Fictitious Play and Reinforcement Learning for Computing Equilibria in Repeated Zero-Sum Games

Ioannis Kasionis Ioannis Koutsoukis

February 17, 2025

Abstract

This report presents a comparative study of learning agents in two-person zero-sum games. We explore both Fictitious Play (FP) and Reinforcement Learning (RL) methods, including Q-Learning, Minimax RL, and a Belief-Based approach. The experiments are carried out in two distinct games: a stochastic Rock-Paper-Scissors game and Matching Pennies. The results are analyzed and visualized, highlighting the performance, convergence behavior, and policy evolution of the agents.

1 Introduction

Two-person zero-sum games provide a rich testbed for studying competitive behaviors and learning algorithms. This report investigates various learning approaches, including Fictitious Play (FP) and Reinforcement Learning (RL) techniques, applied to two well-known games: the stochastic Rock-Paper-Scissors (RPS) and Matching Pennies (MP). The goal is to compare the theoretical underpinnings, practical implementation, and experimental performance of each algorithm.

2 Theoretical Background

2.1 Fictitious Play (FP)

Fictitious Play is an iterative algorithm where each player assumes that their opponent is playing a stationary strategy determined by the historical frequency of actions. At every iteration, a player best-responds to the empirical distribution of the opponent's past actions. FP has been shown to converge to Nash equilibrium in certain classes of games, although convergence is not guaranteed in all scenarios.

2.2 Reinforcement Learning (RL)

Reinforcement Learning (RL) refers to a class of algorithms where agents learn to optimize their actions through trial and error, guided by reward signals. In the context of zero-sum games, RL algorithms aim to discover optimal policies that maximize expected rewards.

2.2.1 Q-Learning

Q-Learning is an off-policy RL algorithm that learns a value function $Q(s, a)$ representing the expected cumulative reward for taking action a in state s and following the optimal policy thereafter.

The update rule is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

where α is the learning rate and γ is the discount factor.

2.2.2 Minimax RL

Minimax RL extends traditional Q-Learning to competitive settings by solving a minimax problem at each stage. In zero-sum games, each agent computes a mixed strategy by solving a linear program, ensuring robust performance against an adversarial opponent.

2.2.3 Belief-Based Learning

The Belief-Based approach assumes that agents have prior knowledge of the payoff matrix. They update their beliefs about the opponent's strategy based on observed actions and select actions that maximize their expected payoff accordingly.

3 Implementation

The algorithms were implemented in Python using a modular design. Two separate scripts were developed for the games:

- **rock_paper_scissors.py**: Implements a stochastic Rock-Paper-Scissors game with two states and four agent types (FP, Q-Learning, Minimax RL, and Belief-Based).
- **matching_pennies.py**: Implements a non-stochastic Matching Pennies game with similar agent setups.

Each script includes:

- **Agent Classes**: Separate classes encapsulate the behavior of each learning algorithm.
- **Environment Dynamics**: The games are modeled with state transition rules (stochastic for RPS and fixed for MP) and payoff matrices.
- **Simulation Function**: A simulation loop that runs multiple episodes and trials, recording rewards, cumulative scores, policy evolution, Q-value updates, and joint action frequencies.
- **Plotting and Data Export**: The scripts generate plots (e.g., moving average rewards, cumulative scores, state evolution, joint action frequency heatmaps) and export processed data to CSV files.

Additionally, a dashboard was built using `Dash` and `Plotly` to interactively visualize and compare experimental results.

4 Results & Discussion

This section presents our experimental findings for two zero-sum games: Rock–Paper–Scissors (RPS) and Matching Pennies (MP). We analyze each pairwise agent matchup with respect to policy evolution, joint action frequencies, Q-value convergence, and cumulative scores. Figures throughout this section illustrate how different learning algorithms adapt and either converge to equilibrium play or systematically exploit an opponent.

4.1 Rock–Paper–Scissors (RPS)

Q-Value Convergence. Rock–Paper–Scissors is a zero-sum game with a well-known mixed-strategy Nash equilibrium (each action with probability 1/3). Figures 1 and 2 compare the norm difference between successive Q-tables for Minimax RL and Q-Learning, respectively. Minimax RL quickly drives the difference to near-zero, indicating stable Q-values. Q-Learning, however, shows persistent fluctuations (a higher, noisy baseline), suggesting it is continually adapting to a non-stationary opponent rather than converging to a fixed solution.

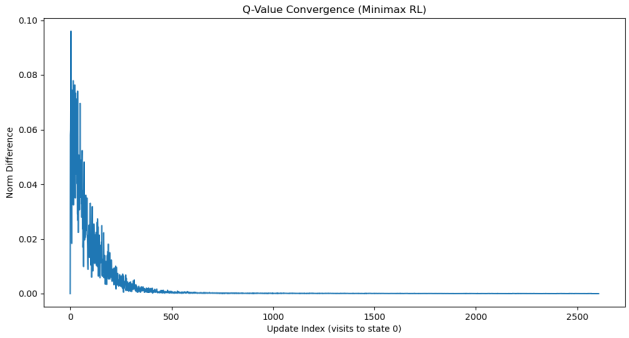

Figure 1: Q-Value Convergence for Minimax RL in RPS. The norm difference quickly drops to near-zero, reflecting fast convergence to an approximate equilibrium policy.

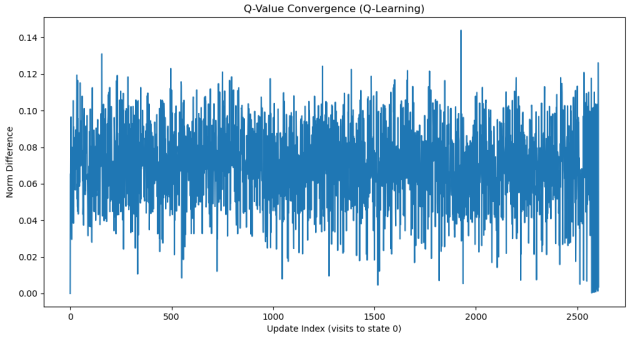

Figure 2: Q-Value Convergence for Q-Learning in RPS. The norm difference remains noisy and never fully settles, as Q-Learning continuously chases the opponent's changing strategy.

Policy Evolution and Joint Action Frequencies. Figure 3 shows how a Fictitious Play (FP) agent's action probabilities evolve over time. In many trials, FP ends up near the 1/3-1/3-1/3 distribution, consistent with the RPS equilibrium. Joint action frequency heatmaps (e.g., Figure 4) reveal that, for well-adapting agents, each of the nine possible (Rock, Paper, Scissors) combinations occurs with probability close to $1/9 \approx 0.11$. Small deviations reflect finite sample effects, exploration, or local biases.

Moving averages & Cumulative Scores. Figure 5 shows sample cumulative score traces for different matchups. Some pairs (e.g., Minimax RL vs. Fictitious Play) hover around zero or oscillate,

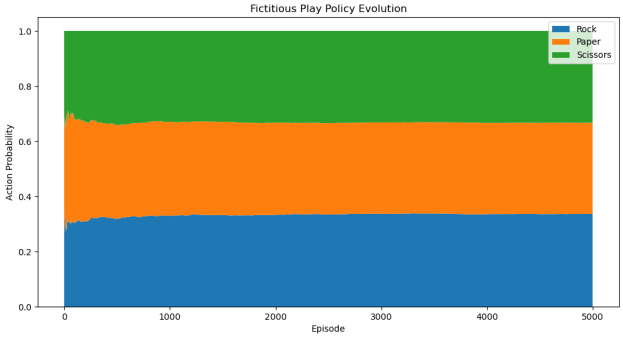

Figure 3: Fictitious Play policy evolution in RPS, showing action probabilities for Rock, Paper, and Scissors. The agent often converges near an even mix (one-third each).

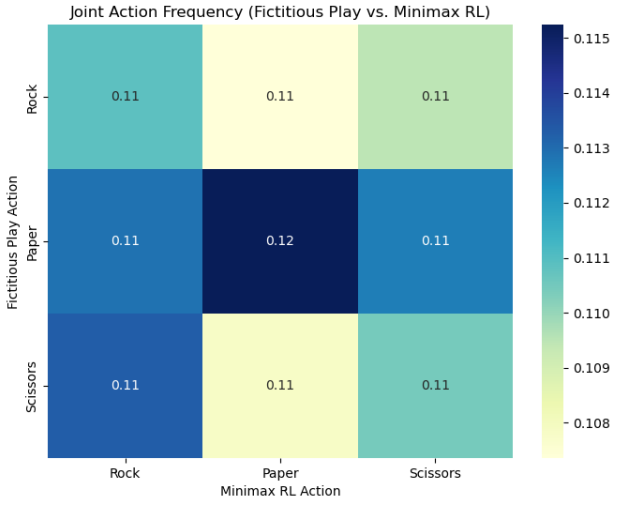

Figure 4: Joint Action Frequency for a pairwise matchup in RPS. Each cell indicates how often (Action_A, Action_B) occurs. Probabilities around 0.11 per cell suggest near-uniform randomization.

indicating near-equilibrium play. Others (e.g., Q-Learning vs. Belief-Based) exhibit diverging lines, meaning one agent exploits the other over time. Such divergences imply that one agent discovered a predictable bias in the opponent's actions and adapted to exploit it Figure 6.

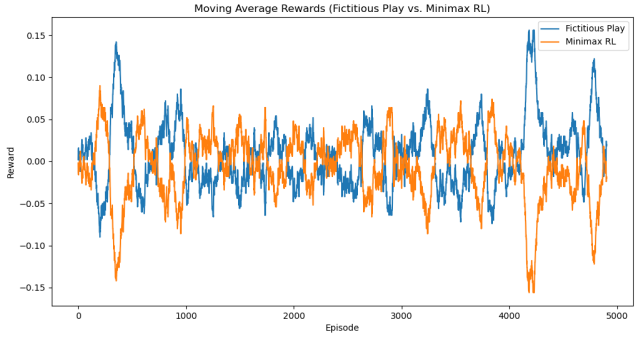

Figure 5: Representative moving average score in RPS. When both agents learn robust mixed strategies, scores fluctuate near zero. Monotonic divergence implies one agent systematically exploits the other.

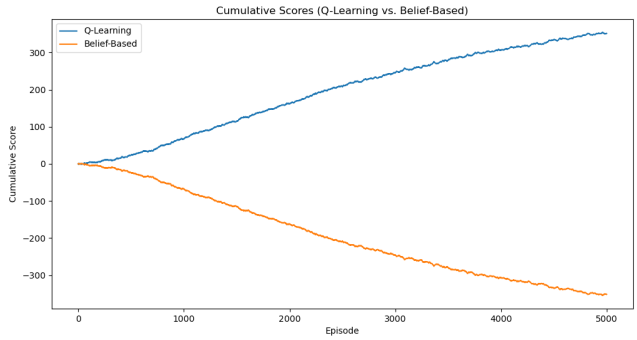

Figure 6: Representative cumulative score evolution in RPS. When one agent exploits the other over time. Such divergences imply that one agent discovered a predictable bias in the opponent's actions and adapted to exploit it.

4.2 Matching Pennies (MP)

Matching Pennies is another strictly competitive, zero-sum game but with only two actions: Heads or Tails. The unique Nash equilibrium requires each player to randomize 50–50.

Q-Value Convergence. Figures 7 and 8 again contrast Minimax RL vs. Q-Learning. Minimax RL converges quickly to stable Q-values, while Q-Learning shows persistent norm-difference oscillations. As in RPS, Q-Learning's inability to account for a non-stationary opponent prevents it from "locking in" a stable solution.

Policy Evolution and Joint Frequencies. Figures like 9 show how Fictitious Play in Matching Pennies often approaches near 50–50 mixing if the opponent also mixes effectively. However, if an

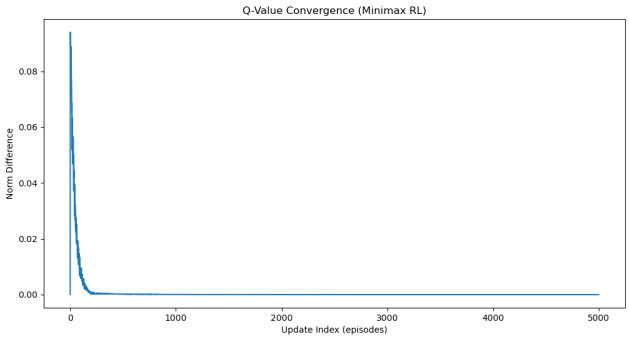

Figure 7: Q-Value Convergence for Minimax RL in Matching Pennies. The agent rapidly converges to an approximate 50–50 strategy.

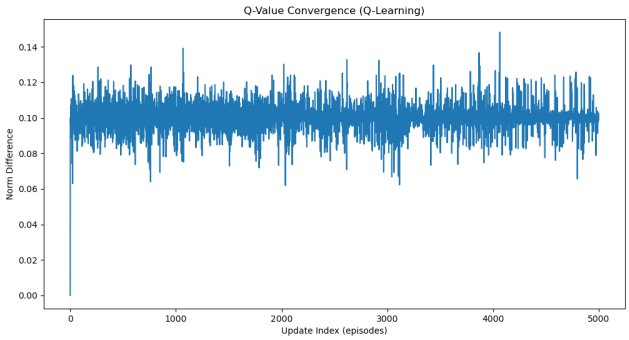

Figure 8: Q-Value Convergence for Q-Learning in Matching Pennies. Norm differences remain higher and noisy, indicating ongoing adaptation to the opponent's shifting policy.

opponent remains predictably biased (e.g. Belief-Based with slow adaptation), Fictitious Play can lock in a counter-bias, leading to a skewed distribution. Joint action frequency heatmaps, such as Figure 10, reveal whether agents are truly randomizing (near 0.25 per cell in a 2×2 matchup) or getting stuck in more deterministic patterns.

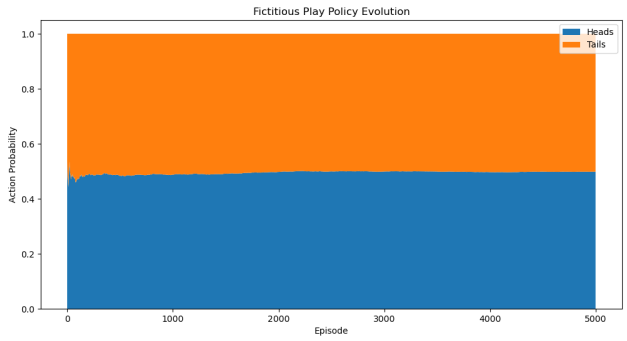

Figure 9: Fictitious Play policy evolution in Matching Pennies. Equilibrium demands 50% Heads and 50% Tails, but slight off-equilibrium biases can persist if the opponent remains predictable.

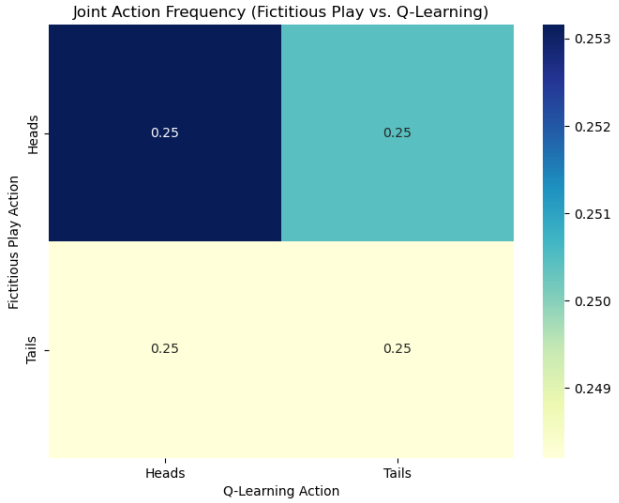

Figure 10: Joint Action Frequency in a 2×2 (Heads/Tails) matchup. Ideal equilibrium mixing would yield 0.25 in each cell. Large deviations indicate that one or both agents are systematically favoring certain actions.

Moving Averages & Cumulative Scores. Figure 11 and Figure 12 shows sample cumulative score traces in Matching Pennies. Because the game is strictly zero-sum, perfectly randomizing players would average zero. However, if one agent fails to correct a predictable pattern, the other agent's cumulative score diverges positively while the former's drops. For instance, Q-Learning often exploits slow-adapting Belief-Based (resulting in a large positive slope), whereas Minimax RL vs. an adaptive agent might hover around zero or show moderate oscillations.

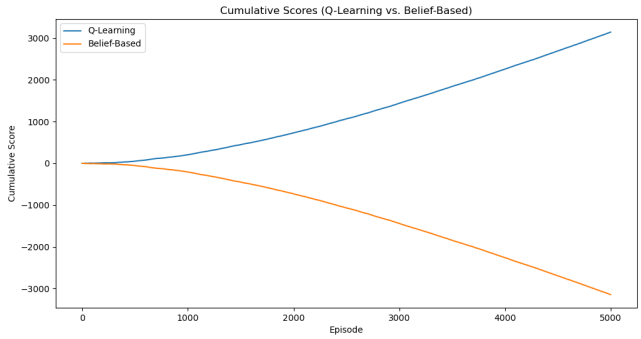

Figure 11: Cumulative scores in Matching Pennies. Large, monotonic separations imply one agent exploits the other's biased play. Near zero or oscillatory outcomes suggest both agents approximate the 50–50 equilibrium.

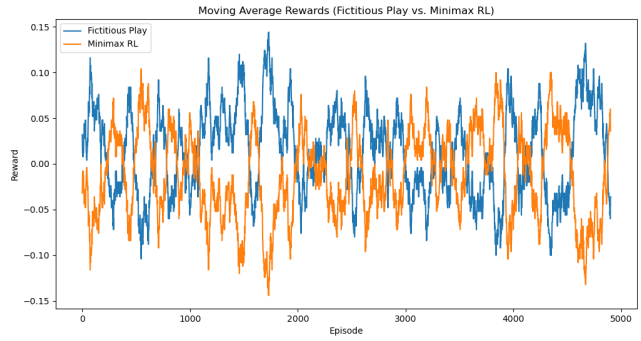

Figure 12: Minimax RL vs. Fictitious Play hover around zero.

4.3 Overall Observations

- **Minimax RL** leverages the zero-sum nature of these games, converging quickly to near-equilibrium policies.
- **Q-Learning** exhibits persistent variability because it views the opponent as part of a non-stationary environment, preventing stable convergence in many runs.
- **Fictitious Play** and **Belief-Based** can do well if the opponent is sufficiently predictable; otherwise, they may remain off-equilibrium and be exploited by more adaptive strategies.
- When both agents effectively randomize near equilibrium (1/3 each in RPS or 1/2 each in MP), *expected* cumulative scores hover around zero.
- Large divergences in cumulative scores generally mean one agent discovered and exploited the other's systematic bias.

In summary, these plots confirm that specialized approaches like Minimax RL quickly home in on robust mixed strategies in strictly competitive games, while Q-Learning's standard update rule struggles with the non-stationarity introduced by another learning opponent. Agents like Fictitious Play or Belief-Based may do quite well against certain opponents but can be exploited if they fail to adapt to changing opponent distributions.