Iterated Traveler's Dilemma: Analysis of Individual and Team Performances and Challenges Ahead

Predrag T. Tošić and Philip Dasler

Department of Computer Science, University of Houston
Houston, Texas, USA
pedja.tosic@gmail.com, *philip.dasler@gmail.com*

Abstract. We study the iterated version of *Travelers Dilemma* (TD). TD is a two-player, non-zero sum game that offers plenty of incentives for cooperation. Our goal is to gain deeper understanding of iterated two-player games whose structures are *far from zero-sum.* Our experimental study and analysis of Iterated TD is based on a round-robin tournament we have recently designed, implemented and analyzed. Our round-robin tournament involves 38 distinct participating strategies, and is motivated by the seminal work by Axelrod et al. on iterated Prisoners Dilemma. In this paper, we first motivate and define the strategies competing in our tournament, followed by a summary of the tournament results with respect to individual strategies. We then extend the comparison-and-contrast of the relative performances of different *individual* strategies in the tournament, and carefully analyze how *groups of closely related strategies* perform when each such group is viewed as a "team". We draw some interesting lessons from the analyses of individual and team performances, and outline our ongoing and future work.

Keywords: *game theory, non-zero-sum games, travelers dilemma, iterated games, bounded rationality, tournaments*

1 Introduction

Game theory provides mathematical foundations for modeling interactions among, in general, *self-interested rational agents* that may need to combine competition and cooperation with each other in order to meet their individual objectives [17, 19, 23]. An example of such interactions is the iterated *Prisoner's Dilemma* (PD) [1, 2], a classical two-person non-zero-sum game that has been extensively studied by psychologists, sociologists, economists, political scientists, applied mathematicians and computer scientists.

In this paper, we study an interesting and rather complex 2-player non-zero sum game, the (Iterated) *Traveler's Dilemma* [7, 8, 12, 16]. In TD, each player has a large number of possible actions or moves. In the iterated context, many possible actions per round imply, for games of many rounds, an astronomic number of possible strategies overall. We are interested in the Iterated TD because its structure defies the usual prescriptions of the classical game theory insofar as what constitutes “optimal” play. We first define Travelers Dilemma, motivate its relevance and summarize most relevant prior art. We then pursue a detailed analysis of the “baseline” variant of the game that is based on a round-robin, many-round tournament that we have recently designed, implemented and run. In our analysis, we first summarize our main findings on the relative performances of various individual strategies with respect to the “bottom line” metric (which is, essentially, appropriately normalized dollar amount). We then focus on “team performance” of several selected groups of closely related strategies. We draw a number of interesting conclusions based on our extensive experimentation and analyzes of the individual and team performances. Finally, we outline some promising ways forward in this ongoing quest of gaining deeper insights into Iterated TD and other “far-from-zero-sum” iterated two-person games.

2 Traveler’s Dilemma

Traveler’s Dilemma was originally introduced in [4]. The motivation behind the game was to show the limitations of classical game theory [14], and in particular the notions of *individual rationality* that stem from game-theoretic notions of “optimal play” based on *Nash equilibria* [4, 5, 23]. The original version of TD, which we will treat as the “default” variant of this game, is defined as follows:

An airline loses two suitcases belonging to two different travelers. Both suitcases happen to be identical and contain identical items. The airline is liable for a maximum of \$100 per suitcase. The two travelers are separated so that they cannot communicate with each other, and asked to declare the value of their lost suitcase and write down (i.e., bid) a value between \$2 and \$100. If both claim the same value, the airline will reimburse each traveler the declared amount. However, if one traveler declares a smaller value than the other, this smaller number will be taken as the true dollar valuation, and each traveler will receive that amount along with a bonus/malus: \$2 extra will be paid to the traveler who declared the lower value and a \$2 deduction will be taken from the person who bid the higher amount. So, what value should a rational traveler (who wants to maximize the amount she is reimbursed) declare?

A tacit assumption in the default formulation of TD is that the bids have to be integers. That is, the *bid granularity* is \$1, as this amount is the smallest possible difference between two non-equal bids.

This default TD game has some very interesting properties. The game’s unique Nash equilibrium (NE), the action pair $(p, q) = (\$2, \$2)$, is actually rather bad for both players, under the usual assumption that the level of the players’ well-being is proportional to the dollar amount they individually receive. The choice of actions corresponding to NE results in a very low payoff for each player. The NE actions also minimize *social welfare*, which for us is simply the sum of the two players’ individual payoffs. However, it has been argued [4, 8, 11] 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, would not deviate from it (since unilaterally deviating from a Nash equilibrium presumably can be expected to result in decreasing one's own payoff). In contrast, the non-equilibrium pair of strategies ($\$100, \$100$) results in each player earning \$100, very near the best possible individual payoff for each player. Hence, the early studies of TD concluded that this game demonstrates a woeful inadequacy of the classical game theory, based on Nash (or similar notions of) equilibria [5]. Interestingly, it has been experimentally shown that humans (both game theory experts and laymen) tend to play far from the TD's only equilibrium, at or close to the maximum possible bid, and therefore fare much better than if they followed the classical game-theoretic approach [7].

We note that adopting one of the alternative notions of game equilibrium found in the "mainstream" literature does not appear to help, either. For example, it is argued in [12] that the action pair ($\$2, \$2$) is also the game's only *evolutionary equilibrium.* Similarly, seeking *sub-game perfect equilibria* (SGPE) [15] of Iterated TD is also not very promising: the set of a game's SGPEs is a subset of that game's full set of Nash equilibria in the mixed strategies. We also note that Iterated TD is structurally rather different from the *Centipede Game* [15] (a game brought to our attention by anonymous reviewers of an earlier paper of ours on ITD); in particular, the Centipede Game has multiple pure strategy Nash Equilibria (NE) and *infinitely many* NE in mixed strategies, whereas our game has a unique (pure strategy) NE and no additional mixed strategy NE.

We study the generalized iterated TD in [20]; there, the impact of the ratio between the bonus and the bid granularity on the game structure (the Nash equilibria, the Pareto optimal strategy pairs etc.) is studied in detail. For experimental studies on how human behavior changes with a change in the bonus, see Capra et al. [8] and Goeree and Holt [11].

3 The iterated travelers dilemma tournament

Our Iterated Traveler's Dilemma tournament has been inspired by, and is in form similar to, Axelrod's *Iterated Prisoner's Dilemma* tournament [3]. In particular, it is a round-robin tournament where each strategy plays against every other strategy as follows: each agent plays N matches against each other agent, incl. one's own "twin". A match consists of T rounds. The agents do not know T or N and cannot tweak their strategies with respect to the duration of the encounter. Similarly, the strategies are not allowed to use any other assumptions (such as, e.g., the general or specific nature of the opponent they are playing against in a given match). Indeed, the only data available to the learning and adaptable strategies in our "pool" of tournament participants (see below) is what they can learn and infer about the future rounds, against a given opponent, based on the bids and outcomes of the prior rounds of the current match against that opponent; no other knowledge of meta-knowledge of any kind is available to the strategies participating in our tournament.

In every round, each agent must select a valid bid. Thus, the *action space* of an agent in the tournament is $A = \{2, 3, \dots, 100\}$. Let C denote the set of strategies that play one-against-one matches with each other, that is, the set of agents competing in the tournament. Agents' actions are defined as follows: $x_t =$ the bid traveler x makes on round t; and $x_{nt} =$ the bid traveler x makes on round t of match n.

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 $M(x, y) = \sum_{t=1}^{T} R(x_t, y_t)$. In a sequence of matches, the reward received by agent x in the n^{th} match against y is denoted as $M_n(x, y)$. In order to make a reasonable baseline comparison, we use the same classes of strategies as in [9], ranging from rather simplistic to moderately complex. Summary of the strategy classes follows; for a more detailed description, see [9].

The "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 intervals: $\{2, 3, ..., 100\}$ and $\{99, 100\}$.

The "Simpletons": The second extremely simple class of strategies which choose the exact same dollar value in every round. The values we used in the tournament were $x_t = 2$ (the lowest possible), $x_t = 51$ ("median"), $x_t = 99$ (slightly below maximal possible; would result in maximal individual payoff should the opponent consistently play the highest possible action, which is $100), 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 famous name for a very simple, yet very effective, strategy for the iterated prisoner's dilemma [1–3, 18]. The idea behind *Tit-for-Tat* (TFT) is simple: cooperate on the first round, then "do to thy neighbor" (that is, opponent) exactly what he did to you on the previous round. We note that the baseline PD can be viewed as a special case of our TD, when the action space of each agent in the latter game is reduced to just two actions: $\{BidLow, BidHigh\}$. However, unlike iterated PD, even in the baseline version iterated TD as defined above, each agent has many actions at his disposal. In general, bidding high values in ITD can be viewed as an approximation of "cooperating" in IPD, whereas playing low values is an approximation of "defecting". We define several Tit-for-Tat-like strategies for ITD. These strategies can be roughly grouped into two categories. One are the *simple* TFT strategies bid value ϵ below the bid made by the opponent in the last round, where we restricted $\epsilon \in \{1, 2\}$. The second category are the *predictive* TFT strategies that compare whether their last bid was lower than, equal to, or higher than that of the other agent. Then a bid is made similar to the simple TFT strategies, 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 the agent strategizing along the TFT lines himself. In essence, the complex TFT strategies are attempting to predict the opponent's next bid based on the bids in the previous round and, given that prediction, they attempt to outsmart the opponent. A variant of TFT was the overall winner of a similar (but much smaller and simpler) iterated prisoner's dilemma round-robin tournament in [1]. Given the differences between traveler's and prisoner's dilemmae, we were very curious to see how well would various TFT-based strategies do in the iterated TD context.

"Mixed": The mixed strategies combine up to three pure strategies based on a probability mass. For each mixed strategy, a pure strategy $\sigma \in C$ is selected from one of the other strategies defined in the competition for each round according to a specified probability distribution (see *Table 1*). 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 allows for greater transparency when attempting to decipher the causes of a particular strategy's performance.

The notation in *Table 1* is *Mixed* followed by up to three $(Strategy, Probability)$ pairs, where each such pair represents a strategy and the probability that that strategy is selected for any given round. Simpleton strategies are represented simply by their bid, e.g. $(100, 20\%)$. Random strategies are represented by the letter R followed by their range, e.g. $(R[99, 100], 20\%)$. TFT strategies come in two varieties: simple and complex. In *Mixed* strategies, a Simple TFT used in the "mix" is represented by $TFT(y - n)$, where n is the value to bid below the opponent's bid (that is, the value of y). Complex TFTs used in a given "mix" are represented with L, E, and H indicators (denoting *Lower*, *Equal* and *Higher*), followed by the bid policy. Bid policies are based on either the opponent's previous bid (y) or this agent's own previous bid (x). Details can be found in [9]. An example (see *Table 1*) will hopefully clarify this somewhat cumbersome notation:

Mixed: $(L(y-g)E(x-g)H(x-g), 80\%); (100, 10\%); (2, 10\%)$ denotes a complex mixed strategy according to which an agent:

- plays a complex TFT strategy 80% of the time, in which it bids: (i) the opponent's last bid minus the granularity if this strategies last bid was *lower* than its opponent's; (ii) this strategies last bid minus the granularity if this strategies last bid was *equal* to its opponent's; and (iii) strategy's last bid minus the granularity if this strategies last bid was *higher* than its opponent's;
- 10% of the time simply bids \$100, that is, plays the *Simpleton \$100* strategy;
- the remaining 10% of the time bids \$2 (i.e., plays the *Simpleton \$2* strategy).

In the version of ITD reported in this paper, the value of bid granularity is $g = 1$ throughout.

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 that has been 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 recent tied-for-the-lead bucket wins. The strategy then bids the highest possible value strictly below the predicted opponent's bid. (If the opponent bids the lowest possible value, which in our baseline version of TD is \$2, then the deterministic bucket agent bids that lowest value, as well.) An instance of each tie breaking method above competed as a different bucket-based strategy in the tournament.

Buckets - Probability Mass Function: As with deterministic buckets, this strategy class counts instances of the opponent's bids and uses them to predict opponent's 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 assign greater weights to the more recent data than to the older data; we've selected a *retention rate* $(0 \leq \gamma \leq 1)$ to specify the speed of memory decay. We have entered into our tournament several instances of this strategy using the following rate of retention values γ: 1.0, 0.8, 0.5, and 0.2. The strategy bids the largest value strictly below the predicted value of the opponent's next bid (so, in the default version, it is the "one under" the predicted opponent's bid). We note that the "bucket" strategies based on probability mass buckets are quite similar to a learning model in [8].

Simple Trending: 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 θ. If the trend has a positive slope greater than θ, then the agent will continue to play the same bid it has been as the rewards are increasing. If the slope is negative and $|slope| > \theta$, then the system is trending toward the Nash equilibrium and, thus, the smaller rewards. In this case, the agent will attempt to entice the opponent to collaboration and will start playing \$100. Otherwise, the system of bidding and payouts is relatively stable and the agent will play the adversarial"one under" strategy that attempts to outsmart the other player. We have implemented instances of this strategy with an arbitrary θ of 0.5 and the following values of k: 3, 10, and 25, where larger values of k mean, trending is determined over a longer time-window. In particular, we have incorporated a simple explicit mechanism to push the player away from the "bad" NE: "simple trenders" share the adversarial philosophy of TFT as long as the rewards are high, but unilaterally move into collaboration-inviting, high-bidding behavior when the rewards are low (presumably, hoping that an adaptable opponent would follow the suit in the subsequent rounds).

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* [22] as a way of predicting the best action at time $(t + 1)$ based on the action selections and payoffs at times $[1, ..., t]$. This is similar to the Friend-or-Foe Q-learning method [13], 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 handful of meaningful classes, namely, based on just three states and three actions, as follows:

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

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

Recall that *actions* are defined for just a single time-step. The actual implementation treats the state as a collection of moves by the opponent over the last k rounds. We have decided to use $k = 5$ as an intuitively reasonable (but admittedly fairly arbitrary) value for k as it allows us to capture some history without data sizes becoming unmanageable. We are implementing this basic Q-learning algorithm with the learning rates of 0.8, 0.5 and 0.2.

Zeuthen Strategies: A Zeuthen Strategy [24] calculates the level of risk of each agent, and makes *concessions* accordingly. Risk is the ratio of loss from accepting the opponent's proposal vs. the loss of forcing the *conflict deal* (the deal made when no acceptable proposal can be found). While ITD is strictly speaking not a negotiation, one can still treat each bid (i.e. x_t and y_t) to be a proposal: if $x_t = i$, then agent x is proposing to agent y the pair $(i, i+1)$ as the next action pair. For TD, we consider the conflict deal to be the N.E. at $(\$2, \$2)$. Given the proposals of each agent, a risk comparison is done. An agent continues making the same bid as long as its risk is greater than or equal its opponent's. Otherwise, the agent will make the *minimal sufficient concession*: the agent adjusts its proposal so that (i) its risk is higher than opponent's risk and (ii) the opponent's utility increases as little as possible. Due to the peculiar structure of TD, it is possible that a "concession" actually leads to a loss of utility for the opponent. This, however, goes against the very notion of *making a concession.* Thus, we have implemented two Zeuthen strategies: one that allows counter-intuitive negative concessions and one that does not.

The metric that we use to evaluate relative performances of various strategies is essentially "the bottom line", that is, appropriately normalized dollar amount that a player would win if she engaged in the prescribed number of plays against a particular (fixed) opponent. More specifically, the metric U_1 below is the sum of all payoffs gained by an agent, normalized by the total number of rounds played and the maximum allowable reward:

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

where $R^\star$ 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 per each match, and $|C|$ is the number of competitors in the tournament. In experiments discussed in this paper, $R^\star = \$101$, $N = 100$, $T = 1000$ and $|C| = 38$. We note that some other candidate metrics for measuring performance in ITD, and analyzes of performances of various strategies w.r.t. those alternative metrics, can be found in [**?**].

4 Results for the individual strategies

The Traveler's Dilemma Tournament with which we have experimented involves a total of 38 competitors (distinct strategies), playing 100 head-to-head matches made of 1000 rounds each. The final rankings with respect to the (normalized) "bottom-line" metric U_1 are given in *Table 1* below.

We briefly summarize our main findings. First, the top three performers in our tournament turn out to be three "dumb" strategies that always bid high values. These three strategies are greedy in a very literal, simplistic sense, and are all utterly oblivious to what their opponents do – yet they outperform, and by a relatively considerable margin, the adaptable strategies such as the Q-learners and the "buckets". The strategy which always bids the maximum possible value ($100 in our case) and the strategy which always bids "one under" the maximum possible value are both outperformed by the strategy which randomly alternates between the two: "Random{99, 100}" picks to bid either $99 or $100 with equal probabilities, and without any consideration for the opponent's bids or previous outcomes.

The Zeuthen strategy that does not allow for negative "concessions" performs quite well, and is the highest performer among all "smart" and adaptable strategies in the tournament. The first work (as far as we are aware) that proposed the use of negotiation-inspired Zeuthen strategies in the game strategy for ITD context (see [9]) encountered some stern criticism on the grounds that playing an ITD-like game has little or nothing in common with multi-agent negotiation. However, ITD is a game ripe for collaboration among self-interested yet adaptable agents, and an excellent performance of a strategy such as Zeuthern-Positive, that is willing to sacrifice its short-term payoff in order to entice the other agent to more collaborative (i.e., systematic higher bidding) behavior in the subsequent rounds, validates our initial argument that highly collaborative, non-greedy (insofar as "outsmarting" the opponent) adaptable strategies should actually be expected to do quite well against a broad pool of other adaptable strategies. (See also our comments on the poor showing of TFT-based strategies, as well as the next section in which we summarize the group or team performances across main classes of adaptable strategies, where each such class is viewed as a team.)

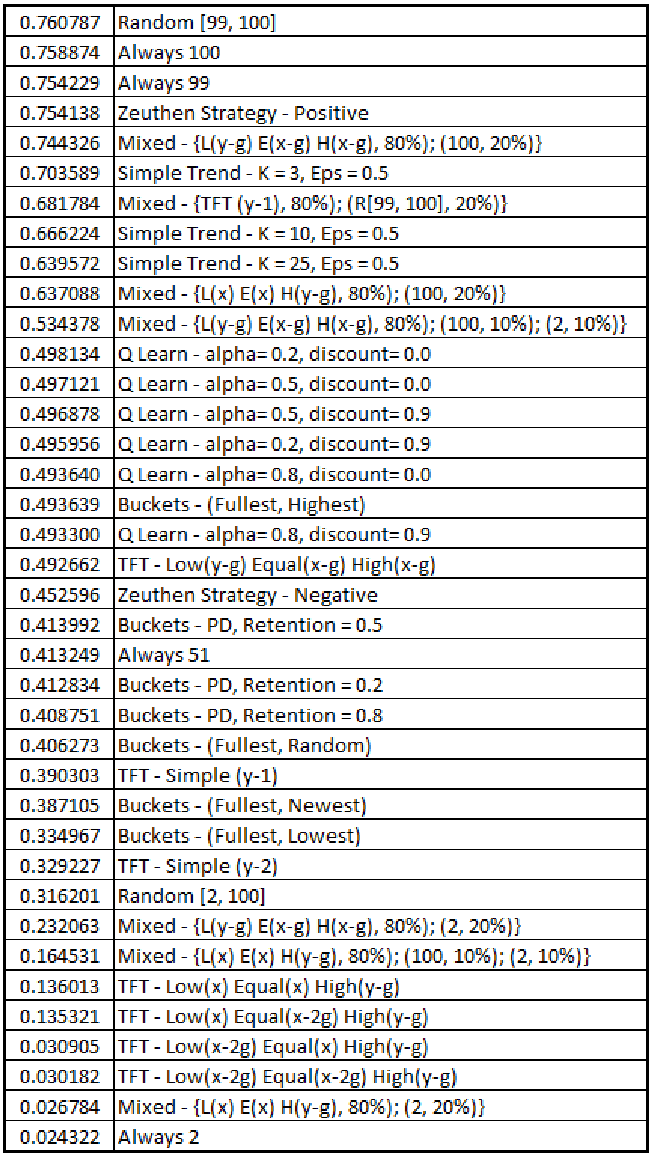
We find it rather interesting that (i) TFT-based strategies, in general, do fairly poorly, and (ii) their performances vary considerably depending on the exact details of the bid prediction method. In [9], it is reported that a relatively complex TFT-based strategy that, in particular, (a) makes a nontrivial model of the other agent's behavior and (b) "mixes in" some randomization, is among the top performers, whereas other TFT-based strategies exhibit mediocre (or worse) performance. In our analysis of individual performances, the top pure TFT based performer, which bids "one under" the opponent if the opponent made a lower bid than our TFT agent on the previous round, and lowers its own bid in the previous round in other scenarios, shows a mediocre performance with respect to the rest of the tournament participants. The best simple TFT strategy simply always bids "two below" the opponent's bid on the previous round. All other pure TFT-based strategies, simple and complex (i.e., predictive) alike, perform poorly, and some of the sophisticated predictive TFT strategies are among the very worst performers among all adaptable strategies in the tournament. This is in stark contrast to Axelrod's famous IPD tournament, where the original TFT strategy ended up the overall winner [1, 2].

We observe that the probabilistic bucket strategies perform decently overall, as long as the retention rate is strictly less than 1; with the retention rate of 1, guessing the opponent turns out to be abysmally poor and is by far the worst adaptable strategy in the tournament. We have therefore restricted our further analysis only to the bucket strategies with $\gamma < 1$ (and have eliminated the latter from the tournament table and further analysis). Probabilistic bucket based strategies with $\gamma < 1$ all perform fairly similarly to each other, that is, the exact value of the retention rate γ does not seem to make too much of a difference, *as long as* $\gamma < 1$.

Some additional observations on the individual performances of various strategies in our tournament include:

- Overall, there seems to be an overbid bias. The difference in payout is only two times the bonus/malus (which, heretofore, we will simply refer to as 'bonus' for brevity, with an understanding that it can come with a + or - sign). We'd expect that strategies that are more careful about bidding below their opponents would do better when the magnitude of the bonus is increased. Impact of the relative size of bonus, δ, with respect to the granularity of bids g, on TD game structure is the subject of [20].
- It is not clear whether *reinforcement learning* (at least in the manner in which we have captured it) really helps in becoming a top performer in the Iterated TD; in particular, the Q-learning based strategies show fairly mediocre performance.
- It is far from clear whether more complex models of the other agent really help insofar as bidding better, and hence performing better, in the long run. The best adaptable strategies (with an exception of Positive-Zeuthen) are the very simple ones: linear approximators ("simple trenders") and bucket-based ones, whereas Q-learners and elaborate adaptable TFT-inspired strategies in general perform considerably below the simple trenders and simple buckets.
- Not all TFT-based strategies in the TD are "born equal"; in fact, performance of different TFT variants tends to vary broadly with respect to all four of our metrics. This observation opens up interesting questions from the meta-learning [21] 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.

Last but not least, the single worst performer w.r.t. the normalized dollar-amount metric is, not surprisingly, the always-bid-lowest-possible strategy. This strategy can be viewed as the ultimate adversarial strategy that tries to always underbid, and hence outperform, the opponent – regardless of the actual payoff earned. (By bidding the lowest possible value, one ensures to never be out-earned by the opponent; some examples of such behavior from politics and economic markets can be readily found, and are discussed elsewhere.) Recall that "always bid $2" also happens to be the unique NE strategy that, according to classical, NE-based game theory, a rational agent that assumes a rational opponent should actually make this strategy her strategy of choice.



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

Table 1: Final rankings of individual strategies w.r.t. metric U_1

5 Team performance analysis

Perhaps the greatest conceptual problem with an experimental study of iterated games based on a round-robin tournament is the sensitivity of results with respect to the choice of participants in the tournament. While our choice of the final 38 competing strategies was made after a great deal of deliberation and careful surveying of prior art, we are aware that both absolute and relative performances of various strategies in the tournament might have been rather different had those strategies encountered a different set of opponents. The types of strategies we implemented (the *Randoms*, the *Simpletons*, Simple Trenders, Tit-For-Tat, Q-learners, etc.) have been extensively studied in the literature, and are arguably fairly "representative" of various relatively cognitively simple (and hence requiring only a modest computational effort) approaches to playing iterated PD, iterated TD and similar games. Within the selected classes of strategies, we admittedly made several fairly arbitrary choices of the critical parameters (such as, e.g., the learning rates in Q-learning). It is therefore highly desirable to be able to claim *robustness* of our findings irrespective of the exact parameter values in various parameterized types of strategies.

The "team performance" study summarized in this section has been undertaken for two main reasons. One, we'd like to reduce as much as possible the effects of some fairly arbitrary choices of particular parameter values for types of strategies. Two, given the opportunities for collaboration that Iterated TD offers, yet the complex structure of this game, we would like to see which pairs of strategy types, when matched against each other, *mutually reinforce* and therefore benefit each other; this analysis also applies to "self-reinforcement" as strategies of the same type are also matched up "against" each other. For example, we want to investigate how well the Q-learners get to do, with time, if playing Iterated TD "against" themselves.

Figure 1 summarizes relative performances of each strategy class against a given type of opponents, with the U1 score against the uniformly random strategy *Random*[2...100] used as the yardstick (hence normalized to 1). For each given "team", the contributions of individual strategies within the team all count equally. Here is how is the plot *Figure 1* to be read. Consider the second leftmost cluster of twelve adjacent bars, corresponding to 12 groups of strategies. The very leftmost one is the performance against the random strategy (in this particular case, it's the mix made of two Randoms vs. itself); the bar indicates that "mixed randoms vs. mixed randoms" score about 35% higher than against the yardstick, which is defined as normalized score against *Random*[2...100] alone. The next bar (second from the left) in the same group shows that the same mix of random strategies scores about 36% higher against the "mix" or team of four different "always bid the same value" strategies (see previous section) than against the yardstick *Random*[2...100]. The highest bar in this cluster shows that the mix of random strategies scores against the complex, predictive TFTs nearly two and a half times higher than against the uniformly random "yardstick" opponent, etc.

We now summarize our most interesting findings for this particular set of strategy classes or groups. Overall, Simple Trending seems to be the best general class of strategies against the given pool of opponents. The simple trenders are overall most consistent group of adaptable strategies: each of them performs quite well individually against (see *Table 1*). Therefore, after the simplistic "always bid very high", the simple trenders offer the best tradeoff between simplicity and underlying computational effort on one hand, and performance, on the other. Among the simple trenders, longer "memory window" of the recent previous runs leads to relatively poorer performance. One possible explanation is that, with a fairly long-term memory (such as for $K = 25$), the “uphill” and ”downhill” trends tend to average out, resulting in smaller slopes (in the absolute value) of the linear trend approximator, and hence slower adjustments in the simple trenders’ bidding.

Essentially adversarial in a game that is far from zero-sum and generally rewards cooperation, predictive TFT strategies “bury themselves into the ground” quite literally: their performance against themselves is among the worst of all team performance pairs, and is the “safest” way of getting to and then staying at the Nash equilibrium ($2, $2). In stark contrast, however, TFT-based strategies and Zeuthen strategies work well together; that is, Zeuthens’ initial “generosity” in order to encourage the opponent to move towards higher bids, in the long run, benefits TFT-based strategies when matched against the Zeuthens. Another interesting result about TFT strategies: when some randomization is added to a TFT-based strategy, esp. of a kind where very high bids are made in randomly selected rounds, the overall performance improves dramatically, as evidenced by the high scores of the group TFT-Mixed in comparison to both simple and complex “pure” TFT strategies. In fact, the mixed TFT strategies (that do include some randomization) are, together with simple trenders, the best “team” overall. In particular, mixed TFTs do very well when matched against any adaptable opponent in our tournament. In contrast, the predictive complex TFTs that don’t use any randomization are by far the worst “team” of strategies overall.

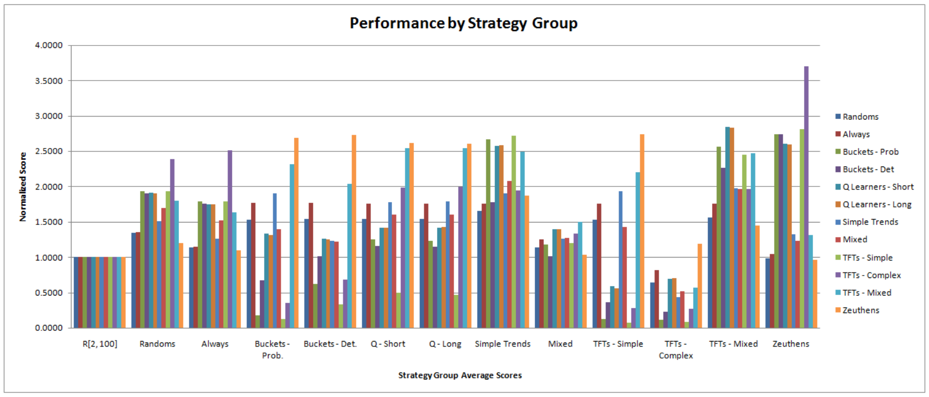


Figure 1: Relative group performances for the selected classes

Q-learners handle TFT based strategies quite well. Furthermore, Q-Learners and Simple Trenders rather nicely reinforce each other, i.e., when matched up “against” each other, both end up doing quite well. Similar *mutual reinforcement* of rewarding collaborative play can be observed when buckets (both probabilistic and deterministic) are matched up with Randomized TFTs and Zeuthens. One very striking instance of mutual reinforcement is what Zeuthens do for complex predictive TFTs (the variants without random bids), and in the process also for themselves, when matched against predictive TFTs. In contrast to these examples of mutual reinforcement, neither short- nor long-term memory Q-learners perform particularly impressively against themselves. We suspect that this in part is due to high sensitivity to the bid choices in the initial round; this sensitivity to initial behavior warrants further investigation. Moreover (see also *Table 1*), choice of the learning rate α seems to make a fairly small difference: all Q-learning based strategies show similar performances to each other against most types of opponents.

6 Summary and future work

We study Iterated Traveler's Dilemma, an interesting two-player non-zero sum game. We analyze this game by designing, implementing and analyzing a round robin tournament with 38 participating strategies. Our study of the performance of various strategies with respect to the "bottom-line" metric has corroborated that, for an iterated game whose structure is far from zero-sum, the traditional game-theoretic notions of rationality, based on the concept(s) of Nash equilibria, are rather unsatisfactory [9]. We have also learned that (i) common-sense unselfish greedy behavior ("bid high") generally tends to be rewarded in ITD, (ii) not all adaptable/learning strategies are necessarily successful, even against simple opponents, (iii) more complex models of an opponent's behavior may but need not result in better performance, (iv) exact choices of critical parameters may have a great impact on performance (such as with various bucket-based strategies) or hardly any impact at all (e.g., the learning rate in Q-learners), and (v) collaboration via *mutual reinforcement* between considerably different adaptable strategies appears to often be much better rewarded than self-reinforcement between strategies that are very much alike each other.

Our analysis also raises several interesting questions, among which we are particularly keen to further investigate (i) to what extent other variations of cognitively simple models of learning can be expected to help performance, (ii) to what extent complex models of the other agent really help an agent increase its payoff in the iterated play, and (iii) assuming that this phenomenon occurs more broadly than what we have investigated so far, what general lessons can be learned from the observed higher rewards for heterogeneous mutual reinforcement than for homogeneous self-reinforcement?

Last but not least, in order to be able to draw general conclusions less dependent on the selection of strategies in a tournament, we are also pursuing *evolving a population of strategies* akin to approach found in [6]. We hope to report new results along those lines in the near future.

Bibliography

[1] R. Axelrod. Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(1):3 –25, Mar. 1980.

[2] R. Axelrod. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

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

[4] K. Basu. The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395, May 1994.

[5] K. Basu. The traveler's dilemma. *Scientific American Magazine*, June 2007.

[6] B. Beaufils, J.-P. Delahaye, and P. Mathieu. Complete classes of strategies for the classical iterated prisoner's dilemma. In *Evolutionary Programming*, pages 33–41, 1998.

[7] T. Becker, M. Carter, and J. Naeve. Experts playing the traveler's dilemma. Technical report, Department of Economics, University of Hohenheim, Germany, Jan. 2005.

[8] C. M. Capra, J. K. Goeree, R. Gmez, and C. A. Holt. Anomalous behavior in a traveler's dilemma? *The American Economic Review*, 89(3):678–690, June 1999.

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

[10] Dasler, P., Tosic, P.: Playing challenging iterated two-person games well: A case study on iterated travelers dilemma. In: Proc. of WorldComp *Foundations of Computer Science* FCS'11; to appear (July 2011)

[11] J. K. Goeree and C. A. Holt. Ten little treasures of game theory and ten intuitive contradictions. *The American Economic Review*, 91(5):1402–1422, Dec. 2001.

[12] S. Land, J. van Neerbos, and T. Havinga. Analyzing the traveler's dilemma Multi-Agent systems project, 2008.

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

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

[15] M. Osborne. *An introduction to game theory.* Oxford University Press, New York, 2004.

[16] 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*, pages 194–200, 2009.

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

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

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

[20] P. T. Tosic, P. Dasler. "How To Play Well in Non-Zero Sum Games: Some Lessons from Generalized Traveler's Dilemma", in Proc. IEEE Active Media technology (AMT-2011), Springer LNCS series, 2011 (to appear)

[21] 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.

[22] C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

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

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