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

Nikoleta E. Glynatsi[a, 1] **and Vince A. Knight**[a, 2]

[a]Cardiff University, School of Mathematics, Cardiff, United Kingdom

This manuscript was compiled on October 30, 2019

Memory-one strategies are a set of Iterated Prisoner's Dilemma strategies that have been acclaimed for their mathematical tractability and performance against single opponents. This manuscript investigates *best responses* to a collection of memory-one strategies as a multidimensional optimisation problem. Though extortionate memory-one strategies have gained much attention, we demonstrate that best response memory-one strategies do not behave in an extortionate way, and moreover, 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 their limited memory in multi agent interactions and can be out performed by longer memory strategies.

Prisoner's Dilemma | zero-determinant | best responses | memory-one | extortionate |

The Prisoner's Dilemma (PD) is a two player game used in understanding the evolution of cooperative behaviour, formally introduced in (1). 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:

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

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 $T > R > P > S$ and $2R > T + S$, and the most common values used in the literature are $(R, P, S, T) = (3, 1, 0, 5)$ (2). 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).

Memory-one strategies are a set of IPD strategies that have been studied thoroughly in the literature (3, 4), however, they have gained most of their attention when a certain subset of memory-one strategies was introduced in (5), the zero-determinants. In (6) it was stated that "Press and Dyson have fundamentally changed the viewpoint on the Prisoner's Dilemma". Zero-determinants are a special case of memory-one and extortionate strategies. They choose their actions so that a linear relationship is forced between the players' score ensuring that they will always receive at least as much as their opponents. Zero-determinants are indeed mathematically unique and are proven to be robust in pairwise interactions, however, their true effectiveness in tournaments and evolutionary dynamics has been questioned (7–12).

In a similar fashion to (5) the purpose of this work is to consider a given memory-one strategy; however, whilst (5) 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 given group of opponents. In particular, this work presents a compact method of identifying the best response memory-one strategy against a given set of opponents, and evaluates whether it behaves extortionately, similar to zero-determinants. This is also done for evolutionary settings. Moreover, we introduce a well designed framework that allows the comparison of an optimal memory one strategy and a more complex strategy which has a larger memory and an identification of conditions for which defection is known to be stable; thus identifying environments where cooperation will not occur.

Methods and Results

Utility. One specific advantage of memory-one strategies is their mathematical tractability. They can be represented completely as an element of $\mathbb{R}^4_{[0,1]}$. This originates from (13) 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; both players cooperated (CC), the first player cooperated whilst the second player defected (CD), the first player defected whilst the second player cooperated (DC) and both players defected (DD). 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 (13) it was shown that it is not necessary to simulate the play of a strategy p against a memory-one opponent q. Rather this exact behaviour can be modeled as a stochastic process, and more specifically as a Markov chain whose corresponding transition matrix M is given by (2). The long run steady state probability vector v, which is the solution to $vM = v$, can be combined with the payoff matrices of (1) to give the expected payoffs for each player. More specifically, the utility for a memory-one strategy p against an opponent q, denoted as $u_q(p)$, is given by (3).

$$M = \begin{bmatrix} p_1q_1 & p_1(-q_1+1) & q_1(-p_1+1) & (-p_1+1)(-q_1+1) \\ p_2q_3 & p_2(-q_3+1) & q_3(-p_2+1) & (-p_2+1)(-q_3+1) \\ p_3q_2 & p_3(-q_2+1) & q_2(-p_3+1) & (-p_3+1)(-q_2+1) \\ p_4q_4 & p_4(-q_4+1) & q_4(-p_4+1) & (-p_4+1)(-q_4+1) \end{bmatrix} \qquad [2]$$

$$u_q(p) = v \cdot (R, S, T, P). \qquad [3]$$

Please provide details of author contributions here.

Please declare any competing interests here.

[1]A.O.(Author One) contributed equally to this work with A.T. (Author Two) (remove if not applicable).

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

This manuscript has explored the form of $u_q(p)$, to the authors knowledge no previous work has done this, and it proves that $u_q(p)$ is given by a ratio of two quadratic forms (14), (Theorem 2):

$$u_q(p) = \frac{\frac{1}{2}pQp^T + cp + a}{\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}}, \quad [4]$$

where $Q = Q(q), \bar{Q} = \bar{Q}(q) \in \mathbb{R}^{4\times4}$, $c = c(q)$ and $\bar{c} = \bar{c}(q)$ $\in \mathbb{R}^{4\times1}$, $a = a(q)$ and $\bar{a} = \bar{a}(q) \in \mathbb{R}$.

This 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:

$$\frac{1}{N}\sum_{i=1}^{N} {u_q}^{(i)}(p) = \frac{1}{N} \quad [5]$$

$$\frac{\sum\limits_{i=1}^{N}(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}) \prod\limits_{\substack{j=1 \\ j \neq i}}^{N} (\frac{1}{2}p\bar{Q}^{(j)}p^T + \bar{c}^{(j)}p + \bar{a}^{(j)})}{\prod\limits_{i=1}^{N}(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)})}. \quad [6]$$

Estimating the utility of a memory-one strategy against any number of opponents without simulating the interactions is the main result used in this manuscript. It will be used to define best response memory-one strategies, in tournaments and evolutionary dynamics, and to explore the conditions under which defection dominates cooperation.

Stability of defection. An immediate result from our formulation can be obtained by evaluating the sign of the utility's (5) derivative at $p = (0,0,0,0)$. If at that point the derivative is negative, then the utility of a player only decreases if they were to change their behaviour, and thus **defection at that point is stable**.

Lemma 1. *In a tournament of N players* $\{q^{(1)}, q^{(2)}, \dots, q^{(N)}\}$ *for* $q^{(i)} \in \mathbb{R}^4_{[0,1]}$ *defection is stable if the transition probabilities of the opponents satisfy conditions (7) and (8).*

$$\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \leq 0 \quad [7]$$

while,

$$\sum_{i=1}^{N}\bar{a}^{(i)} \neq 0 \quad [8]$$

Proof. For defection to be 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 (19) gives:

$$\sum_{i=1}^{N}\frac{(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)})}{(\bar{a}^{(i)})^2} \quad [9]$$

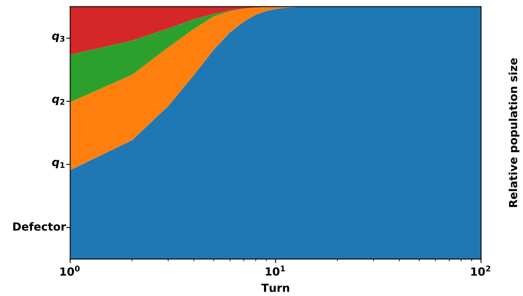

Fig. 1. For opponents $q_1 = (\frac{371}{1250}, \frac{4693}{25000}, \frac{4037}{50000}, \frac{18461}{25000})$, $q_2 = (\frac{48841}{100000}, \frac{30587}{50000}, \frac{76591}{100000}, \frac{25921}{50000})$ and $q_3 = (\frac{22199}{100000}, \frac{87073}{100000}, \frac{646}{3125}, \frac{91861}{100000})$ conditions (7) and (8) hold and Defector takes over the population.

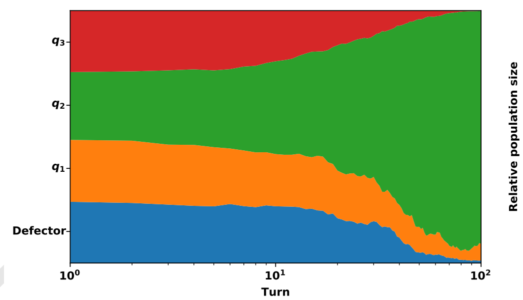

Fig. 2. For opponents $q_1 = (\frac{69773}{100000}, \frac{21609}{100000}, \frac{97627}{100000}, \frac{623}{100000})$, $q_2 = (\frac{12649}{50000}, \frac{43479}{100000}, \frac{38969}{50000}, \frac{19769}{100000})$ and $q_3 = (\frac{96703}{100000}, \frac{54723}{100000}, \frac{24317}{25000}, \frac{35741}{50000})$ (7) fails and (8) holds and Defector does not take over the population.

The sign of the numerator $\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)})$ can vary based on the transition probabilities of the opponents. The denominator can not be negative, and otherwise is always positive. Thus the sign of the derivative is negative if and only if $\sum_{i=1}^{N}(c^{(i)T}\bar{a}^{(i)} - \bar{c}^{(i)T}a^{(i)}) \leq 0$. □

Consider a population for which defection is known to be stable. In that population all the members will over time adopt the same behaviour; thus in such population cooperation will never take over. This is demonstrated in Fig. 1 and Fig. 2.

Best response memory-one strategies. As briefly discussed zero-determinants have been acclaimed for their robustness against a single opponent. Zero-determinants are evidence that extortion works in pairwise interactions, their behaviour ensures that the strategies will never lose a game. However, this paper argues that in multi opponent interactions, where the payoffs matter, strategies trying to exploit their opponents will suffer. Compared to zero-determinants, best response memory-one strategies which have a theory of mind of their opponents, utilise their behaviour in order to gain the most from their interactions. The question that arises then is whether best response strategies are optimal because they behave in an extortionate way.

To answer this very question, we initially define *memory-one best response* strategies as a multi dimensional optimisation problem given by:

$$\max_p : \sum_{i=1}^{N} {u_q}^{(i)}(p) \tag{10}$$
$$\text{such that} : p \in \mathbb{R}_{[0,1]}$$

A *best response* is a strategy which corresponds to the most favorable outcome (15), thus a memory-one best response to a set of opponents $q^{(1)}, q^{(2)}, \dots, q^{(N)}$ corresponds to a strategy p^* for which (5) is maximised.

Optimising this particular ratio of quadratic forms is not trivial. It can be verified empirically for the case of a single opponent that there exists at least one point for which the definition of concavity does not hold. The non concavity of $u(p)$ indicates multiple local optimal points. This is also intuitive. The best response against a cooperator, $q = (1, 1, 1, 1)$, is a defector $p^* = (0, 0, 0, 0)$. The strategies $p = (\frac{1}{2}, 0, 0, 0)$ and $p = (\frac{1}{2}, 0, 0, \frac{1}{2})$ are also best responses. The approach taken here is to introduce a compact way of constructing the candidate set of all local optimal points, and evaluating which corresponds to the best response strategy. The approach is given in Theorem 3.

Finding best response memory-one strategies is analytically feasible using the formulation of Theorem 3 and resultant theory (16). However, for large systems building the resultant quickly becomes intractable. As a result, best responses will be estimated heuristically using a numerical method taking advantage of the structure called Bayesian optimisation (17).

This is extended to evolutionary settings. In these settings self interactions are key. Self interactions can be incorporated in the formulation that has been used so far. More specifically, the optimisation problem of (10) is extended to include self interactions:

$$\max_p : \frac{1}{N} \sum_{i=1}^{N} {u_q}^{(i)}(p) + u_p(p) \tag{11}$$
$$\text{such that} : p \in \mathbb{R}_{[0,1]}$$

For determining the memory-one best response in an evolutionary setting, an algorithmic approach is considered, called *best response dynamics*. The best response dynamics approach used in this manuscript is given by Algorithm 1.

$p^{(t)} \leftarrow (1, 1, 1, 1)$;
while $p^{(t)} \neq p^{(t-1)}$ **do**
  $p^{(t+1)} = \text{argmax} \frac{1}{N} \sum\limits_{i=1}^{N} {u_q}^{(i)}(p^{(t+1)}) + u_p^{(t)}(p^{(t+1)})$;
end

Algorithm 1: Best response dynamics Algorithm

The results of this section rely on estimating best response memory-one strategies. Bayesian optimisation is used to generated a data set of best response memory-one strategies, in tournaments and evolutionary dynamics whilst $N = 2$. The data set is available at (21). It contains a total of 1000 trials corresponding to 1000 different instances of a best response strategy in tournaments and evolutionary dynamics. For each trial a set of 2 opponents is randomly generated and the memory-one best responses against them is found.

The source code used in this manuscript has been written in a sustainable manner. It is open source (https://github.com/Nikoleta-v3/Memory-size-in-the-prisoners-dilemma) and tested which ensures the validity of the results. It has also been archived and can be found at.

	mean	std	5%	50%	95%	max	median	skew	kurt
tournament	0.34	0.40	0.028	0.17	1.05	2.47	0.17	1.87	3.60
evolutionary setting	0.17	0.23	0.01	0.12	0.67	1.53	0.12	3.42	1.92

Table 1. SSE of best response memory-one when $N = 2$

In order to investigate whether best responses behave in an extortionate matter the SSE method as described in (18) is used. The SSE is defined as how far a strategy is from behaving as a ZD and thus a highand a high SSE implies a non extortionate behaviour. The SSE method has been applied to the data set. The distribution of SSE for the best response is given in A statistics summary of the SSE distribution for the best response in tournaments and evolutionary dynamics is given in Table 1.

For the best response in tournaments the distribution of SSE is skewed to the left, indicating that the best response does exhibit extortionate behaviour, however, the best response is not uniformly extortionate. A positive measure of skewness and kurtosis indicates a heavy tail to the right. Therefore, in several cases the strategy is not trying to extort its the opponents. Similarly the evolutionary best response strategy does not behave uniformly extortionately. A larger value of both the kurtosis and the skewness of the SSE distribution indicates that in evolutionary settings a memory-one best response is even more adaptable.

The difference between best responses in tournaments and in evolutionary settings are further explored by Fig. 3. Though, no statistically significant differences has been found, from Fig. 3, it seems that evolutionary best response has a higher p_2 median. Thus, they more likely to forgive after being tricked. This is due to the fact that they could be playing against themselves, and they need to be able to forgive so that future cooperation can occur.

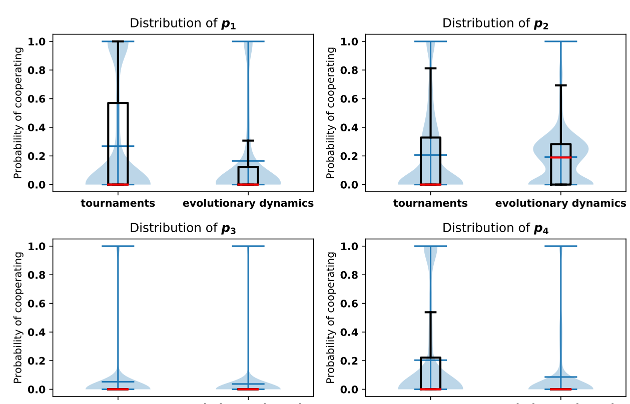

Fig. 3. Distributions of p^* for both best response and evo memory-one strategies.

Longer memory best response. This section focuses on the memory size of strategies. The effectiveness of memory in the IPD has been previously explored in the literature, however, none of the works mentioned here have compared the performance of longer-memory strategies to memory-one best responses.

In (19), a strategy called *Gambler* which makes probabilistic decisions based on the opponent's n_1 first moves, the opponent's m_1 last moves and the player's m_2 last moves was introduced. In this manuscript Gambler with parameters: $n_1 = 2, m_1 = 1$ and $m_2 = 1$ is used as a longer-memory strategy. By considering the opponent's first two moves, the opponents last move and the player's last move, there are only 16 ($4 \times 2 \times 2$) possible outcomes that can occur, furthermore, Gambler also makes a probabilistic decision of cooperating in the opening move. Thus, Gambler is a function $f : \{\text{C, D}\} \to [0,1]_{\mathbb{R}}$. This can be hard coded as an element of $[0,1]_{\mathbb{R}}^{16+1}$, one probability for each outcome plus the opening move. Hence, compared to (10), finding an optimal Gambler is a 17 dimensional problem given by:

$$\max_{p} : \sum_{i=1}^{N} {U_q}^{(i)}(f) \quad \text{such that} : f \in \mathbb{R}_{[0,1]}^{17} \qquad [12]$$

Note that (5) can not be used here for the utility of Gambler, and actual simulated players are used. This is done using (20) with 500 turns and 200 repetitions, moreover, (12) is solved numerically using Bayesian optimisation.

Similarly to previous sections, a large data set has been generated with instances of an optimal Gambler and a memory-one best response, available at (21). Estimating a best response Gambler (17 dimensions) is computational more expensive compared to a best response memory-one (4 dimensions). As a result, the analysis of this section is based on a total of 130 trials. For each trial two random opponents have been selected. The 130 pair of opponents are a sub set of the opponents used in the previous section.

The utilities of both strategies are plotted against each other in Fig. 4. It is evident from Fig. 4 that Gambler always performs as well as the best response memory-one or better. This seems to be at odd with the result of (5) that against a memory-one opponent having a longer memory will not give a strategy any advantage. However, against two memory-one opponents, Gambler's performance is better than the optimal memory-one strategy. This is evidence that in the case of two opponents having a shorter memory is limiting and this is potentially another example of the advantages of adaptability.

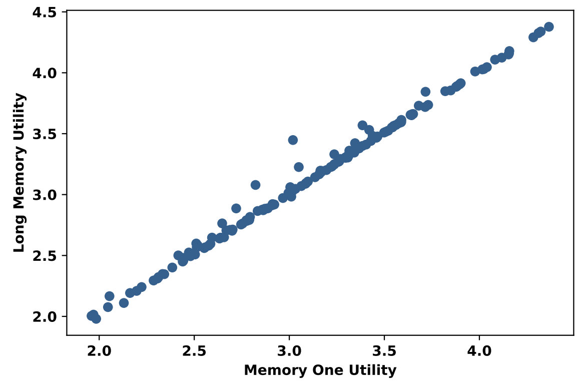

Fig. 4. Utilities of Gambler and best response memory-one strategies for 130 different pair of opponents.

Discussion

This manuscript has considered *best response* strategies in the IPD game, and more specifically, *memory-one best responses*. It has proven that there is a compact way of identifying a memory-one best response to a group of opponents, and moreover, that there exists a condition for which in an environment of memory-one opponents defection is the stable choice. The later parts of this paper focused on a series of empirical results, where it was shown that the performance and the evolutionary stability of memory-one strategies rely not on extortion but on adaptability. Finally, it was shown that memory-one strategies' performance is limited by their memory in cases where they interact with multiple opponents.

Following the work described in (13), where it was shown that the utility between two memory-one strategies can be estimated by a Markov stationary state, we proved that the utilities can be written as a ration of two quadratic forms in R^4, Theorem 2. This was extended to include multiple opponents, as the IPD is commonly studied in such situations. This formulation allowed us to introduce an approach for identifying memory-one best responses to any number of opponents; Theorem 3. This does not only have game theoretic novelty, but also a mathematical novelty of solving quadratic ratio optimisation problem where the quadratics are non concave. The results of were also used to define a condition for which defection is known to be stable.

This manuscript presented several experimental results. All data for the results is archived in (21). These results were mainly to investigate the behaviour of memory-one strategies and their limitations. A large data set which contained best responses in tournaments and in evolutionary settings for $N = 2$ was generated. This allowed us to investigate their respective behaviours, and whether it was extortionate acts that made them the most favorable strategies. However, it was shown that it was not extortion but adaptability that allowed the strategies to gain the most from their interactions. In evolutionary settings it was specifically shown that being adaptable and being able to forgive after being tricked were key factors. Moreover, the performance of memory-one strategies was put against the performance of a longer memory strategy called Gambler. There were several cases where Gambler would outperform the memory-one strategy, however, a memory-one strategy never managed to outperform a Gambler. This result occurred whilst considering a Gambler with a sufficiently larger memory but not a sufficiently larger amount of information regarding the game.

All the empirical results presented in this manuscript have been for the case of $N = 2$. In future work we would consider larger values of N, however, we believe that for larger values of N the results that have been presented here would only be more evident. In addition, we would investigate potential theoretical results for the evolutionary best responses dynamics algorithm discussed.

Appendix

Theorem 2. *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:*

$$u_q(p) = \frac{\frac{1}{2}pQp^T + cp + a}{\frac{1}{2}p\bar{Q}p^T + \bar{c}p + \bar{a}}, \qquad [13]$$

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

$$Q = \begin{bmatrix} 0 & -(q_1 - q_3)(q_2 - 5q_4 - 1) & q_3(q_1 - q_2) & -5q_3(q_1 - q_4) \\ -(q_1 - q_3)(q_2 - 5q_4 - 1) & 0 & (q_2 - q_3)(q_1 - 3q_4 - 1) & (q_3 - q_4)(5q_1 - 3q_2 - 2) \\ q_3(q_1 - q_2) & (q_2 - q_3)(q_1 - 3q_4 - 1) & 0 & 3q_3(q_2 - q_4) \\ -5q_3(q_1 - q_4) & (q_3 - q_4)(5q_1 - 3q_2 - 2) & 3q_3(q_2 - q_4) & 0 \end{bmatrix}, \qquad [14]$$

$$\bar{Q} = \begin{bmatrix} 0 & -(q_1 - q_3)(q_2 - q_4 - 1) & (q_1 - q_2)(q_3 - q_4) & (q_1 - q_4)(q_2 - q_3 - 1) \\ -(q_1 - q_3)(q_2 - q_4 - 1) & 0 & (q_2 - q_3)(q_1 - q_4 - 1) & (q_1 - q_2)(q_3 - q_4) \\ (q_1 - q_2)(q_3 - q_4) & (q_2 - q_3)(q_1 - q_4 - 1) & 0 & -(q_2 - q_4)(q_1 - q_3 - 1) \\ (q_1 - q_4)(q_2 - q_3 - 1) & (q_1 - q_2)(q_3 - q_4) & -(q_2 - q_4)(q_1 - q_3 - 1) & 0 \end{bmatrix}. \qquad [15]$$

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

$$c = \begin{bmatrix} q_1(q_2 - 5q_4 - 1) \\ -(q_3 - 1)(q_2 - 5q_4 - 1) \\ -q_1q_2 + q_2q_3 + 3q_2q_4 + q_2 - q_3 \\ 5q_1q_4 - 3q_2q_4 - 5q_3q_4 + 5q_3 - 2q_4 \end{bmatrix}, \qquad [16]$$

$$\bar{c} = \begin{bmatrix} q_1(q_2 - q_4 - 1) \\ -(q_3 - 1)(q_2 - q_4 - 1) \\ -q_1q_2 + q_2q_3 + q_2 - q_3 + q_4 \\ q_1q_4 - q_2 - q_3q_4 + q_3 - q_4 + 1 \end{bmatrix}, \qquad [17]$$

and the constant terms $a, \bar{a}$ are defined as $a = -q_2 + 5q_4 + 1$ and $\bar{a} = -q_2 + q_4 + 1$.

Proof is given in (SI).

Theorem 3. *The optimal behaviour of a memory-one strategy player $p^* \in \mathbb{R}^4_{[0,1]}$ against a set of N opponents $\{q^{(1)}, q^{(2)}, \dots, q^{(N)}\}$ for $q^{(i)} \in \mathbb{R}^4_{[0,1]}$ is given 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:

$$S_q = \left\{ p \in \mathbb{R}^4 \left| \begin{array}{l} \bullet \quad p_j \in \{0,1\} \quad \textit{and} \quad \dfrac{d}{dp_k} \displaystyle\sum_{i=1}^{N} u_q^{(i)}(p) = 0 \\ \qquad\qquad \textit{for all} \quad j \in J \quad \& \quad k \in K \quad \textit{for all} \quad J, K \\ \qquad\qquad \textit{where} \quad J \cap K = \text{‰} \quad \textit{and} \quad J \cup K = \{1,2,3,4\}. \\ \bullet \quad p \in \{0,1\}^4 \end{array} \right. \right\}. \qquad [18]$$

Note that there is no immediate way to find the zeros of $\frac{d}{dp}\sum_{i=1}^{N} u_q(p)$ where,

$$\frac{d}{dp}\sum_{i=1}^{N} {u_q}^{(i)}(p) = \sum_{i=1}^{N} \frac{\left(pQ^{(i)} + c^{(i)}\right)\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)}{\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)^2} - \frac{\left(p\bar{Q}^{(i)} + \bar{c}^{(i)}\right)\left(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}\right)}{\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)^2} \qquad [19]$$

For $\frac{d}{dp}\sum_{i=1}^{N} u_q(p)$ to equal zero then:

$$\sum_{i=1}^{N} \left(pQ^{(i)} + c^{(i)}\right)\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right) \qquad [20]$$

$$- \left(p\bar{Q}^{(i)} + \bar{c}^{(i)}\right)\left(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}\right) \quad = 0, \quad \textit{while} \qquad [21]$$

$$\sum_{i=1}^{N} \frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)} \neq 0. \qquad [22]$$

Proof is given in (SI).

ACKNOWLEDGMENTS. A variety of software libraries have been used in this work, the Axelrod library for IPD simulations (20), the Scikit-optimize library for an implementation of Bayesian optimisation (22), the Matplotlib library for visualisation (23), the SymPy library for symbolic mathematics (24) and the Numpy library for data manipulation (25).

1. MM Flood, Some experimental games. *Manag. Sci.* **5**, 5–26 (1958).
2. R Axelrod, WD Hamilton, The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
3. M Nowak, K Sigmund, The evolution of stochastic strategies in the prisoner's dilemma. *Acta Appl. Math.* **20**, 247–265 (1990).
4. M Nowak, K Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56 (1993).
5. WH Press, FJ Dyson, Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proc. Natl. Acad. Sci.* **109**, 10409–10413 (2012).
6. AJ Stewart, JB Plotkin, Extortion and cooperation in the prisoner's dilemma. *Proc. Natl. Acad. Sci.* **109**, 10134–10135 (2012).
7. C Adami, A Hintze, Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nat. communications* **4**, 2193 (2013).
8. C Hilbe, M Nowak, K Sigmund, Evolution of extortion in iterated prisoner's dilemma games. *Proc. Natl. Acad. Sci.*, 201214834 (2013).
9. C Hilbe, M Nowak, A Traulsen, Adaptive dynamics of extortion and compliance. *PLOS ONE* **8**, 1–9 (2013).
10. C Hilbe, A Traulsen, K Sigmund, Partners or rivals? strategies for the iterated prisoner's dilemma. *Games Econ. Behav.* **92**, 41 – 52 (2015).
11. V Knight, M Harper, NE Glynatsi, O Campbell, Evolution reinforces cooperation with the emergence of self-recognition mechanisms: An empirical study of strategies in the moran process for the iterated prisoner's dilemma. *PLOS ONE* **13**, 1–33 (2018).
12. C Lee, M Harper, D Fryer, The art of war: Beyond memory-one strategies in population games. *PLOS ONE* **10**, 1–16 (2015).
13. M Nowak, K Sigmund, Game-dynamical aspects of the prisoner's dilemma. *Appl. Math. Comput.* **30**, 191–213 (1989).
14. J Kepner, J Gilbert, *Graph algorithms in the language of linear algebra.* (SIAM), (2011).
15. S Tadelis, *Game theory: an introduction.* (Princeton University Press), (2013).
16. G Jonsson, S Vavasis, Accurate solution of polynomial equations using macaulay resultant matrices. *Math. computation* **74**, 221–262 (2005).
17. J Močkus, On bayesian methods for seeking the extremum in *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974.* (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 400–404 (1975).
18. VA Knight, M Harper, NE Glynatsi, J Gillard, Recognising and evaluating the effectiveness of extortion in the iterated prisoner's dilemma. *CoRR* **abs/1904.00973** (2019).
19. M Harper, et al., Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PLOS ONE* **12**, 1–33 (2017).
20. The Axelrod project developers , Axelrod: 4.4.0 (2016).
21. NE Glynatsi, Raw data for: "Stability of defection, optimisation of strategies and the limits of memory in the Prisoner's Dilemma." (2019).
22. T Head, et al., scikit-optimize/scikit-optimize: v0.5.2 (2018).
23. JD Hunter, Matplotlib: A 2D graphics environment. *Comput. In Sci. & Eng.* **9**, 90–95 (2007).
24. A Meurer, et al., Sympy: symbolic computing in python. *PeerJ Comput. Sci.* **3** (2017).
25. S Van Der Walt, SC Colbert, G Varoquaux, The NumPy array: a structure for efficient numerical computation. *Comput. Sci. & Eng.* **13**, 22–30 (2011).