Pure Strategy Best Responses to Mixed Strategies in Repeated Games

Shiheng Wang and **Fangzhen Lin**
Hong Kong University of Science and Technology
{swangbv, flin}@cse.ust.hk

Abstract

Repeated games are difficult to analyze, especially when agents play mixed strategies. We study one-memory strategies in iterated prisoner's dilemma, then generalize the result to k-memory strategies in repeated games. Our result shows that there always exists a pure strategy best response, which can be computed with SMT or MDP solvers. However, there may not exist such pure strategy best response in multi-agent tournaments. All source code is released for verification(see additional files).

1 Introduction

Repeated games are hard to analyze. It's well known that Nash equilibrium is the solution concept in one-shot games, whereas there exist infinite Nash equilibria in repeated games according to the folk theorem. In most repeated games, a player responses to the previous actions of the other player, thus the player's strategy need to be analysed with dynamic game theory [Han, 2018]. When the game is infinitely repeated, one cannot simply add up the payoff of each round, which in general will be infinite. Average rewards or discounted rewards are usually used to evaluate a strategy, and we focus on the former in this paper.

In repeated games, a player's strategy is a function from histories of interactions to actions. Sometimes one restricts strategies to some specific forms, such as Turing machines [Chen and Tang, 2015; Knoblauch, 1994; Megiddo and Wigderson, 1986], finite automata [Rubinstein, 1986; Ben-Porath, 1990; Gilboa, 1988; Zuo and Tang, 2015], ones with limited memories [Hauert and Schuster, 1997; Lindgren, 1992; Chen *et al.*, 2017], and other forms of bounded rationality (e.g. [Osborne and Rubinstein, 1994; Shoham and Leyton-Brown, 2008]).

A mixed strategy maps histories to a probability distribution of actions. [Press and Dyson, 2012] concluded that shortest memory sets the rule of the game, then [Chen *et al.*, 2017] pointed out the best response to a k-memory strategy in infinitely repeated games should also be k-memory. We further explore the best response to mixed strategies. Our result shows that there always exists a pure strategy best response in two-agent infinitely repeated games, but it's not the case in multi-agent tournaments.

Table 1: Prisoner's Dilemma

	c	d
c	(R,R)	(S,T)
d	(T,S)	(P,P)

In the first place we study a concrete repeated game, namely the iterated prisoner's dilemma (IPD). It involves two agents playing repeatedly the Prisoner's Dilemma (PD) in table 1. In the PD, each player can choose between Cooperate (c) and Defect (d). If both choose c, they receive a payoff of R (rewards); If both choose d, they receive a payoff of P (penalty); If one chooses c and the other d, the defector receives a payoff of T (temptation to defect) and the cooperator receives a payoff of S (sucker's payoff). The assumption is that $T > R > P > S$ and that $2 * R > T + S$, which makes (d, d) the only equilibrium in one-shot game, but cooperation provides more utility in the long run.

When both agents take one-memory mixed strategies, the IPD can be modeled as a Markov chain, whose stationary distribution can be computed according to [Press and Dyson, 2012] and the best response can be solved using an SMT solver like Z3 [De Moura and Bjørner, 2008]. We give a method to compute the best response and then summarize the best response to some popular strategies. In the meanwhile, our analysis explains the behavior of evolutionary agents. The solves a problem in [Press and Dyson, 2012] as we can now formally prove the behaviour of these agents.

Some of our results on one-memory strategies can be generalized to k-memory mixed strategies. In order to compute the best response to a completely mixed strategy, we build a Markov decision process (MDP) and compute its optimal policy. Later we prove that this is a communicating MDP, which has a pure optimal policy independent of the initial state. This shows that there always exists a pure strategy best response to any completely mixed strategy in a repeated game.

However, this result does not hold for tournaments with more than two agents. We use a similar Markov chain model to compute the best strategy in a tournament of several one-memory agents, and show that the best response can not be a pure strategy.

In this paper, when a strategy $\mathbf{q}$ is given, we calculate its best response $\mathbf{p}$. Most calculations or proofs are too laboursome to be done manually, so we show how state-of-art com-

puter solvers assist us in solving this classic economic problem and bringing us new insights. In most cases, our program solved the problems in seconds. We release all our source code for verification.

The rest of this paper is organized as follows. First we study one-memory strategies of the iterated prisoner's dilemma. Then we model k-memory problems into MDP and solve it with existing algorithms. Later we analyze a multi-agent tournament which serves as a counter example. Finally we give some discussion and concluding remarks.

2 One-Memory Strategies in IPD

2.1 One-Memory Strategies

In the section we consider a concrete example where both players play one-memory strategies in the iterated prisoner's dilemma (IPD).

The PD game in Table 1 is played by two players X and Y for infinite rounds. The score (or payoff) of either player is calculated by average score in the limit (cf. [Shoham and Leyton-Brown, 2008]). Given an infinite sequence of scores $s_i^{(1)}, s_i^{(2)}, ...$ for player $i \in \{X, Y\}$, the average score of i is

$$s_i = \lim_{k\to\infty} \frac{\sum_{j=1}^{k} s_i^{(j)}}{k} \tag{1}$$

Although a player can have unlimited memory and decide what to do based on the entire history of the interactions so far, one-memory strategies base their response only on the outcome of the previous round. Press and Dyson [Press and Dyson, 2012] proved that shortest-memory player sets the rules of the game, in the sense that, for any strategy of the longer-memory player Y, X's score is exactly the same as if Y had played a certain shorter memory strategy, disregarding any history in excess of that shared with X. This conclusion enables us to focus on one-memory strategies.

A one-memory strategy consists of an initial state p_0 (the probability to cooperate in the first round) and a vector $\mathbf{p} = (p_1, p_2, p_3, p_4) = (p_{cc}, p_{cd}, p_{dc}, p_{dd})$ where p_z is the probability of playing *Cooperate* when the outcome z occurred in the previous round.

If X uses the initial probability p_0 and strategy $\mathbf{p} = (p_1, p_2, p_3, p_4)$, Y uses the initial probability q_0 and strategy $\mathbf{q} = (q_1, q_2, q_3, q_4)$, then the probability distribution of the first iteration is $\mathbf{v}^1 = (p_0 q_0, p_0(1-q_0), (1-p_0)q_0, (1-p_0)(1-q_0))$ and the successive outcomes follow a Markov chain with transition matrix given by:

$$\mathbf{M} = \begin{pmatrix} p_1 q_1 & p_1(1-q_1) & (1-p_1)q_1 & (1-p_1)(1-q_1) \\ p_2 q_3 & p_2(1-q_3) & (1-p_2)q_3 & (1-p_2)(1-q_3) \\ p_3 q_2 & p_3(1-q_2) & (1-p_3)q_2 & (1-p_3)(1-q_2) \\ p_4 q_4 & p_4(1-q_4) & (1-p_4)q_4 & (1-p_4)(1-q_4) \end{pmatrix}$$

The probability distribution in r-th iteration $\mathbf{v}^r$ over the set of outcomes is a non-negative vector with unit sum, indexed by four states, $\mathbf{v}^r = (v^r_{cc}, v^r_{cd}, v^r_{dc}, v^r_{dd}) = (v^r_1, v^r_2, v^r_3, v^r_4)$.

Notice that we define outcome from each player's perspective, for example, if the outcome is cd from X's perspective, it is dc from Y's. $\mathbf{v}$ is defined from player X's perspective, meaning that v_{cd} refers to X plays C and Y plays D. If the previous outcome is cd, X's probability of cooperation is p_2 while Y's is p_3. See Table 2 for this correspondence.

Table 2: Outcome and Strategy

Outcome*	cc	cd	dc	dd
Strategy of $\mathbf{p}$	p_1	p_2	p_3	p_4
Strategy of $\mathbf{q}$	q_1	q_3	q_2	q_4

* Outcome is defined from $\mathbf{p}$'s perspective

Thus each entry of $\mathbf{M}$ represents the probability of transition between different states, which satisfies

$$\mathbf{M}\mathbf{v}^r = \mathbf{v}^{r+1} \tag{2}$$

In accordance with [Akin, 2016], we will call $\mathbf{M}$ convergent when $\mathbf{M}$ has a unique stationary distribution of $\mathbf{v}$, which satisfies,

$$\mathbf{M}\mathbf{v} = \mathbf{v}$$

[Press and Dyson, 2012] gave a determinant representation of player's payoff when it is calculated by the limit of average. For an arbitrary four-vector $\mathbf{f} = (f_1, f_2, f_3, f_4)$, let

$$D(\mathbf{p}, \mathbf{q}, \mathbf{f}) = \begin{vmatrix} p_1 q_1 - 1 & p_1 - 1 & q_1 - 1 & f_1 \\ p_2 q_3 & p_2 - 1 & q_3 & f_2 \\ p_3 q_2 & p_3 & q_2 - 1 & f_3 \\ p_4 q_4 & p_4 & q_4 & f_4 \end{vmatrix} \tag{3}$$

Then the average payoff of s_X and s_Y can be calculated with payoff vector $\mathbf{S_X} = (R, S, T, P)$ and $\mathbf{S_Y} = (R, T, S, P)$.

$$s_X = \frac{D(\mathbf{p}, \mathbf{q}, \mathbf{S_X})}{D(\mathbf{p}, \mathbf{q}, \mathbf{1})}, \quad s_Y = \frac{D(\mathbf{p}, \mathbf{q}, \mathbf{S_Y})}{D(\mathbf{p}, \mathbf{q}, \mathbf{1})} \tag{4}$$

2.2 Ergodic Markov Chain

According to [Akin, 2016], the following statements on stationary distribution are equivalent.

- There is a unique stationary distribution $\mathbf{v}$ in accordance with $\mathbf{M}$.
- The stationary distribution $\mathbf{v}$ is independent of the initial distribution $\mathbf{v}^1$, and thus p_0 and q_0.
- There is no absorbing states (trapped states) in the Markov chain.
- $D(\mathbf{p}, \mathbf{q}, \mathbf{1}) \neq 0$

To avoid $D(\mathbf{p}, \mathbf{q}, \mathbf{1}) = 0$, we assume $\mathbf{q}$ plays a **completely mixed** strategy, that is, $q_i \in (0, 1)$. For the convenience of defining best response, we assume $p_i \in [0, 1]$ and $\mathbf{p} \neq (1, 1, 0, 0)$ (namely strategy *Repeat*). The last assumption makes sense because if $\mathbf{p}$ plays $Cooperate$ in the first round, strategy *Repeat* is equivalent to *Always Cooperate*, and otherwise to *Always Defect*. Under these assumptions, we can prove $D(\mathbf{p}, \mathbf{q}, \mathbf{1}) \neq 0$ with the SMT solver Z3 [De Moura and Bjørner, 2008].

Theorem 1. *Assume* $p_i \in [0, 1]$ *and* $q_i \in (0, 1)$*, and that* $\mathbf{p} \neq (1, 1, 0, 0)$*, then* $D(\mathbf{p}, \mathbf{q}, \mathbf{1}) < 0$.

Proof. The negation of Theorem 1 is, $\exists \mathbf{p}, \mathbf{q}$

$$\Big(\bigwedge_{i=1,2,3,4} 0 \le p_i \le 1\Big) \quad \wedge \quad \Big(\bigwedge_{j=1,2,3,4} 0 < q_j < 1\Big)$$
$$\wedge \quad \big(\mathbf{p} \neq (1,1,0,0)\big) \quad \wedge \quad \big(D(\mathbf{p},\mathbf{q},\mathbf{1}) \ge 0\big) \tag{5}$$

Z3 returns "unsatisfiable" to (5), meaning that its value is always *False*. Thus Theorem 1 is always *True*. □

In practice we accelerate this proof with domain knowledge. We first prove that $D(\mathbf{p},\mathbf{q},\mathbf{1})$ is monotonic to p_i by calculating derivatives(see also proof of Thm 2), then calculate extrema by letting $p_i \in \{0,1\}$. Since all extrema are less than zero, the determinant is less than zero.

2.3 Symbolic Calculation

Theorem 2 (Monotonicity)**.** *Suppose $p_i \in [0,1]$ and $q_i \in (0,1)$, $\mathbf{p} \neq (1,1,0,0)$, s_X, s_Y are defined in equation (4), and $z \in \{p_1,...,p_4,q_1,...,q_4\}$.*
When all variables in $\{p_1,...,p_4,q_1,...,q_4\}\backslash\{z\}$ are fixed, s_X is monotonic to z.

Proof. According to Theorem 1, $D(\mathbf{p},\mathbf{q},\mathbf{1}) \neq 0$.

$$s_X = \frac{D(\mathbf{p},\mathbf{q},\mathbf{S_X})}{D(\mathbf{p},\mathbf{q},\mathbf{1})}$$

For simplicity, let

$$M = D(\mathbf{p},\mathbf{q},\mathbf{S_X}), N = D(\mathbf{p},\mathbf{q},\mathbf{1})$$

Then

$$s_X = M/N$$

Without loss of generality, get partial derivative to p_1,

$$\frac{\partial s_X}{\partial p_1} = \frac{\partial M/\partial p_1 \cdot N - \partial N/\partial p_1 \cdot M}{N^2} \tag{6}$$

Let

$$U = \partial M/\partial p_1 \cdot N - \partial N/\partial p_1 \cdot M$$

Again, get partial derivative of U to p_1,

$$\partial U/\partial p_1 = 0$$

In Eq. (6), since $N^2 > 0$, and the numerator is not a function of p_1, we can draw the conclusion that, when $\{p_2,p_3,p_4,q_1,q_2,q_3,q_4\}$ are fixed, s_X is monotonic in terms of p_1.

Similarly, s_X is also monotonic to any variable in p_i, q_i when other variables are fixed. □

As symbolic calculations of determinant are laboursome, we use a *Python* library *SymPy* [Meurer *et al.*, 2017] and double checked with *MATLAB*.

Given a fixed completely mixed strategy $\mathbf{q}$, we want to compute its best response $\mathbf{p}$. Theorem 2 implies that there exists a pure one-memory strategy best response. We only need to enumerate all pure strategy $\mathbf{p}$, and one of them must be a best response to $\mathbf{q}$.

Theorem 3. *There always exists an one-memory pure strategy best response to completely mixed one-memory strategies.*

[Press and Dyson, 2012] did some experiments to show that evolutionary players who evolve to get better payoff will finally reach the optimal reward but they didn't prove it analytically. Actually this property follows from Theorem 2. No matter where a strategy begins, it will finally evolve to the best response. In other words, there is no local maxima.

We are now able to compute all extreme values of s_X by letting $p_i = 0$ or $p_i = 1$. The result is denoted as F_k, where $k = \sum_{i=1}^{4} p_i * 2^{4-i}$. For instance, when $\mathbf{p} = (1,1,1,1)$,

$$F_{15} = \frac{-R*q_3 + S*q_1 - S}{q_1 - q_3 - 1}$$

Some other F_k has a much more complex expression. We'll not show them due to limited space but readers can refer to our source code for details. Notice that F_{12} is not considered as it corresponds to *Repeat* strategy, $\mathbf{p} = (1,1,0,0)$. Some formulas turn out to have the same value, that is,

$$F_0 = F_4 = F_8, \quad F_{13} = F_{14} = F_{15}$$

The corresponding strategies always receive the same payoff whatever the co-player's strategy is, we call such strategies equivalent.

Theorem 4 (Equivalent)**.** *When playing agaisnt completely mixed strategies, the following strategies are equivalent.*

$$(0,0,0,0) \equiv (0,1,0,0) \equiv (1,0,0,0)$$
$$(1,1,0,1) \equiv (1,1,1,0) \equiv (1,1,1,1)$$

Duplicates will not be considered in the rest of this paper, and the set of distinct expressions is defined as,

$$\mathcal{F} = \{F_0,F_1,F_2,F_3,F_5,F_6,F_7,F_9,F_{10},F_{11},F_{15}\} \tag{7}$$

2.4 Theorem Discovery

With concrete values of (R,S,T,P), when strategy $\mathbf{q}$ is given, we can simply compute all values in $\mathcal{F}$ so that the strategy corresponding to the largest value is a best response. Our experiments show that concrete values of (R,S,T,P) really matter. When the concrete setting of the IPD varies, the best responses to some strategies also change. But some general theorems are discovered with SMT solver Z3.

The knowledge base is defined according to the constraints of the IPD and mixed one-memory strategies. In this section, $\supset$ means "logical imply" , $\equiv$ means "logical equivalent" and $\neg$ means "logical not". Theorems are proved by refutation.

Definition 1 (Knowledge Base)**.**

$$\mathcal{KB} \equiv \Big(\bigwedge_{i=1,2,3,4} 0 \le p_i \le 1\Big) \wedge \Big(\bigwedge_{i=1,2,3,4} 0 < q_i < 1\Big)$$
$$\wedge (T > R > P > S) \wedge (2R > T + S) \tag{8}$$

If a strategy is less cooperative under a better situation, there is no reason to cooperate with him. As $T > R > P > S$ is the setting of the prisoner's dilemma, column player's preference is $CD \succ CC \succ DD \succ DC$. If his strategy $\mathbf{q}$ satisfies $q_3 \le q_1 \le q_4 \le q_2$, then there is no chance of reciprocity. See Table 2 for this correspondence.

Theorem 5. *If strategy $\mathbf{q} = \{q_1,q_2,q_3,q_4\}$ satisfies $q_3 \le q_1 \le q_4 \le q_2$, then* Always Defect *is its best response .*

Proof. First, we give a formal representation of this theorem. $\forall \mathbf{p}, \mathbf{q}, R, S, T, P$,

$$\mathcal{KB} \wedge (q_3 \leq q_1 \leq q_4 \leq q_2) \supset \bigwedge_{F' \in \mathcal{F} \setminus \{F_0\}} F_0 > F' \quad (9)$$

Eq. (9) is equivalent to, $\forall \mathbf{p}, \mathbf{q}, R, S, T, P$,

$$\neg\big(\mathcal{KB} \wedge (q_3 \leq q_1 \leq q_4 \leq q_2)\big) \vee \bigwedge_{F' \in \mathcal{F} \setminus \{F_0\}} F_0 > F' \quad (10)$$

This theorem can be proved by conclusion refutation, and the negation of Eq. (10) is, $\exists \mathbf{p}, \mathbf{q}, R, S, T, P$,

$$\big(\mathcal{KB} \wedge (q_3 \leq q_1 \leq q_4 \leq q_2)\big) \wedge \bigvee_{F' \in \mathcal{F} \setminus \{F_0\}} F_0 \leq F' \quad (11)$$

which is equivalent to,

$$\bigvee_{F' \in \mathcal{F} \setminus \{F_0\}} \big(\mathcal{KB} \wedge (q_3 \leq q_1 \leq q_4 \leq q_2) \wedge F_0 \leq F'\big) \quad (12)$$

Z3 returns "unsatisfiable" to equation (12), implying that its negation(i.e. Theorem 5) is always true. □

A natural corollary of Theorem 5 is, if $q_1 = q_2 = q_3 = q_4$, then *Always Defect* is a best response. One famous example of such strategy is *Random*, i.e. $\mathbf{q} = (0.5, 0.5, 0.5, 0.5)$. This corollary illustrates that if some player is indifferent to the outcome of previous round, there is no reason to cooperate with him. More discussion on this property and backward induction can be found in Section 4.3.

Some other interesting theorems are discovered on the neighbor of *Mischief* strategy defined by [Press and Dyson, 2012]. Such strategy is also called equalizer. We only show the theorems together with their formal representations as the proofs are very similar.

Definition 2 (MisChief). *A MisChief strategy is an one-memory strategy* $\mathbf{q} = (q_1, \overline{q_2}, \overline{q_3}, q_4)$ *defined as,*

$$\mathcal{MC} \equiv \quad \overline{q_2} = \frac{q_1(T-P) - (1+q_4)(T-R)}{R-P} \quad \wedge \quad \overline{q_3} = \frac{(1-q_1)(P-S) + q_4(R-S)}{R-P} \quad (13)$$

Theorem 6 (Mischief). *While playing with MisChief strategy* $\mathbf{q} = (q_1, q_2, q_3, q_4)$*, every strategy of* $\mathbf{p}$ *receives the same average payoff. Formally,* $\forall \mathbf{p}, \mathbf{q}, R, S, T, P$

$$\mathcal{KB} \wedge \mathcal{MC} \wedge (q_2 = \overline{q_2}) \wedge (q_3 = \overline{q_3}) \supset \bigwedge_{F_i, F_j \in \mathcal{F}} F_i = F_j$$

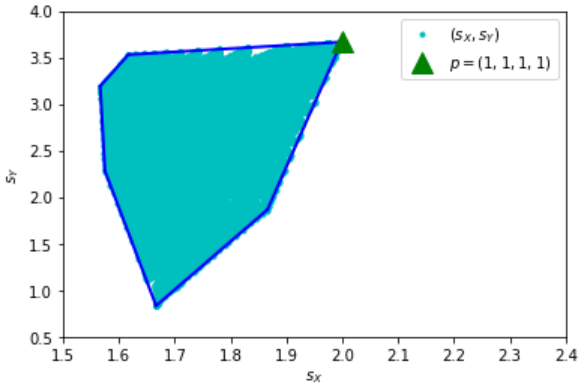
Theorem 7 (MisTort). *While playing with strategy* $\mathbf{q} = (q_1, q_2, q_3, q_4)$ *s.t.* $q_2 < \overline{q_2}$, $q_3 = \overline{q_3}$*, Always Cooperate is a best response. Formally,* $\forall \mathbf{p}, \mathbf{q}, R, S, T, P$,

$$\mathcal{KB} \wedge \mathcal{MC} \wedge (q_2 < \overline{q_2}) \wedge (q_3 = \overline{q_3}) \supset \bigwedge_{F' \in \mathcal{F} \setminus \{F_{15}\}} F_{15} > F'$$

Theorem 8. *While playing with strategy* $\mathbf{q} = (q_1, q_2, q_3, q_4)$ *s.t.* $q_2 = \overline{q_2}$,$q_3 > \overline{q_3}$*, Always Defect is a best response. Formally,* $\forall \mathbf{p}, \mathbf{q}, R, S, T, P$,

$$\mathcal{KB} \wedge \mathcal{MC} \wedge (q_2 = \overline{q_2}) \wedge (q_3 > \overline{q_3}) \supset \bigwedge_{F' \in \mathcal{F} \setminus \{F_0\}} F_0 > F'$$

Figure 1: MisTort Strategy $\mathbf{q} = (0.9, 0.5, 0.2, 0.1)$



We call the strategy in Theorem 7 *MisTort* because it's on the neighbor of *MisChief* strategy and shares the property of extortionate strategies in [Press and Dyson, 2012], to which the best response is *Always Cooperate*. Due to similar property, readers may guess that one may be the a subclass of the other. However, there turns out to be no intersection between *MisTort* and extortionate strategies.

Theorem 9. *There is no intersection between extortionate strategy and MisTort strategy.*

Proof. Extortionate strategy($\mathcal{EX}$) is defined in [Press and Dyson, 2012],

$$\begin{aligned}\mathcal{EX} \equiv \quad & q_1 = 1 - \phi(\chi - 1)\frac{R-P}{P-S} \\ \wedge \quad & q_2 = 1 - \phi(1 + \chi\frac{T-P}{P-S}) \\ \wedge \quad & q_3 = \phi(\chi + \frac{T-P}{P-S}) \quad \wedge \quad q_4 = 0 \\ \wedge \quad & 0 < \phi \leq \frac{(P-S)}{(P-S) + \chi(T-P)} \quad \wedge \quad \chi > 1\end{aligned}$$

The negation of Theorem 9 is, $\exists \mathbf{q}$,

$$\Big(\bigwedge_{i=1,2,3,4} 0 \leq q_i \leq 1\Big) \wedge \mathcal{EX} \wedge \mathcal{MC} \wedge (q_2 < \overline{q_2}) \wedge (q_3 = \overline{q_3}) \wedge (T > R > P > S) \wedge (2R > T + S) \quad (14)$$

Z3 returns "unsat", meaning Theorem 9 is always true. □

One specific instance of *MisTort* strategy can be calculated by letting $q_1 = 0.9, q_4 = 0.1$ and $(R, S, T, P) = (3, 0, 5, 1)$. From Eq. (14) we can calculate out $q_2 < 0.7, q_3 = 0.2$. A possible *MisTort* strategy is, $\mathbf{q} = (0.9, 0.5, 0.2, 0.1)$. We enumerate a number of strategy $\mathbf{p}$ to play against this $\mathbf{q} = (0.9, 0.5, 0.2, 0.1)$, and the average payoff of both players is shown in Figure 1. All pairs of payoffs (s_X, s_Y) form a compact convex hull, and the case that $\mathbf{p} = (1, 1, 1, 1)$ is receives highest payoff. The figure implies that *Always Cooperate* is a best response to $\mathbf{q}$, leading to a payoff where $s_X = 2.0, s_Y = 3.67$.

3 Repeated Games and MDP

3.1 Repeated Game and K-memory Strategy

A finite, 2-person normal form game is a tuple (N, A, u), where

- $N = \{1, 2\}$ is the set of two players.
- $A = A_1 \times A_2$, where A_i is a finite set of actions available to player i. Each vector $a = (a_1, a_2) \in A$ is called an action profile (or outcome).
- $u = (u_1, u_2)$, where $u_i : A \mapsto \mathbb{R}$ is a real-valued utility (or payoff) function for player i.

This stage game is played for infinite rounds. A player's payoff is defined as the average payoff of the stage game in the limit [Shoham and Leyton-Brown, 2008]. Similar to Eq. (1), given an infinite sequence of payoffs $u_i^{(1)}$, $u_i^{(2)}$,... for player i, the average payoff of i is $\lim_{k\to\infty} \sum_{j=1}^{k} u_i^{(j)}/k$.

A k-memory mixed strategy of player i is a function that maps k action profiles to a distribution on actions,

$$p_i : A^k \mapsto \Delta(A_i) \tag{15}$$

Given previous k actions profiles $s \in A^k$, the probability of taking action $a_i \in A_i$ is written as $p_i(s, a_i)$. In the start of the game, there aren't k action profiles, in this case we fill empty action profile with action c. Our conclusions don't rely on the outcomes of the first k rounds, see Thm 11.

3.2 Markov Decision Process

Now we compute the best response to a k-memory strategy. The first question is how much memory is required to be a best response. According to [Press and Dyson, 2012], shortest memory sets the rule of the game, so that it needs to be at most k-memory. On the other hand, any strategy with less than k-memory can be represented by a k-memory strategy. Then we know k-memory strategy should be exactly enough. This problem is also discussed in [Chen *et al.*, 2017].

Both players take k-memory strategies. Assume player 2 plays a completely mixed strategy p_2, namely $\forall s \in S, a_2 \in A_2, p_2(s, a_2) > 0$, then from player 1's point of view, his action and payoff can be modeled as the following MDP.

- Actions: same as A_1 in the repeated game.
- States: $S = A^k$, the set of vectors of k action profiles, indexed by 1,2,...,k, from farthest to most recent. Any $s_{1:k} \in S$ refers to k continuous action profiles. When there comes a new action profile (a_1, a_2), it goes to a new state $s' = s_{2:k} + (a_1, a_2)$, which drops the farthest action profile s_1 and appends (a_1, a_2).
- Transition Function. When player 1 takes action a_1, the states transits from state s to $s' = s_{2:k} + (a_1, a_2)$ with probability $T(s, a_1, s') = p_2(s, a_2)$. At every state s, for any action a_1, $\sum_{s'} T(s, a_1, s') = 1$.
- Reward $R(s, a_1) = u_1(s_k), \forall a_1 \in A_1$, meaning player 1's utility of the most recent action profile.

A mixed policy π assigns a probability distribution of actions to each state. Formally, $\pi : S \mapsto \Delta(A_1)$. For each $s \in S$, the probability of taking action $a_1 \in A_1$ is written as $\pi(s, a_1)$. A completely mixed policy is $\pi(s, a_1) > 0, \forall s \in S, a_1 \in A_1$. Π^+ is the set of all completely mixed policies, which is a subset of all policies Π. A pure policy is that, for each $s \in S$, there is exactly one $a_1 \in A_1$ satisfying $\pi(s, a_1) = 1$. From the definition of state S and strategy in Eq. (15), we can easily see that every policy π corresponds to a strategy p_1.

Player 1 walks among states for infinite rounds, starting from any state $s \in S$. At each round, it takes an action according to current state and its policy π, then it goes into a new state. This agent receives a sequence of rewards, namely $R^1, R^2, ..., R^j, ...$, and the final score $\rho^\pi(s)$ is calculated by the expectation of average, formally,

$$\rho^\pi(s) = \lim_{k\to\infty} E^\pi\{\frac{\sum_{j=0}^{k} R^j}{k}\}$$

Our goal is to find an optimal policy π^* that maximizes the expectation of average payoff, $\forall \pi, \rho^{\pi^*}(s) \geq \rho^\pi(s)$. We will show that the initial state s doesn't matter, for our MDP.

3.3 Solving Average-Payoff MDPs

There is a detailed explanation on average-payoff MDPs in textbook [Puterman, 2014]. Our results in this part are mainly based on propositions and theorems in [Filar and Schultz, 1988], which are introduced as lemmas. To be specific, Lemma 1 is from paragraph 2, section 1; Definition 3 is from Definition 1.1, and Lemma 2 is from Theorem 2.1 in [Filar and Schultz, 1988].

Lemma 1. *While computing optimal policy, it is sufficient to consider pure policies.*

A policy π induces a Markov chain on states S with transition matrix $\mathbf{M}(\pi)$, whose entries $\mathbf{M}_{st}(\pi)$ denote the probability of transition from state s to state t when policy π is followed. For τ being a nonnegative integer, $\mathbf{M}^\tau(\pi)$ denotes the τ-th power of square matrix $\mathbf{M}(\pi)$, and $\mathbf{M}^\tau_{st}(\pi)$ denotes an entry in it.

Definition 3. *An MDP is communicating if, for every pair of states $s, t \in S$, there exists a pure policy π and an integer $\tau \geq 1$ such that $\mathbf{M}^\tau_{st}(\pi)$ is strictly positive.*

Lemma 2. *Let Π^+ be the set of completely mixed policies. The following two conditions are equivalent.*

- *An MDP is communicating.*
- *Every policy $\pi^+ \in \Pi^+$ induces an irreducible $\mathbf{M}(\pi^+)$.*

Since we have assumed that player 2 takes a completely mixed strategy p_2, for every $\pi^+ \in \Pi^+$, the induced matrix $\mathbf{M}(\pi^+)$ is irreducible. In other words, when both players take completely mixed strategies, there is no transient state in the corresponding Markov chain, because it's possible to append any action profile to any state.

Theorem 10. *The MDP in section 3.2 is communicating. Its optimal policy can be calculated by linear programming, which is independent of starting state. If π^* is an optimal policy starting from state s, then it is also an optimal policy starting from any other state s'.*

Then there is a theorem in repeated games.

Theorem 11. *There always exists a k-memory pure strategy best response to k-memory completely mixed strategy in infinitely repeated games, which is independent of the initial k outcomes.*

Such pure strategy best response can be computed with existing MDP solvers, e.g. MDPtoolbox [Chadès *et al.*, 2014]. As an example, we compute the best response to Stochastic Tif-for-2-Tats(*STF2T*). The game is IPD in Table 1 with (R,S,T,P)=(3,0,5,1). *STF2T* cooperates with probability 0.1 when the other player defects for continuous two rounds, and cooperates with probability 0.9 otherwise. MDPtoolbox solves this model, the best response is to play c and d alternatively, and the average payoff is 2.67. See our source code for details.

4 Discussion

4.1 Multi-agent Tournament

In the previous sections we have discussed how to calculate the best response in two-agent repeated games. Our result mainly relies on the fact that there always exists a pure strategy best response. Now we consider the best response in multi-agent tournaments, which may not be a pure strategy.

Due to the complexity of symbolic calculations, we conduct experiments of a multi-agent IPD tournament instead of giving an analytical proof. Consider a tournament of eleven one-memory agents, namely one $\mathbf{p}$, nine $\mathbf{q}$ and one $\mathbf{u}$. Suppose $(R, S, T, P) = (3, 0, 5, 1)$, $\mathbf{q} = (0.9, 0.5, 0.2, 0.1)$ and $\mathbf{u} = (0.4, 0.8, 0.2, 0.6)$, we want to compute the optimal $\mathbf{p}$ that maximizes his payoff. Strategy $\mathbf{p}$ plays with every $\mathbf{q}$ and $\mathbf{u}$ respectively and the final score takes the average of all games, which can be calculated according to Eq. (4),

$$s_{\mathbf{p}} = 0.9 * \frac{D(\mathbf{p}, \mathbf{q}, \mathbf{S_X})}{D(\mathbf{p}, \mathbf{q}, \mathbf{1})} + 0.1 * \frac{D(\mathbf{p}, \mathbf{u}, \mathbf{S_X})}{D(\mathbf{p}, \mathbf{u}, \mathbf{1})} \tag{16}$$

In two-agent games, there always exists a pure strategy best response. As is shown in Theorem 7 and Theorem 5, the best response to $\mathbf{q}$ is *Always Cooperate*, while the best response to $\mathbf{u}$ is *Always Defect*. Now we consider whether there is a pure strategy one-memory best response $\mathbf{p}^*$ in this tournament .

First we let $\mathbf{p}$ be a pure strategy and compute the values of $s_{\mathbf{p}}$. There are $2^4 - 1 = 15$ pure strategies, where *Repeat* strategy $\mathbf{p} = (1, 1, 0, 0)$ is excluded. Among them, strategy *Tit-for-Tat* $\mathbf{p} = (1, 0, 1, 0)$ receives highest payoff of 1.90. However, after we consider mixed strategies, we found that strategy $\mathbf{p} = (1, 0.9, 0, 0.1)$ reaches a higher payoff of 2.02.

This specific example shows that there may not exist a pure strategy best response in general multi-agent tournaments.

4.2 Limitation of *Z3*

Although *Z3* is effective in solving linear constraints and successfully prove our theorems, it sometimes fails to give a solution in several hours when the constraint is a non-linear combination of several variables. We take comprehensive measures to overcome this challenge. (1) **Avoid quantifiers**. A mixture of $\forall$ and $\exists$ prevents us from efficient proof. We avoid using both quantifiers and manually simplify some formulas. (2) **Simplify formulas**. We take advantage of domain knowledge to remove fractions, such as in the proof of Theorem 1 we first prove monotonicity and prove all extreme values are less than zero. (3) **Break down.** We solve one clause of conjunction or disjunction at one time. Usually a theorems requires every clause has the same value of *true* or *false*, such as Eq. (12). (4) **Remove variables**. When SMT cannot solve a theorem, we have to replace some variables with concrete values to get some conjectures or a weaker theorem. In section 4.1 we compute a counter-example instead of prove it analytically.

4.3 Backward Induction

There has been many debates on backward induction [Kreps *et al.*, 1982] [Binmore, 1997]. For finite n round IPD, playing defect is the best strategy in the one-shot game at n-th round because there is no further possibility of reciprocity. Since both rational players will play defect in the n-th round, there is no reason to cooperate in the (n-1)-th round. By backward induction, *Always Defect* is the only equilibrium of finite iterated prisoner's dilemma.

To explain the emergence of cooperation, it is usually assumed that the game is played for infinite rounds, or an unknown number of rounds[Axelrod and Hamilton, 1981]. Although this assumption invalidates backward induction, there still lacks an explanation of cooperation. Actually, when people conduct backward induction, they implicitly assume that the outcome of previous round has no effect on the decision of current round. According to the corollary of Theorem 5, when someone is indifferent to the outcome of previous round, there is no reason to cooperate with him. Therefore, we hold the view that the emergence of cooperation cannot be explained without taking previous history into consideration.

5 Conclusion

In this paper we compute best responses to mixed strategies in repeated games. Our main result shows that there always exists a pure strategy best response in two-agent repeated games. Based on this result, we analyze one-memory strategies, give the method to compute best response, and discover new theorems in the iterated prisoner's dilemma. The work enhances our comprehension of the IPD and explains the evolutionary behavior left over in [Press and Dyson, 2012].

We generalize this result to the best response to k-memory strategies. Such problem is modeled as MDP and solved with existing solvers. In a multi-agent tournament, however, there may not exist a pure strategy best response. As a result, computing the best strategy in a tournament where an agent should take the same strategy to all other agents is still an open question.

As most calculations and proofs are conducted by computer programs, we release all our source code for verification. Source code is uploaded as additional files.

References

[Akin, 2016] Ethan Akin. The iterated prisoner's dilemma: good strategies and their dynamics. *Ergodic Theory, Advances in Dynamical Systems*, pages 77–107, 2016.

[Axelrod and Hamilton, 1981] Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *science*, 211(4489):1390–1396, 1981.

[Ben-Porath, 1990] Elchanan Ben-Porath. The complexity of computing a best response automaton in repeated games with mixed strategies. *Games and Economic Behavior*, 2(1):1–12, 1990.

[Binmore, 1997] Ken Binmore. Rationality and backward induction. *Journal of Economic Methodology*, 4(1):23–41, 1997.

[Chadès *et al.*, 2014] Iadine Chadès, Guillaume Chapron, Marie-Josée Cros, Frédérick Garcia, and Régis Sabbadin. Mdptoolbox: a multi-platform toolbox to solve stochastic dynamic programming problems. *Ecography*, 37(9):916–920, 2014.

[Chen and Tang, 2015] Lijie Chen and Pingzhong Tang. Bounded rationality of restricted turing machines. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1673–1674. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[Chen *et al.*, 2017] Lijie Chen, Fangzhen Lin, Pingzhong Tang, Kangning Wang, Ruosong Wang, and Shiheng Wang. K-memory strategies in repeated games. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1493–1498. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[De Moura and Bjørner, 2008] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.

[Filar and Schultz, 1988] Jerzy A Filar and Todd A Schultz. Communicating mdps: equivalence and lp properties. *Operations Research Letters*, 7(6):303–307, 1988.

[Gilboa, 1988] Itzhak Gilboa. The complexity of computing best-response automata in repeated games. *Journal of economic theory*, 45(2):342–352, 1988.

[Han, 2018] *Handbook of dynamic game theory*. Springer, Cham, Switzerland, 2018.

[Hauert and Schuster, 1997] CH Hauert and Heinz Georg Schuster. Effects of increasing the number of players and memory size in the iterated prisoner's dilemma: a numerical approach. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1381):513–519, 1997.

[Knoblauch, 1994] Vicki Knoblauch. Computable strategies for repeated prisoner's dilemma. *Games and Economic Behavior*, 7(3):381–389, 1994.

[Kreps *et al.*, 1982] David M Kreps, Paul Milgrom, John Roberts, and Robert Wilson. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic theory*, 27(2):245–252, 1982.

[Lindgren, 1992] Kristian Lindgren. Evolutionary phenomena in simple dynamics. In *Artificial life II*, pages 295–312, 1992.

[Megiddo and Wigderson, 1986] Nimrod Megiddo and Avi Wigderson. On play by means of computing machines: preliminary version. In *Proceedings of the 1986 Conference on Theoretical aspects of reasoning about knowledge*, pages 259–274. Morgan Kaufmann Publishers Inc., 1986.

[Meurer *et al.*, 2017] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.

[Osborne and Rubinstein, 1994] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

[Press and Dyson, 2012] William H Press and Freeman J Dyson. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences*, 109(26):10409–10413, 2012.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Rubinstein, 1986] Ariel Rubinstein. Finite automata play the repeated prisoner's dilemma. *Journal of economic theory*, 39(1):83–96, 1986.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[Zuo and Tang, 2015] Song Zuo and Pingzhong Tang. Optimal machine strategies to commit to in two-person repeated games. In *AAAI*, pages 1071–1078, 2015.