval = 100
- do test = True

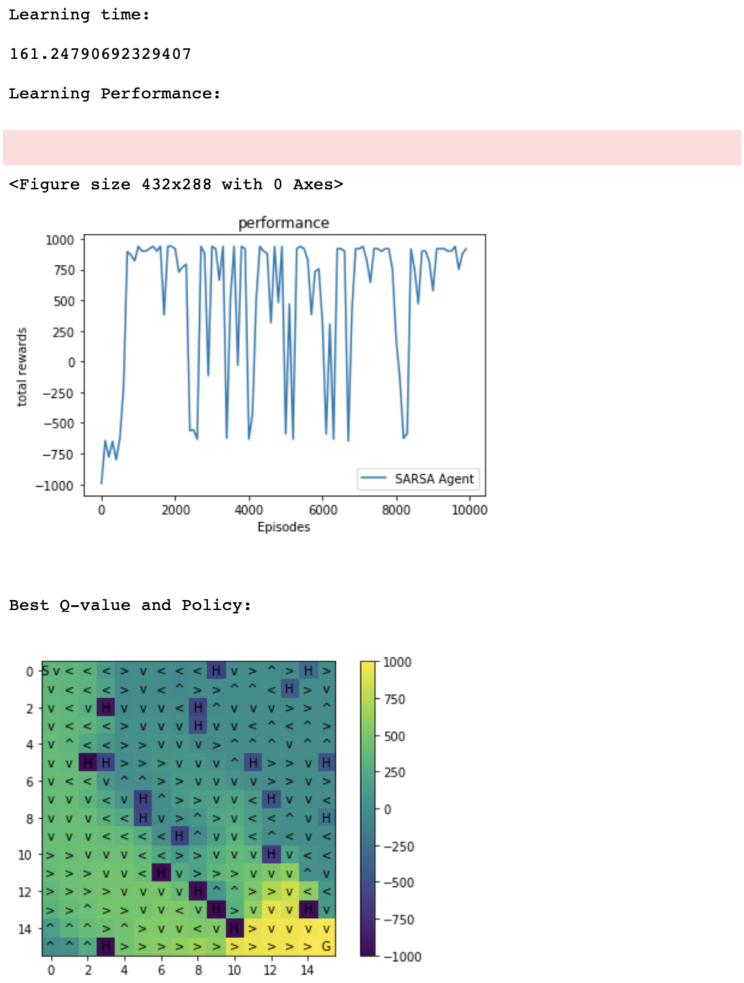


```
Learning time:

67.33977174758911

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

```
S> >  >  >  >  v  v  <  <  H  >  v  >  <  H  ^
>  >  ^  ^  >  v  v  <  <  v  >  >  ^  H  v  <
^  ^  ^  H  ^  >  v  ^  H  v  <  v  >  <  ^  <
^  ^  >  v  v  v  v  v  H  v  v  >  v  ^  ^  v
v  <  >  >  >  >  >  v  v  v  <  <  <  ^  <  <
v  ^  H  H  ^  ^  >  v  >  v  v  H  >  >  ^  H
v  v  v  >  ^  >  >  >  v  >  v  <  ^  v  >  >
v  v  v  v  ^  H  >  >  >  >  v  v  H  v  v  <
v  v  v  v  v  H  >  >  >  v  v  <  <  v  <  H
>  >  >  >  v  <  v  H  v  >  v  <  <  <  v  v
^  >  >  >  >  >  >  >  >  >  v  v  H  v  v  v
<  >  ^  v  v  ^  H  >  ^  ^  >  v  >  v  >  v
>  v  v  >  v  v  <  ^  H  ^  ^  >  v  <  >  v
>  >  >  >  >  v  <  >  v  H  >  >  v  <  H  v
v  >  >  >  >  v  >  >  v  <  H  >  v  v  v  v
^  v  v  H  >  >  >  >  >  >  >  >  >  >  >  G
```

Fig. 9: Q learning.



```
Learning time:

161.24790692329407

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

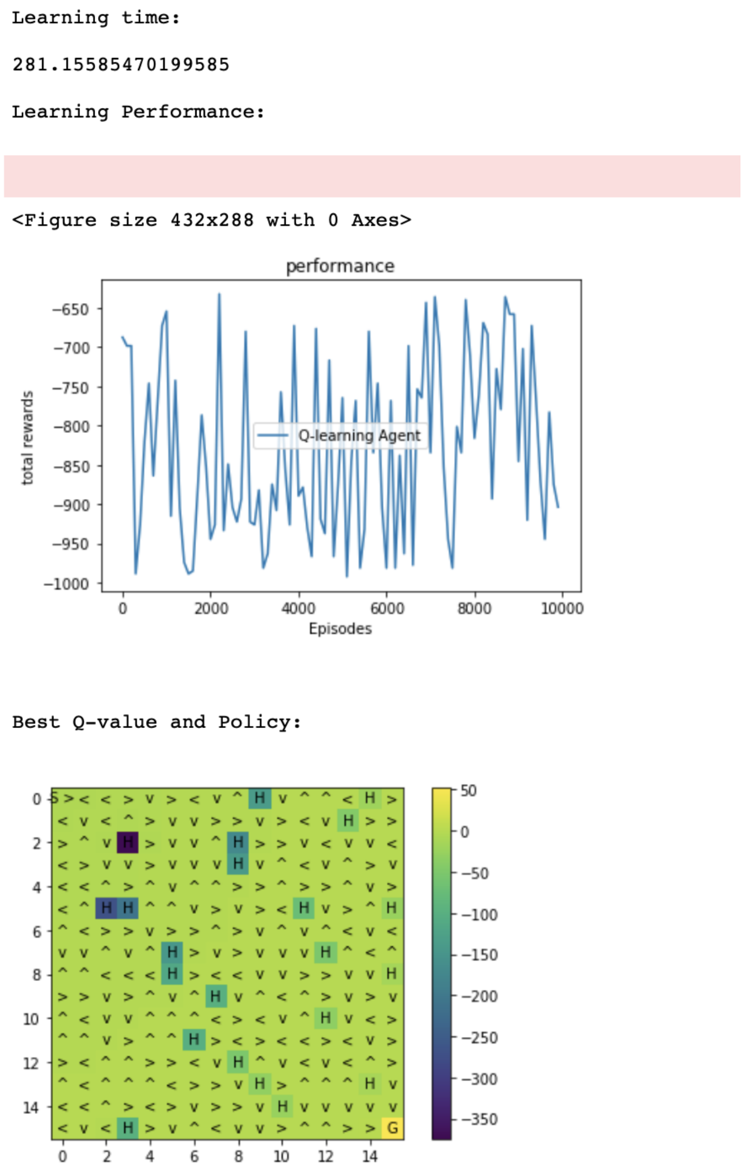
```
Best Q-value and Policy:
```

```
Sv <  <  <  >  v  <  <  <  H  v  >  ^  >  H  >
v  <  <  <  >  v  <  ^  >  >  ^  ^  <  H  >  v
v  <  v  H  v  v  v  <  H  ^  v  v  v  >  >  ^
v  <  <  <  >  v  v  v  H  v  v  <  ^  <  ^  >
v  ^  <  <  >  >  v  v  v  >  ^  ^  ^  v  ^  ^
v  v  H  H  >  >  >  v  v  v  ^  H  >  >  v  H
v  <  <  v  ^  ^  >  >  v  v  v  v  >  >  v  >
v  v  v  <  v  H  ^  >  >  v  v  <  H  v  v  <
v  v  v  <  <  H  v  >  ^  >  v  <  <  ^  v  H
v  v  v  <  <  <  <  H  ^  v  v  <  ^  <  v  <
>  >  v  v  v  <  <  >  >  v  v  v  H  v  <  <
>  >  >  v  v  <  H  v  >  >  >  v  v  v  ^  v
>  >  >  >  v  v  v  v  H  ^  ^  >  >  v  <  <
>  >  ^  >  >  v  v  <  v  H  >  v  v  v  H  v
^  ^  ^  >  ^  >  v  v  <  v  H  >  v  v  v  v
^  ^  ^  H  >  >  >  >  >  >  >  >  >  >  >  G
```

Fig. 10: SARSA.

Experiment 2 Parameters used:

- epsilon = 0.3
- learning rate = 0.001
- learning episodes = 100000
- test interval = 100
- do test = True

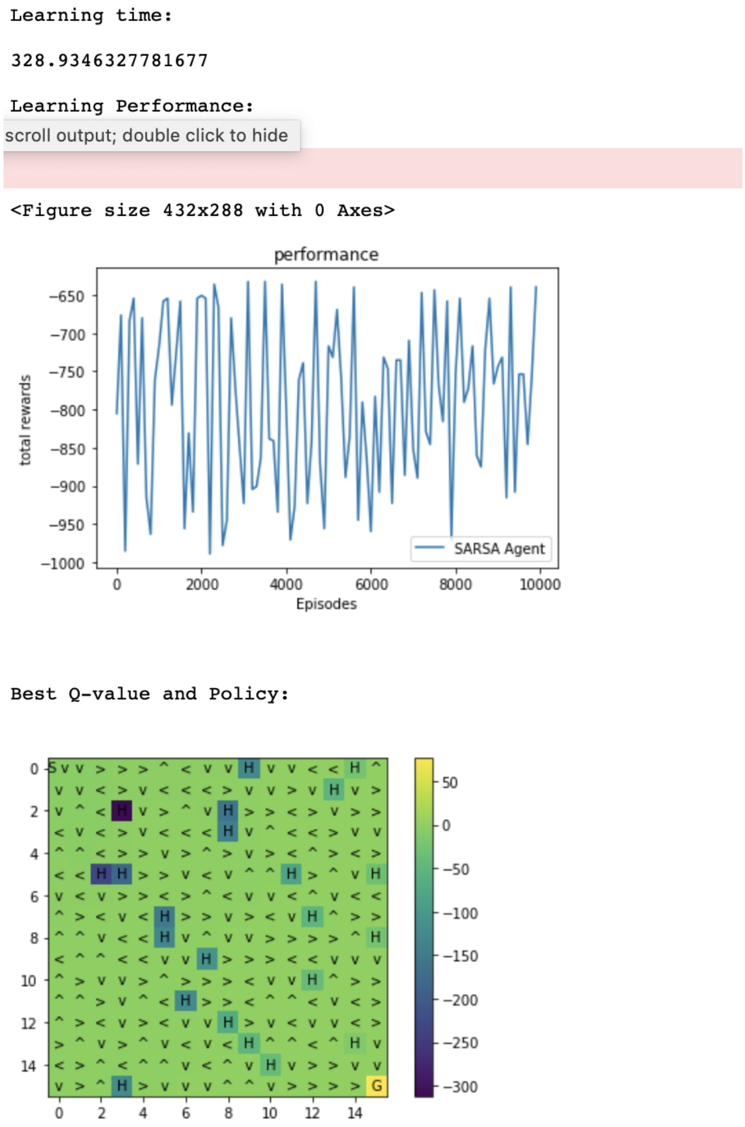


```
Learning time:

281.15585470199585

Learning Performance:


<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

```
S> < < > v > < v ^ H v ^ ^ < H >
<  v < ^ > v v > > v > < v H > >
>  ^ v H > v v ^ H > > v < v v <
<  > v v > v v v H v ^ < v ^ > v
<  < ^ > ^ v ^ ^ > > ^ > > ^ v >
<  ^ H H ^ ^ v > v > < H v > ^ H
^  < > > v > > ^ > v ^ v ^ < v <
v  v ^ v ^ H > v > v v v H ^ < ^
^  ^ < < < H > < < v v > > v v H
>  > v > ^ v ^ H v ^ < ^ > v > v
^  < v v ^ ^ ^ < > < v ^ H v < >
^  ^ v > ^ ^ H > < > < < > < v >
>  < ^ ^ > > < v H ^ v < v < ^ >
^  < ^ ^ ^ < > > v H > ^ ^ ^ H v
<  < ^ > < > v > > v H v v v v v
<  v < H > v ^ < v v > ^ ^ > > G
```

Fig. 11: Q learning.



```
Learning time:

328.9346327781677

Learning Performance:
```

scroll output; double click to hide

```
<Figure size 432x288 with 0 Axes>
```

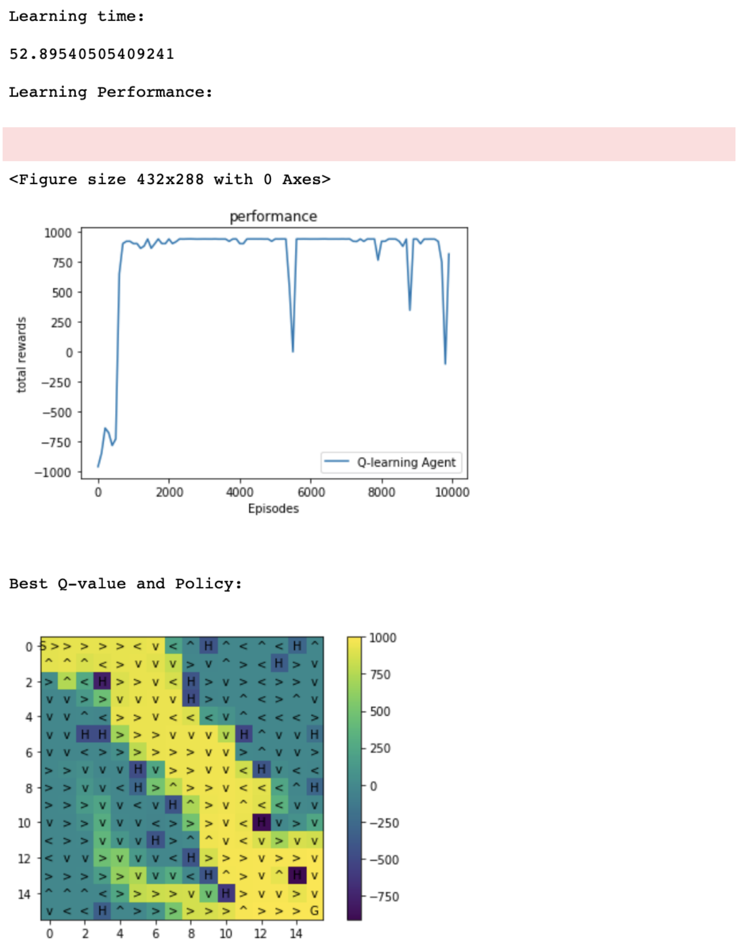
```
Best Q-value and Policy:
```

0	Sv	v	>	>	>	^	<	v	v	H	v	v	<	<	H	^
	v	v	<	>	v	<	<	<	>	v	v	>	v	H	v	>
2	v	^	<	H	v	>	^	v	H	>	>	<	>	v	>	>
	<	v	<	>	v	<	<	<	H	v	^	<	<	>	v	v
4	^	^	<	>	>	v	>	^	>	v	>	<	^	>	<	>
	<	<	H	H	>	>	v	<	v	^	^	H	>	^	v	H
6	v	<	v	>	>	<	>	^	<	v	v	<	^	v	<	<
	^	>	<	v	<	H	>	>	v	>	<	v	H	^	>	>
8	^	^	v	<	<	H	v	^	v	v	>	>	>	>	^	H
	<	^	^	<	<	v	v	H	>	>	>	<	<	v	v	v
10	^	>	v	v	>	^	>	>	>	<	v	v	H	^	>	>
	^	^	>	v	^	<	H	>	>	<	^	^	<	v	<	>
12	^	>	<	v	>	<	v	v	H	>	v	<	v	v	<	>
	>	^	v	>	^	v	<	v	<	H	^	^	<	^	H	v
14	<	>	^	<	^	^	v	<	^	v	H	v	>	>	v	v
	v	>	^	H	>	v	v	v	^	^	v	>	>	>	>	G
	0		2		4		6		8		10		12		14	

Fig. 12: SARSA.

Experiment 3 Parameters used:

- epsilon = 0.05
- learning rate = 0.1
- learning episodes = 100000
- test interval = 100
- do test = True



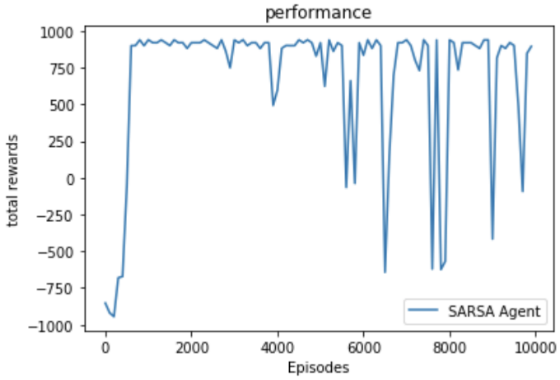
```
Learning time:

52.89540505409241

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

0	S>	>	>	>	>	<	v	<	^	H	^	<	^	<	H	^
	^	^	^	<	>	v	v	v	>	v	^	>	<	H	>	v
2	>	^	<	H	>	>	v	<	H	>	v	>	<	>	>	v
	v	v	>	>	v	v	v	v	H	>	v	^	<	>	^	v
4	v	v	^	<	>	>	v	<	<	<	v	^	<	<	<	>
	v	v	H	H	>	>	>	v	v	v	v	H	^	v	v	H
6	v	v	<	>	>	>	>	>	>	v	v	>	^	v	v	>
	>	>	v	v	v	H	v	>	>	v	v	<	H	v	<	<
8	>	>	v	v	<	H	>	^	>	>	v	<	<	<	^	H
	>	>	>	v	v	<	v	H	^	>	v	^	<	<	v	v
10	v	>	>	v	v	v	<	>	>	>	v	<	H	v	>	v
	<	>	>	v	v	v	H	>	^	^	v	<	v	>	v	v
12	<	v	v	>	v	v	v	<	H	>	>	>	v	>	>	v
	>	>	>	>	>	v	v	v	<	H	^	>	v	^	H	v
14	^	^	^	<	>	>	>	>	v	v	H	>	v	v	>	v
	v	<	<	H	^	>	>	>	>	>	>	^	>	>	>	G
	0		2		4		6		8		10		12		14	

Fig. 13: Q learning.

```
Learning time:

117.68178462982178

Learning Performance:


<Figure size 432x288 with 0 Axes>
```



```
Best Q-value and Policy:
```

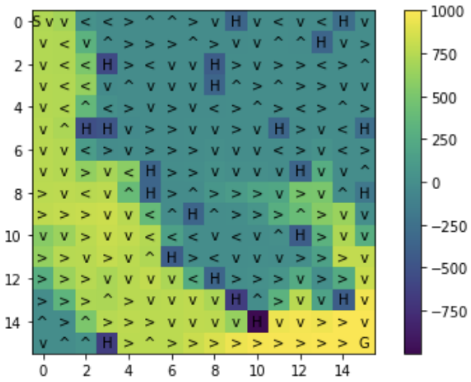
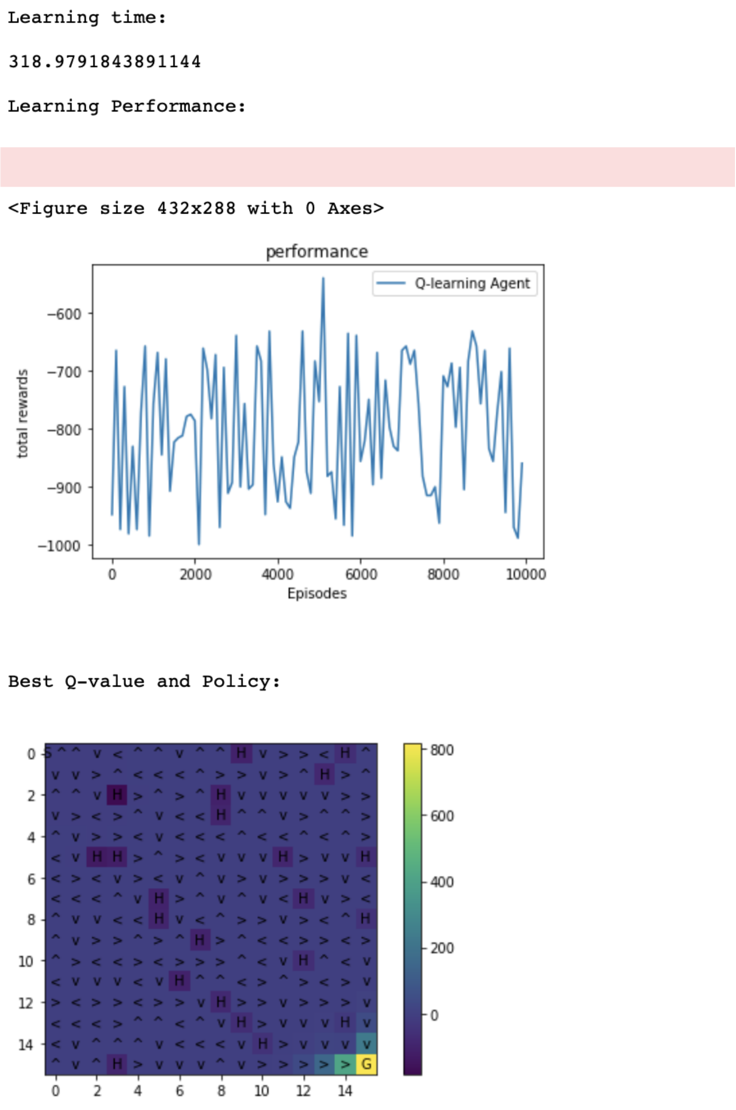


Fig. 14: SARSA.

Experiment 4 Parameters used:

- epsilon = 0.05
- learning rate = 0.001
- learning episodes = 100000
- test interval = 100

– do test = True

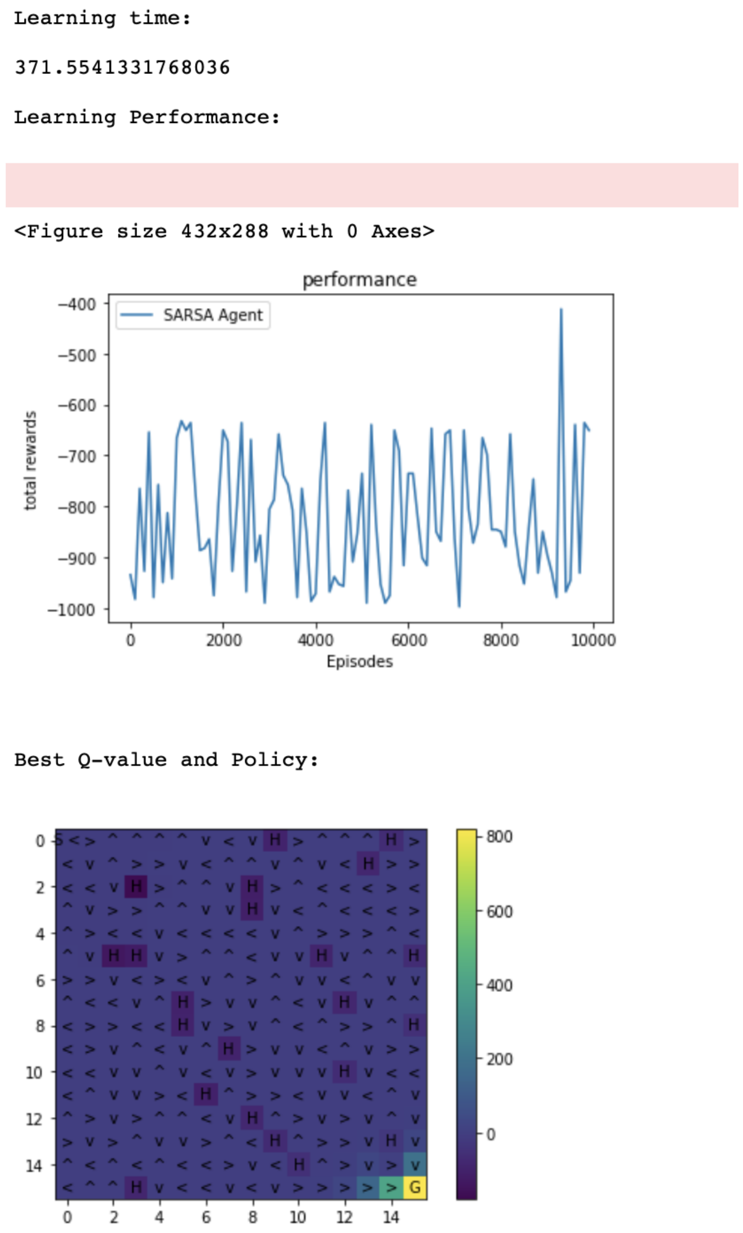


```
Learning time:

318.9791843891144

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

```
S^ ^ v < ^ ^ v ^ ^ H v > > < H ^
v  v > ^ < < < ^ > > v > ^ H > ^
^  ^ v H > ^ > ^ H v v v v v > >
v  > < > ^ v < < H ^ ^ v > ^ ^ >
^  v > > < v < < < ^ < < ^ < ^ >
<  v H H > ^ > < v v v H > v v H
<  > < v > < v ^ v > v > > > v <
<  < < ^ v H > ^ v ^ v < H v > <
^  v v < < H v < ^ > > v > < ^ H
^  v > > ^ > ^ H > ^ < < > > < >
^  > < < > < > > > ^ < v H ^ < v
<  v v v < v H ^ ^ < > ^ > < > v
>  < > > < > > v H > > v > > > v
<  < < > ^ ^ < ^ v H > v v v H v
<  v ^ ^ ^ v < < < v H > v v v v
^  v ^ H > v v v ^ v > > > > > G
```

Fig. 15: Q learning.



```
Learning time:

371.5541331768036

Learning Performance:

<Figure size 432x288 with 0 Axes>
```

```
Best Q-value and Policy:
```

```
 0  S< >  ^  ^  ^  ^  v  <  v  H  >  ^  ^  ^  H  >
    <  v  ^  >  >  v  <  ^  ^  v  ^  v  <  H  >  >
 2  <  <  v  H  >  ^  ^  v  H  >  ^  <  <  <  >  <
    ^  v  >  >  ^  ^  v  v  H  v  <  ^  <  <  <  >
 4  ^  >  <  <  v  <  <  <  <  v  ^  >  >  >  ^  <
    ^  v  H  H  v  >  ^  ^  <  v  v  H  v  ^  ^  H
 6  >  >  v  <  >  <  v  ^  >  ^  v  v  <  ^  v  v
    ^  <  <  v  ^  H  >  v  v  ^  <  v  H  v  ^  ^
 8  <  >  >  <  <  H  v  >  v  ^  <  ^  >  >  ^  H
    <  >  v  ^  <  v  ^  H  >  v  v  <  ^  v  >  >
10  <  <  v  v  ^  v  <  v  >  v  v  v  H  v  <  <
    <  ^  v  v  >  <  H  ^  >  >  <  v  v  <  ^  v
12  ^  >  v  >  ^  ^  <  v  H  ^  >  v  >  v  ^  v
    >  v  >  ^  v  v  >  ^  <  H  ^  >  >  v  H  v
14  ^  <  ^  <  ^  <  <  >  v  <  H  ^  >  v  >  v
    <  ^  ^  H  v  <  <  v  <  v  >  >  >  >  >  G
    0     2     4     6     8     10    12    14
```

Fig. 16: SARSA.

1.3 Dangerous Halway: Distributed

Experiment 1 Parameters used:

- epsilon = 0.1
- learning rate = 0.001
- learning episodes = 30000
- Number of Collector workers = 8
- Number of Evaluator workers = 4
- test interval = 100
- do test = True

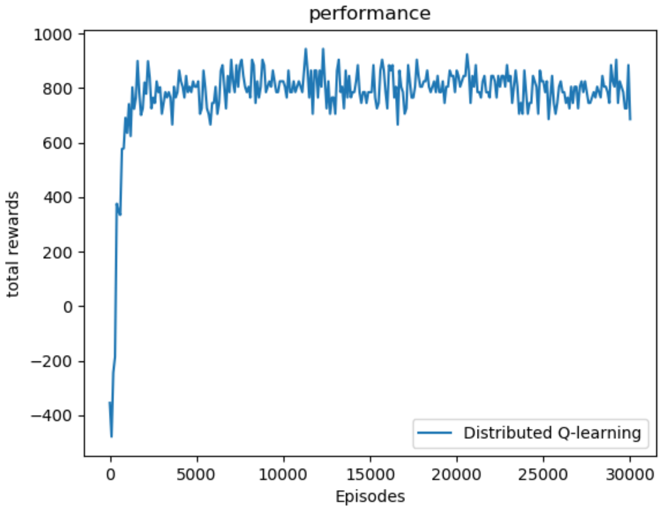


Fig. 17: Performance curve - Q learning.

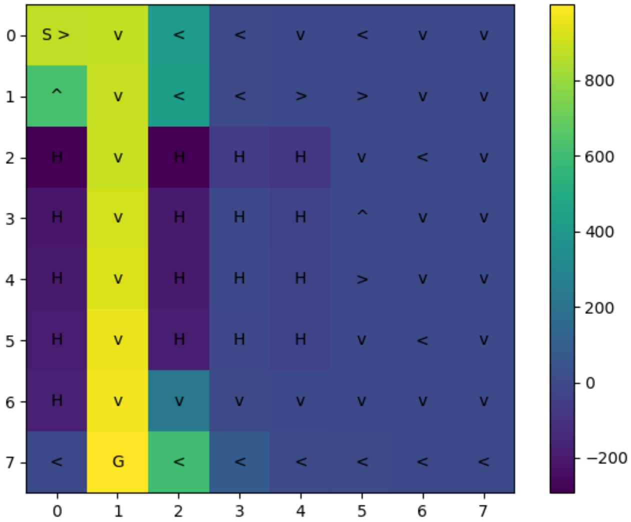


Fig. 18: Q table - Q learning.

Experiment 2 Parameters used:

– epsilon = 0.1

– learning rate = 0.001

– learning episodes = 30000

– Number of Collector workers = 4

– Number of Evaluator workers = 4

– test interval = 100

– do test = True

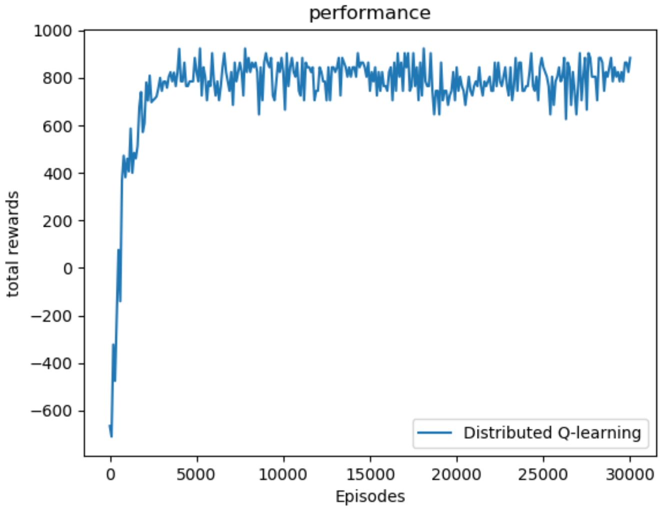


Fig. 19: Performance curve

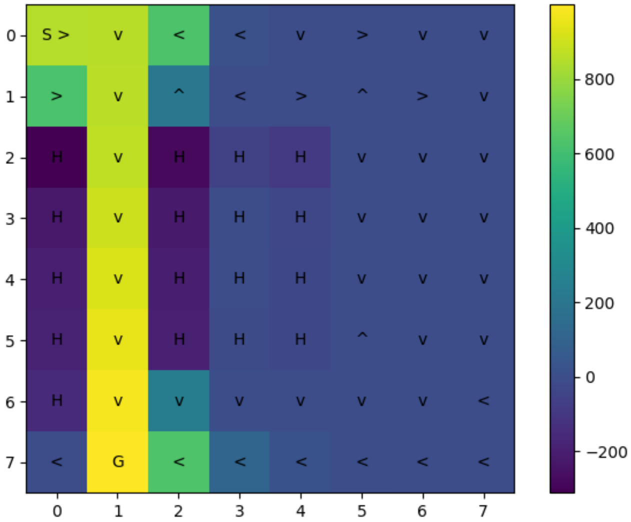


Fig. 20: Q table

Experiment 3 Parameters used:

- epsilon = 0.1
- learning rate = 0.001
- learning episodes = 30000
- Number of Collector workers = 2
- Number of Evaluator workers = 4
- test interval = 100
- do test = True

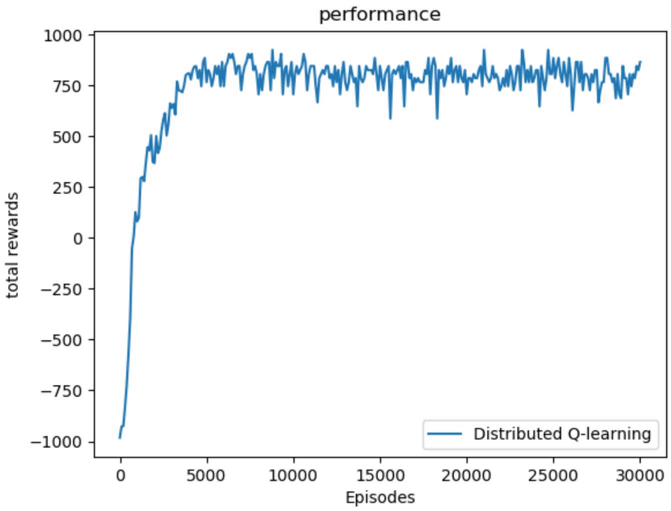


Fig. 21: Performance curve

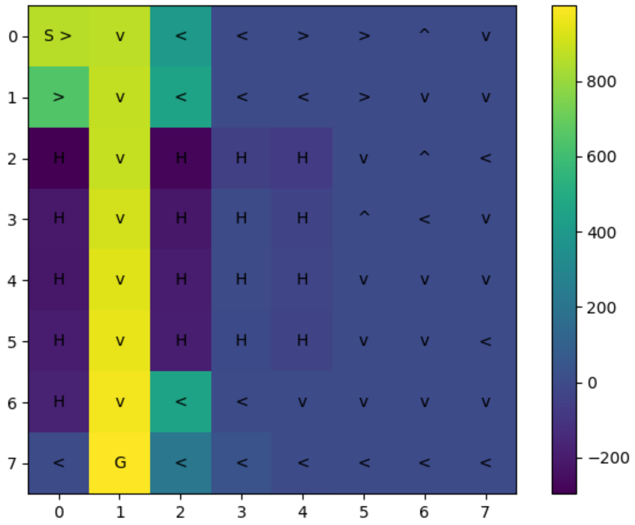


Fig. 22: Q table

1.4 Map 16: Distributed

Experiment 1 Parameters used:

- epsilon = 0.1
- learning rate = 0.001
- learning episodes = 10000
- Number of Collector workers = 8
- Number of Evaluator workers = 4
- test interval = 100
- do test = True

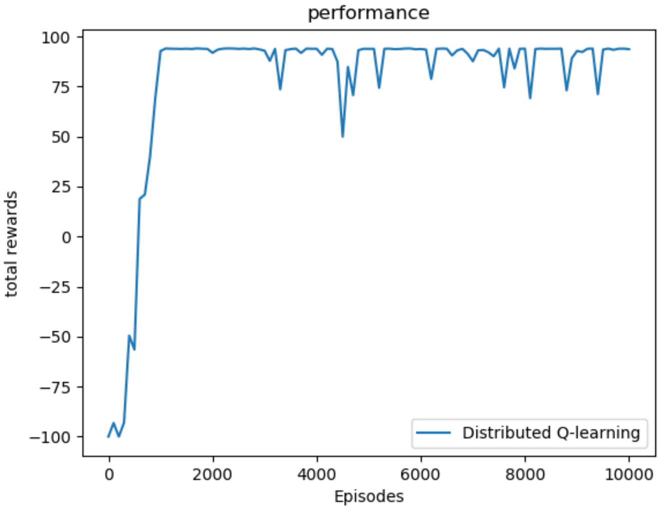
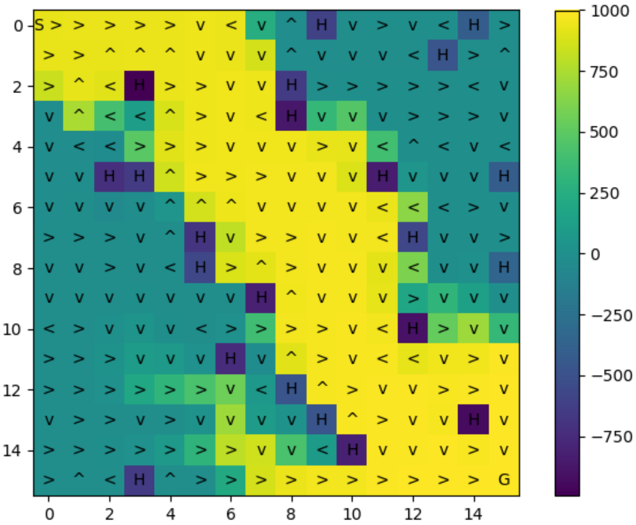


Fig. 23: Performance curve



S >	>	>	>	>	v	<	v	^	H	v	>	v	<	H	>
>	>	^	^	^	v	v	v	^	v	v	v	<	H	>	^
>	^	<	H	>	>	v	v	H	>	>	>	>	>	<	v
v	^	<	<	^	>	v	<	H	v	v	v	>	>	>	v
v	<	<	>	>	>	v	v	v	>	v	<	^	<	v	<
v	v	H	H	^	>	>	>	v	v	v	H	v	v	v	H
v	v	v	v	^	^	^	v	v	v	v	<	<	<	>	v
>	>	>	v	^	H	v	>	>	v	v	<	H	v	v	>
v	v	>	v	<	H	>	^	>	v	v	v	<	v	v	H
v	v	v	v	v	v	v	H	^	v	v	v	>	v	v	v
<	>	v	v	v	<	>	>	>	>	v	<	H	>	v	v
>	>	>	v	v	v	H	v	^	>	v	<	<	v	>	v
>	>	>	>	>	>	v	<	H	^	>	v	v	>	>	v
v	>	>	v	>	v	v	v	v	H	^	>	v	v	H	v
>	>	>	>	>	>	>	v	v	<	H	v	v	v	>	v
>	^	<	H	^	>	>	>	>	>	>	>	>	>	>	G

Fig. 24: Q table

Experiment 2 Parameters used:

- epsilon = 0.1

- learning rate = 0.001

- learning episodes = 10000

- Number of Collector workers = 4

- Number of Evaluator workers = 4

- test interval = 100

- do test = True

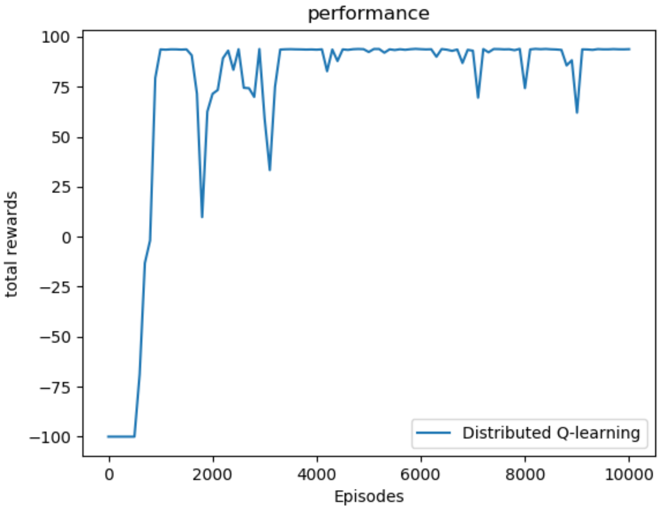


Fig. 25: Performance curve

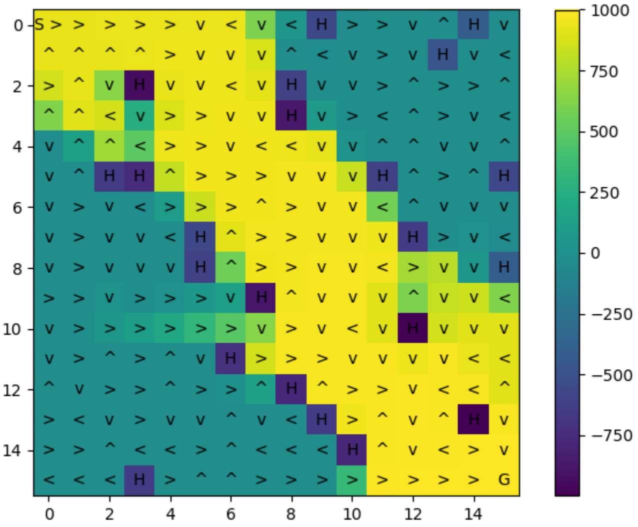


Fig. 26: Q table

Experiment 3 Parameters used:

- epsilon = 0.1
- learning rate = 0.001
- learning episodes = 10000
- Number of Collector workers = 2
- Number of Evaluator workers = 4
- test interval = 100
- do test = True

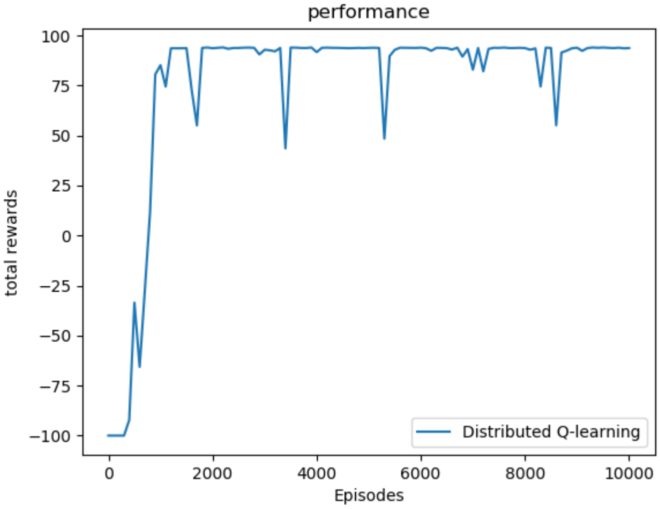


Fig. 27: Performance curve

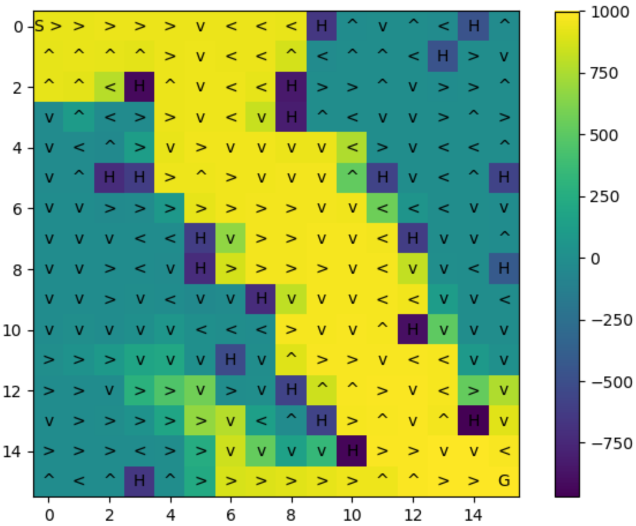


Fig. 28: Q table

2 Did you observe differences for SARSA when using the two different learning rates? If there were significant differences, what were they and how can you explain them?

On using a high learning rate, SARSA converges faster but it keeps on diverging for some of the iterations. The convergence is slow for lower learning rate but it converges to a stable optima. For the higher learning rate, the convergence is much faster but does not stay stable at the optima.

3 Did you observe differences for Q-learning when using the two different learning rates? If there were significant differences, what were they and how can you explain them?

On using a high learning rate, Q-learning converges faster but it keeps on diverging for some of the iterations. The divergence is worse than for SARSA as its off policy. The convergence is slow for lower learning rate but it converges to a stable optima. For the higher learning rate, the convergence is much faster but does not stay stable at the optima.

4 Did you observe differences for SARSA when using different values of ϵ ? If there were significant differences, what were they and how do you explain them?

Using lower value of epsilon reduces the exploration. Hence SARSA is not able to find optimal solution and converges at a suboptimal solution.

5 Did you observe differences for Q-learning when using different values of ϵ ? If there were significant differences, what were they and how do you explain them?

Using lower value of epsilon reduces the exploration. Hence Q-learning is not able to find optimal solution and converges at a suboptimal solution. However, the exploration is better for Q-learning than for SARSA since Q-learning is off policy, thus inherently inducing an exploration.

6 For the map ”Dangerous Hallway” did you observe differences in the policies learned by SARSA and Q-Learning for the two values of epsilon (there should be differences between Q-learning and SARSA for at least one value)? If you observed a difference, give your best explanation for why Q-learning and SARSA found different solutions.

Q-learning converges to a policy corresponding to a Q-table which yields high q-values to the dangerous hallway in between. SARSA does not exploit the dangerous hallway and gives more uniform q-value distribution depending on distance of states from goal state.