Part 1: The Exploration-Exploitation Trade-Off (Epsilon Decay)

The following parameters were locked/fixed throughout this exercise:

- GAMMA = 0.9
- BUFFER_CAPACITY = 10000
- LEARNING_RATE = 0.0005
- BATCH_SIZE = 64
- EPS_START = 1.0 # starting epsilon value
- EPS_END = 0.01 # ending epsilon value
- NUM_EPISODES = 5000 # number of episodes
- TARGET_UPDATE_FREQ = 500
- MAX_STEPS = 100 # cap per episode so rewards do not explode

Reward Structure:

Goal Reached = +50 when the agent arrives at the goal
Time Step Penalty = -0.1 on each move to discourage slow/unnecessary movement
Crash Penalty = -10 and immediate episode termination when the agent hits a wall
Reward Shaping = A small adjustment proportional to the change in distance to the goal.

- Positive bonus when a move reduces the distance to the goal, and negative when a move increases the distance to the goal.

High Epsilon Decay Rate

- **Strategy**: Rapidly favors exploration
- **Expected Outcome**: Settles quickly, but potentially sub-optimally. The agent commits to a path before fully exploring other potentially better ones.
- **Explanation**: A high decay rate means that the epsilon value drops very quickly. The agent risks settling on a sub-optimal path because it stops randomly searching for better routes too soon. Once it finds a decent path, it rarely deviates to check if a better one exists.

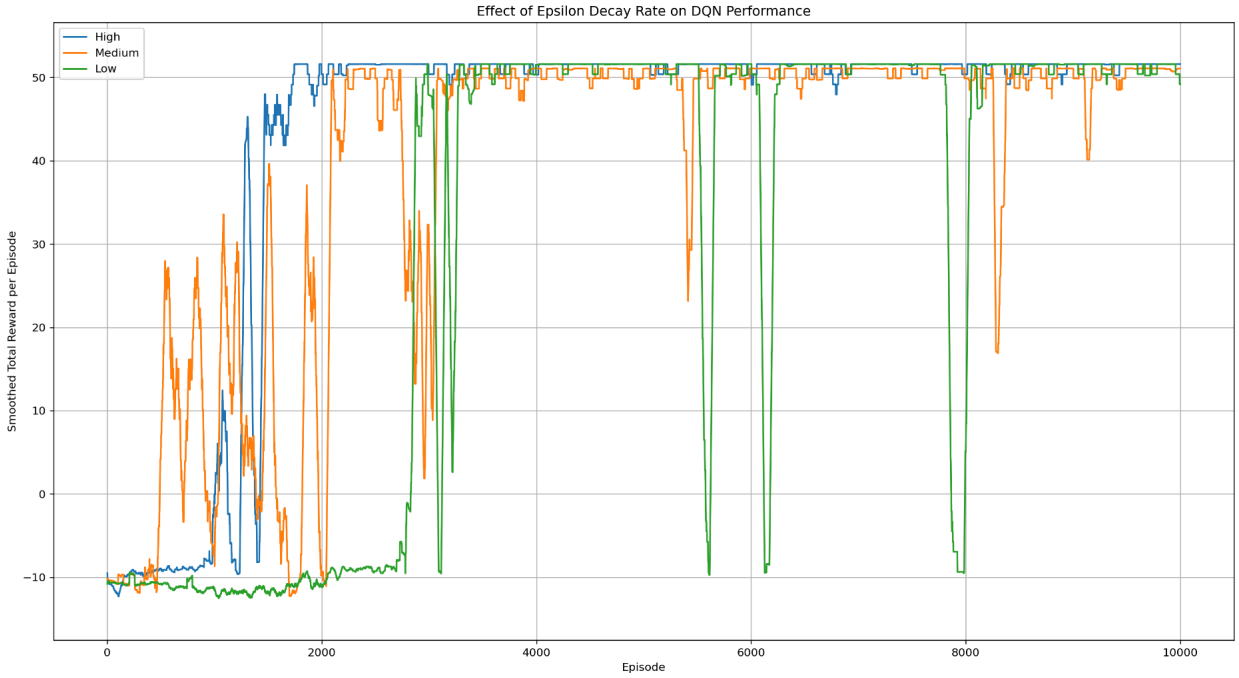
Medium Epsilon Decay Rate

- **Strategy**: Balanced
- **Expected Outcome**: Good trade-off. It finds a solid path relatively efficiently.

Low Epsilon Decay Rate

- **Strategy**: Slowly favors exploitation.
- **Expected Outcome**: Takes the longest to stabilize, but finds the optimal path. The agent keeps exploring longer, giving it the best chance to discover the absolute shortest route.
- **Explanation**: A low decay rate means epsilon stay high for many episodes. The agent offers the ebay chance of finding the absolute optimal path because it continues to explore randomly for a much longer time, testing many more state-action combinations. However, it takes the longest to stabilize and its performance (Average Reward) will be lower in the early stages due to the excessive amount of random, non-productive exploration.

Effect of Epsilon Decay on DQN Performance:



The plot above shows how different epsilon decay rates are able to influence learning behavior of a DQN agent attempting to navigate through a grid. When a high epsilon decay rate was used, there was a rapid shift from exploration to exploitation. This causes its performance to improve quickly and reach a very strong path after around 2,000 episodes. Once it does, there is very little fluctuation, showing that the agent commits early and rarely deviates from the path it discovered. However, because of the sudden/early reduction in exploration, it may overlook other potentially better routes, which fits the expectation that a high decay rate causes convergence to happen fast, but it could also lead to settling on a sub-optimal route.

The medium decay strikes a more balanced strategy, as its early performance is slower to improve than when the high decay rate was used. When the medium decay rate is used, we see a gradual improvement as it becomes more consistent and competitive. This aligns with the expected trade-off we were expecting to see, as there is a moderate decay rate that allows for it to not try to exploit too early but also avoids exploring for too long. Even after stabilizing, the medium decay curve shows minor dips (small downward fluctuations) likely caused by occasional exploratory decisions still being attempted before the agent fully settles.

The low decay rate also behaves as predicted due to its long exploration phase. It takes much more time before reaching higher rewards (around 3,000 episodes) because of its desire to continue testing random combinations for longer. Although this slows down the learning, the extended exploration would theoretically provide a better chance of finding the most optimal path through the grid. Once it confirms that path, its performance stabilizes, matching the ideas

that low decay offers the strongest chance of discovering the best route. Once the green line stabilizes, it still shows some sharp drops at several points. These sudden drops highlight that even late in training, the agent is still waiting to explore, sometimes risking a crash or inefficient path simply because its epsilon value remains higher than it would be at the same stage under a medium or high decay rate.

These three behaviors lead directly into where our results deviated slightly from the ideal theoretical expectation. In theory, the extra exploration in the low decay case should eventually allow it to outperform the others. However, in this experiment, the high decay agent often finishes with slightly higher average rewards once all of three strategies stabilize. This outcome is driven by the environment's reward structure. Because of it, a crash or inefficient wandering produces a large penalty (negative reward), so even one risky action late in training can make an episode much worse. Both the medium and low decay agents continue to explore even when they are late into the training process. This means that they are left vulnerable to these costly mistakes. The high decay agent, on the other hand, stops exploring early and avoids those penalties altogether. The difference between a "good route" and the absolute best route in this grid is relatively small, which means that avoiding late crashes ends up being more valuable than continuing to search for a possibly better route.

Overall, this experiment aligns very closely with the expected epsilon decay behavior. High decay learns quickly but risks missing the best path, medium decay provides a balanced progression, and low decay takes the longest to stabilize due to heavy exploration. The only deviation in the observed behavior was that the extended exploration (high epsilon decay agent) does not produce the highest final reward. This is because of the penalty structure, where safety becomes more valuable than more exploration once a "good route" is found.

Part 2: The Planning Horizon

In this experiment, I investigated how different gamma values influence a DQN agent's ability to navigate the grid-based driving course. Gamma controls how strongly the agent values future rewards over immediate ones. As a result, this value shapes the planning horizon of the agent as it aims to find a path through the grid.

Since the high epsilon decay rate performed the best in part 1, that was used in this experiment. All of the other parameters were kept the same.

Expectations:

High Gamma

- **Interpretation of Gamma**: Far-sighted. Future rewards are valued almost as much as immediate ones.
- **Expected Outcome**: Finds the shortest path. Prioritizes the final, large goal reward.
- **Explanation**: The Goal reward is discounted with each step. The agent is not strongly motivated to move quickly toward the goal, which can lead it to find a longer than optimal path or even get "stuck" if a local state offers less penalty than a necessary stepping stone.

Medium Gamma

- **Interpretation of Gamma**: Balanced
- **Expected Outcome**: Good performance. It balances immediate and future rewards.

Low Gamma

- **Interpretation of Gamma**: Short-sighted. Future rewards are heavily discounted.
- **Expected Outcome**: Finds a longer than optimal path. Prioritizes immediate rewards/penalties over the distant goal.
- **Explanation**: The future Goal reward is discounted very little. The agent has the best chance of finding the shortest possible path, as it prioritizes the long-term payout over the immediate movement penalty.

The tested Gamma values and the results are listed below:

```
Summary across gamma values

Case     Gamma   Successes  Success rate  Avg steps to goal  Shortest path
-----------------------------------------------------------------------
Low      0.30    35/1000    3.5%          30.7               9
Medium   0.60    583/1000   58.3%         12.7               9
High     0.99    166/1000   16.6%         19.8               9
```

One important detail in assessing performance is that the theoretical optimal route from the start to the goal requires 9 actions, due to the grid layout. This gives a benchmark for evaluating navigation quality.

Several outcomes aligned with theoretical expectations. A higher gamma should encourage longer-term planning since the agent aims to prioritize the final goal reward instead of immediate penalties. As expected, the low gamma agent behaved in a short-sighted manner, which produced the lowest success rate among the three and frequently became stuck or inefficient paths. The high gamma agent outperformed the low gamma case in terms of success rate. This shows that discounting the future less leads to better overall task completion. The optimal 9-step solution was discovered by all three cases, which means that each gamma setting is technically capable of achieving the best possible path to take through the grid.

However, the results also reveal something that was unexpected. The high-gamma agent did not produce the best performance, but instead, the medium gamma agent (with a value of 0.60) achieved the strongest results (by far) overall. It reached the goal in 58.3% of the trials and was able to take the fewest average steps to reach the goal. This improvement reflects consistent learning (as opposed to random luck) since the agent repeatedly produced near-optimal paths rather than stumbling into them by chance. This means that the medium gamma agent not only succeeds more often, but it also reaches the goal using more efficient paths when it does. The medium agent, on the other hand, averaged about 12.7 steps to reach the goal, which is the closest of our agents to the theoretical optimal path of 9 steps. The low and high gamma agents took more steps than this, on average. This shows that the medium gamma is able to find a balance by reaching the goal strongly enough to avoid looping or drifting unnecessarily and penalizing wasted actions heavily enough to ensure that excessive detours are avoided too.

Even though high gamma is a far-sighted configuration, it often tolerated inefficient moves as long as it believed it would eventually obtain a large reward at the goal state. The low gamma agent struggled more so, and it rarely valued the future goal enough to plan a successful path through the grid. At times, this agent found a 9-step path, but this appears to have occurred due to random exploration. These two cases make clear the idea that with a low gamma, the rare 9-step success(es) occurred only due to exploratory randomness, whereas with a medium gamma value, these efficient paths were discovered and pursued due to actual learned behavior.

All in all, these results show that each agent is capable of finding a path that is 9-steps (sometimes by random chance), but only the medium agent learns to approach this level of performance reliably. By balancing long-term planning with immediate penalties, the medium gamma configuration produced the most stable navigation behavior, the highest success rate, and the most efficient average path length toward the goal.