

## Experiments with Q-Learning Parameters

### 1. Exploration-Exploitation Balance: Varying Epsilon ( $\epsilon$ )

- **Experiment:** Test different values of epsilon (e.g., 0.1, 0.3, 0.5, 0.8).
- **Objective:** Observe how more or less exploration impacts the agent's learning speed and path efficiency.
- **Questions:**
  - What happens when epsilon is high (more exploration)?
  - Does a low epsilon (more exploitation) lead to faster or slower learning?

### 2. Learning Rate Impact: Varying Alpha ( $\alpha$ )

- **Experiment:** Test different values of alpha (e.g., 0.01, 0.1, 0.5, 0.9).
- **Objective:** See how a higher or lower learning rate affects how quickly the agent updates its Q-values and learns the optimal path.
- **Questions:**
  - What effect does a low alpha have on learning speed?
  - Does a high alpha cause the agent to learn faster but risk instability?

### 3. Discount Factor Impact: Varying Gamma ( $\gamma$ )

- **Experiment:** Test different values of gamma (e.g., 0.1, 0.5, 0.9, 0.99).
- **Objective:** Explore how the discount factor influences the agent's focus on short-term versus long-term rewards.
- **Questions:**
  - How does a low gamma (more focus on immediate rewards) affect the path?
  - How does a high gamma (focus on long-term rewards) affect the agent's decisions?

### 4. Number of Episodes: Increasing Training Time

- **Experiment:** Run the model with varying numbers of episodes (e.g., 100, 500, 1000, 2000).
- **Objective:** Examine how training duration impacts the agent's policy stability and optimal path formation.
- **Questions:**
  - Does a higher episode count improve path consistency?
  - At what point does additional training show diminishing returns?

---

## Experiments with Environment Settings

### 5. Adding More Obstacles

- **Experiment:** Place additional obstacles in the grid-world (e.g., increase obstacle count to 5 or 6).
- **Objective:** Observe how the agent adapts to navigate a more challenging environment with more negative rewards.
- **Questions:**
  - How does the agent's path change with additional obstacles?
  - Does it take longer for the agent to learn the optimal path in a more challenging grid?

## 6. Moving the Goal Position

- **Experiment:** Change the goal position to different locations in the grid (e.g., top-right corner, center of the grid).
- **Objective:** Test if the agent's learned policy can adapt quickly to new goal locations.
- **Questions:**
  - How does a goal position change affect learning?
  - Does the agent take longer to adapt when the goal is moved farther from the original position?

## 7. Using Higher Penalties for Wrong Moves

- **Experiment:** Increase the penalty for hitting obstacles (e.g., from -10 to -20 or -50).
- **Objective:** See if higher penalties discourage risky moves more effectively and impact the agent's learned policy.
- **Questions:**
  - How do higher penalties affect the path taken?
  - Does a higher penalty encourage the agent to take safer routes?

## 8. Reward Adjustments for Steps Taken

- **Experiment:** Change the step penalty from -1 to 0 (or even +1).
- **Objective:** Test how the agent's behavior changes when it's not penalized for taking extra steps, or is rewarded for each move.
- **Questions:**
  - Does removing the step penalty lead to longer paths?
  - Does a positive step reward encourage the agent to explore more of the grid?

---

## Comparative Experiments

### 9. Q-Learning vs. SARSA

- **Experiment:** Implement SARSA instead of Q-learning and compare the resulting paths.
- **Objective:** See how on-policy (SARSA) vs. off-policy (Q-learning) learning methods influence the agent's strategy.
- **Questions:**
  - How does the agent's learned path differ between SARSA and Q-learning?
  - Does one algorithm handle exploration or obstacles more effectively?

### 10. Different Grid Sizes

- **Experiment:** Test the model on a larger grid (e.g., 10x10 or 15x15).
  - **Objective:** Explore how increasing grid size impacts the agent's learning time and path complexity.
  - **Questions:**
    - How does a larger grid affect learning time?
    - Does the agent struggle to reach the goal in larger grids, especially with more obstacles?
-

## Extension Ideas for Advanced Exploration

### 11. Dynamic Goal Position

- **Experiment:** Move the goal position randomly at each episode and observe how well the agent adapts.
- **Objective:** Test the model's flexibility in dynamic environments.
- **Questions:**
  - Can the agent still learn effective policies when the goal is unpredictable?
  - How does this affect the stability of the Q-values?

### 12. Changing Epsilon Over Time (Epsilon Decay)

- **Experiment:** Use a decaying epsilon value to gradually reduce exploration as episodes increase (e.g.,  $\epsilon = \epsilon \cdot \text{decay\_rate}$  after each episode).
- **Objective:** Observe how gradually shifting from exploration to exploitation impacts learning.
- **Questions:**
  - Does decaying epsilon result in a more efficient learning path?
  - Is there an ideal rate of decay to balance learning speed and path optimization?

## Key Metrics for Each Experiment

To help students understand the effects of their manipulations, here's a tailored list of metrics that align with each experiment:

---

### 1. Exploration-Exploitation Balance: Varying Epsilon ( $\epsilon$ )

- **Primary Metrics:**
  - **Exploration vs. Exploitation:** Track epsilon values and observe how the agent's exploration behavior changes.
  - **Episode Length:** Shorter episodes often suggest better exploitation as the agent learns the environment.
- **Secondary Metrics:**
  - **Cumulative Reward:** Monitor whether cumulative rewards improve as the agent finds effective strategies.

---

### 2. Learning Rate Impact: Varying Alpha ( $\alpha$ )

- **Primary Metrics:**
    - **Cumulative Reward:** Check if higher learning rates lead to faster accumulation of rewards or if they destabilize the learning process.
    - **Policy Stability:** Observe if the agent's actions become inconsistent with high alpha values, indicating unstable learning.
  - **Secondary Metrics:**
    - **Episode Length:** Lower episode lengths with stable cumulative rewards suggest that the agent is efficiently converging.
-

### 3. Discount Factor Impact: Varying Gamma ( $\gamma$ )

- **Primary Metrics:**
    - **Cumulative Reward:** Observe if higher gamma values lead to higher cumulative rewards, as they encourage the agent to consider long-term outcomes.
    - **Episode Length:** Lower episode lengths with high cumulative rewards suggest the agent is planning well.
  - **Secondary Metrics:**
    - **Policy Stability:** A stable policy with high cumulative rewards indicates effective long-term planning.
- 

### 4. Number of Episodes: Increasing Training Time

- **Primary Metrics:**
    - **Average Reward per Episode:** Observe how rewards stabilize over episodes to understand the agent's learning progress.
    - **Time to Convergence:** Check when rewards stop increasing significantly, showing that the agent's policy has stabilized.
  - **Secondary Metrics:**
    - **Policy Stability:** Observe if actions in the animation become consistent, showing the agent has settled on a strategy.
- 

### 5. Adding More Obstacles

- **Primary Metrics:**
    - **Episode Length:** Longer episodes may indicate the agent is struggling to find paths around obstacles.
    - **Success Rate:** Monitor if the agent consistently reaches the goal as the environment becomes more complex.
  - **Secondary Metrics:**
    - **Cumulative Reward:** Lower rewards with additional obstacles might indicate difficulty adapting to the new environment.
- 

### 6. Moving the Goal Position

- **Primary Metrics:**
    - **Cumulative Reward:** Watch how cumulative rewards adjust as the agent learns to reach a new goal.
    - **Episode Length:** Longer initial episodes after the goal move indicate the agent is adjusting its path.
  - **Secondary Metrics:**
    - **Time to Convergence:** Observe how long it takes for the agent's performance to stabilize after moving the goal.
- 

### 7. Using Higher Penalties for Wrong Moves

- **Primary Metrics:**
  - **Episode Length:** If penalties discourage risky moves, episode length may decrease as the agent learns safer paths.

- **Cumulative Reward:** Observe if the agent avoids penalties and maintains or improves cumulative rewards.
  - **Secondary Metrics:**
    - **Policy Stability:** Look for consistency in actions as the agent finds ways to avoid penalties.
- 

## 8. Reward Adjustments for Steps Taken

- **Primary Metrics:**
    - **Episode Length:** Observe if step penalties or rewards encourage the agent to find shorter or longer paths.
    - **Cumulative Reward:** A higher cumulative reward with positive step rewards suggests the agent is taking optimal paths while still maximizing rewards.
  - **Secondary Metrics:**
    - **Policy Stability:** A stable policy with high cumulative rewards suggests the agent has learned an efficient path.
- 

## 9. Q-Learning vs. SARSA

- **Primary Metrics:**
    - **Success Rate:** Observe if SARSA's cautious approach leads to higher success in environments with risks, compared to Q-learning.
    - **Policy Stability:** Check if SARSA produces more stable paths, while Q-learning may appear more aggressive or direct.
  - **Secondary Metrics:**
    - **Cumulative Reward:** Higher rewards may indicate which method is learning an optimal path.
- 

## 10. Different Grid Sizes

- **Primary Metrics:**
    - **Episode Length:** Larger grids may increase episode length initially as the agent learns to navigate.
    - **Time to Convergence:** Track if it takes longer for the agent to find stable policies in larger grids.
  - **Secondary Metrics:**
    - **Success Rate:** Verify if the agent can consistently reach the goal even with a larger state space.
- 

## Summary of Key Metrics Across Experiments

- **Cumulative Reward:** Central to most experiments, as it indicates whether the agent is effectively learning to maximize rewards.
- **Episode Length:** Essential for understanding efficiency and showing if the agent is finding shorter, more effective paths.
- **Exploration vs. Exploitation:** Vital for experiments involving epsilon, where observing exploration behavior helps students understand the effects of exploration settings.

- **Policy Stability:** Important for all experiments, as it indicates if the agent is finding a consistent, reliable policy over time.

### 1. Cumulative Reward

- **Definition:** The total reward accumulated by the agent in each episode.
  - **What to Look For:**
    - **Increasing Cumulative Reward:** A general upward trend indicates that the agent is learning to achieve its goal (maximizing rewards) more effectively. This suggests the agent is avoiding penalties (obstacles) and reaching the goal consistently.
    - **Plateau:** When cumulative reward levels off, it usually means the agent has stabilized in its learning and found an optimal or near-optimal path.
    - **Fluctuations:** Frequent ups and downs in cumulative reward may suggest that the agent's policy is unstable or that the environment is challenging (e.g., many obstacles). This could also mean that parameters like alpha or epsilon need adjustment.
  - **Questions to Ask:**
    - Does the cumulative reward increase steadily, or does it fluctuate?
    - How long does it take for the reward to stabilize? Does a particular setting lead to faster stabilization?
- 

### 2. Episode Length (Steps per Episode)

- **Definition:** The number of steps the agent takes to complete an episode, ideally reaching the goal.
  - **What to Look For:**
    - **Decreasing Episode Length:** As the agent learns, you should see a trend toward shorter episodes. This indicates that the agent is finding more efficient paths and avoiding unnecessary moves.
    - **Stabilization:** When episode length levels off, it often suggests that the agent has learned an optimal or stable path to the goal.
    - **High Episode Length:** Consistently high episode lengths may indicate that the agent is struggling to find efficient routes, possibly due to high exploration (epsilon), a low learning rate (alpha), or a challenging environment (e.g., lots of obstacles).
  - **Questions to Ask:**
    - Does the agent find shorter paths as learning progresses?
    - How quickly does episode length stabilize, and does a particular setting lead to shorter episodes?
- 

### 3. Exploration vs. Exploitation Balance

- **Definition:** The proportion of exploratory actions (random choices) versus exploitative actions (choosing the best-known option).
- **What to Look For:**

- **High Exploration Early:** At the beginning, you should see a high exploration rate as the agent tries to learn about the environment. This allows it to gather information on different states and actions.
  - **Shift to Exploitation:** Over time, the exploration rate should decrease as the agent starts to exploit known strategies. This indicates that the agent is becoming more confident in its choices and focusing on maximizing rewards.
  - **Persistent Exploration:** If exploration remains high throughout training, it may mean the agent isn't effectively settling on a policy. This can happen if epsilon is too high or if there's no epsilon decay. Alternatively, too little exploration can mean the agent isn't learning enough about the environment.
  - **Questions to Ask:**
    - Does the exploration rate decrease over time as expected?
    - How does exploration-exploitation balance affect other metrics (like cumulative reward and episode length)?
- 

#### 4. Policy Stability

- **Definition:** The number of changes in the agent's policy (chosen actions in each state) over time. High stability suggests that the agent's behavior is becoming consistent.
  - **What to Look For:**
    - **Initial Policy Changes:** At the beginning, it's normal to see frequent changes in policy as the agent explores different actions.
    - **Reduction in Policy Changes:** Over time, as the agent learns an effective strategy, the number of policy changes should decrease. This stabilization reflects that the agent is consistently selecting actions that lead to higher rewards.
    - **Persistent Policy Changes:** If policy changes remain high, it may indicate instability in learning, possibly due to a high learning rate (alpha) or a lack of convergence. This could also mean the agent is "unlearning" previous strategies, which may suggest a need for adjustment in alpha or gamma.
  - **Questions to Ask:**
    - Do policy changes decrease over time, showing the agent is settling on a consistent path?
    - Are there sudden increases in policy changes? If so, consider if certain parameters need adjustment.
- 

#### How Each Metric Relates to Parameter Adjustments

Here's a quick guide to how each metric may change with different parameters, helping students understand the impact of their manipulations:

- **Learning Rate (alpha):**
  - High alpha can lead to **fast learning** but may cause **instability** (frequent policy changes) if it's too high.
  - Low alpha slows down learning, possibly resulting in more stable but **slower cumulative reward growth** and **longer episode lengths**.
- **Discount Factor (gamma):**

- Higher gamma encourages **long-term planning**, often resulting in **lower episode lengths** as the agent finds efficient paths.
- Lower gamma means the agent favors immediate rewards, which can lead to **shorter-term paths** that might not be optimal in complex environments.
- **Exploration Rate (epsilon):**
  - Higher epsilon encourages **exploration**, which is useful early on but may lead to **longer episode lengths** as the agent explores.
  - Lower epsilon (or epsilon decay) should lead to **increasing cumulative rewards** as the agent shifts to exploiting known strategies, resulting in **shorter episode lengths** over time.