EXPERIMENTING WITH LEARNING POLICIES ON SIMPLE REINFORCEMENT POLICIES USING DIFFERENT METHODS

by

MARY DARKWAA NTIM

RAHMAT EFUA ETUAFUL

ASSIGNMENT 2

MSC DATA SCIENCE

STUDENT ID's: 202398287 and 202485757

Part 1- The Gridworld Problem

The environment consists of a 5 × 5 gridworld, where each of the 25 cells represents a possible state. The agent can move one step up, down, left, or right at each time step. If the agent attempts to move off the grid, it remains in the same position.

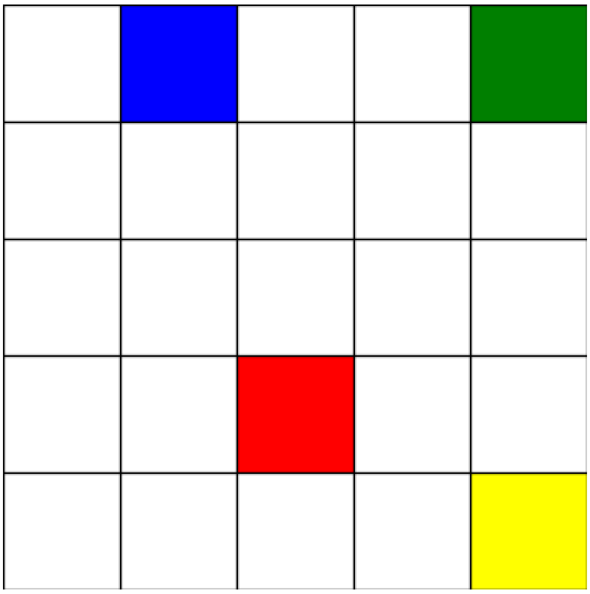


Figure 1: Gridworld environment showing colored special states: blue, green, red, and yellow squares as described.

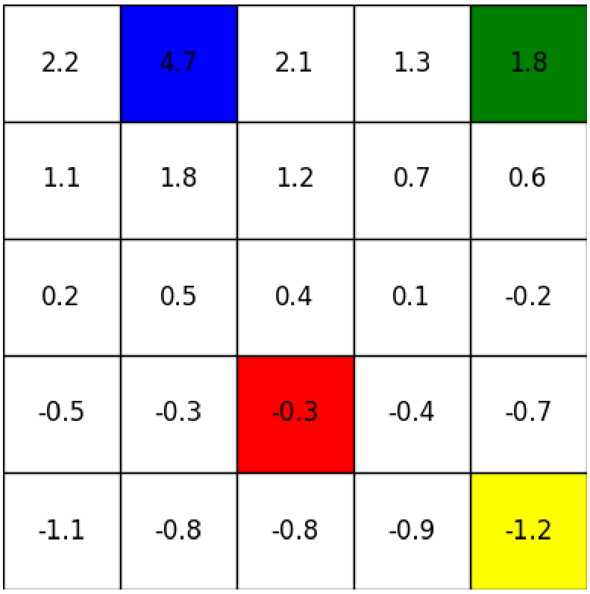
There are four special states in the grid:

- The **blue** square: Any action from this state yields a reward of 5 and moves the agent instantly to the **red** square.
- The **green** square: Any action yields a reward of 2.5 and teleports the agent to either the **yellow** or **red** square, each with probability 0.5.
- The **red** and **yellow** squares: Actions from these states yield no immediate reward unless the agent attempts to move off the grid from a white or yellow square, which results in a penalty of -0.5.

The discount factor $\gamma = 0.95$ is used. The policy under evaluation selects any of the four actions with equal probability 0.25.

Solving Bellman's equations explicitly

The value function $V(s)$ computed by explicitly solving the system of Bellman equations is shown below (visualized on the grid):



2.2	4.7	2.1	1.3	1.8
1.1	1.8	1.2	0.7	0.6
0.2	0.5	0.4	0.1	-0.2
-0.5	-0.3	-0.3	-0.4	-0.7
-1.1	-0.8	-0.8	-0.9	-1.2

Figure 2: Estimated state-value function overlayed on the colored gridworld states.

Interpretation and Analysis

- The highest value, approximately 4.73, is located at the **blue** square (second cell). This reflects the immediate reward of 5 and the subsequent teleportation to the red square, yielding a high expected return.
- The **green** square (fifth cell) has a high value of about 1.78, representing the 2.5 reward and equal chance teleportation to the yellow or red square.
- The **red** square (eighteenth cell) and **yellow** square (last cell) have much lower values

(close to or below zero), consistent with no immediate positive rewards and penalties for moving off the grid from the yellow square.

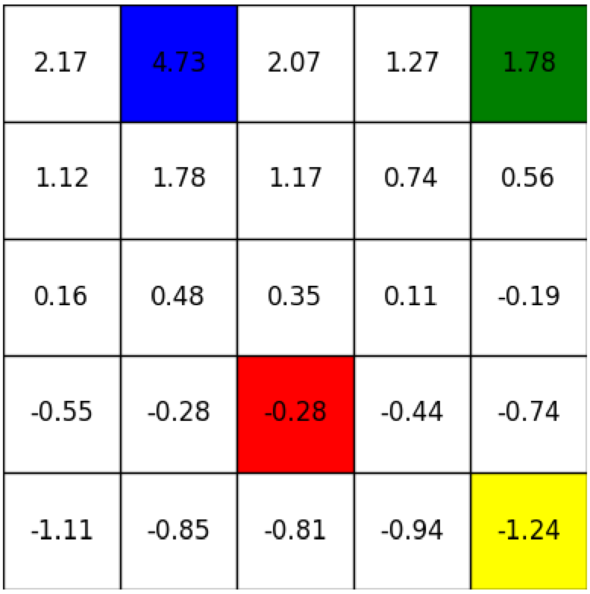
- Values tend to decrease gradually with distance from these rewarding squares, confirming the effect of discounting and the random uniform policy choice.
- Negative values, especially near the bottom row and edges, reflect the penalty of -0.5 incurred when attempting to move off the grid, lowering expected returns in those states.

Overall, the results align with the intuition that the states offering high immediate rewards and beneficial teleportation yield the highest expected values. The policy's randomness results in a gradual decay of values as the states become less reachable or less rewarding.

The next step involves iterative policy evaluation to verify these findings.

Iterative Policy Evaluation

Using the same reward discount $\gamma = 0.95$ and equiprobable random policy (equal probability of 0.25 for each action), iterative policy evaluation was performed to estimate the value function for each state.



2.17	4.73	2.07	1.27	1.78
1.12	1.78	1.17	0.74	0.56
0.16	0.48	0.35	0.11	-0.19
-0.55	-0.28	-0.28	-0.44	-0.74
-1.11	-0.85	-0.81	-0.94	-1.24

Interpretation: The value function obtained through iterative policy evaluation reveals that the blue square (2nd cell, index 1) holds the highest value (4.73) on the grid due to its large immediate reward of 5 and the transition to the red square. The green square (5th cell, index 4) also shows relatively high value (1.78) given its immediate reward of 2.5 and probabilistic transition to the red and yellow squares.

The red square (18th cell, index 17) and yellow square (25th cell, index 24) exhibit low or negative values (-0.28 and -1.24 respectively), consistent with their lack of immediate rewards and penalties associated with boundary conditions.

Values decrease from top to bottom rows, reflecting the discounted future rewards, while states at edges and corners receive penalties for attempted moves off the grid, leading to more negative values.

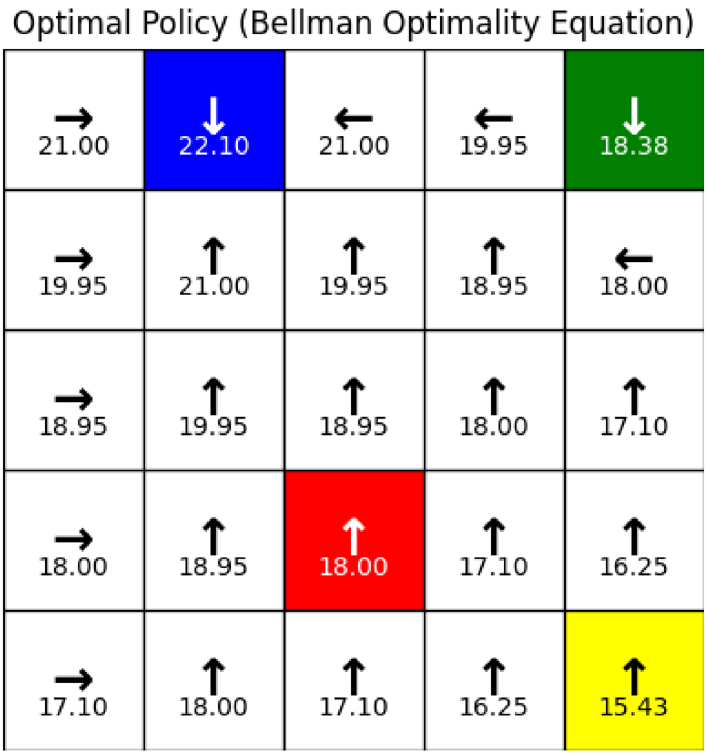
This pattern aligns with expectations under an equiprobable random policy where actions are chosen uniformly at random and the discount factor $\gamma = 0.95$ reduces the weight of distant rewards.

Determining the Optimal Policy for the Gridworld

The optimal policy for the 5x5 gridworld environment was determined using three methods:

1. Explicitly solving the Bellman optimality equation.
2. Policy iteration with iterative policy evaluation.
3. Policy improvement with value iteration.

All methods converged to the same optimal policy and value function.



Optimal Policy (Bellman Optimality Equation)

→ 21.00	↓ 22.10	← 21.00	← 19.95	↓ 18.38
→ 19.95	↑ 21.00	↑ 19.95	↑ 18.95	← 18.00
→ 18.95	↑ 19.95	↑ 18.95	↑ 18.00	↑ 17.10
→ 18.00	↑ 18.95	↑ 18.00	↑ 17.10	↑ 16.25
→ 17.10	↑ 18.00	↑ 17.10	↑ 16.25	↑ 15.43

Figure 3: Optimal policy and corresponding state values for the gridworld environment. The arrows indicate the best action at each state, and the numbers represent the optimal state values.

Interpretation

The optimal policy guides the agent predominantly upwards in the grid, reflecting a strategy to move towards states that yield higher cumulative rewards. Specifically, the agent is encouraged to reach the blue and green special states, which provide immediate rewards and beneficial transitions.

The value function shows the highest values at and near the blue square (2nd state), which offers a reward of 5, and the green square (5th state), offering a reward of 2.5. The discounted returns diminish gradually as the agent moves away from these states, consistent with the discount factor $\gamma = 0.95$.

In addition, strategic lateral moves on the top row suggest an efficient path planning towards the rewarding states. The red and yellow squares, being terminal or absorbing states, have moderate values reflecting their transition roles without additional rewards.

These results confirm that the agent optimally balances immediate rewards with future expected returns, following a policy that maximizes cumulative discounted rewards in this gridworld.

Conclusion

The consistent optimal policy and value function obtained through all methods validate the correctness of the solution and provide a clear baseline for reinforcement learning benchmarks in this environment.

Part 2: Modified Gridworld with Terminal States

Problem Description

In this part, the gridworld environment is modified to include terminal states where the episode ends immediately upon arrival. The agent interacts with the environment as follows:

- Certain states are designated as terminal, causing episode termination upon entry.
- Moving from most non-terminal states to any adjacent state yields a reward of -0.2.
- Attempting to move off the grid results in a penalty of -0.5.
- The discount factor for future rewards is set to $\gamma = 0.95$.

The exact positions and colors of the special and terminal states are depicted in the grid below.

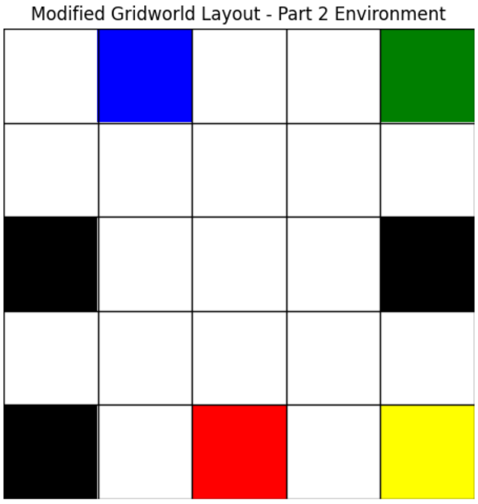
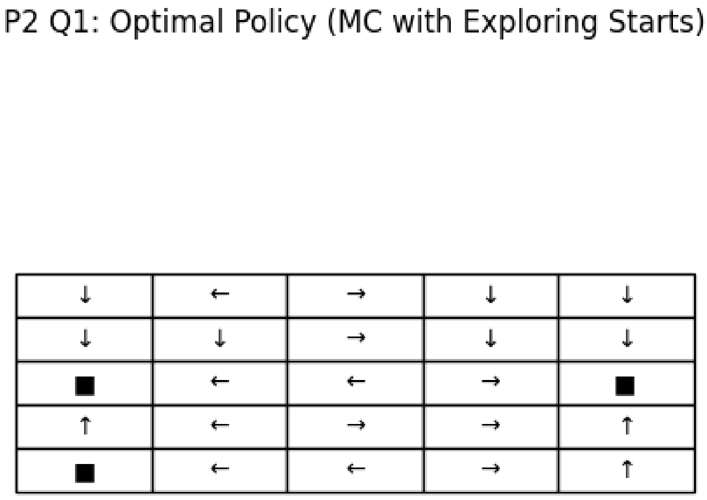


Figure 4: Modified gridworld layout with special states colored as specified. Black squares indicate terminal states.

Monte Carlo Control with Exploring Starts

Method: Exploring starts means each episode begins from a randomly selected state-action pair (excluding terminal states). This ensures broad exploration of the state-action space. The agent acts following the current policy, and the action-value function $Q(s,a)$ is updated from observed returns. The policy is improved greedily with respect to Q.



P2 Q1: Optimal Policy (MC with Exploring Starts)

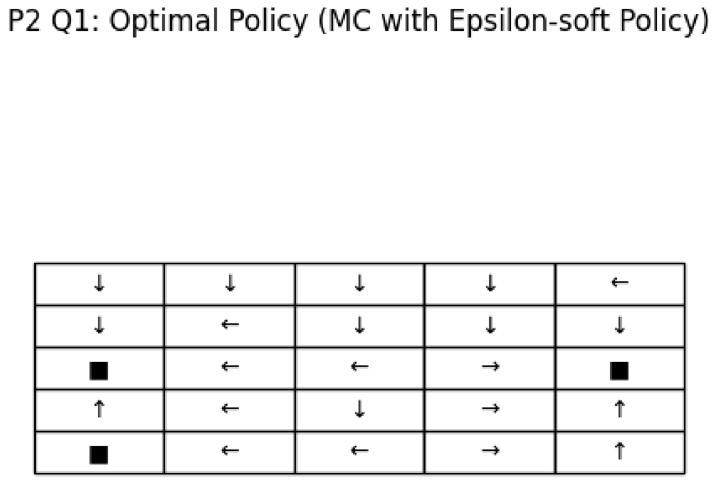
↓	←	→	↓	↓
↓	↓	→	↓	↓
■	←	←	→	■
↑	←	→	→	↑
■	←	←	→	↑

Figure 5: Optimal policy learned with Monte Carlo control using exploring starts. Arrows indicate the best action per state; black squares are terminal states.

Interpretation: The learned policy mostly directs the agent downward and towards terminal states, which is expected due to the position of high-value terminal squares and penalties for invalid moves. Exploring starts enable sufficient coverage of state-action pairs, resulting in a robust and intuitive policy.

Monte Carlo Control with ε-soft Policy (No Exploring Starts)

Method: Here, episodes always start from a fixed initial state. The policy is ε-soft: with probability $1-\varepsilon$, the agent selects the greedy action; with probability ε, it explores by choosing a random action. This encourages exploration during episodes without varying start states.



↓	↓	↓	↓	←
↓	←	↓	↓	↓
■	←	←	→	■
↑	←	↓	→	↑
■	←	←	→	↑

Figure 6: Optimal policy learned with Monte Carlo control using ε-soft policy without exploring starts.

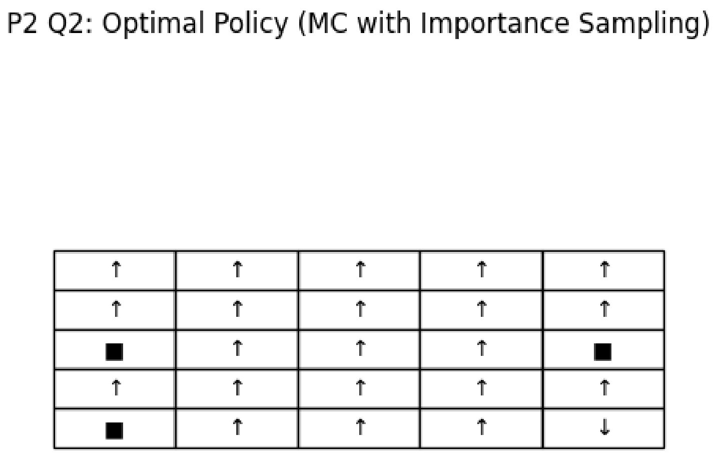
Interpretation: The policy differs from exploring starts, exhibiting more rightward moves in some areas. Fixed starting states limit exploration of the entire grid, leading to less reliable Q-estimates in states rarely visited. The ε-soft strategy partially mitigates this but cannot fully replace the benefits of exploring starts.

Monte Carlo Control with Importance Sampling (Off-policy)

Task: Use a behaviour policy with equiprobable moves to learn an optimal policy. The exact environment dynamics enable computation of importance weights needed for off-policy learning.

Method:

Using importance sampling, we evaluate and improve a greedy target policy based on episodes generated from an equiprobable behaviour policy. Importance weights correct for the distribution mismatch between behaviour and target policies.



↑	↑	↑	↑	↑
↑	↑	↑	↑	↑
■	↑	↑	↑	■
↑	↑	↑	↑	↑
■	↑	↑	↑	↓

Figure 7: Optimal policy learned with Monte Carlo control using importance sampling.

Interpretation: This policy predominantly favors upward moves, contrasting with the other methods. The difference arises from the variance and bias introduced by importance sampling due to the mismatch between behaviour and target policies. While theoretically enabling off-policy learning, practical challenges can cause such policies to be less intuitive and possibly suboptimal.

Overall Discussion

- **Exploring Starts** provides thorough state-action exploration, leading to robust policies at the cost of randomized episode starts.
- **ε-soft without exploring starts** simplifies implementation but risks insufficient state-space coverage, affecting policy optimality.
- **Importance Sampling** enables learning from off-policy data but suffers from high variance and possible bias, reflected in unexpected policy directions.

These methods illustrate key trade-offs between exploration, sample efficiency, and stability in Monte Carlo reinforcement learning.

Conclusion

This part demonstrated Monte Carlo control techniques on a modified gridworld with terminal states. Exploring starts yielded the most reliable policy. The ε-soft method is simpler but less effective in full exploration. Importance sampling highlights challenges in off-policy learning, often producing less stable results in practice.