CT5130 CT5134 Reinforcement Learning Assignment-2

Name: Krishan D Radadia

ID: 22239157

Analysis for Basic Hyperparameters

The analysis indicates the agent is making progress in learning. The plot of total rewards shows a clear improvement from very low initial values, suggesting the agent is finding better ways to navigate the environment. However, the presence of fluctuations after the initial rise suggests the learning process isn't entirely smooth. This could be due to the agent still exploring different options or the environment itself being particularly challenging.

Looking deeper at the Q-table, we see a mix of positive, negative, and zero values. Positive values correspond to actions the agent has learned to be beneficial, while negative values likely represent choices leading to penalties. The presence of zero values suggests there might be parts of the environment the agent hasn't explored yet, or areas that result in an immediate end state with no reward change.

The provided hyperparameters offer some insight into the agent's learning strategy. The high learning rate (alpha) indicates the agent prioritizes new information heavily, while the discount factor (gamma) suggests a strong emphasis on future rewards. The exploration rate (epsilon) seems balanced, with the agent choosing to explore 10% of the time and exploit its learned knowledge 90% of the time.

In conclusion, while the agent is demonstrably learning, there's room for improvement. The unstable learning curve suggests potential benefits from fine-tuning the hyperparameters or employing a more sophisticated exploration strategy. The Q-table confirms the agent's ability to differentiate between good and bad actions, although there might be unexplored areas that could hold further rewards.

Analysis of Epsilon Decay Strategy

The revised plot suggests a more controlled learning process compared to the basic approach. Initial penalties are lower, and rewards stabilize around a narrower range (-5 to -10). This indicates less extreme outcomes and potentially more consistent performance by the agent.

While there's still no clear upward trend in rewards, the agent seems to be avoiding very low rewards seen earlier. This suggests some improvement, possibly by focusing on actions that maintain a certain performance level.

The Q-table remains similar, with positive values for beneficial actions and negative values for penalties. The highest value (10) suggests the agent has learned a particularly effective action. Zero values still exist, indicating areas that remain unexplored or actions with neutral reward outcomes.

The key difference lies in the epsilon decay parameter. Though the initial value (0.10) is the same as before, the title "epsilon_decay" implies this value decreases over time. This means the agent starts with a higher exploration rate (trying random actions) but gradually prioritizes exploiting the knowledge it has gained (choosing actions based on learned Q-values). This shift likely explains the plot's reduced variability and suggests the epsilon decay strategy promotes more stable learning.

In essence, the epsilon decay approach seems to lead to a more consistent learning process with less fluctuation in rewards. The agent is differentiating between good and bad actions, and the decaying exploration rate likely contributes to the improved stability observed in the plot.

Analysis of Custom Hyperparameters

The plot reveals a learning process with mixed results. While rewards are slightly higher than the epsilon decay strategy (ranging from 2.5 to -17.5), they also exhibit much greater variability. This is evident from the numerous outliers, indicating episodes where the agent performed significantly better or worse than average. There's no clear improvement trend, suggesting the agent's performance is inconsistent across episodes.

The Q-table reflects this inconsistency. Positive values indicate the agent has learned favorable actions in some states, but negative values persist alongside zeros. This suggests the agent is still differentiating between good and bad actions, but the learning process might not be fully converged.

The custom hyperparameters likely contribute to this. The high learning rate (alpha = 0.7) prioritizes very recent experiences, potentially causing the agent to overfit to specific situations and forget valuable long-term knowledge. The lower discount factor (gamma = 0.8) emphasizes immediate rewards more than future ones, which might explain the lack of consistent improvement over time. Finally, the increased exploration rate (epsilon = 0.15) allows the agent to explore a wider range of actions, leading to the observed high variance in rewards.

In essence, the custom hyperparameters seem to introduce more exploration but potentially at the cost of stable learning. The agent is making some progress, but the learning curve is more erratic compared to the epsilon decay strategy.

Reinforcement Learning Strategies: Performance Analysis

This part analyzes three reinforcement learning strategies (Basic, Epsilon Decay, Custom) based on their reward curves over 10,000 episodes.

- **Basic Strategy (Orange Line):** Consistent decline suggests deteriorating performance. A fixed exploration rate might limit exploration and learning.
- **Epsilon Decay Strategy (Blue Line):** Upward trend after initial slump indicates successful learning. Dynamic exploration likely facilitates this by transitioning from exploration to exploitation.
- **Custom Strategy (Green Line):** Highest final reward but with significant volatility. Increased exploration and focus on recent experiences allow discovery of high-rewarding actions but introduce variance.

Conclusion:

Each strategy presents trade-offs. Basic is ineffective. Epsilon Decay excels with balanced exploration. Custom achieves high rewards with increased volatility. Further exploration of hyperparameters and advanced techniques could yield improved results.