A theory of mind: Best responses to memory-one strategies. The limitations of extortion and restricted memory.

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

[1] *Cardiff University, School of Mathematics, Cardiff, United Kingdom*
[*] *Corresponding author: Nikoleta E. Glynatsi, glynatsine@cardiff.ac.uk*

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 a theory of mind: *best response* memory-one strategies, as a multidimensional optimisation problem. We add to the literature that has shown that extortionate play is not always optimal by showing that optimal play is often not extortionate. We 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.

The Prisoner's Dilemma (PD) is a two player game used in understanding the evolution of cooperative behaviour, formally introduced in [6]. 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 $(R, P, S, T) = (3, 1, 0, 5)$ [4]. 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 [24, 25], however, they have gained most of their attention when a certain subset of memory-one strategies was introduced in [27], the zero-determinant strategies (ZDs). In [28] 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 [3, 12, 13, 14, 18, 20].

In [27] 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 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 [29, 30] and that longer-memory strategies have a certain form of advantage over short memory strategies [11, 26].

In particular, this work presents a closed form algebraic expression for the utility of a memory-one strategy against a given set of opponents, 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 [19]. This is done in tournaments with and without self interactions. 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 [5]. 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 [8].

1 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 [23] 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 [23] 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 Eq. 2. The long run steady state probability vector v, which is the solution to $vM = v$, can be combined with the payoff matrices of Eq. 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 Eq. 3.

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

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

This manuscript has explored the form of $u_q(p)$, to the 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 [17].

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:

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

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

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

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

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

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 Appendix 5.1. 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:

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

Eq. (9) is the average score (using Eq. (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.

2 Results

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:

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

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 Eq. 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:

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

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 Appendix 5.2. Finding best response memory-one strategies is analytically feasible using the formulation of Theorem 2 and resultant theory [16]. 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 [22].

2.1 Limitations of extortionate behaviour

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 [7]. It contains a total of 1000 trials corresponding to 1000 different instances of a best response strategy in tournaments with and without self interactions. 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 [19] is used. More specifically, in [19] 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):

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

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 of SSE for the best response in tournaments ($N = 2$) with and without self interactions (with $K = 1$) are given in Figure 1. Moreover, a statistical summary of the SSE distributions 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 $N = 2$

For the best response in tournaments 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

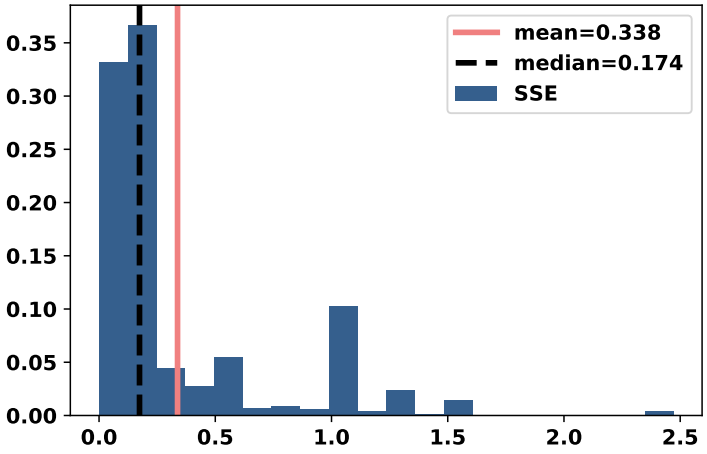

Figure 1: SEE distribution for best response in tournaments without self interactions.

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.

2.2 Limitations of memory size

The other main finding presented in [?] 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 [9], 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 Eq. 27, finding an optimal Gambler is a 17 dimensional problem given by:

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

Note that Eq. 9 can not be used here for the utility of Gambler, and actual simulated players are used. This is done using [1] with 500 turns and 200 repetitions, moreover, Eq. 18 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 [7]. 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 Fig. 2. It is evident from Fig. 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.

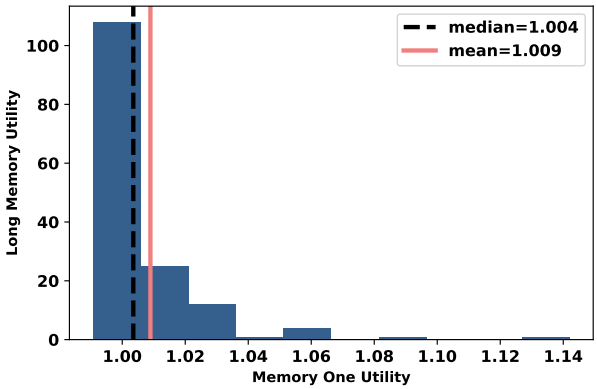

Figure 2: The ratio between the utilities of Gambler and best response memory-one strategy for 152 different pair of opponents.

2.3 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 [18]. 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 Eq. 27. 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 2.