Playing Challenging Iterated Two-Person Games Well: A Case Study on the Iterated Traveler's Dilemma

Philip Dasler, Predrag T. Tošić
Department of Computer Science, University of Houston, Houston, Texas, USA
{philip.dasler, pedja.tosic}@gmail.com

Abstract— *We study an interesting 2-player game known as the* Iterated Traveler's Dilemma*, a non-zero sum game in which there is a large number of possible actions in each round and therefore an astronomic number of possible strategies overall. What makes the Iterated TD particularly interesting is that it defies the usual prescriptions of classical game theory insofar as what constitutes an "optimal" strategy. In particular, TD has a* single Nash equilibrium*, yet that equilibrium corresponds to a very low payoff, essentially minimizing* social welfare*. We propose a number of possible strategies for ITD and perform a thorough comparison via a round-robin tournament in the spirit of Axelrod's well-known work on the Prisoner's Dilemma. We motivate the choices of "players" that comprise our tournament and then analyze their performance with respect to several metrics. Finally, we share some interesting conclusions and outline directions for future work.*

Keywords: game theory, two-person non-zero-sum games, bounded rationality, decision making under uncertainty, tournaments

1. Introduction

Theoretical computer science, mathematical economics and AI research communities have extensively studied strategic interactions among two or more autonomous agents from a game-theoretic standpoint. Game theory provides mathematical foundations for modeling interactions among, in general, *self-interested* autonomous agents that may need to combine competition and cooperation in non-trivial ways in order to meet their objectives [1–3]. A classical example of such interactions is the *iterated prisoner's dilemma* [4, 5], a two-person non-zero sum game that has been extensively studied by psychologists, sociologists, economists and political scientists, as well as mathematicians and computer scientists.

We have been studying an interesting and complex 2-player game known as the *Iterated Traveler's Dilemma* [6–8]. The *Traveler's Dilemma* (TD) is a non-zero sum game in which each player has a large number of possible actions. In the iterated context, this means many possible actions in each round and therefore (for games with many rounds) an astronomic number of possible strategies overall. What makes Iterated TD particularly interesting, is that its structure defies the usual prescriptions of classical game theory insofar as what constitutes an "optimal" strategy. There are two fundamental problems to be addressed in this context. One is finding an optimal, or close to optimal, strategy from the standpoint of an individual intelligent agent. This is the "default" problem in classical game theory: finding the best "play" for each agent participating in the game. The second core problem is identifying *pairs of strategies* that would result in an overall desirable behavior, such as maximizing a joint utility function of some sort (i.e., appropriately defined "social welfare"). We have been investigating both these problems in the context of the Iterated Traveler's Dilemma, which has thus far received only modest attention by the computer science research communities. In this paper, we shed some light on the first problem above using a round-robin, many-round tournament and several different performance metrics. We also draw several interesting (and possibly controversial) conclusions based on our extensive experimentation and analyses.

The rest of this paper is organized as follows. We first describe the Traveler's Dilemma (TD) game and motivate why we find it interesting. We also briefly survey the prior art. We then describe the Iterated TD round-robin tournament that we have devised, implemented and experimented with. In that context, we focus on the actual strategies we have chosen as the participants in this tournament, and on why these strategies are good examples of the kinds of strategies one would expect to be "reasonable". We then describe several *metrics* that we have used as yardsticks of performance of the various strategies involved in our round-robin tournament. Next, we summarize our tournament results and discuss our main findings, both those that we expected and those that we honestly find fairly surprising. Finally, we draw conclusions based on our analytical and experimental results to date and outline some promising directions for future research.

2. Traveler's Dilemma (TD)

The Traveler's Dilemma was originally introduced by K. Basu in 1994 [9]. The motivation behind the game was to show the limitations of classical game theory [10], and in particular the notions of *individual rationality* that stem from game theoretic notions of "solution" or "optimal play"

based on well-known mathematical concepts such as *Nash equilibria* [3, 9, 11]. TD is defined with the following parable:

"An airline loses two suitcases belonging to two different travelers. Both suitcases happen to be identical and contain identical antiques. An airline manager tasked to settle the claims of both travelers explains that the airline is liable for a maximum of $100 per suitcase, and in order to determine an honest appraised value of the antiques the manager separates both travelers so they can't confer, and asks them to write down the amount of their value at no less than $2 and no larger than $100. He also tells them that if both write down the same number, he will treat that number as the true dollar value of both suitcases and reimburse both travelers that amount. However, if one writes down a smaller number than the other, this smaller number will be taken as the true dollar value, and both travelers will receive that amount along with a bonus/malus: $2 extra will be paid to the traveler who wrote down the lower value and a $2 deduction will be taken from the person who wrote down the higher amount. The challenge is: what strategy should both travelers follow to decide the value they should write down?"

This game has several interesting properties. Perhaps the most striking among them is that its *unique* Nash equilibrium, the action pair $(p, q) = (\$2, \$2)$, is actually rather bad for both players. This choice of actions results in:

- very low payoff to each player individually (basically, only slightly above the absolutely worst possible, which is $0); and, moreover,
- it minimizes *social welfare*, if we understand social welfare to simply be the sum of the two players' individual utilities.

Yet, it has been argued [7, 9, 12] that a *perfectly rational* player, according to classical game theory, would "reason through" and converge to choosing the lowest possible value, $2. Given that the TD game is symmetric, each player would reason along the same lines and, once selecting $2, not deviate from it, as *unilaterally* deviating from a Nash equilibrium is presumably bad "by definition". However, the non-equilibrium pair of strategies $(\$100, \$100)$ results in each player earning $100, very near the best possible individual payoff. Hence, the early studies of TD concluded that this game demonstrates a woeful inadequacy of classical, Nash Equilibrium based notions of rational behavior. It has also been shown that humans (both game theory experts and laymen) tend to play far from the Nash equilibrium [6], and therefore fare much better than they would if they followed the classical approach.

In general, basing the notion of a "solution" to a game on *Nash equilibria* (NE) has been known to be tricky. Among other things, a game may fail to have any NE (in pure strategies), or it may have multiple Nash equilibria. TD is interesting precisely because it has a *unique* pure-strategy Nash equilibrium, yet this NE results in nearly as low a payoff as one can get. The situation is further complicated by the fact that the game's only "stable" strategy pair is easily seen to be nowhere close to *Pareto optimal*; there are many obvious ways of making both players much better off than if they play the equilibrium strategies. In particular, while neither stable nor an equilibrium, $(\$100, \$100)$ is the unique strategy pair that maximizes social welfare (understood as the sum of individual payoffs), and is, in particular, Pareto optimal. So, the fundamental question arises, how can agents learn to sufficiently trust each other so that they end up playing this optimal strategy pair in the Iterated TD or similar scenarios?

3. The Iterated TD Tournament

Our Iterated Traveler's Dilemma tournament is similar to Axelrod's Iterated Prisoner's Dilemma tournament [13]. In particular, it is a round-robin tournament where each agent plays N matches against each other agent and its own "twin". A match consists of T rounds. In order to have statistically significant results (esp. given that many of our strategies involve randomization in some way), we have selected $N = 100$ and $T = 1000$. In each round, both agents must select a valid bid within the *action space*, defined as $A = \{2, 3, \ldots, 100\}$.

The method in which an agent chooses its next action for all possible histories of previous rounds is known as a strategy. A *valid strategy* is a function S that maps some set of inputs to an action: $S : \cdot \rightarrow A$. In general, the input may include the entire history of prior play, or, in the case of *bounded rationality* models, an appropriate summary of the past histories.

We next define the participants in the tournament, that is, the set of strategies that play one-against-one matches with each other. Let C be the set of agents competing in the tournament: $C = \{c : (c \in S) \wedge (c \text{ is in the tournament})\}$.

We specify a pair of agents competing in a match as $(x, y) \in C$. While we refer to agents as *opponents* or *competitors*, this need not necessarily imply that the agents act as each other's adversaries.

We define agents' actions as follows:

$x_t =$ the bid traveler x makes on round t.

$x_{nt} =$ the bid traveler x makes on round t of match n.

We next define the *reward function* that describes agent payoffs. *Reward per round*, $R : A \times A \rightarrow \mathbb{Z} \in [0, 101]$, for action α against action β, where $\alpha, \beta \in A$, is defined as $R(\alpha, \beta) = min(\alpha, \beta) + 2 \cdot sgn(\beta - \alpha)$, where $sgn(x)$ is the usual *sign* function. Therefore, the total reward $M : S \times S \rightarrow \mathbb{R}$ received by agent x in a match against y is defined as:

$$M(x, y) = \sum_{t=1}^{T} R(x_t,\ y_t)$$

Within a sequence of matches, the reward received by agent x in the n^{th} match against y shall be denoted as $M_n(x, y)$.

4. Strategies in the Tournament

In order to make a reasonable baseline comparison, we chose to utilize the same strategies used by [14], ranging from rather simplistic to relatively complex. What follows is a brief description of the 9 classes of strategies. For a more in depth description see [14].

Randoms: The first, and simplest, class of strategies play a random value, uniformly distributed across a given interval. We have implemented two instances using the following integer intervals: $[2, 100]$ and $[99, 100]$.

Simpletons: The second extremely simple class of strategies are agents that choose the same dollar amount in every round. The \$ values we used in the tournament were $x_t = 2$ (the lowest possible), $x_t = 51$ ("median"), $x_t = 99$ (one below maximal possible, resulting in maximal payoff should the opponent make the highest possible bid), and $x_t = 100$ (the highest possible).

Tit-for-Tat, in spirit: The next class of strategies are those that can be viewed as *Tit-for-Tat-in-spirit*, where Tit-for-Tat is the famously simple, yet effective, strategy for the iterated prisoner's dilemma [4, 5, 13, 15]. The idea behind *Tit-for-Tat* is simple: cooperate on the first round, then do exactly what your opponent did on the previous round. In the iterated TD, each agent has many actions at his disposal, hence there are different ways of responding appropriately in a Tit-for-Tat manner. In general, playing high values can be considered as an approximate equivalent of "cooperating", whereas playing low values is an analogue of "defecting". Following this basic intuition, we have defined several Tit-for-Tat-like strategies for the iterated TD. These strategies can be roughly grouped into two categories. First, the simple Tit-for-Tat strategies bid some value ϵ below the bid made by the opponent in the last round, where $\epsilon \in \{1, 2\}$. Second, the predictive Tit-for-Tat strategies compare whether their last bid was lower than, equal to, or higher than that of their opponent. Then a bid is made similar to the simple TFT strategy, i.e. some value ϵ below the bid made by competitor c in the last round, where $c \in [x, y]$ and $\epsilon \in \{1, 2\}$. The key distinction is that a bid can be made relative to either the opponent's last bid or the bid made by one's own strategy itself. In essence, this strategy is predicting that the opponent may make a bid based on the agent's own last move and, given that prediction, it attempts to "outsmart" the opponent.

Mixed: The mixed strategies probabilistically combine up to three strategies. For each mixed strategy, a strategy σ is selected from one of the other strategies defined in the competition (i.e., $\sigma \in C$) for each round according to a probability mass distribution. Once a strategy has been selected, the value that σ would bid at time step t is bid. We chose to use only mixtures of the TFT, Simpleton, and Random strategies. This allowed for greater transparency when attempting to interpret and understand the causes of a particular strategy's performance.

Buckets – Deterministic: These strategies keep a count of each bid by the opponent in an array of "buckets". The bucket that is most full (i.e., the value bid most often) is used as the predicted value, with ties being broken by one of the following methods: the highest valued bucket wins, the lowest valued bucket wins, a random bucket wins, and the most recently tied-for-the-largest bucket wins. The strategy then bids the next lowest value below the predicted value. An instance of each tie breaking method competed in the tournament.

Buckets – Probability Mass Function based: As above, this strategy class counts instances of the opponent's bids and uses them to predict the agent's own next bid. Rather than picking the value most often bid, the buckets are used to define a probability mass function from which a prediction is randomly selected. Values in the buckets decay over time in order to emphasize newer data over old and we have set a retention rate ($0 \leq \gamma \leq 1$) to determine the rate of decay. We have entered into our tournament several instances of this strategy using the following rate of retention values γ: 0.8, 0.5, and 0.2. As above, the strategy bids the next lowest value below the predicted value. We observe that the "bucket" strategies based on probability mass buckets are quite similar to a learning model in [7].

Simple Trend: This strategy looks at the previous k time steps, creates a line of best fit on the rewards earned, and compares its slope to a threshold δ. If the trend has a positive slope greater than δ, then the agent will continue to play the same bid it has been as the rewards are increasing. If the slope is negative and $|slope| > \delta$, then the system is trending toward the Nash equilibrium and, thus, the smaller rewards. In this case, the agent will attempt to maximize social welfare and play 100. Otherwise, the system of bidding and payouts is relatively stable and the agent will play the "one under" strategy. We have implemented instances of this strategy with $\delta = 0.5$ and the following values of k: 3, 10, and 25. While our choice of δ intuitively makes sense, we admit that picking δ "half-way" between 0.0 and 1.0 is fairly arbitrary.

Q-learning: This strategy uses a learning rate α to emphasize new information and a discount rate γ to emphasize future gains. In particular, the learners in our tournament are simple implementations of *Q-learning* [16] as a way of predicting the best action at time $(t + 1)$ based on the action selections and payoffs at times $\{1, 2, ..., t\}$. This is similar to the Friend-or-Foe Q-learning method [17] without the limitation of having to classify the allegiance of one's opponent.

Due to scaling issues, our implementation of Q-learning does not capture the entire state/action space but rather divides it into a small number of meaningful classes, namely,

three states and three actions, as follows:

State: The opponent played higher, lower, or equal to our last bid.

Action: Play one higher than, one lower than, or equal to our previous bid.

Recall that *actions* are defined for a single round. Our implementation treats each state as a collection of moves by the opponent over the last k rounds. We have decided to use 5 as an arbitrary value for k as it allows us to capture some history without data sizes becoming unmanageable. We have implemented this basic Q-learning algorithm with the learning rates of 0.8, 0.5 and 0.2, and with discount rates of 0.0 and 0.9, for a total of 6 different variations of Q-learning.

Zeuthen Strategies: A Zeuthen Strategy [18] calculates the level of risk of each agent, and makes *concessions* accordingly. Risk is the ratio of loss from accepting the opponent's proposal to the loss of forcing the conflict deal (the deal made when no acceptable proposal can be found). While ITD is not a strict negotiation, we treat each bid as a proposal. If $x_t = i$, then X is proposing $(i, i+1)$ be the next action pair. For the Traveler's Dilemma, we consider the conflict deal to be the Nash Equilibrium at $(\$2, \$2)$.

Given the proposals of each agent, a risk comparison is done. An agent will continue to make the same proposal while its risk is greater than or equal to its opponent's. Otherwise, the agent will make the *minimal sufficient concession*, i.e. the agent adjusts its proposal so that (i) the agent's risk is higher than that of its opponent and (ii) the opponent's utility increases as little as possible. Due to the peculiar structure of the Iterated TD game, it is possible for the "concession" to actually lead to a *loss of utility* for the opponent. We have therefore implemented two variations: one that allows a negative concession and one that does not.

5. Utility metrics

In order to classify a particular strategy as better than another, one needs to define the metric used to make this determination. Our experimentation and subsequent analysis were performed with respect to four distinct utility metrics. The first, U_1, treats the actual dollar amount as the direct payoff to the agent. This is the most common metric in the game theory literature; prior art on Iterated TD generally considers only this metric or some variant of it. In contrast, U_2 is a "pairwise victory" metric: an agent strives to beat its opponent, regardless of the actual dollar amount it receives. Finally, we introduce two additional metrics, U_3 and U_3', that attempt to capture both the payoff (dollar amount) that an agent has achieved, and the "opportunity lost" due to not playing differently. In a sense, both of these metrics attempt to quantify how much an agent wins compared to how much an *omniscient agent* (i.e., one that always correctly predicts the other agent's bid) would be able to win. To be clear, the assumption here is one of omniscience, not omnipotence: an "ideal" omniscient agent is still not able to actually force what the other agent does.

Total reward: U_1

This metric captures the overall utility rewarded to the agent. It is simply a sum of all money gained, normalized by the total number of rounds played and the maximum allowable reward. It is defined as follows:

$$U_1(x) = \frac{1}{|C|} \sum_{j \in C} \left[\frac{1}{max(R)NT} \sum_{n=1}^{N} M_n(x, j) \right]$$

where:

- $max(R)$ is the maximum possible reward given in one round;
- N is the number of matches played between each pair of competitors;
- T is the number of rounds to be played in each match; and
- $|C|$ is the number of competitors in the tournament.

Pairwise Victory Count : U_2

This metric captures the agent's ability to do better than its opponents on a match per match basis. The metric itself is essentially the difference between matches won and matches lost. The result is normalized and 0.5 is added in order to bring all values inside the range $[0.0, 1.0]$. It is defined as follows:

$$U_2(x) = 0.5 + \frac{1}{2|C|} \sum_{j \in C} \left[\frac{1}{N} \sum_{n=1}^{N} sgn(M_n(x, j) - M_n(j, x)) \right]$$

where:

- N is the number of matches played between each pair of competitors;
- $|C|$ is the number of competitors in the tournament.

The intent of this metric is to capture a strategy's ability to "outsmart" its opponent, regardless of the possibility of a Pyrrhic victory.

Perfect Score Proportion : U_3

This metric attempts to capture the level of optimality of an agent, where both the achieved payoff and the missed opportunity for yet higher payoff (based on what the opposing agent does) are taken into account. The metric captures these two aspects of performance by keeping a running total of the *lost reward* ratio. This is the ratio of the reward received to the best possible reward, given what the opponent has played. The resulting sum is then normalized by the total number of rounds played. The metric is formally defined as follows:

$$U_3 = \frac{1}{|C|} \sum_{j \in C} \left[\frac{1}{N} \sum_{m=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \frac{R(x_{mt}, j_{mt})}{R(max(2, (j_{mt} - 1)), j_{mt})} \right) \right]$$

where:

- N is the number of matches played between each pair of competitors
- T is the number of rounds to be played in each match
- $|C|$ is the number of competitors in the tournament

During the course of our work, it has been observed that this metric tends to be biased in favor of strategies that *overbid*. When overbidding, the difference between the reward received and the optimal reward is at worst 4. Thus, regardless of by how much an agent overbids, the lost reward ratio remains relatively small.

Perfect Bid Proportion : U_3'

This metric is another attempt to capture the level of optimality of the agent, but without the overbid bias. It does so by keeping a running total of the bid imperfection ratio. This is the difference between the agent's bid and the ideal bid, given what the opponent has played, divided by the greatest possible difference in bids. Since we want to look favorably upon a smaller difference, this value is subtracted from 1. This sum is then normalized using the total number of rounds played. The metric is defined as follows:

$$U_3' = \frac{1}{|C|} \sum_{j \in C} \left[\frac{1}{N} \sum_{m=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \left[1 - \frac{|x_{mt} - (j_{mt} - 1)|}{max(A) - min(A)} \right] \right) \right]$$

where:

- N is the number of matches played between each pair of competitors;
- T is the number of rounds to be played in each match;
- $|C|$ is the number of competitors in the tournament.

Table 1: Ranking Based on U1

0.760787	Random [99, 100]
0.758874	Always 100
0.754229	Always 99
0.754138	Zeuthen Strategy - Positive
0.744326	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 20%)
0.703589	Simple Trend - K = 3, Eps = 0.5
0.681784	Mixed - TFT (y-1), 80%); (R[99, 100], 20%)
0.666224	Simple Trend - K = 10, Eps = 0.5
0.639572	Simple Trend - K = 25, Eps = 0.5
0.637088	Mixed - L(x) E(x) H(y-g), 80%); (100, 20%)
0.534378	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 10%); (2, 10%)
0.498134	Q Learn - alpha= 0.2, discount= 0.0
0.497121	Q Learn - alpha= 0.5, discount= 0.0
0.496878	Q Learn - alpha= 0.5, discount= 0.9
0.495956	Q Learn - alpha= 0.2, discount= 0.9
0.493640	Q Learn - alpha= 0.8, discount= 0.0
0.493639	Buckets - (Fullest, Highest)
0.493300	Q Learn - alpha= 0.8, discount= 0.9
0.492662	TFT - Low(y-g) Equal(x-g) High(x-g)
0.452596	Zeuthen Strategy - Negative
0.413992	Buckets - PD, Retention = 0.5
0.413249	Always 51
0.412834	Buckets - PD, Retention = 0.2
0.408751	Buckets - PD, Retention = 0.8
0.406273	Buckets - (Fullest, Random)
0.390303	TFT - Simple (y-1)
0.387105	Buckets - (Fullest, Newest)
0.334967	Buckets - (Fullest, Lowest)
0.329227	TFT - Simple (y-2)
0.316201	Random [2, 100]
0.232063	Mixed - L(y-g) E(x-g) H(x-g), 80%); (2, 20%)
0.164531	Mixed - L(x) E(x) H(y-g), 80%); (100, 10%); (2, 10%)
0.136013	TFT - Low(x) Equal(x) High(y-g)
0.135321	TFT - Low(x) Equal(x-2g) High(y-g)
0.030905	TFT - Low(x-2g) Equal(x) High(y-g)
0.030182	TFT - Low(x-2g) Equal(x-2g) High(y-g)
0.026784	Mixed - L(x) E(x) H(y-g), 80%); (2, 20%)
0.024322	Always 2

6. Results and Analysis

The Traveler's Dilemma Tournament that we have experimented with involves a total of 38 competitors (i.e., distinct strategies). Each competitor plays each other competitor (including its own "twin") 100 times. Each match is played for 1000 rounds. No meta-knowledge or knowledge of the future is allowed: learning and adaptation of those agents whose strategies are adaptable takes place exclusively based on the previous rounds in a match against a given opponent without *a priori* knowledge of that opponent. For definitions of the shorthand notation used in the sequel, see [14]. Throughout the rest of the paper, we assume the default version of ITD: the space of allowable bids is the interval of integers [2, 100], granularity of bids is 1, and the Bonus/Malus is equal to 2.

[Note: due to space constraints, we do not include the tabulated tournament results with respect to metric U_3'.]

The top three performers in our tournament, with respect to the earned dollar amount as the bottom line (metric U_1), are three "dumb" strategies that always bid very high. Interestingly enough, randomly alternating between the highest possible bid ($100) and "one under" the highest bid ($99) slightly outperforms both "always max. possible" and "always one under max. possible" strategies. We find it somewhat surprising that the performance of Tit-for-Tat-based strategies varies so greatly depending on the details of bid prediction method and metric choice. So, while a relatively complex TFT-based strategy that, in particular, (i) makes a nontrivial model of the other agent's behavior and (ii) "mixes in" some randomization, is among the top performers with respect to metric U_1, other TFT-based strategies have fairly mediocre performance with respect to the same metric, and are, indeed, scattered all over the tournament table. In contrast, if metric U_3 is used, then the simplest, deterministic "one under what the opponent did on the previous round" TFT strategy, which is a direct analog of the famous TFT in Axelrod's Iterated Prisoner's Dilemma, is the top performer among all 38 strategies in the tournament – while more sophisticated TFT strategies, with considerably more complex models of the opponent's behavior and/or randomization involved, show fairly average performance. Moreover, if U_3 is used as the yardstick, then (i) 3 out of the top 4 performers overall are TFT-based strategies, and (ii) all simplistic TFT strategies outperform all more sophisticated ones.

Not very surprisingly, the top (and bottom) performers with respect to metric U_1 and those with respect to U_2 are practically inverted; so, for example, the very best performer

Table 2: Ranking Based on U2

0.984342	Always 2
0.924474	TFT - Low(x-2g) Equal(x-2g) High(y-g)
0.915263	TFT - Low(x-2g) Equal(x) High(y-g)
0.887500	Mixed - L(x) E(x) H(y-g), 80%); (2, 20%)
0.845132	TFT - Simple (y-2)
0.839868	TFT - Low(x) Equal(x-2g) High(y-g)
0.832368	TFT - Low(x) Equal(x) High(y-g)
0.791842	TFT - Simple (y-1)
0.727105	Buckets - PD, Retention = 0.2
0.681842	Mixed - L(x) E(x) H(y-g), 80%); (100, 20%)
0.669079	Mixed - L(y-g) E(x-g) H(x-g), 80%); (2, 20%)
0.653816	Buckets - PD, Retention = 0.5
0.629605	Mixed - L(x) E(x) H(y-g), 80%); (100, 10%); (2, 10%)
0.622632	TFT - Low(y-g) Equal(x-g) High(x-g)
0.616711	Buckets - PD, Retention = 0.8
0.558158	Mixed - TFT (y-1), 80%); (R[99, 100], 20%)
0.557368	Buckets - (Fullest, Newest)
0.539342	Simple Trend - K = 25, Eps = 0.5
0.528421	Buckets - (Fullest, Lowest)
0.491842	Random [2, 100]
0.483684	Simple Trend - K = 10, Eps = 0.5
0.480789	Buckets - (Fullest, Random)
0.463816	Buckets - (Fullest, Highest)
0.455000	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 10%); (2, 10%)
0.407500	Simple Trend - K = 3, Eps = 0.5
0.303158	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 20%)
0.260263	Q Learn - alpha= 0.5, discount= 0.0
0.255658	Q Learn - alpha= 0.8, discount= 0.0
0.253289	Q Learn - alpha= 0.8, discount= 0.9
0.252763	Q Learn - alpha= 0.2, discount= 0.0
0.249605	Q Learn - alpha= 0.5, discount= 0.9
0.247237	Q Learn - alpha= 0.2, discount= 0.9
0.200000	Always 51
0.183289	Zeuthen Strategy - Negative
0.092368	Always 99
0.066711	Random [99, 100]
0.040789	Zeuthen Strategy - Positive
0.013289	Always 100

Table 3: Ranking Based on U3

0.973118	Buckets - PD, Retention = 0.2
0.970587	Buckets - PD, Retention = 0.5
0.970356	TFT - Simple (y-1)
0.968923	Simple Trend - K = 10, Eps = 0.5
0.967860	TFT - Low(y-g) Equal(x-g) High(x-g)
0.965654	TFT - Simple (y-2)
0.962212	Simple Trend - K = 3, Eps = 0.5
0.959252	Buckets - PD, Retention = 0.8
0.955886	Simple Trend - K = 25, Eps = 0.5
0.953725	Buckets - (Fullest, Newest)
0.945405	Buckets - (Fullest, Random)
0.945222	Buckets - (Fullest, Lowest)
0.943694	Buckets - (Fullest, Highest)
0.919699	Mixed - TFT (y-1), 80%); (R[99, 100], 20%)
0.908562	Mixed - L(y-g) E(x-g) H(x-g), 80%); (2, 20%)
0.899511	Mixed - L(x) E(x) H(y-g), 80%); (100, 20%)
0.864914	TFT - Low(x) Equal(x) High(y-g)
0.863831	TFT - Low(x) Equal(x-2g) High(y-g)
0.823397	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 20%)
0.822670	Random [99, 100]
0.820128	Always 99
0.818674	Always 100
0.817728	Zeuthen Strategy - Positive
0.803646	Q Learn - alpha= 0.2, discount= 0.9
0.801725	Q Learn - alpha= 0.5, discount= 0.0
0.801380	Q Learn - alpha= 0.2, discount= 0.0
0.800006	Q Learn - alpha= 0.5, discount= 0.9
0.798992	TFT - Low(x-2g) Equal(x) High(y-g)
0.798681	TFT - Low(x-2g) Equal(x-2g) High(y-g)
0.798417	Always 2
0.798402	Q Learn - alpha= 0.8, discount= 0.9
0.798277	Mixed - L(x) E(x) H(y-g), 80%); (2, 20%)
0.797728	Q Learn - alpha= 0.8, discount= 0.0
0.758573	Mixed - L(x) E(x) H(y-g), 80%); (100, 10%); (2, 10%)
0.751044	Mixed - L(y-g) E(x-g) H(x-g), 80%); (100, 10%); (2, 10%)
0.741721	Always 51
0.521901	Zeuthen Strategy - Negative
0.518840	Random [2, 100]

with respect to U_2 is the strategy "always bid \$2" (which also happens to be the only non-dominated strategy in the classical game theoretic-sense). On the other hand, the three best performers with respect to U_1 are all among the four bottom performers with respect to U_2, with the only strategy that *may* maximize social welfare (bidding \$100 against a collaborative opponent) falling at the rock bottom of the U_2 ranking. The main conclusion we draw from this performance inversion is that *when a two-player game has a structure that makes it very far from being zero-sum, the traditional precepts from classical game theory on what constitutes good strategies are more likely to fail.* This does not mean to suggest that classical game theory is useless; rather, we'd argue that the appropriate quantitative, mathematical models of rationality for zero-sum, or nearly zero-sum, encounters aren't necessarily the most appropriate notions for games that are rather far from being zero-sum.

Returning to our tournament results, what we have found *very surprising* is the relative mediocrity of the learning based strategies: Q-learning based strategies did not excel with respect to any of the four metrics we studied. On the other hand, it should be noted that the adaptability of Q-learning based strategies apparently ensures that they do not do too badly overall, regardless of the choice of metric. Furthermore, the choice of the learning rate seems to make very little difference: for each of the four metrics, all three Q-learning based strategies show similar performance, and hence end up ranked adjacently or almost adjacently.

Another interesting result is the performance of Zeuthen-based strategies. ITD as a strategic encounter is not of a negotiation nature, hence we have been criticized for even considering Zeuthen-like strategies as legitimate contenders in our tournament. However, excellent performance of the Zeuthen strategy with positive concessions only (at least w.r.t. the "bottom line" metric U_1) validates our approach. Interestingly enough, the same strategy does not perform particularly well w.r.t. metric U_3. It is worth noting, however, that the only three strategies that outperform Zeuthen-positive with respect to U_1 perform similarly to Zeuthen-positive with respect to U_3. Those strategies perform only slightly better than Zeuthen and way below the best performers with respect to U_3, namely, the bucket-based, simplistic TFT-based, and simple-trend-based strategies.

Finally, given the performance of Zeuthen-negative (the variant allowing negative "concessions") with respect to all metrics, it appears that "enticing" the opponent to behave differently indeed does not work when the "concessions" are not true concessions. Intuitively, this makes a perfect sense;

our simulation results provide an experimental validation of that common-sense intuition.

Due to space constraints, further analysis of our tournament results is left for the future work.

7. Summary and Future Work

We study the *Iterated Traveler's Dilemma* two-player game by designing, implementing and analyzing a round robin tournament with 38 distinct participating strategies. Our detailed analysis of the performance of various strategies with respect to several different metrics has corroborated that, for a game whose structure is far from zero-sum, the traditional game-theoretic notions of rationality and optimality may turn out to be rather unsatisfactory. Our investigations raise several interesting questions, among which we are particularly keen to further investigate the following:

(i) To what extent simple models of reinforcement learning, such as Q-learning, can be really expected to help performance?

(ii) To what extent complex models of the other agent really help an agent increase its payoff in the repeated play?

(iii) Why are performances of various TFT-based strategies so broadly different from each other? This opens up interesting questions from meta-learning [19, 20] and meta-reasoning standpoints: how can one design TFT-based strategies that are likely to do well across tournaments (that is, choices of opponents) and across performance metrics.

(iv) What effects on strategies and their performance would an adjustment in the bonus/malus have? For prior research on how human behavior changes with a change in bonus/malus, see [7] and [12].

In our future work, in addition to more detailed analysis of the existing strategies and study of some new ones, we plan to pursue a systematic comparative analysis of how groups of closely related strategies perform against each other when viewed as *teams*. We also plan to further investigate other notions of game equilibria, and try to determine which such notions adequately capture what our intuition would tell us constitutes good ways of playing the iterated TD and other 'far-from-zero-sum" two-player games.

References

[1] J. S. Rosenschein and G. Zlotkin, *Rules of encounter: designing conventions for automated negotiation among computers*. MIT Press, 1994.

[2] S. Parsons and M. Wooldridge, "Game theory and decision theory in Multi-Agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 5, pp. 243–254, 2002.

[3] M. Wooldridge, *An Introduction to MultiAgent Systems*. John Wiley and Sons, 2009.

[4] R. Axelrod, "Effective choice in the prisoner's dilemma," *Journal of Conflict Resolution*, vol. 24, no. 1, pp. 3 –25, Mar. 1980.

[5] ——, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[6] T. Becker, M. Carter, and J. Naeve, "Experts playing the traveler's dilemma," Department of Economics, University of Hohenheim, Germany, Tech. Rep., Jan. 2005.

[7] C. M. Capra, J. K. Goeree, R. Gómez, and C. A. Holt, "Anomalous behavior in a traveler's dilemma?" *The American Economic Review*, vol. 89, no. 3, pp. 678–690, Jun. 1999.

[8] M. Pace, "How a genetic algorithm learns to play traveler's dilemma by choosing dominated strategies to achieve greater payoffs," in *Proc. of the 5th international conference on Computational Intelligence and Games*, 2009, pp. 194–200.

[9] K. Basu, "The traveler's dilemma: Paradoxes of rationality in game theory," *The American Economic Review*, vol. 84, no. 2, pp. 391–395, May 1994.

[10] J. V. Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton Univ. Press, 1944.

[11] K. Basu, "The traveler's dilemma," *Scientific American Magazine*, Jun. 2007.

[12] J. K. Goeree and C. A. Holt, "Ten little treasures of game theory and ten intuitive contradictions," *The American Economic Review*, vol. 91, no. 5, pp. 1402–1422, Dec. 2001.

[13] R. Axelrod, *The evolution of cooperation*. Basic Books, 2006.

[14] P. Dasler and P. Tosic, "The iterated traveler's dilemma: Finding good strategies in games with "bad" structure: Preliminary results and analysis," in *8th Euro. Workshop on Multi-Agent Systems, EUMAS'10*, Dec 2010.

[15] A. Rapoport and A. M. Chammah, *Prisoner's Dilemma*. Univ. of Michigan Press, Dec. 1965.

[16] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[17] M. L. Littman, "Friend-or-Foe q-learning in General-Sum games," in *Proc. of the 18th Int'l Conf. on Machine Learning*. Morgan Kaufmann Publishers Inc., 2001, pp. 322–328.

[18] F. F. Zeuthen, *Problems of monopoly and economic warfare / by F. Zeuthen ; with a preface by Joseph A. Schumpeter*. London: Routledge and K. Paul, 1967, first published 1930 by George Routledge & Sons Ltd.

[19] R. Sun, "Meta-Learning processes in Multi-Agent systems," *Proceedings of Intelligent Agent Technology*, pp. 210—219, 2001.

[20] P. T. Tosic and R. Vilalta, "A unified framework for reinforcement learning, co-learning and meta-learning how to coordinate in collaborative multi-agent systems," *Procedia Computer Science*, vol. 1, no. 1, pp. 2211–2220, May 2010.