CS5180 Fall (2023)
Pengkai Lyu
Sep 15, 2023

# EX0

For **Q1** & **Q2** Coding part, please check in Jupyter Notebook (ex0.ipynb)

**Q3: How do you think this compares with your manual policy? (You do not have to run your manual policy for $10^4$ steps!) What are some reasons for the difference in performance?**

Plot is shown as Figure 1. Apparently, the agent can reach $10^4$ step easily in a few seconds using random policy. But if we look at the mean reward line, it took the agent 1000 steps, approximately, to make it arrived at goal, rewarding +1. In contrast, if we control the agent, using manual policy, we can reach goal state within 25 or 40 steps because we had specific strategy to direct the agent, not taking the steps randomly. However, it is hard for human to control the agent over 10000 times. The reason is that, while directing the agent, it will sacrificing our time to come up with strategies. The steps we made were very effective but it took much more time to think (compute) which direction is better. Therefore, we need to balance its time and computation, learning various algorithm to train the agent to get the best strategy (policy)

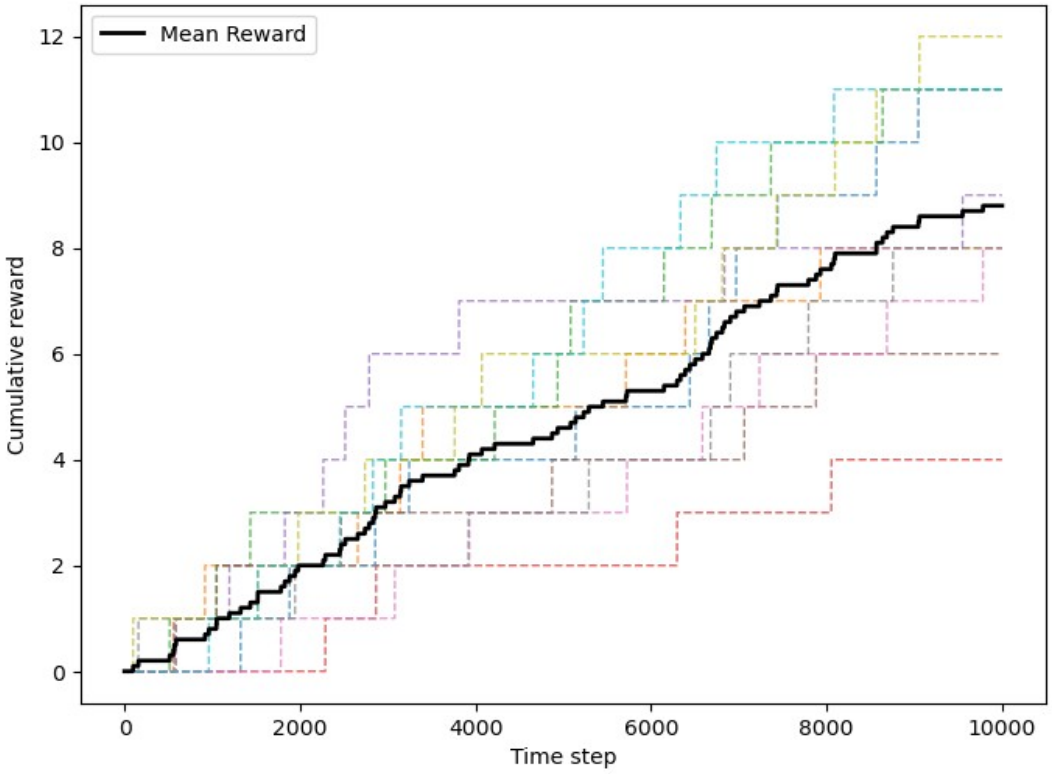


Figure 1

**Q4: Describe the strategy each policy uses, and why that leads to generally worse/better performance.**

A better policy is implemented, shown as Figure 2(a). Another one, Figure 2(b) is for worse policy. Because the goal is at the top-right corner, simply increase the probabilities of moving upwards and to the right from 25% to 40%, which will help us get a better policy. In contrast, we decreased the probabilities of that two actions from 25% to 20% to get the worse policy.

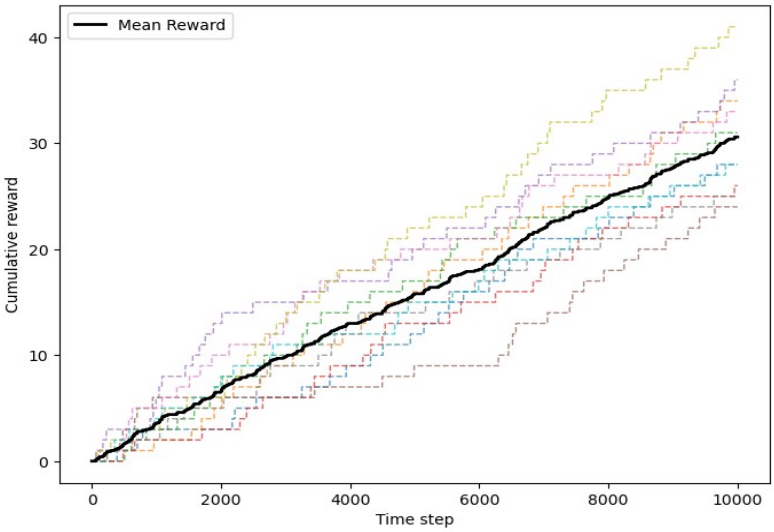


Figure 2 (a)

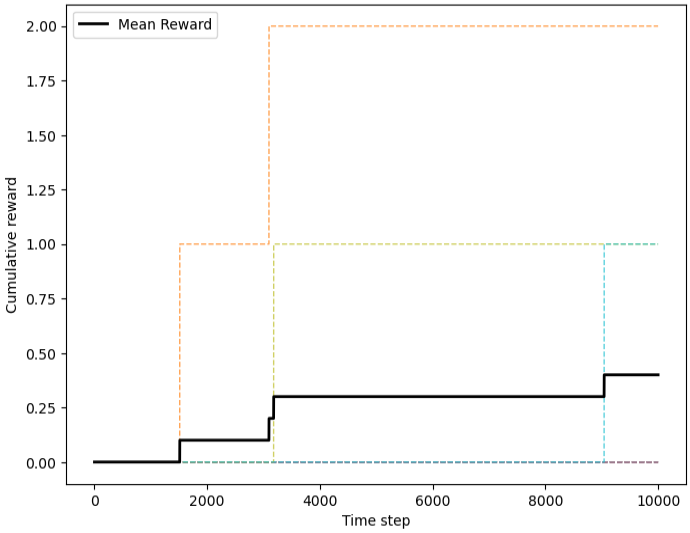


Figure 2 (b)