INDIAN INSTITUTE OF TECHNOLOGY,
MADRAS

REINFORCEMENT LEARNING PROGRAMMING
ASSIGNMENT 2

CS6700

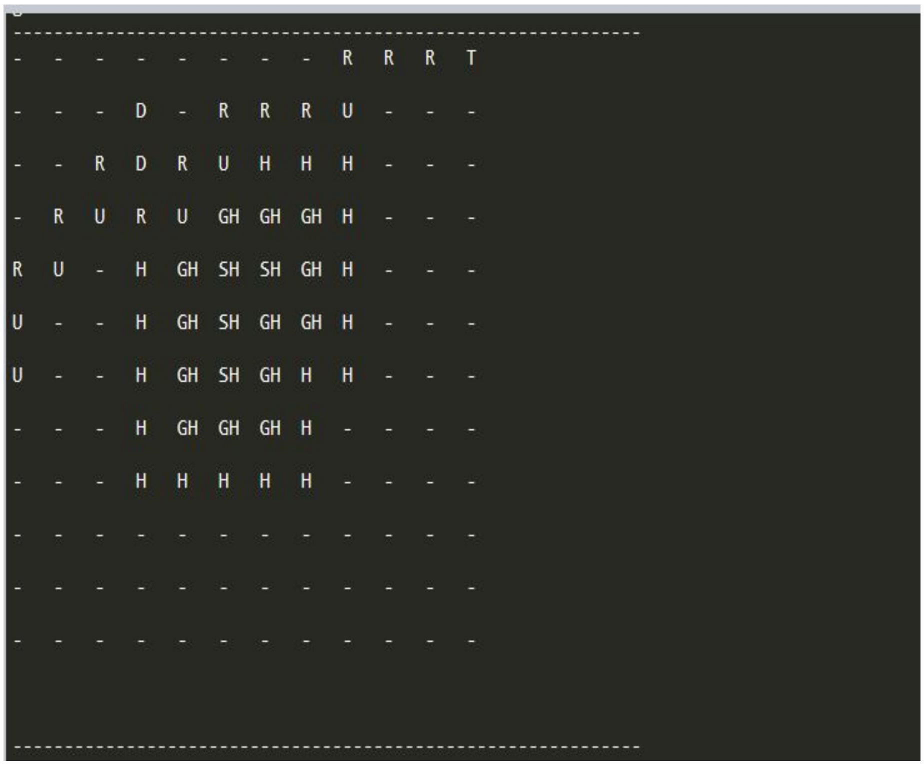
SARSA,Qlearning,Policy Gradient

Author	*Roll Number*
DODIYA KEVAL	CS19M023

March 15, 2020

1 Problem 1

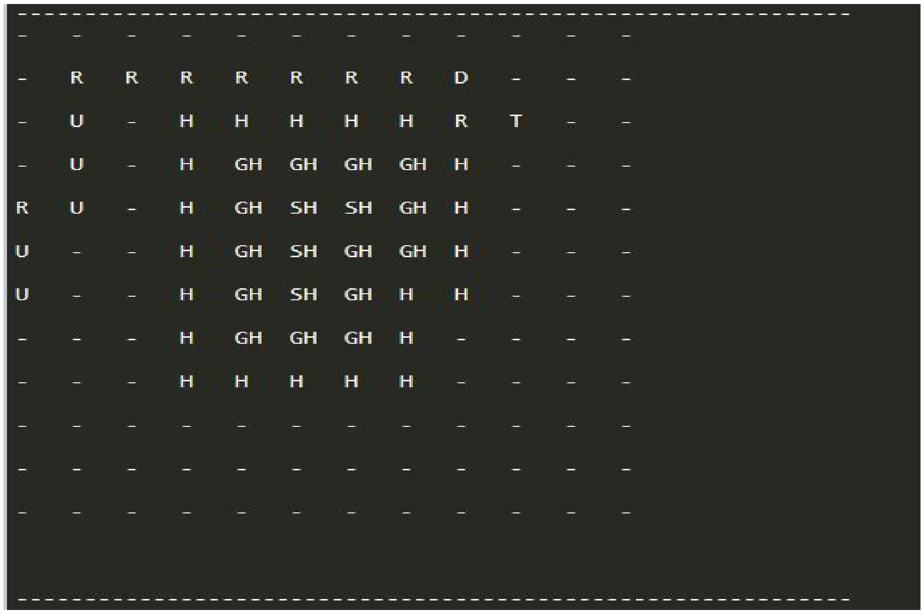
1.1 Q Learning



-	-	-	-	-	-	-	-	R	R	R	T
-	-	-	D	-	R	R	R	U	-	-	-
-	-	R	D	R	U	H	H	H	-	-	-
-	R	U	R	U	GH	GH	GH	H	-	-	-
R	U	-	H	GH	SH	SH	GH	H	-	-	-
U	-	-	H	GH	SH	GH	GH	H	-	-	-
U	-	-	H	GH	SH	GH	H	H	-	-	-
-	-	-	H	GH	GH	GH	H	-	-	-	-
-	-	-	H	H	H	H	H	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-

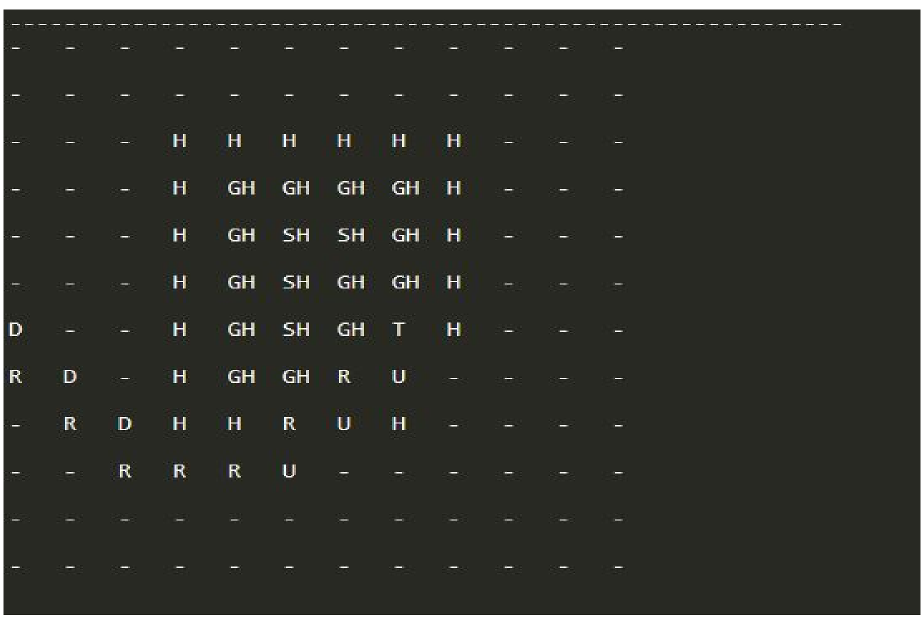
Figure 1: Qlearning optimal policy for goal A

here i'm assuming H = hole with reward -1,GH = hole with reward -2, SH = hole with reward -3, and U = upper direction, L = left, R = right, D = down,T = terminal and '-' = empty slot.



```
------------------------------------------------------------
-   -   -   -   -   -   -   -   -   -   -   -
-   R   R   R   R   R   R   R   D   -   -   -
-   U   -   H   H   H   H   H   R   T   -   -
-   U   -   H   GH  GH  GH  GH  H   -   -   -
R   U   -   H   GH  SH  SH  GH  H   -   -   -
U   -   -   H   GH  SH  GH  GH  H   -   -   -
U   -   -   H   GH  SH  GH  H   H   -   -   -
-   -   -   H   GH  GH  GH  H   -   -   -   -
-   -   -   H   H   H   H   H   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
------------------------------------------------------------
```

Figure 2: Qlearning optimal policy for goal B



```
------------------------------------------------------------
-   -   -   -   -   -   -   -   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
-   -   -   H   H   H   H   H   H   -   -   -
-   -   -   H   GH  GH  GH  GH  H   -   -   -
-   -   -   H   GH  SH  SH  GH  H   -   -   -
-   -   -   H   GH  SH  GH  GH  H   -   -   -
D   -   -   H   GH  SH  GH  T   H   -   -   -
R   D   -   H   GH  GH  R   U   -   -   -   -
-   R   D   H   H   R   U   H   -   -   -   -
-   -   R   R   R   U   -   -   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
-   -   -   -   -   -   -   -   -   -   -   -
```

Figure 3: Qlearning optimal policy for goal C

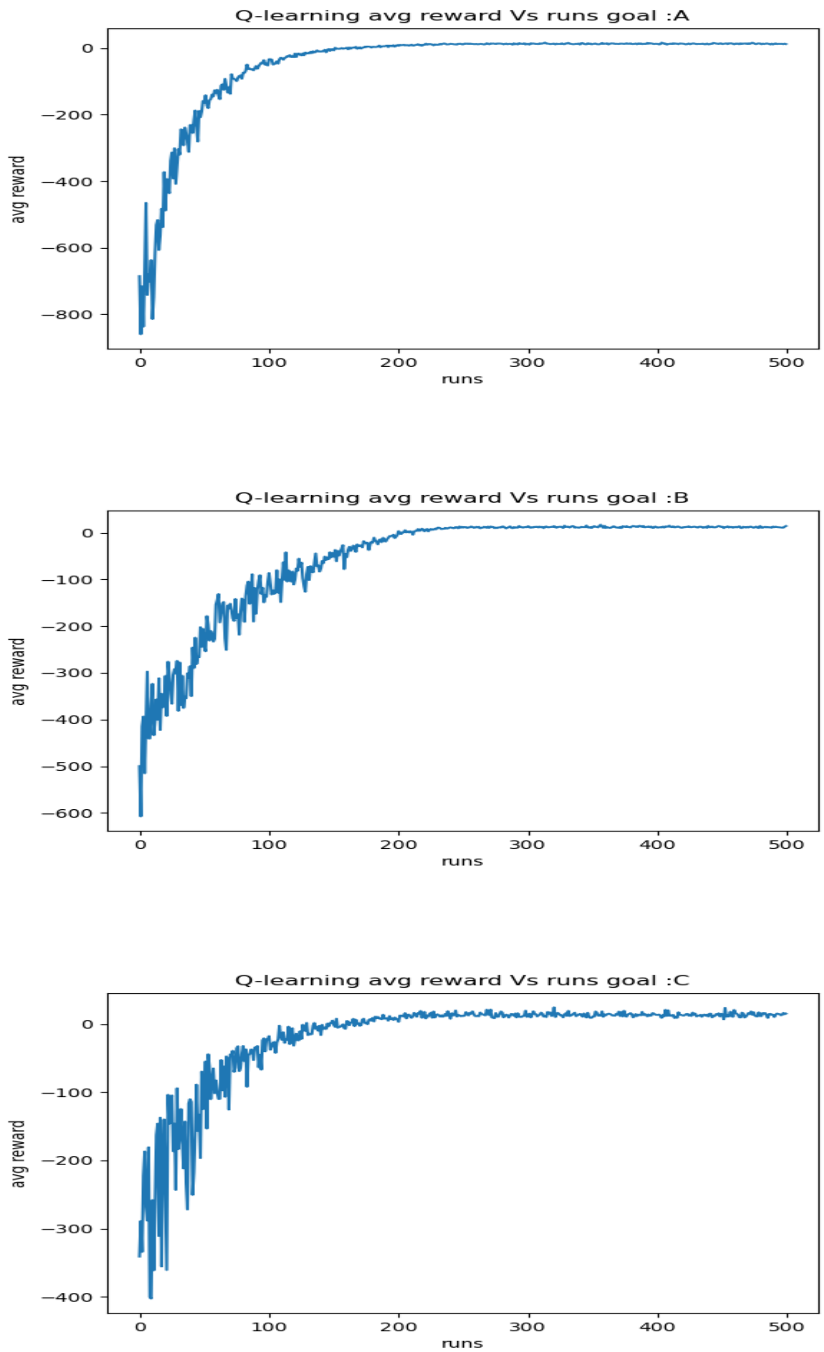
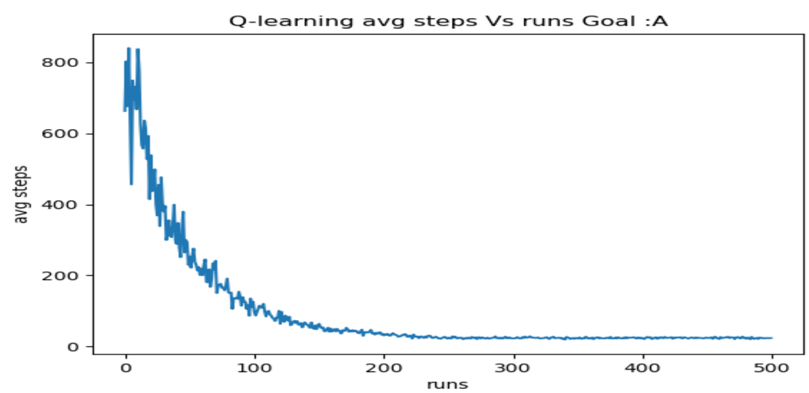
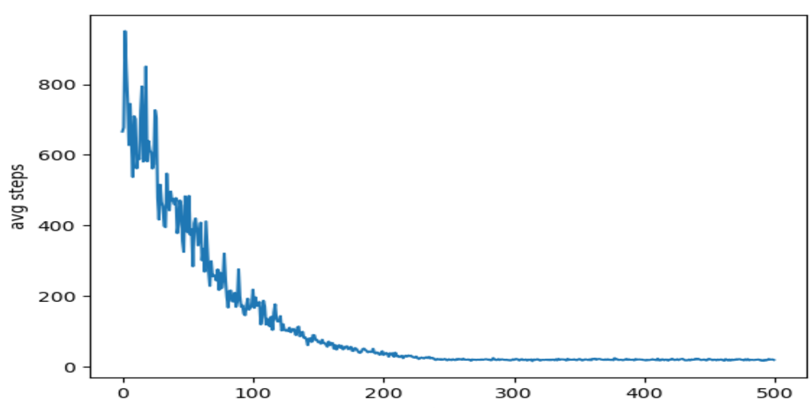


Figure 4: QLearning Reward VS runs with different Goals

- I counted offgrid Reward as -5 and 0 reward for normal action hence minimum possible reward goes very low. and i implemented a e-greedy with 1 as initial epsilon so that algorithm explore more in intial stage then it slowly decreases to 0. so i'm increasing exploitation steps.
- for Q learning initially it explores and eventually it learns optimal policy hence graph of reward increases as shows in graph.
- for initial steps variance in reward is more because of exploration and then after it converges(finding best policy of states) to best rewards
- i ran 50 independent experiment with 500 episodic task and again i register only one gridworld with different goals.
- showing policy is optimal policy for all runs





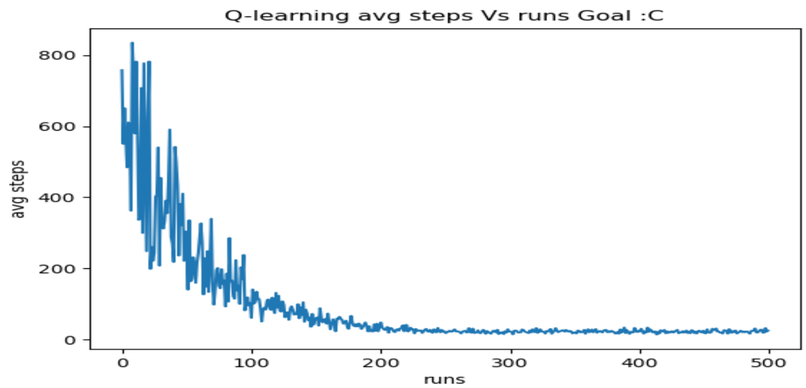


Figure 5: QLearning steps VS runs with different Goals

- here initially i got maximum steps 800 around then with number of runs increases that eventually steps decreases.
- and variance of goal C is more that is because of random behaviour and exploration of Q learning is more hence its take more steps

1.2 SARSA :

```
------------------------------------------------------------
-   -   -   -   -   -   -   -   R   R   R   T

-   -   R   R   R   R   R   R   U   -   -   -

-   R   U   H   H   H   H   H   H   -   -   -

-   U   U   H   GH  GH  GH  GH  H   -   -   -

R   U   -   H   GH  SH  SH  GH  H   -   -   -

U   -   -   H   GH  SH  GH  GH  H   -   -   -

U   -   -   H   GH  SH  GH  H   H   -   -   -

-   -   -   H   GH  GH  GH  H   -   -   -   -

-   -   -   H   H   H   H   H   -   -   -   -

-   -   -   -   -   -   -   -   -   -   -   -

-   -   -   -   -   -   -   -   -   -   -   -

-   -   -   -   -   -   -   -   -   -   -   -


------------------------------------------------------------
```

Figure 6: SARSA optimal policy for goal A

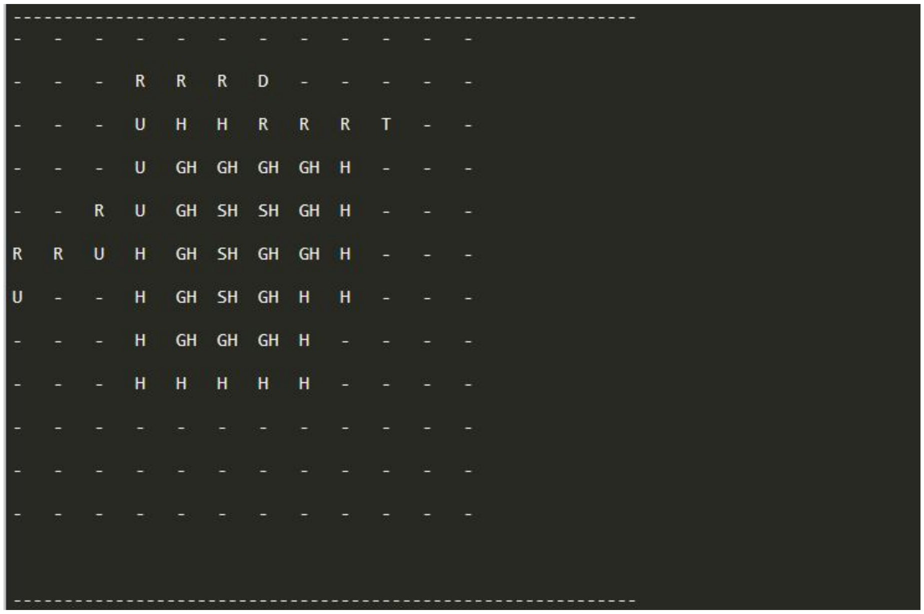


Figure 7: SARSA optimal policy for goal B

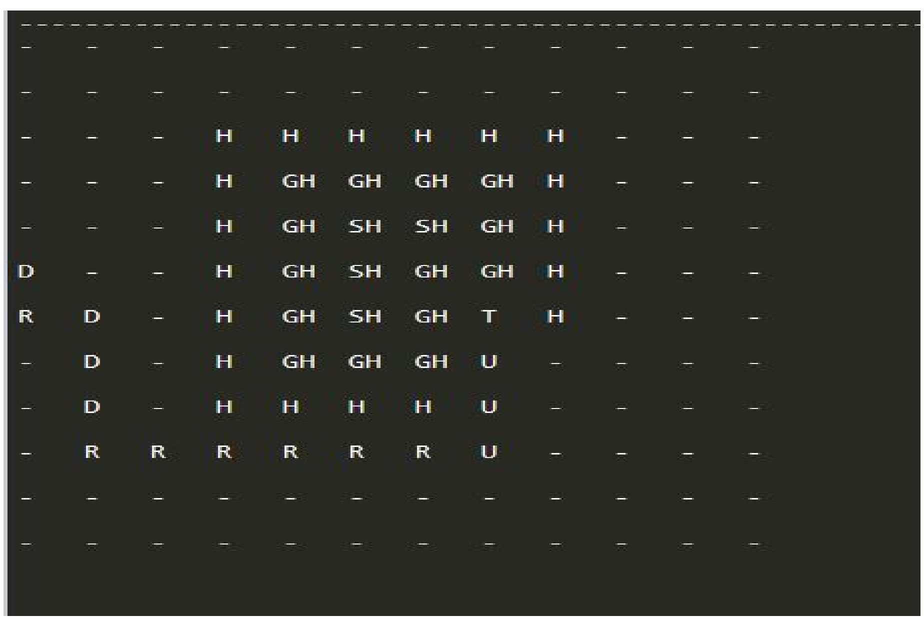


Figure 8: SARSA optimal policy for goal C

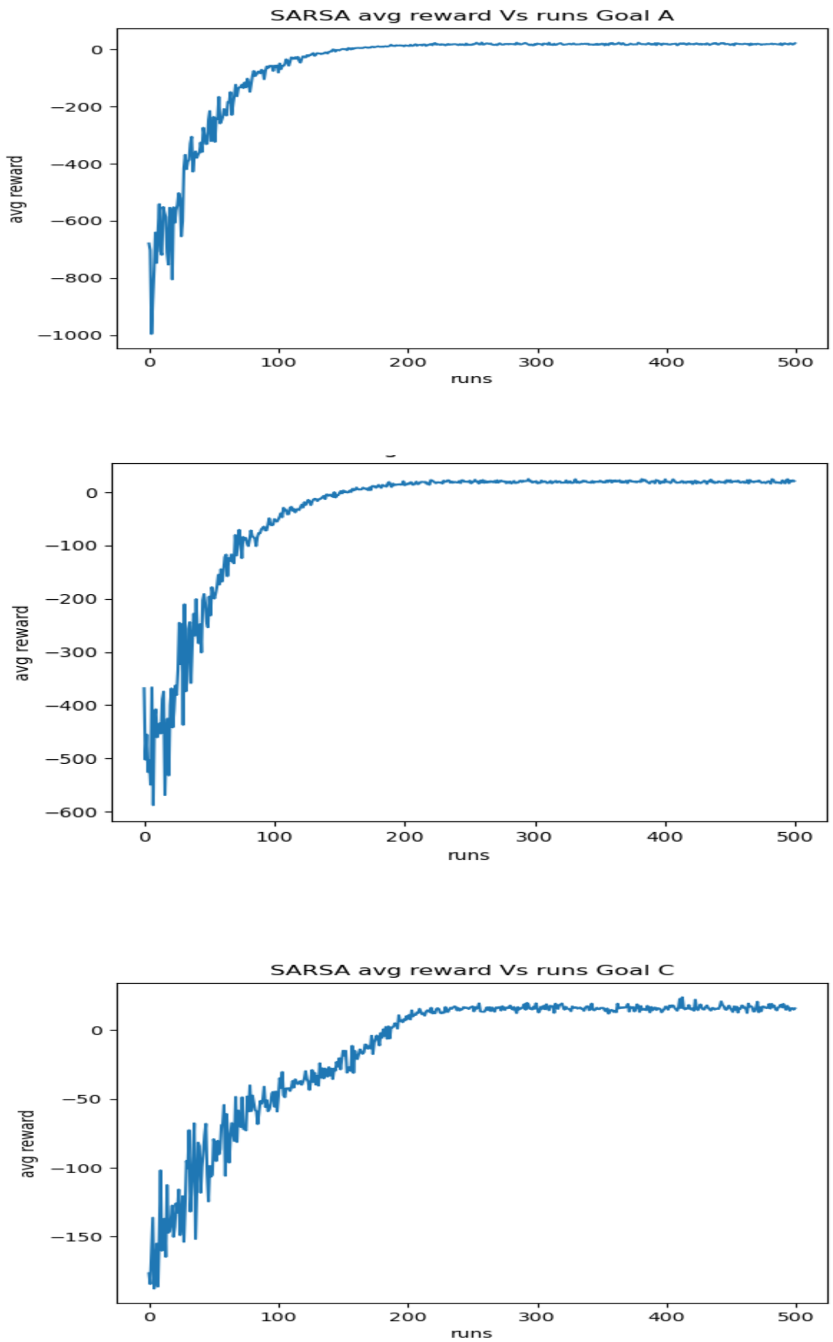


Figure 9: SARSA avg Reward VS runs with different Goals with runs of 500 and avg over 50 experiments

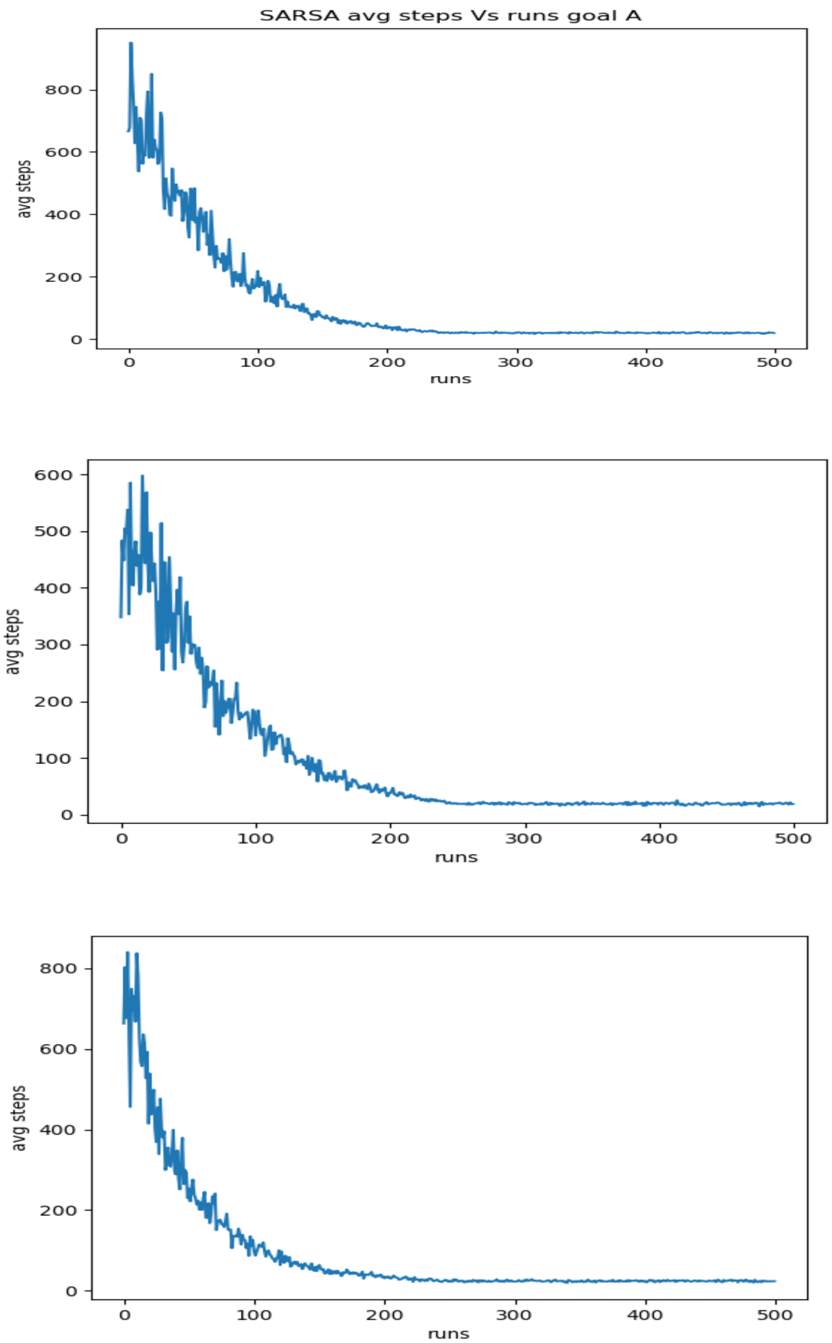


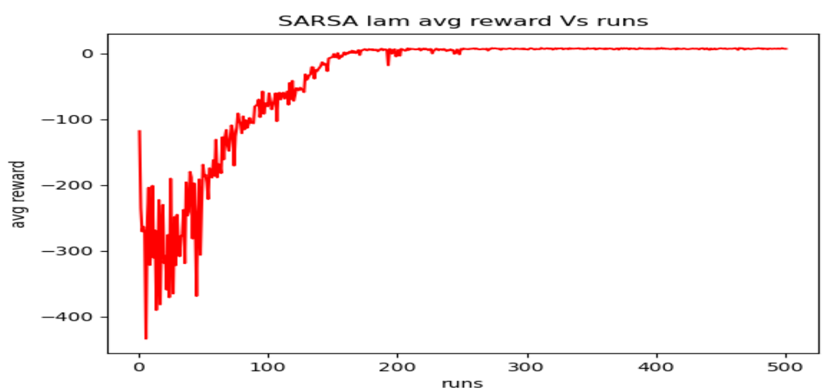
Figure 10: SARSA steps VS runs with different Goals first graph is for goal A, second graph is for goal B,and then goal C

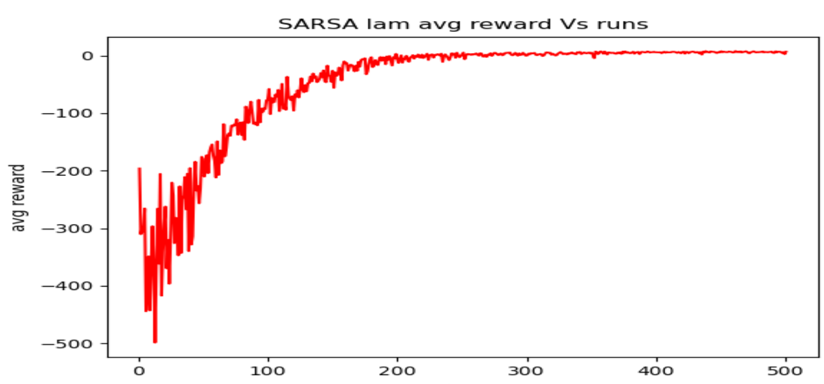
- SARSA uses its own policy to select next action's action state value hence SARSA is safer than Q learning. i put offgrid reward as -5 hence SARSA and Q learning both performing like a same algorithm with minor difference.

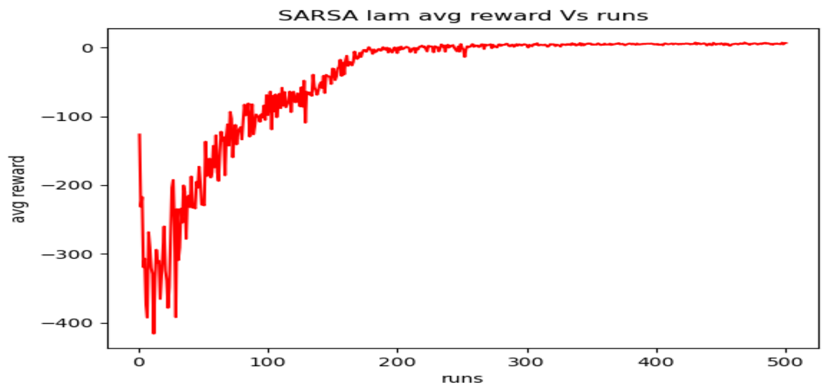
1.3 SARSA λ:

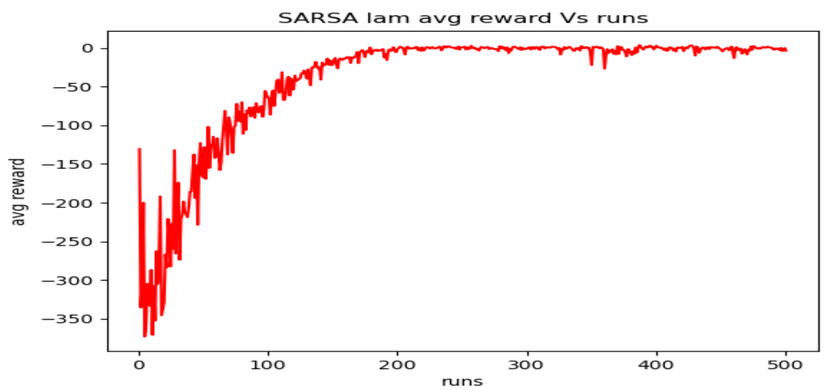
1.3.1 Goal A:

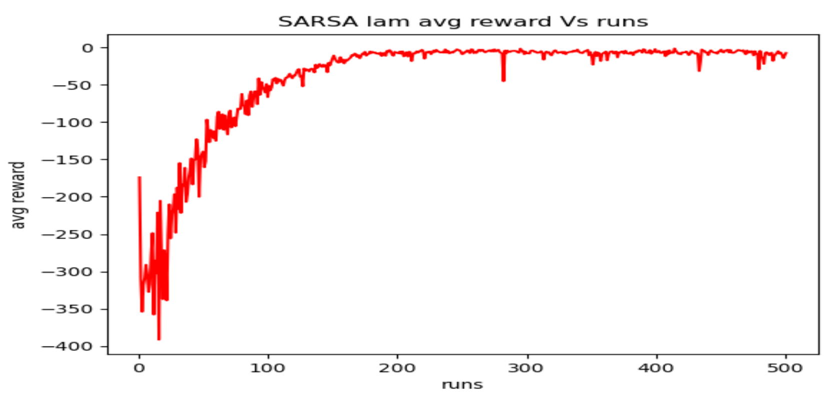
- following plots are of avg reward VS episodes(runs) of SARSA λ with different λs. and we can observed that rewards are near to same.
- epsilon value starting from 0.7 means more exploration then i'm decreasing epsilon(increment of exploitation) with increment of episodes. hence reward goes tens to zero and above not in range 8-10.






SARSA lam avg reward Vs runs
avg reward
runs
0
−100
−200
−300
−400
0
100
200
300
400
500


SARSA lam avg reward Vs runs
avg reward
runs
0
−50
−100
−150
−200
−250
−300
−350
0
100
200
300
400
500


SARSA lam avg reward Vs runs
avg reward
runs
0
−50
−100
−150
−200
−250
−300
−350
−400
0
100
200
300
400
500

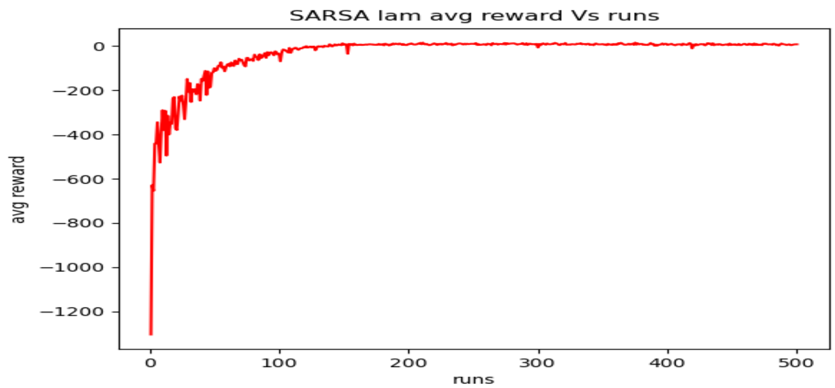
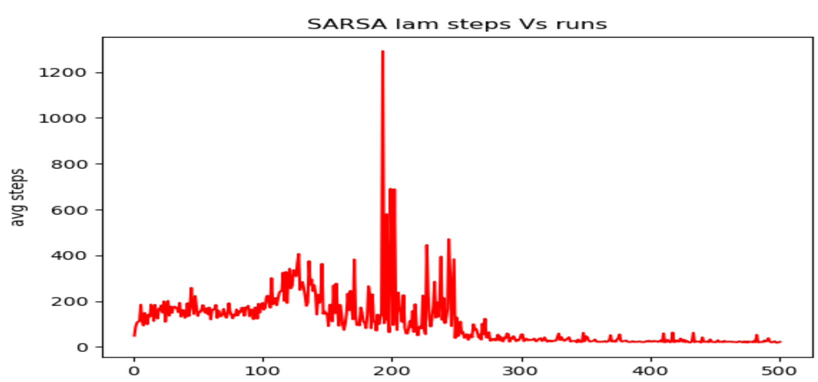
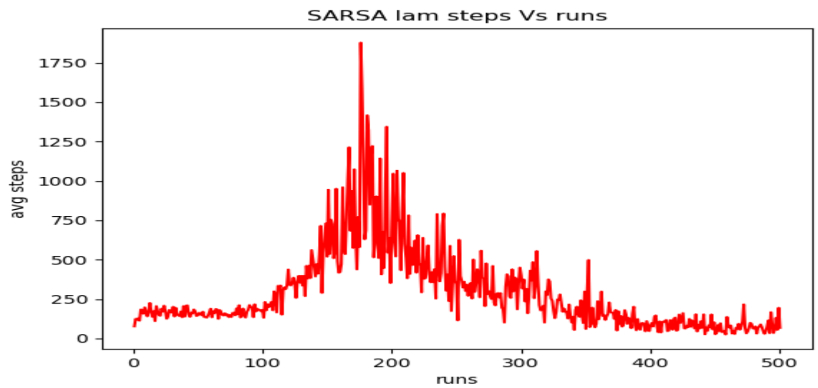
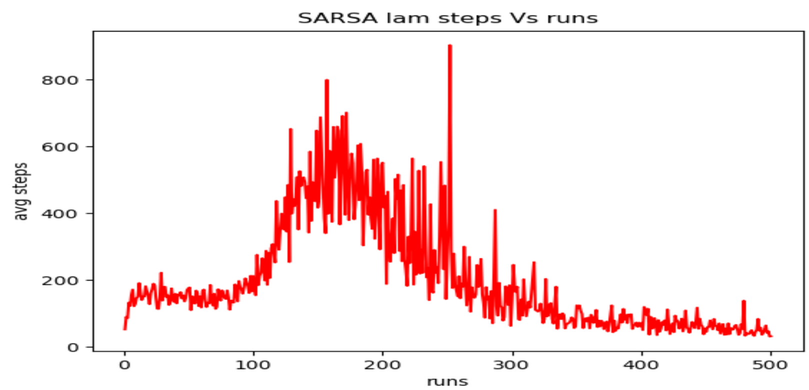


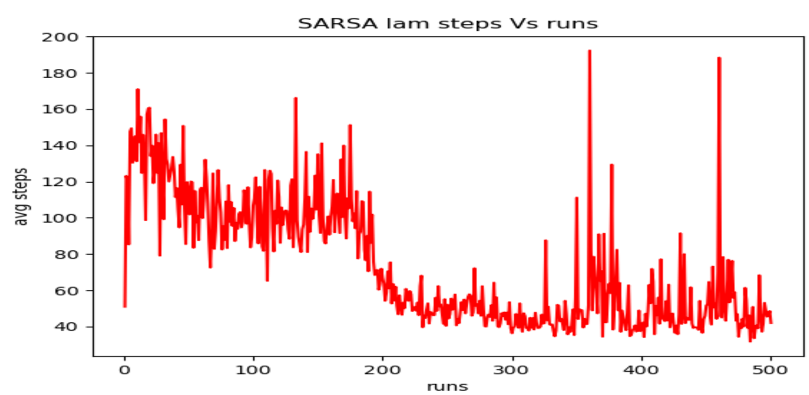
Figure 11: SARSA lambada reward VS episodes with different lambada over 500 episodes and 50 runs with goal A where $\lambda = [0, 0.3, 0.5, 0.9, 0.99, 1]$

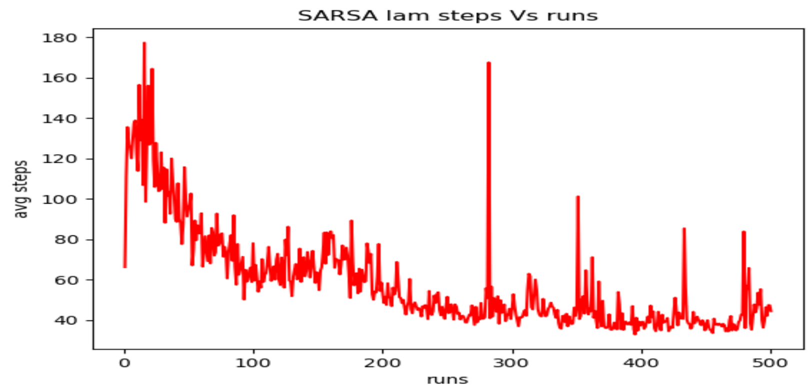
- in SARSA λ whenever lambada value increases we can observe(see below plots) that avg steps size also got decreases.
- all reward points are converges around 500 episodes.
- I have taken offgrid reward as -5 hence minimum reward amount is very low(around -400).
- following plots are of avg steps VS episodes(runs) of SARSA λ with different λs. and we can observed that avg steps are decreasing with λ
- In steps Vs episodes graph we can see more variance it's because of exploration. exploration may take more steps that is varies in all runs.




SARSA lam steps Vs runs
avg steps
1750
1500
1250
1000
750
500
250
0
0
100
200
300
400
500
runs


SARSA lam steps Vs runs
avg steps
800
600
400
200
0
0
100
200
300
400
500
runs


SARSA lam steps Vs runs
avg steps
200
180
160
140
120
100
80
60
40
0
100
200
300
400
500
runs



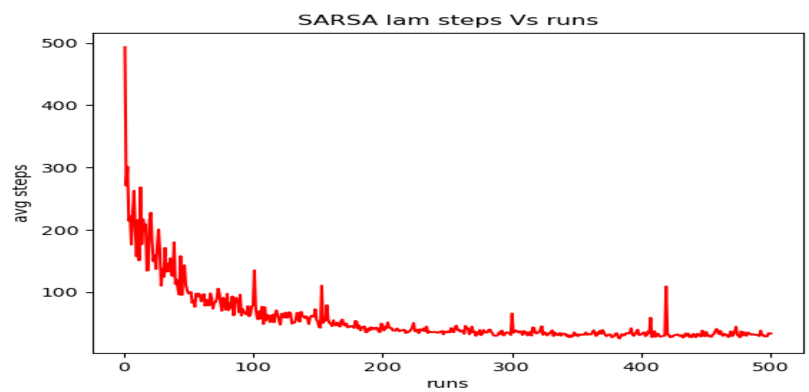
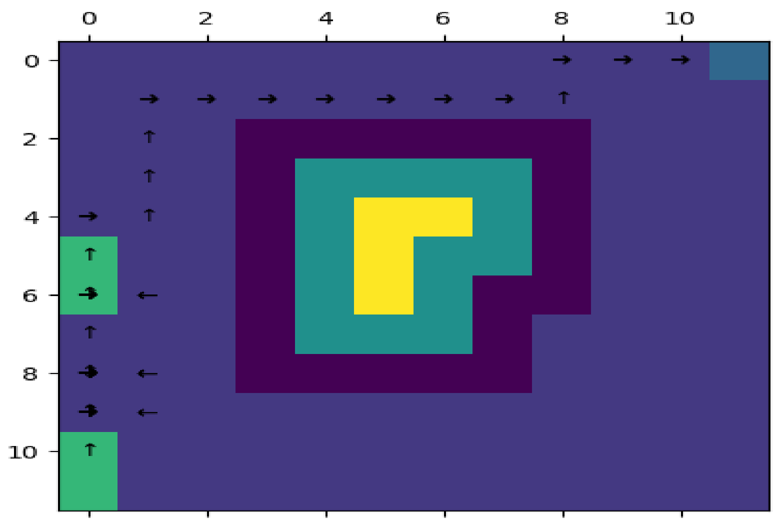
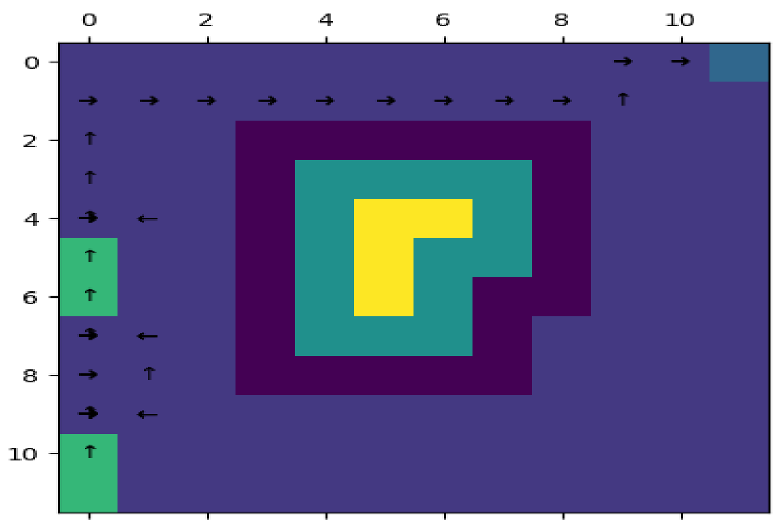


Figure 12: SARSA lambada avg steps VS episodes with different lambada over 500 episodes and 50 runs with goal A where λ =[0, 0.3, 0.5, 0.9, 0.99, 1]

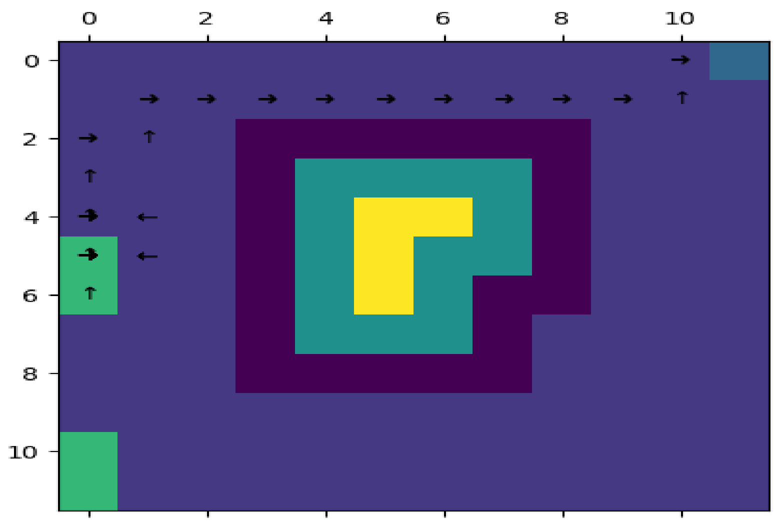
- following policies are optimal policies i got with different λs with 500 episodes and 50 runs. i uploaded policy which is optimal of all 50 runs with 500 episodes so depends on starting state policies may take different path with maximum reward and minimum steps.
- in below mentioned policies arrow indiactes directions and gridworld with different colors have different reward value (e.x. yellow color describes -3 reward,magenta color = -1,skyblue = -2,dark blue = 10(terminal) ,light green = strating states)
- some gridbox have more than one direction discribes more than one actions in same gridbox when it return back to same state.



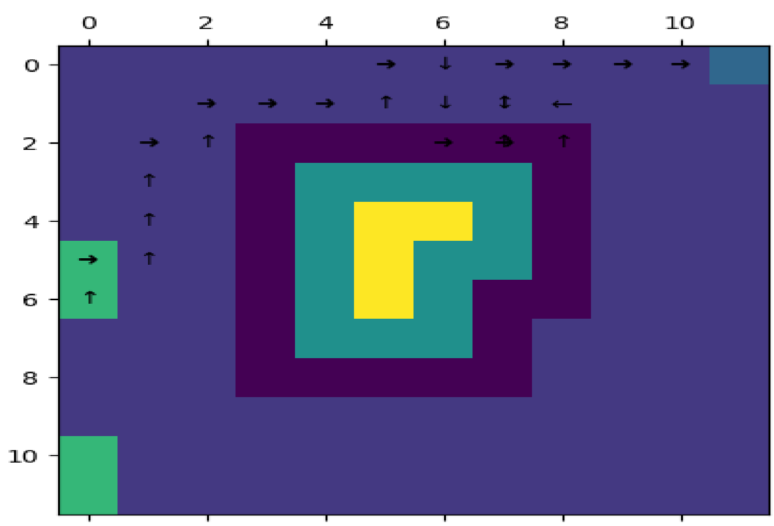
0
2
4
6
8
10



0
2
4
6
8
10



0
2
4
6
8
10



0
2
4
6
8
10

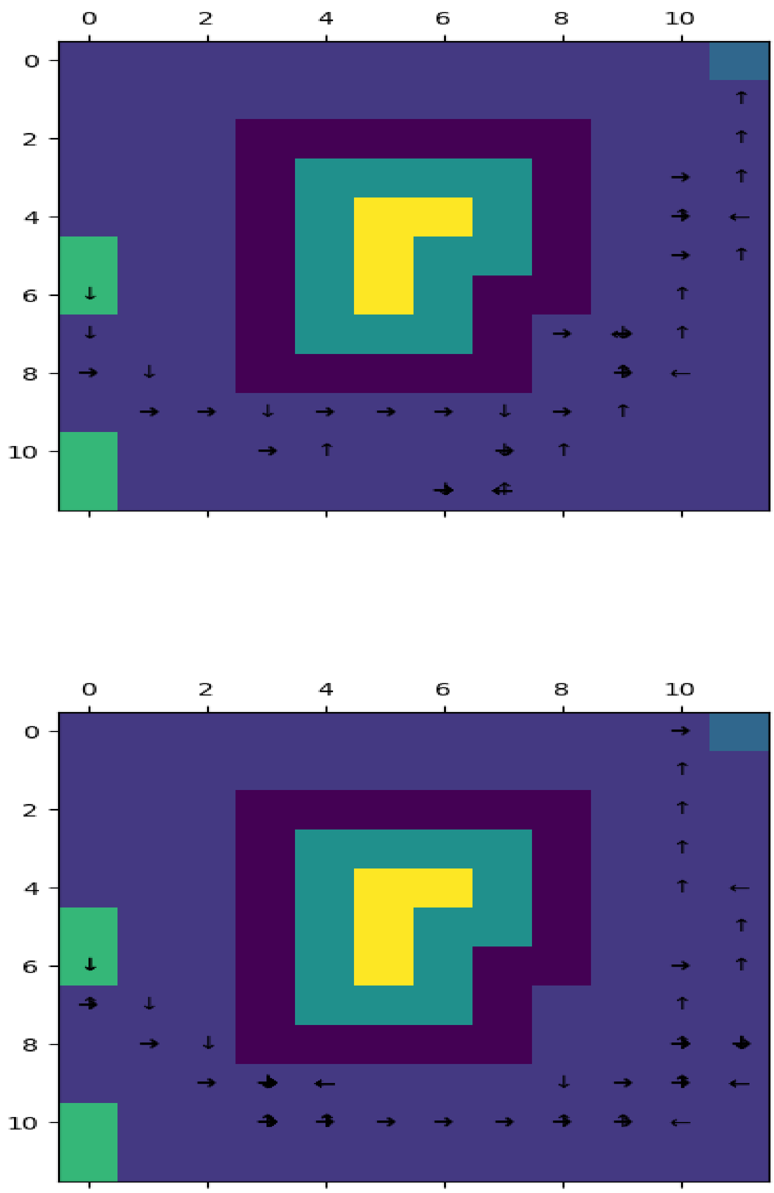
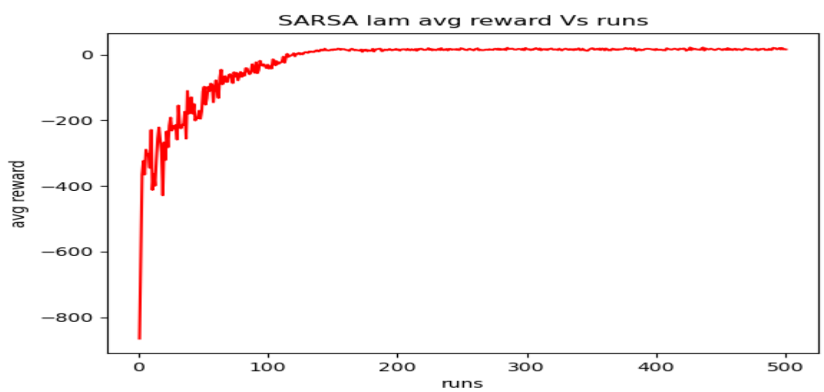
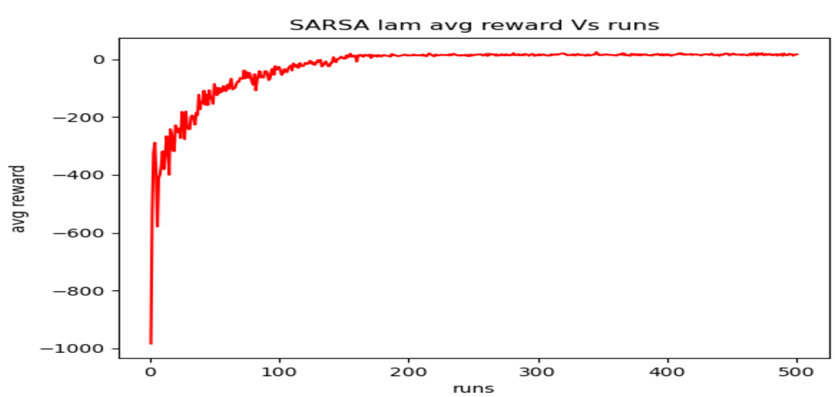


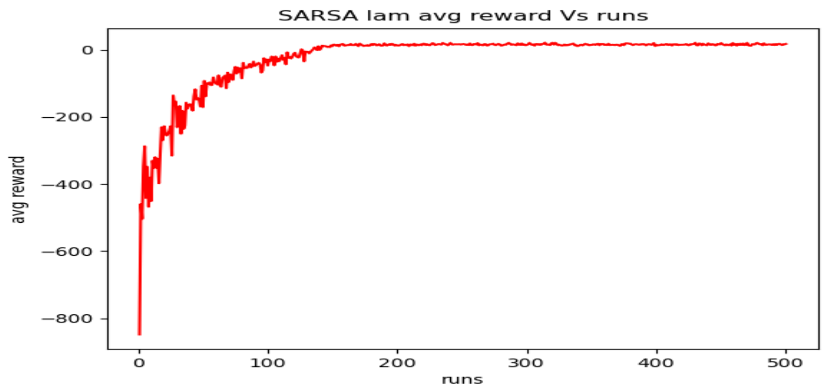
Figure 13: SARSA lambada optimal policies with different lambada over 500 episodes and 50 runs with goal A .where λ =[0, 0.3, 0.5, 0.9, 0.99, 1]

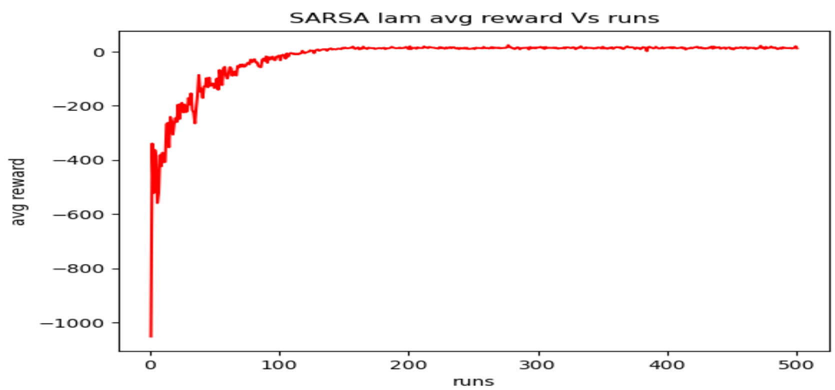
1.3.2 Goal B:

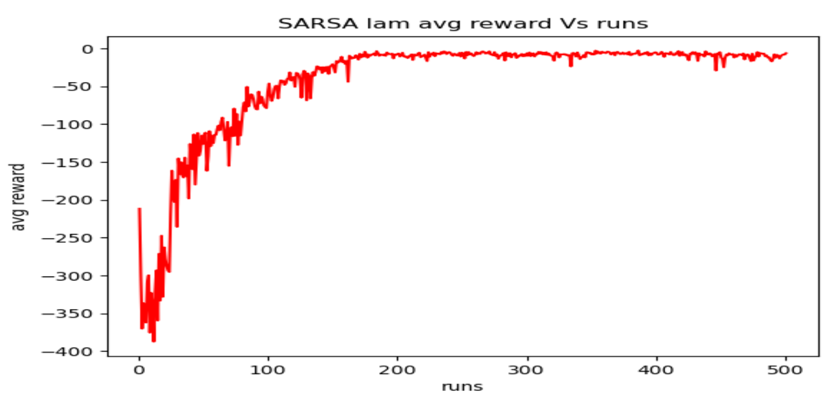
- following plots are of avg reward VS episodes(runs) of SARSA λ with different λs. and we can observed that rewards are near to same.
- epsilon value starting from 0.7 means more exploration then i'm decreasing epsilon(increment of exploitation) with increment of episodes. hence reward goes tens to zero and above not in range 8-10.






SARSA lam avg reward Vs runs
avg reward
runs
0
−200
−400
−600
−800
0
100
200
300
400
500


SARSA lam avg reward Vs runs
avg reward
runs
0
−200
−400
−600
−800
−1000
0
100
200
300
400
500


SARSA lam avg reward Vs runs
avg reward
runs
0
−50
−100
−150
−200
−250
−300
−350
−400
0
100
200
300
400
500

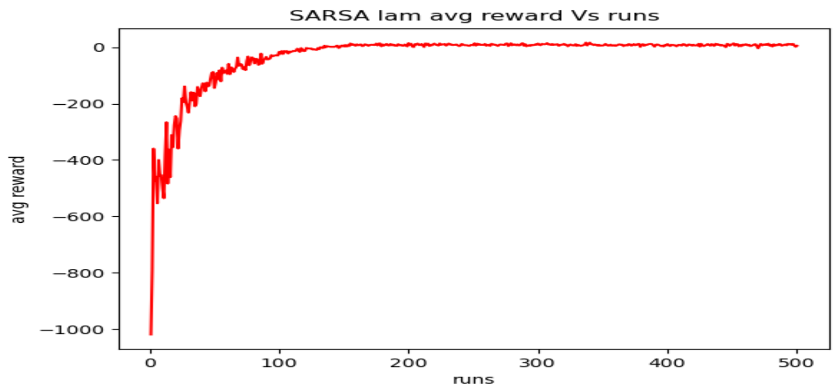
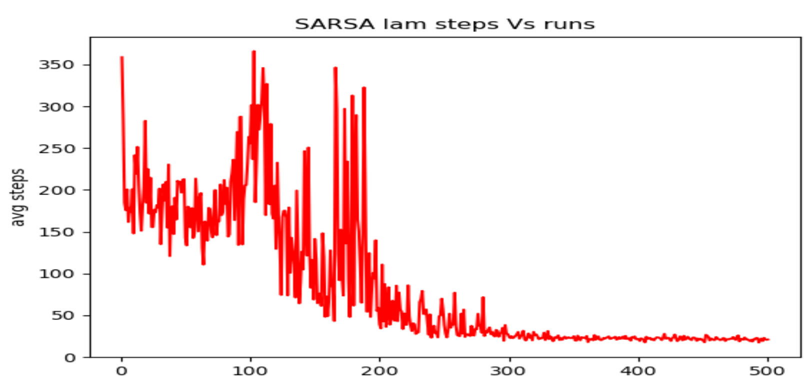
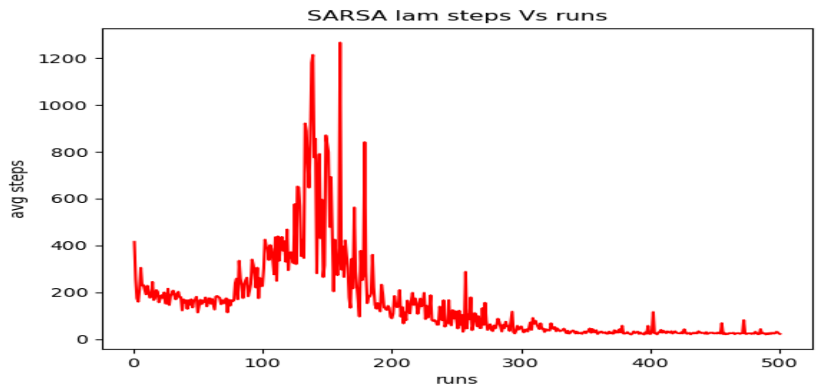
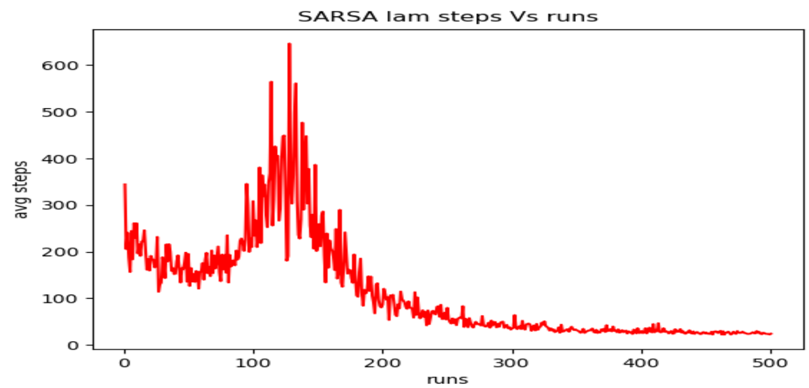


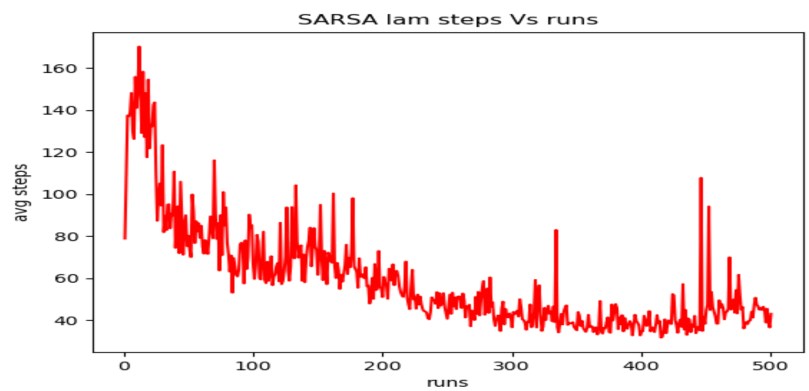
Figure 14: SARSA lambada reward VS episodes with different lambada over 500 episodes and 50 runs with goal B where $\lambda = [0, 0.3, 0.5, 0.9, 0.99, 1]$

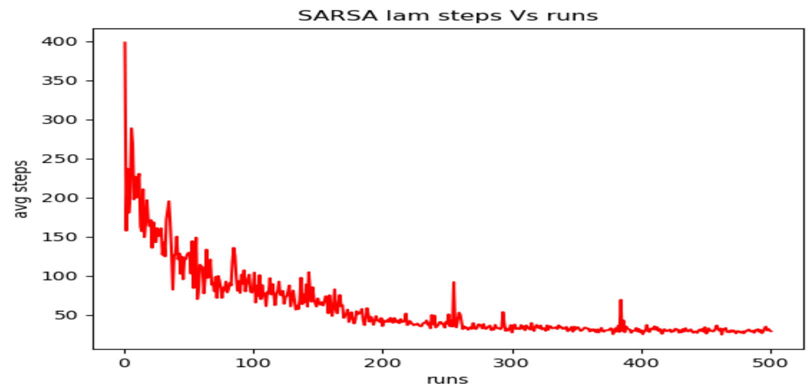
- in SARSA λ whenever lambada value increases we can observe(see below plots) that avg steps size also got decreases.
- all reward points are converges around 500 episodes.
- I have taken offgrid reward as -5 hence minimum reward amount is very low(around -400).
- following plots are of avg steps VS episodes(runs) of SARSA λ with different λs. and we can observed that avg steps are decreasing with λ
- In steps Vs episodes graph we can see more variance it's because of exploration. exploration may take more steps that is varies in all runs.




SARSA lam steps Vs runs
avg steps
runs
0
200
400
600
800
1000
1200
100
300
500


SARSA lam steps Vs runs
avg steps
runs
0
100
200
300
400
500
600


SARSA lam steps Vs runs
avg steps
runs
40
60
80
100
120
140
160
0
200
300
400
500



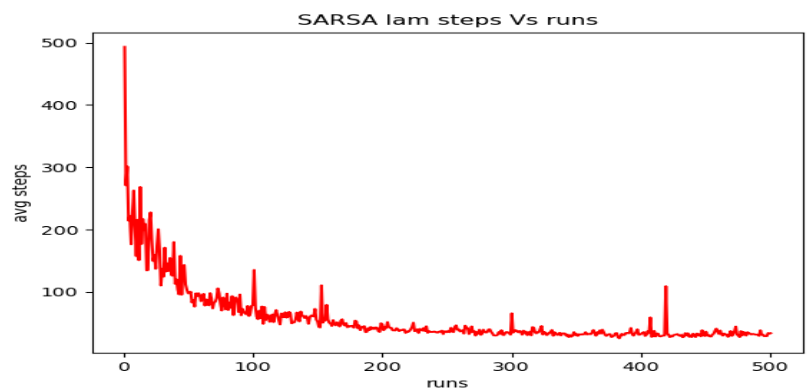
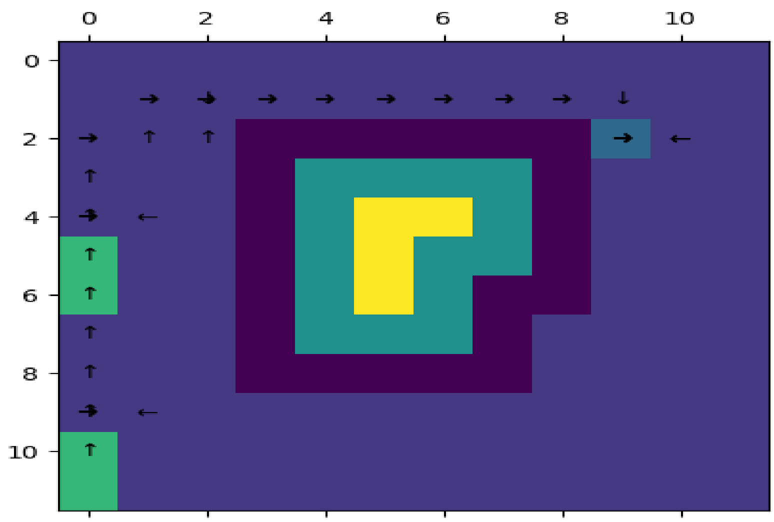
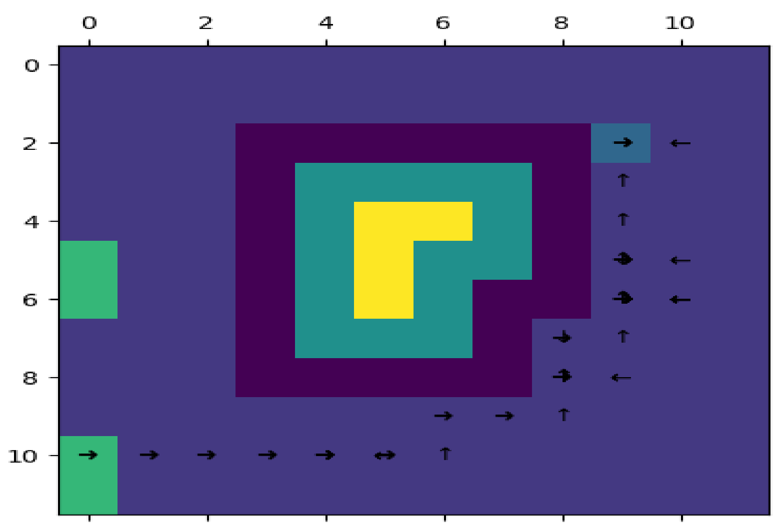
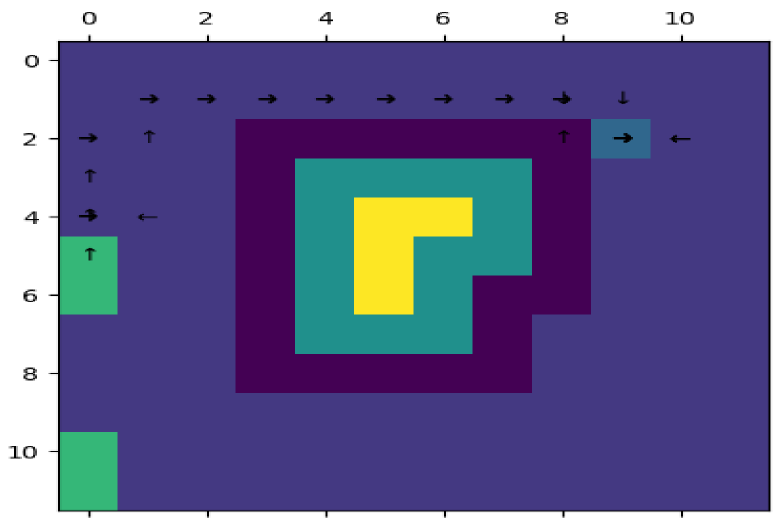


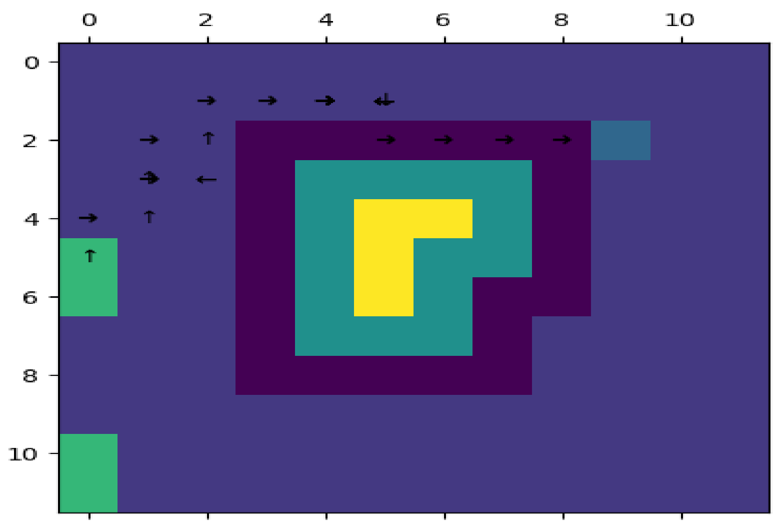
Figure 15: SARSA lambada avg steps VS episodes with different lambada over 500 episodes and 50 runs with goal B where λ =[0, 0.3, 0.5, 0.9, 0.99, 1]

- following policies are optimal policies i got with different λs with 500 episodes and 50 runs. i uploaded policy which is optimal of all 50 runs with 500 episodes so depends on starting state policies may take different path with maximum reward and minimum steps.
- in below mentioned policies arrow indiactes directions and gridworld with different colors have different reward value (e.x. yellow color describes -3 reward,magenta color = -1,skyblue = -2,dark blue = 10(terminal) ,light green = strating states)
- some gridbox have more than one direction discribes more than one actions in same gridbox when it return back to same state.


0 2 4 6 8 10
0 2 4 6 8 10


0 2 4 6 8 10
0 2 4 6 8 10


0 2 4 6 8 10
0 2 4 6 8 10


0 2 4 6 8 10
0 2 4 6 8 10

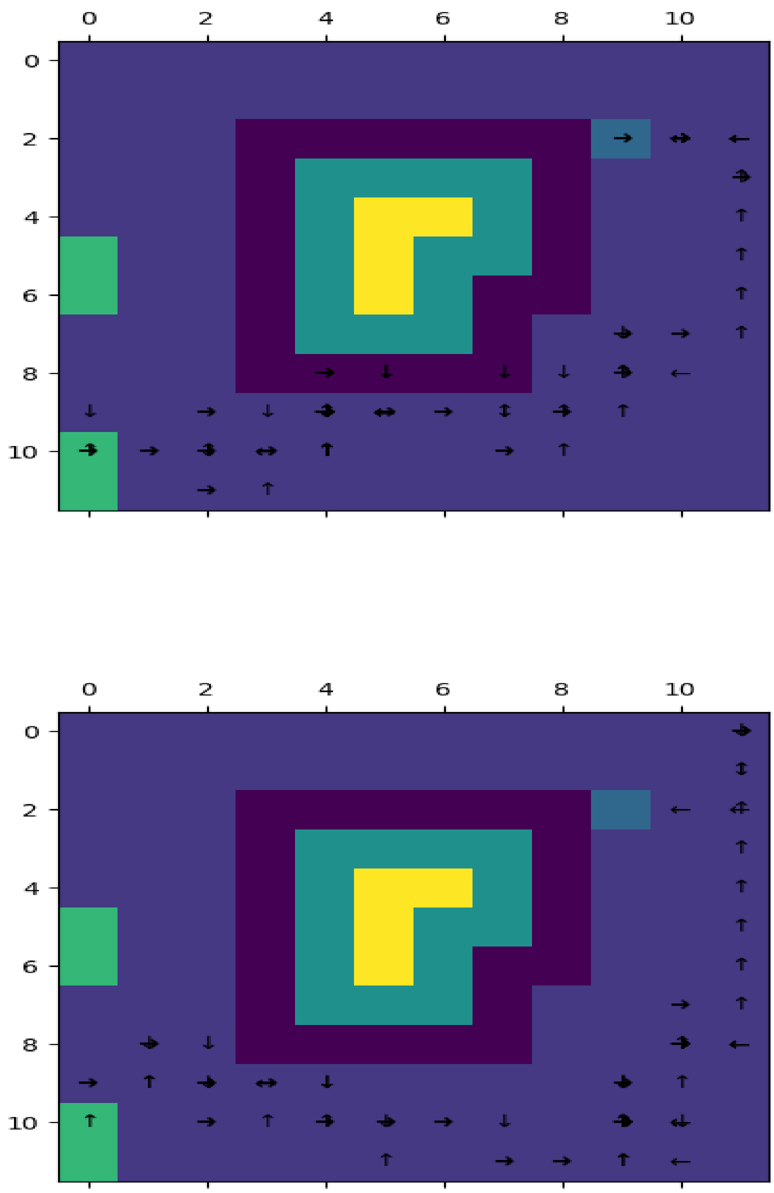


Figure 16: SARSA lambada optimal policies with different lambada over 500 episodes and 50 runs with goal B .where λ =[0, 0.3, 0.5, 0.9, 0.99, 1]

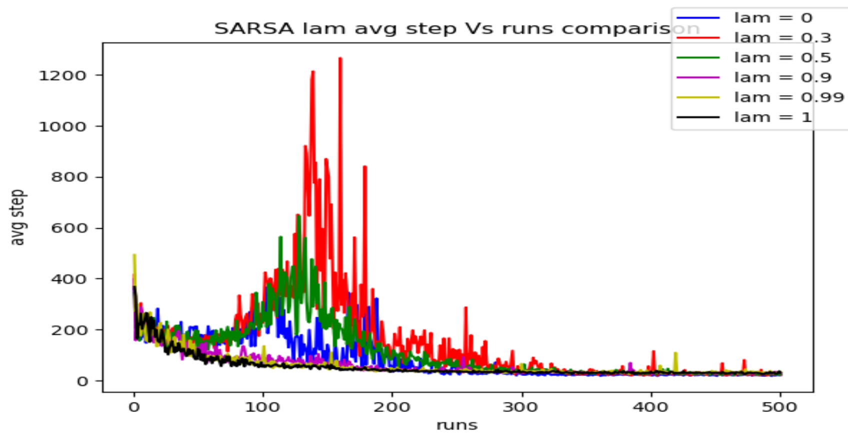


Figure 18: SARSA lambada comparison plots(avg steps VS epsiodes) with different lambada over 500 episodes and 50 runs with goal B .where λ =$[0, 0.3, 0.5, 0.9, 0.99, 1]$

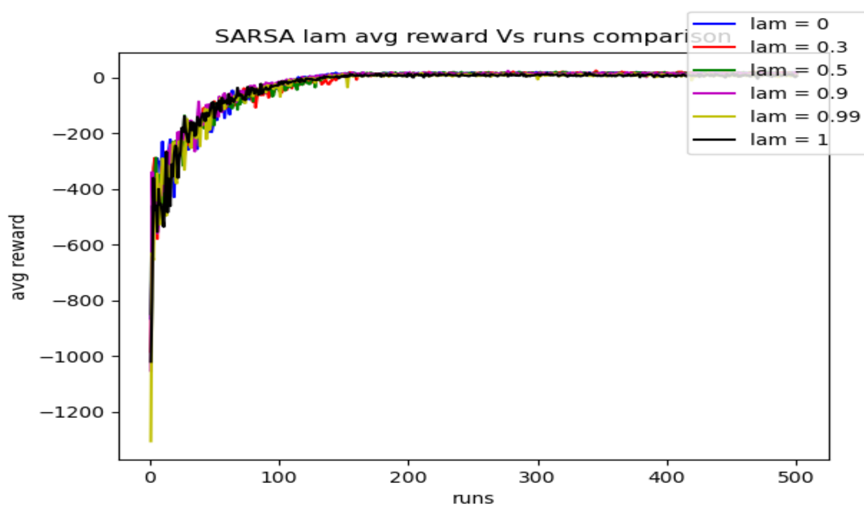


Figure 17: SARSA lambada comparison plots(avg reward VS epsiodes) with different lambada over 500 episodes and 50 runs with goal B .where λ =$[0, 0.3, 0.5, 0.9, 0.99, 1]$