COMP579 Assignment 2 Report: Evaluation of Learning Parameters in Softmax Exploration for MDPs

Antoine Voyer
260989010
antoine.voyer@mail.mcgill.ca
Dept. of Mechanical Engineering
McGill University

Simon Li
260984998
xi.yang.li@mcgill.ca
Dept. of Electrical Engineering
McGill University

***Abstract*—This paper explores the effect of learning rate and temperature in a Softmax-based exploration scheme for a reinforcement learning agent in an MDP environment. We evaluate different hyperparameter combinations and analyze their impact on final training and test performance. Additionally, we solve a small three-state MDP both analytically and numerically to determine the optimal policy. Our findings suggest that temperature plays a significant role in balancing exploration and exploitation, affecting convergence speed and final performance.**

I. Introduction

Reinforcement learning agents must balance exploration and exploitation to maximize cumulative rewards. In this experiment, we evaluate the performance of an agent using Softmax (Boltzmann) exploration across different learning rates (α) and temperatures (T). Additionally, we solve a simplified three-state MDP using policy evaluation and value iteration to find the optimal policy.

II. Experimental Setup

We consider the FrozenLake environment, where the agent uses Softmax exploration to balance action selection. The experiment follows these steps:

- Evaluate final training performance as a function of α, with different values of T.
- Evaluate final training performance as a function of T, with different values of α.
- Plot training return over episodes for different parameter combinations.
- Plot test return over episodes for different parameter combinations.

Each trial consists of 500 segments, with 10 training episodes per segment followed by 1 testing episode using the greedy policy. The tested hyperparameters include:

- Learning rates: $\alpha \in \{0.1, 0.5, 0.9\}$
- Temperatures: $T \in \{0.1, 1.0, 5.0\}$

III. Results and Discussion

A. Effect of Learning Rate (α) on Final Training Performance

Fig. 1 shows how the learning rate affects final training performance across different temperature settings. The trend suggests that lower temperatures ($T = 0.1$) lead to more stable learning but decreased final rewards as α increases. Moderate temperatures ($T = 1.0$) exhibit an optimal learning rate near $\alpha = 0.5$, while higher temperatures ($T = 5.0$) result in lower final training performance.

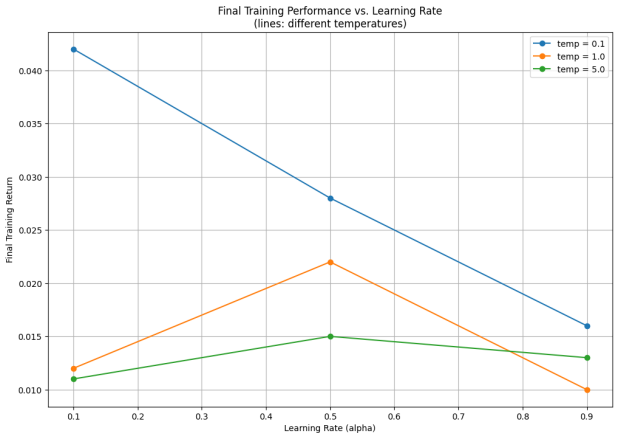

Fig. 1. Final Training Performance vs. Learning Rate.

B. Effect of Temperature (T) on Final Training Performance

Fig. 2 displays the effect of temperature on final training performance. Lower temperatures generally result in higher returns at lower learning rates. However, at higher α, performance deteriorates for all temperatures. This indicates that an optimal balance between T and α is necessary for effective learning.

C. Training and Test Performance over Episodes

Fig. 3 shows how training return evolves over episodes for different parameter combinations. We observe that higher

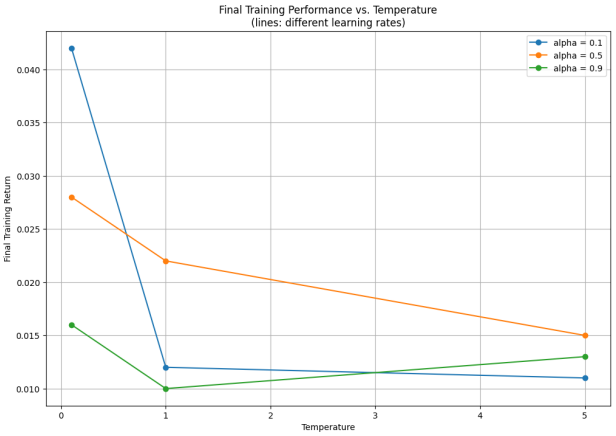

Fig. 2. Final Training Performance vs. Temperature.

temperatures lead to more erratic behavior, as they encourage exploration at the cost of immediate reward. Similarly, Fig. 4 illustrates the test performance, where the best-performing combinations stabilize at higher rewards over time.

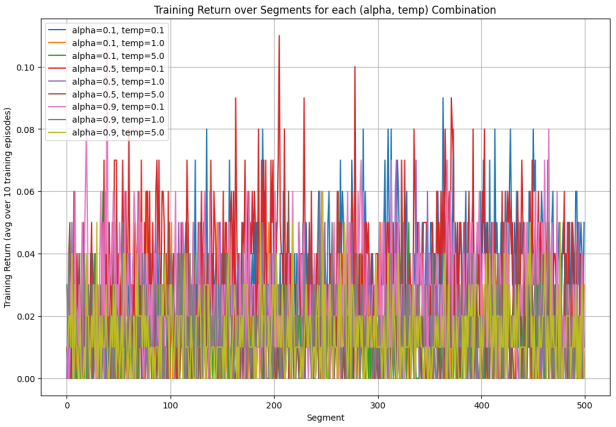

Fig. 3. Training Return Over Episodes for Different Hyperparameter Combinations.

IV. MDP Policy Evaluation and Optimization

A. Policy Evaluation

We solve the Bellman equations for the given policy $\pi(s_1) = a_1, \pi(s_2) = a_1$ both analytically and iteratively. The results confirm convergence to:

$$v(s_1) = 2.9000, \quad v(s_2) = 1.0000, \quad v(s_3) = 0. \tag{1}$$

B. Finding the Optimal Policy

Using value iteration, we derive the optimal value function and policy:

$$v^*(s_1) = 5.7895, \quad v^*(s_2) = 4.2105, \quad v^*(s_3) = 0. \tag{2}$$

The optimal policy is determined as:

- $\pi^*(s_1) = a_1$
- $\pi^*(s_2) = a_2$

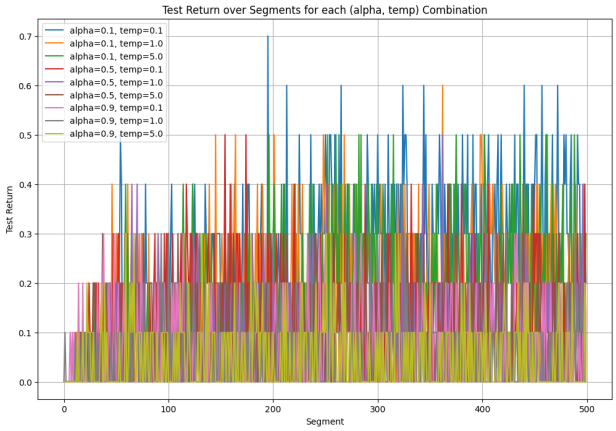

Fig. 4. Test Return Over Episodes for Different Hyperparameter Combinations.

These results were obtained using both analytical methods and value iteration, confirming their correctness.

V. Conclusion

This experiment analyzed the impact of learning rate and temperature in Softmax exploration for reinforcement learning in an MDP. We found that a balance between T and α is crucial for achieving stable and high-performance learning. Additionally, solving the three-state MDP validated our numerical methods against analytical solutions. These insights are useful for designing reinforcement learning agents with optimal exploration strategies.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

[2] R. Bellman, "Dynamic Programming," *Science*, vol. 153, no. 3731, pp. 34-37, 1966.