CSCI/ROBO 7000/4830:
Deep Reinforcement Learning and Robotics
Fall 2025

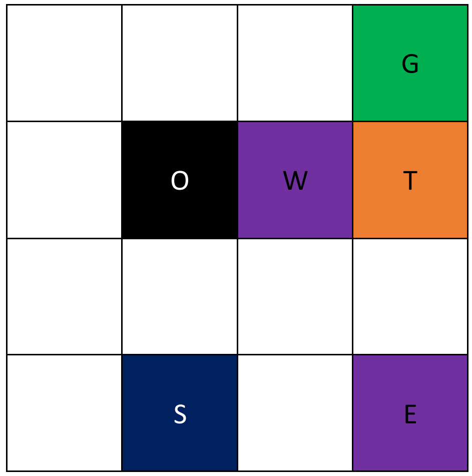
Homework #1: The Wormhole Grid

Introduction

This assignment builds on the concepts of policies, value functions, and the Bellman equations within the framework of a Markov Decision Process (MDP). You will analyze a modified grid world that introduces more complex dynamics, requiring you to not only calculate optimal policies but also reason about the trade-offs involved in an agent's decision-making process.

Problem Description

We will use a 4x4 grid world with a few new features. The agent can move in the four cardinal directions (North, South, East, West).



The grid contains the following cells:

- S (Start): The agent begins each episode in this blue cell, located at (3, 1).
- G (Goal): The green cell at (0, 3). Reaching this cell gives a large positive reward and terminates the episode.
- T (Trap): The tangerine-colored cell at (1, 3). Entering this cell gives a large negative reward and terminates the episode.
- O (Obstacle): The black cell at (1, 1). The agent cannot enter this cell.
- W (Wormhole Entry): The purple cell at (1, 2). Entering this cell instantly teleports the agent to the Wormhole Exit.
- E (Wormhole Exit): The purple cell at (3, 3). This is the destination of the wormhole.

Dynamics & Rewards: The world is fully **deterministic**. An action to move in a certain direction successfully moves the agent one step in that direction.

If the agent attempts an illegal move (into a wall or an obstacle), it remains in its current state.

Unless otherwise specified, the rewards are:

- Entering the Goal (G): R=+50
- Entering the Trap (T): R=−50
- For any other transition (into white, S, W, or E squares): R=w=−1 per step.

The discount factor is γ=0.9 unless otherwise specified.

The value of a state s, denoted V(s), is the total cumulative discounted reward an agent expects to receive starting from state s and following a specific policy π. The optimal value $V_*(s)$ is the maximum possible value achievable by any policy.

PART 1: POLICY AND VALUE ANALYSIS [40 POINTS]

In this section, we will analyze and compare a simple, pre-defined policy with the optimal policy.

1. POLICY EVALUATION [15/40 POINTS]

Consider the following simple, "go-up-and-right" policy, π_{simple}: in every state, the agent attempts to move North. If North is blocked, it tries to move East. If both are blocked, it moves South.

A) Calculate the state-value function, $V_{\pi_{simple}}(s)$, for this policy for all states. Fill in your values on a 4x4 grid.

B) Show the setup for your calculations for at least two non-terminal states.

Hint: This is a policy evaluation problem! You can solve the system of Bellman expectation equations.

2. OPTIMAL VALUE AND POLICY [15/40 POINTS]

Now, find the optimal state-value function $V_*(s)$, and the corresponding optimal policy, value $\pi_*(a|s)$, for this grid world using the default rewards and discount factor (γ=0.9, w=−1).

- Fill in the optimal values $V_*(s)$ for all states in one grid.
- Draw the optimal policy $\pi_*(a|s)$ (the arrows indicating the best action(s)) in a separate grid.

Hint: This is a control problem! You can use Value Iteration.

3. JUSTIFICATION [10/40 POINTS]

Find at least one state where π_{simple} πsimple and π_* disagree. Using the values you calculated in Q1 and Q2, provide a brief written explanation for why the optimal policy is superior in that state. Your justification must reference the Bellman Optimality Equation and compare the expected returns of the different actions.

PART 2: REWARD ENGINEERING [30 POINTS]

A core challenge in RL is designing a reward function that elicits the desired behavior. Here, you will act as the reward engineer. Your goal is to find a reward for the white squares, w, that makes the agent explicitly avoid the wormhole.

4. Designing for Risk Aversion [30/30 Points]

Your task is to find the range of values for the step reward w such that the optimal policy will never choose to enter the wormhole W from any adjacent state. All other rewards (G=50, T=−50) and the discount factor (γ=0.9) remain the same.

- Identify the state(s) from which an agent could choose to enter the wormhole.
- For one of these states, write down the Bellman Optimality Equation that defines its value, $V_*(s)$. This equation should contain w as a variable.
- Set up an inequality where the value of taking the action leading to the wormhole is less than the value of taking the best alternative action(s).
- Solve this inequality to find the range of values for w that guarantees the wormhole is never used. Provide a clear justification for your steps.

Part 3: The Role of the Discount Factor [30 Points]

The discount factor γ determines how "patient" an agent is. Let's explore its impact. For this section, reset the step reward to $w = -1$.

5. Impatient vs. Patient Agents [30/30 Points]

- Calculate the optimal policy π_* for the grid world under two different scenarios:
- Scenario A (Impatient Agent): $\gamma = 0.5$
- Scenario B (Patient Agent): $\gamma = 0.99$
- Draw the resulting optimal policy for each scenario in a separate grid.
- Is the optimal policy different in these two scenarios? Provide a concise explanation for any observed differences (or lack thereof). Your explanation should focus on how the discount factor influences the agent's evaluation of short-term vs. long-term rewards, especially concerning the trade-off between a longer path and using the wormhole.