Report on the 'Lazy' Strategy Problem in the Original Q-Learning Code

Introduction

In the original implementation of the Q-Learning algorithm, the agent tends to adopt a 'lazy' strategy. This occurs when the agent finds a path or a sequence of actions that yields consistent rewards, but does not further explore other potentially better strategies. This behavior is typically the result of the Q-values stabilizing around a suboptimal strategy, leading the agent to exploit this strategy rather than explore new options.

Problem Overview

The 'lazy' strategy problem is characterized by the following issues:

- Stagnation in Q-values: The Q-values for certain state-action pairs become stable, leading the agent to repeatedly choose the same actions without improvement.
- Lack of Exploration: The agent primarily exploits known strategies without adequately exploring other potential actions, which limits the possibility of finding more optimal paths.
- Suboptimal Rewards: While the agent might achieve consistent rewards, these rewards are often suboptimal compared to what could be achieved with a more exploratory strategy.

Evidence of the Problem

During the training episodes, the agent frequently chooses the same actions in similar states, as evidenced by repeated outputs of 'Exploiting: Chose action ...'. This repetitive behavior indicates that the agent is not sufficiently exploring other possible actions. The Q-values also show minimal updates, suggesting that the agent is stuck in a local optimum.

Conclusion

The 'lazy' strategy problem in the original code limits the agent's ability to discover optimal strategies. To address this issue, modifications are needed to encourage more exploration and prevent the agent from becoming fixated on suboptimal paths. These modifications can include adjusting the reward function, increasing the exploration rate (epsilon), and considering long-term rewards by adjusting the discount factor (gamma).