Rajas Gautam Upalekar CS5180

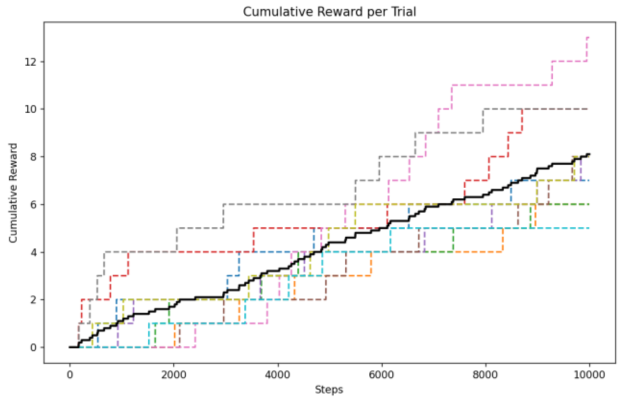
Reinforcement Learning

Ex0

1. The behavior of a goal-state and walled, stochastic grid-world environment is captured by the ‘simulate’ function. It specifies the movement rules of the agent, manages probabilistic slippage, and abides by obstacles and boundaries. After the goal is reached, it resets the agent and distributes rewards based on where the agent is about the objective.

2. By directly interacting with the agent's surroundings through the implementation of a manual policy, the dynamics at work could be grasped practically. Using an interactive tool, the manual policy verified the simulation's functionality and the way the environment responded to user input. For extensive testing, however, this approach proved to be unsustainable and labor-intensive. It brought attention to the need for automated policies in reinforcement learning, which make environment-based exploration and learning more effective. Developing intelligent agents that can learn and adapt on their own through trial and error without human assistance requires a significant step change from manual to automated policies.

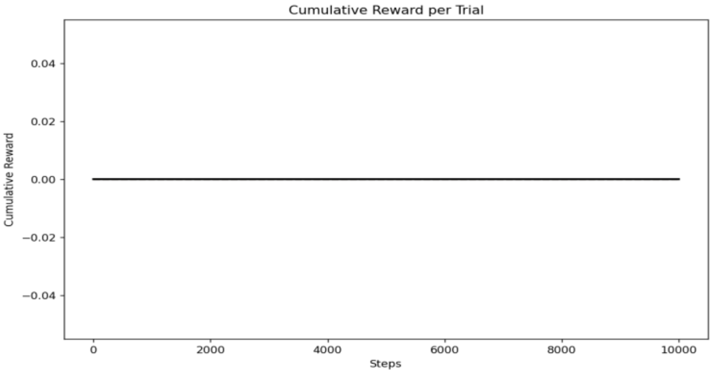
3.



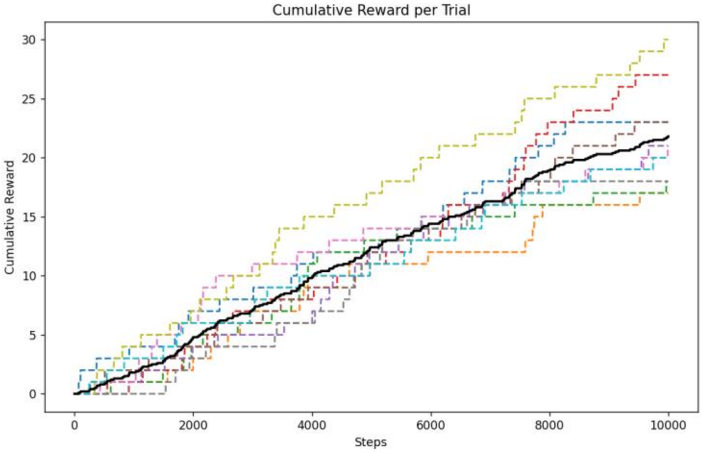
The provided graph illustrates the outcomes of utilizing a random policy over ten trials, with each trial consisting of 10,000 steps. In this strategy, the agent selects actions with equal probability, without considering the state or past results. The dotted lines represent the cumulative rewards for each individual trial, displaying the inherent variability of a random policy. The solid black line indicates the mean cumulative reward across all trials, which progresses more gradually than what might be expected from a policy with a more systematic approach.

When compared to a manual policy, where decisions are likely influenced by knowledge and strategy, the random policy is expected to underperform due to its lack of purposeful direction and learning capability. A manual policy, guided by human understanding, could potentially demonstrate a steeper and more consistent increase in cumulative reward, reflecting a more directed effort towards the objective. In contrast, the random policy's performance fluctuates widely, with some trials possibly achieving success by chance, but on average, advancing towards the goal at a slower pace due to the absence of intentionality in the action selection process.

4. **Worse Policy Result:**



Better Policy Result:



Worse Policy Strategy: It chooses a dependable but inefficient course of action, like traveling in a single direction that doesn't lead to the objective, which may cause it to repeatedly collide with walls or other obstacles.

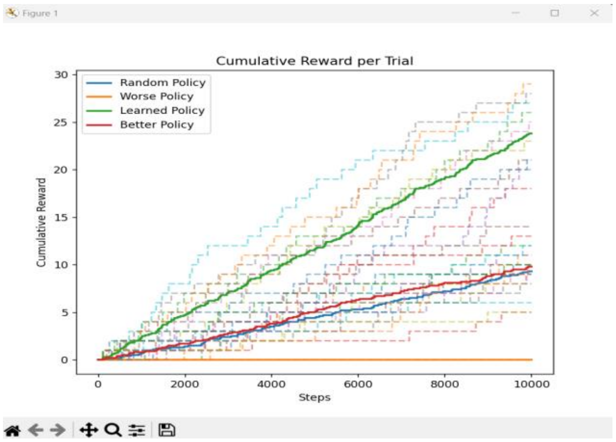
Worse Policy Performance: This policy performs worse than random actions because it frequently stops moving or becomes stuck because it ignores the state of the environment.

Better Policy Strategy: It prioritizes movements that shorten the distance to the goal by taking the agent's position about the goal into account when making decisions.

Better Policy Performance: This policy should perform better than random selections by taking a more methodical approach to reducing the distance to the goal.

Implementation and Comparison: The agent's framework is used to implement both policies, and a plot is used to compare each one's performance against the random policy and show which is more effective. The cumulative reward curve for the worse policy is significantly lower than that for the better policy, indicating less progress toward the goal over the same number of steps.
The better policy's curve shows a consistent upward trend, suggesting systematic and progressive achievement of rewards, whereas the worse policy's curve remains flat or has minimal progress, reflecting a lack of effective strategy.
In contrast, the better policy displays not just higher average cumulative rewards but also higher variability, likely due to more strategic actions that can occasionally result in significantly higher rewards.
The average cumulative reward (thick solid line) for the better policy is much higher and steeper, which indicates that on average, the better policy is more successful at achieving rewards quickly compared to the worse policy.

5.



The graph displays the cumulative rewards achieved per trial by four distinct policies: random, worse, learned, and better. The random policy, as expected, shows variability in its performance, with each trial's cumulative reward progressing in a stochastic manner. The worse policy underperforms significantly when compared to the random policy, as it likely makes systematically poor decisions that do not lead to the reward, such as moving in the opposite direction of the goal or making moves that result in negative outcomes.

The learned policy, while not explicitly using a learning algorithm, demonstrates a markedly improved performance over the random baseline. This policy presumably uses a heuristic or strategy that aligns more closely with the goal's position, even though it does not adapt based on the reward signal. It might prioritize certain actions over others based on the agent's current state, which consistently leads to higher rewards.

Lastly, the better policy outperforms all others, including the learned policy. This suggests a more refined strategy, perhaps one that combines heuristics with a form of state evaluation that helps to determine the most promising direction to move in, without direct knowledge of the goal's location.

Conclusively, the better policy demonstrates improved performance not through learning from the reward signal, but through a static yet effective set of rules or heuristics. These guiding principles enable the agent to make decisions that are more aligned with achieving the goal, even without explicit knowledge of the goal's location, resulting in a consistently higher reward accumulation compared to other policies.