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

Nikoleta E. Glynatsi   Vincent A. Knight

Abstract

1   Introduction

The Prisoner's Dilemma (PD) is a two player game used in understanding the evolution of co-operative behaviour, formally introduced in [9]. 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 $(3, 1, 0, 5)$ [3].

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

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

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 [4, 5] it attracted the attention of the scientific community. In [4] and [5], 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* [20] 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. As the name suggests memory one strategies have a memory of size 1.

Several successful reactive strategies and memory one are found in the literature, such as Generous Tit For Tat [21] and Pavlov [17]. However, memory one strategies generated a small shock in the game theoretic community ([25] stated that "Press and Dyson have fundamentally changed the viewpoint on the Prisoner's Dilemma") when a cernent set of memory one strategies was introduced in [23]. 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.

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. The main findings are:

- A compact method of identifying the best memory one strategy against a given set of opponents.

- 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 [20] 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 modelled as a 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 $u_q(p)$.
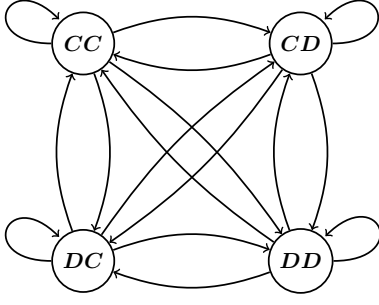
$$M = \begin{bmatrix} p_1 q_1 & p_1 (-q_1 + 1) & q_1 (-p_1 + 1) & (-p_1 + 1)(-q_1 + 1) \\ p_2 q_3 & p_2 (-q_3 + 1) & q_3 (-p_2 + 1) & (-p_2 + 1)(-q_3 + 1) \\ p_3 q_2 & p_3 (-q_2 + 1) & q_2 (-p_3 + 1) & (-p_3 + 1)(-q_2 + 1) \\ p_4 q_4 & p_4 (-q_4 + 1) & q_4 (-p_4 + 1) & (-p_4 + 1)(-q_4 + 1) \end{bmatrix}$$

Figure 1: markov

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

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

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

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

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

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

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

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

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. Two examples are graphically represented in Figure 2 and show 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].
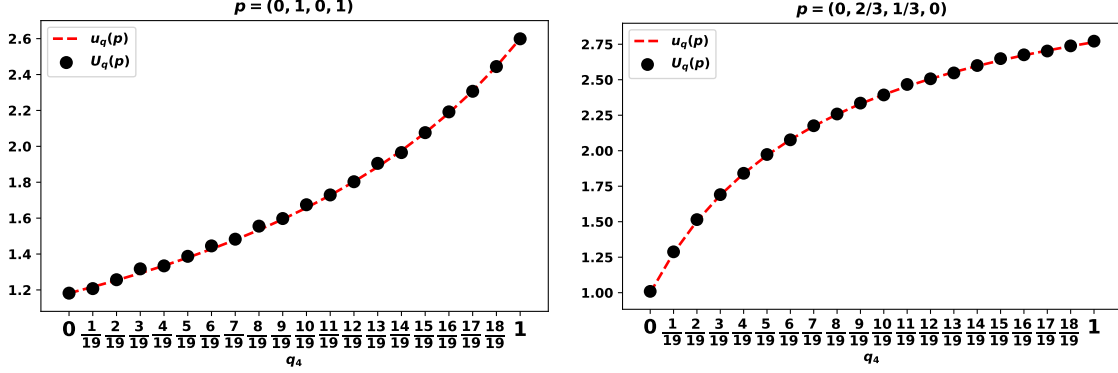

Figure 2: Differences between simulated and analytical results for $q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, q_4)$.

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:*

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

As an illustration, Theorem 2 is used to calculate the 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]; the names of the strategic rules are given by Table 1. Figure 3 provides evidence that 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.

Note that it was explored whether the the utility against a group of strategies could be captured by the utility against the mean opponent. Thus, finding the best response to a given group of opponents correspond to finding the best response to a single player, the mean player formed by the probabilities of that given group. The hypothesis however fails, as:

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

4

	Name	Memory one representation	Reference
1	Cooperator	$(1, 1, 1, 1)$	[3]
2	Defector	$(0, 0, 0, 0)$	[3]
3	Random	$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$	[3]
4	Tit for Tat	$(1, 0, 1, 0)$	[3]
5	Grudger	$(1, 0, 0, 0)$	[16]
6	Generous Tit for Tat	$(1, \frac{1}{3}, 1, \frac{1}{3})$	[21]
7	Win Stay Lose Shift	$(1, 0, 0, 1)$	[17]
8	ZDGTFT2	$(1, \frac{1}{8}, 1, \frac{1}{4})$	[25]
9	ZDExtort2	$(\frac{8}{9}, \frac{1}{2}, \frac{1}{3}, 0)$	[25]
10	Hard Joss	$(\frac{9}{10}, 0, \frac{9}{10}, 0)$	[25]

Table 1: List of strategies used in the tournament described in [25].
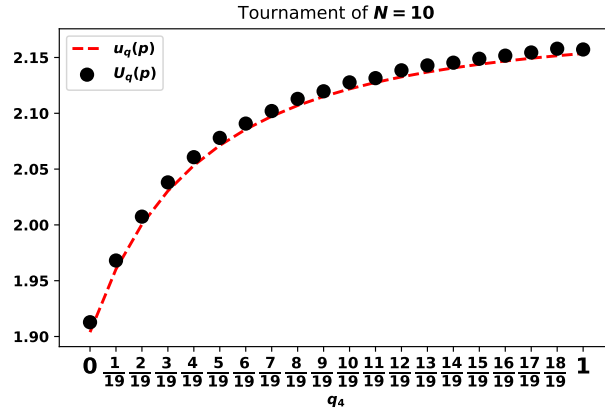

Figure 3: Results of memory one strategies against the strategies in Table 1.
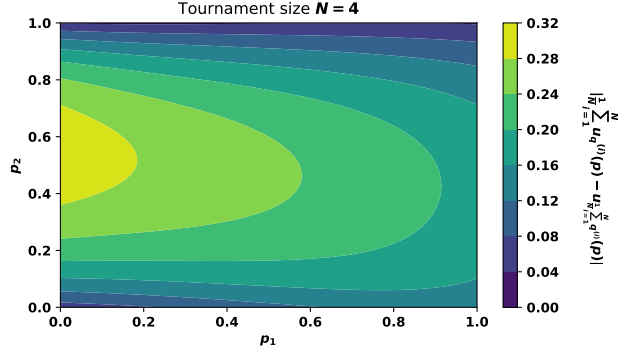
which is captured by Figure 4.

Figure 4: The difference between the average utility against 10 opponents and the utility against the average player of these 10 opponents is plotted. It is clear that the hypothesis fails for this case as the absolute difference is manly positive for the example. The hypothesis fails fro at least one case, thus it can not be assumed to be true.

In the following section best responses are introduced and explored for the case of memory one strategies. Moreover, in the following sections several other theoretical results and presented and the advantages of analytical formulation of Theorem 2 become evident.

3   Best responses to memory one players

In game theory a *best response* is the strategy which corresponds to the most favourable outcome. In this manuscript a best response memory one strategy corresponds to the p^* for which $\sum u_{q^{(i)}}(p^*)$ for $i \in \{1, \dots, N\}$ 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).

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

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 [6, 8], in these both the numerator and the denominator of the fractional problem were concave or that the denominator was greater than zero. Both assumptions fails for (12). 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.*

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.

6

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 $0 < \lambda < 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{13}$$

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 (13) does not hold as: $1.49 \geq 1.48$.

In [2] 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:

$$|A|_i \leq 0 \text{ for } i \text{ is odd and } |A|_i \geq 0 \text{ for } i \text{ is even.} \tag{14}$$

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 (14) fails for both quadratic forms. $\qquad\square$

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 [7] and Karush-Kuhn-Tucker conditions [10] are based on this. The Karush-Kuhn-Tucker conditions are used here because the constraints are inequalities. 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:*

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

where the set S_q is defined as

$$S_q = \left\{ \bar{p} \in \mathbb{R}^4 \ \middle| \ \begin{array}{c} \bar{p}_k \in \{0, 1\} \text{ for all } k \in \{1, 2, 3, 4\} \text{ or } F(\bar{p}) = 0 \\ \prod_{i=1}^{N} Q_D^{(i)} \neq 0 \end{array} \right\}$$

where,

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

(15)

and,

$$Q_N^{(i)} = \frac{1}{2} p Q^{(i)} p^T + c^{(i)} p + a^{(i)},$$

$$Q_N^{(i)'} = p Q^{(i)} + c^{(i)},$$

$$Q_D^{(i)} = \frac{1}{2} p \bar{Q}^{(i)} p^T + \bar{c}^{(i)} p + \bar{a}^{(i)},$$

$$Q_D^{(i)'} = p \bar{Q}^{(i)} + \bar{c}^{(i)}.$$

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,

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

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

(16)

For equation 16 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 (10). 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 $p^* \in \mathbb{R}_{[0,1]}^4$ against a given opponent $q \in \mathbb{R}_{[0,1]}^4$ is given by:*

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

where the set S_q is defined as

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

8

where any $\bar{p}$ satisfy conditions:

$$(\bar{p}Q + c)(\frac{1}{2}\bar{p}\bar{Q}\bar{p}^T + \bar{c}\bar{p} + \bar{a}) - (\bar{p}\bar{Q} + \bar{c})(\frac{1}{2}\bar{p}Q\bar{p}^T + c\bar{p} + a) = 0 \tag{17}$$

and

$$\frac{1}{2}\bar{p}\bar{Q}\bar{p}^T + \bar{c}\bar{p} + \bar{a} \neq 0 \tag{18}$$

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,

$$\frac{du_q(p)}{dp} = \frac{(pQ + c)(\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}) - (p\bar{Q} + \bar{c})(\frac{1}{2}pQp^T + cp + a)}{(\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a})^2} \tag{19}$$

For equation 16 to be zero, the numerator must fall to zero and the denominator can not be zero. □

Equation (17) is 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

4 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. This manuscript provides an identification of conditions for which defection is known to be a best response; thus identifying environments where cooperation can not occur, 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:*

$$\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \leq 0 \tag{20}$$

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 (16) which gives:

9

$$\sum_{i=1}^{N} (c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \prod_{\substack{j=1 \\ j \neq i}}^{N} (\bar{a}^{(i)})^2 \tag{21}$$

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 $\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \leq 0$. $\qquad\square$

A further constrained version of Lemma 6, is for single interactions while the opponent is a reactive player. Defection is known to be stable in such interactions by the condition given in Lemma 7.

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:*

$$4q_1 - 5q_2 - 1 > 0 \tag{22}$$

Proof. Initially, consider equation (19) for $p = (0, 0, 0, 0)$,

$$\frac{du}{dp}_{|p=(0,0,0,0)} = \frac{c\bar{a} - \bar{c}a}{\bar{a}^2}. \tag{23}$$

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

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

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,

$$q_4 \left(4q_1q_2 - 3q_2^2 - 4q_2q_3 + 3q_2q_4 + 4q_3 - 5q_4 - 1\right) \tag{24}$$
$$-(q_2 - 1)\left(4q_1q_4 - 3q_2q_4 + q_2 - 4q_3q_4 + 4q_3 + 3q_4^2 - 6q_4 - 1\right) \tag{25}$$

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 (24) and (25) vary. There are cases that they have the same sign and cases that they do not,

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 (24) and (25) are now re written as follow,

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

$\square$

5   Numerical experiments

As described in Section 3, the optimisation problem of (12). The results of the previous section allow for the quick characterisation of a strategy and indeed a Nash equilibria. Here the optimisation problem of (12) is maximized using Bayesian optimisation. Bayesian optimisation is a global optimisation algorithm, introduced in [18], which has proven to outperform many other popular algorithms [12].

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 [11] 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 [11] 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.

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 2.

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.

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

6 Best response dynamics

As we briefly discussed in Section 2, the IPD is commonly studied in Moran Process, 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 we have be used so far. The utility of a memory one strategy in an evolutionary setting is given by,

$$\frac{1}{N} \sum_{i=1}^{N} u_q^{(i)}(p) + u_p(p). \tag{26}$$

and respectively the optimisation problem is now re written as,

$$\max_p : \frac{1}{N} \sum_{i=1}^{N} u_q^{(i)}(p) + u_p(p) \tag{27}$$
$$\text{such that} : p \in \mathbb{R}_{[0,1]}$$

Due to the new term being added to the utility, the assumption that has been made so far regarding the form of the utility now fails. The utility is not a ratio of two quadratic forms any more. Furthermore, a new method for identifying an evolutionary best response is composed in this Section. The method considered is called *best response dynamics*, and the algorithm describing the method is given by Algorithm 1.

Best response dynamics are commonly used in evolutionary game theory. Best response dynamics represent a class of strategy updating rules, where players strategies in the next round are determined by their best responses to some subset of the population, whether this might be in a large population model such as Moran Processes [15] or in a spatial model [22]. Moreover, in the theory of potential games, best response dynamics refers to a way of finding a pure Nash equilibrium by computing the best response for every player [19]. Here we defined a combination of the two methods.

**Algorithm 1** Best response dynamics algorithm

---
1: **procedure** Approximate best evolutionary response

2:     $\tilde{p} \leftarrow (0, 0, 0, 0)$

3:     $p^* \leftarrow (1, 1, 1, 1)$

4:     **while** $p^* \neq \tilde{p}$ **do**:

5:         $\text{temp} \leftarrow \tilde{p}$

6:         $\tilde{p} \leftarrow p^*$

7:         $p^* = \text{argmax}_{p^*} \left(\sum_{i=1}^{N} u_q(p^*) \right) + u_{\text{temp}}(p^*))$

8:     **return** p^*;

Numerical simulations have been carried out to visualise how the algorithm convergences after a few iterations. In Figure 5, the results are illustrated. The algorithm have been set to start from the point $(1, 1, 1, 1)$. A more optimal point could be considered, but it has been shown that the algorithm converges to same optimal solution for different initial starts. Moreover, in Figure 5 it is shown that the algorithm stops once the shame point has been evaluated.

7 Limitation of memory

References

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

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

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.
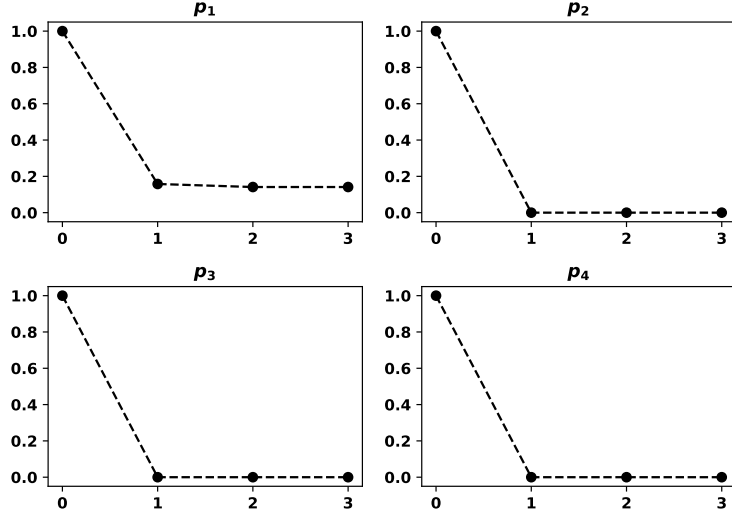
(a) Best response dynamics with $N = 2$. More specifically, for $q^{(1)} = (0.9560, 0.9478, 0.0565, 0.0848)$ and $q^{(2)} = (0.8354, 0.7359, 0.6697, 0.3081)$.
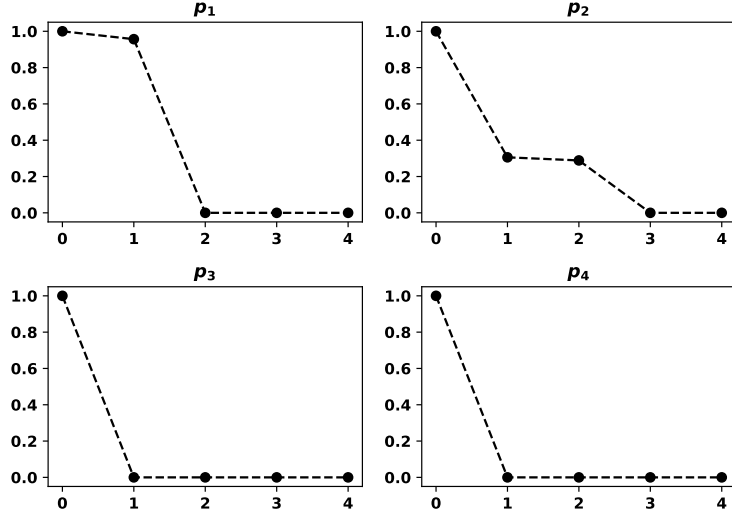
(b) Best response dynamics with $N = 2$. More specifically, for $q^{(1)} = (0.2360, 0.1031, 0.3960, 0.1549)$ and $q^{(2)} = (0.0665, 0.4015, 0.9179, 0.8004)$.
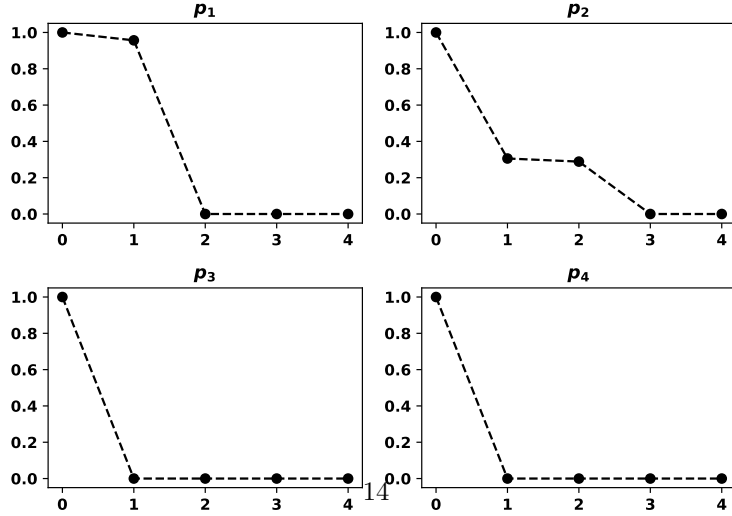
(c) Best response dynamics with $N = 2$. More specifically, for $q^{(1)} = (0.4630, 0.3733, 0.1385, 0.8665)$ and $q^{(2)} = (0.0064, 0.5027, 0.8982, 0.0808)$.

Figure 5: Best response dynamics numerical experiments.

[11] 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.

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

[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] Vincent Knight, Marc Harper, Nikoleta E. Glynatsi, and Owen Campbell. Evolution reinforces cooperation with the emergence of self-recognition mechanisms: An empirical study of strategies in the moran process for the iterated prisoners dilemma. *PLOS ONE*, 13(10):1–33, 10 2018.

[16] 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.

[17] 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.

[18] 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.

[19] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*. Cambridge University Press, 2007.

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

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

[22] Martin A Nowak and Robert M May. Evolutionary games and spatial chaos. *Nature*, 359(6398):826, 1992.

[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.