CS 5033 - RL Project Report

Airi Shimamura Khoi Trinh

1. Introduction

For our RL project this semester, we want to build an RL agent that can easily beat the CartPole game.

In this checkpoint document, we will provide details and updates on our current experiments, as well as mention any difficulties we encountered, and list any future work to be done. A slight change from the proposal, Airi is implementing Q-Learning, while Khoi will be implementing SARSA learning instead of TD-Learning.

2. Literature Review

2.1. Reinforcement Learning

The goal of reinforcement learning (RL) is not simply to play a game. Rather, the driving force is to create systems that can be adaptive in the real world (Arulkumaran et al., 2017).

The essence of RL is learning through interaction and reward-driven behavior. The RL agent interacts with the environment and the results of its actions; it can then adjust its behavior in response to the reward(s) received (Arulkumaran et al., 2017)

This trial-and-error behavior can be considered the root of RL.

2.2. SARSA

2.3. Q-Learning

3. Hypothesis

Hypothesis 1: We epxect the Q Learning algorithm to perform better overall than the SARSA algorithm over 10000 episodes.

Hypothesis 2: For both algorithms, we expect exponential decay of ϵ to perform the best, as in, have the highest average reward over 10000 episodes.

Both Airi and Khoi's experiments will be performed in accordance to this hypothesis.

4. Experiments Done

4.1. Airi's Experiments - Q-Learning

4.2. Khoi's Experiments - SARSA Learning

The SARSA algorithm is a popular reinforcement learning algorithm. At each time step, the agent selects an action according to its policy, observes the next state and reward, and updates the estimated value of the current state-action pair using a learning rate and discount factor. The algorithm maintains a Q-value function that estimates the expected future rewards for each state-action pair. The agent selects an action using an exploration-exploitation strategy and takes the selected action in the environment.

To this end, Khoi is trying to implement an ϵ greedy method for this SARSA algorithm. He also assumed that the agent will successfully be trained in 1000 episodes. For his setup, the hyperparameters are: $\epsilon = 0.5$; $\gamma = 0.9$; and $\alpha = 0.5$

For the criteria related to the CartPole environment, the agent will play the game for a set number of episodes, and in each episode, the last 100 steps will have their rewards recorded (if the number of steps is less than 100, then the average will be over however many steps was taken for that episode) in order to generate an average. If any of the failing conditions is met, a penalty of -10 is given, otherwise, a reward of +1 is given. In the plot below, notice that a few episodes have taken around 300 steps, and in general, the average reward stays around the 50 mark.

For one run of 10000 episodes, with ϵ kept static at 0.5, here are his current results. Note that due to the random nature of taking actions, each run will produce a different graph.

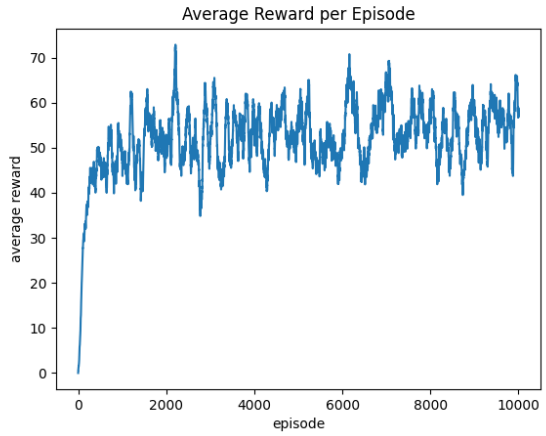

Figure 1. SARSA Results Over 10000 Episodes, Episode-related decay

With ϵ decaying with the function $\epsilon * (\frac{1}{episodes})$, a run will look like so:

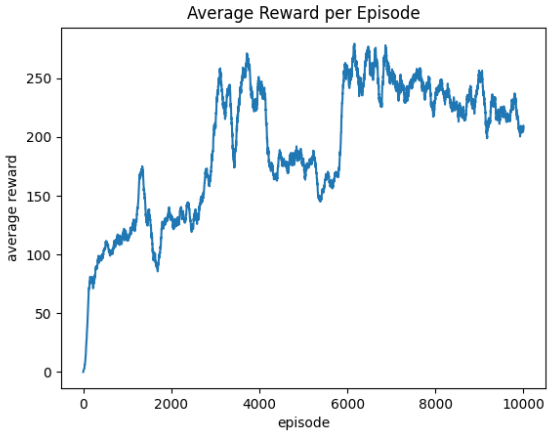

Figure 2. SARSA Results Over 10000 Episodes, 1/Episodes Decay

And finally, with exponential decay of ϵ, a run looks like so:

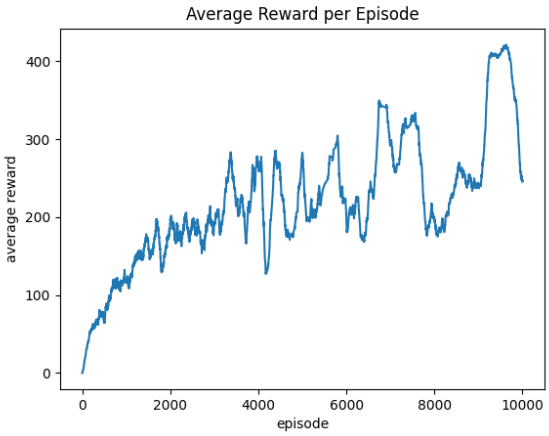

Figure 3. SARSA Results Over 10000 Episodes, Exponential decay

5. General Analysis

5.1. Hypothesis 1

5.2. Hypothesis 2

For SARSA, comparing the three methods of ϵ decay, yields the following results:

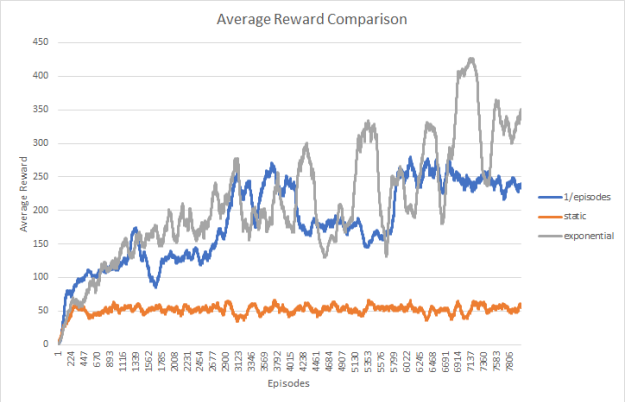

Figure 4. Average Reward of SARSA, Comparing DIfferent ϵ Decay

Additionally, the mean and standard deviation for each method are as follows:

Method	Mean	Standard Deviation
Static	52.07	7.545
1/Episodes	171.9	15.454
Exponential	207.45	80.459

Table 1. Mean of Average Reward and Standard Deviation

As these results show, for the SARSA algorithm, exponential decay of ϵ performed the best.

6. Comparison to Other Methods

7. Future Work

Reference