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

Nikoleta E. Glynatsi Vincent A. Knight

Abstract

In this manuscript we build upon a framework provided in 1989 to study best responses in the well known memory one strategies of the Iterated Prisoner's Dilemma. The aim of this work is to construct a compact way of identifying best responses of short memory strategies and to show their limitations in multi-opponent interactions. A number of theoretic results are presented.

1 Introduction

The Prisoner's Dilemma (PD) is a two player game used in understanding the evolution of co-operative behaviour, formally introduced in [11]. 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 $(3, 1, 0, 5)$ [4].

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

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

The PD is a one shot game, however it is commonly studied in a manner where the history of the interactions matter. The repeated form of the game is called the Iterated Prisoner's Dilemma (IPD) and in the 1980s, following the work of[5, 6] it attracted the attention of the scientific community. In [5] and [6], the first well known computer tournaments of the IPD were performed. A total of 13 and 63 strategies were submitted in computer code and competed against each other in a round robin tournament. All contestants competed against each other, a copy of themselves and a random strategy. The winner was decided on the average score a strategy achieved and not the total number of wins. How many turns of history that a strategy would use,

the memory size, was a result of the particular strategic decisions made by the author. The winning strategy of both tournaments was a strategy called Tit for Tat. Tit for Tat is a strategy which starts by cooperating and then mimics the last move of it's opponent. This is a strategy which makes use of the previous move of the opponent only, this type of strategy is called *reactive*. In [21] a framework for studying such strategies was introduced. This was later used to introduce well known reactive strategies such as Generous Tit For Tat [22].

Reactive strategies are a subset of so called *memory one* strategies. Memory one strategies similarly are only concerned with the previous turn. However, they take into consideration both players' recent moves to decide on an action. Several successful memory one strategies are found in the literature, for example Pavlov [19].

A well known set of memory on strategies was introduced in [23], these were called zero determinant (ZD) strategies. The ZD strategies manage to force a linear relationship between the score of the strategy and the opponent. According to [23] the ZD strategies can dominate any evolutionary opponent in pairwise interactions by using a single turn of memory. Thus, in a sense this work questioned the importance of memory and the sophistication required/necessary for evolutionary fixation.

These ZD strategies attracted a lot of attention. It was stated that "Press and Dyson have fundamentally changed the viewpoint on the Prisoner's Dilemma" [25]. In [25] a very similar tournament to Axelrod's tournament is run including ZD strategies and a new set of ZD strategies. One specific advantage of memory one strategies is their mathematical tractability. As described in Section 2 they can be represented completely as an vector of $\mathbb{R}^4$.

Even so, ZD and memory one strategies have also received criticism. In [18], the 'memory of a strategy does not matter' statement was questioned. A set of more complex strategies, strategies that take in account the entire history set of the game, were trained and proven to be more robust against multiple opponents.

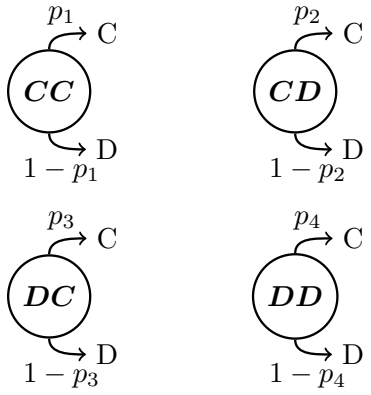
The purpose of this work is to consider a given memory one strategy in a similar fashion to [23]. However whilst [23] found a way for a player to manipulate a given opponent, this work will consider a multidimensional optimisation approach to identify the best response to a group of opponents. In essence the aim is to produce a compact method of identifying the best memory one strategy against a given opponent.

In the second part of this manuscript we explore the limitation of these best response strategies. This is achieved by comparing the performance of an optimal memory one strategy, for a given environment, with the performance of a more complex strategy that has a larger memory. One particular benefit of this analysis is the identification of conditions for which defection is a best response. Thus, identifying environments for which cooperation can not occur.

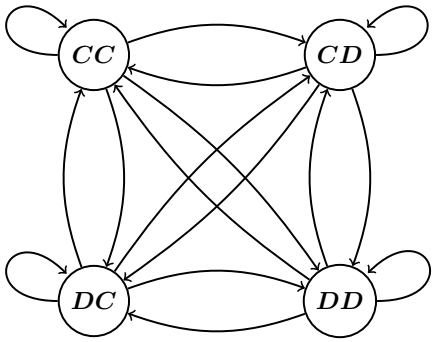
2 The utility

In [23] 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. These are CC, CD, DC, CC. A memory one strategy is denoted by the probabilities of cooperating after each of these states, $p = (p_1, p_2, p_3, p_4) \in \mathbb{R}^4_{[0,1]}$. A diagrammatic representation of such a strategy is given in Figure 1a.

Moreover, if two memory one players are moving from state to state this can be modelled as a Markov process. A diagrammatic representation of the Markov chain is shown in Figure 1b. The corresponding transition matrix M given by:



(a) Diagrammatic representation of a memory one strategy.



(b) Markov chain on a PD game.

$$M = [[p_1*q_1p_1*(-q_1+1)q_1*(-p_1+1)(-p_1+1)*(-q_1+1)][p_2*q_3p_2*(-q_3+1)q_3*(-p_2+1)(-p_2+1)*(-q_3+1)][p_3*q_2p_3*(-q_2+1)q_2* \tag{4}$$

The long run steady state probability v is the solution to $vM = v$ This corresponds to finding the unit eigenvector of M and a closed form for v can be found (given in the Appendix). Combining the stationary vector v with the payoff matrices of equation (1) allows us to retrieve the expected payoffs for each player. Thus, the utility for player p against q, denoted as $u_q(p)$, is defined by,

$$u_q(p) = v \times (R, P, S, T). \tag{5}$$

Here, to our knowledge, we present the first theoretical result which concerns the form of $u_q(p)$. That is that $u_q(p)$ is given by a ratio of two quadratic forms [16], as presented by Theorem 1.

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

$$u_q(p) = \frac{\frac{1}{2}pQp^T + cp + a}{\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}}, \tag{6}$$

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

$$\boldsymbol{Q} = \begin{bmatrix} \mathbf{0} & \mathbf{-(q_1-q_3)(q_2-5q_4-1)} & \mathbf{q_3(q_1-q_2)} & \mathbf{-5q_3(q_1-q_4)} \\ \mathbf{-(q_1-q_3)(q_2-5q_4-1)} & \mathbf{0} & \mathbf{(q_2-q_3)(q_1-3q_4-1)} & \mathbf{(q_3-q_4)(5q_1-3q_2-2)} \\ \mathbf{q_3(q_1-q_2)} & \mathbf{(q_2-q_3)(q_1-3q_4-1)} & \mathbf{0} & \mathbf{3q_3(q_2-q_4)} \\ \mathbf{-5q_3(q_1-q_4)} & \mathbf{(q_3-q_4)(5q_1-3q_2-2)} & \mathbf{3q_3(q_2-q_4)} & \mathbf{0} \end{bmatrix}, \tag{7}$$

$$\bar{\boldsymbol{Q}} = \begin{bmatrix} \mathbf{0} & \mathbf{-(q_1-q_3)(q_2-q_4-1)} & \mathbf{(q_1-q_2)(q_3-q_4)} & \mathbf{(q_1-q_4)(q_2-q_3-1)} \\ \mathbf{-(q_1-q_3)(q_2-q_4-1)} & \mathbf{0} & \mathbf{(q_2-q_3)(q_1-q_4-1)} & \mathbf{(q_1-q_2)(q_3-q_4)} \\ \mathbf{(q_1-q_2)(q_3-q_4)} & \mathbf{(q_2-q_3)(q_1-q_4-1)} & \mathbf{0} & \mathbf{-(q_2-q_4)(q_1-q_3-1)} \\ \mathbf{(q_1-q_4)(q_2-q_3-1)} & \mathbf{(q_1-q_2)(q_3-q_4)} & \mathbf{-(q_2-q_4)(q_1-q_3-1)} & \mathbf{0} \end{bmatrix}. \tag{8}$$

c and $\bar{c} \in \mathbb{R}^{4\times 1}$ are similarly defined by:

$$c = \begin{bmatrix} q_1\left(q_2 - 5q_4 - 1\right) \\ -\left(q_3 - 1\right)\left(q_2 - 5q_4 - 1\right) \\ -q_1q_2 + q_2q_3 + 3q_2q_4 + q_2 - q_3 \\ 5q_1q_4 - 3q_2q_4 - 5q_3q_4 + 5q_3 - 2q_4 \end{bmatrix}, \tag{9}$$

$$\bar{c} = \begin{bmatrix} q_1\left(q_2 - q_4 - 1\right) \\ -\left(q_3 - 1\right)\left(q_2 - q_4 - 1\right) \\ -q_1q_2 + q_2q_3 + q_2 - q_3 + q_4 \\ q_1q_4 - q_2 - q_3q_4 + q_3 - q_4 + 1 \end{bmatrix}. \tag{10}$$

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

Proof: The proof is given in Appendix.

Figure 2 indicates that the formulation of $u_q(p)$ as a quadratic ratio successfully captures the simulated behaviour. A data set offering further validation is available at. The simulated utility, denoted as $U_q(p)$ has been calculated using [1], an open research framework for the study of the IPD and is described in [17]. Note that when referring to $U_q(p)$ here onwards we mean the simulated utility calculated with [1].

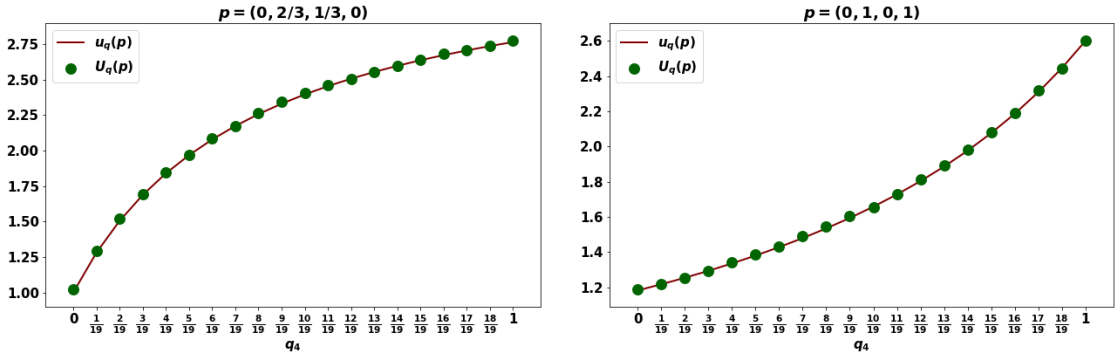


Figure 2: Differences between simulated and analytical results for $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, q_4)$.

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 is then given by the average payoffs the 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:*

$$\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) = \frac{1}{N}\frac{\sum_{i=1}^{N}(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}) \prod_{\substack{j=1 \\ j\neq i}}^{N}(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)})}{\prod_{i=1}^{N}(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)})}. \tag{11}$$

As an illustration, Theorem 2 is used to calculate he theoretical payoffs of several memory one strategies against a set of 10 opponents. The opponents used are the memory one strategies for the tournament conducted in [25]. This tournament is also simulated as before The names and a small explanation of the strategic rules are given in Appendix A. The values of $\frac{1}{N}\sum_{i=1}^{N} u_{q^{(i)}}(p)$ and $\frac{1}{N}\sum_{i=1}^{N} U_{q^{(i)}}(p)$ match (Table 7),

	p_1	p_2	p_3	p_4	$\frac{1}{10}\sum_{i=1}^{10} u_q^{(i)}(p)$	$\frac{1}{10}\sum_{i=1}^{10} U_q^{(i)}(p)$
0	0.0	$\frac{1}{3}$	$\frac{1}{3}$	1.0	2.158	2.166
1	0.0	0.0	$\frac{1}{3}$	1.0	2.165	2.173
2	0.0	$\frac{1}{3}$	1.0	1.0	2.149	2.157
3	0.0	$\frac{1}{3}$	$\frac{2}{3}$	1.0	2.139	2.149
4	0.0	0.0	0.0	$\frac{2}{3}$	2.191	2.200
5	0.0	$\frac{1}{3}$	1.0	$\frac{2}{3}$	2.157	2.167
6	0.0	0.0	$\frac{2}{3}$	1.0	2.156	2.166
7	0.0	0.0	$\frac{2}{3}$	$\frac{2}{3}$	2.145	2.156
8	0.0	$\frac{1}{3}$	0.0	$\frac{2}{3}$	2.199	2.211
9	0.0	0.0	1.0	$\frac{1}{3}$	2.186	2.198

Table 1: Results of memory one strategies against the strategies in Table 7.

Note that the utility against a group of strategies can not be captured by the utility against the mean opponent.

$$\frac{1}{N}\sum_{i=1}^{N} u_q{}^{(i)}(p) \neq u_{\frac{1}{N}\sum_{i=1}^{N} q^{(i)}}(p). \tag{12}$$

This is illustrated in Figure 3.

The analytical formulation of Theorem 2 will be used in the following sections to explore the best response to memory one strategies.

3 Best responses to memory one players

Identifying the **best response** to a group of memory one strategies will be considered as a multi dimensional optimisation problem, where a memory one strategy p aims to optimise $\frac{1}{N}\sum u_{q^{(i)}}(p)$ against a set opponents $\{q^{(1)}, q^{(2)}, \ldots, q^{(N)}\}$.

The decision variable is the vector p and the solitary constraint is that $p \in \mathbb{R}^4_{[0,1]}$. The optimisation problem is given by (13).

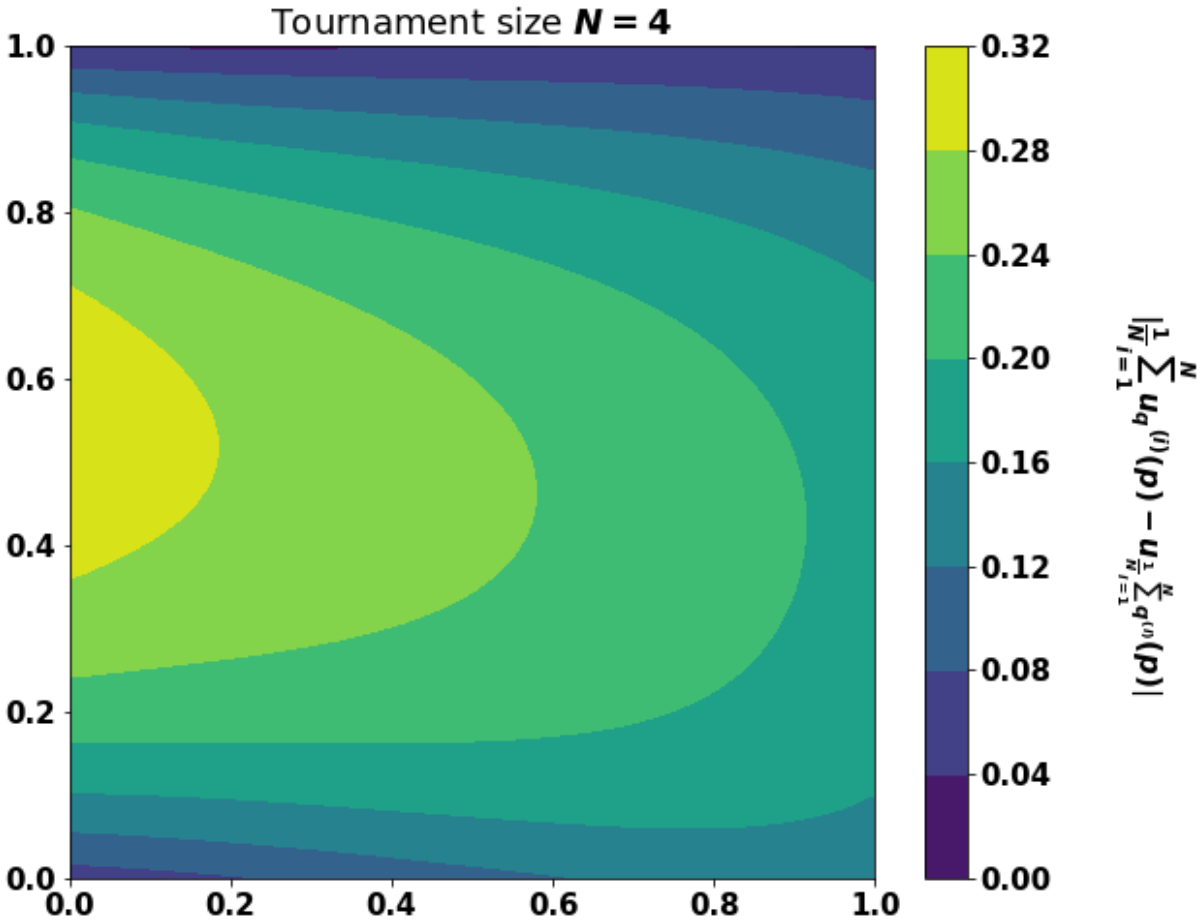


Figure 3: Plotting the difference of the average utility against ten opponents versus the utility against the average of the ten strategies.

$$\max_{p} : \sum_{i=1}^{N} {u_q}^{(i)}(p) \quad \text{such that} : p \in \mathbb{R}_{[0,1]} \tag{13}$$

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.

There is some work on the optimisation on non concave ratios of of quadratic forms [7, 9], however in these both the numerator and the denominator of the fractional problem were concave which is not true for 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 (6), is not concave. Furthermore neither the numeration or the denominator of (6), are concave.*

Proof. A function $f(x)$ is said to be concave on an interval $[a, b]$ if, for any points x_1 and $x_2 \in [a, b]$, the function $-f(x)$ is convex on that interval.

A function $f(x)$ is convex on an interval $[a, b]$ if for any two points x_1 and x_2 in $[a, b]$ and any λ where $0 < \lambda < 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{14}$$

Let f be $u_{(\frac{1}{3},\frac{1}{3},\frac{1}{3},\frac{1}{3})}$ it can be shown that for $x_1 = (\frac{1}{4}, \frac{1}{2}, \frac{1}{5}, \frac{1}{2}), x_2 = (\frac{8}{10}, \frac{1}{2}, \frac{9}{10}, \frac{7}{10})$ and $\lambda = \frac{1}{10}$ condition (14) does not hold as: $1.49 \geq 1.48$.

In [3] it is stated that a quadratic form will be concave if and only if it's symmetric matrix is negative semi definite. A matrix$A$ is semi-negative definite if:

$$|A|_i \leq 0 \text{ for } i \text{ is odd and } |A|_i \geq 0 \text{ for } i \text{ is even.} \tag{15}$$

For both Q and $\bar{Q}$ it is exhibited that for $i = 2$ (odd):

$$|Q|_2 = -\left(q_1 - q_3\right)^2 \left(q_2 - 5q_4 - 1\right)^2,$$
$$|\bar{Q}|_2 = -\left(q_1 - q_3\right)^2 \left(q_2 - q_4 - 1\right)^2$$

both determinants are negative, thus the concavity condition (15) fails for both quadratic forms. □

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 (11) 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 [8] and Karush-Kuhn-Tucker conditions [12] are based on this. The Karush-Kuhn-Tucker conditions are used here because the constraints are inequalities.

The above discussion leads to Lemma 4 which presents the best response 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:*

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

where the set S_q is defined as

$$S_q = \left\{ p \in \mathbb{R}^4 \mid \begin{array}{c} p_i \in 0, 1 \text{ or } F_i(p_i) = 0 \\ \prod_{i=1}^{N} Q_D^{(i)} \neq 0 \end{array} \quad \text{for all } i \in \{1,2,3,4\} \right\}$$

where,

$$F = (\sum_{i=1}^{N} Q_N^{(i)'} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)} + \sum_{i=1}^{N} Q_D^{(i)'} \sum_{\substack{j=1 \\ j\neq i}}^{N} Q_N^{(i)} \prod_{j\neq\{i,j\}}^{N} Q_D^{(i)}) \times \prod_{i=1}^{N} Q_D^{(i)} - (\sum_{i=1}^{N} Q_D^{(i)'} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)}) \times (\sum_{i=1}^{N} Q_N^{(i)} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)}) \quad (16)$$

and,

$$\begin{aligned} Q_N^{(i)} &= \frac{1}{2} p Q^{(i)} p^T + c^{(i)} p + a^{(i)}, \\ Q_N^{(i)'} &= p Q^{(i)} + c^{(i)}, \\ Q_D^{(i)} &= \frac{1}{2} p \bar{Q}^{(i)} p^T + \bar{c}^{(i)} p + \bar{a}^{(i)}, \\ Q_D^{(i)'} &= p \bar{Q}^{(i)} + \bar{c}^{(i)}. \end{aligned}$$

Proof. The best response of a memory one strategy against a group of memory one strategies can captured by a candidate set of behaviours. This candidate set is constructed by considering behaviours where any or all of p_1, p_2, p_3, p_4 are $\in \{0, 1\}$ and the rest or all of p_1, p_2, p_3, p_4 are given by roots of the partial derivatives.

Note that for $p_i \in \{0, 1\}$ we consider the roots of the partial derivatives for $p_j \neq p_i$ for $i, j \in [1, 4]$.

The derivatives, $\frac{d \sum u}{dp}$, are given by,

$$\frac{d}{dp} \sum_{i=1}^{N} {u_q}^{(i)}(p) = \\ = \frac{(\sum_{i=1}^{N} Q_N^{(i)'} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)} + \sum_{i=1}^{N} Q_D^{(i)'} \sum_{\substack{j=1 \\ j\neq i}}^{N} Q_N^{(i)} \prod_{j\neq\{i,j\}}^{N} Q_D^{(i)}) \times \prod_{i=1}^{N} Q_D^{(i)} - (\sum_{i=1}^{N} Q_D^{(i)'} y - vk \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)}) \times (\sum_{i=1}^{N} Q_N^{(i)} \prod_{\substack{j=1 \\ j\neq i}}^{N} Q_D^{(i)})}{(\prod_{i=1}^{N} Q_D^{(i)})^2} \quad (17)$$

For equation 17 to be zero, the numerator must fall to zero and the denominator can not nullified.

One the candidate set is constructed each point is evaluated using equation (11). The point with the maximum utility is selected. □

A special case of Lemma 4 is for $N = 1$, thus when a strategy plays against a single opponent. In this case the formulation of Theorem 1 is used and the best response is captured by Lemma 5.

Lemma 5. *The optimal behaviour of a memory one strategy player $p^* \in \mathbb{R}^4_{[0,1]}$ against a given opponent $q \in \mathbb{R}^4_{[0,1]}$ is given by:*

$$p^* = \text{argmax}(u_q(p)),\ p \in S_q,$$

where the set S_q is defined as

$$S_q = \{0, \bar{p}_i, 1\}^4 \text{ for } i \in \mathbb{R},$$

*where any $\bar{p}$ satisfy condition (**??**). Note that now the numerators of the partial derivatives, (16), are given by*

$$(pQ + c)(\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}) - (p\bar{Q} + \bar{c})(\frac{1}{2}pQp^T + cp + a) \tag{18}$$

*and (**??**) is re-written as:*

$$\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a} \neq 0 \tag{19}$$

Proof. The best response of a memory one strategy against another memory one strategy can captured by a candidate set of behaviours. This candidate set is constructed by considering behaviours where any or all of p_1, p_2, p_3, p_4 are $\in \{0, 1\}$ and the rest or all of p_1, p_2, p_3, p_4 are given by roots of the partial derivatives.

Note that for $p_i \in \{0, 1\}$ we consider the roots of the partial derivatives for $p_j \neq p_i$ for $i, j \in [1, 4]$.

The derivatives, $\frac{du}{dp}$, are given by,

$$\frac{du_q(p)}{dp} = \frac{(pQ + c)(\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}) - (p\bar{Q} + \bar{c})(\frac{1}{2}pQp^T + cp + a)}{(\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a})^2} \tag{20}$$

For equation 17 to be zero, the numerator must fall to zero and the denominator can not be zero. □

Equations (16) and (18) are systems of at most 4 polynomials and the degree of the polynomials is gradually increasing every time an extra opponent is taken into account. Solving system of polynomials corresponds to the calculation of a resultant and for large systems these quickly become intractable. Because of that no further analytical consideration is given to problems described here.

Theorems **??** can also be used to identify if a strategy is a best response.

Lemma 4 and Theorem 2 will now be used to give a list of particular best response results.

4 Stability of defection

In this section the stability of defection is explored. 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. In this manuscript we try to provide a condition for when defection is the best response in the IPD, thus when it is known that cooperation is not dominant.

Initially, let us consider equation (20) for $p = (0, 0, 0, 0)$,

$$\frac{du}{dp_{|p=(0,0,0,0)}} = \quad \frac{c\bar{a} - \bar{c}a}{\bar{a}^2}. \tag{21}$$

The numerator $\bar{c}a - c\bar{a}$ is given by,

$$\begin{bmatrix} 0 \\ 0 \\ q_4\left(4q_1q_2 - 3q_2^2 - 4q_2q_3 + 3q_2q_4 + 4q_3 - 5q_4 - 1\right) \\ -\left(q_2 - 1\right)\left(4q_1q_4 - 3q_2q_4 + q_2 - 4q_3q_4 + 4q_3 + 3q_4^2 - 6q_4 - 1\right) \end{bmatrix}$$

and the denominator $\bar{a}^2 = (-q_2 + q_4 + 1)^2$, which is always positive. In order for defection to be the best response the derivative must have a negative sign at the point $p = (0, 0, 0, 0)$. That means that the utility is only decreasing after $p = (0, 0, 0, 0)$.

Because $\bar{a}^2$ is always positive the sign of the derivative is given by $\bar{c}a - c\bar{a}$. More specifically from equations,

$$q_4\left(4q_1q_2 - 3q_2^2 - 4q_2q_3 + 3q_2q_4 + 4q_3 - 5q_4 - 1\right) \tag{22}$$

$$-\left(q_2 - 1\right)\left(4q_1q_4 - 3q_2q_4 + q_2 - 4q_3q_4 + 4q_3 + 3q_4^2 - 6q_4 - 1\right) \tag{23}$$

Both signs of the partial derivatives must be negative in order for the overall function to be decreasing ensuring defection is a best response. The signs of equations (22) and (23) vary. There are cases that they have the same sign and cases that they do not,

For a tournament setting we substitute $p = (0, 0, 0, 0)$ in equation (17) which gives:

$$\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \prod_{\substack{j=1 \\ j\neq i}}^{N} (\bar{a}^{(i)})^2 \tag{24}$$

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, as discussed above. A condition that must hold in order for defection to be stable in a tournament is that each term of the sum must be negative. The results are exhibited in Lemma 6.

Lemma 6. *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:*

$$\sum_{i=1}^{N} (c^{(i)T} \bar{a}^{(i)} - \bar{c}^{(i)T} a^{(i)}) \leq 0 \tag{25}$$

Moreover lets us consider a constrained version of the problem once again. Lets us assume that in an pairwise interaction the opponent is a reactive player $q = (q_1, q_2, q_1, q_2)$. By substituting $q_3 = q_1$ and $q_4 = q_2$ equations (22) and (23) are now re written as follow,

$$\begin{bmatrix} -q_2 \left(4q_1 - 5q_2 - 1\right) \\ \left(q_2 - 1\right) \left(4q_1 - 5q_2 - 1\right) \end{bmatrix}$$

Lemma 7. *Defection is the best responses of a memory one player p against a reactive player q if the transition probabilities of the opponent satisfy the condition:*

$$4q_1 - 5q_2 - 1 > 0 \tag{26}$$

5 Optimisation against memory one strategies

The results of the previous section allow for the quick characterisation of a strategy and indeed a Nash equilibria. Here the optimisation problem of (13) is maximized using Bayesian optimisation. Bayesian optimisation is a global optimisation algorithm, introduced in [20], which has proven to outperform many other popular algorithms [15].

Bayesian optimisation constructs a probabilistic model for f and then exploits this model to make decisions about where in the bounded set to next evaluate the function. It relies on the prior information and does not simply rely on local gradient and Hessian approximations. This allows the algorithm to optimise a non concave function with relatively few evaluations, at the cost of performing more computation to determine the next point to try [24].

The open source package [14] offers an implementation of bayesian optimisation and is used in this paper to compute a large number of best memory one responses against sets of random memory one strategies. The implementation of bayesian in [14] allows us to perform the algorithm for a different combination of parameters. The parameters explored here are:

- Number of calls, maximum number of calls to the objective function.
- Number of random starts, number of evaluations of the objective function with random points before approximating it.

The different parameters' combinations and their respective values are laid out in Table 2.

	number of calls	number of random starts
1	20	10
2	30	20
3	40	20
4	45	20
5	50	20

Table 2: Bayesian optimisation sets of parameters' values.

A total of 9900 different memory one opponents were randomly generated and Bayesian optimisation was used to find the optimal reactive strategy. The results were compared to that of Lemma **??** which is used to obtain the exact optimal. The results of this comparison are presented by Table 3.

Difference	Labels
0.026991	calls: 20, random starts: 10
0.018916	calls: 30, random starts: 20
0.007298	calls: 40, random starts: 20
0.005552	calls: 45, random starts: 20
0.005114	calls: 50, random starts: 20

Table 3: Difference of $u_q(p)$ for $p \in \mathbb{R}^2_{[0,1]}$. The difference was calculated as exact $u_q(p^*)$ minus Bayesian $u_q(\tilde{p}^*)$.

The combination that was chosen to carry out the empirical trials is that of 50 number of calls and 20 number of random starts. This set of parameters was determined to be the most efficient using best reactive responses as an experimental case.

Bayesian optimisation finds the optimal behaviour of reactive strategies. The very same set of parameters will now be used to optimise memory one strategies against single opponents and against sets of $N = 2$ opponents. $N = 2$ was chosen because is the smallest N for which there is a multi opponent interaction. For each $N = 1$ and $N = 2$ opponents a total of 1022 best responses of memory one strategies have been captured. This data has been archived in.

Note that another global optimization algorithm called differential evolution [26] was also evaluated. Bayesian optimization was chosen over differential evolution due to a lower computational cost and comparable results.

6 Limitation of memory

The third and final part of this paper focuses on proving that short memory strategies have limitations. Though it has been proven [23] that there exists a set of memory one strategies that can outperform any opponent, this was done only for the case of $N = 1$. 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 long memory strategies are trained using reinforcement learning through Bayesian optimisation similarly to Section 5. The trained strategy used is a strategy called Gambler, introduced and discussed in [13], and the objective function is the average performance in a tournament of 200 turns and 50 repetitions. Thus, the player learns by playing a number of players, building a Bayesian landscape of it's parameters and updating them accordingly.

6.1 Gambler

Several ways of representing IPD strategies have been used over the years. In [13] several of those 'archetypes' are presented and used to train different successful strategies. One of the archetypes firstly introduced in that paper is Gambler. Gambler is based on a lookup table and maps the opponent's first n_1 moves, the opponent's last m_1 moves, and the players last m_2 moves to a probability of cooperation.

Several variants of Gambler have been trained for this work (Table 4).

	n_1	m_1	m_2
1	1	1	2
2	2	2	0
3	2	2	1
4	2	2	2
5	4	4	4

Table 4: Variants of Gambler used.

6.2 Empirical Results

The performance of the memory one and Gamblers strategies are compared for cases of $N = 1$ and $N = 2$. The following steps are taken:

1. An N number of random opponents are generated: this gives the environment.
2. Using (13) and Bayesian optimisation p^* s obtained for the environment.
3. A Gambler type (each variant of Table 4) is trained for the same environment.
4. Both utilities are compared.

A large data set containing the opponents as well as the optimised/trained behaviours can be found in. The number of experimental cases for each Gambler are displayed in Table 5. Note that a number of 1022 trials corresponds to 1022 trials for $N = 1$ and 1022 trials for $N = 2$.

	Gamblers	Number of trials
0	Gambler 1_1_2	1022
1	Gambler 2_2_0	1043
2	Gambler 2_2_1	1074
3	Gambler 2_2_2	821
4	Gambler 4_4_4	22

Table 5: Number of trials, for $N = 1$ and $N = 2$, for each Gambler instance.

The results are explored by studying the difference between $\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p^*)$ and $\frac{1}{N}\sum_{i=1}^{N} {U_q}^{(i)}(G)$, where $U(G)$ represents the utility of a Gambler. The results are shown in Figure 4.

For the cases of Gambler $n_1 = 1, m_1 = 1, m_2 = 2$, $n_1 = 2, m_1 = 2, m_2 = 0$ and $n_1 = 2, m_1 = 2, m_2 = 1$, though there are few edges cases, the difference distribution is congregated around zero. For the rest of the Gambler's types there nce is mainly worse. This could be a result of the Gamblers not being trained for long enough. Thus a larger number of calls should be used.

Furthermore, there is no significant difference between the distributions of $N = 1$ and $N = 2$. This was checked by performing $T-$test for the means of two samples. The calculated $p-$ values are presented in Table 6.

There appears to be no significant difference between complex and memory one strategies. The difference in performance is mainly congregated around zero, and that is true for both cases of $N = 1$ and $N = 2$. However, that there is indication that complex strategies can outperform memory one strategies for $N = 2$.There are cases that they have a difference in score of 0.5.

	Gamblers	$p-$ values
0	Gambler 1_1_2	0.242
1	Gambler 2_2_0	0.214
2	Gambler 2_2_1	0.179
3	Gambler 2_2_2	0.629
4	Gambler 4_4_4	0.141

Table 6: $p-$ values for the means of $N = 1$ to $N = 2$ using $T-$tests.

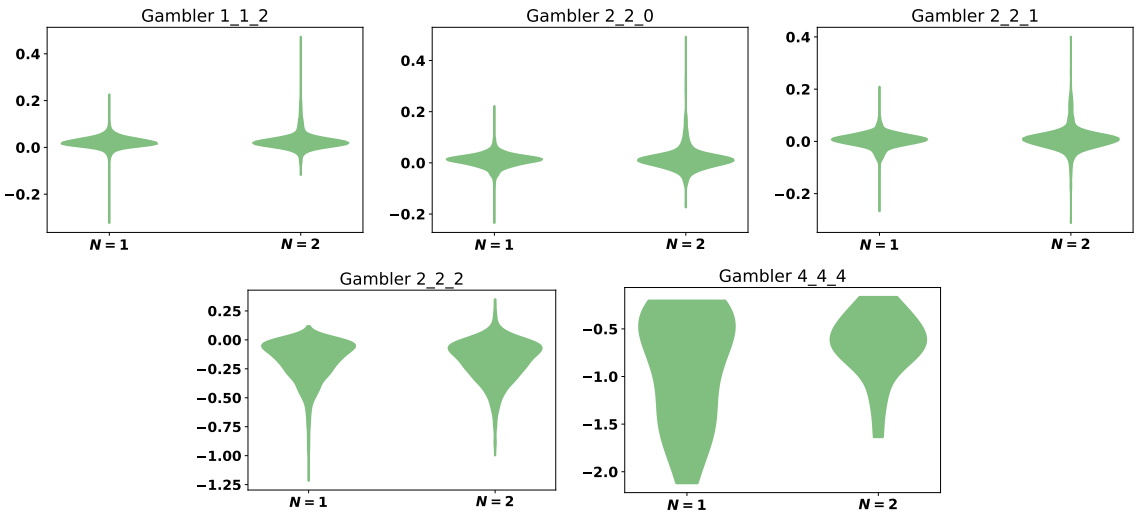


Figure 4: Difference between $\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p^*)$ and $\frac{1}{N}\sum_{i=1}^{N} {U_q}^{(i)}G$ for $N = 1$ and $N = 2$.

7 Discussion

In this framework, memory one strategies for the well known game the IPD were studied. These are strategies that utilize a single slot of memory to define their next action. An analytical formulation for retrieving the payoffs of memory one strategies against memory one strategies was used here. Though the analytical formulation has been previously use, this manuscript is the first to prove that the payoff of a such a player p has a compact form and proved that is a non concave function. Furthermore, best memory one responses were exploit as an optimisation problem of a ratio of quadratic forms.

We have managed to prove that for reactive and purely random strategies that best responses can be captured analytically. This was done using using algebraic approaches such as companion matrices and resultant theory. We investigated the stability of defection and proved that environments for which cooperation will never emerge can be recognised immediately by the transitions of the opponents.

Finally, we generated a large date set of bests memory one responses for $N = 1$ and $N = 2$. The limitations of memory were tried to be shown by comparing the performance of best memory one strategies to that of more complex strategies. Though there are indications that complex strategies indeed perform better, the significant of the difference is in question. More experimental trials and exploration will be carried out.

A Appendix Tables

The memory one strategies used in the computer tournament described in [25] are given by Table 7.

References

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

	Name	Memory one representation	Explanation
1	Cooperator	$(1, 1, 1, 1)$	Always chooses C.
2	Defector	$(0, 0, 0, 0)$	Always chooses D.
3	Random	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	Randomly chooses between C and D with a probability of 0.5.
4	Tit for Tat	$(1, 0, 1, 0)$	Start with a C and then mimics the opponent's last move.
5	Grudger	$(1, 0, 0, 0)$	Starts by cooperating however will defect if at any point the opponent has defected.
6	Generous Tit for Tat	$(1, \frac{1}{3}, 1, \frac{1}{3})$	A more generous version of Tit for Tat.
7	Win Stay Lose Shift	$(1, 0, 0, 1)$	Starts with a C and then repeats it's previous move only if it was awarded with a payoff of R or T.
8	ZDGTFT2	$(1, \frac{1}{8}, 1, \frac{1}{4})$	A generous zero determinant strategy introduced in [25]
9	ZDExtort2	$(\frac{8}{9}, \frac{1}{2}, \frac{1}{3}, 0)$	An extortionate zero determinant strategy introduced in [25]
10	Hard Joss	$(\frac{9}{10}, 0, \frac{9}{10}, 0)$	Cooperates with probability $\frac{9}{10}$ when the opponent cooperates, otherwise emulates Tit for Tat.

Table 7: The memory one strategies from [25].

[2] Alkiviadis G. Akritas. *Sylvester's form of the Resultant and the Matrix-Triangularization Subresultant PRS Method*, pages 5–11. Springer New York, New York, NY, 1991.

[3] Howard Anton and Chris Rorres. *Elementary Linear Algebra: Applications Version*. Wiley, eleventh edition, 2014.

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

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

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

[7] 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.

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

[9] 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.

[10] Alan Edelman and H Murakami. Polynomial roots from companion matrix eigenvalues. *Mathematics of Computation*, 64(210):763–776, 1995.

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

[12] 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.

[13] 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.

[14] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Z Vincius, cmmalone, Christopher Schrder, nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, rene rex, Kejia (KJ) Shi, Justus Schwabedal, carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, Loc Estve, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch. scikit-optimize/scikit-optimize: v0.5.2, March 2018.

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

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

[17] 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.

[18] Christopher Lee, Marc Harper, and Dashiell Fryer. The art of war: Beyond memory-one strategies in population games. *PLOS ONE*, 10(3):1–16, 03 2015.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[25] 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.

[26] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.