Reinforcement Learning Platform - Technical Report

Project: RL Algorithms Visualizer

Version 1.0

Adham Ashraf 202200953

December 2025

Executive Summary

This platform implements a comprehensive suite of reinforcement learning (RL) algorithms with interactive visualization capabilities. The system supports 7 distinct RL algorithms across 8 different environments, featuring advanced parameter tuning and intelligent visualization for both small and large state spaces.

Contents

1 Implemented Algorithms

1.1 Dynamic Programming (DP)

Value Iteration: Iteratively updates state values using the Bellman optimality equation: $V(s) \leftarrow \max_a \sum P(s'|s,a)[R(s,a,s') + \gamma V(s')]$.
Policy Iteration: Alternates between Evaluation (V^π) and Improvement until the policy π is stable.

1.2 Monte Carlo (MC) Methods

MC Control: Learns from complete episodes. $Q(s,a) \leftarrow Q(s,a) + \alpha[G - Q(s,a)]$. Supports First-Visit and Every-Visit variants.

1.3 Temporal Difference (TD) Learning

Q-Learning: Off-policy learning. $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$.
SARSA: On-policy learning. $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$.
TD(0): Updates state values $V(s)$.
n-step TD: Bridges MC and TD(0) using an n-step lookahead return.

2 Implemented Environments

2.1 Grid-Based Environments

- **GridWorld:** size × size grid. Reward: Goal (+1.0), Step (-0.01).
- **Maze:** Procedural generation with wall density settings ('easy', 'medium', 'hard').

2.2 Classic Control (Gymnasium)

- **FrozenLake:** Stochastic (33% slip chance). High difficulty.
- **CliffWalking:** Deterministic. Reward: Goal (0), Cliff (-100), Step (-1).
- **CartPole:** Discretized 4D continuous space ($\sim 10,000$ states).
- **Mountain Car:**
  - **State Space:** Discretized 2D (Position: $[-1.2, 0.6]$, Velocity: $[-0.07, 0.07]$).
  - **Reward:** -1.0 for every step until the goal is reached.
  - **Goal:** Reach the flag at position 0.5.

2.3 Game Environments

- **TicTacToe:** Discrete space (3^9 states). Play against a random opponent.
- **Blackjack:**
  - **State Space:** Player sum (12-21), Dealer card (1-10), Usable Ace (0-1).
  - **Reward:** Win (+1), Loss (-1), Draw (0).

3 Parameter Adjustment Capabilities

3.1 Universal Parameters

- **Discount Factor (γ):** Range $[0.0, 1.0]$. Default: 0.99. Controls weight of future rewards.
- **Convergence Threshold (θ):** Range $[0.0001, 0.1]$. Default: 0.001. Used in DP for stopping criteria.

3.2 Learning Parameters (Model-Free)

- **Learning Rate (α):** Range $[0.001, 1.0]$. Default: 0.1. Small α is stable; large α is fast.
- **Exploration Rate (ϵ):** Range $[0.0, 1.0]$. Default: 0.1. Probability of taking a random action. Recommendation: Start high and decay.

3.3 Training Parameters

- **Number of Episodes:** Range $[10, 10000]$. Recommended: 1000 for GridWorld, 5000+ for FrozenLake/Blackjack.
- **Max Iterations (DP):** Range $[10, 1000]$. Default: 100.
- **n-steps (n-step TD):** Range $[1, 20]$. Default: 3. $n = 1$ is TD(0), $n = \infty$ is MC.

3.4 Algorithm-Specific Parameters

- **Monte Carlo Type:** 'FV' (First-Visit) vs 'EV' (Every-Visit).
- **Fixed Alpha:** If `False`, uses $1/N(s)$ averaging; if `True`, uses a constant step size.

4 Visualization Techniques

The platform differentiates between small and large state spaces (Threshold: 100 states).

- **Small Spaces:** Annotated Heatmaps (RdYlGn), Arrow Grids for policies, and Multi-Action Q-Value grids.
- **Large Spaces:** 6-Panel Statistical Dashboards including Q-value distributions (histograms), action frequency bar charts, and sampled state-action pair rankings.
- **Metrics:** Moving average reward plots, episode length tracking, and ϵ-decay visualization.

5 Performance Characteristics

Table 1: Training Time and Complexity Benchmarks

Environment	**State Space**	**Typical Episodes**	**Training Time**
GridWorld 4x4	16	1000	$\sim$ 5s
FrozenLake	16	5000	$\sim$ 15s
CliffWalking	48	1000	$\sim$ 7s
Blackjack	704	5000	$\sim$ 12s
Mountain Car	$\sim$ 2500	2000	$\sim$ 18s
CartPole	$\sim$ 10000	1000	$\sim$ 15s
TicTacToe	19683	2000	$\sim$ 20s

6 Conclusion

This platform provides a production-ready environment for experimenting with tabular Reinforcement Learning. With support for complex tasks like **Mountain Car** and high-dimensional spaces like **TicTacToe**, users can observe how hyperparameter tuning $(\alpha, \gamma, \epsilon)$ directly impacts convergence and policy stability.