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

Ioannis Kasionis Ioannis Koutsoukis

February 14, 2025

1 Introduction

The computation of equilibria in games is a central problem in game theory and multi-agent systems. Repeated zero-sum games, where the gain of one player is exactly balanced by the loss of the other, provide an ideal test-bed for studying learning dynamics in adversarial settings. In this work, we explore iterative learning algorithms that allow players to converge toward equilibrium strategies without requiring complete analytical solutions. We focus on comparing Fictitious Play (FP) with reinforcement learning (RL) methods—including Q-Learning, Minimax RL, and Belief-Based approaches—in two well-known games: Rock-Paper-Scissors and Matching Pennies.

2 Theoretical Background

2.1 Repeated Zero-Sum Games

Repeated zero-sum games consist of a stage game that is played over several episodes. In each round, both players select actions simultaneously. The payoff for one player is the negative of the other player's payoff, ensuring that the total payoff sums to zero. The repetition of the game allows players to learn from past outcomes, adapt their strategies, and potentially converge to a Nash equilibrium or a minimax solution. These games serve as a foundational model for adversarial interactions in economics, security, and machine learning.

2.2 Learning Agents

Several learning algorithms have been proposed to compute equilibria in repeated games:

- **Fictitious Play (FP):** Players assume that opponents will play according to the empirical frequency of their past actions. They then choose the best response to these beliefs.
- **Q-Learning (QL):** An RL method where agents learn the expected utility of actions using a Q–table and update it iteratively via temporal difference learning.
- **Minimax RL:** A variation of Q–Learning adapted for adversarial settings, where agents aim to maximize the minimum gain against a worst-case opponent.
- **Belief-Based Methods:** Agents update probabilistic beliefs about the opponent's actions and choose their best response accordingly.

3 Implementation

3.1 Environment Setup

Two game environments were implemented in Python:

- **Stochastic Rock-Paper-Scissors (RPS):** In this environment, the game is played in two states with different payoff matrices. A state transition function introduces stochasticity, altering the stage game dynamically.
- **Matching Pennies (MP):** A non-stochastic repeated zero-sum game with a fixed payoff matrix, where the payoff for one player is the negative of the other.

Each environment is encapsulated within a Python class that defines the state, payoff matrices, and state transition rules.

3.2 Experimental Design

Experiments were conducted by pitting every pair of the four agents (FP, QL, Minimax RL, Belief-Based) against each other in both game environments. For each experiment, multiple trials were executed, each spanning thousands of episodes. During these simulations, various performance metrics were recorded, including:

- Moving average rewards.
- Cumulative scores over episodes.
- Q–value evolution and convergence.
- Policy evolution and learning stability.
- Joint action frequency.

The simulation results were exported to CSV files and subsequently visualized using an interactive Python Dash dashboard.

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.

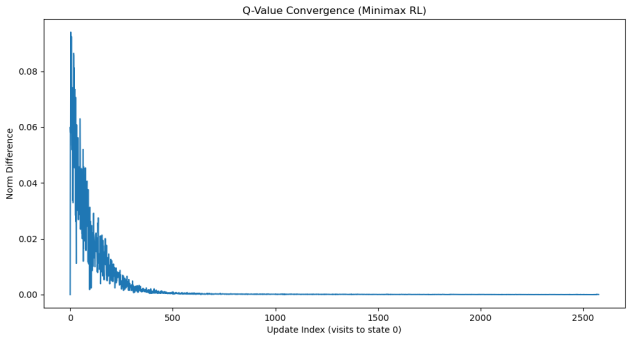

Figure 1: Q-Value Convergence for Minimax RL in RPS. The norm difference quickly drops to near-zero, reflecting fast convergence to an approximate equilibrium policy.

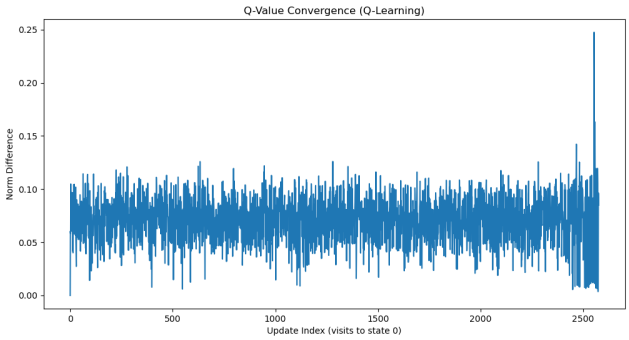

Figure 2: Q-Value Convergence for Q-Learning in RPS. The norm difference remains noisy and never fully settles, as Q-Learning continuously chases the opponent's changing strategy.

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.

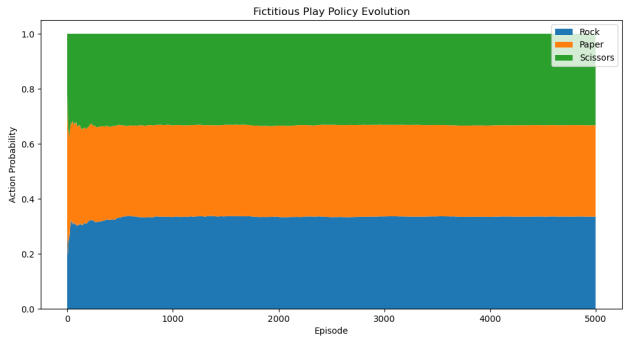

Figure 3: Fictitious Play policy evolution in RPS, showing action probabilities for Rock, Paper, and Scissors. The agent often converges near an even mix (one-third each).

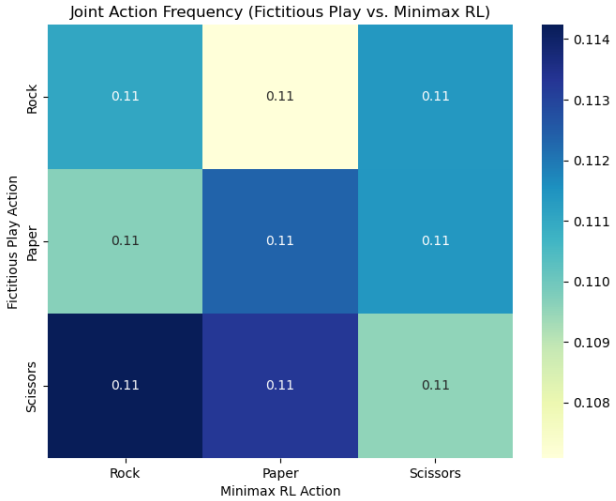

Figure 4: Joint Action Frequency for a pairwise matchup in RPS. Each cell indicates how often (Action_A, Action_B) occurs. Probabilities around 0.11 per cell suggest near-uniform randomization.

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, 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.

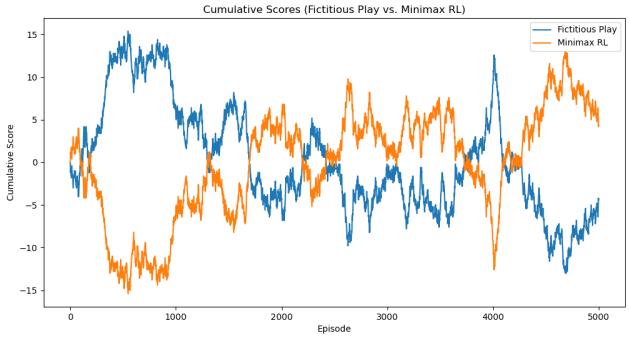

Figure 5: Representative cumulative score evolution in RPS. When both agents learn robust mixed strategies, scores fluctuate near zero. Monotonic divergence implies one agent systematically exploits the other.

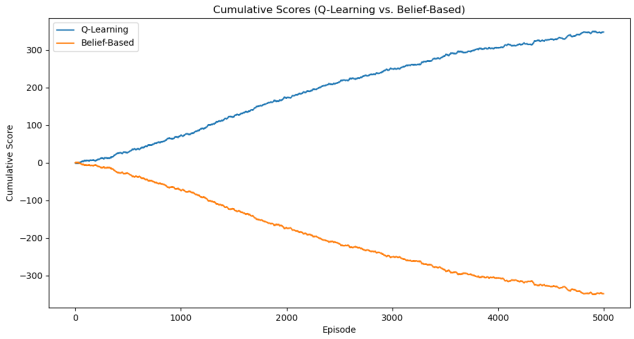

Figure 6: Representative cumulative score evolution in RPS. When one agent exploits the other over time. Such divergences imply that one agent discovered a predictable bias in the opponent's actions and adapted to exploit it.

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.

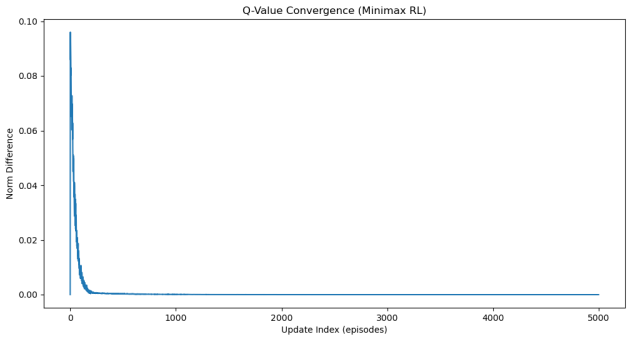

Figure 7: Q-Value Convergence for Minimax RL in Matching Pennies. The agent rapidly converges to an approximate 50–50 strategy.

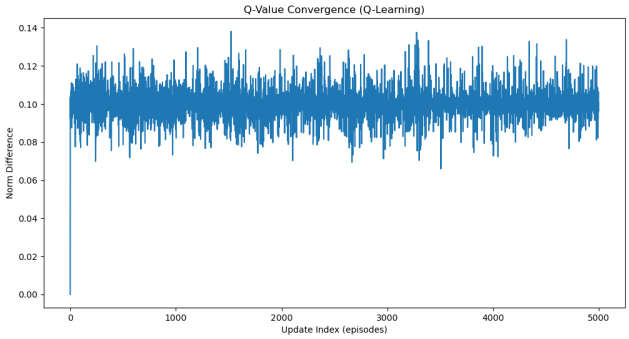

Figure 8: Q-Value Convergence for Q-Learning in Matching Pennies. Norm differences remain higher and noisy, indicating ongoing adaptation to the opponent's shifting policy.

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 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.

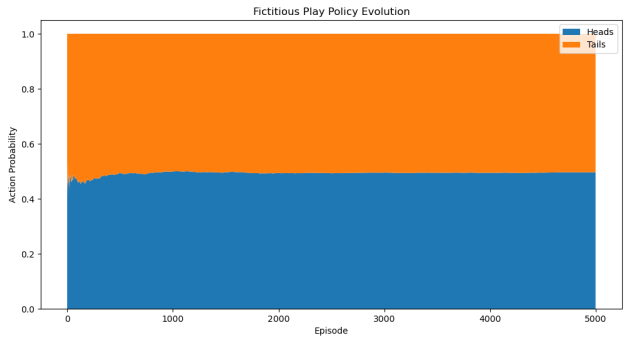

Figure 9: Fictitious Play policy evolution in Matching Pennies. Equilibrium demands 50% Heads and 50% Tails, but slight off-equilibrium biases can persist if the opponent remains predictable.

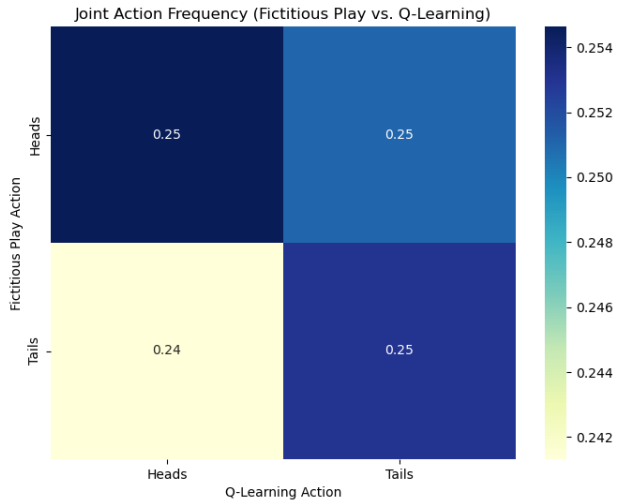

Figure 10: Joint Action Frequency in a 2×2 (Heads/Tails) matchup. Ideal equilibrium mixing would yield 0.25 in each cell. Large deviations indicate that one or both agents are systematically favoring certain actions.

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.

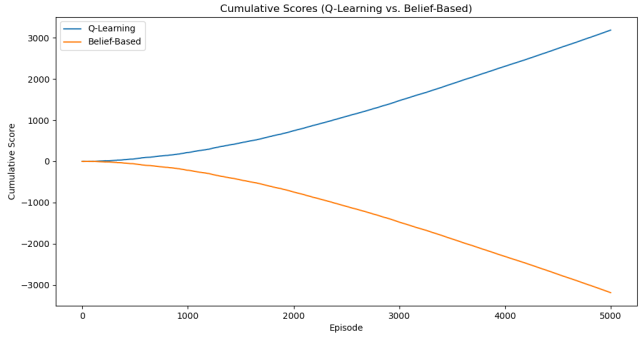

Figure 11: Cumulative scores in Matching Pennies. Large, monotonic separations imply one agent exploits the other's biased play. Near zero or oscillatory outcomes suggest both agents approximate the 50–50 equilibrium.

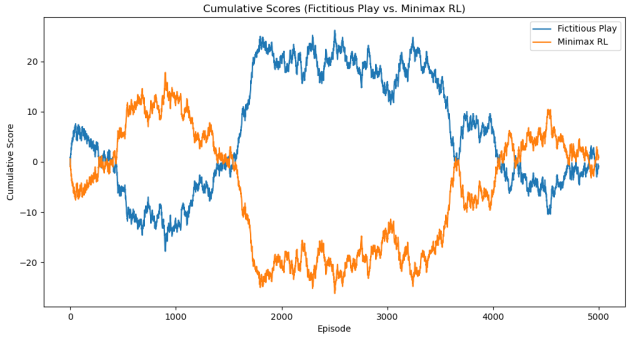

Figure 12: Minimax RL vs. Fictitious Play hover around zero.

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.