National Taiwan University
Department of Electrical Engineering
Name: Ching-Hsiang, Wu
ID: R11921080

113-1 (Fall 2024)
Reinforcement Learning

Assignment 1 Report

Q1. What methods have you tried for async DP? Compare their performance.

3 kinds of methods mentioned in the class have been tried for async DP. Overall, the In-place DP has the best performance in these methods. The steps cost in three methods are list below,

- In-place DP: 968 steps
- Prioritized sweeping: 2651 steps
- Real-time DP: 1112 steps

It's worth mentioning that Prioritized sweeping and Real-time DP method even have a worse performance than the synchronous update, which may result from that I couldn't figure out a better criterion to determine whether all states have converged.

Currently, in In-place DP and Prioritized sweeping, I assign my current values as old value every traverse of states, after each traverse, I calculate the new updated value and calculate the norm of the discrepancy of new updated value and old value.

In Real-time DP, the current values will be assigned as old value before each episode starts. Once the episode finished, the new updated value will be reassigned as old value and recalculate the discrepancy norm until it converges.

Q2. What is your final method? How is it better than other methods you've tried?

My final method is In-place DP. Compared to the synchronous updated method - value iteration and policy iteration, the In-place DP outperforms the other 2 methods. Their performances are listed below,

- Policy iteration: 1320 steps
- Value iteration: 1056 steps

Overall, In-place DP still traverses all states and compares the value discrepancy after each traverse to determine whether to stop the algorithm. In contrast, the Prioritized sweep may yield more steps from updating bellman error in traversing and updating all the predecessors' values due to the coding problem.

On the other hand, I noticed that the trap reward has the same value as the step reward. I think it would be better to find the best policy if we can make the trap reward be worse than step reward.