IIT Bombay WiDS Project: Reinforcement Learning

Final Summary Report

Sanidhya

February 1, 2026

1 Introduction

This report summarizes the four-week Reinforcement Learning (RL) project completed under the Winter in Data Science (WiDS) program at IIT Bombay. The project progressed from basic decision-making under uncertainty to complex sequential problem-solving.

2 Week 1: Multi-Armed Bandits

The objective was to explore the exploration-exploitation trade-off.

- **Methods:** Implemented ϵ-greedy, UCB, and Optimistic Initial Values.
- **Findings:** Purely greedy agents often converge on sub-optimal actions. ϵ-greedy ($\epsilon = 0.1$) maintains a balance that ensures the optimal action is eventually identified.

3 Week 2 & 3: MDPs and Dynamic Programming

We transitioned to full RL environments using Markov Decision Processes.

- **Formulation:** Modeled environments using the tuple (S, A, P, R, γ).
- **Algorithms:** Implemented **Policy Iteration** for Jack's Car Rental and **Value Iteration** for the Gambler's Problem.
- **Mathematical Foundation:** Solved the Bellman Equation:

$$V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$$

4 Week 4: Final Project - 15-Puzzle Solver

The capstone project involved solving the 15-Puzzle using RL techniques.

- **Problem Space:** A combinatorial puzzle with over 10 trillion possible states.
- **Implementation:** Developed an agent that learns the value of tile configurations. The reward structure penalized each move (-1) to encourage the shortest path to the goal state.
- **Success:** The agent successfully learns to navigate the state space, mimicking or exceeding human-level heuristic performance.

5 Conclusion

The project provided a comprehensive foundation in RL, moving from tabular methods to large-scale state space problems. This experience has equipped me with the skills to model complex systems as MDPs and apply iterative algorithms to find optimal strategies.