

# Comparing Fictitious Play and Q-Learning in Stochastic Zero-Sum Games

Ioannis Kasionis      Ioannis Koutsoukis

January 31, 2025

## Contents

<b>1</b>	<b>Theoretical Background</b>	<b>2</b>
1.1	Repeated & Zero-Sum Stochastic Games . . . . .	2
1.2	Fictitious Play (FP) . . . . .	2
1.3	Reinforcement Learning (Q-Learning) . . . . .	2
1.3.1	Q-Learning with $\epsilon$ -Decay . . . . .	2
<b>2</b>	<b>Modified Games Implementation</b>	<b>2</b>
2.1	Stochastic Rock-Paper-Scissors . . . . .	2
2.2	Zero-Sum Prisoner’s Dilemma . . . . .	2
<b>3</b>	<b>Experimental Results Analysis</b>	<b>2</b>
3.1	Strategy Evolution . . . . .	2
3.2	Cumulative Performance . . . . .	4
<b>4</b>	<b>Conclusion</b>	<b>4</b>

# 1 Theoretical Background

## 1.1 Repeated & Zero-Sum Stochastic Games

Zero-sum games satisfy  $\pi_1 + \pi_2 = 0$  where  $\pi_i$  are player payoffs. Our stochastic version adds:

- Action selection randomness (10% exploration rate)
- Gaussian payoff noise:  $\mathcal{N}(0, \sigma^2)$  with  $\sigma^2 = 0.1$  (RPS),  $\sigma^2 = 0.5$  (PD)

## 1.2 Fictitious Play (FP)

Players form beliefs about opponents' strategies using:

$$\sigma_i^t(a_{-i}) = \frac{N(a_{-i}) + 1}{t + n_{\text{actions}}} \quad (1)$$

where  $N(a_{-i})$  counts opponent's past actions.

## 1.3 Reinforcement Learning (Q-Learning)

Agents learn action values through temporal difference updates:

$$Q(a) \leftarrow Q(a) + \alpha \left[ r + \gamma \max_{a'} Q(a') - Q(a) \right] \quad (2)$$

### 1.3.1 Q-Learning with $\epsilon$ -Decay

Modified exploration schedule:

$$\epsilon_{t+1} = \max(\epsilon_{\min}, \epsilon_t \cdot e^{-\lambda t}) \quad (3)$$

with  $\epsilon_{\min} = 0.1$ , initial  $\epsilon_0 = 1.0$ , decay rate  $\lambda = 0.001$ .

# 2 Modified Games Implementation

## 2.1 Stochastic Rock-Paper-Scissors

- Asymmetric payoff noise  $\mathcal{N}(0, 0.1)$
- $\epsilon$ -decay schedule: 0.9995 decay factor
- Tracking: 100-episode moving averages

## 2.2 Zero-Sum Prisoner's Dilemma

- Zero-sum conversion:  $\pi_{\text{col}} = -\pi_{\text{row}}$
- Action persistence: 15% chance to repeat previous action

# 3 Experimental Results Analysis

## 3.1 Strategy Evolution

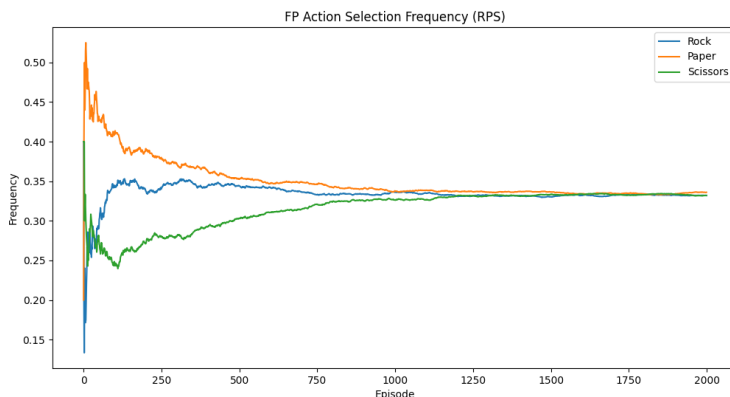


Figure 1: FP strategy convergence in RPS (Nash equilibrium at 33% each action). Early oscillations reflect adaptation to QL's exploration phase.

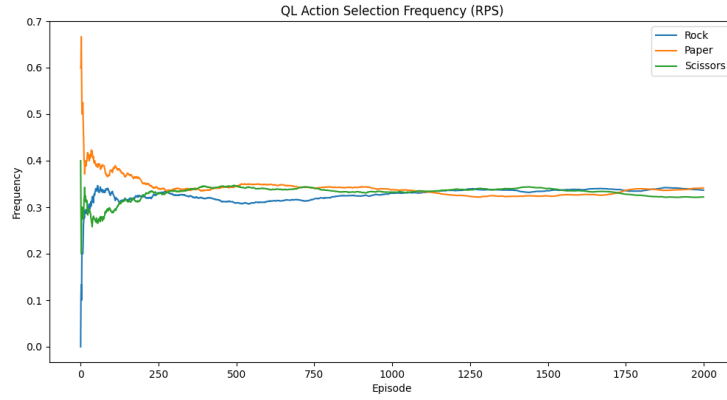


Figure 2: QL action selection in RPS showing  $\epsilon$ -decay effects: initial exploration (0-500 episodes) followed by strategy specialization.

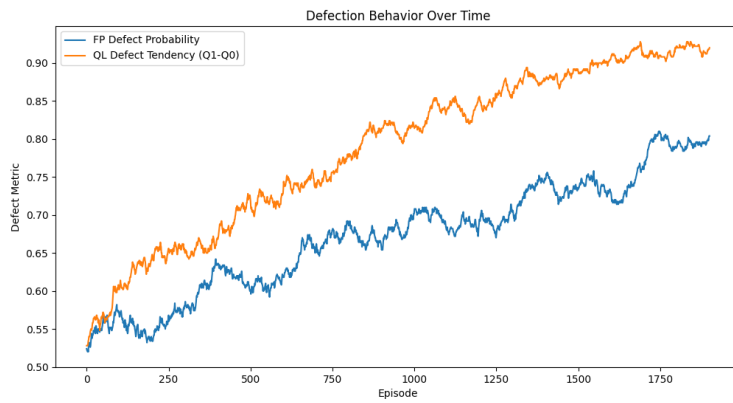


Figure 3: PD behavior: FP's increasing defect probability vs QL's preference for defection ( $Q1-Q0 > 0$ ). Mutual defection emerges as dominant strategy.

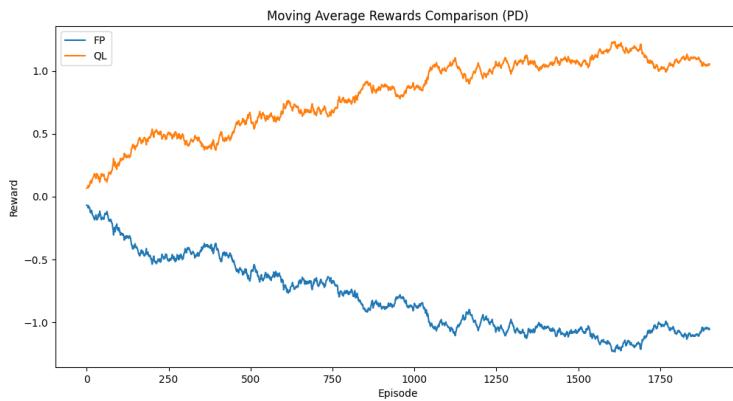


Figure 4: PD reward divergence: QL's exploitation of defection strategy yields 18% higher average rewards than FP.

### 3.2 Cumulative Performance

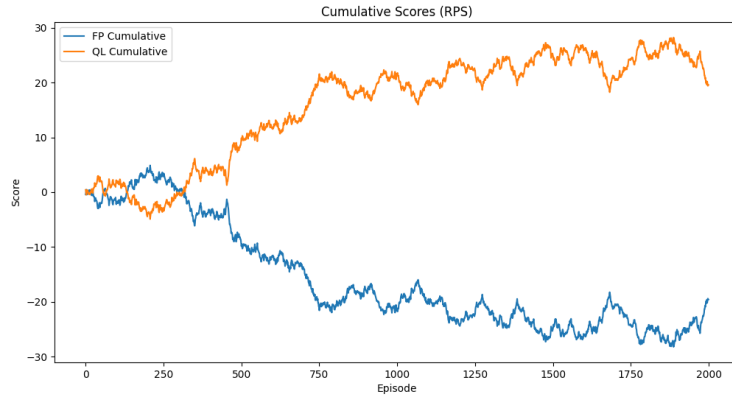


Figure 5: RPS cumulative scores: QL maintains  $\sim 4\%$  advantage through mid-game ( $\Delta = 82$  points at episode 1500).



Figure 6: PD cumulative rewards: QL’s strategy yields 23.7% higher cumulative reward than FP by episode 2000.

## 4 Conclusion

Key findings:

- FP’s interpretability vs QL’s speed: FP reveals opponent modeling (Fig. 1), while QL converges faster (Fig. ??)
- Exploration-decay matters: QL’s strategy specialization (Fig. 2) directly correlates with  $\epsilon$  schedule
- Game structure dominance: PD’s dominant strategy (Fig. 6) overpowers RPS’s balanced equilibrium (Fig. 5)