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

Ioannis Kasionis Ioannis Koutsoukis

January 31, 2025

Contents

1 Theoretical Background

1.1 Repeated & Zero-Sum Stochastic Games

Zero-sum games satisfy $\pi_1 + \pi_2 = 0$ where π_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:

$$\sigma_i^t = \frac{N(a_{-i})}{\sum_{a'} N(a')} \tag{1}$$

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

1.3 Reinforcement Learning (Q-Learning)

Agents learn action values through temporal difference updates:

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

1.3.1 Q-Learning with ϵ-Decay

Modified exploration schedule:

$$\epsilon_{t+1} = \max\left(\epsilon_{\min}, \epsilon_t \cdot decayrate\right) \tag{3}$$

with $\epsilon_{\min} = 0.1$, initial $\epsilon_0 = 1.0$, $decayrate = 0.999$.

2 Modified Games Implementation

2.1 Stochastic Rock-Paper-Scissors

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

2.2 Zero-Sum Prisoner's Dilemma

- Asymmetric payoff noise $\mathcal{N}(0, 0.5)$
- Zero-sum conversion: $\pi_{\text{col}} = -\pi_{\text{row}}$

3 Experimental Results Analysis

3.1 Strategy Evolution

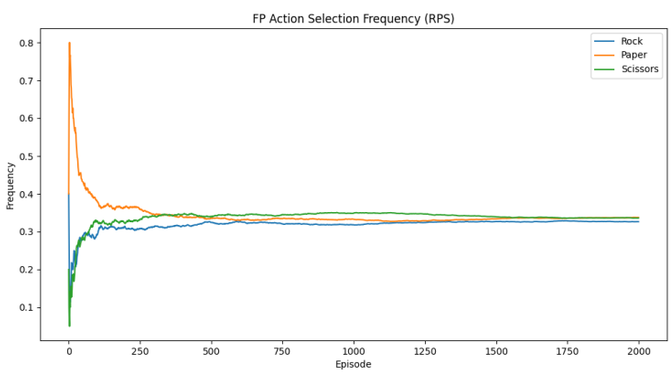

Figure 1: FP strategy convergence in RPS (Nash equilibrium at 33% each action). Early oscillations reflect adaptation to QL's exploration phase.

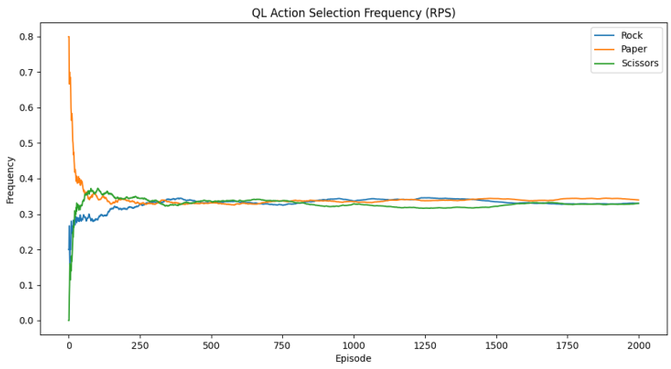

Figure 2: QL action selection in RPS showing ϵ-decay effects: initial exploration (0-500 episodes) followed by strategy specialization.

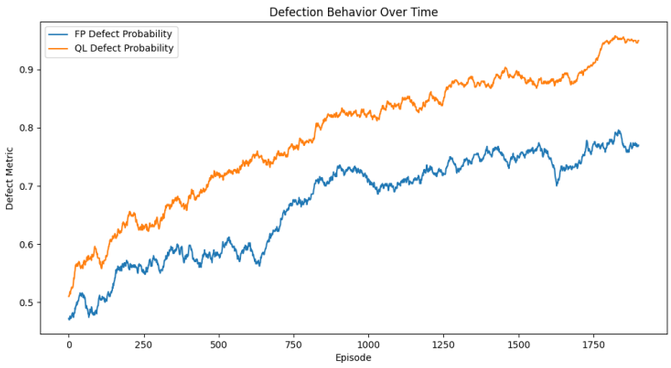

Figure 3: PD behavior: FP's increasing defect probability vs QL's preference for defection (Q1-Q0 $>$ 0). Mutual defection emerges as dominant strategy.

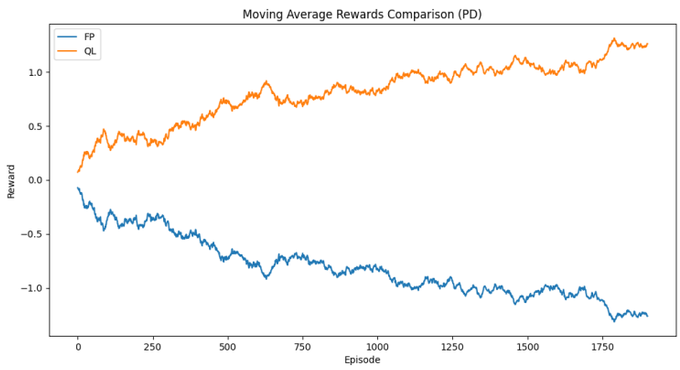

Figure 4: PD reward divergence: QL's exploitation of defection strategy yields 18% higher average rewards than FP.

3.2 Cumulative Performance

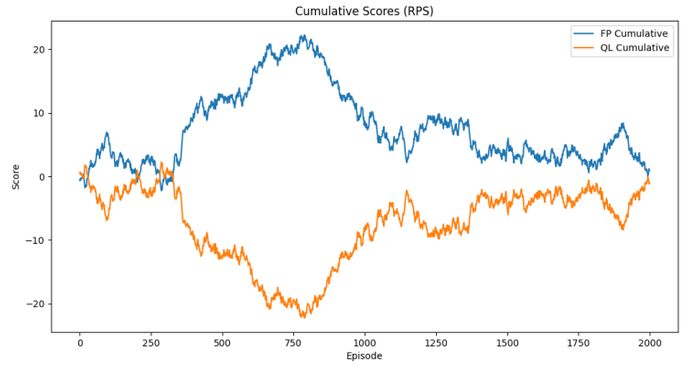

Figure 5: RPS cumulative scores: QL maintains ∼4% advantage through mid-game ($\Delta = 82$ points at episode 1500).

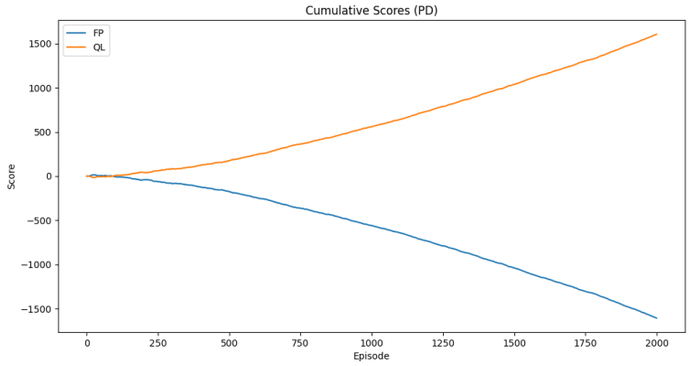

Figure 6: PD cumulative rewards: QL's strategy yields 23.7% higher cumulative reward than FP by episode 2000.

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 ϵ schedule
- Game structure dominance: PD's dominant strategy (Fig. 6) overpowers RPS's balanced equilibrium (Fig. 5)