Using a theory of mind to find best responses to memory-one strategies.

Nikoleta E. Glynatsi, Vincent A. Knight

September 22, 2020

We would like to open this response by thanking the reviewers for their thoughtful comments and suggestions. We have fully taken their comments on board and made significant modifications and additions to the manuscript. We feel this has greatly improved the work. Given the nature of these changes, we hope that the editorial team are willing to overlook the fact that they stipulated that if any further modification are required the manuscript would be rejected. Hopefully if further minor modifications are needed we will be given the opportunity to make them.

Both reviewers commented on different aspects of the paper and so our modifications naturally fit in to two categories (which we will discuss in detail comment by comment):

- The first reviewer suggested that the manuscript lacked discussion of noise and of the literature on the theory of mind. We address this including noise in our formulation (in the appendix) and by discussing relevant literature on the theory of mind (in the introduction).
- The second reviewer questioned the claims on the evolutionary stability of the best response and expressed that the manuscript would benefit from such experiments and discussion. To address this, we have have obtained expressions for the fixation probability of a best response strategy in a Moran process that adapts dynamically to the population.

We will now take each comment of the reviewers in turn and highlight our efforts to improve the work.

1 Response

Regarding the comments of Reviewer 1:

> ‘‘ Noise or error (in implementing a move, e.g. due to trembling hand or fuzzy mind) is a crucial aspect in the context of the IPD. Given that noise is unavoidable in real-world applications of the IPD, it’s therefore important to take them into account. A relevant question is whether similar observations will be seen in noisy IPD? Authors should consider or at least discuss this important issues of IPD.’’

We agree with the reviewer that noise is an important aspect of research. Noise is now included in the discussion, specifically pointing at the appendix where we have obtained expressions for the utilities with noise.

> ‘‘The paper lacks sufficient discussion of previous works, especially regarding the literature of theory-of-mind and complex strategies in repeated games. Authors should consider discussing this highly literature to improve the relevance of the paper.’’

We have discussed several articles that have been proposed by the reviewer.

Regarding the comments of Reviewer 2:

> ‘‘The structure despite having been improved quite a bit, still leaves something to be desired. For example, the contents of Section 1.2 were not fully moved into the appendix, but instead partially pop up in the discussion section. This interrupts the flow of the paper, as the discussion section is not the place for a new lemma. It would be better to make a reference to this result as a corollary to Theorem 1, and leave all details and Figure 4 to the appendix.’’

We have made changes in order to improved the narrative. This includes moving Section 1.2 to the appendix, changing the order of our results and removing the discussion of tournaments with self interactions.

> On that note, the lack of evolutionary results is somewhat disappointing, and the current results give no intuition if such a best response strategy, as described in the text, would arise in the evolutionary trajectory. The original manuscript was somewhat misleading in this regard, and even the revised version is a bit confusing, by mixing best response *dynamics*, best response strategies, and vague hints to learning/stochastic processes such as the Moran process into the results section without actually presenting results on this issue. Self-interactions do not equal an evolutionary setting, and best response does not automatically equal evolutionary stability. It would be crucial to clear up any confusion there, starting with further purging misleading references to any "evolutionary setting" from the text (e.g. the authors missed this in the caption for Fig.2).

This particular comment proved fruitful for us as it lead to many discussions and some interesting new work. Thank to the reviewer.

We note that we have not addressed whether or not a best response strategy would arise in the evolutionary trajectory. This is because the best response strategy is obtained using the quadratic ratio, essentially transforming the problem to a finite dimensional optimisation problem. As such, an evolutionary algorithm could be used to find the best response. However, in our particular case Bayesian optimisation is the tool chosen. This does not offer insight as to whether or not the best response strategy is on the evolutionary trajectory. We have added a comment about this in the conclusion as it would prove to be an interesting avenue for future work.

We have addressed the latter parts of this comment with a significant addition of work to the manuscript. We no longer refer vaguely to Moran processes but instead explicitly consider a Moran process where the best response player is able to dynamically adapt to the population distribution (and a classic Moran process). As well as the theoretic expressions obtained (modifications of the standard results on Moran processes) this is coupled with a data set of experiments on the Moran process. We feel that this is a strong addition to the manuscript that ensures there is no further confusion.

2 Marked changes made to the manuscript

We are also attaching a copy of the revised manuscript where we have highlighted the changes made to the originally submitted manuscript.

Using a theory of mind to find best responses to memory-one strategies.

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

[1]Cardiff University, School of Mathematics, Cardiff, CF24 4AG, United Kingdom
[2]Max Planck Institute for Evolutionary Biology, Plön, 24 306, Germany
[*]glynatsi@evolbio.mpg.de

ABSTRACT

Memory-one strategies are a set of Iterated Prisoner's Dilemma strategies that have been praised for their mathematical tractability and performance against single opponents. This manuscript investigates *best response* memory-one strategies with a theory of mind for their opponents. The results add to the literature that has shown that extortionate play is not always optimal by showing that optimal play is often not extortionate. They also provide evidence that memory-one strategies suffer from their limited memory in multi agent interactions and can be out performed by optimised strategies with longer memory. We have developed a theory that has allowed to explore the entire space of memory-one strategies. The framework presented is suitable to study memory-one strategies in the Prisoner's Dilemma, but also in evolutionary processes such as the Moran process. Furthermore, results on the stability of defection in populations of memory-one strategies are also obtained.

Introduction

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:

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

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 $T > R > P > S$ and $2R > T + S$, and the most common values used in the literature are $(R,P,S,T) = (3,1,0,5)$[2]. The numerical experiments of our manuscript are carried out using these payoff values. 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-determinant strategies (ZDs). In[6] it was stated that "Press and Dyson have fundamentally changed the viewpoint on the Prisoner's Dilemma". A special case of ZDs are extortionate strategies that 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. ZDs 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 [5] the authors stated that "Only a player with a theory of mind about his opponent can do better, in which case Iterated Prisoner's Dilemma is an Ultimatum Game". ~~The purpose of this work is to ~~investigate the first part of this sentence, more specifically, to identify the best response strategy~~ reinforce the literature on the limitations of extortionate strategies by considering a new approach. More specifically, by considering best response memory-one strategies with a theory of mind of ~~a given group of opponents. The outcomes of our work reinforce known results, namely that memory-one strategies must be forgiving to be evolutionarily stable [?,?] and that longer-memory strategies have a certain form of advantage over short memory strategies [?,?]~~their opponents. There are several works in the literature that have considered strategies with a theory of mind [5,6,13–16]. These works defined "theory of mind" as intention recognition [13–16] and as the ability of a strategy to realise that their actions can influence opponents [6]. Compared to these works, theory of mind is defined here as the ability of a strategy to know the behaviour of their opponents and alter their own behaviour in response to that.

~~In particular, this work presents~~ We present a closed form algebraic expression for the utility of a memory-one strategy against a given set of opponents ~~,~~ and a compact method of identifying it's best response to that given set of opponents~~essentially:~~

~~a theory of mind~~. The aim is to evaluate whether a best response memory-one strategy behaves in a zero-determinant way which in turn indicates whether it can be extortionate. We do this using a linear algebraic approach presented in[17]. This is done in tournaments with ~~and without self interactions~~two opponents. Moreover, we introduce a framework that allows the comparison of an optimal memory-one strategy and an optimised strategy which has a larger memory.

To illustrate the analytical results obtained in this manuscript a number of numerical experiments are run. The source code for these experiments has been written in a sustainable manner[18]. 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 ~~?~~[19].

Methods

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[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; 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[20] 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 Equation (2). The long run steady state probability vector v, which is the solution to $vM = v$, can be combined with the payoff matrices of Equation (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 Equation (3).

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

$$u_q(p) = v \cdot (R, S, T, P). \tag{3}$$

This manuscript has explored the form of $u_q(p)$, to the best of the authors knowledge no previous work has done this, and Theorem 1 states that $u_q(p)$ is given by a ratio of two quadratic forms[21].

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

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

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

$$Q = \begin{bmatrix} 0 & -(q_1-q_3)(Pq_2-P-Tq_4) & (q_1-q_2)(Pq_3-Sq_4) & (q_1-q_4)(Sq_2-S-Tq_3) \\ -(q_1-q_3)(Pq_2-P-Tq_4) & 0 & (q_2-q_3)(Pq_1-P-Rq_4) & -(q_3-q_4)(Rq_2-R-Tq_1+T) \\ (q_1-q_2)(Pq_3-Sq_4) & (q_2-q_3)(Pq_1-P-Rq_4) & 0 & (q_2-q_4)(Rq_3-Sq_1+S) \\ (q_1-q_4)(Sq_2-S-Tq_3) & -(q_3-q_4)(Rq_2-R-Tq_1+T) & (q_2-q_4)(Rq_3-Sq_1+S) & 0 \end{bmatrix}, \tag{5}$$

$$\bar{Q} = \begin{bmatrix} 0 & -(q_1-q_3)(q_2-q_4-1) & (q_1-q_2)(q_3-q_4) & (q_1-q_4)(q_2-q_3-1) \\ -(q_1-q_3)(q_2-q_4-1) & 0 & (q_2-q_3)(q_1-q_4-1) & (q_1-q_2)(q_3-q_4) \\ (q_1-q_2)(q_3-q_4) & (q_2-q_3)(q_1-q_4-1) & 0 & -(q_2-q_4)(q_1-q_3-1) \\ (q_1-q_4)(q_2-q_3-1) & (q_1-q_2)(q_3-q_4) & -(q_2-q_4)(q_1-q_3-1) & 0 \end{bmatrix}. \tag{6}$$

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

$$c = \begin{bmatrix} q_1\left(Pq_2 - P - Tq_4\right) \\ -\left(q_3 - 1\right)\left(Pq_2 - P - Tq_4\right) \\ -Pq_1q_2 + Pq_2q_3 + Pq_2 - Pq_3 + Rq_2q_4 - Sq_2q_4 + Sq_4 \\ -Rq_2q_4 + Rq_4 + Sq_2q_4 - Sq_2 - Sq_4 + S + Tq_1q_4 - Tq_3q_4 + Tq_3 - Tq_4 \end{bmatrix}, \tag{7}$$

$$\bar{c} = \begin{bmatrix} q_1\left(q_2 - q_4 - 1\right) \\ -\left(q_3 - 1\right)\left(q_2 - q_4 - 1\right) \\ -q_1q_2 + q_2q_3 + q_2 - q_3 + q_4 \\ q_1q_4 - q_2 - q_3q_4 + q_3 - q_4 + 1 \end{bmatrix}, \tag{8}$$

and the constant terms $a, \bar{a}$ are defined as $a = -Pq_2 + P + Tq_4$ and $\bar{a} = -q_2 + q_4 + 1$.

The proof of Theorem 1 is given in the Supplementary Information. 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 N opponents is given by:

$$\frac{1}{N}\sum_{i=1}^{N} u_q{}^{(i)}(p) = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}\right)\prod_{\substack{j=1 \\ j\neq i}}^{N}\left(\frac{1}{2}p\bar{Q}^{(j)}p^T + \bar{c}^{(j)}p + \bar{a}^{(j)}\right)}{\prod_{i=1}^{N}\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)}. \tag{9}$$

Equation (9) is the average score (using Equation (4)) against the set of opponents.

Estimating the utility of a memory-one strategy against any number of opponents without simulating the interactions is the main result used in the rest of this manuscript. It will be used to obtain best response memory-one strategies in tournaments ~~with and without self interactions~~ in order to explore the limitations of extortion and restricted memory.

Results

~~Here we~~ The formulation as presented in Theorem 1 can be used to define *memory-one best response* strategies as a multi dimensional optimisation problem given by:

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

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 discrete candidate set of all local optimal points, and evaluating the objective function Equation (9). This gives the best response memory-one strategy. The approach is given in Theorem 2.

Theorem 2 *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 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:

$$S_q = \left\{ p \in \mathbb{R}^4 \;\middle|\; \begin{array}{l} \bullet \quad p_j \in \{0,1\} \quad \text{and} \quad \frac{d}{dp_k}\sum_{i=1}^{N} u_q^{(i)}(p) = 0 \\ \qquad\qquad\qquad\quad \text{for all} \quad j \in J \quad \& \quad k \in K \quad \text{for all} \quad J,K \\ \qquad\qquad\qquad\quad \text{where} \quad J \cap K = \quad \text{and} \quad J \cup K = \{1,2,3,4\}. \\ \bullet \quad p \in \{0,1\}^4 \end{array} \right\}. \tag{11}$$

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

$$\frac{d}{dp}\sum_{i=1}^{N} u_q{}^{(i)}(p) = \sum_{i=1}^{N} \frac{\left(pQ^{(i)} + c^{(i)}\right)\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)}{\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)^2} - \frac{\left(p\bar{Q}^{(i)} + \bar{c}^{(i)}\right)\left(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}\right)}{\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right)^2} \tag{12}$$

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

$$\sum_{i=1}^{N}\left(pQ^{(i)} + c^{(i)}\right)\left(\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)}\right) - \left(p\bar{Q}^{(i)} + \bar{c}^{(i)}\right)\left(\frac{1}{2}pQ^{(i)}p^T + c^{(i)}p + a^{(i)}\right) = 0, \quad \textit{while} \tag{13}$$

$$\sum_{i=1}^{N}\frac{1}{2}p\bar{Q}^{(i)}p^T + \bar{c}^{(i)}p + \bar{a}^{(i)} \neq 0. \tag{14}$$

The proof of Theorem 2 is given in the Supplementary Information. Finding best response memory-one strategies is analytically feasible using the formulation of Theorem 2 and resultant theory[22]. However, for large systems building the resultant becomes intractable. As a result, best responses will be estimated heuristically using a numerical method, suitable for problems with local optima, called Bayesian optimisation[23].

~~In several evolutionary settings such as Moran Processes self interactions are key. Previous work has identified interesting results such as the appearance of self recognition mechanisms when training strategies using evolutionary algorithms in Moran processes [11]. This aspect of reinforcement learning can be done for best response memory-one strategies by incorporating the strategy itself in the objective function as shown in Equation (10). K is the number of self interactions that will take place.~~

Limitations of extortionate behaviour

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

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

~~$p^{(t)} \leftarrow (1,1,1,1)$ Best response dynamics Algorithm~~

~~Using this approach it would be possible to create a memory-one best response strategy that updates on every generation of a Moran process to recalculate the optimal behaviour given the population. This extension of the "theory of mind" is an interesting avenue for future work.~~

In multi opponent settings, where the payoffs matter, strategies trying to exploit their opponents will suffer. Compared to ZDs, 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.

The results of this section use Bayesian optimisation to generate a data set of best response memory-one strategies for $N = 2$ opponents. The data set is available at[24]. It contains a total of 1000 trials corresponding to 1000 different instances of a best response strategy in tournaments with ~~and without self interactions~~$N = 2$. For each trial a set of 2 opponents is randomly generated and the memory-one best response against them is found. In order to investigate whether best responses behave in an extortionate matter the SSE method described in [17] is used. More specifically, in[17] the point x^*, in the space of memory-one strategies, that is the nearest extortionate strategy to a given strategy p is given by,

$$x^* = \left(C^T C\right)^{-1} C^T \bar{p} \tag{15}$$

where $\bar{p} = (p_1 - 1, p_2 - 1, p_3, p_4)$ and

$$C = \begin{bmatrix} R-P & R-P \\ S-P & T-P \\ T-P & S-P \\ 0 & 0 \end{bmatrix}. \tag{16}$$

Once this closest ZDs is found, the squared norm of the remaining error is referred to as sum of squared errors of prediction (SSE):

$$\text{SSE} = \bar{p}^T \bar{p} - \bar{p}C \left(C^T C\right)^{-1} C^T \bar{p} = \bar{p}^T \bar{p} - \bar{p}Cx^* \tag{17}$$

Thus, SSE is defined as how far a strategy is from behaving as a ZD. A high SSE implies a non extortionate behaviour. The ~~distributions~~ distribution of SSE for the best response in tournaments ($N = 2$) ~~with and without self interactions (with $K = 1$) are~~ is given in Figure 1. Moreover, a statistical summary of the SSE ~~distributions~~ distribution is given in Table 1.

mean	std	5%	50%	95%	max	median	skew	kurt
0.34	0.40	0.028	0.17	1.05	2.47	0.17	1.87	3.60

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

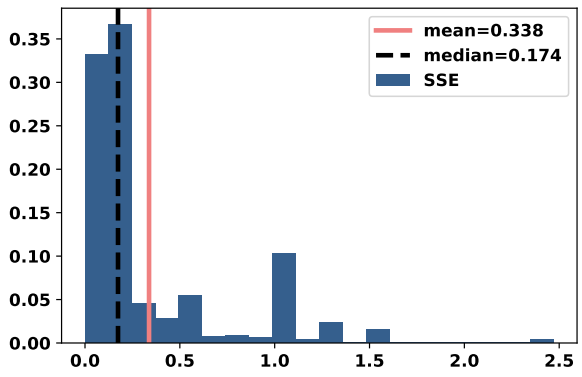

Figure 1. SEE distribution for best response in tournaments ~~without self interactions~~with $N = 2$. ~~SEE distribution for best response in tournaments with self interactions. SEE distributions for best response in tournaments without and with self interactions.~~

For the best response in tournaments ~~that do not include self interactions~~ with $N = 2$ the distribution of SSE is skewed to the left, indicating that the best response does exhibit ZDs behaviour and so could be extortionate, however, the best response is not uniformly a ZDs. A positive measure of skewness and kurtosis, and a mean of 0.34 indicate a heavy tail to the right. Therefore, in several cases the strategy is not trying to extort its opponents. ~~In [25] a similar behaviour is refereed to as the *partner strategy*. The partner strategy aims to share the payoff for mutual cooperation, but it is ready to fight back when being exploited. The partner strategy was designed, but the best responses which are defined by their opponents seem to exhibit the same behaviour.~~

~~Similarly, when considering self interactions, the distribution of SSE for~~ This highlights the importance of adaptability since the best response strategy ~~has skewness and kurtosis that indicate a heavy tail to the right. This indicates that evolutionary stable memory-one strategies need to more adaptable than a ZDs, and aim for mutual cooperation as well as exploitation which is in line with the results of [25] where their strategy was designed to adapt and was shown to be evolutionary stable. The findings of this work show that an optimal strategy acts in the same way.~~ against an opponent is rarely (if ever) a unique ZDs.

The difference between best responses in tournaments without and with self interactions is further explored by Figure **??**. Though, no statistically significant differences have been found, from Figure **??**, it seems that the best response that incorporate self interactions has a higher median p_2, which corresponds to the probability of cooperating after receiving a defection. Thus, they are 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.

Limitations of memory size

Distributions of p^* for best responses in tournaments and evolutionary settings. The medians, denoted as $\bar{p}^*$, for tournaments are $\bar{p}^* = (0,0,0,0)$, and for evolutionary settings $\bar{p}^* = (0,0.19,0,0)$.

The other main finding presented in[5] was that short memory of the strategies was all that was needed. We argue that the second limitation of ZDs in multi opponent interactions is that of their restricted memory. To demonstrate the effectiveness of memory in the IPD we explore a best response longer-memory strategy against a given set of memory-one opponents, and compare it's performance to that of a memory-one best response.

In[26], 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 $f : \{\mathrm{C}, \mathrm{D}\} \to [0,1]_{\mathbb{R}}$. 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 Equation (10), finding an optimal Gambler is a 17 dimensional problem given by:

$$\begin{aligned} &\max_p : \sum_{i=1}^{N} {U_q}^{(i)}(f) \\ &\text{such that} : f \in \mathbb{R}_{[0,1]}^{17} \end{aligned} \tag{18}$$

Note that Equation (9) can not be used here for the utility of Gambler, and actual simulated players are used. This is done using[27] with 500 turns and 200 repetitions, moreover, Equation (18) is solved numerically using Bayesian optimisation.

Similarly to previous sectionsthe previous section, a large data set has been generated with instances of an optimal Gambler and a memory-one best response, available at[24]. 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 152 trials. As before, for each trial $N = 2$ random opponents have been selected.

The ratio between Gambler's utility and the best response memory-one strategy's utility has been calculated and its distribution in given in Figure 2. It is evident from Figure 2 that Gambler always performs as well as the best response memory-one strategy and often performs better. There are no points where the ratio value is less than 1, thus Gambler never performed less than the best response memory-one strategy and in places outperforms it. 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 multiple opponents, having a shorter memory is limiting.

Dynamic best response player

In several evolutionary settings such as Moran Processes self interactions are key. Previous work has identified interesting results such as the appearance of self recognition mechanisms when training strategies using evolutionary algorithms in Moran processes [11]. This aspect of reinforcement learning can be done for best response memory-one strategies, as presented in this manuscript, by incorporating the strategy itself in the objective function as shown in Equation (10). Where K is the number of self interactions that will take place.

$$\begin{aligned} &\max_p : \frac{1}{N} \sum_{i=1}^{N} {u_q}^{(i)}(p) + K u_p(p) \\ &\text{such that} : p \in \mathbb{R}_{[0,1]} \end{aligned} \tag{19}$$

For determining the memory-one best response with self interactions, an algorithmic approach called *best response dynamics* is proposed. The best response dynamics approach used in this manuscript is given by Algorithm 1.

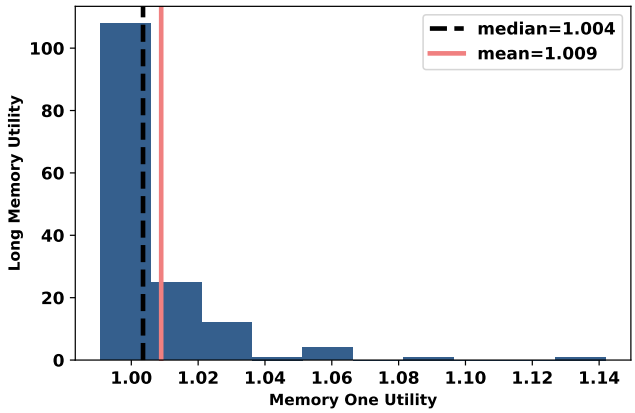

Figure 2. The ratio between the utilities of Gambler and best response memory-one strategy for 152 different pair of opponents.

Algorithm 1: Best response dynamics Algorithm

$p^{(t)} \leftarrow (1,1,1,1)$; **while** $p^{(t)} \neq p^{(t-1)}$ **do**

$p^{(t+1)} = \text{argmax} \frac{1}{N} \sum_{i=1}^{N} {u_q}^{(i)}(p^{(t)}) + K u_{p^{(t)}}(p^{(t)})$;

end

To investigate the effectiveness of this approach, more formally a Moran process will be considered. If a population of n total individuals of two types is considered, with K individuals of the first type and $n-K$ of the second type. The probability that the individuals of the first type will take over the population (the fixation probability) is denoted by x_K and is known to be [28]:

$$x_K = \frac{1 + \sum_{j=1}^{K-1} \prod_{i=1}^{j} \gamma_i}{1 + \sum_{j=1}^{n-1} \prod_{i=1}^{j} \gamma_i}$$

where:

$$\gamma_i = \frac{p_{K,K-1}}{p_{K,K+1}}.$$

To evaluate the formulation proposed here the best response player (taken to be the first type of individual in our population) will be allowed to act dynamically: adjusting their probability vector at every generation. In essence using the theory of mind to find the best response to not only the opponent but also the distribution of the population. Thus for every value of K there is a different best response player.

Considering the dynamic best response player as a vector $p \in \mathbb{R}^4_{[0,1]}$ and the opponent as a vector $q \in \mathbb{R}^4_{[0,1]}$, the transition probabilities depend on the payoff matrix $A^{(K)}$ where:

- $A^{(K)}_{11} = u_p(p)$ is the long run utility of the best response player against itself.
- $A^{(K)}_{12} = u_q(p)$ is the long run utility of the best response player against the opponent.
- $A^{(K)}_{11} = u_p(q)$ is the long run utility of the opponent against the best response player.
- $A^{(K)}_{11} = u_q(q)$ is the long run utility of the opponent against itself.

The matrix $A^{(K)}$ is calculated using Equation (4). For every value of K the best response dynamics algorithm (Algorithm 1) is used to calculate the best response player.

The total utilities/fitnesses for each player can be written down:

$$f_1^{(K)} = (K-1)A_{11}^{(K)} + (n-K)A_{12}^{(K)}$$

$$f_2^{(K)} = (K)A_{21}^{(K)} + (n-K-1)A_{22}^{(K)}$$

where $f_1^{(K)}$ is the fitness of the best response player, and $f_2^{(K)}$ is the fitness of the opponent.
Using this:

$$p_{K,K-1} = \frac{(n-K)f_2^{(K)}}{Kf_1^{(K)} + (n-K)f_2^{(K)}} \frac{K}{n}$$

and:

$$p_{K,K+1} = \frac{Kf_1^{(K)}}{Kf_1^{(K)} + (n-K)f_2^{(K)}} \frac{(n-K)}{n}$$

which are all that are required to calculate x_K.

Figure 3 shows the results of an analysis of x_K for dynamically updating players. This is obtained over 182 Moran process against 122 randomly selected opponents. For each Moran process the fixation probabilities for $K \in \{1,2,3\}$ are collected. As well as recording x_K, $\tilde{x}_K$ is measured where $\tilde{x}_K$ represents the fixation probability of the best response player calculated for a given K but not allowing it to dynamically update as the population changes. The ratio $\frac{x_K}{\tilde{x}_K}$ is included in the Figure. This is done to be able to compare to a high performing strategy that has a theory of mind of the opponent but not of the population density. The ratio shows a relatively high performance compared to a non dynamic best response strategy. The mean ratio over all values of K and all experiments is 1.044. In some cases this dynamic updating results in a 25% increase in the absorption probability.

As denoted before it is clear that the best response strategy in general does not have a low SSE (only 25% of the data is below .923 and the average is .454) this is further compounded by the ratio being above one showing that in many cases the dynamic strategy benefits from its ability to adapt. This indicates that memory-one strategies that perform well in Moran processes need to more adaptable than a ZDs, and aim for mutual cooperation as well as exploitation which is in line with the results of [25] where their strategy was designed to adapt and was shown to be evolutionary stable. The findings of this work show that an optimal strategy acts in the same way.

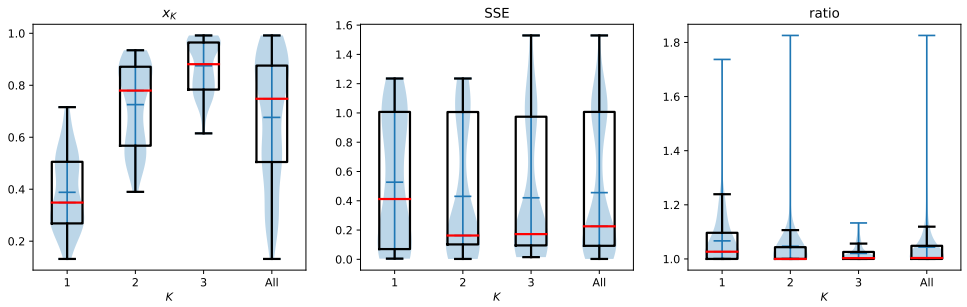

Figure 3. Results for the best response player in a dynamic Moran process. The ratio is taken as the ratio of x_k of the dynamically updating player to the fixation probability of a best response player that does not update as the population density changes.

Discussion

This manuscript has considered *best response* strategies in the IPD game, and more specifically, *memory-one best responses*. It has proven that:

- The utility of a memory-one strategy against a set of memory-one opponents can be written as a sum of ratios of quadratic forms (Theorem 1).
- There is a compact way of identifying a memory-one best response to a group of opponents through a search over a discrete set (Theorem 2).

~~Note that Theorem 2 which~~ There is one further theoretical result that can be obtained from Theorem 1, which allows the identification of environments for which cooperation cannot occur (Details are in the Supplementary Information). Moreover, Theorem 2 does not only have game theoretic novelty, but also the mathematical novelty of solving quadratic ratio optimisation problems where the quadratics are non concave.

~~Moreover Theorem 1, allows us to obtain a condition for which in an environment of~~ The empirical results of the manuscript investigated the behaviour of memory-one ~~opponents defection is the stable choice, based only on the coefficients of the opponents, as stated in Lemma **??**.~~

~~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 Equation **(??)** and Equation **(??)**.~~

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

~~while,~~

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

~~The proof of Lemma **??** is given in the Supplementary Information and a numerical simulation demonstrating the result is given in Figure **??**.~~

~~A. For $q_1 = (0.22199, 0.87073, 0.20672, 0.91861)$, $q_2 = (0.48841, 0.61174, 0.76591, 0.51842)$ and $q_3 = (0.2968, 0.18772, 0.08074, 0.73$ Equation **(??)** and Equation **(??)** hold and Defector takes over the population. B. For $q_1 = (0.96703, 0.54723, 0.97268, 0.71482)$, $q_2 = (0.69773, 0.21609, 0.97627, 0.0062)$ and $q_3 = (0.25298, 0.43479, 0.77938, 0.19769)$, Equation **(??)** fails and Defector does not take over the population. These results have been obtained by using [27] an open source research framework for the study of the IPD.~~

strategies and their limitations. The empirical results have shown that the performance ~~and the evolutionary stability~~ of memory-one strategies rely on adaptability and not on extortion, and that memory-one strategies' performance is limited by their memory in cases where they interact with multiple opponents.

~~These results were mainly to investigate the behaviour of~~ These relied on two bespoke data sets of 1000 and 152 pairs of memory-one ~~strategies and their limitations. A large data set which contained best responses in tournaments whilst including or not self interactions for $N = 2$~~ opponents equivalently, archived at [24].

A further set of results for Moran processes with a dynamically updating best response player was generated and is archived in [24]. ~~This allowed us to investigate their respective behaviours, and whether it was extortionate acts that made them the most favorable strategies. It was shown that it was not extortionbut adaptability that allowed the strategies to gain the most from their interactions. In settingswith self interactions there is some evidence that it is more likely to forgive after being tricked~~[29]. This confirmed the previous results which is that high performance requires adaptability and not extortion. It also provides a framework for future stability of optimal behaviour in evolutionary settings.

~~All the empirical results presented in this manuscript have been for the case of $N = 2$~~In the interactions we have considered here the players do not make mistakes; their actions were executed with perfect accuracy. Mistakes, however, are relevant in the reasearch of repeated games [4,30–32]. 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 best responses dynamics algorithm discussed. Another interesting avenue~~ interactions with "noise". Noise can be incoroporated into our formulation and it can be shown that the utility remains a ratio of quadratic forms

(Details see the Supplementary Information). Another avenue of investigation would be to ~~study the Moran process with a dynamically updating best response~~ understand if and/or when an evolutionary trajectory leads to a best response strategy.

By specifically exploring the entire space of memory-one strategies to identify the best strategy for a variety of situations, this work adds to the literature casting doubt on the effectiveness of ZDs, highlights the importance of adaptability and provides a framework for the continued understanding of these important questions.

References

1. Flood, M. M. Some experimental games. *Manag. Sci.* **5**, 5–26, DOI: 10.1287/mnsc.5.1.5 (1958).
2. Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396, DOI: 10.1126/science.7466396 (1981).
3. Nowak, M. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Appl. Math.* **20**, 247–265, DOI: 10.1007/BF00049570 (1990).
4. Nowak, M. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**, 56, DOI: 10.1038/364056a0 (1993).
5. Press, W. H. & Dyson, F. J. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proc. Natl. Acad. Sci.* **109**, 10409–10413, DOI: 10.1073/pnas.1206569109 (2012).
6. Stewart, A. J. & Plotkin, J. B. Extortion and cooperation in the prisoner's dilemma. *Proc. Natl. Acad. Sci.* **109**, 10134–10135, DOI: 10.1073/pnas.1208087109 (2012).
7. Adami, C. & Hintze, A. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nat. Commun.* **4**, 2193 (2013).
8. Hilbe, C., Nowak, M. & Sigmund, K. Evolution of extortion in iterated prisoner's dilemma games. *Proc. Natl. Acad. Sci.* 6913–6918, DOI: 10.1073/pnas.1214834110 (2013).
9. Hilbe, C., Nowak, M. & Traulsen, A. Adaptive dynamics of extortion and compliance. *PLOS ONE* **8**, 1–9, DOI: 10.1371/journal.pone.0077886 (2013).
10. Hilbe, C., Traulsen, A. & Sigmund, K. Partners or rivals? strategies for the iterated prisoner's dilemma. *Games Econ. Behav.* **92**, 41 – 52, DOI: 10.1016/j.geb.2015.05.005 (2015).
11. Knight, V., Harper, M., Glynatsi, N. E. & Campbell, O. 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, DOI: 10.1371/journal.pone.0204981 (2018).
12. Lee, C., Harper, M. & Fryer, D. The art of war: Beyond memory-one strategies in population games. *PLOS ONE* **10**, 1–16, DOI: 10.1371/journal.pone.0120625 (2015).
13. Han, T. A., Pereira, L. M. & Santos, F. C. Intention recognition promotes the emergence of cooperation. *Adapt. Behav.* **19**, 264–279, DOI: 10.1177/1059712311410896 (2011).
14. De Weerd, H., Verbrugge, R. & Verheij, B. How much does it help to know what she knows you know? an agent-based simulation study. *Artif. Intell.* **199**, 67–92, DOI: 10.1016/j.artint.2013.05.004 (2013).
15. Devaine, M., Hollard, G. & Daunizeau, J. Theory of mind: did evolution fool us? *PloS One* **9**, e87619, DOI: 10.1371/journal.pone.0087619 (2014).
16. Han, T. A., Pereira, L. M. & Santos, F. C. Corpus-based intention recognition in cooperation dilemmas. *Artif. Life* 365–383, DOI: 10.1162/ARTL_a_00072 (2012).
17. Knight, V. A., Harper, M., Glynatsi, N. E. & Gillard, J. Recognising and evaluating the effectiveness of extortion in the iterated prisoner's dilemma. *Preprint at* https://arxiv.org/abs/1904.00973 (2019).
18. Benureau, F. C. & Rougier, N. P. Re-run, repeat, reproduce, reuse, replicate: transforming code into scientific contributions. *Front. neuroinformatics* **11**, 69, DOI: 10.3389/fninf.2017.00069 (2018).
19. Glynatsi, N. E. & Knight, V. A. Nikoleta-v3/Memory-size-in-the-prisoners-dilemma: initial release. *Zenodo* https://doi.org/10.5281/zenodo.3533146 (2019).
20. Nowak, M. & Sigmund, K. Game-dynamical aspects of the prisoner's dilemma. *Appl. Math. Comput.* **30**, 191–213, DOI: 10.1016/0096-3003(89)90052-0 (1989).
21. Kepner, J. & Gilbert, J. *Graph algorithms in the language of linear algebra* (SIAM, 2011).

22. Jonsson, G. & Vavasis, S. Accurate solution of polynomial equations using macaulay resultant matrices. *Math. computation* **74**, 221–262 (2005).

23. Močkus, J. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk*, 400–404 (Springer Berlin Heidelberg, 1975).

24. Glynatsi, N. E. & Knight, V. A. Raw data for: "A theory of mind. Best responses to memory-one strategies. The limitations of extortion and restricted memory.". *Zenodo* https://doi.org/10.5281/zenodo.3533094 (2019).

25. Hilbe, C., Chatterjee, K. & Nowak, M. A. Partners and rivals in direct reciprocity. *Nat. Hum. Behav.* **2**, 469–477, DOI: 10.1038/s41562-018-0320-9 (2018).

26. Harper, M. *et al.* Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PLOS ONE* **12**, 1–33, DOI: 10.1371/journal.pone.0188046 (2017).

27. The Axelrod project developers. Axelrod: 4.4.0. *Zenodo* https://doi.org/10.5281/zenodo.1168078 (2016).

28. Nowak, M. A. *Evolutionary dynamics: exploring the equations of life* (Harvard University Press, 2006).

29. Glynatsi, N. E. & Knight, V. A. Raw data Moran Experiments: "A theory of mind. Best responses to memory-one strategies. The limitations of extortion and restricted memory". *Zenodo* https://doi.org/10.5281/zenodo.4036427 (2020).

30. Boyd, R. Mistakes allow evolutionary stability in the repeated prisoner's dilemma game. *J. Theor. Biol.* **136**, 47–56, DOI: 10.1016/s0022-5193(89)80188-2 (1989).

31. Imhof, L. A., Fudenberg, D. & Nowak, M. A. Tit-for-tat or win-stay, lose-shift? *J. Theor. Biol.* **247**, 574–580, DOI: 10.1016/j.jtbi.2007.03.027 (2007).

32. Wu, J. & Axelrod, R. How to cope with noise in the iterated prisoner's dilemma. *J. Confl. Resolut.* **39**, 183–189, DOI: 10.1177/0022002795039001008 (1995).

33. T., H. *et al.* scikit-optimize/scikit-optimize: v0.5.2. *Zenodo* https://doi.org/10.5281/zenodo.1207017 (2018).

34. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. In Sci. & Eng.* **9**, 90–95, DOI: 10.1109/MCSE.2007.55 (2007).

35. Meurer, A. *et al.* Sympy: symbolic computing in python. *PeerJ Comput. Sci.* **3**, DOI: 10.7717/peerj-cs.103 (2017).

36. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. & Eng.* **13**, 22–30, DOI: 10.1109/MCSE.2011.37 (2011).

Acknowledgements

A variety of software libraries have been used in this work:

- The Axelrod library for IPD simulations[27].
- The Scikit-optimize library for an implementation of Bayesian optimisation[33].
- The Matplotlib library for visualisation[34].
- The SymPy library for symbolic mathematics[35].
- The Numpy library for data manipulation[36].

Author contributions statement

N.G. and V.K. conceived the idea. N.G. conducted the experiments, N.G. and V.K. analysed the results. All authors reviewed the manuscript.

Additional information

Competing interests. The author(s) declare no competing interests.