COMP 579 A2

Group 7 - Mahmood Hegazy, Sebastian Sabry

February 2025

1 Report: SARSA and Expected SARSA on FrozenLake Environment

1.1 Experimental Setup and Parameter Choices

Our study investigates the performance of SARSA and Expected SARSA algorithms on the FrozenLake environment, focusing on how different combinations of learning rates and temperature parameters affect learning dynamics. The FrozenLake environment presents unique challenges due to its slippery surface mechanics, which introduce stochasticity into state transitions, and its sparse reward structure, where rewards are only received upon reaching the goal.

1.2 Parameter Selection Rationale

We carefully selected our parameters to explore a specific region of the hyperparameter space:

Learning Rates (α): $[0.1, 0.2, 0.5]$

- The low value ($\alpha = 0.1$) enables stable but slow learning
- The medium value ($\alpha = 0.2$) offers a balance between stability and learning speed
- The high value ($\alpha = 0.5$) allows for more aggressive value updates

Temperature Values (T): $[0.005, 0.01, 0.02]$

- These unusually low temperature values were chosen because the FrozenLake environment has a small action space (4 actions) and requires precise navigation
- Even small temperature values create sufficient exploration due to the environment's inherent stochasticity
- The chosen range allows for fine-grained control over exploration behavior

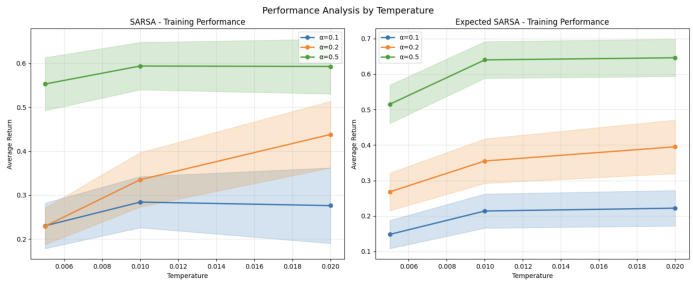

Figure 1: SARSA and Expected Training Performance by Learning Rate

1.3 Analysis of Results

We created four distinct visualizations to understand different aspects of the learning process:

1.3.1 Training Performance by Learning Rate

The first visualization shows how different learning rates affect the training process while varying temperatures. The results for SARSA and Expected SARSA can be seen in Figure 1. We observed that:

- Lower learning rates ($\alpha = 0.1$) showed slight improvement with incresing temperate initially but regressed when temperature was increased further
- The medium learning rate ($\alpha = 0.2$) achieved the best improvement curve with a clear increase in performance as temperature increases.
- The highest learning rate ($\alpha = 0.5$) showed consistent performance throughout with just a slight increase in the beginning.

1.3.2 Training Performance by Temperature

The second visualization, organizing the results by temperature, as we see in Figure 2 revealed that:

- Even our low temperature values (0.005-0.02) provided sufficient exploration due to the environment's stochastic nature but performance was still always less than the higher temperatures throughout as expected.
- The middle temperature ($T = 0.01$) performed best at the lowest learning rate and join best with highest temperature ($T = 0.02$) and the highers learning rate.
- Highest temperature ($T = 0.02$) consistently performed best across different learning rates except for lowest learning rate.

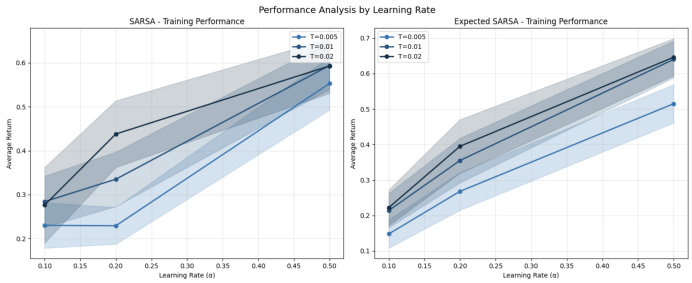

Figure 2: SARSA and Expected Training Performance by Temperature

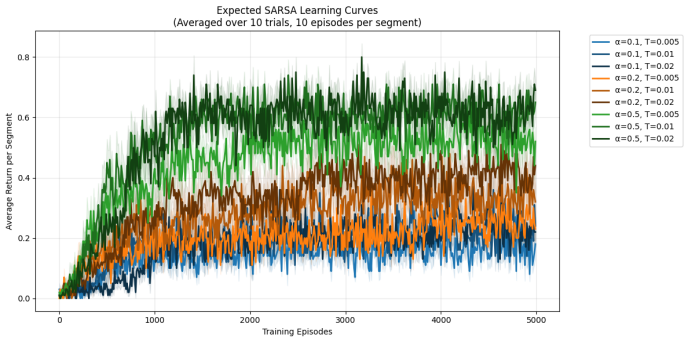

Figure 3: Expected SARSA Training Learning Curve

1.3.3 Learning Curves Over Time

The third visualization, showing the progression of learning over episodes (Figures 4 and 3), demonstrated the following.

- Both algorithms required approximately 2000-3000 episodes to reach peak performance
- The combination of the highest learning rate of $\alpha = 0.5$ and either temperature of $T >= 0.01$ produced the fastest learning with the best performance over time

1.3.4 Final Testing Performance

The fourth visualization, focusing on testing performance, seen in Figures 5 and 6, revealed:

- The best-performing parameter combinations during training also performed well during testing
- Expected SARSA showed slightly better generalization in testing, particularly with higher learning rates

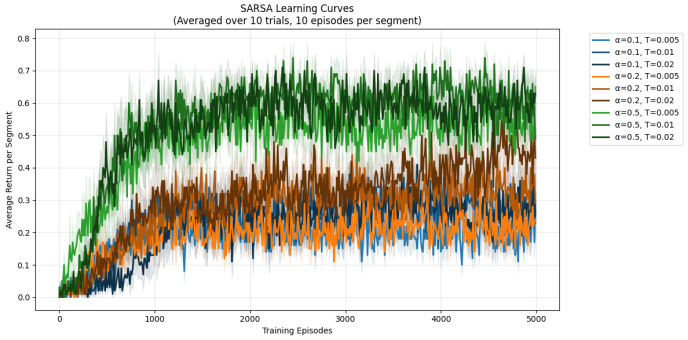

Figure 4: SARSA Training Learning Curve

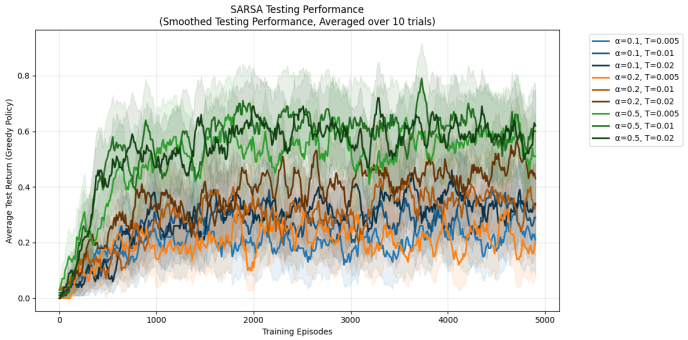

Figure 5: SARSA Testing Learning Curve

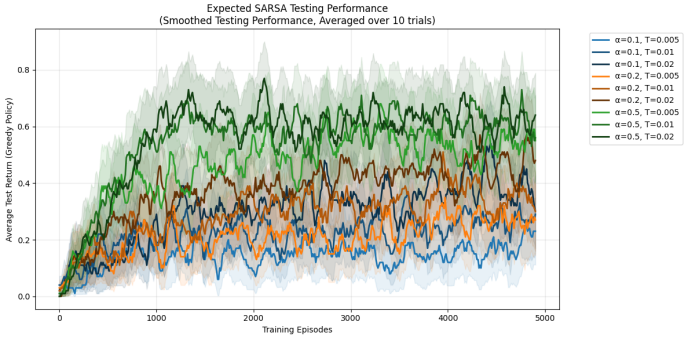

Figure 6: Expected SARSA Testing Learning Curve

1.4 Key Findings and Conclusions

1.4.1 Parameter Sensitivity

The FrozenLake environment showed high sensitivity to both learning rate and temperature parameters. The optimal combination ($\alpha = 0.5$, $T = 0.01$) struck a good balance between learning stability and exploration efficiency.

1.4.2 Algorithm Comparison

Expected SARSA demonstrated slightly better final performance compared to regular SARSA, particularly in testing scenarios. This advantage was most pronounced with higher learning rates, where Expected SARSA's ability to consider all possible next actions helped mitigate the volatility typically associated with higher learning rates.

1.4.3 Exploration Dynamics

Despite using very low temperature values, the environment's inherent stochasticity provided sufficient exploration. This suggests that perhaps in environments with significant built-in randomness, the temperature parameter should be set lower than in deterministic environments.

1.5 Recommendations for Implementation

Based on our findings, we recommend:

- Using Expected SARSA over regular SARSA
- Using a temperature of 0.01 for this environment, as it provides adequate exploration without compromising exploitation
- Implementing a longer training period (at least 3000 episodes) to ensure convergence

.

Q2. Policy Evaluation (Given Policy)

a) Let $v_\pi = \sum_a \pi(a|s) \sum s', r p(s', r|s, a)[r + \gamma v_\pi(s')]$
Since we have a given policy that is deterministic for state s_1 we have:

$$\begin{aligned} v_\pi(s_1) &= p(s_2|s_1, a_1)[r(s_1, a_1) + 0.9 v_\pi(s_2)] \\ &= 2 + 0.9 v_\pi(s_2) \end{aligned}$$

Similarly for state s_2 we have:

$$\begin{aligned} v_\pi(s_2) &= p(s_3|s_2, a_1)[r(s_2, a_1) + 0.9 v_\pi(s_3)] \\ &= 1 \end{aligned}$$

So the solution to the bellman equations for v_π are:

$$\begin{aligned} v_\pi(s_1) &= 2.9 \\ v_\pi(s_2) &= 1 \\ v_\pi(s_2) &= 0 \end{aligned}$$

Q3. Finding the Optimal Policy

a) Let $v_* = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_*(s')]$

For state s_1 we have:

$$\begin{aligned} v_*(s_1) &= \max_a [p(s_3|s_2, a_1)[r(s_2, a_1) + 0.9 v_*(s_3)], p(s_2|s_1, a_1)[r(s_1, a_1) + 0.9 v_*(s_2)]] \\ &= \max_a [2 + 0.9 v_*(s_2), 5] \end{aligned}$$

For state s_2 we have:

$$\begin{aligned} v_*(s_2) &= \max_a [p(s_1|s_2, a_2)[r(s_2, a_2) + 0.9 v_*(s_1)], p(s_3|s_2, a_1)[r(s_2, a_1) + 0.9 v_*(s_3)]] \\ &= \max_a [-1 + 0.9 v_*(s_1), 1] \end{aligned}$$

To solve this analytically we have to check each case. Suppose $v_*(s_1) = 5$

$$\begin{aligned} v_*(s_2) &= \max_a [1, 0.9(0.5) - 1] \\ &= \max_a [1, 3.5] = 3.5 \end{aligned}$$

We then check for contradictions. Since we assumed $v_*(s_1) = 5$ and thus $2 + 0.9v_*(s_2) \leq 5$ we check if this is the case. $2 + 0.9(3.5) \leq 5$ which is not true. Thus our initial assumption was wrong. Thus we have:

$$v_*(s_1) = 2 + 0.9v_*(s_2) \tag{1}$$

We can substitute equation one in the max of state 2.

$$\begin{aligned} v_*(s_2) &= \max_a[-1 + 0.9v_*(s_1), 1] \\ &= \max_a[-1 + 0.9(2 + 0.9v_*(s_2)] \\ &= \max_a(1, 0.8 + 0.81v_*(s_2)] \end{aligned}$$

Suppose that $v_*(s_2) = 1$ we have an obvious contradiction. $1 > 0.8 + 0.81(1)$ which is not true. Thus, we have the following optimal values for each state:

$$\begin{aligned} v_*(s_1) &= 5.789 \\ v_*(s_2) &= 4.21 \\ v_*(s_3) &= 0 \end{aligned}$$