

Report: Solving a 3D Gridworld Problem using MDP & RL

Group Members:

K Yaswanth Reddy - 2022A7PS0109P

Due Date: 03/10/2025

● Part A: MDP & RL Formulation

- We formulate the 3D Gridworld problem as a Markov Decision Process (S, A, P, R, γ) as follows:

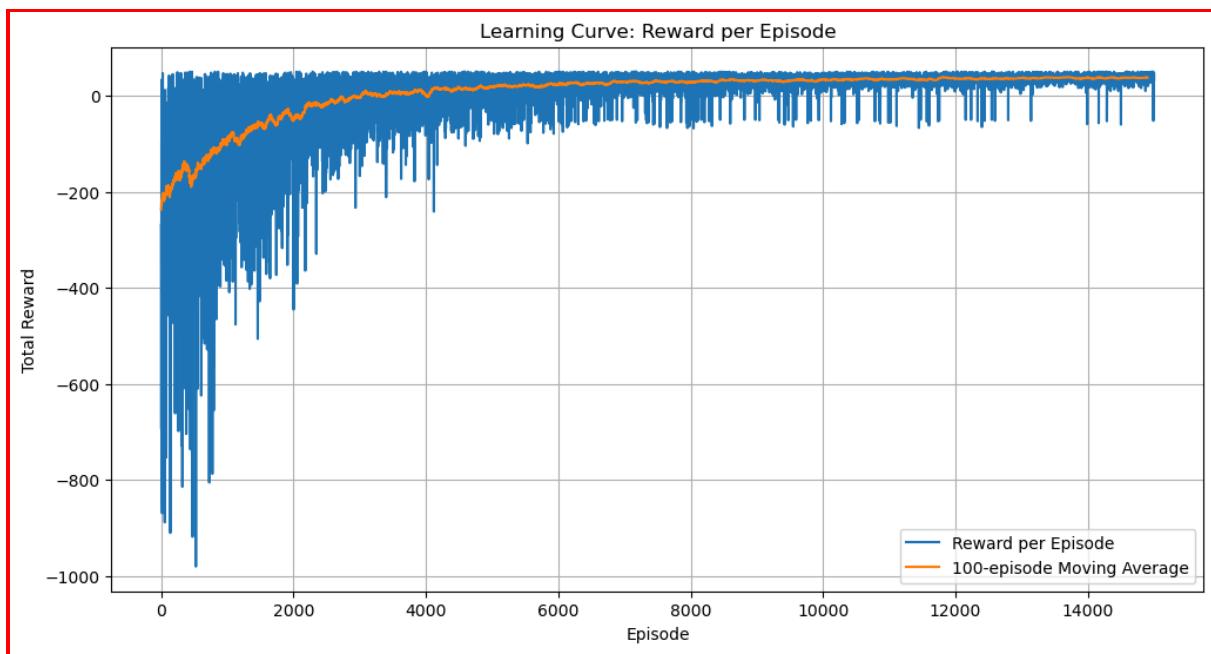
1. **States (S):** The set of all possible locations for the agent. The grid is of size $H \times W \times D$ ($6 \times 6 \times 6$). A state ' s ' is represented by a tuple '(x, y, z)' where ' $0 \leq x, y, z < 6$ '. The state space consists of all such coordinates that are not obstacles.
2. **Actions (A):** The set of 6 possible actions the agent can take, corresponding to movement along the three axes:
 $A = \{+x \text{ (East)}, -x \text{ (West)}, +y \text{ (North)}, -y \text{ (South)}, +z \text{ (Up)}, -z \text{ (Down)}\}$
3. **Rewards (R):** The reward function ' $R(s, a, s')$ ' defines the reward received after transitioning from state ' s ' to ' s' ' by taking action ' a '.
Goal: +50 (entering the goal at (5,5,5)).
Pit: -50 (entering the pit at (2,2,2)).
Step Cost: -1 for any other valid move.
4. **Discount Factor (γ):** The discount factor is given as ' $\gamma = 0.95$ '. This prioritizes more immediate rewards over distant ones.
5. **Transition Probabilities (P):** The transition function ' $P(s' | s, a)$ ' defines the probability of landing in state ' s' ' after taking action ' a ' in state ' s '. Let the intended direction of movement for action ' a ' be ' $d_{intended}$ '.
 - a. with probability ' p ' (e.g., ' $p=0.8$ '), the agent moves one step in the intended direction ' $d_{intended}$ '.
 - b. With probability ' $1-p$ ', the agent "slips" and moves one step in one of the 4 perpendicular directions. Each perpendicular direction has a uniform probability of ' $(1-p)/4$ '.
 - c. If a move (intended or slip) would result in hitting a boundary or an obstacle, the agent remains in its current state ' s '.

- Part C & D: Q-learning, Policy Evaluation & Comparison

We implemented the tabular Q-learning algorithm with an ϵ -greedy exploration strategy. The agent was trained for 15,000 episodes.

1. Learning Curve:

The learning curve below shows the total reward obtained per episode. It demonstrates that the agent's performance improves over time and converges, indicating that it has successfully learned a policy to reach the goal.



2. Policy Evaluation:

After training, we extracted the learned greedy policy. We evaluated this policy over 100 test episodes (with no exploration) and compared its average return against a baseline random policy.

- Learned Policy Average Return: 40.80
- Random Policy Average Return: -204.20

The significantly higher average return of the learned policy confirms its effectiveness and superiority over random wandering.

- Part E: Experiments & Analysis

We conducted experiments by varying the discount factor (γ) and slip probability.

1. Varying Gamma (γ):

- A lower γ (e.g., 0.8) makes the agent more "short-sighted," potentially choosing paths that are quicker but riskier.

- b. A higher γ (e.g., 0.99) makes the agent more "far-sighted," valuing future rewards more highly and finding a more globally optimal path. Convergence was generally stable across these values.

2. Varying Slip Probability ($1-p$):

- a. Low Slip (e.g., 0.05): The environment is more deterministic. The agent learns a very direct and optimal path quickly.
- b. High Slip (e.g., 0.5): The environment is highly stochastic. The learned policy becomes more conservative, avoiding states near obstacles or the pit, as a slip could be catastrophic. The agent takes longer to converge and the final policy is less direct.

- Part F: Visualization

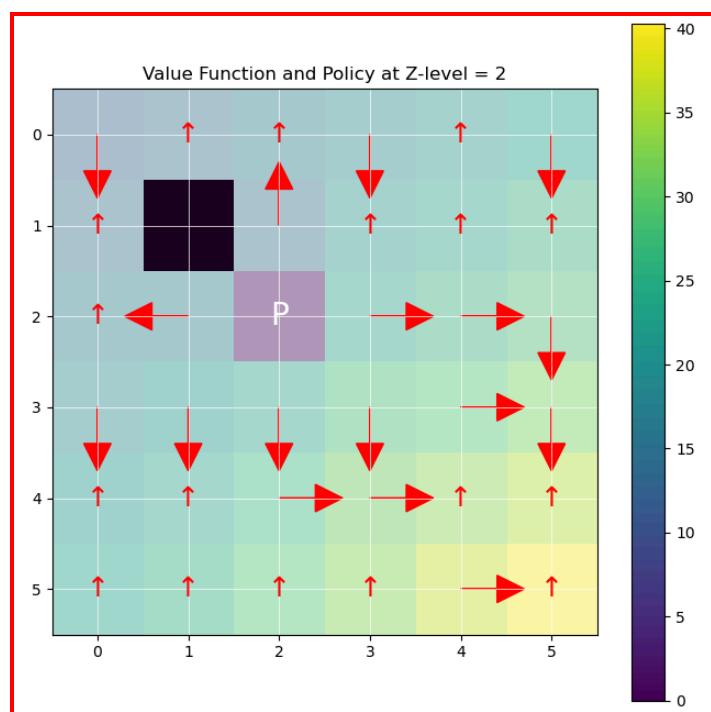
The following heatmaps visualize the learned value function (`max_a Q(s,a)`) for three different z-levels (slices) of the 3D grid. The arrows represent the learned greedy policy at each state.

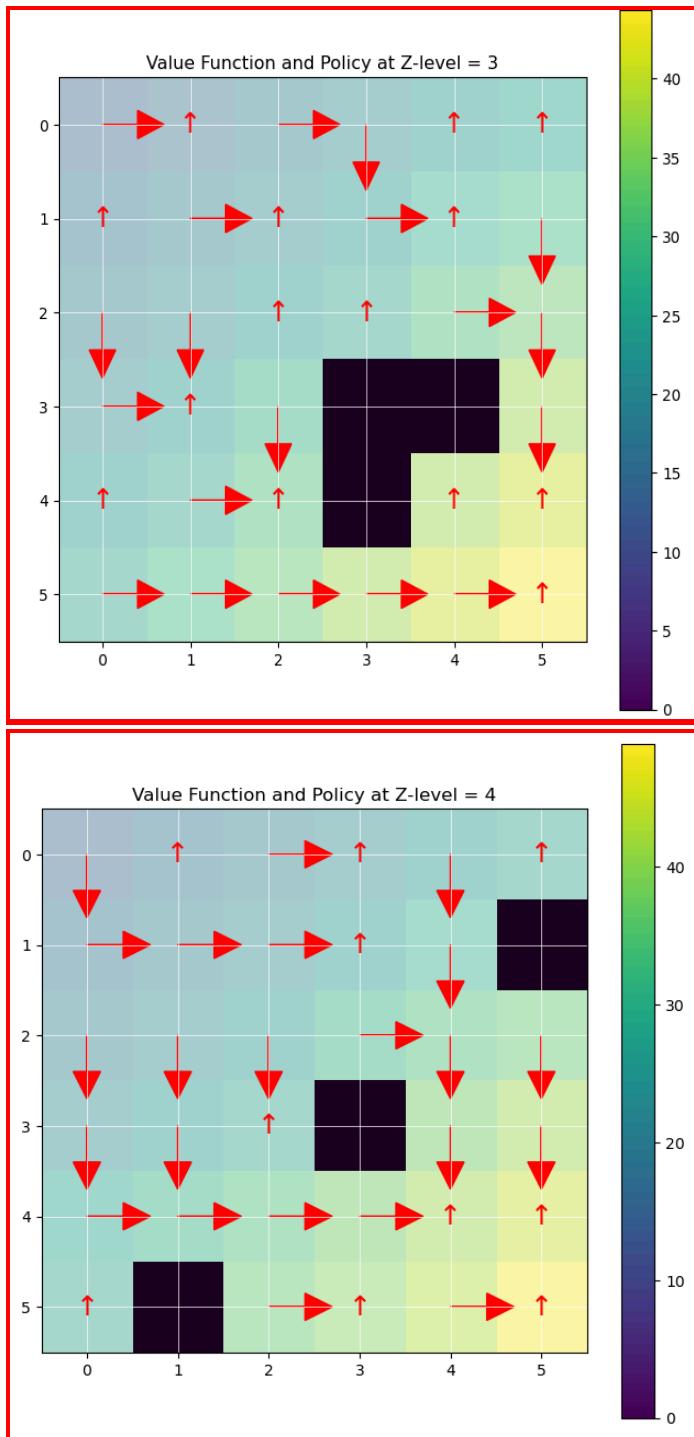
Brighter cells indicate higher value (i.e., it's good to be in those states).

Darker cells indicate lower value.

Obstacles are shown in black.

The Goal (G) and Pit (P) are marked.





The visualizations clearly show that the value is highest near the goal state (5,5,5) and radiates outwards. The policy arrows consistently point in directions that lead towards states with higher values, effectively showing a path from any free cell towards the goal while avoiding the pit and obstacles. The value of states near the pit is very low, and the policy arrows direct the agent away from it, which is the desired behavior.