Reinforcement Learning

Assignment-IV - Monte Carlo - Blackjack Policy Evaluation and Temporal Difference - Cliff Walking with SARSA Report

Student: Adisu Bezu

Submitted to: Natnael Argaw(PhD)

ID: GSR/5572/15

Addis Ababa University

Addis Ababa Institute of Technology

Department of Artificial Intelligence

February 09, 2025

Implementation Report

This report presents the implementation and analysis of two reinforcement learning exercises: Monte Carlo Blackjack Policy Evaluation and Temporal Difference Learning using SARSA in Cliff Walking.

Monte Carlo Blackjack Policy Evaluation

In the first exercise, Monte Carlo methods are applied to estimate the state-value function in the Blackjack environment. A simple policy is used where the agent sticks if its sum is 18 or greater and hits otherwise. The value function is estimated using first-visit Monte Carlo, where rewards obtained from various states are averaged over multiple episodes.

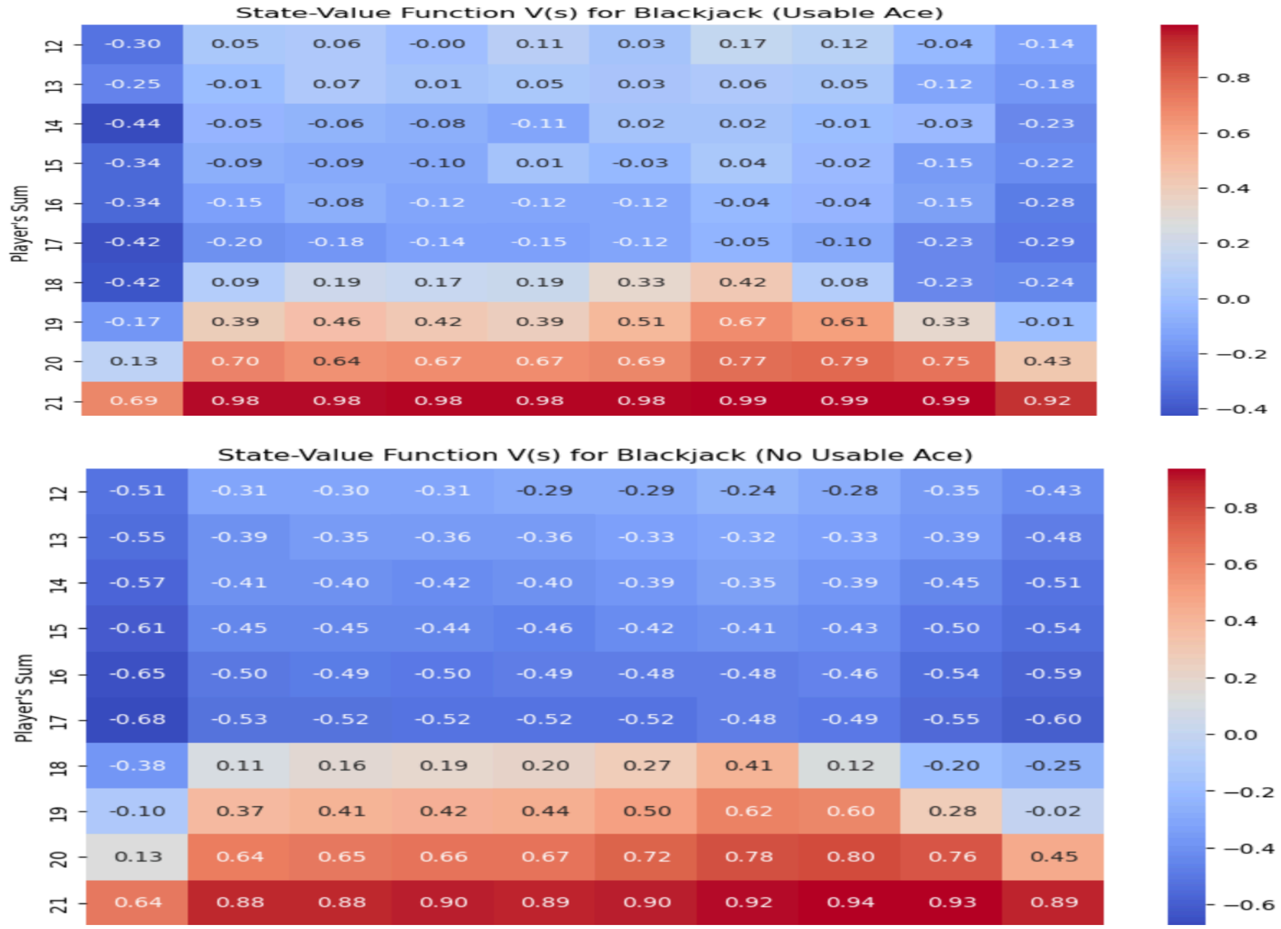
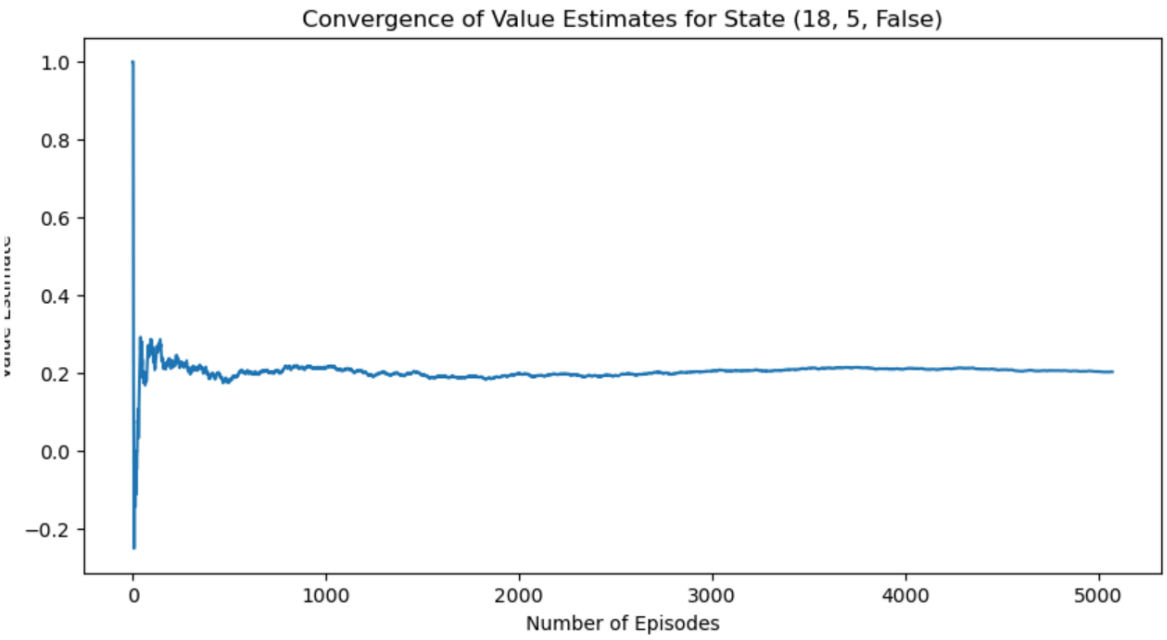


Fig: State Value Function for Monte Carlo Blackjack Policy Evaluation



Convergence of value estimates Monte Carlo Blackjack Policy Evaluation

The provided plot represents the results of Monte Carlo value estimation for the Blackjack game under different conditions. The first two heatmaps display the estimated state-value function for Blackjack with and without a usable ace. The colors indicate the expected returns from each state, with redder values representing higher expected rewards and bluer values denoting negative or less favorable outcomes. The presence of a usable ace generally leads to higher value estimates due to the increased flexibility in decision-making.

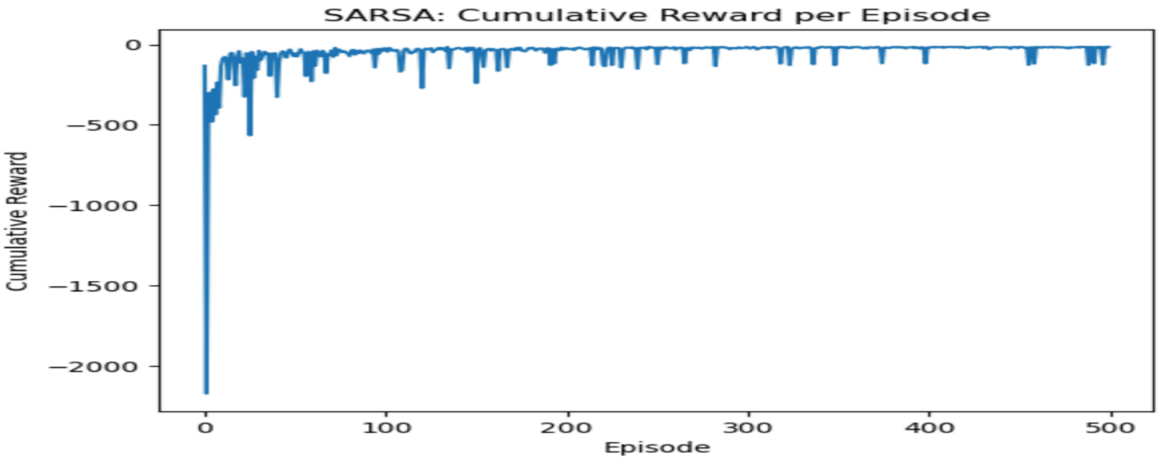
The third plot illustrates the convergence of value estimates for a specific state, (18, 5, False). Initially, the estimate fluctuates significantly, which is expected as the agent explores different game trajectories. Over time, the estimates stabilize, suggesting that the Monte Carlo method effectively learns the long-term expected reward for this state. The overall convergence pattern indicates that the learning process is functioning correctly, though some variance remains due to the stochastic nature of the game.

The observed differences between the two heatmaps highlight the impact of having a usable ace. States with a usable ace tend to have higher values, particularly in scenarios where the player's total is close to 21. This aligns with Blackjack strategy principles, where a usable ace provides more favorable outcomes. The findings suggest that reinforcement learning techniques like

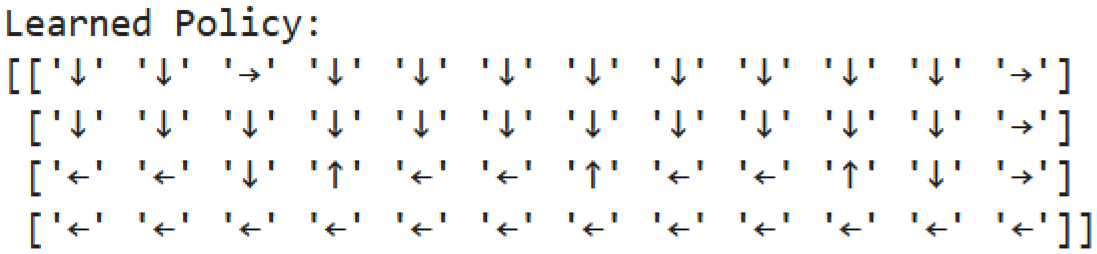
Monte Carlo estimation can effectively approximate optimal strategies for Blackjack, demonstrating their applicability in solving similar decision-making problems.

Temporal Difference Learning using SARSA in Cliff Walking

The second exercise involves implementing the SARSA algorithm to solve the Cliff Walking problem. This problem consists of a 4x12 grid where stepping off a cliff incurs a heavy penalty. SARSA, an on-policy learning method, updates the Q-table based on the agent's interactions with the environment using an epsilon-greedy policy. The training process tracks total reward per episode and refines the learned policy over multiple iterations. Initially, the agent explores randomly, accumulating high penalties. Over time, it learns to navigate towards the goal while avoiding the cliff. Unlike Q-learning, SARSA incorporates exploration into its updates, leading to a more cautious policy.



The SARSA algorithm in the Cliff Walking environment demonstrates a gradual improvement in cumulative reward over episodes, as seen in the plot. Initially, the cumulative reward is low, indicating that the agent is exploring and occasionally falling off the cliff, resulting in penalties. As training progresses, the agent learns to navigate the environment more effectively, avoiding the cliff and reaching the goal with fewer penalties. By around 500 episodes, the cumulative reward stabilizes, showing that the agent has converged to a near-optimal policy. This reflects SARSA's ability to balance exploration and exploitation, learning a safe path while considering the stochastic nature of the environment.



```
Learned Policy:
[['↓' '↓' '→' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '→']
 ['↓' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '↓' '→']
 ['←' '←' '↓' '↑' '←' '←' '↑' '←' '←' '↑' '↓' '→']
 ['←' '←' '←' '←' '←' '←' '←' '←' '←' '←' '←' '←']]
```

Learned policy for SARSA

The learned policy in the given image represents a navigation strategy for Cliff Walking. The repeated patterns suggest consistent actions in certain states, indicating that the agent has identified a stable path or policy. However, the presence of varied sequences implies some exploration or adaptation to specific state transitions, reflecting the agent's attempt to optimize its actions while avoiding penalties or hazards.

The analysis highlights the strengths of both approaches in different reinforcement learning scenarios. Monte Carlo methods are well-suited for episodic tasks like Blackjack, where rewards are obtained after complete episodes. In contrast, SARSA excels in environments requiring sequential decision-making, where learning occurs incrementally based on immediate feedback. The convergence trends observed in both exercises validate their effectiveness, with Monte Carlo stabilizing long-term value estimates and SARSA balancing exploration and exploitation to improve short-term decision-making.

Overall, the findings reinforce fundamental reinforcement learning principles such as value estimation, policy evaluation, and convergence behavior. The experiments demonstrate how different algorithms can be tailored to various problem settings, with Monte Carlo offering robust policy evaluation for stochastic environments and SARSA providing practical decision-making strategies in dynamic settings. Both methods showcase the potential of reinforcement learning in optimizing sequential decision-making processes.