Stability of defection, optimisation of strategies and the limits of memory in the Prisoner's Dilemma.

Nikoleta E. Glynatsi Vincent A. Knight

Abstract

Memory one strategies are a set of iterated prisoners dilemma strategies that have been praised for their mathematical tractability and robustness. This manuscript explores *best response* memory one strategies and studies them as a multidimensional optimisation problem. Though extortionate memory one strategies have gained much attention, we prove that best response memory one strategies do not behave in an extortionate way. For memory one strategies to be evolutionary robust they need to be able to behave in a forgiving way. We also provide evidence that memory one strategies suffer from the limitation of their memory and can be out performed in multi agent interactions by longer memory strategies.

1 Introduction

The Prisoner's Dilemma (PD) is a two player game used in understanding the evolution of co-operative behaviour, formally introduced in [8]. Each player has two options, to cooperate (C) or to defect (D). The decisions are made simultaneously and independently. The normal form representation of the game is given by:

$$S_p = \begin{pmatrix} R & S \\ T & P \end{pmatrix} \quad S_q = \begin{pmatrix} R & T \\ S & P \end{pmatrix} \tag{1}$$

where S_p represents the utilities of the row player and S_q the utilities of the column player. The payoffs, (R, P, S, T), are constrained by equations (2) and (3). Constraint (2) ensures that defection dominates cooperation and constraint (3) ensures that there is a dilemma; the sum of the utilities for both players is better when both choose to cooperate. The most common values used in the literature are $(3, 1, 0, 5)$ [2].

$$T > R > P > S \tag{2}$$

$$2R > T + S \tag{3}$$

The PD is a one shot game, however it is commonly studied in a manner where the history of the interactions matters. The repeated form of the game is called the Iterated Prisoner's Dilemma (IPD) and in the 1980s, following the work of [3, 4] it attracted the attention of the scientific community. In [3] and [4], the first well

known computer tournaments of the IPD were performed. A total of 13 and 63 strategies were submitted respectively in the form of computer code. The contestants competed against each other, a copy of themselves and a random strategy. The winner was then decided on the average score a strategy achieved (not the total number of wins). The contestants were given access to the entire history of a match, however, how many turns of history a strategy would incorporate, refereed to as the *memory size* of a strategy, was a result of the particular strategic decisions made by the author.

The winning strategy of both tournaments was the strategy called Tit for Tat. Tit for Tat starts by cooperating and then mimics the last move of it's opponent, more specifically, it is a strategy that considers only the previous move of the opponent. These type of strategies are called *reactive* [18] and are a subset of so called *memory one* strategies. Memory one strategies similarly only consider the previous turn, however, they incorporate both players' recent moves.

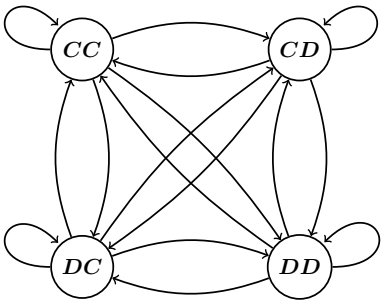
Several successful reactive and memory one strategies are found in the literature, such as Generous Tit For Tat [19] and Pavlov [16]. However, memory one strategies generated a small shock in the game theoretic community ([21] stated that "Press and Dyson have fundamentally changed the viewpoint on the Prisoner's Dilemma") when a curtain set of memory one strategies was introduced in [20]. These strategies are called zero determinate (ZD) and they chose their actions so that a linear relationship is forced between their score and that of the opponent. ZD strategies are indeed mathematically unique and are proven to be robust in pairwise interactions. Their true effectiveness in tournament interactions and evolutionary dynamics has been questioned by several works.

The purpose of this work is to consider a given memory one strategy in a similar fashion to [20], however whilst [20] found a way for a player to manipulate a given opponent, this work will consider a multidimensional optimisation approach to identify the best response memory one to a group of opponents. The main questions we raise are concerned with:

- A compact method of identifying the best response memory one strategy against a given set of opponents.
- The behaviour of a best response memory one strategy and whether it behaves extortionate, similar to [20].
- The factors that make a best response memory one strategy evolutionary robust.
- A well designed framework that allows the comparison of an optimal memory one strategy, and a more complex strategy that has a larger memory and was obtained through contemporary reinforcement learning techniques.
- An identification of conditions for which defection is known to be a best response; thus identifying environments where cooperation can not occur.

2 The utility

One specific advantage of memory one strategies is their mathematical tractability. They can be represented completely as a vector of $\mathbb{R}^4$. This originates from [18] where it is stated that if a strategy is concerned with only the outcome of a single turn then there are four possible 'states' the strategy could be in; CC, CD, DC, CC. Therefore, a memory one strategy can be denoted by the probability vector of cooperating after each of these states; $p = (p_1, p_2, p_3, p_4) \in \mathbb{R}^4_{[0,1]}$. In an IPD match two memory one strategies are moving from state to state, at each turn with a given probability. This exact behaviour can be modeled as a



$$M = \begin{bmatrix} p_1q_1 & p_1\left(-q_1+1\right) & q_1\left(-p_1+1\right) & \left(-p_1+1\right)\left(-q_1+1\right) \\ p_2q_3 & p_2\left(-q_3+1\right) & q_3\left(-p_2+1\right) & \left(-p_2+1\right)\left(-q_3+1\right) \\ p_3q_2 & p_3\left(-q_2+1\right) & q_2\left(-p_3+1\right) & \left(-p_3+1\right)\left(-q_2+1\right) \\ p_4q_4 & p_4\left(-q_4+1\right) & q_4\left(-p_4+1\right) & \left(-p_4+1\right)\left(-q_4+1\right) \end{bmatrix}$$

Figure 1: markov

stochastic process, and more specifically as a Markov chain (Figure 1). The corresponding transition matrix M of Figure 1 is given below,

The long run steady state probability v is the solution to $vM = v$. The stationary vector v can be combined with the payoff matrices of equation (1) and the expected payoffs for each player can be estimated without simulating the actual interactions. More specifically, the utility for a memory one strategy p against an opponent q, denoted as $u_q(p)$, is defined by,

$$u_q(p) = v \times (R, P, S, T). \tag{4}$$

In Theorem 1, the first theoretical results of the manuscript is presented, that is that $u_q(p)$ is given by a ratio of two quadratic forms [13]. To the authors knowledge this is the first work that has been done on the form of $u_q(p)$.

Theorem 1. *The expected utility of a memory one strategy $p \in \mathbb{R}^4_{[0,1]}$ against a memory one opponent $q \in \mathbb{R}^4_{[0,1]}$, denoted as $u_q(p)$, can be written as a ratio of two quadratic forms:*

$$u_q(p) = \frac{\frac{1}{2}pQp^T + cp + a}{\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}}, \tag{5}$$

where $Q, \bar{Q} \in \mathbb{R}^{4\times4}$ are hollow matrices defined by the transition probabilities of the opponent q_1, q_2, q_3, q_4 as follows:

$$\boldsymbol{Q} = \begin{bmatrix} \mathbf{0} & \mathbf{-\left(q_1-q_3\right)\left(q_2-5q_4-1\right)} & \mathbf{q_3\left(q_1-q_2\right)} & \mathbf{-5q_3\left(q_1-q_4\right)} \\ \mathbf{-\left(q_1-q_3\right)\left(q_2-5q_4-1\right)} & \mathbf{0} & \mathbf{\left(q_2-q_3\right)\left(q_1-3q_4-1\right)} & \mathbf{\left(q_3-q_4\right)\left(5q_1-3q_2-2\right)} \\ \mathbf{q_3\left(q_1-q_2\right)} & \mathbf{\left(q_2-q_3\right)\left(q_1-3q_4-1\right)} & \mathbf{0} & \mathbf{3q_3\left(q_2-q_4\right)} \\ \mathbf{-5q_3\left(q_1-q_4\right)} & \mathbf{\left(q_3-q_4\right)\left(5q_1-3q_2-2\right)} & \mathbf{3q_3\left(q_2-q_4\right)} & \mathbf{0} \end{bmatrix}, \tag{6}$$

$$\bar{\boldsymbol{Q}} = \begin{bmatrix} \mathbf{0} & \mathbf{-\left(q_1-q_3\right)\left(q_2-q_4-1\right)} & \mathbf{\left(q_1-q_2\right)\left(q_3-q_4\right)} & \mathbf{\left(q_1-q_4\right)\left(q_2-q_3-1\right)} \\ \mathbf{-\left(q_1-q_3\right)\left(q_2-q_4-1\right)} & \mathbf{0} & \mathbf{\left(q_2-q_3\right)\left(q_1-q_4-1\right)} & \mathbf{\left(q_1-q_2\right)\left(q_3-q_4\right)} \\ \mathbf{\left(q_1-q_2\right)\left(q_3-q_4\right)} & \mathbf{\left(q_2-q_3\right)\left(q_1-q_4-1\right)} & \mathbf{0} & \mathbf{-\left(q_2-q_4\right)\left(q_1-q_3-1\right)} \\ \mathbf{\left(q_1-q_4\right)\left(q_2-q_3-1\right)} & \mathbf{\left(q_1-q_2\right)\left(q_3-q_4\right)} & \mathbf{-\left(q_2-q_4\right)\left(q_1-q_3-1\right)} & \mathbf{0} \end{bmatrix}. \tag{7}$$

c and $\bar{c} \in \mathbb{R}^{4\times1}$ are similarly defined by:

$$c = \begin{bmatrix} q_1\,(q_2 - 5q_4 - 1) \\ -\,(q_3 - 1)\,(q_2 - 5q_4 - 1) \\ -q_1q_2 + q_2q_3 + 3q_2q_4 + q_2 - q_3 \\ 5q_1q_4 - 3q_2q_4 - 5q_3q_4 + 5q_3 - 2q_4 \end{bmatrix}, \tag{8}$$

$$\bar{c} = \begin{bmatrix} q_1\,(q_2 - q_4 - 1) \\ -\,(q_3 - 1)\,(q_2 - q_4 - 1) \\ -q_1q_2 + q_2q_3 + q_2 - q_3 + q_4 \\ q_1q_4 - q_2 - q_3q_4 + q_3 - q_4 + 1 \end{bmatrix}. \tag{9}$$

and $a = -q_2 + 5q_4 + 1$ *and* $\bar{a} = -q_2 + q_4 + 1$.

The proof of Theorem 1 is given in Appendix.

Numerical simulations have been carried out to validate the formulation of $u_q(p)$ as a quadratic ratio, a data set is available at. Figure 2 shows that the formulation successfully captures the simulated behaviour. The simulated utility, which is denoted as $U_q(p)$, has been calculated using [1] an open source research framework for the study of the IPD. The project is described in [14]. All of the aforementioned simulated results have been estimated using [1].

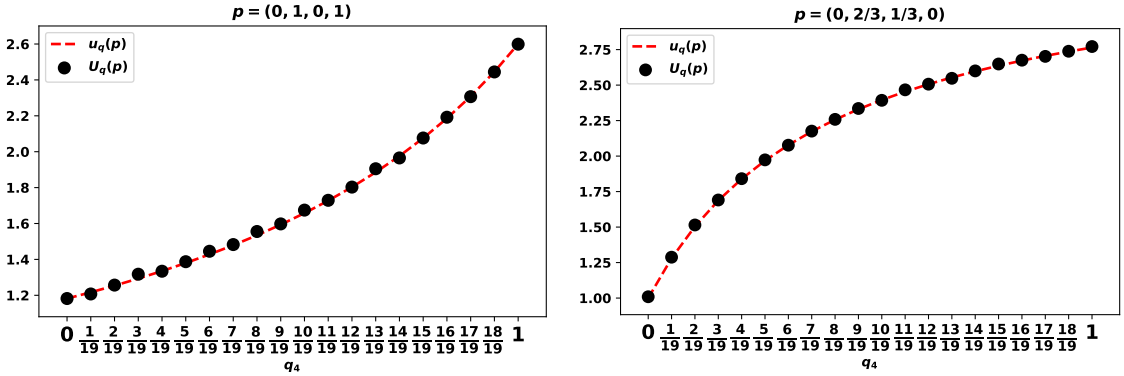


Figure 2: Differences between simulated and analytical results for $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, q_4)$.

Theorem 1 can be extended to consider multiple opponents. The IPD is commonly studied in tournaments and/or Moran Processes where a strategy interacts with a number of opponents. The payoff of a player in such interactions is given by the average payoff the player received against each opponent. More specifically the expected utility of a memory one strategy against a N number of opponents is given by Theorem 2.

Theorem 2. *The expected utility of a memory one strategy* $p \in \mathbb{R}^4_{[0,1]}$ *against a group of opponents* $q^{(1)}, q^{(2)}, \ldots, q^{(N)}$, *denoted as* $\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p)$ *is given by:*

$$\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) = \frac{1}{N}\frac{\sum_{i=1}^{N}(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}) \prod_{\substack{j=1 \\ j\neq i}}^{N}(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)})}{\prod_{i=1}^{N}(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)})}. \tag{10}$$

Theorem 2 is validated against the strategies used in [21], Figure 3. The list of strategies from [21] alonside their original reference from the are given by Table 4 in the Appendix.

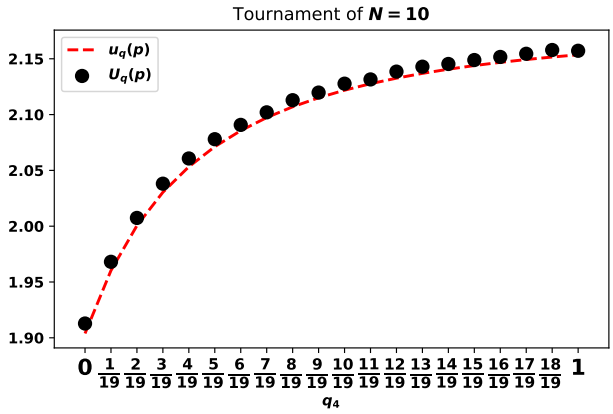


Figure 3: Results of memory one strategies against the strategies in Table 4.

Furthermore, using the same list of strategies the hypothesis the utility against a group of strategies could be captured by the utility against the mean opponent, thus:

$$\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) = u_{\frac{1}{N}\sum_{i=1}^{N} q^{(i)}}(p), \tag{11}$$

has been checked. The hypothesis fails and numerical evidence are given by Figure 4.

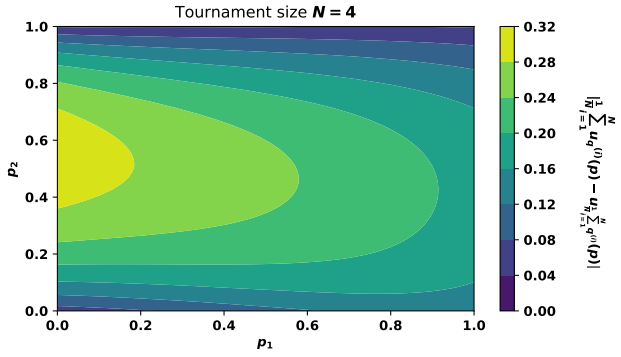


Figure 4: The difference between the average utility and against the utility against the average player of the strategies in [21]. A positive difference indicates that the condition (11) does not hold.

Theorem 2 can be used to identify best responses in the case of memory one strategies. In the following sections several theoretical results are presented and the advantages of analytical formulation of become evident.

3 Best responses to memory one players

A *best response* is the strategy which corresponds to the most favourable outcome. A best response memory one strategy corresponds to the p^* for which $\sum u_{q^{(i)}}(p^*)$ for $i \in \{1, \ldots, N\}$ is maximized. This is considered as a multi dimensional optimisation problem where the decision variable is the vector p, the solitary constraint is that $p \in \mathbb{R}^4_{[0,1]}$ and the objective function is a sum of quadratic ratios. The optimisation problem is formally given by (12).

$$\begin{aligned} &\max_p : \sum_{i=1}^{N} {u_q}^{(i)}(p) \\ &\text{such that} : p \in \mathbb{R}_{[0,1]} \end{aligned} \tag{12}$$

Optimising this particular ratio of quadratic forms is not trivial. It can be verified empirically for the case of a single opponent that there exist at least one point for which the definition of concavity does not hold. Though [5, 7] are also concerned with a non concave ratios of quadratic forms, in both the numerator and the denominator of the fractional problem were concave or that the denominator was greater than zero; both assumptions fail here. These results are established in Theorem 3.

Theorem 3. *The utility of a player p against an opponent q, $u_q(p)$ given by (5), is not concave. Furthermore neither the numeration or the denominator of (5), are concave.*

The non concavity of $u(p)$ indicates multiple local optimal points. The approach taken here is to introduce a compact way of constructing the candidate set of all local optimal points. Once the set is defined the point that maximises (10) corresponds to the best response strategy, this approach transforms the continuous optimisation problem in to a discrete problem. The problem considered is a bounded because $p \in \mathbb{R}^4_{[0,1]}$. The candidate solutions will exist either at the boundaries of the feasible solution space, or within that space. The method of Lagrange Multipliers [6] and Karush-Kuhn-Tucker conditions [9] are based on this.

These lead to Lemma 4 which presents the best response memory one strategy to a group of opponents.

Lemma 4. *The optimal behaviour of a memory one strategy player $p^* \in \mathbb{R}^4_{[0,1]}$ against a set of N opponents $\{q^{(1)}, q^{(2)}, \ldots, q^{(N)}\}$ for $q^{(i)} \in \mathbb{R}^4_{[0,1]}$ is established by:*

$$p^* = \text{argmax}(\sum_{i=1}^{N} u_q(p)),\ p \in S_q.$$

The set S_q is defined as all the possible combinations of:

- *Any one, two, three, four or non of the transition probabilities of p are 0,*
- *while any one, two, three, four or non of the transition probabilities of p are 1,*
- *while any one, two, three, four or non of the transition probabilities of p are the roots of $\frac{d}{dp}\sum_{i=1}^{N} u_q(p)$.*

The derrivate $\frac{d}{dp}\sum_{i=1}^{N} u_q(p)$ is given by:

$$\frac{d}{dp}\sum_{i=1}^{N} {u_q}^{(i)}(p) =$$

$$= \frac{(\sum_{i=1}^{N} {Q_N^{(i)}}' \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)} + \sum_{i=1}^{N} {Q_D^{(i)}}' \sum_{\substack{j=1 \\ j\neq i}}^{N} Q_N^{(i)} \prod_{\substack{j=1 \\ j\neq\{i,j\}}}^{N} Q_D^{(i)}) \times \prod_{i=1}^{N} Q_D^{(i)} - (\sum_{i=1}^{N} {Q_D^{(i)}}' y - vk \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)}) \times (\sum_{i=1}^{N} Q_N^{(i)} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)})}{(\prod_{i=1}^{N} Q_D^{(i)})^2} \tag{13}$$

Proof in the Appendix.

Constructing the subset S_q is analytical possible. The points for any or none of $p_i \in \{0, 1\}$ for $i \in 1, 2, 3, 4$ are trivial. Finding the roots of the partial derivates $\frac{d}{dp}\sum_{i=1}^{N} u_q(p)$ is feasible using resultant theory. Resultant theory [12] allow us to solve systems of polynomials by the calculation of a resultant. However, for large systems these quickly become intractable. Because of that no further analytical consideration is given to problems described here.

So far we have provided an analytical formulation that can estimate best response memory one strategies against a number of opponents. This will be revisited and solved numerically in Section [**?**]. In the following subsection we present a theoretical results which has been possible due to the formulation discussed here.

3.1 Stability of defection

Defection is known to be the dominant action in the PD and it can be proven to be the dominant strategy for the IPD for given environments. Even so, several works have proven that cooperation emerges in the IPD and many studies focus on the emergence of cooperation.

An immediaty results from the our formulation of best response memory one strategies is that we can provide an identification of conditions for which defection is known to be a best response; thus identifying environments where cooperation can not occur.

The results are presented in Lemma 5.

Lemma 5. *In a tournament of N players where $q^{(i)} = (q_1^{(i)}, q_2^{(i)}, q_3^{(i)}, q_4^{(i)})$ defection is a best response if the transition probabilities of the opponents satisfy the condition:*

$$\sum_{i=1}^{N} (c^{(i)T} \bar{a}^{(i)} - \bar{c}^{(i)T} a^{(i)}) \leq 0 \tag{14}$$

Proof. For defection to be evolutionary stable the derivative of the utility at the point $p = (0, 0, 0, 0)$ must be negative. This would indicate that the utility function is only declining from that point onwards.

Substituting $p = (0, 0, 0, 0)$ in equation (13) which gives:

$$\sum_{i=1}^{N} (c^{(i)T} \bar{a}^{(i)} - \bar{c}^{(i)T} a^{(i)}) \prod_{\substack{j=1 \\ j\neq i}}^{N} (\bar{a}^{(i)})^2 \tag{15}$$

The second term $\prod_{\substack{j=1 \\ j \neq i}}^{N} (\bar{a}^{(i)})^2$ is always positive, however, the sign of the first term $\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)})$ can vary based on the transition probabilities of the opponents. Thus the sign of the derivative is negative if and only if $\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \leq 0$. □

4 Numerical experiments

In this section best responses are explored numerically. Best responses are estimated using the Bayesian optimisation algorithm, which is a global optimisation algorithm, introduced in [17], that has proven to outperform many other popular algorithms [11]. Differential evolution had also been considered, however it was not chosen due to Bayesian being computationally faster.

Bayesian optimisation tries to find values for the decision variables for which the utility of a player is maximised, over a given time of calls. Consider the problem of (12) where $N = 2$ and p^* is being estimated. Figure 5 illustrates the change of the utility function over number of calls. The default number of iterations that have been used in this work is 60. After the 60 calls the convergence of the utility is checked. If it is not then the calls are increased by 20, this step is then repeated until utility reaches convergence.

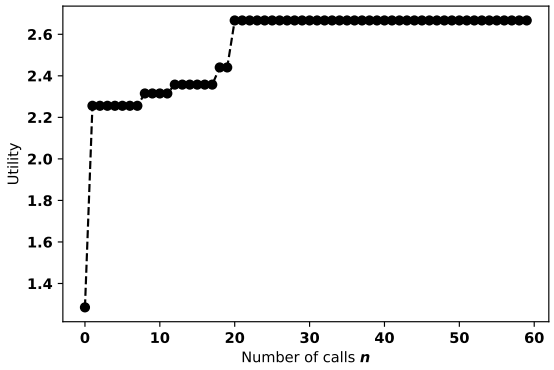


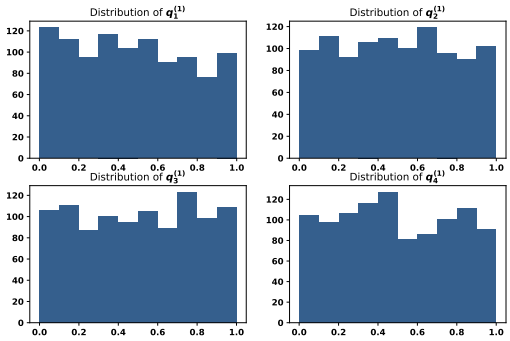
Figure 5: Utility over time of calls using Bayesian optimisation. The opponents are $q^{(1)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $q^{(2)} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

We will be Bayesian optimisation not only to estimate best response memory one strategies but a series of cases are going to be explored. Evolutionary memory one and longer memory best responses are also considered here. This is done so we can gain a better understanding of memory one strategies, their behaviour, robustness and limitations.

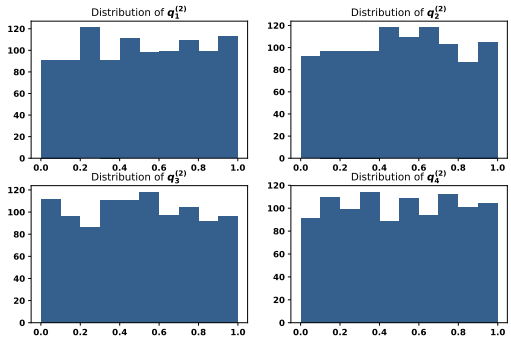
4.1 Best response memory one strategies for $N = 2$

The first case considered is that of best response memory one strategies in tournaments in order to understand whether best respones behave in an extortionate way. A large data set of best response memory one strategies when $N = 2$ has been generated and is available here. $N = 2$ has been chosen as it's the smallest size of a tournament.

The data set contains a total of 1000 trials and 1000 different best responses. For each trial a set of 2 opponents is randomly generated, the memory one best response against them is estimated and it's behaviour is being recorded. Though the probabilities q_i of the opponents are randomly generated, Figures 6a and 6b they are uniformly distributed over the trial. Thus, the space full space of possible opponents has been covered.



(a) Distributions of first opponents' probabilities.



(b) Distributions of second opponents' probabilities.

It was briefly discussed in Section 1 that ZD strategies have received praised for their robustness against a single opponent. By forcing a linear relationship between the scores ZD strategies will always manage to receive a higher payoff than their opponents. In tournament setting the winner is defined by the average score a strategy received, thus winning against your opponent at each interaction does not guaranty a strategy's overall win. This manuscript argues that by trying to exploit their opponents ZD strategies suffer in multi opponent interaction where the payoffs matter. Compared to ZD best response memory one strategies utilise their behaviour to gain the most from their interactions.

In [Knight 2019] the authors provied a method of measuring the extortionate behaviour of a strategy based on it's estimated probabilities. The method estimates the error of behaving as a ZD strategy defined as SSerror. The SSerror method is applied on the data set that has been generated in order to gain an understanding whether best reponses memory one strategies behave in an extortionate way. The distribution of the SSerror is shown in Figure 9 and a statistics summary by Table 2. Only the 30% of the best responses have a SSerror less than a 0.10 and a positive measure of skewness (= 1.96), indicates that the distribution is skewed to the right.

	SSerror
count	1023.000000
mean	0.322116
std	0.388394
min	0.000000
25%	0.076284
30%	0.102444
35%	0.128244
50%	0.167952
max	2.470588

Table 1: Summary statistics SSerror

Overall only a very small percentage of the best responses seem to behaving in an extortionate way, confirming our original hypothesis. To gain the most of our environment you would avoid being extortionate. The following section the second experiment and the result of memory one best responses in evolutionary dynamics are presented.

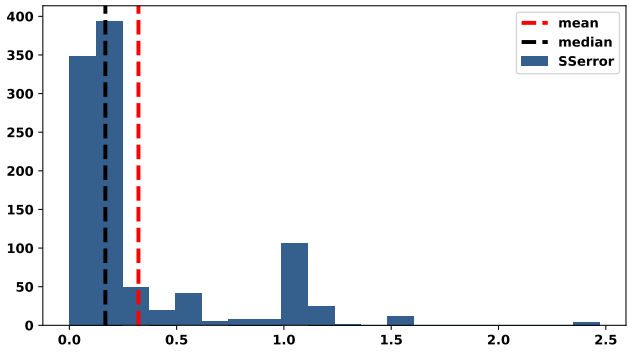


Figure 7: Distribution of sserrors for memory one best responses, when $N = 2$.

4.2 Memory one best responses in evolutionary dynamics

As briefly discussed in Section 2, the IPD is commonly studied in Moran Processes, and generally in evolutionary processes. In evolutionary processes, a finite population is assumed where the strategies that compose the population can adapt and change their behaviour based on the outcomes of their interactions at each turn. A key in successfully being an evolution stable strategy (ESS) is self interactions. An ESS must be a best response not only to the opponents in the population, but also it has to be a best response to it's self.

Self interactions can easily be incorporated in the formulation that has been used in this paper. The utility of a memory one strategy in an evolutionary setting is given by,

$$\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) + u_p(p). \tag{16}$$

and respectively the optimisation problem of (12) is now re written as,

$$\begin{aligned} &\max_p : \frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) + u_p(p) \\ &\text{such that} : p \in \mathbb{R}_{[0,1]} \end{aligned} \tag{17}$$

We suggest an algorithmic approach for estimating the evolutionary best response memory one strategy (evo) called the algorithm *best response dynamics*, given by Algorithm 1.

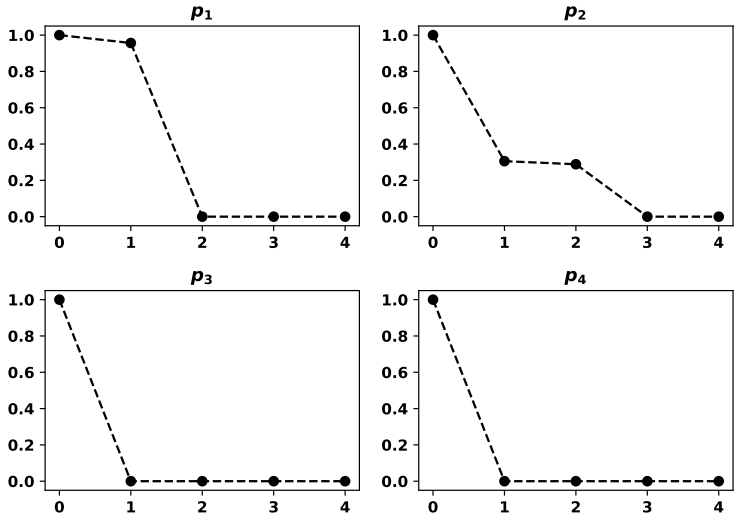


Figure 8: Best response dynamics with $N = 2$. More specifically, for $q^{(1)} = (0.2360, 0.1031, 0.3960, 0.1549)$ and $q^{(2)} = (0.0665, 0.4015, 0.9179, 0.8004)$.

The algorithm starts by setting an initial solution $p^{(1)} = (1, 1, 1, 1)$. Though a more optimal set of probabilities could be considered for an initial solution, it has been shown that the algorithm converges to same optimal solution even with more optimal starts. Once the initial solution is given then $p^{(2)}$ is estimated as the best response memory one to the N opponents plus to $(1, 1, 1, 1)$. The current solution then changes to be the new best response memory one, $p^{(2)}$. In the next step, $p^{(3)}$ is the best response memory one to the N opponents plus to $p^{(2)}$. This is repeated until the same algorithms returns a solution that has already been evaluated. This is done in order to avoid cycles. Figure 8 illustrates a numerical examples. The algorithm stops once it evaluates the same point again.

$p^{(t)} \leftarrow (1, 1, 1, 1)$;
while $p^{(t)}$ *not converged* **do**
  $p^{(t+1)} =$
  $\text{argmax} \frac{1}{N} \sum_{i=1}^{N} {u_q}^{(i)}(p^{(t+1)}) +$
  $u_p^{(t)}(p^{(t+1)})$;
end

Algorithm 1: Best response dynamics Algorithm

For each pair of opponents, from the data set described in Section 4.1 we have also recorder the evo strategy. Thus, a total of 1000 different evos have been estimated. Similarly, to previous results, the evos do not appear to behave in an extortionate way either. Only 30% have an SSerror less than a 0.1 and the distribution of error has a large positive skewness = 3.33 indicating a longer tail to the right. More and more strategies behave less and less extortionate (Figure 9 and Table 2).

The majority of neither best response and evolutionary best response memory one strategies are behaving in extortionate way. In order to understand the difference between the two set of strategies we consider the distributions of their respective transition probabilities, Figure 10. Though there is no significant difference between the medians of each distribution, Table 3, it is evident from Figure 10 that there is variation in the behaviors.

Except from the case of p_3:

- If a strategy manages to get away with a defection, and receives a temptation payoff, the strategy will try another defection in the next round. That is true for both best response memory one and evo.

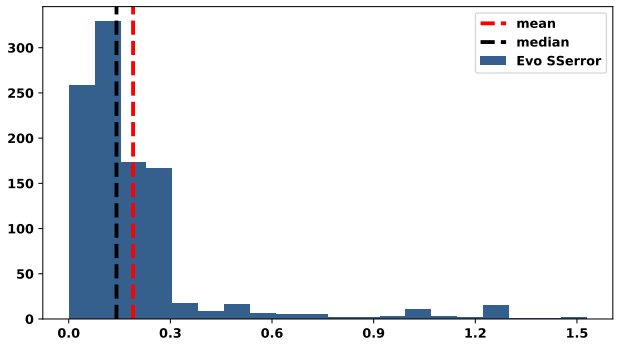


Figure 9: Distribution of sserrors for memory one best responses, when $N = 2$

	Evo SSerror
count	1023.000000
mean	0.189929
std	0.219760
min	0.000000
25%	0.075918
30%	0.097995
35%	0.113404
50%	0.140975
max	1.529412

Table 2: Summary statistics SSerror

	Memory one Median	Evo Median	p-values
Distribution p_1	0.0	0.000000	0.0
Distribution p_2	0.0	0.174359	0.0
Distribution p_3	0.0	0.000000	0.0
Distribution p_4	0.0	0.000000	0.0

Table 3: A non parametric test, Wilcoxon Rank Sum, has been performed to tests the difference in the medians. A non parametric test is used because is evident that out data are skewed.

For cases of p_1, p_2 and p_4:

- After the CC state where a mutual cooperation has occur, best responses memory one strategies are either going to cooperate of defect, with very high probabilities. The are more likely however, to defect and break the cycle of cooperation. In comparison evos are more stochastic. The two extreme peaks do not appear in the case of evos, and overall there is a slightly bigger insensitive for cooperation from evos.
- In cases of CD, a state that a strategy has been tricked, best response memory one strategies are very quick at punishing the strategy, they are interacting with. On the other hand evos are slightly more likely to cooperate again, to forgive their opponent. This could be a result of self interaction and evos and trying to not punish themselves.
- Finally, in cases that a mutual defection has occur, evos probability cooperating again is practically zero. Evos are not forgiving after a mutual defection, whereas best response strategies are.

The difference of between the transition probabilities of evos and best responses, at each trial, is also given, Figure 11 to further our results.

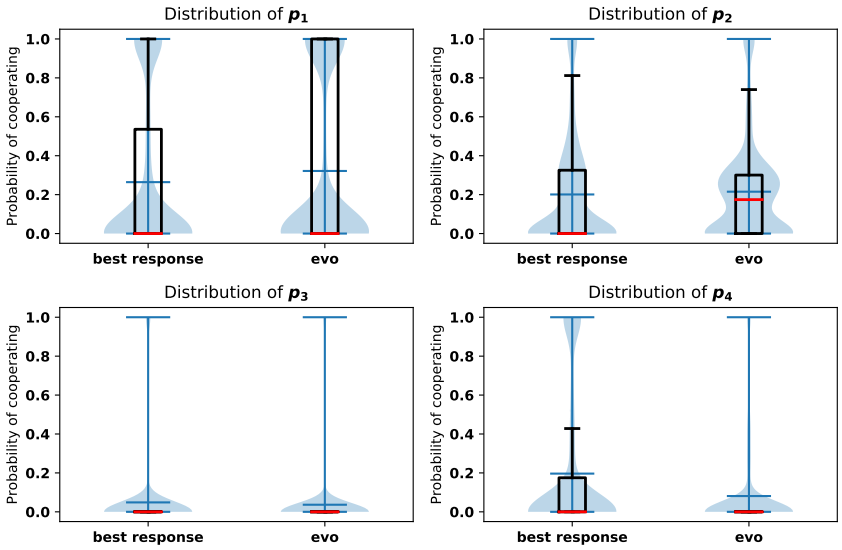


Figure 10: Distributions of p for both best response and evo memory one strategies.

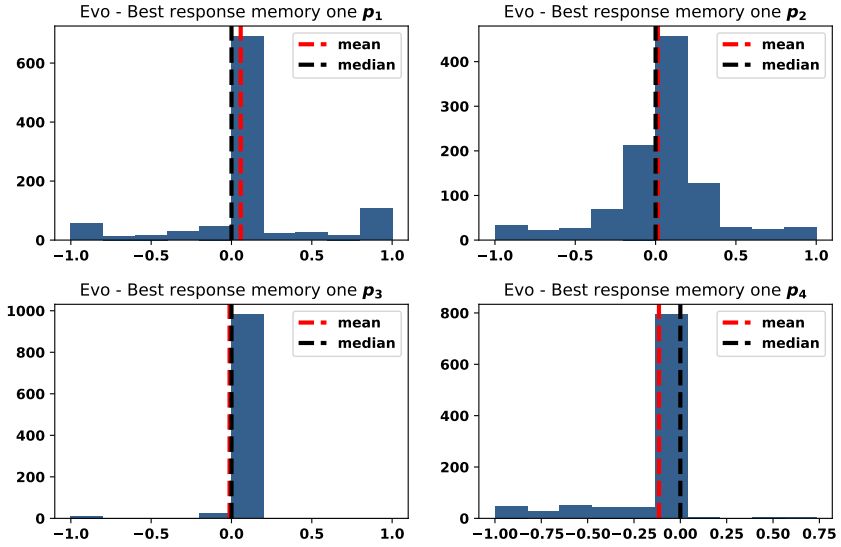


Figure 11: Differences of $p_(i)$ for $i \in 1, 2, 3, 4$ between evos and best responses at each trial.

5 Longer memory best response

The third and final case considered in this paper focuses on proving that short memory strategies have limitations. In this section we introduce several empirical results that show that more complex strategies can indeed perform better in cases of $N = 2$. This is achieved by comparing the performance of an optimised memory one strategy to that of a trained long memory one.

The longer memory strategy we have chosen is a strategy called Gambler, introduced and discussed in [10]. A Gambler strategy makes probabilistic decisions based on the:

- Opponent's first moves, n_1.
- Opponent's last moves, m_1.
- Player's last moves, m_2

This manuscript considers a Gambler($n_1 = 2, m_1 = 1, m_2 = 1$). By considering the opponent's first more, the opponents last move and two of our own, there are only 16 possible outcomes that can occur.

16 {
$C, C, C, C \rightarrow ?$

$C, C, C, D \rightarrow ?$

$\cdots$

$D, D, D, C \rightarrow ?$

$D, D, D, D \rightarrow ?$
}

Bayesian optimisation is used to estimate the cooperating probabilities for a Gambler after each possible state plus the probability of starting with a cooperation. Gambler's utility is the estimated using [1] as the tournament result against two random opponents. The tournament is run for. A total of 70 trails have been recorded and the data can be found here.

6 Conclusion

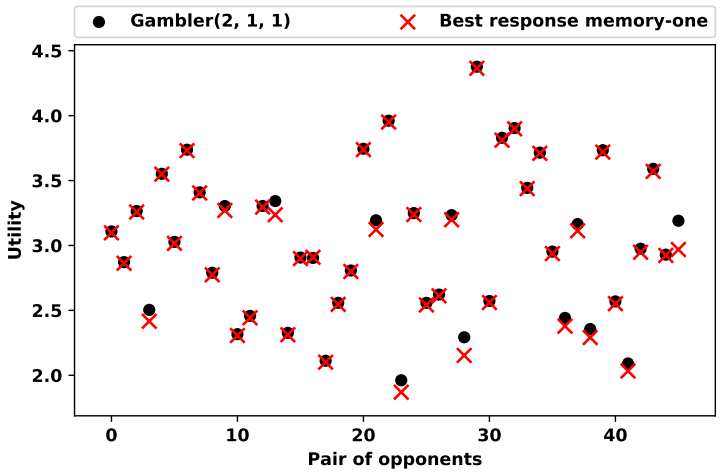
References

[1] The Axelrod project developers . Axelrod: ¡release title¿, April 2016.

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

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

[4] Robert Axelrod. More effective choice in the prisoner's dilemma. *The Journal of Conflict Resolution*, 24(3):379–403, 1980.



[5] Amir Beck and Marc Teboulle. A convex optimization approach for minimizing the ratio of indefinite quadratic functions over an ellipsoid. *Mathematical Programming*, 118(1):13–35, 2009.

[6] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[7] Hongyan Cai, Yanfei Wang, and Tao Yi. An approach for minimizing a quadratically constrained fractional quadratic problem with application to the communications over wireless channels. *Optimization Methods and Software*, 29(2):310–320, 2014.

[8] Merrill M. Flood. Some experimental games. *Management Science*, 5(1):5–26, 1958.

[9] Giorgio Giorgi, Bienvenido Jiménez, and Vicente Novo. Approximate karush—kuhn—tucker condition in multiobjective optimization. *J. Optim. Theory Appl.*, 171(1):70–89, October 2016.

[10] Marc Harper, Vincent Knight, Martin Jones, Georgios Koutsovoulos, Nikoleta E. Glynatsi, and Owen Campbell. Reinforcement learning produces dominant strategies for the iterated prisoners dilemma. *PLOS ONE*, 12(12):1–33, 12 2017.

[11] Donald R Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.

[12] Gubjorn Jonsson and Stephen Vavasis. Accurate solution of polynomial equations using macaulay resultant matrices. *Mathematics of computation*, 74(249):221–262, 2005.

[13] Jeremy Kepner and John Gilbert. *Graph algorithms in the language of linear algebra*. SIAM, 2011.

[14] Vincent Knight, Owen Campbell, Marc Harper, Karol Langner, James Campbell, Thomas Campbell, Alex Carney, Martin Chorley, Cameron Davidson-Pilon, Kristian Glass, Tomáš Ehrlich, Martin Jones, Georgios Koutsovoulos, Holly Tibble, Müller Jochen, Geraint Palmer, Paul Slavin, Timothy Standen, Luis Visintini, and Karl Molden. An open reproducible framework for the study of the iterated prisoner's dilemma. 1(1), 2016.

[15] Jiawei Li, Philip Hingston, and Graham Kendall. Engineering design of strategies for winning iterated prisoner's dilemma competitions. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(4):348–360, 2011.

[16] Frederick A Matsen and Martin A Nowak. Win–stay, lose–shift in language learning from peers. *Proceedings of the National Academy of Sciences*, 101(52):18053–18057, 2004.

	Name	Memory one representation	Reference
1	Cooperator	$(1, 1, 1, 1)$	[2]
2	Defector	$(0, 0, 0, 0)$	[2]
3	Random	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	[2]
4	Tit for Tat	$(1, 0, 1, 0)$	[2]
5	Grudger	$(1, 0, 0, 0)$	[15]
6	Generous Tit for Tat	$(1, \frac{1}{3}, 1, \frac{1}{3})$	[19]
7	Win Stay Lose Shift	$(1, 0, 0, 1)$	[16]
8	ZDGTFT2	$(1, \frac{1}{8}, 1, \frac{1}{4})$	[21]
9	ZDExtort2	$(\frac{8}{9}, \frac{1}{2}, \frac{1}{3}, 0)$	[21]
10	Hard Joss	$(\frac{9}{10}, 0, \frac{9}{10}, 0)$	[21]

Table 4: List of strategies used in the tournament described in [21].

[17] J. Močkus. On bayesian methods for seeking the extremum. In G. I. Marchuk, editor, *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.

[18] Martin Nowak and Karl Sigmund. Game-dynamical aspects of the prisoner's dilemma. *Applied Mathematics and Computation*, 30(3):191–213, 1989.

[19] Martin Nowak and Karl Sigmund. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Applicandae Mathematicae*, 20(3):247–265, 1990.

[20] William H. Press and Freeman J. Dyson. Iterated prisoners dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

[21] Alexander J. Stewart and Joshua B. Plotkin. Extortion and cooperation in the prisoners dilemma. *Proceedings of the National Academy of Sciences*, 109(26):10134–10135, 2012.

A Appendix Tables

The memory one strategies used in the computer tournament described in [21] are given by Table 4.