CS6700 : Reinforcement Learning
Programming Assignment #2
Report

Sai Vinay G
CE17B019
Indian Institute of Technology, Madras

April 14, 2019

Problem 1

Puddle World environment

- For problems 2 to 5 we have a grid of size [12,12] and for problem 6 a grid of size 1000 time's previous one, we consider it as [120,1200].
- It has methods for resetting the environment, getting the start positions, getting rewards, selecting actions, performing state transition and methods for making large grid with similar properties of smaller grid.
- The start position are common for all problems except last one, $S = [[6, 0], [7, 0], [10, 0], [11, 0]]$

Problem 2

In this problem we Implement SARSA with 3 goals A,B,C on the grid world with a gamma = 0.9 and plot the average rewards and average number of steps taken to reach the goal, by performing 50 experiments

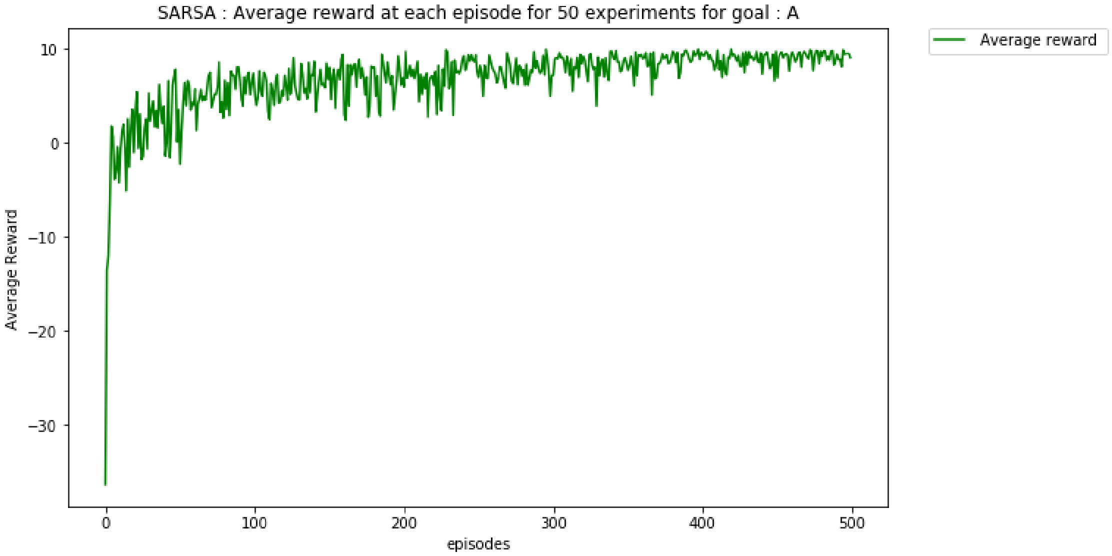


Figure 1: Average performance of SARSA. These data are averages of rewards over 50 experiments with 500 episodes each, for goal 'A'

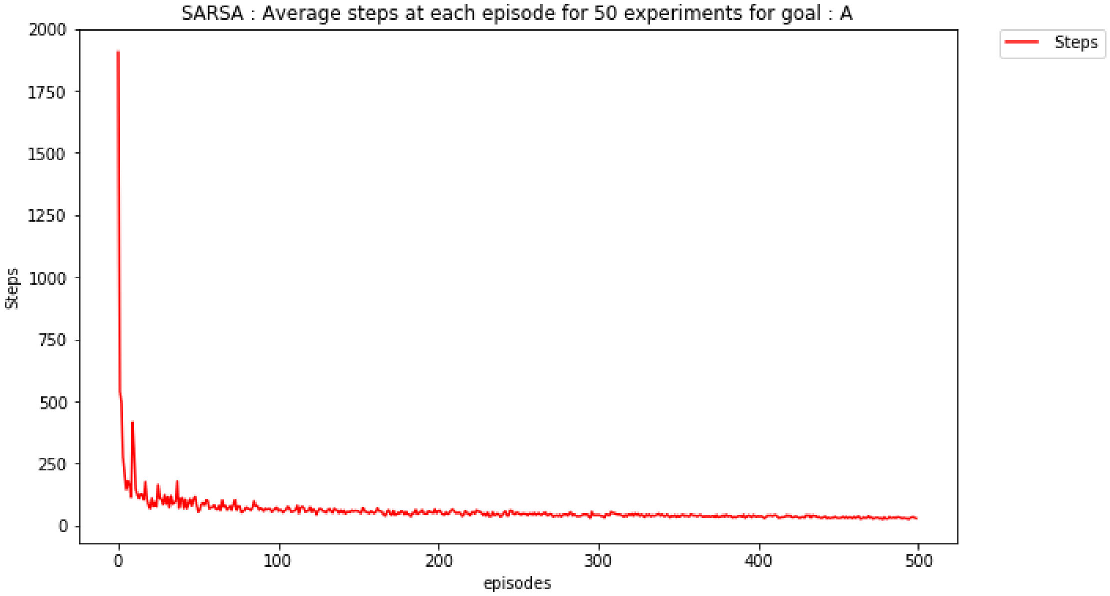
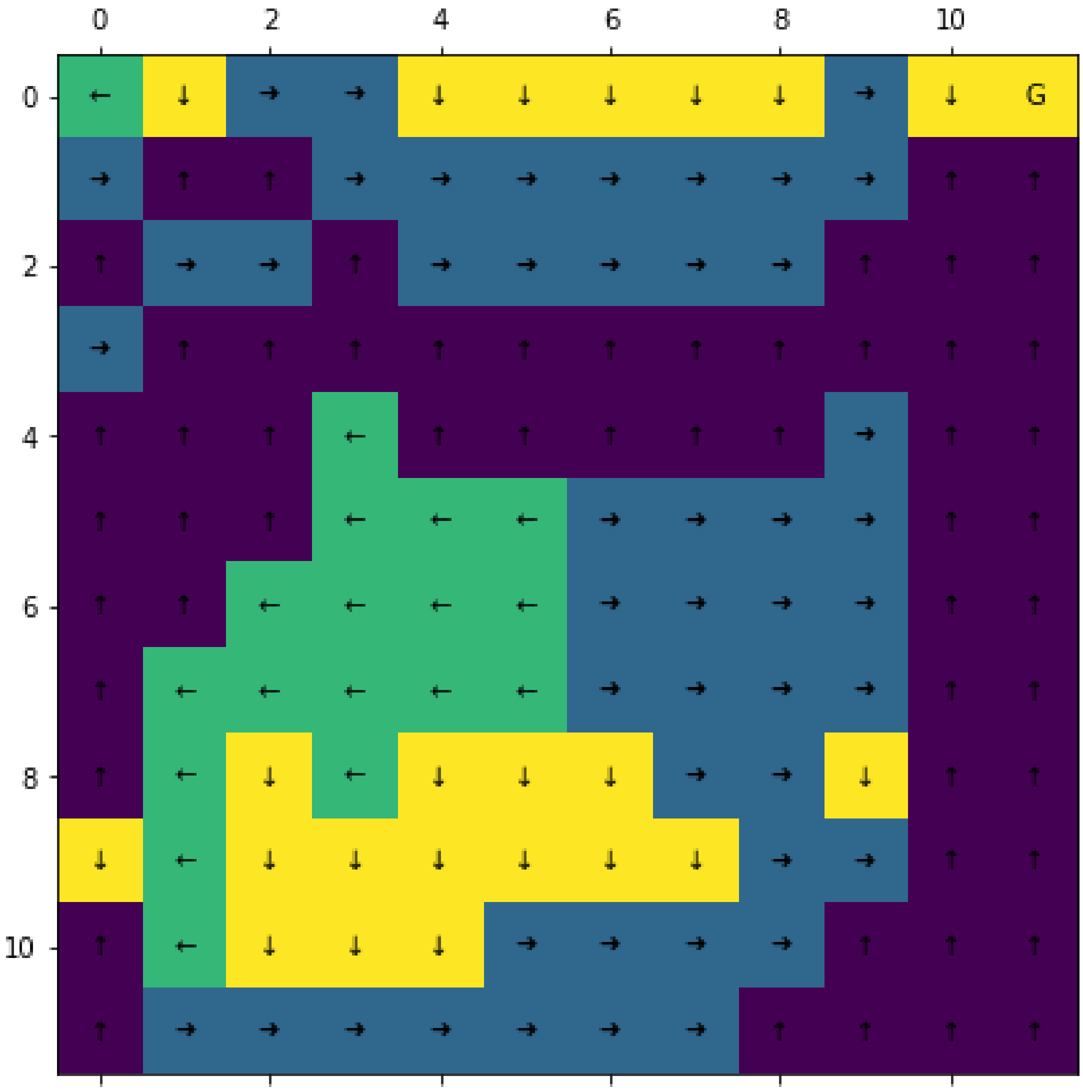


Figure 2: Average performance of SARSA. These data are averages of steps over 50 experiments with 500 episodes each, for goal 'A'



←	↓	→	→	↓	↓	↓	↓	↓	→	↓	G
→	↑	↑	→	→	→	→	→	→	→	↑	↑
↑	→	→	↑	→	→	→	→	→	↑	↑	↑
→	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
↑	↑	↑	←	↑	↑	↑	↑	↑	→	↑	↑
↑	↑	↑	←	←	←	→	→	→	→	↑	↑
↑	↑	←	←	←	←	→	→	→	→	↑	↑
↑	←	←	←	←	←	→	→	→	→	↑	↑
↑	←	↓	←	↓	↓	↓	→	→	↓	↑	↑
↓	←	↓	↓	↓	↓	↓	↓	→	→	↑	↑
↑	←	↓	↓	↓	→	→	→	→	↑	↑	↑
↑	→	→	→	→	→	→	→	↑	↑	↑	↑

Figure 3: SARSA : Policy obtained at the end for goal ’A’

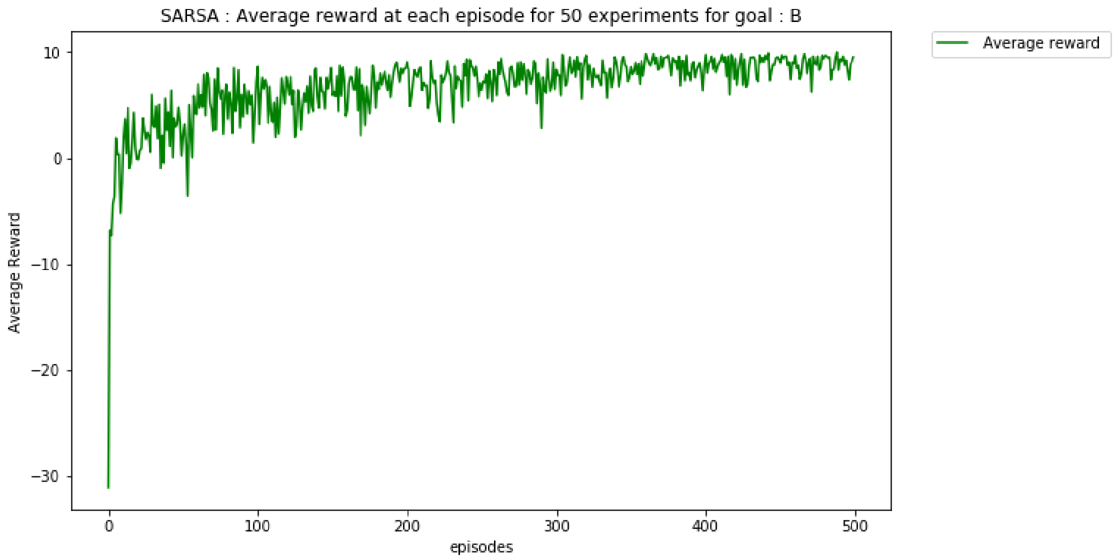


Figure 4: Average performance of SARSA. These data are averages of rewards over 50 experiments with 500 episodes each, for goal 'B'

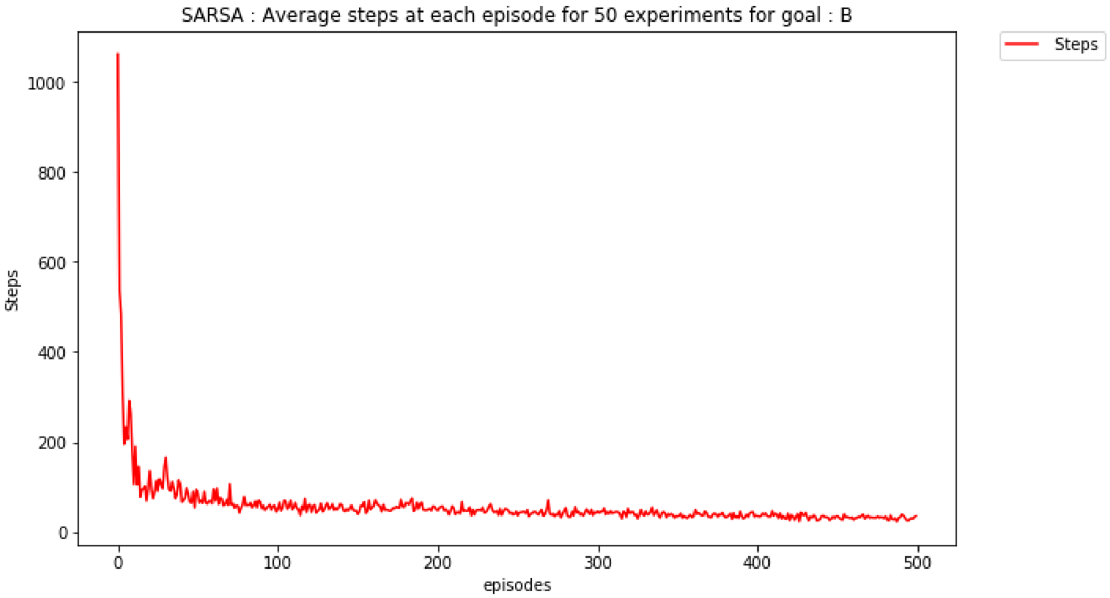


Figure 5: Average performance of SARSA. These data are averages of steps over 50 experiments with 500 episodes each, for goal 'B'

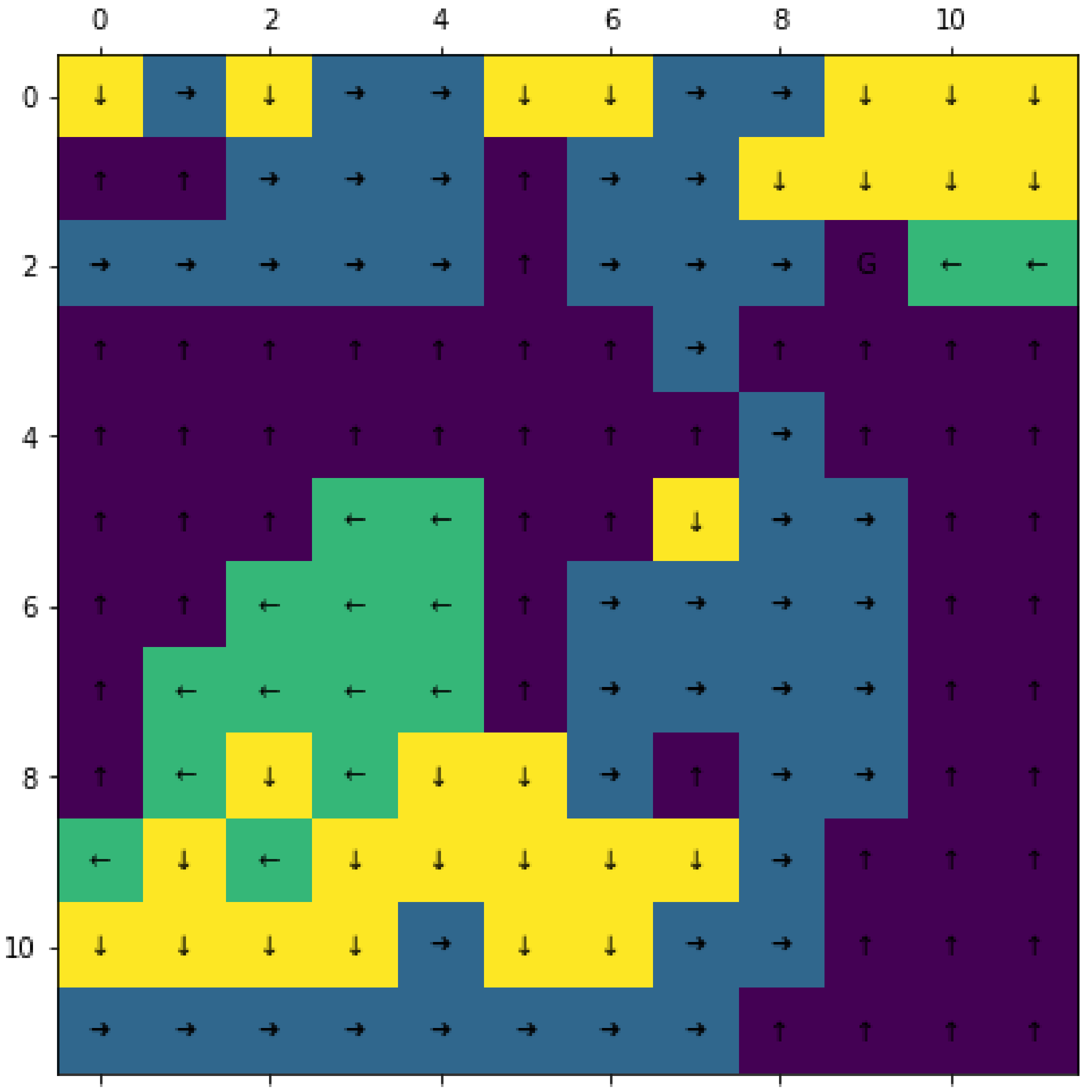


Figure 6: SARSA : Policy obtained at the end for goal ’B’

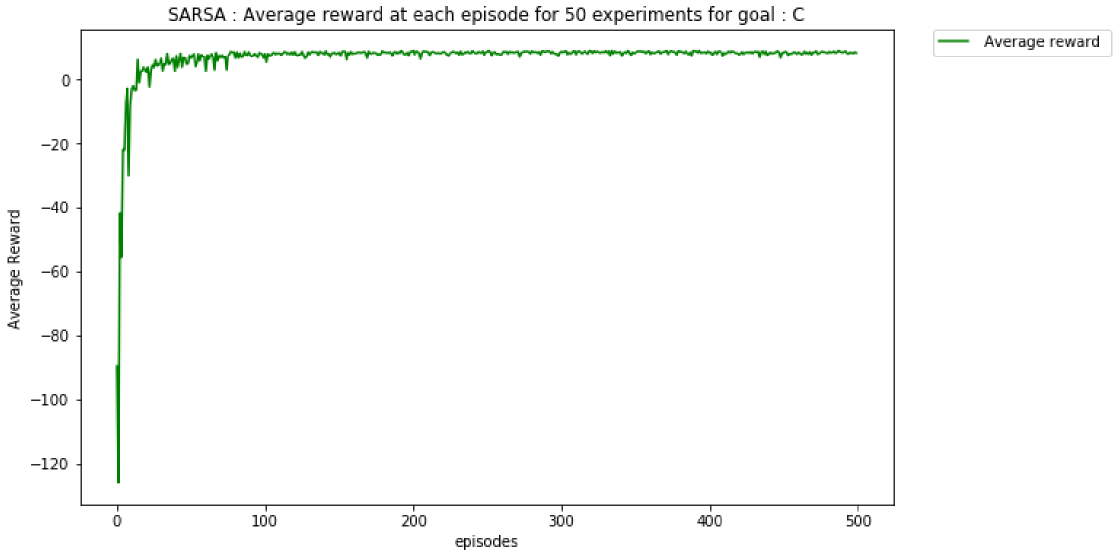


Figure 7: Average performance of SARSA. These data are averages of rewards over 50 experiments with 500 episodes each, for goal 'C'

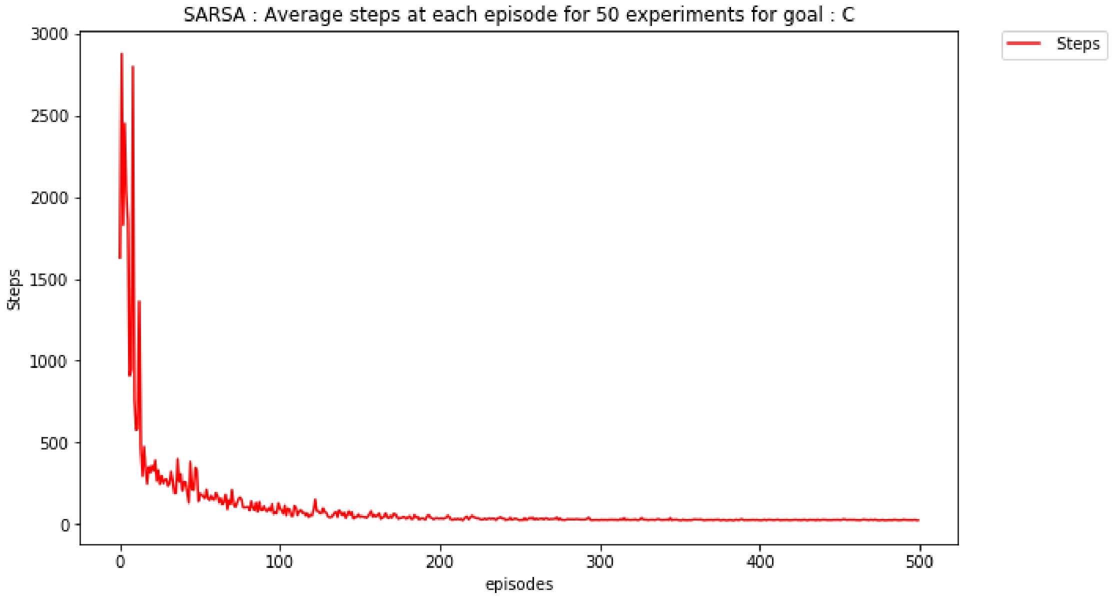


Figure 8: Average performance of SARSA. These data are averages of steps over 50 experiments with 500 episodes each, for goal 'C'

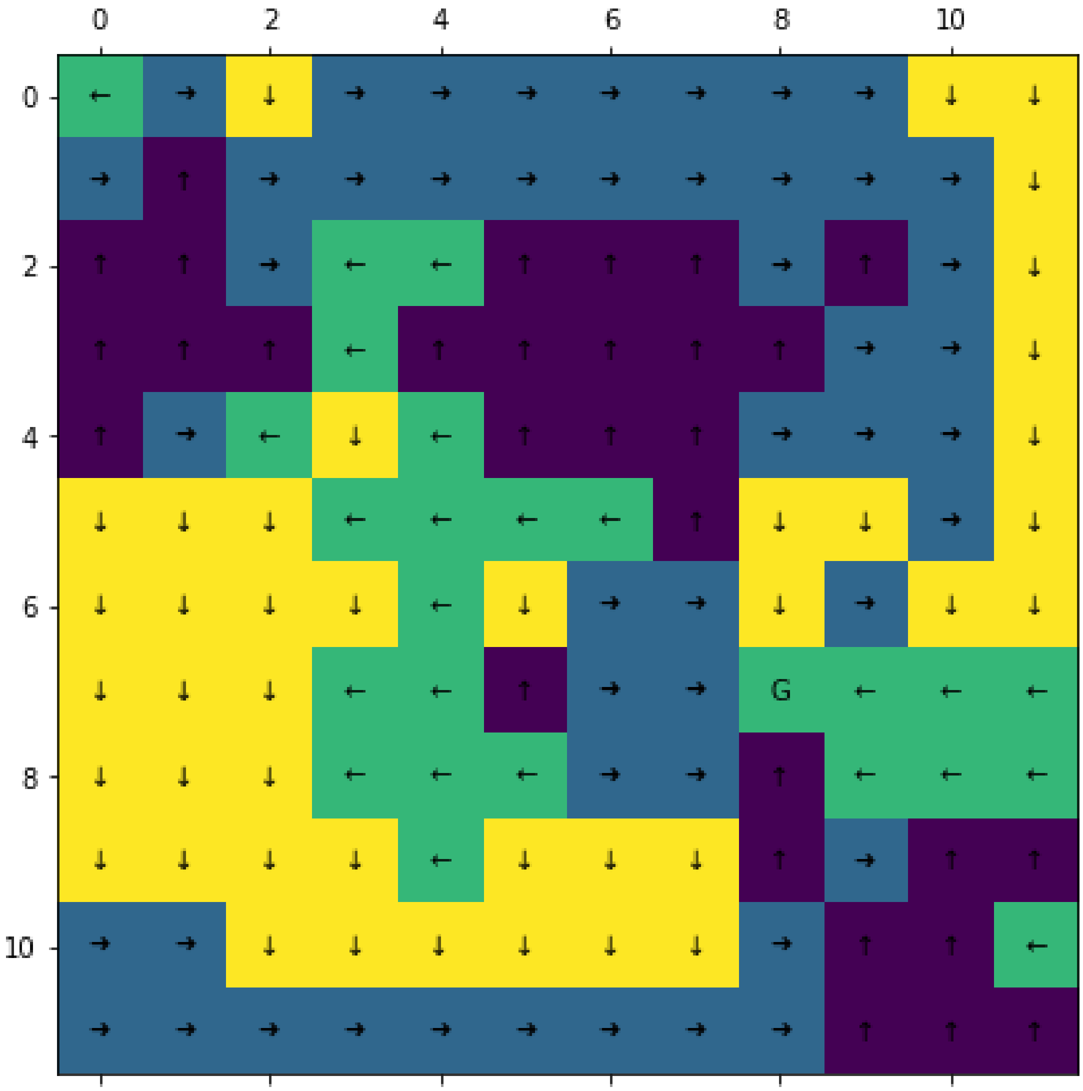


Figure 9: SARSA : Policy obtained at the end for goal 'C'

- We can see that the agent has learned to avoid the puddle and goes around the puddle.
- The reward obtained when moving towards goal C is a bit less as sometimes it falls into the puddle
- We can also see that it has also learn't to counter the westerly winds by taking a larger path to reach goal.

Problem 3

In this problem we implement the SARSA(λ). For $\lambda = [0; 0.3; 0.5; 0.9; 0.99; 1.0]$.

- Here given are the plots of average reward vs episodes, obtained by performing 50 experiments with 500 episodes, goal as 'A'. Here we consider accumulating traces.
- We can see that there is not much variation

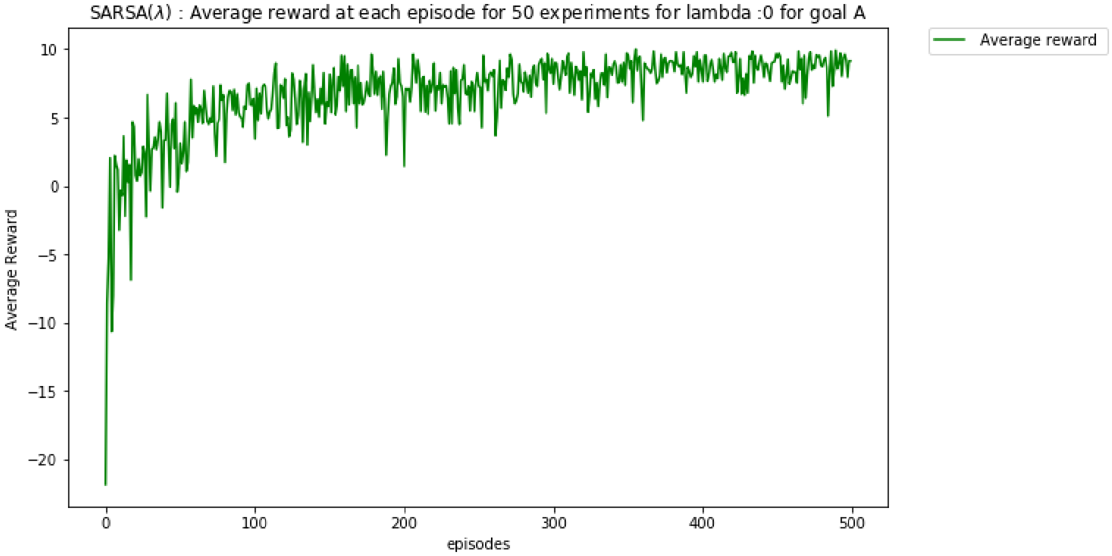


Figure 10: Average performance of SARSA(λ) for $\lambda = 0$. These data are average rewards over 50 experiments with different 500 episodes each

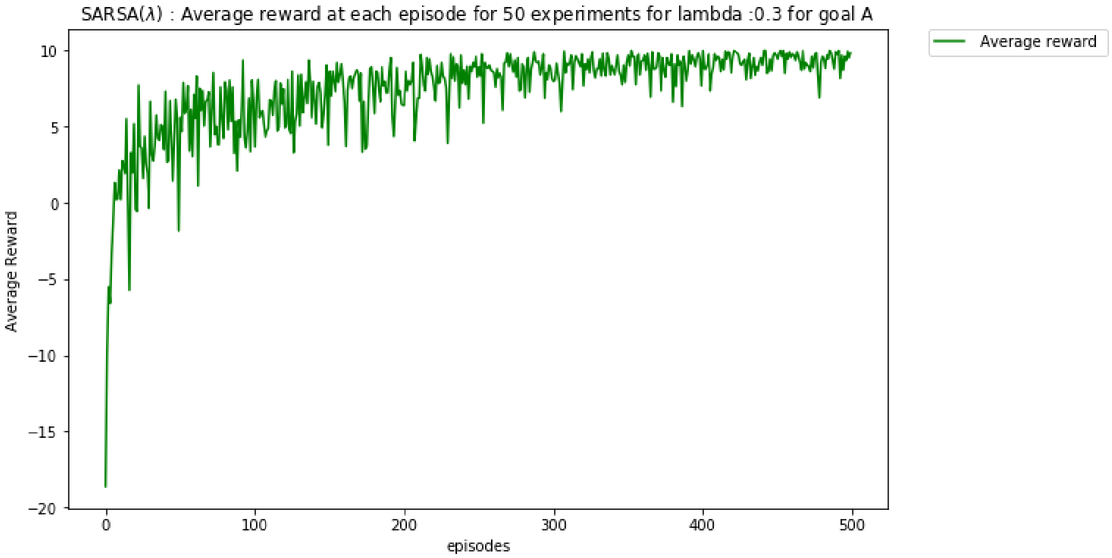


Figure 11: Average performance of SARSA(λ) for $\lambda = 0.3$. These data are average rewards over 50 experiments with different 500 episodes each

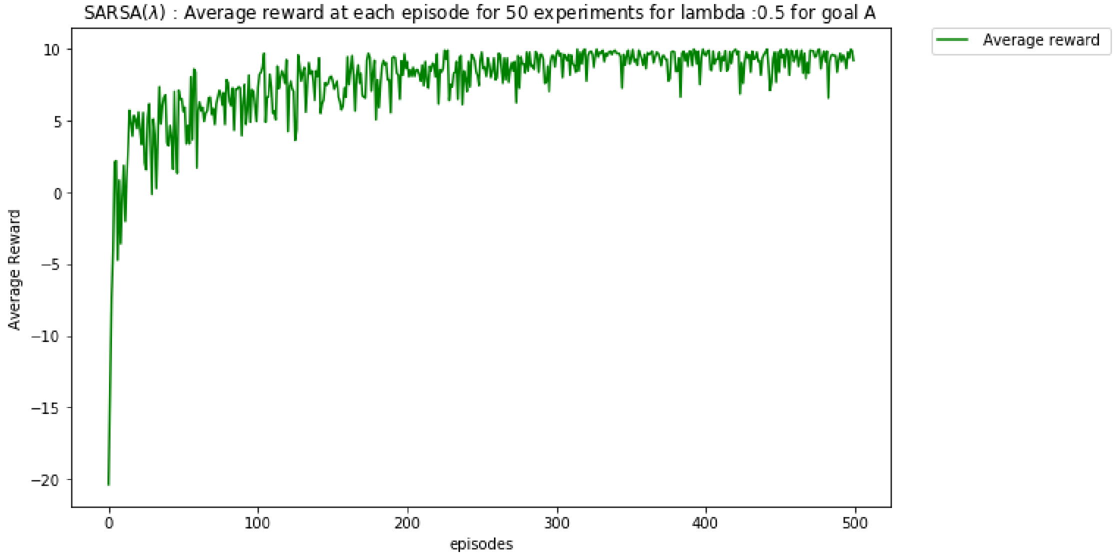


Figure 12: Average performance of SARSA(λ) for $\lambda = 0.5$. These data are average rewards over 50 experiments with different 500 episodes each

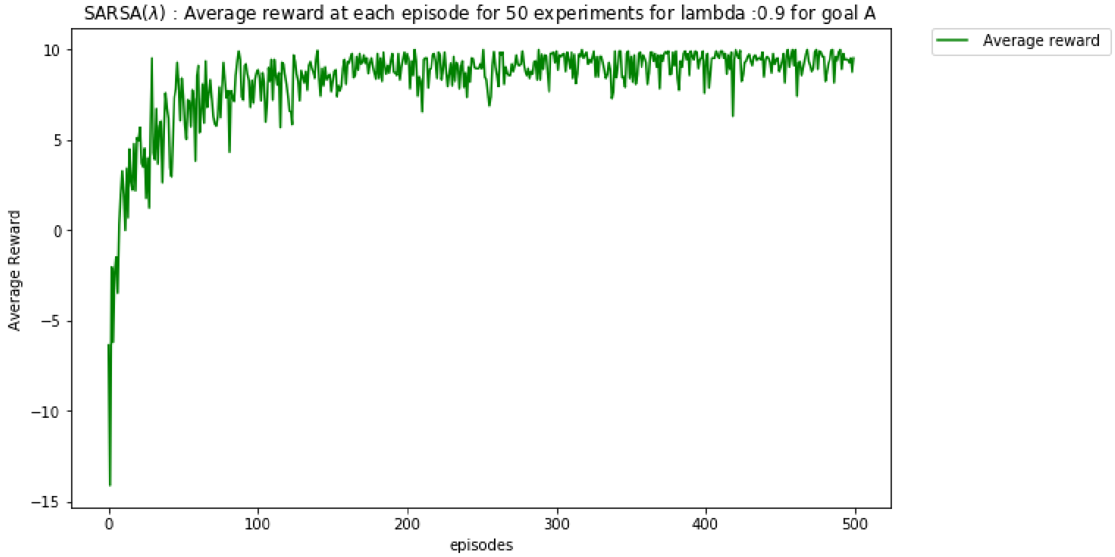


Figure 13: Average performance of SARSA(λ) for $\lambda = 0.9$. These data are average rewards over 50 experiments with different 500 episodes each

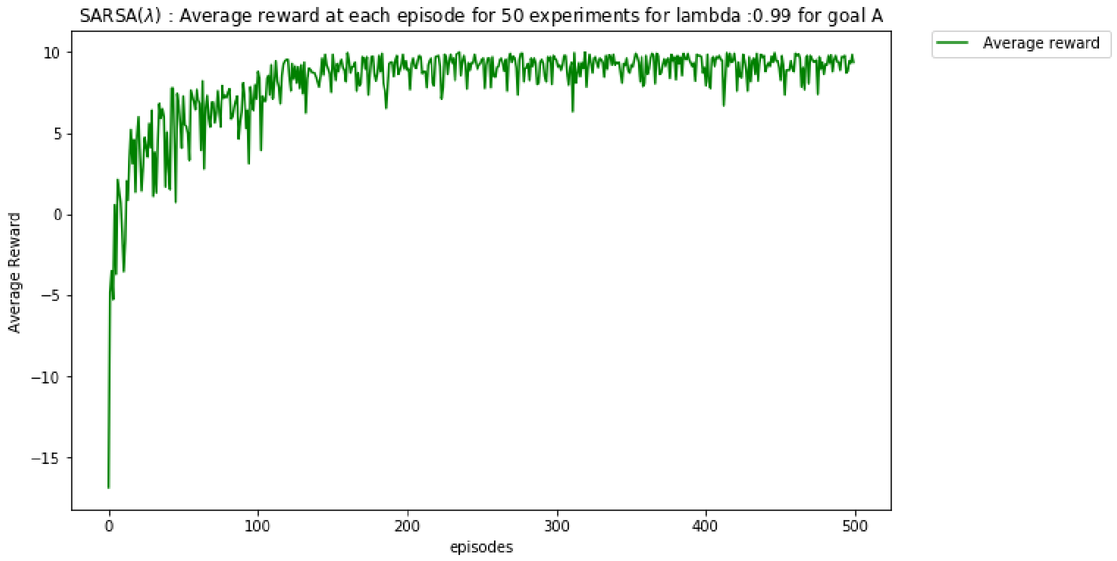


Figure 14: Average performance of SARSA(λ) for $\lambda = 0.99$. These data are average rewards over 50 experiments with different 500 episodes each

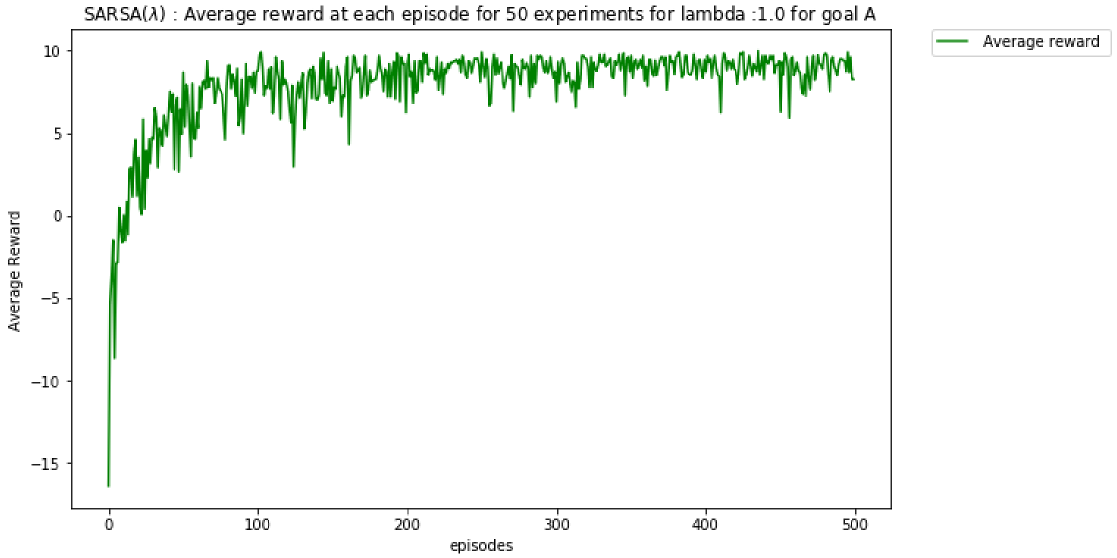


Figure 15: Average performance of SARSA(λ) for $\lambda = 1$. These data are average rewards over 50 experiments with different 500 episodes each

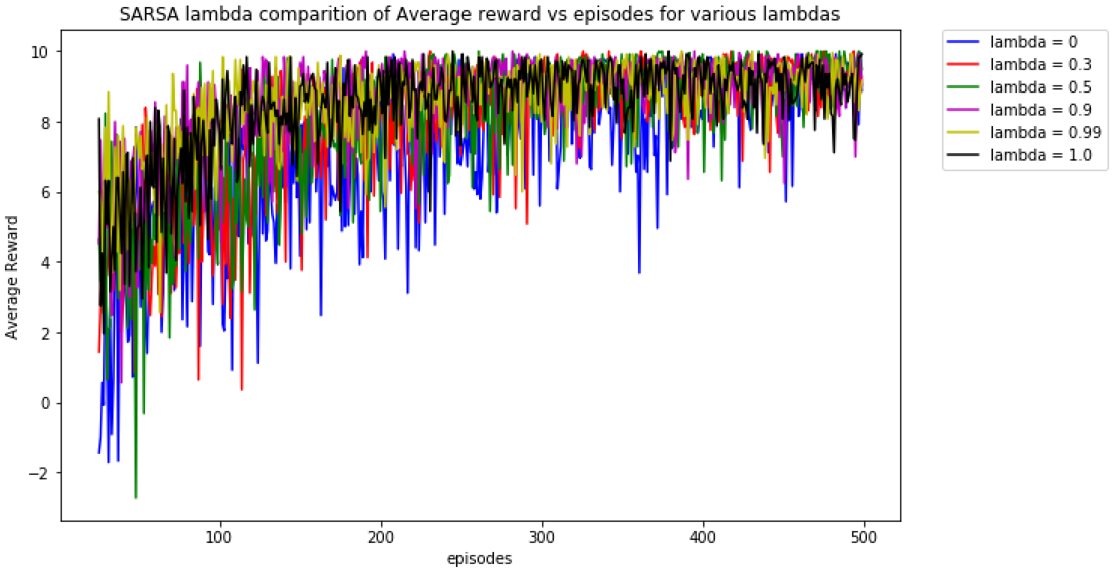


Figure 16: Average performance of SARSA(λ) for various lambdas. These data are average rewards over 50 experiments with different 500 episodes each

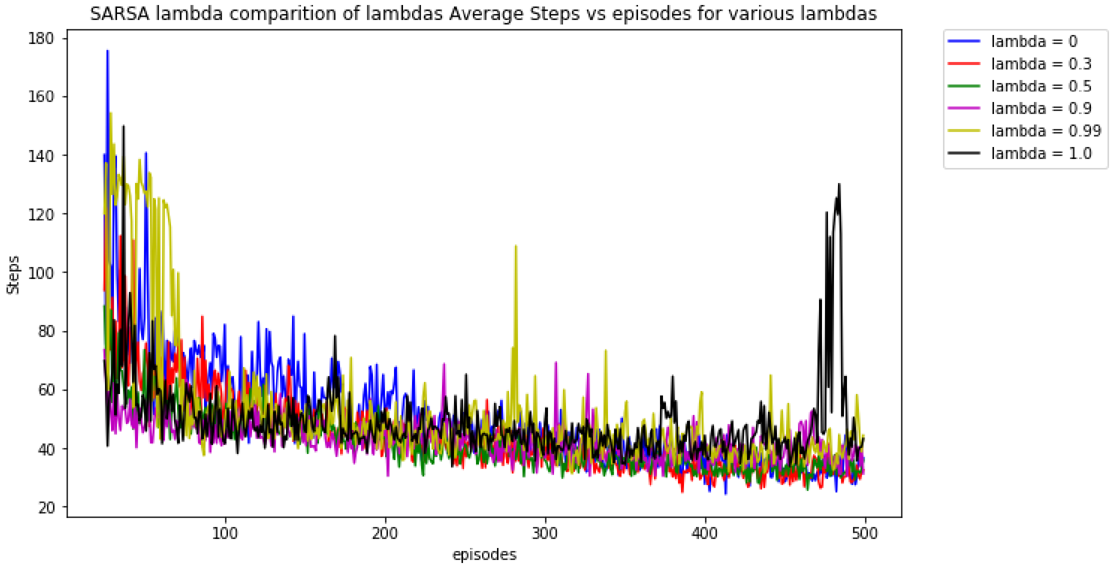


Figure 17: Average performance of SARSA(λ) for various lambdas. These data are the number of steps over 50 experiments with different 500 episodes each

- Here we show the comparison of all lambdas in one plot for average reward vs episodes and steps vs episodes.
- We can see there is not much variation for the goal 'A'
- But we can expect there to be a variation if we plot for goal 'C' as for λ close to 1 may not perform well.

Problem 4

In this problem we implement the Monte carlo policy gradient.with $\alpha = 0.1, \gamma = 0.9$

- Here given are the plots of average reward vs episodes, obtained by performing 50 experiments with 500 episodes, goal as 'A, B, C'.

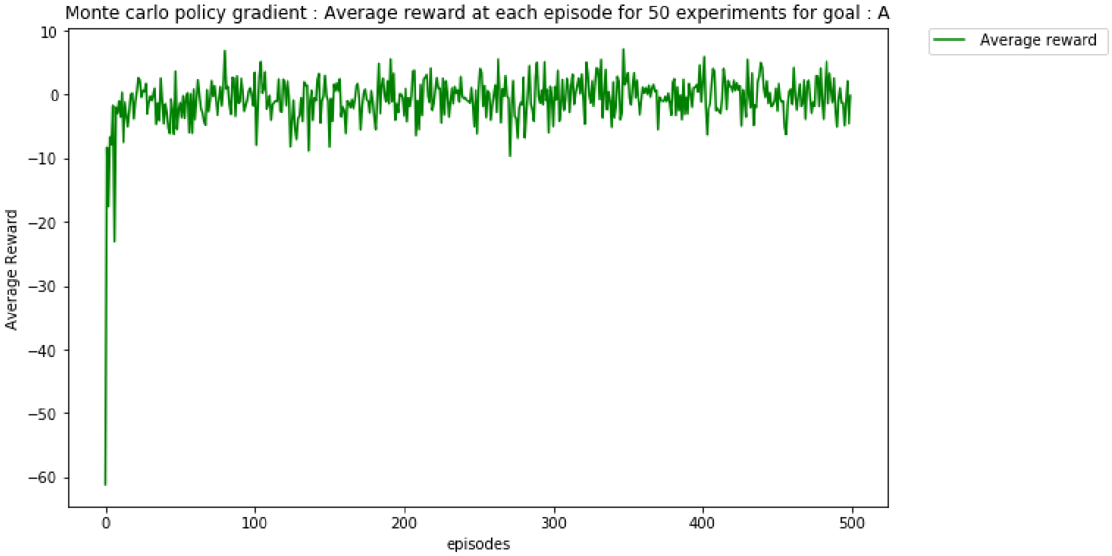


Figure 18: Average performance of Monte carlo policy gradient. These data are averages of rewards over 50 experiments with 500 episodes each, for goal 'A'

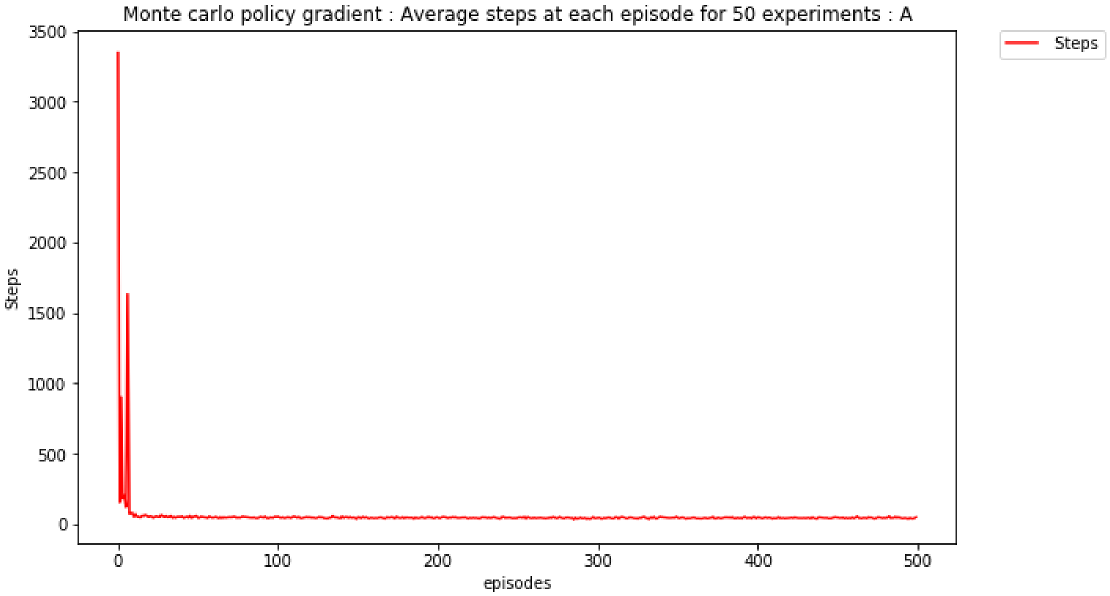
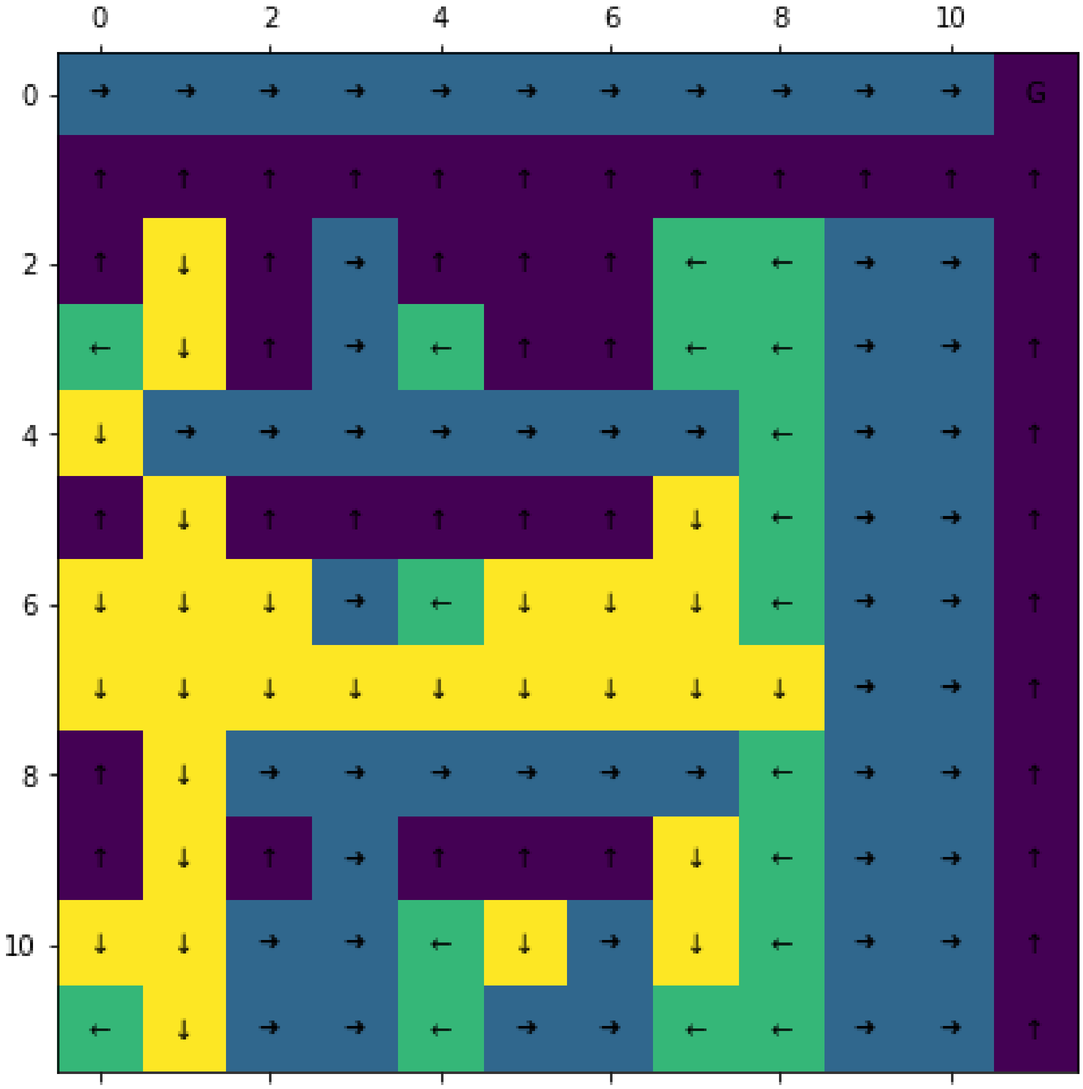


Figure 19: Average performance of Monte carlo policy gradient. These data are averages of steps over 50 experiments with 500 episodes each, for goal 'A'



	0		2		4		6		8		10	
0	→	→	→	→	→	→	→	→	→	→	→	G
	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
2	↑	↓	↑	→	↑	↑	↑	←	←	→	→	↑
	←	↓	↑	→	←	↑	↑	←	←	→	→	↑
4	↓	→	→	→	→	→	→	→	←	→	→	↑
	↑	↓	↑	↑	↑	↑	↑	↓	←	→	→	↑
6	↓	↓	↓	→	←	↓	↓	↓	←	→	→	↑
	↓	↓	↓	↓	↓	↓	↓	↓	↓	→	→	↑
8	↑	↓	→	→	→	→	→	→	←	→	→	↑
	↑	↓	↑	→	↑	↑	↑	↓	←	→	→	↑
10	↓	↓	→	→	←	↓	→	↓	←	→	→	↑
	←	↓	→	→	←	→	→	←	←	→	→	↑

Figure 20: Monte carlo policy gradient : Policy obtained at the end for goal ’A’

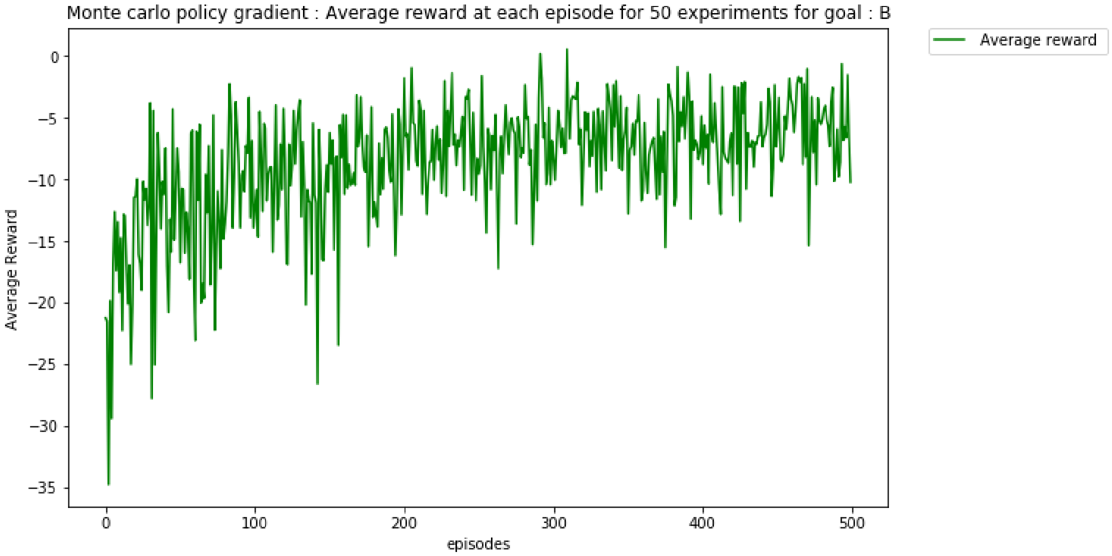


Figure 21: Average performance of Monte carlo policy gradient. These data are averages of rewards over 50 experiments with 500 episodes each, for goal 'B'

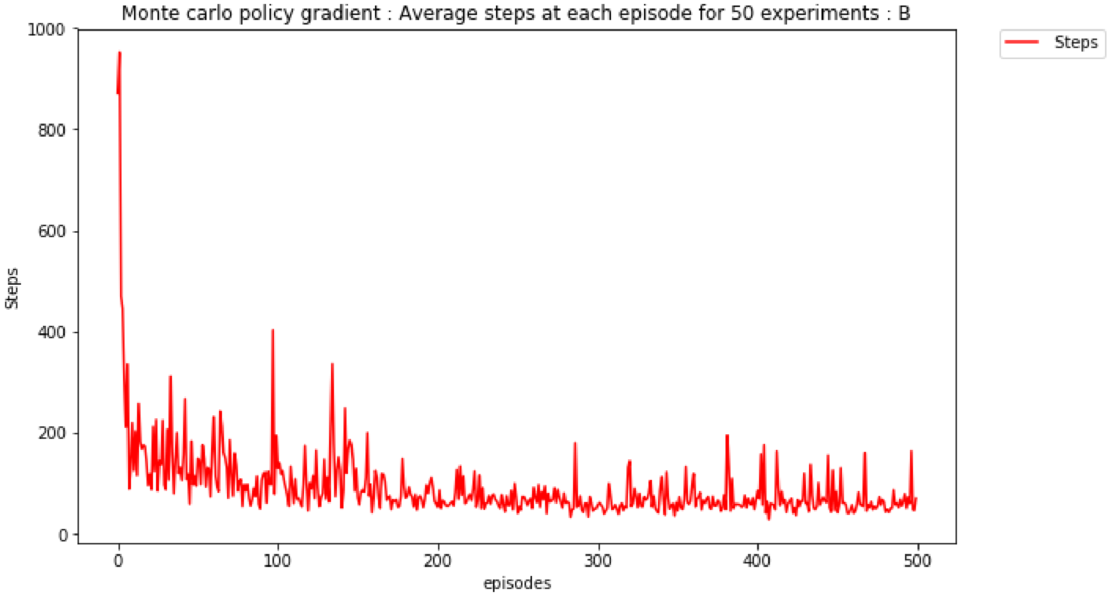


Figure 22: Average performance of Monte carlo policy gradient. These data are averages of steps over 50 experiments with 500 episodes each, for goal 'B'

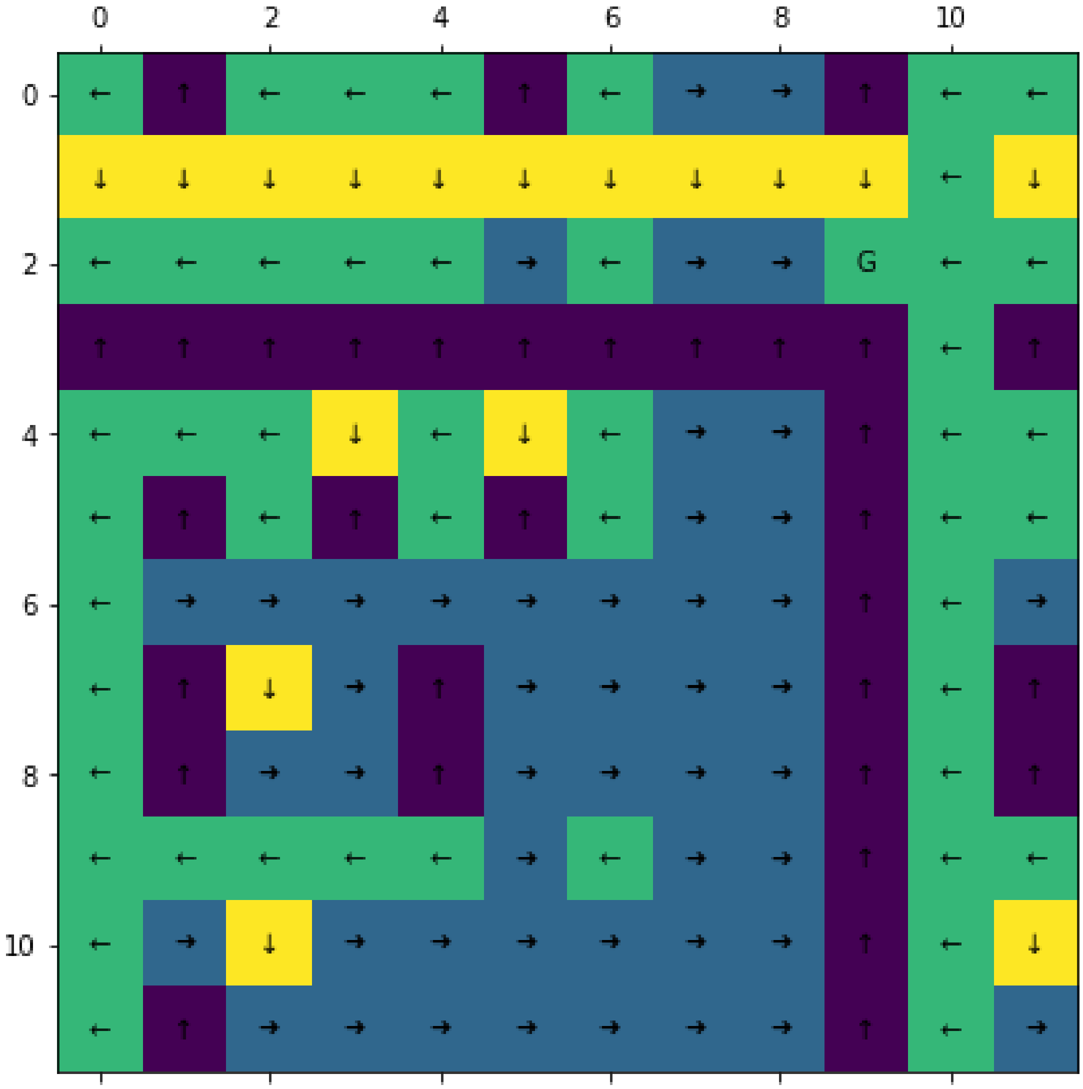


Figure 23: Monte carlo policy gradient : Policy obtained at the end for goal 'B'

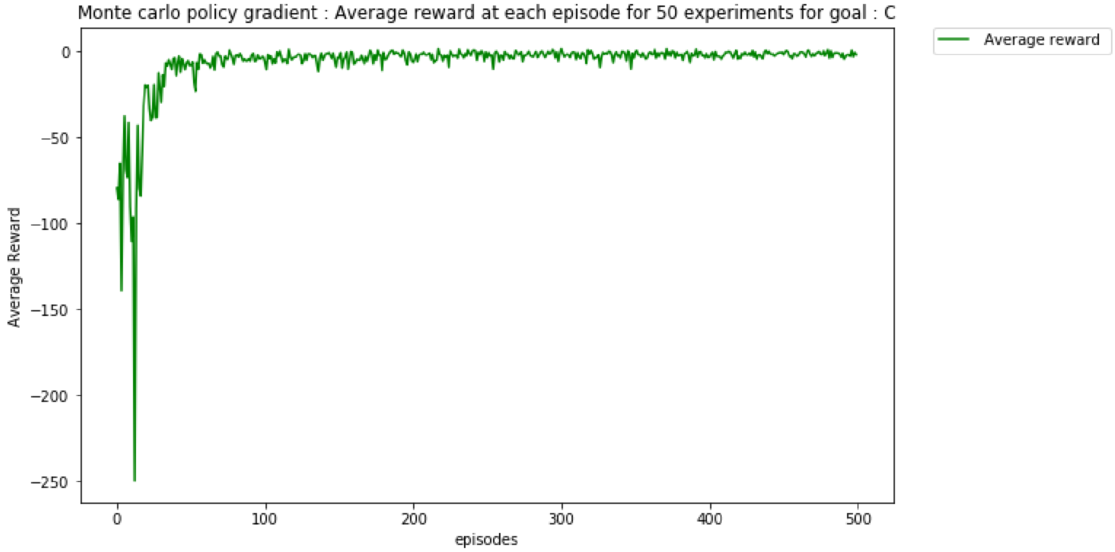


Figure 24: Average performance of Monte carlo policy gradient. These data are averages of rewards over 50 experiments with 500 episodes each, for goal ’C’

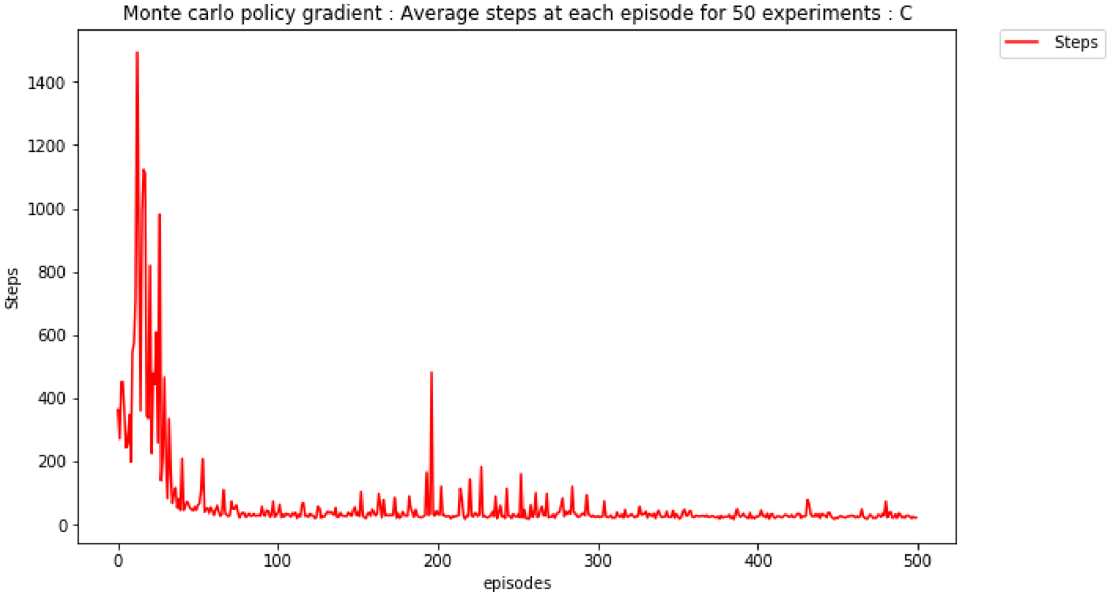


Figure 25: Average performance of Monte carlo policy gradient. These data are averages of steps over 50 experiments with 500 episodes each, for goal ’C’

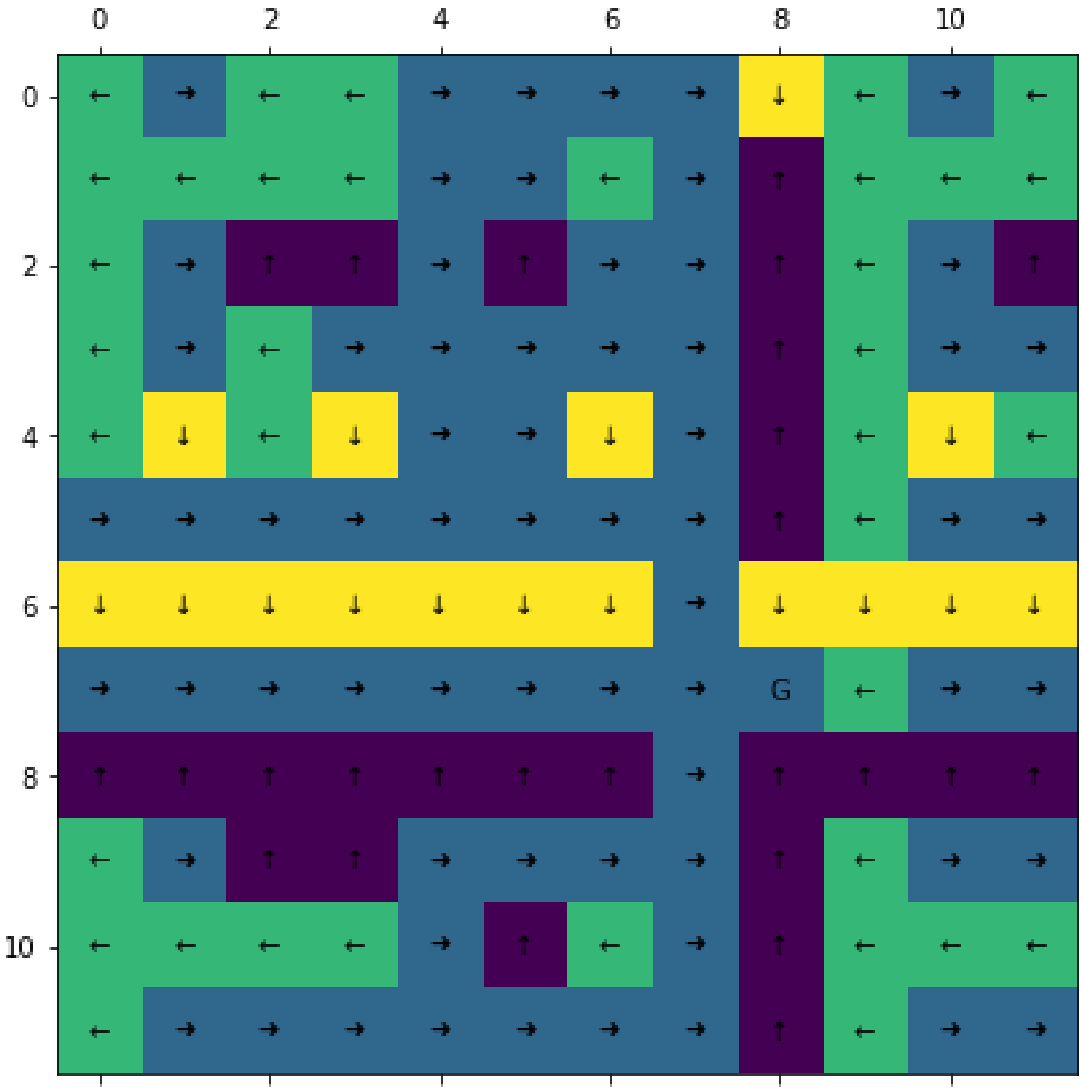


Figure 26: Monte carlo policy gradient : Policy obtained at the end for goal 'B'

- Here we can see that unlike SARSA, the agent is trying to go through the puddle to reach the goal for goals 'B' and 'C' , which are close to puddle.

Problem 5

(a) Monte Carlo Policy Gradient :

We here first consider a policy parameterised by θ i.e, $\pi(a|s, \theta)$, differentiable for every action in action action, state in state space and theta belongs to real numbers.

The update equation is :

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

where,

$$J(\theta) = \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi_\theta(a|s) q^{\pi}(s,a)$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} q^{\pi}(s,a) \nabla_\theta \pi_\theta(a|s) \\
&= \sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi_\theta(a|s) q^{\pi}(s,a) \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \\
&= E_\pi[Q^{\pi}(s,a) \nabla_\theta \ln \pi\theta(a|s)]
\end{aligned}$$

As $q^{\pi}(s,a)$ is the expected value of G_t i.e, $q^{\pi}(s,a) = E[G_t|s_t, a_t]$, the expectation of sample gradient will be equal to the actual gradient.
We can write above equation as

$$\nabla_\theta J(\theta) = E_\pi[G_t \nabla_\theta \ln \pi_\theta(a|s)]$$

Therefore ,

$$\theta \leftarrow \theta + \alpha G_t \nabla_\theta \ln \pi_\theta(a|s)$$

As, our policy is softmax, with sum of two parameters for the preference of an action, we can use chain rule to obtain the update equation of each parameter.
Consider update for one parameter of chosen action. let it be north at state i,j.

$$\begin{aligned}
\frac{\partial}{\partial \theta_x(N,i)}(\ln e^{\theta_x(N,i)+\theta_y(N,j)}) - (\ln \sum_{a,p,q} e^{\theta_x(a,p)+\theta_y(a,q)}) &= \frac{\partial}{\partial(\theta_x(N,i)+\theta_y(N,j))}\Bigg(\ln e^{\theta_x(N,i)+\theta_y(N,j)} \\
- \ln \sum_{a,p,q} e^{\theta_x(a,p)+\theta_y(a,q)}\Bigg) * \frac{\partial \theta_x(N,i) + \theta_y(N,j)}{\partial \theta_x(N,i)} \\
&= (1 - \frac{e^{\theta_x(N,i)+\theta_y(N,j)}}{\sum_{a,p,q} e^{\theta_x(a,p)+\theta_y(a,q)}}) \\
&= (1 - \pi_\theta(a|s,\theta)) * 1
\end{aligned}$$

(b) From the above plots we can see that it doesn't as good as SARSA(λ). So we can conclude that this parameterization is not good for grid world .

Problem 6

In this problem we perform SARSA(λ) with $\lambda = [0.1, 0.5]$ on a 1000x previous grid world.

- In this case we increase the gamma, so that we can update the state which are far away from the goal.

- We also explore more in the start and decrease the ϵ as we proceed as the grid is very big.

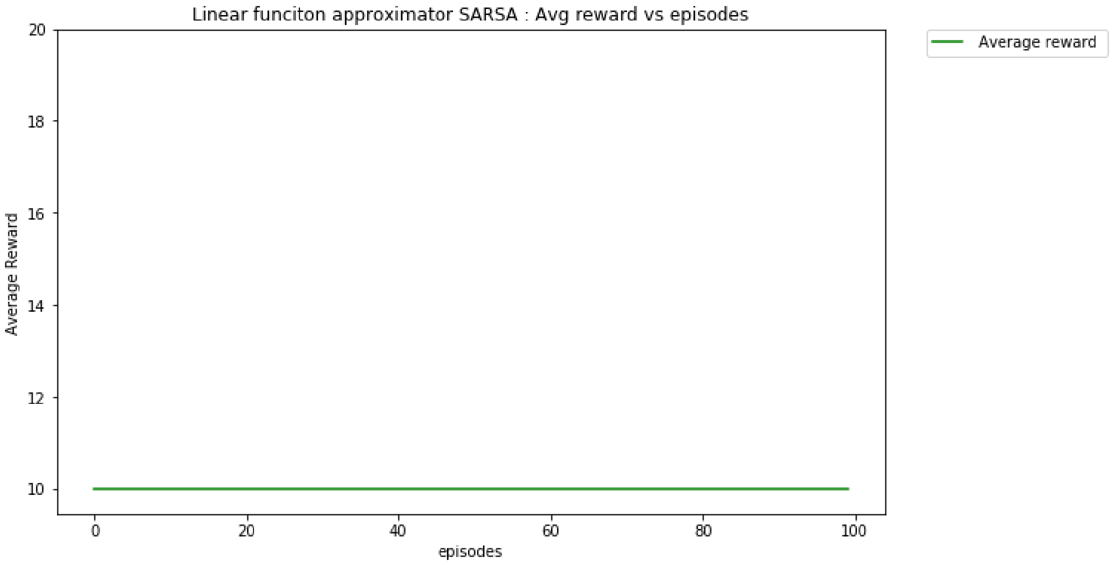


Figure 27: Linear function approximation SARSA for $\lambda = 0.1$: Average Reward Vs episodes in large grid world

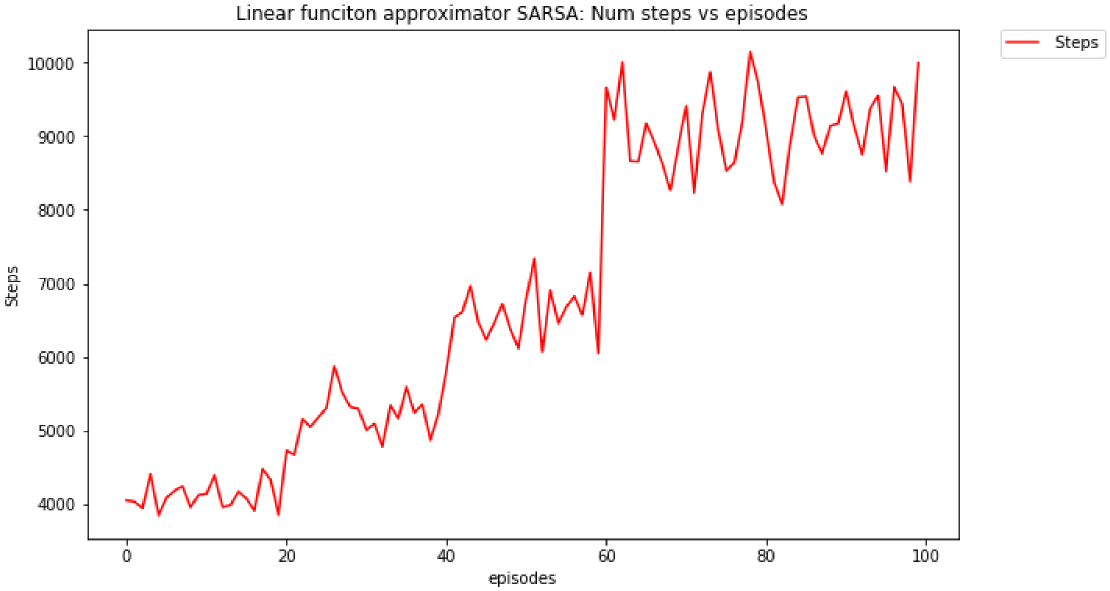


Figure 28: Linear function approximation SARSA for $\lambda = 0.1$: Steps Vs episodes in large grid world

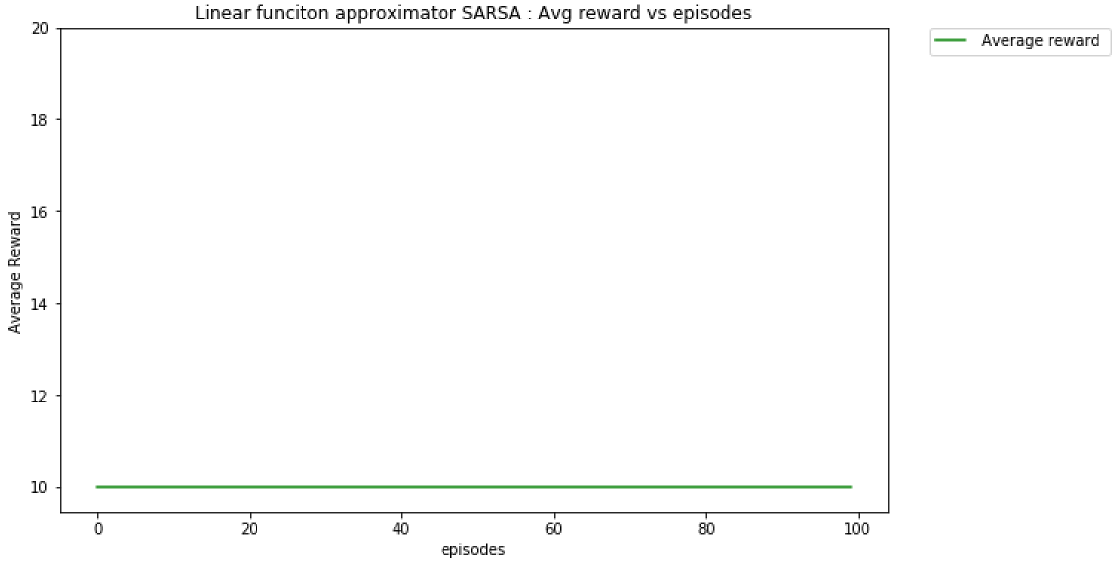


Figure 29: Linear function approximation SARSA for $\lambda = 0.5$: Average Reward Vs episodes in large grid world

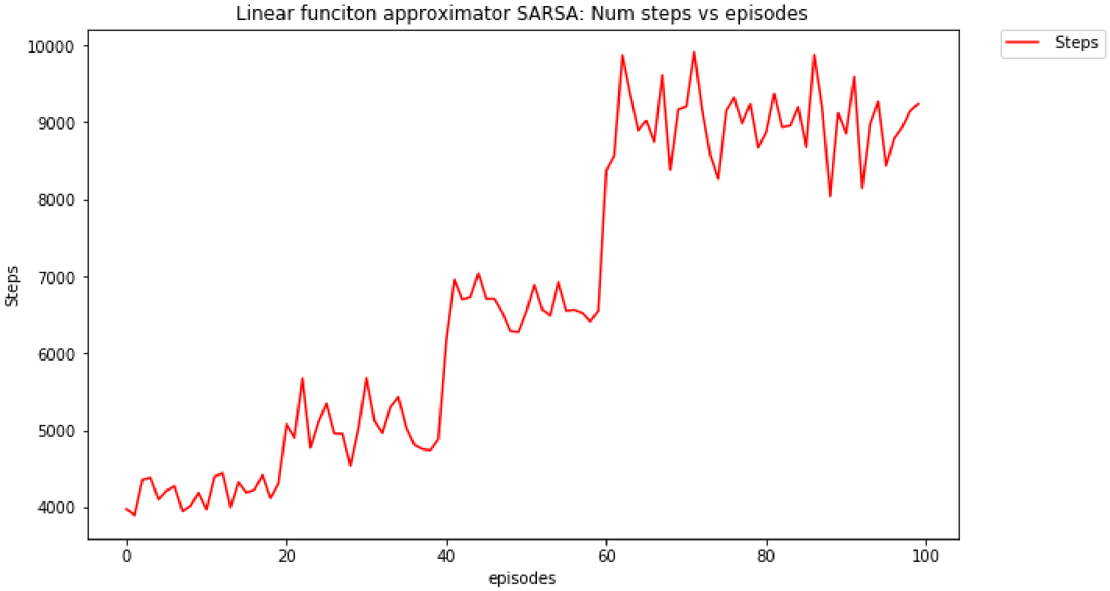


Figure 30: Linear function approximation SARSA for $\lambda = 0.5$: Steps Vs episodes in large grid world

- From the graph's its difficult to tell about convergence. And whiling experimenting, we can find that gamma value effects a lot in this case, for lower values of gamma, it takes more than 100000 steps to reach the goal.

References

[1] matplotlib. https://matplotlib.org/tutorials/introductory/pyplot.html.

[2] Numpy. http://www.numpy.org/.

[3] lilianweng. Gradient-descent methods. [Online]. Available: https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html

[4] D. silver. Value function approximation. [Online]. Available: http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching$_{f}iles/FA.pdf$

[5] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning.* MIT press Cambridge, 1998, vol. 135.