**A STUDY OF THE PERFORMANCE OF DIFFERENT ALGORITHMS IN GYMNASIUM ENVIRONMENTS**

Raunak Nandkumar More

*Department of Computer Science*

*University of South Dakota*

**ABSTRACT:**

Reinforcement learning (RL) algorithms enable agents to learn how to act in different environments. The SARSA and Q-learning algorithms are two popular RL algorithms, each with unique features. In this paper, we evaluate and compare the performance of SARSA and Q-learning algorithms on two gymnasium environments, namely CliffWalker-v0 and Blackjack-v1, based on convergence rate and reward received. We also investigate the impact of different Epsilon values on the performance of the two algorithms. Our results show that SARSA and Q-learning algorithms both learn the optimal policy for the CliffWalker-v0 environment. However, SARSA outperforms Q-learning in terms of efficiency, requiring fewer steps to reach convergence and less time to learn the optimal policy.

**INTRODUCTION:**

Reinforcement learning (RL) techniques enable agents to learn how to act in different environments by trial and error. The agent is rewarded or penalized based on its interaction with the environment, and the objective is to learn a policy that maximizes the overall reward. SARSA and Q-learning are two well-known RL algorithms, each with different characteristics. SARSA modifies its policy after every action, whereas Q-learning modifies its policy based on the predicted rewards of all possible actions, not just the chosen action.

In this study, we compare the performance of SARSA and Q-learning algorithms on two gymnasium environments. The CliffWalker-v0 environment requires the agent to move from the starting point to the goal in a 12x4 grid world without falling off the cliff, earning a reward of -1 for each step taken and a penalty of -100 for falling off the cliff. The Blackjack-v1 environment requires the agent to obtain cards that sum closer to 21 than the dealer’s cards without going over 21.

**MOTIVATION:**

The objective of this study is to evaluate the performance of the SARSA and Q-learning algorithms in a specific environment and determine which algorithm is more suitable for the task. Initially, we used the Blackjack environment, but inconsistent results prompted us to switch to the CliffWalker environment.

**SETUP:**

Formula used for Epsilon Decay is as follows:

**epsilon\_decay = start\_epsilon / (n\_episodes / 2)**

We conducted our experiments using the following parameters:

\* Learning rate: 0.01

\* Discount factor: 0.95

\* Number of episodes: 2000(Cliffwalker) and 50000(Blackjack)

\*Epsilon: 0.2,0.1,0.05,0.01

We evaluated the performance of SARSA and Q learning in terms of the following metrics:

\* Total reward received

\* Path taken

\* Convergence Rate

**METHODOLOGY:**

We will compare the performance of SARSA and Q learning on the CliffWalker-v0 Gym environment. CliffWalker-v0 is a simple environment that consists of a grid of squares, with a cliff in the center. The agent's goal is to reach the goal square at the bottom right corner of the grid without falling off the cliff.

Blackjack v1 is a card game where the goal is to beat the dealer by obtaining cards that sum to closer to 21 (without going over 21) than the dealer’s cards.

We change epsilon with 4 values 0.2,0.1,0.05,0.01 and run each algorithm on Cliff walker 20 times for Epsilon Decay set to False and for 5 times when its set to True and plot a graph to see how it converges and which one converges faster.

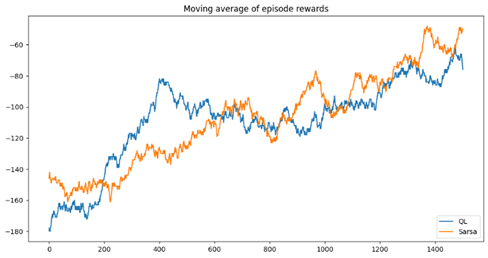
**EXPERIMENT & RESULTS:**

**BlackJack-v1(Epsilon Decay set to True)**

A picture containing screenshot, text, line, plot

Description automatically generated

* On running both algorithms for 2000 episodes and plotting reward vs episodes the plots overlapped; as a result, it was difficult to gain info on the convergence.
* We used moving average to gain insights.
* The moving average was taken for 500 previous data points.
* For this environment, we saw Q Learning converged faster than SARSA.
* SARSA had higher rewards compared to Q learning when considering 2000 episodic run.

Graphical user interface, chart, line chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

* However, when considering 100000 episodes with 5000 episodes of moving average we didn’t see any consistent results. This could be attributed to the fact that Blackjack environment is very random.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

* The rewards remained in negative, meaning the agent lost more games than it won.

**BlackJack-v1(Epsilon Decay set to False)**

Chart, line chart

Description automatically generated

* For the Blackjack-v1 environment, both algorithms were run for 3000 episodes for each Epsilon being run 5 times.
* The reward vs. episode plots showed overlapping results, making it difficult to identify the convergence rate.

Chart, line chart

Description automatically generated

**Cliffwalker-v0(Epsilon Decay set to True)**

Chart

Description automatically generated with medium confidence

Chart

Description automatically generated with medium confidence

Chart

Description automatically generated

* 2000 episodes were run for this environment
* For this environment SARSA converged faster
* Both algorithms ended with having same rewards
* It was also observed that Path taken by Q learning was more aggressive/Unsafe than that taken by SARSA
* Optimal/Shortest Path was taken by Q learning which was why it had lower rewards as sometimes the agent would fall down the cliff incurring a steep -150 reward.
* Middle Path was taken by SARSA

**Cliffwalker-v0(Epsilon Decay set to False)**

* We used 4 differnet Epsilon values 0.2,0.1,0.05,0.01
* For this environment, we ran both algorithms for 500 episodes and plotted graph with 20 sample runs for each Epsilon and Algorithm
* We observed that there was higher variance with Q learning than that of SARSA
* Epsilon being 0.2 led to more random fluctuation which is intuitive as higher Epsilon means the algorithm would try to explore more than exploit.

Chart, bar chart

Description automatically generated

* As seen below in Q Learning keeping the moving average to 100 as we decrease e from 0.2 to 0.01, the convergence rate increases which makes sense as the algorithm prefers exploiting over exploration

Chart, line chart

Description automatically generated

Chart, bar chart

Description automatically generated

* In SARSA however we see that e=0.2 does generally well with getting faster convergence which is stark opposite of what was observed in Q learning.

**Chart, line chart

Description automatically generated**

**CONCLUSION:**

Our results show that SARSA outperforms Q learning on CliffWalker-v0. In BlackJack-v1, we see that This suggests that SARSA is a more effective RL algorithm than Q learning for some environments.

**REFERENCES:**

\* Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

\* OpenAI Gym. (n.d.). Retrieved February 25, 2023, from https://gym.openai.com/