Q-LEARNING IN AXELROD'S IPD TOURNAMENT

ABSTRACT. In this paper I replicate and advance the experimental work done by Sandholm, Tuomas W., and Robert H. Crites (Biosystems 37.1 (1996): 147-166) into evaluating the use of reinforcement learning techniques in the iterated prisoner's dilemma (IPD). Specifically, I further their investigation of the impact of varying the learning rate, discount rate, and exploration schemes of q-learning strategies for the IPD. Additionally, where their work stopped at examining IPD engagement between two individual strategies, I have extended my experimentation to cover the more general setting of an IPD-tournament as popularized by Robert Axelrod in his 1984 book: The evolution of Cooperation

Source code for implementation available at
https://github.com/petercwill/IPD_Q_learning_exp.

1. MOTIVATION

There is an especially interesting phenomenon in group dynamics in which self-rational behavior for any individual dictates one course of action, however, mass adoption of this policy results in a sub-optimal outcome for all members of the group. Such conflicts between individual and group incentives are sometimes referred to as "collective action" problems. The classical example of this dynamic is the so-called Tragedy of the Commons. In this hypothetical, there is a shared parcel of land that all members of the community may use to graze their livestock upon. Each member of the community thus derives an individual benefit from use of the land, and is selfishly incentivized to use the resource as much as possible. However, if the land is overused, then it will be depleted to the detriment of the entire community. Thus, every individual has a choice: forgo use of the resource and risk having their altruism exploited by more selfish members of the community, or, act self-rationally with the knowledge that if others do as well, the resource will be ruined.

These sorts of collective action problems are common enough that they have attracted much study across many different domains. Robert Axelrod, published a seminal paper in 1981 that studied how so-called stable-cooperative strategies could emerge from a general environment of purely dominant selfish strategies leading to inferior equilibriums [1]. One of the most interesting aspects of the paper was a simulation Axelrod created, wherein multiple different strategies played a round-robin style tournament of the iterated prisoner’s dilemma (IPD). Each game between strategies consisted of 200 iterations of the prisoners dilemma game, and the winning strategy was the one that managed to amass the most amount of points across all tournament pairings. While, there is no single ”best strategy” - as the success of any agent is dependent upon the behavior of other agents in the tournament - the strategy known as ”Tit-for-Tat” generally emerged as a consistent winner in these simulations. This strategy begins by cooperating with the opponent, and then continues by mirroring its opponent’s previous action.

This tournament attracted much notice and spurred much investigation into the IPD game [2]. Today, we examine this setting, to see if it is possible to design agents that can learn to play well within this framework. We note that the constraints of the IPD tournament seem to provide a natural context for reinforcement learning techniques. The IPD is a Markov decision process, in which the current game-state and rewards are purely a function of the agents’ preceding actions. The central task of any agent is to maximize their discounted future payoff, by selecting actions based upon information received from the environment in the form of historical game-states. Additionally, while the payoffs each agent receives for a given game-state are deterministic and fully known, the environment model is not. Thus, it is necessary to estimate the discounted value of game-state action pairs. To do this, I chose to implement a q-learning model[3].

Implementation details for this model are given in section 3, and the results of several experiments examining its performance are given in sections 4 through 6. Overall, I found that when paired against a stable-strategy such as Tit-for-Tat, the q-learners converged to optimal play across a wide range in parameter values. Additionally, when q-learners were paired against each other, the q-learner with a longer game-state memory and exploration schedule managed to outplay the learner with shorter memory and exploration schedule. Finally, in an example tournament I found that the q-learners were successful in finding beneficial strategies against a range of opponents, and often accumulated the most overall points.

2. PRELIMINARIES

The actual prisoners dilemna payoff matrix is shown below. In this table, the first player’s move determines the row, and the second the column. The payoff for their joint action is the tuple (X,Y) with the first player receiving X and the second Y.

In order for the payoffs displayed above to simulate the prisoner’s dilemma, two constraints are required. First, the best outcome for individual players

		Player 2	
		C	D
Player 1	C	(R, R)	(S, T)
	D	(T, S)	(P, P)

must be to exploit the cooperation of another player by defecting, therein earning the temptation payoff T. Conversely, the worst outcome must occur when a player is themselves exploited, and receives the sucker's payoff S. Additionally, the reward for mutual cooperation, R, must be higher than that for mutual defection P. Thus, we require that

$$T > R > P > S$$

Additionally, players should not be able to solve the dilemma by taking turns exploiting each other. Consequently, the payoff for mutual cooperation must exceed the averaged payoff for T and S.

$$R > \frac{T+S}{2} > P$$

While, any set of values that satisfy the above inequalities are valid for the prisoner's dilemma, for my experimentation I choose to use the canonical payoffs of:

$$T = 5, R = 3, P = 1, S = 0$$

3. Q-learning Implementation

Since the environmental model for the IPD game is not known, it is necessary for a learner to estimate optimal state-action values at each iteration of the game. I chose to do this through a stochastic formulation of value iteration. The algorithm for updating $Q(s, a)$ can be simply stated as:

(1) From the current state, s, select an action, a based upon policy π.
(2) Receive a reward r' and new state s'
(3) Update the state-action value for state s by:

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

(4) set the current state to s' and return to (1)

In step (3), the constants α and γ represent the learning rate and discount rate. Experimentally, I found that $\alpha = .2$ produced reasonable results for the conditions examined in experimentation. I investigated the effect of changing γ, intuitively, one would expect that higher values of this tend to make learners more interested in pursuing cooperative strategies as they seek to maximize longer term payoffs. When not specified, I choose to set the discount rate equal to .95, as this is the value specified in Axelrod's original paper.

For the policy, π, I chose to use Boltzmann exploration in which the probability of choosing an action, a at iteration t is given by:

$$p_t(a|s,Q) = \frac{e^{\frac{Q(s,a)}{\tau_t}}}{\sum_{a' \in A} e^{\frac{Q(s,a')}{\tau_t}}}$$

In which the temperature τ_t was annealed as:

$$\tau_t = a \cdot b^t$$

While I experimentally examined the effect of changing annealing constants a and b, typically I found that representative values of $a = 50$ and $b \in [.9, .999]$ worked well. Unless specified, experiments were run under these conditions. I chose to halt learning when the value of τ_t was less than .01, as the probability of not choosing the higher Q-value at this point became vanishingly small, and faced difficulties machine-precision representation.

Finally, the last design consideration I examined in my experimentation was the order of the historical game-state stored in each q-learner's memory. Since the value-iteration scheme given above, works by associating an estimated payoff with each state-action pair, it's natural to wonder how varying the number of games that compose a "state" may impact play. While, theoretically, the state is only limited by the number of games played in the IPD so far, for computational reasons I choose to limit my analysis to orders of between 1-10 games.

4. Experiment 1: Q-learning against Tit-for-Tat

The first experiment I ran involved pitting the implemented Q-learners against the steady strategy Tit-for-tat. Tit-for-tat was chosen due to its prominence in the original Axelrod tournaments. Additionally, Tit-for-tat has another nice property in that optimal play against it is solely a function of the discount rate.

$$\begin{array}{rlll}
\text{Always cooperate:} & V_C = \dfrac{R}{1-\gamma} & \text{for} & \dfrac{2}{3} \leq \gamma \leq 1 \\
\text{Cycle between defect and cooperate:} & V_M = \dfrac{T+\gamma S}{1-\gamma^2} & \text{for} & \dfrac{1}{4} \leq \gamma < \dfrac{2}{3} \\
\text{Always Defect:} & V_D = T + \dfrac{\gamma P}{1-\gamma} & \text{for} & 0 \leq \gamma < \dfrac{1}{4}
\end{array}$$

order	γ	α	**C**	**D**	**M**
1	0.9	0.2	20	0	0
2	0.9	0.2	20	0	0
4	0.9	0.2	20	0	0
8	0.9	0.2	8	0	12
1	0.5	0.2	0	0	20
2	0.5	0.2	0	0	20
4	0.5	0.2	0	0	20
8	0.5	0.2	0	0	20
1	0.1	0.2	0	20	0
2	0.1	0.2	0	20	0
4	0.1	0.2	0	20	0
8	0.1	0.2	0	20	0

TABLE 1. Table showing convergence counts to steady cooperation (C), steady defection (D), and alternating Cooperation and Delectations (M) for IPD games of 10,000 trials

Table 1 shows the results for Q-learners with different discount rates and order parameters playing against against Tit-for-tat in 20 independent IPD trials of 10,000 iterations. The values of γ were chose to be representative of the three styles of optimal play against Tit-for-Tat. As is evident, the Q-learners converged to the optimal style of play in all occasions, with the exception of the 8th-order learner with a discount rate of .9. I believe this anomaly is a result of the learner's relatively long state-memory. Since, it takes 8 games, the learner must store values for a total of $2^9 = 512$ state-action pairs. Since each trial was only 10,000 iterations, it is possible that states were not explored a sufficient number of times for the state-action value estimates to approximate actual state-action values. Additionally, the annealing schedule may have been to aggressive, and the learner may have stopped exploring states, before "good" estimates were obtained.

Finally Figure 1 shows representative payoffs obtained by a q-learner against Tit-for-Tat as it converges into a steady always-cooperate strategy. Here, the first game is shown in the lower left corner, and iterations proceed left-to-right, bottom-to-top. The color of each grid square corresponds to the pay off obtained.

5. EXPERIMENT 2: Q-LEARNER AGAINST Q-LEARNER

The next experiment I considered, was Q-learner against Q-learner. All of these IPD games resulted in either convergence to mutual defection, or

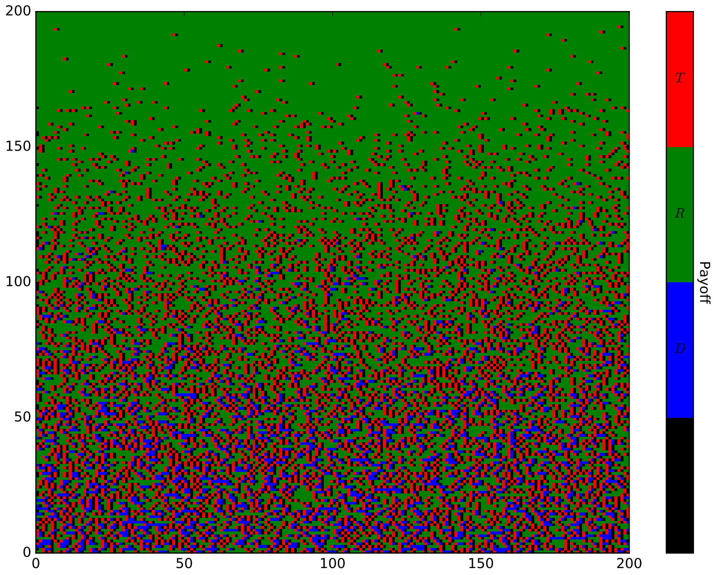

FIGURE 1. Last 40,000 iterations of IPD play until 1000 (R,R) game streak threshold between a 4th-order Q-Learner with discount rate set to .9, learning rate set to .2 and annealing constants a=20, b=.999, against tit-for-tat player

a steady cycle between alternating payoff values. In no case was mutual cooperation observed. When two Q-learners with similar parameters played each other, the IPD almost always converged to cycles in which the two agents traded off an equal number of points with one another. This behavior is shown in Figures 2 and 3

Potentially more interesting, are the results for when differently paramatized q-learners play one another. I consistently found that for IPD games of sufficient length, learners with higher-order memories and longer learning (exploration) policies were able to outperform those with lower-order memories and shorter learning policies. Results for a representative matchup are show in Figures 4 and 5. In Figure 5 the final cycle converged upon, involves the higher-order learner getting the temptation payout, followed by the mutual defection payout.

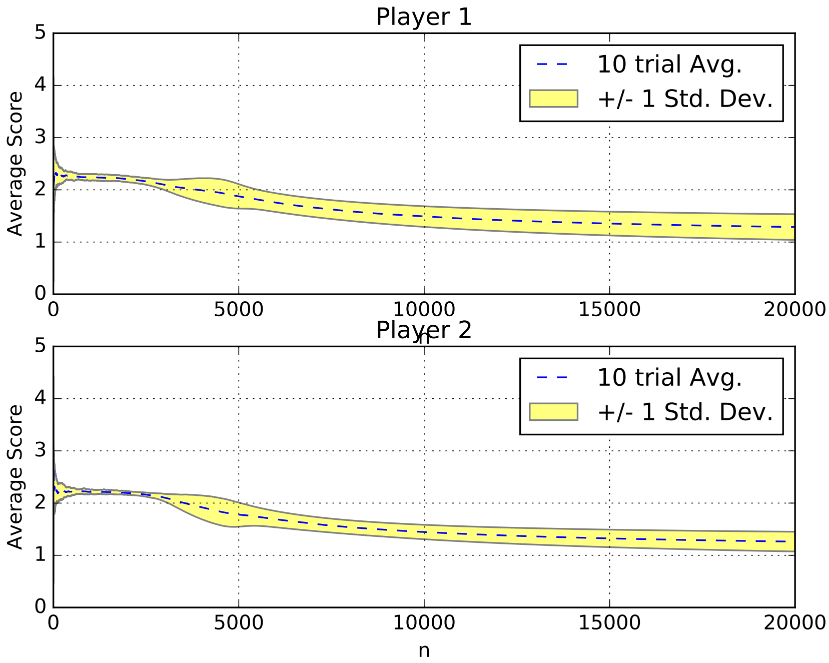

FIGURE 2. Average score for first 20,000 iterations of an IPD game, for two Q-Learners with identical parameters (order = 3, α = .2, γ=.9, a=50, b=.999). Blue dashed line shows the mean value taken across 10 IPD games, yellow shaded region denotes ± one standard deviation

6. EXPERIMENT 3: TOURNAMENT

The final experiment considered in my project is the full Axelrod-style IPD tournament. Here an arbitrary number of players, employing different strategies, play each other in exhaustive round-robin pairings. At the end of each tournament I ranked the players by the total number of points they were able to accumulate across all IPD pairings. Here, it becomes challenging to offer generalized conclusions because the number and kinds of opponents enter into the state space. That is, strategies that perform well against a tournament that consists of one set of opponents, won't generally perform as well against a different set. Nonetheless, as an effort to offer some first results, I considered a tournament that consisted of the following 16 players.

- 4 Tit-for-Tat Players
- 4 Players that always defected (Mean)
- 4 Players that always cooperated (Nice)

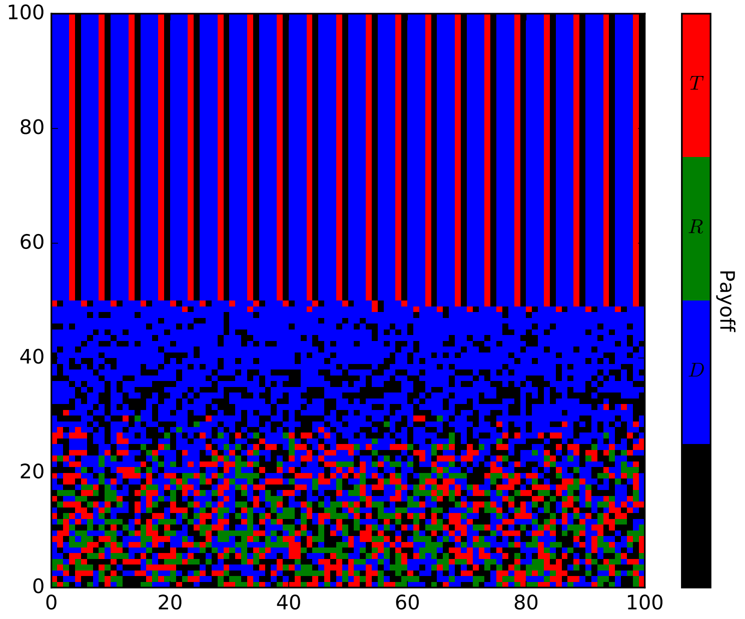

FIGURE 3. 10,000 iterations of IPD play for two Q-learners with identical parameters (order = 3, α = .2, γ=.9, a=50, b=.999). Shown from Player 1's perspective. Convergence to steady-state T,S,P,P,P cycle shown.

- 4 Q-learners
  - (Order = 1, $\alpha = .2$, $\gamma = .95$, $a = 50$, $b = .8$)
  - (Order = 2, $\alpha = .2$, $\gamma = .95$, $a = 50$, $b = .9$)
  - (Order = 3, $\alpha = .2$, $\gamma = .95$, $a = 50$, $b = .99$)
  - (Order = 4, $\alpha = .2$, $\gamma = .95$, $a = 50$, $b = .999$)

Each IPD game consisted of 10,000 iterations. Figure 6 shows the results of 10 independent trials for this tournament setup. For each trial, the rank placement of each strategy is shown. As is evident, the "nice" players consistently did the worst. In match ups against the "mean" players they were immediately and continually exploited. Additionally, the Q-learners eventually learned this style of optimal play against them as well. The "mean" players typically came in second-to-last. While, they represent the individually dominant strategy of always defecting, this meant that they forced the Tit-for-Tat players into steady mutual defection, meaning neither agent accumulated many points. Additionally, the Q-learners eventually converged

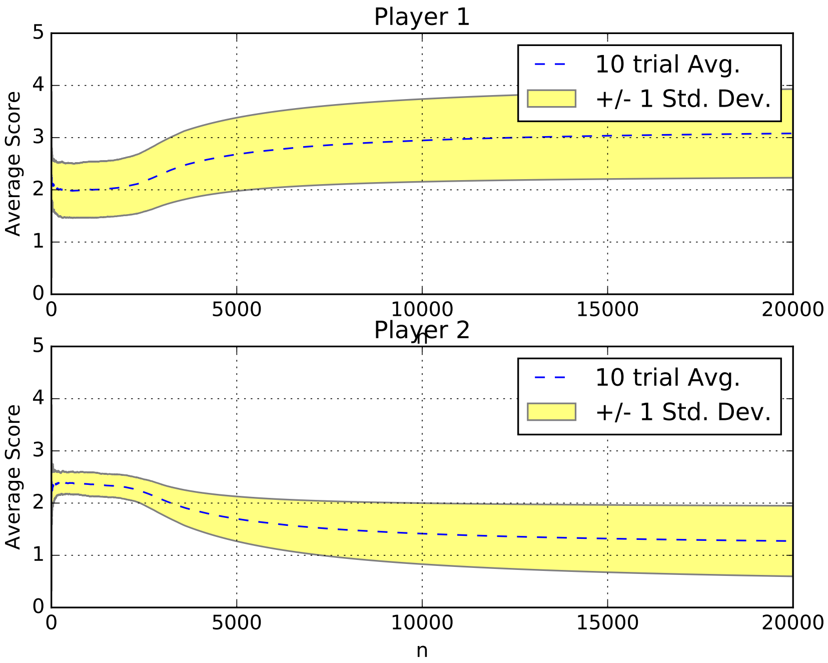

FIGURE 4. Average score for first 20,000 iterations of an IPD game, for two Q-Learners with different parameters. **Player 1** with: (order = 5, α = .2, γ=.9, a=50, b=.999). **Player 2** with: (order = 2, α = .2, γ=.9, a=50, b=.8) Blue dashed line shows the mean value taken across 10 IPD games, yellow shaded region denotes ± one standard deviation

to this style of play as well. Overall, The Q-learning agents did rather well. The two with the longest learning policies consistently finished in the top 2 spots of the tournament. The one Q-learner with the drastically shorter learning policy did rather poorly, typically finishing around the 11th or 12th position.

REFERENCES

[1] R. Axelrod et al. The evolution of strategies in the iterated prisoners dilemma. *The dynamics of norms*, pages 1–16, 1987.

[2] D. B. Fogel. Evolving behaviors in the iterated prisoner's dilemma. *Evolutionary Computation*, 1(1):77–97, 1993.

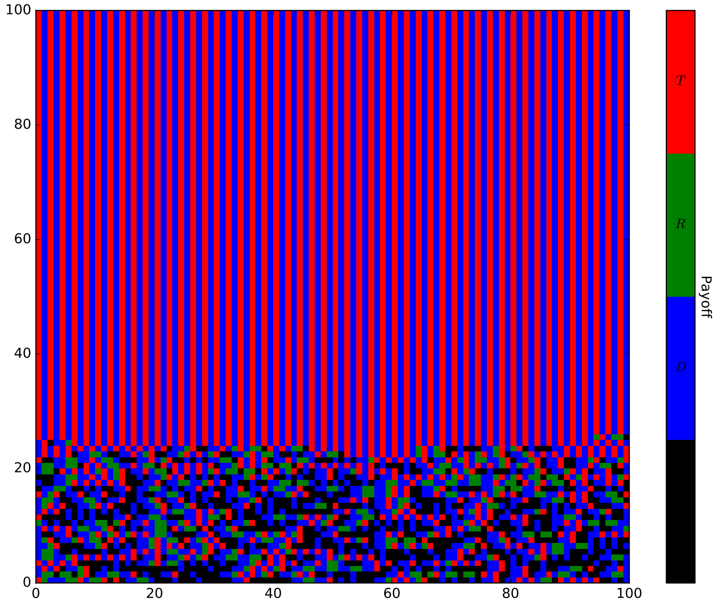

FIGURE 5. 10,000 iterations of IPD play for two Q-learners with different parameters **Player 1** with: (order = 5, α = .2, γ=.9, a=50, b=.999). **Player 2** with: (order = 2, α = .2, γ=.9, a=50, b=.8). Shown from Player 1's perspective. Convergence to steady-state T,P cycle shown.

[3] T. W. Sandholm and R. H. Crites. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1):147–166, 1996.

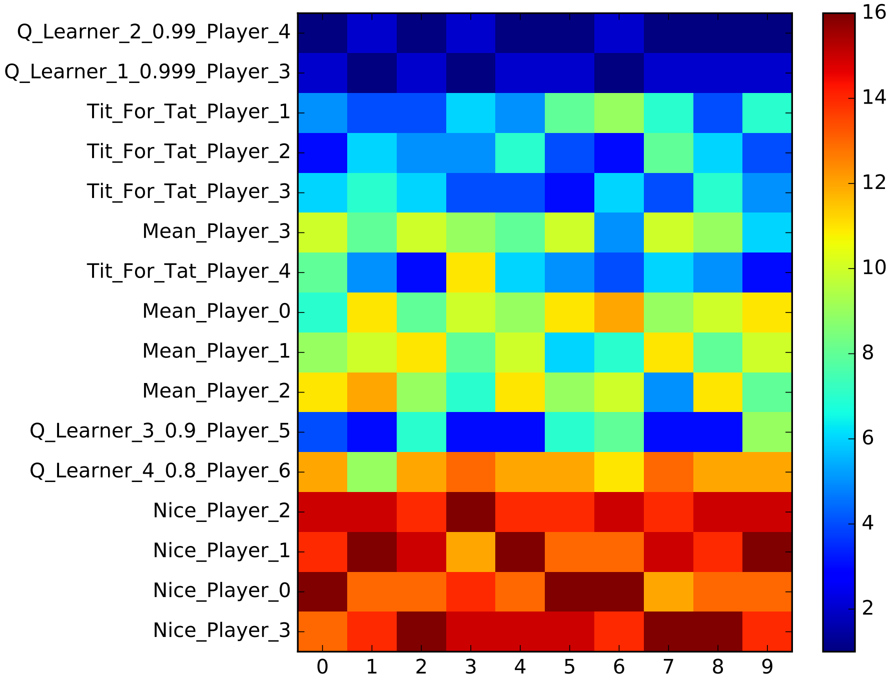

FIGURE 6. Results for 10 independent IDP tournaments shown. For each tournament, every player plays every other player in 10,000 iterations. Heatmap colorbar shows position that the strategy finished the tournament in (blue denoting high-place finishers, red low-place finishers)