REINFORCEMENT LEARNING PROJECT

Taxi learning problem

By:
Najeeb Jilani & Ahmed Atif

TABLE OF CONTENTS

01
Introduction

Introduction

The Problem states:

THE TAXI DRIVES TO THE PASSENGER'S LOCATION, PICKS UP THE PASSENGER, DRIVES TO THE PASSENGER'S DESTINATION, AND THEN DROPS OFF THE PASSENGER BY USING THE SHORTEST OPTIMAL PATH.

02

Q-learning & Sarsa learning

Background understanding

When Reinforcement learning Agent is playing a game, it does two things:

1. Taking action - **Behavior Policy**

2. Learning which actions are good or bad in a given state. Using this learning agent updates its estimates of Q values. The agent has to use a policy to update its estimate of Q-values - **Target Policy**

Background understanding

Behavior Policy:

The policy that the agent uses to determine its action(behaviour) in a given state.

Target Policy

The policy that the agent uses to learn from the rewards received for its actions,
i.e. to determine updated updated Q-value

ON Policy & OFF policy Learner

ON Policy

The target policy is same from the behaviour policy.

OFF Policy

The target policy is different from the behaviour policy.

Q-Learning

- MODEL-FREE REINFORCEMENT LEARNING
- OFF POLICY
- BEHAVIOUR POLICY IS DIFFERENT FROM THE TARGET POLICY
- MAKE A Q-TABLE
- UPDATE Q-TABLE TILL CONVERGENCE

Q-Learning

Updated Q Value | Current Q Value | Target Q Value | Current Q Value

$$Q(s,a) = Q(s,a) + \alpha \left[r + \max_{a'} \gamma Q(s',a') - Q(s,a) \right]$$

α = Learning Rate

Target policy is always Greedy Policy

What is SARSA-Learning

- MODEL-FREE REINFORCEMENT LEARNING
- ON POLICY
- BEHAVIOUR POLICY IS THE SAME AS THE TARGET POLICY
- MAKE A Q-TABLE
- UPDATE Q-TABLE TILL CONVERGENCE

SARSA-Learning

Updated Q Value | Current Q Value | Target Q Value | Current Q Value

$$Q(s,a) = Q(s,a) + \alpha\,[\, r + \gamma Q(s',a') - Q(s,a)]$$

α = Learning Rate

Target Policy is always same as Behaviour Policy

03

Methodology & Experimental Design

OPENAI GYM [TOY TEXT]: Taxi-V3

TRAINING:

- Training the Q-Table in every episode
- Fixed max steps = 99 in each episode
- 1st approach:
  updating the value till convergence in Q-table
- 2nd approach:
  updating till large number of episodes.

OPENAI GYM [TOY TEXT]: Taxi-V3

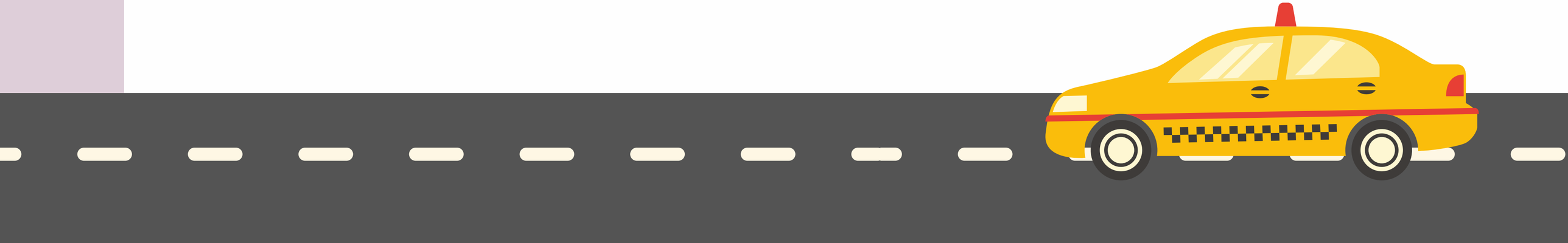
TESTING:

- Test the Q-function on environment
- 2nd approach was more successful as it was giving the optimal policy

Q-learning

BEHAVIOUR POLICY --> EPSILON GREEDY

TARGET POLICY --> GREEDY POLICY



SARSA-learning

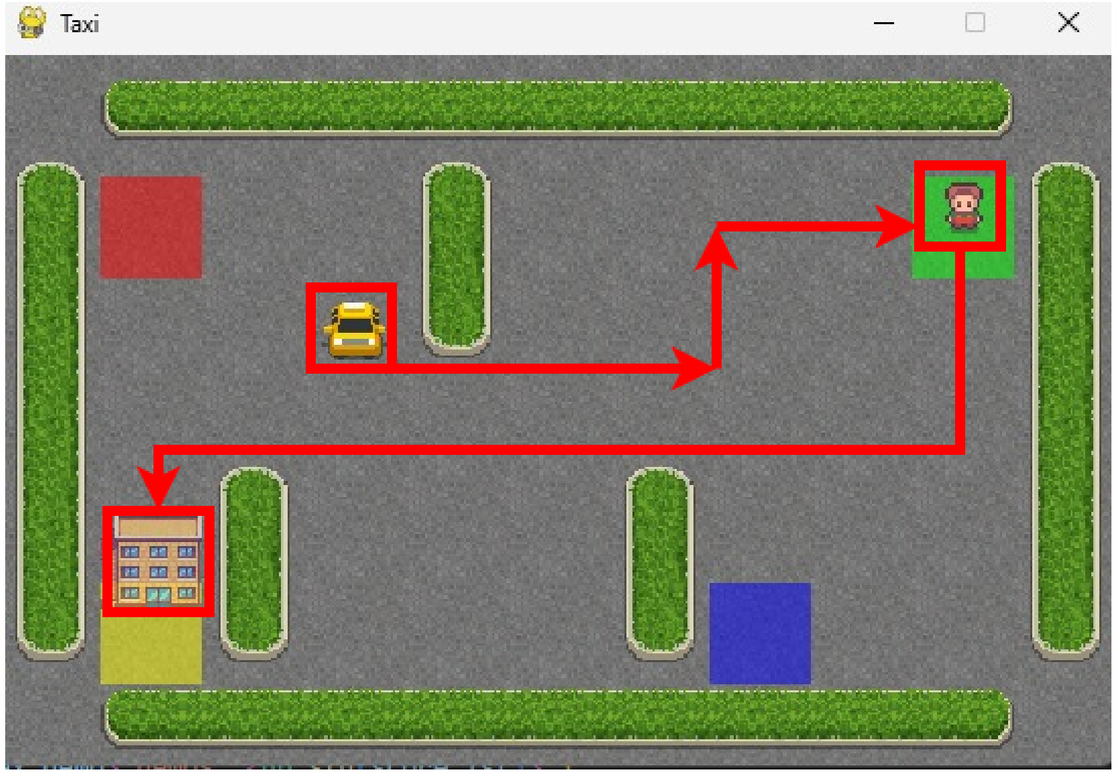
BEHAVIOUR POLICY --> GREEDY POLICY

TARGET POLICY --> GREEDY POLICY

04

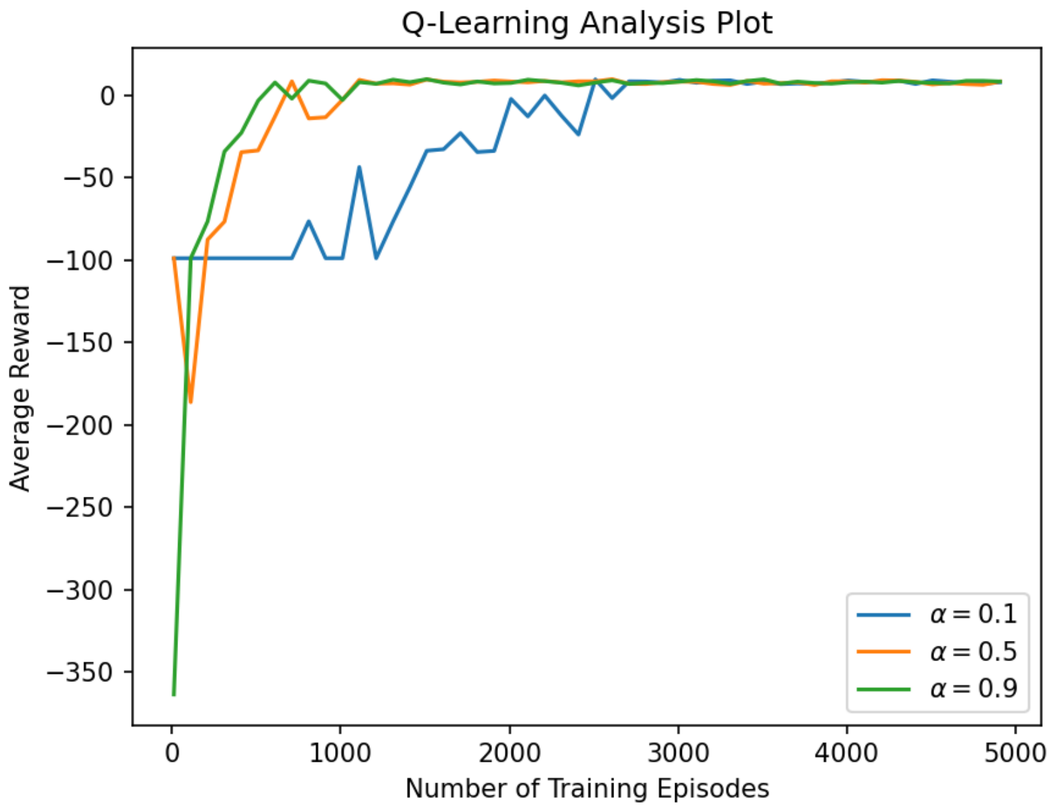
Demo

Demo

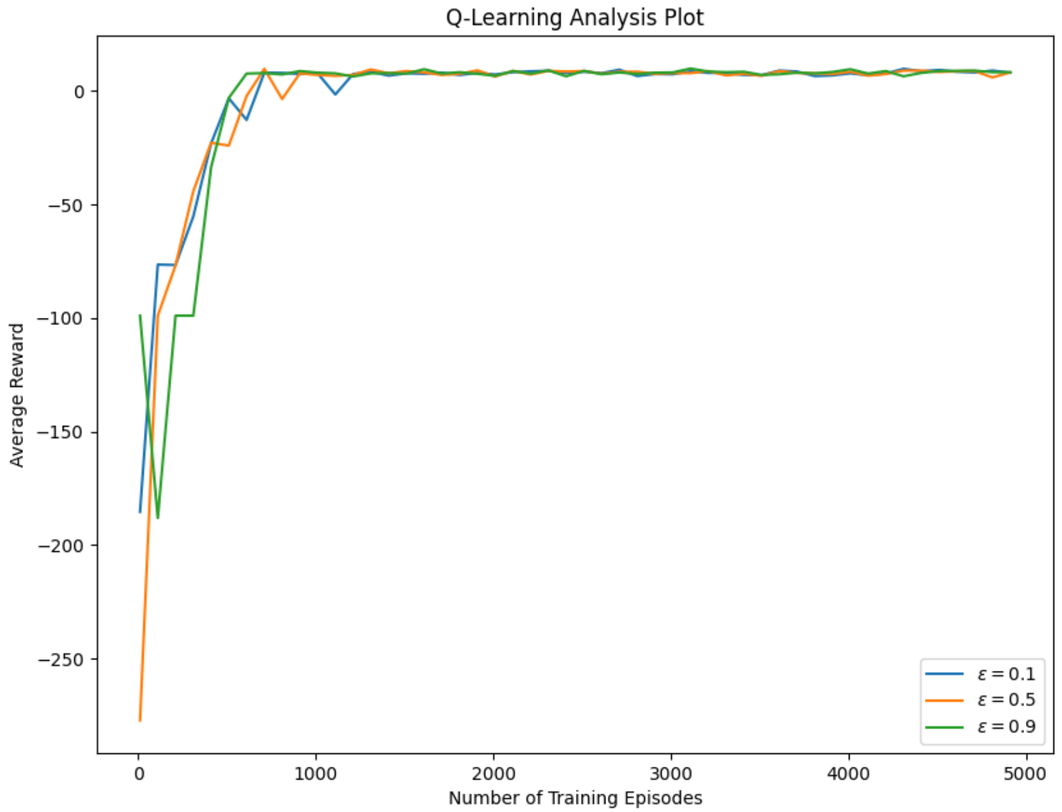


05
Analysis

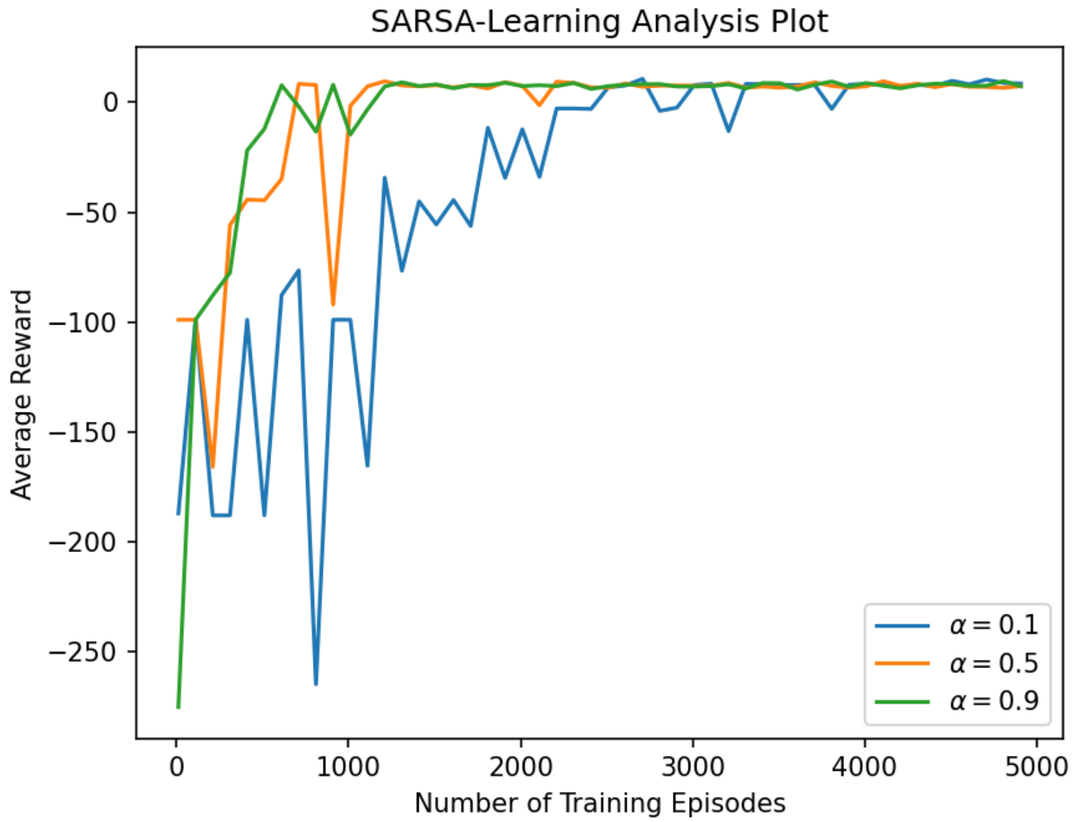
Q-Learning


Q-Learning Analysis Plot
Average Reward
0
−50
−100
−150
−200
−250
−300
−350
0
1000
2000
3000
4000
5000
Number of Training Episodes
α = 0.1
α = 0.5
α = 0.9

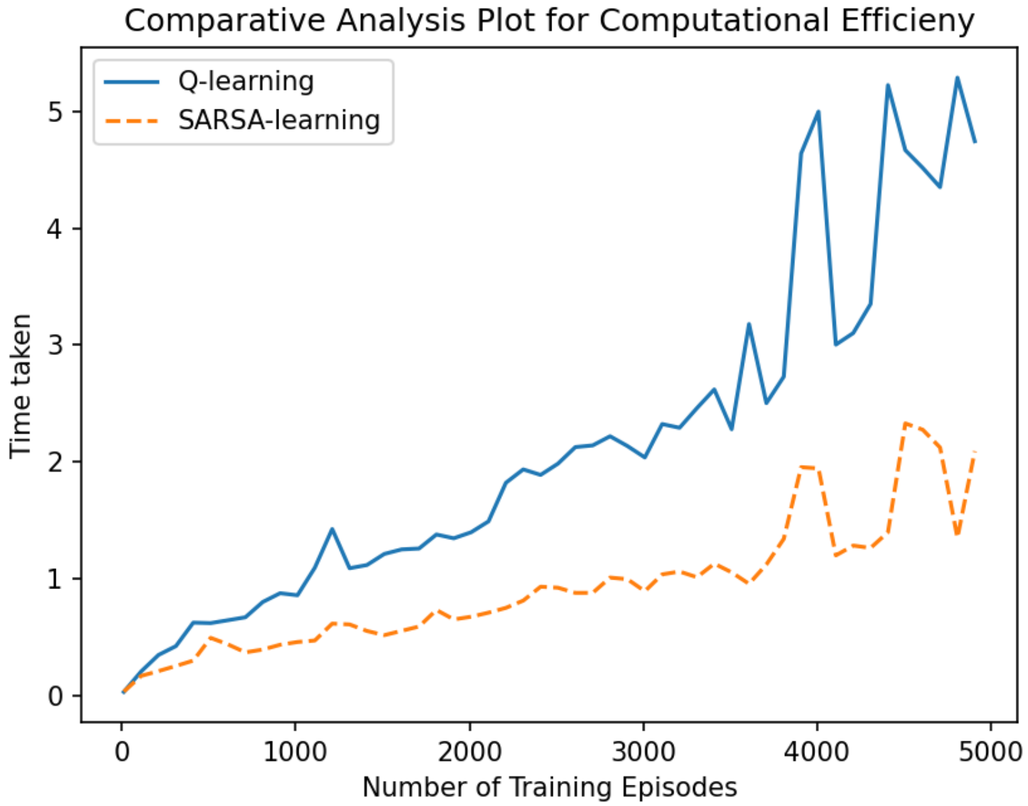
Q-Learning


Q-Learning Analysis Plot
Average Reward
0
−50
−100
−150
−200
−250
0
1000
2000
3000
4000
5000
Number of Training Episodes
ε = 0.1
ε = 0.5
ε = 0.9

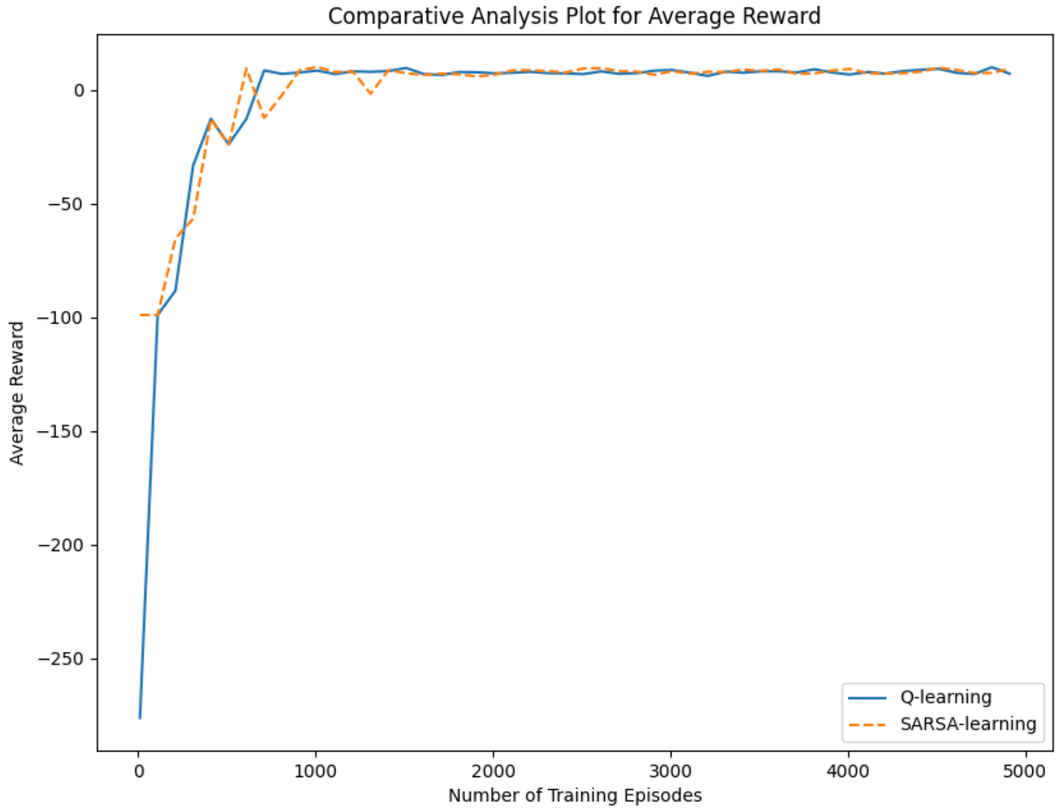
SARSA-Learning



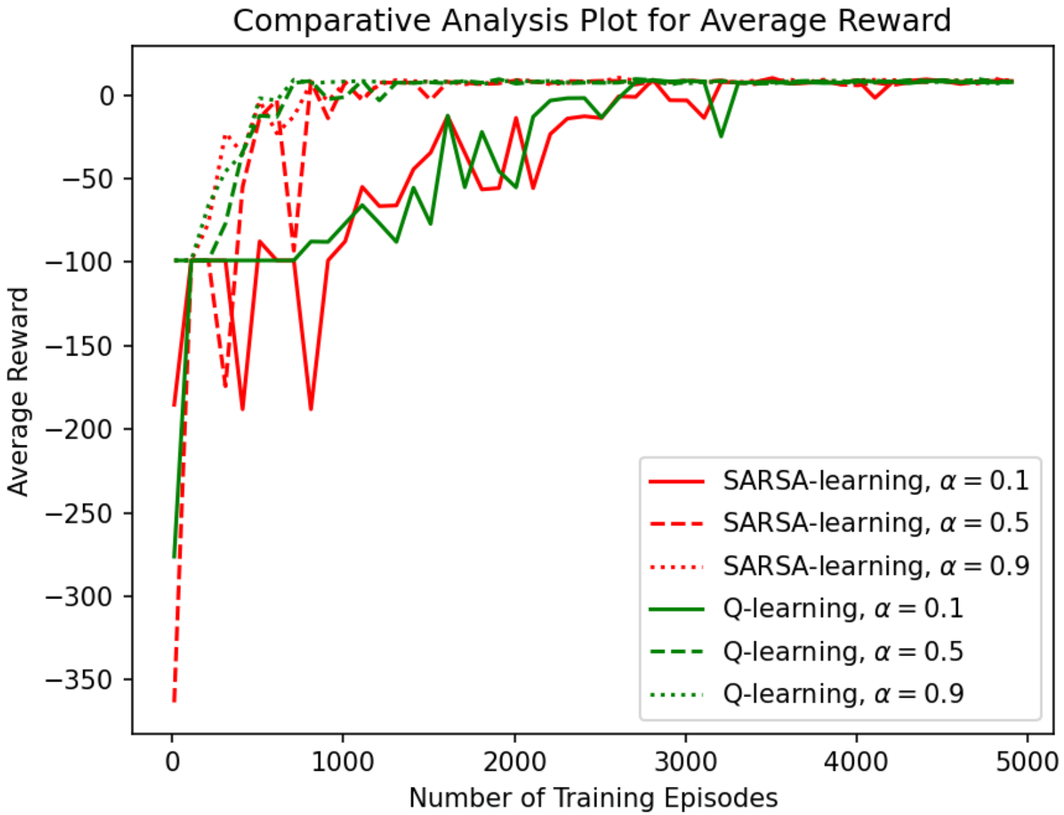
Q-Learning vs SARSA-Learning



Q-Learning vs SARSA-Learning



Q-Learning vs SARSA-Learning



06

Conclusion

CONCLUSION

- SARSA is computationally faster than Q-Learning
- Both Converge faster when the learning rate is high
- The exploration-Exploitation Rate has minimal effect on Convergence
- The Maximum (Average) Rewards are the same for both SARSA and Q-Learning, in all conditions.

REFRENCES

- Sutton, R.S. and Barto, A.G. (1998) Reinforcement Learning: An Introduction. Vol. 1, MIT press, Cambridge
- https://www.gymlibrary.dev/environments/toy_text/taxi/
- https://youtu.be/FhSaHuC0u2M

That's all Folks!

THANK YOU FOR WATCHING