Payoff Control in the Iterated Prisoner's Dilemma

Dong Hao[1], **Kai Li**[2] and **Tao Zhou**[1]
[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Shanghai Jiao Tong University, Shanghai, China
haodong@uestc.edu.cn, kai.li@sjtu.edu.cn, zhutou@ustc.edu

Abstract

Repeated game has long been the touchstone model for agents' long-run relationships. Previous results suggest that it is particularly difficult for a repeated game player to exert an autocratic control on the payoffs since they are jointly determined by all participants. This work discovers that the scale of a player's capability to unilaterally influence the payoffs may have been much underestimated. Under the conventional iterated prisoner's dilemma, we develop a general framework for controlling the feasible region where the players' payoff pairs lie. A control strategy player is able to confine the payoff pairs in her objective region, as long as this region has feasible linear boundaries. With this framework, many well-known existing strategies can be categorized and various new strategies with nice properties can be further identified. We show that the control strategies perform well either in a tournament or against a human-like opponent.

1 Introduction

Understanding what are the best strategies for intelligent agents in a long-run relationship is a fundamental challenge in many disciplines. Repeated games are prevailing tools for modeling and analyzing intelligent agents' long-run relationships [Mailath and Samuelson, 2006], which have been richly studied in economics, artificial intelligence and biology [Kandori, 2002; Claus and Boutilier, 1998; Nowak *et al.*, 1995]. For multi-agent systems, repeated games are widely utilized for understanding how cooperation or competition emerges among agents and how cooperative or winning strategies can be identified. It has been commonly accepted that in such games, it is impossible for a unilateral player to freely control the payoffs and determine the evolutionary route of the game, since the outcomes are jointly determined by all participants.

In this paper, we propose a general framework for payoff control in iterated prisoner's dilemma, which is a conventional model for repeated games. First of all, based on the game's Markov decision process (MDP), the correlation between a single player's strategy and the MDP's joint stationary distribution is derived. Then according to this correlation, we establish a general payoff control framework, under which a control strategy can be easily obtained by solving a system of linear inequalities. Using the payoff control framework, as long as the control objective is feasible, a controller can restrict the relation between her and the opponent's payoffs (represented by a two-tuple) to an arbitrary region with linear boundaries. More specifically, she can (i) unilaterally determine the maximum and minimum values of the opponent's possible payoffs; or (ii) always win the game no matter what the opponent's strategy is, and she can even control her winning probability; or (iii) control the evolutionary route of the game, as long as the opponent is rational and self-optimizing, the controller can enforce the game to finally converge either to a mutual-cooperation equilibrium or to any feasible equilibrium that she wishes. We simulate serval specific strategies generated under the payoff control framework in a tournament similar to that of Axelrod [Axelrod and Hamilton, 1981], it is found that the new payoff control strategies have remarkably good performances.

The discussion of payoff control in games can be traced back to [Boerlijst *et al.*, 1997], in which the authors discovered that, in iterated prisoner's dilemma, one player can set the opponent's payoff to a certain value. However, what is the underlying mechanism for such strategies to exist and how to formally derive them are not thoroughly investigated. In recent years, Press and Dyson's discovery of "zero-determinant (ZD)" strategies illuminates a new starting point for the control [Hao *et al.*, 2014]. They show that in repeated games, it is possible for a player to unilaterally enforce a *linear* relation between her and the opponent's payoff [Press and Dyson, 2012]. This is the first time that the linear payoff relation control is formally investigated, which receives a lot of attention. Thereafter, the linear control on players' payoff relations is discovered in multiplayer games [Pan *et al.*, 2015; Hilbe *et al.*, 2014], games with imperfect information [Chen and Zinger, 2014; Hao *et al.*, 2015] and evolutionary games [Adami and Hintze, 2013; Hilbe *et al.*, 2013]. Furthermore, from a mathematical point of view, Akin formally investigated why such linear control exists in games which can be represented by an MDP and proposed a new payoff control scheme whereby one player can fix the upper bound of the opponent's payoff to the mutual cooperation reward R, and such strategies can enforce the game to converge to a mutual cooperation situation [Akin, 2012]. Extended cases of Akin's cooperation-enforcing control are then studied and special

cases of the nonlinear payoff relation control are identified [Hilbe *et al.*, 2015].

These existing payoff control strategies confront two major problems. The first one is that they only realize special cases of payoff control such as linear control or cooperative control. Zero-determinant based strategies can only realize linear payoff relations, which are very strong and sometimes not easy to use; Akin's method based strategies can only control the upper bound of the opponent's payoff to a mutual-cooperation reward R. However, how to establish a general control framework with multiple and free control objectives is still challenging. The second problem is that these strategies are mostly difficult to obtain. For the zero-determinant based strategies, calculating the determinant of a matrix already has high computational complexity; for strategies based on Akin's method, when one tries to add more objectives other than cooperation-enforcement, the computational complexity increases exponentially and deriving the strategy becomes intractable. In this paper, we propose a general payoff control framework which conquers both of these two problems. In section 2, the repeated game is modeled as an MDP and the relationship between a single player's strategy and the stationary distribution of the MDP is derived. In section 3, we realize a control on the opponent's maximum and minimum payoffs. In section 4, this is extended to a free regional payoff control with multiple linear boundaries, and various types of regional control strategies, especially the cooperation-enforcing control strategies, are identified. To analyze the performances of the payoff control strategies when confronting various famous strategies, in section 5, we simulate control strategies in the Axelrod's tournament. In the last section, to evaluate how payoff control strategies perform in the real world, we simulate them against a reinforcement learning player [Sutton and Barto, 1998].

2 Strategy and Game Convergence

The iterated prisoner's dilemma (IPD) is the canonical example for analyzing the cooperation and competition in agents' long-run relationships. The IPD consists of multiple rounds of stage games. In each stage, player $i \in \{X, Y\}$ adopts an action $a_i \in \{C, D\}$ with a certain probability, where C denotes cooperation and D denotes defection. The space of the outcomes in each stage game is $\Omega = \{CC, CD, DC, DD\}$. If both players cooperate (CC), then each earns a reward R; if one cooperates but the other defects (CD or DC), then the cooperator earns S and the defector earns T; if they both defect (DD), then both get P. The payoff vector of player X over Ω is thus defined as $\mathbf{S}_X = (R, S, T, P)$ and for player Y it is $\mathbf{S}_Y = (R, T, S, P)$. In this paper we consider the case that player X chooses her action conditioning only on the outcome of the previous stage. It is worth noting that, in infinitely repeated games it has been proved that such one-stage memory strategies have no disadvantages as if the opponent has a longer memory [Press and Dyson, 2012]. The strategy of player X is defined as a vector of probabilities $\mathbf{p} = (p_1, p_2, p_3, p_4)$, where each component is a probability that she cooperates with player Y conditioning on the last stage outcomes CC, CD, DC or DD, respectively. Analogously, the strategy of player Y is a vector of probabilities for cooperation $\mathbf{q} = (q_1, q_2, q_3, q_4)$ conditioning on the previous outcomes CC, DC, CD or DD, respectively.

Then the transition matrix over the state space between adjacent stage games is derived as $\mathbf{M}$:

$$\begin{bmatrix} p_1 q_1 & p_1(1-q_1) & (1-p_1) q_1 & (1-p_1)(1-q_1) \\ p_2 q_3 & p_2(1-q_3) & (1-p_2) q_3 & (1-p_2)(1-q_3) \\ p_3 q_2 & p_3(1-q_2) & (1-p_3) q_2 & (1-p_3)(1-q_2) \\ p_4 q_4 & p_4(1-q_4) & (1-p_4) q_4 & (1-p_4)(1-q_4) \end{bmatrix} \quad (1)$$

If this Markov matrix is regular, it has the unique stationary distribution $\mathbf{v} = (v_1, v_2, v_3, v_4)$, which is a probability distribution over the state space Ω and can be calculated as

$$\mathbf{v} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{v}^t, \quad (2)$$

where $\mathbf{v}^t$ is the distribution over Ω at the t-th stage. Then the average expected payoffs for players X and Y are derived as $s_X = \mathbf{v} \cdot \mathbf{S}_X = (v_1, v_2, v_3, v_4) \cdot (R, S, T, P)$ and $s_Y = \mathbf{v} \cdot \mathbf{S}_Y = (v_1, v_2, v_3, v_4) \cdot (R, T, S, P)$, respectively. In the t-th stage, the total probability that player X cooperates is $p_c^t = (1, 1, 0, 0) \cdot \mathbf{v}^t$. And the probability she will cooperate in the next stage game is calculated as $p_c^{t+1} = \mathbf{p} \cdot \mathbf{v}^t$. Deriving the difference between these two probabilities, we have:

$$p_c^{t+1} - p_c^t = (\mathbf{p} - (1, 1, 0, 0)) \cdot \mathbf{v}^t. \quad (3)$$

Denote $\widetilde{\mathbf{p}} = \mathbf{p} - (1, 1, 0, 0)$. Essentially, this vector depicts to what extend of speed player X changes its probability for cooperation. Sum eq. (3) from $t = 1$ to $t = n$, then we have

$$\sum_{t=1}^{n} \widetilde{\mathbf{p}} \cdot \mathbf{v}^t = \sum_{t=1}^{n} p_c^{t+1} - p_c^t = p_c^{n+1} - p_c^1. \quad (4)$$

If the game is infinitely repeated, averaging the above equation when $n \to \infty$ ensures that

$$\widetilde{\mathbf{p}} \cdot \mathbf{v} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \widetilde{\mathbf{p}} \cdot \mathbf{v}^t = \lim_{n\to\infty} \frac{1}{n} \left(p_c^{n+1} - p_c^1\right) = 0, \quad (5)$$

where $\mathbf{v}$ is just the stationary distribution of the game's state transition matrix. Expanding the vector equation $\widetilde{\mathbf{p}} \cdot \mathbf{v} = 0$ leads to:

$$(-1 + p_1) v_1 + (-1 + p_2) v_2 + p_3 v_3 + p_4 v_4 = 0. \quad (6)$$

This relation is firstly discovered in [Press and Dyson, 2012] and then investigated in [Akin, 2012]. It significantly reveals the underlying relationship between the game's transition matrix and the unilateral strategy of a single player.

3 Control the Maximum and Minimum Values of Opponent's Payoff

If player X's objective is to ensure that Y's expected payoff

$$s_Y \le W, \quad (7)$$

where $s_Y = \mathbf{v} \cdot \mathbf{S}_Y = (v_1, v_2, v_3, v_4) \cdot (R, T, S, P)$ and W is a constant, then the objective function eq. (7) becomes $(v_1, v_2, v_3, v_4) \cdot ((R, T, S, P) - (W, W, W, W)) \le 0$. Multiplying both side with a positive factor $1 - p_2$ does

not change the inequality, thus eq. (7) is equivalent to $(s_Y - W)(1 - p_2) \leq 0$. Substituting eq. (6) into it and combining the coefficients of v_i, the objective of player X is finally reduced to an inequality as follows.

$$\alpha_1 v_1 + \alpha_3 v_3 + \alpha_4 v_4 \leq 0, \tag{8}$$

where α_1, α_3 and α_4 are the coefficients and

$$\begin{cases} \alpha_1 = (R - W)(1 - p_2) + (W - T)(1 - p_1), \\ \alpha_3 = (T - W)p_3 + (S - W)(1 - p_2), \\ \alpha_4 = (T - W)p_4 + (P - W)(1 - p_2). \end{cases} \tag{9}$$

One sufficient condition for eq. (8) to hold is that all α_i are non-positive. This further requires that the strategy of player X should fall into the following region:

$$\begin{cases} 0 \leq p_2 < 1, \\ 0 \leq p_1 \leq \min\left(1 - \frac{R-W}{T-W}(1 - p_2), 1\right), \\ 0 \leq p_3 \leq \min\left(\frac{W-S}{T-W}(1 - p_2), 1\right), \\ 0 \leq p_4 \leq \min\left(\frac{W-P}{T-W}(1 - p_2), 1\right). \end{cases} \tag{10}$$

It is shown that there are multiple candidate strategies for player X to control the maximum value of Y's possible payoff. Moreover, how to choose p_1, p_3 and p_4 depends on the value of p_2, which is the probability that X will cooperate after she cooperated but was defected by the opponent. The value of p_2 partially reflects X's bottom line for cooperation and all the other components in $\mathbf{p}$ should be adjusted accordingly. Nevertheless, as long as eqs. (10) has solutions, no matter which level of such a bottom line player X has, it is always possible for her to control the maximum value of player Y's payoff to her objective W.

Based on the above model, besides controlling the maximum value of Y's payoff, player X can also control the the minimum payoff Y can achieve. If X's objective is:

$$s_Y \geq U, \tag{11}$$

then X actually secures a bottom line for Y's payoff. Applying the similar trick as for solving eq. (7), the constraints for eq. (11) becomes:

$$\beta_1 v_1 + \beta_3 v_3 + \beta_4 v_4 \geq 0, \tag{12}$$

where

$$\begin{cases} \beta_1 = (R - U)(1 - p_2) + (U - T)(1 - p_1), \\ \beta_3 = (T - U)p_3 + (S - U)(1 - p_2), \\ \beta_4 = (T - U)p_4 + (P - U)(1 - p_2). \end{cases} \tag{13}$$

Thus one sufficient condition for $s_Y \geq U$ is that all β_i are non-negative, then the solution is:

$$\begin{cases} 0 \leq p_2 < 1, \\ \max\left(0,\ 1 - \frac{R-U}{T-U}(1 - p_2)\right) \leq p_1 \leq 1, \\ \max\left(0,\ \frac{U-S}{T-U}(1 - p_2)\right) \leq p_3 \leq 1, \\ \max\left(0,\ \frac{U-P}{T-U}(1 - p_2)\right) \leq p_4 \leq 1. \end{cases} \tag{14}$$

It is straightforward that X can set W and U simultaneously and sandwich Y's expected payoff s_Y into an intermediate region. She can do this by choosing a strategy $\mathbf{p}$ satisfying both eqs. (10) and eqs. (14). When W and U become closer to each other, the range of Y's possible payoff shrinks. In the extreme case, when controller X sets $W = U$, the region of Y's possible payoffs will be compressed into a line. At this point, $\mathbf{p}$ degenerates to an equalizer strategy, which has been discovered in [Boerlijst *et al.*, 1997] and formally discussed in [Press and Dyson, 2012].

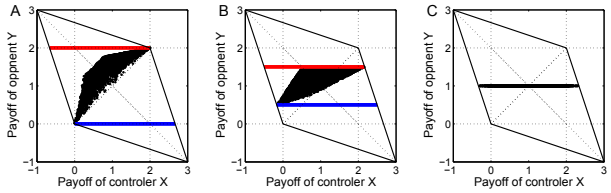

Figure 1: Control on maximum and minimum values of Y's possible payoffs. Each black dot is a possible payoff pair. In (A), (B) and (C), X's strategies are $\mathbf{p} = (1, 0.51, 1, 0)$, $\mathbf{p} = (0.998, 0.994, 0.01, 0.0012)$ and $\mathbf{p} = (0.5, 0, 1, 0.5)$, respectively.

In Figure 1 we show an example of how a payoff control on Y's maximum and minimum payoffs degenerates to a equalizer strategy. The convex hull is the space for the two players' payoff pairs (s_X, s_Y). The x-axis and y-axis are the payoff values for player X and Y, respectively. In each subfigure, X uses a control strategy and Y's strategy is randomly sampled for 5000 times. We use a traditional prisoner's dilemma payoff matrix setting $(R, T, S, P) = (2, 3, -1, 0)$. Each black dot represents a possible payoff pair consisted of X's and Y's average payoffs under the fixed control strategy of X and a random strategy of Y. The upper and lower bounds of Y's possible payoffs are depicted by the red and blue lines, respectively. In 1(A), X's control strategy yields the maximum and minimum values $W = 2$ and $U = 0$ for Y; In 1(B), X sets $W = 1.5$ and $U = 0.5$ and the possible payoff region of Y shrinks; In 1(C), the general regional payoff control finally degenerates to an equalizer strategy, under which Y's payoff is pinned to a fixed value $W = U = 1.0$. We can see that the equalizer/pinning strategy are special cases of control on the maximum and minimum values of opponent's payoffs.

4 Control of Players' Payoff Relations and Cooperation Enforcement

The above framework makes it possible to control the maximum and minimum values of the opponent's possible payoffs. In this section, we show that it is even possible for the controller X to confine the two players' possible payoff pairs in an arbitrary region of which the boundaries are characterized by linear functions. Under such a regional control, the game can be lead to a mutual-cooperation situation.

Assume the controller X wants to establish a relation between the two players' payoffs such that the opponent always obtains less than a linear combination of what she earns:

$$s_Y \leq \frac{1}{\chi} s_X + \kappa, \tag{15}$$

where $\chi \geq 1$ and $(1 - \frac{1}{\chi})R \geq \kappa \geq (1 - \frac{1}{\chi})P$. This objective claims a linear upper bound of Y's payoff and ensures that all possible payoff pairs are under it. Eq. (15) is equivalent to $(\mathbf{S}_X - \chi \mathbf{S}_Y + \chi\kappa \cdot \mathbf{1}) \cdot \mathbf{v} \geq 0$, which further leads to

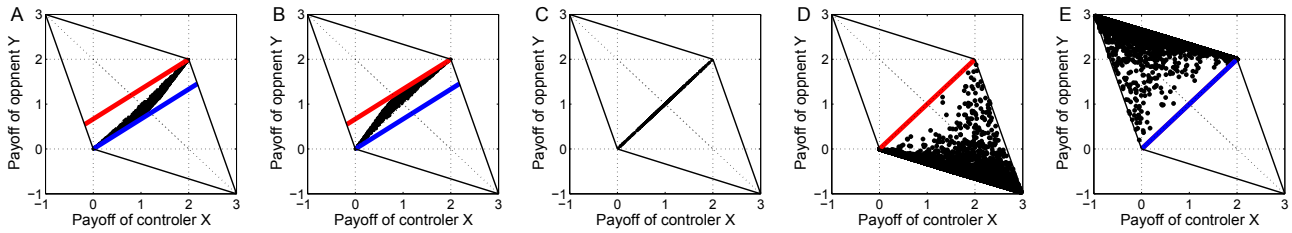

Figure 2: Control on the region of possible payoff pairs. Each black dot is a possible payoff pair. (A) is a normal winning/selfish control $\mathbf{p} = (1, 0.1, 0.75, 0)$. (B) is a normal altruist control $\mathbf{p} = (1, 0.182, 1, 0)$. (C) is a TFT-like strategy with $p_1 = 1, p_4 = 0$ and $p_2 + p_3 = 1$. (D) is an extreme case of selfish control $\mathbf{p} = (1, 0, 0.005, 0)$. (E) is an extreme case of altruist control $\mathbf{p} = (1, 0.995, 1, 0)$. (A), (B), (C) and (D) are all cooperation-enforcing.

$[(R, S, T, P) - \chi(R, T, S, P) + \chi\kappa \cdot \mathbf{1}] \cdot (v_1, v_2, v_3, v_4) \geq 0$. Multiplying both side by $1 - p_2$ and representing v_2 by using v_1, v_3 and v_4, we have:

$$\gamma_1 v_1 + \gamma_3 v_3 + \gamma_4 v_4 \geq 0, \tag{16}$$

where

$$\begin{cases} \gamma_1 = \mu(-1 + p_1) + [(1 - \chi)R + \chi\kappa](1 - p_2), \\ \gamma_3 = \mu p_3 + (T - \chi S + \chi\kappa)(1 - p_2), \\ \gamma_4 = \mu p_4 + [(1 - \chi)P + \chi\kappa](1 - p_2), \end{cases} \tag{17}$$

and $\mu = (S - \chi T + \chi\kappa)$. Making γ_1, γ_3 and γ_4 simultaneously nonnegative is sufficient to ensure that eq. (15) holds. Similarly, if X wants to establish a payoff relation such that Y always obtains more than a linear combination of what she earns, then her objective is

$$s_Y \geq \frac{1}{\chi} s_X + \kappa. \tag{18}$$

This indicates that X sets a linear lower bound of Y's payoff. Such an objective demands a payoff region above the line $s_X - \chi s_Y + \chi\kappa = 0$. To realize this, she just needs to make γ_1, γ_3 and γ_4 in eqs. (17) nonpositive simultaneously. One necessary condition for a control strategy to exist is that the objective region should be *feasible*, which means on the one hand, the objective region of the possible payoff pairs must intersect with the $(P, P) - (S, T)$ line, which is the left boundary of the payoffs in the iterated prisoner's dilemma, depicting the payoff pairs when Y unconditionally defects, i.e., $\mathbf{q} = (0, 0, 0, 0)$; on the other hand, the payoff region must also terminates at some point on the $(R, R) - (T, S)$ line, which is the right boundary depicting the possible payoff pairs when Y unconditionally cooperates, i.e., $\mathbf{q} = (1, 1, 1, 1)$.

Specifically, (1) if X controls by using objective function eq. (15) with $\chi \geq 1$ and $\kappa = (1 - \frac{1}{\chi})P$, we have $(s_X - P) \geq \chi(s_Y - P)$. Any point in the objective region ensures that X's payoff difference to P is *at least* χ times of that of Y. Under such a strategy, X only concerns about herself winning the game regardless of the opponent's outcome. This feature well captures the selfishness in nature, therefore, we call such strategies the "selfish control" strategies. For example, in a game with $P = 0$, if X sets eq. (15) with $\chi = 1.5$ and $\kappa = 0$, she always obtains more than 1.5 times of what Y obtains. $\mathbf{p} = (0.5, 0.5, 0.4, 0)$ is one of such selfish strategies. (2) If X's objective function is eq. (18) with $\chi \geq 1$ and $\kappa = (1 - \frac{1}{\chi})R$, Y's payoff difference to R will be *at most* $\frac{1}{\chi}$ times of that of X, meaning that X is offering a benefit to Y at the expense of hurting her own benefit. Since in biological organisms, altruism can be defined as an individual performing an action which is at a cost to themselves but benefits another individual, we call this strategy the "altruist control" strategies. (3) However, if X controls with constraint $(1 - \frac{1}{\chi})R > \kappa > (1 - \frac{1}{\chi})P$, who can win the game is uncertain, since whether a payoff pair locates below the diagonal line $(P, P) - (R, R)$ still depends on Y's strategy. Thus we call them "contingent control" strategies.

More generally, it is also possible for controller X to set up combinatorial objectives, such that there are multiple linear upper and/or lower bounds of Y's possible payoffs. She can do this by generalizing the constraint coefficients γ to

$$\mathbf{G}\mathbf{v}' \geq \mathbf{0}, \tag{19}$$

where $\mathbf{v}' = (v_1, v_3, v_4)$ and $\mathbf{G}$ is a coefficient matrix with each entry γ_{ij} as the j-th coefficient from the i-th control objective. Following such a regularization, the complex payoff control problem is reduced to a formal linear programming. As long as $\mathbf{G}$ constitutes a feasible payoff region, the combinatorial control objective can be realized. Under this framework of regional control with multiple constraints, various shape of payoff regions can be generated. Especially, ZD strategies are extreme cases of regional control.

If each player has chosen a certain strategy and no one can benefit by changing his strategy while the other players keep theirs unchanged, then the current set of strategy choices and the corresponding payoffs constitute a Nash equilibrium. A strategy p_N of player X is called a Nash strategy if player X can control the upper bound of player Y to R. Thus, under the general payoff control framework in eq.(19), any strategy with $s_Y \leq R$ as a tight constraint is a Nash strategy. According to the definition of Nash equilibrium, it is straightforward that any pair of Nash strategies constitute a Nash equilibrium. However, although a Nash strategy can induce a fixed upper bound R of Y's payoff, it is possible for Y to choose an alternative strategy other than fully cooperating ($\mathbf{q} = \mathbf{1}$), which still yields R for herself but with the payoff for controller X smaller than R. This is why a Nash equilibrium is not necessarily a cooperative equilibrium. So how to select out an cooperation-enforcing Nash strategy is a problem. The con-

troller can enforce cooperation by setting:

$$s_Y \leq \frac{1}{\chi} s_X + (1 - \frac{1}{\chi})R, \quad p_1 = 1, \tag{20}$$

where $\chi \geq 1$. Under such strategies of X, the only best response of Y is to fully cooperate, whereby both players finally receive payoffs R which will lead the game to a win-win situation. We call them "cooperation-enforcing control".

Under the above framework, one can derive arbitrary regional control strategies, as long as the region has feasible linear boundaries. In Figure 2, we show several examples of regional control strategies. In 2(A) and 2(B), X sets same linear upper and lower bounds for the payoff region. The red lines are upper bounds with $\chi = 1.5$ and $\kappa = -1$ and blue lines are lower bounds with $\chi = 1.5$ and $\kappa = 0$. In 2(A), X uses a selfish control where her payoff is always larger than that of Y. In 2(B), X uses an altruist control which always lets Y win. Both these two strategies are cooperation enforcing, leading the game to evolve to a mutual cooperation equilibrium. In 2(C), we shrink the controlled region to an extreme case by setting the upper and lower bounds identically with $\chi = 1.0, \kappa = 0$. The solution shows, as long as $p_1 = 1, p_4 = 0$ and $p_2 + p_3 = 1$, the control strategies have similar effect as the traditional Tit-for-tat (TFT): equalizing the two players' expected payoffs. In 2(D) X uses an extreme case of selfish control while in 2(E) X uses an extreme case of altruist control. These two cases are also investigated as partnership and rival strategies in [Hilbe *et al.*, 2015].

5 Control in Axelrod's Tournaments

Right up to today, it has been a fundamental challenge for many disciplines to understand how various strategies perform in multi-agent interactions, what is the best strategy in repeated games and why it is the best. The most influential experiments for strategy evaluation are established by Robert Axelrod as his iterated prisoner's dilemma computer tournaments [Axelrod and Hamilton, 1981]. Based on the payoff control framework, in this section, we derive several control strategies and simulate them in a tournament.

The simulated tournament is similar as in [Stewart and Plotkin, 2012] but uses a different IPD setting $(R, T, S, P) = (2, 3, -1, 0)$. Besides classic strategies, Stewart and Plotkin implemented two ZD strategies: Extort-2 with $s_X - P = 2(s_Y - P)$ and Generous-2 with $s_X - R = 2(s_Y - R)$. Their simulation shows the best performance is from Generous-2, which is followed by GTFT and TFT. We add four regional control strategies into the tournament, including Altruist_$\mathbf{G}$, Selfish_$\mathbf{G}$, Altruist_TFT and Selfish_TFT. Here $\mathbf{G}$ denotes the control objective matrix from eq. (19). Altruist_$\mathbf{G}$ is derived with respect to two objectives $s_X - R \geq 2(s_Y - R)$ and $s_X - R \leq \frac{4}{3}(s_Y - R)$, while the Selfish_$\mathbf{G}$ is derived with respect to $s_X - P \leq 2(s_Y - P)$ and $s_X - P \geq \frac{4}{3}(s_Y - P)$. Selfish_TFT and Altruist_TFT are using the same strategies as in Figure 2(A) and 2(B), respectively. It is worth noting that Altruist_$\mathbf{G}$ is essentially a regional control expansion based on Generous-2, while Selfish_$\mathbf{G}$ is expanded from Extort-2. Both Selfish_TFT and Altruist_TFT can be viewed as expansions from original TFT.

Name	p	Score	Wins
ALTRUIST_$\mathbf{G}$	$(1, 2/15, 1, 1/3)$	1.66	0
GENEROUS-2	$(1, 2/7, 1, 2/7)$	1.60	0
ALTRUIST_TFT	$(1, 0.182, 1, 0)$	1.52	0
GTFT	$(1, 2/3, 1, 2/3)$	1.46	0
TFT	$(1, 0, 1, 0)$	1.44	0
TF2T		1.43	0
HARD_TF2T		1.37	0
SELFISH_TFT	$(1, 0.1, 0.75, 0)$	1.33	1
WSLS	$(1, 0, 0, 1)$	1.29	0
HARD_PROBE		1.26	8
ALLC	$(1, 1, 1, 1)$	1.18	0
PROBE2		1.11	4
GRIM	$(1, 0, 0, 0)$	1.08	4
HARD_TFT		1.08	4
RANDOM	$(1/2, 1/2, 1/2, 1/2)$	0.92	10
HARD_MAJO		0.91	13
PROBE		0.81	6
CALCULATOR		0.76	12
PROBE3		0.72	10
HARD_JOSS	$(0.9, 0, 1, 0)$	0.72	14
SELFISH_$\mathbf{G}$	$(5/7, 0, 13/15, 0)$	0.64	15
ALLD	$(0, 0, 0, 0)$	0.45	20
EXTORT-2	$(6/7, 1/2, 5/14, 0)$	0.45	19

Table 1: Results of the IPD tournament

Due to the inherent stochasticity of some strategies, the tournament is repeated 1000 times. In a tournament, each strategy in the above set meets each other (including itself) in a perfect iterated prisoner's dilemma (IPD) game, and each IPD game has 200 stages. The average results are shown in Table 1. The shaded rows are for the control strategies derived under our framework. One can see that the Altruist_$\mathbf{G}$ has the best performance. It is better than Generous-2 and has much higher score than either TFT or GTFT. The Altruist_TFT also performs better than TFT and GTFT. The Selfish_TFT is a little tougher than TFT, although it has slightly higher number of wins. Analogously, using the above payoff control framework, one could also generate other regional control strategies which are better than the corresponding ZD strategies. Although no strategy is universally best in such tournaments, because a player's performance depends on the strategies of all her opponents as well as the environment of the game, the control framework still provides us a new perspective to formally quantify new nice strategies.

In perfect environments, TFT has long been recognized as the most remarkable basic strategy. Starting with cooperation, it can constitute a Nash equilibrium strategy that enforces long-run cooperation. Nevertheless, TFT is not flawless. The first drawback of it, which is not apparent in perfect environment, is that if one of the two interacting TFT players faces the problem of trembling hand or imperfect observation, then a false defection will leads to a sequence of alternating cooperation and defection. Then the two players both receive a payoff much less than mutual cooperation. This indicates TFT is not a subgame perfect equilibrium strategy. Another

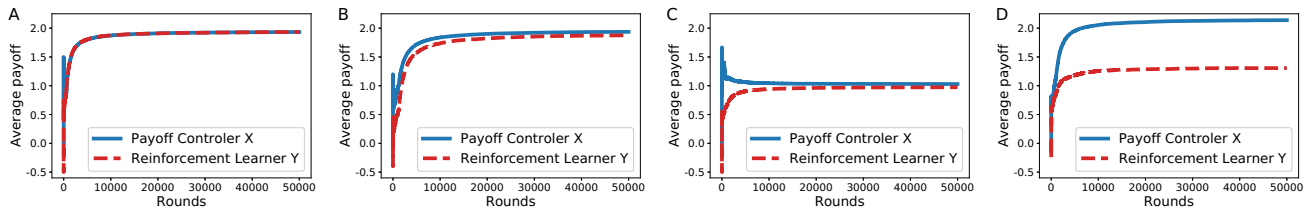

Figure 3: Control against human-like players. In each subfigure, Y uses a reinforcement learning strategy. In (A) X uses a TFT-like strategy $\mathbf{p} = (1, 0.2, 0.8, 0)$. In (B) X uses a selfish control $\mathbf{p} = (1, 0.1, 0.6, 0)$ which makes mutual cooperation $(s_X = R, s_Y = R)$ as the optimal outcome for Y. In (C) X uses selfish control $\mathbf{p} = (0.3, 0, 1, 0)$ which makes a payoff pair $(s_X = 1, s_Y = 1)$ as the optimal outcome for Y. In (D) X uses a selfish control $\mathbf{p} = (5/7, 0, 1, 0)$ which guarantees Y's payoff is much lower than that of X.

weakness of TFT is a population of TFT players can be replaced by ALLC through random drift. Once ALLC has increased to some threshold, ALLD can invade the population. The reason why TFT is vulnerable to noise is that when confronting the opponent's unilateral defect (CD), a TFT player is too vengeful and will fully defect $(p_2 = 0)$. The reason why TFT can be invaded by ALLC players is that it is not greedy at all, when it occasionally takes advantage of the opponent (DC), it completely stops defection and turns back to cooperation $(p_3 = 1)$. To conquer these drawbacks, a nice strategy in a noisy environment should necessarily embody three features: (1) It should be cooperation-enforcing, i.e., its objective payoff region should have a tight upper bound $s_Y \leq \frac{1}{\chi} s_X + (1 - \frac{1}{\chi})R$ and $\chi \geq 1$; (2) It should not be too vengeful, i.e., $p_2 > 0$, meaning its objective payoff region should not be too far from the (S, T) point; (3) It should be somewhat greedy, i.e., $p_3 < 1$, meaning its objective payoff region should not be too far from the (T, S) point.

6 Control against Human-like Players

In the real world, if a player is not aware of any nice strategies, he actually dynamically updates his stage action according to a learned history of the long-run game, and gradually evolves his own optimal plan for interacting with the opponent. In artificial intelligence, this learning and planning procedure is usually investigated by reinforcement learning models, which are state-of-the-art human like plays when agents are confronting with complex environment or strategic opponents. To try our best to understand the performance of the payoff control strategies in a real world, in this section, we simulate several repeated games between the payoff control players and the reinforcement learning players.

Let X be the payoff controller who uses a payoff control strategy obtained beforehand, and let Y be the reinforcement learner who evolves his strategy/plan $\mathbf{q}$ according to the reinforcement learning dynamics. $\mathbf{q}$ is a mapping from the game history to the probabilities of selecting $a = C$. Y's objective is to find an optimal $\mathbf{q}^*$ which maximizes his stage payoff:

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}_{\mathbf{q}} \left[s_Y^t \right] \right\}, \tag{21}$$

where s_Y^t is Y's realized stage payoff at time t and $\mathbb{E}_{\mathbf{q}}$ is an expectation with respect to $\mathbf{q}$. Y's strategy $\mathbf{q}$ is updated according to the following average-reward value function:

$$Q(\omega, a) \leftarrow (1 - \alpha) Q(\omega, a) + \alpha \left[\bar{r} + \max_{a'} Q(\omega', a') \right] \tag{22}$$

where $Q(\omega, a)$ is an evaluation value of player Y choosing action a after stage game outcome ω. $\bar{r} = r(\omega, a, \omega') - r^*$ is difference between the instantiate reward r and the estimated average reward r^*. The instantiation reward $r(\omega, a, \omega')$ is induced by player Y taking action a after outcome ω and transiting the game to a new outcome ω'. α is a free variable for the learning rate. With Q's values updated stage after stage, Y can improve his strategy dynamically [Gosavi, 2004].

We implement the above algorithm, simulate four repeated games and show the results in Figure 3. In each of these four games, player Y uses the reinforcement learning strategy described above. In 3(A) and 3(B), the strategies used by the controller X are both cooperation-enforcing. Under X's TFT-like control strategy in 3(A), X and Y always have almost the same average payoff; While under X's winning yet cooperation-enforcing strategy in 3(B), X dominates Y for a long time but the game finally converges to a mutual cooperation. In 3(C), X's objective is to set $s_X = s_Y = 1$. We can see the game finally converges as X wishes. In 3(D), X uses a very tough selfish control, which means she can win Y a lot. In this situation, when the intelligent agent Y improves his own payoff step by step, he improves that of the controller even more. In a word, when playing against human-like reinforcement learning players, payoff control strategy players can lead the game to evolve to their objective outcomes.

7 Conclusions

We propose a general framework for controlling the linear boundaries of the region where the repeated game players' possible payoff pairs lie. By generating payoff control strategies under this framework, a single player can unilaterally set arbitrary boundaries on the two players' payoff relation and thereby realize her control objective, including limiting the maximum and minimum payoffs of the opponent, always winning the game, offering an altruist share to the opponent, enforcing the game to converge to a win-win outcome, and so on. The idea in this work is not limited to iterated prisoner's dilemma, it can be introduced into other two player repeated games and also can be generalized for repeated multi-player games, such as iterated public goods games. Future

researches on the payoff control in repeated games with imperfect monitoring [Hao *et al.*, 2015], with different memory sizes [Li and Kendall, 2014] and researches investigating better winning or cooperation-enforcing strategies [Crandall *et al.*, 2018; Mathieu and Delahaye, 2015] could be potentially inspired by this work. Furthermore, all the control strategies are based on the important premise that the player is with a theory of mind [Devaine *et al.*, 2014]. Therefore, how to identify more cognitively complex human-like strategies in the context of the IPD, such as intention recognition [Han *et al.*, 2012], is of great value for the future research.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (NNSFC) under Grant No. 71601029 and No. 61761146005. Kai Li also acknowledges the funding from Ant Financial Services Group.

References

[Adami and Hintze, 2013] Christoph Adami and Arend Hintze. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nature Communications*, 4:2193, 2013.

[Akin, 2012] Ethan Akin. Stable cooperative solutions for the iterated prisoner's dilemma. *arXiv preprint*, 1211.0969, 2012.

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

[Boerlijst *et al.*, 1997] Maarten Boerlijst, Martin A Nowak, and Karl Sigmund. Equal pay for all prisoners. *The American Mathematical Monthly*, 104(4):303–305, 1997.

[Chen and Zinger, 2014] Jing Chen and Aleksey Zinger. The robustness of zero-determinant strategies in iterated prisoner's dilemma games. *Journal of Theoretical Biology*, 357:46–54, 2014.

[Claus and Boutilier, 1998] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *In Proceedings of the 15th International Conference on Artificial Intelligence*, 1998:746–752, 1998.

[Crandall *et al.*, 2018] Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. Cooperating with machines. *Nature communications*, 9(1):233, 2018.

[Devaine *et al.*, 2014] Marie Devaine, Guillaume Hollard, and Jean Daunizeau. Theory of mind: did evolution fool us? *PloS One*, 9(2):e87619, 2014.

[Gosavi, 2004] Abhijit Gosavi. Reinforcement learning for long-run average cost. *European Journal of Operational Research*, 155(3):654–674, 2004.

[Han *et al.*, 2012] The Anh Han, Luis Moniz Pereira, and Francisco C Santos. Corpus-based intention recognition in cooperation dilemmas. *Artificial Life*, 18(4):365–383, 2012.

[Hao *et al.*, 2014] Dong Hao, Zhihai Rong, and Tao Zhou. Zero-determinant strategy: An underway revolution in game theory. *Chinese Physics B*, 23(7):078905, 2014.

[Hao *et al.*, 2015] Dong Hao, Zhihai Rong, and Tao Zhou. Extortion under uncertainty: Zero-determinant strategies in noisy games. *Physical Review E*, 91(5):052803, 2015.

[Hilbe *et al.*, 2013] Christian Hilbe, Martin A Nowak, and Karl Sigmund. Evolution of extortion in iterated prisoner's dilemma games. *Proceedings of the National Academy of Sciences*, 110(17):6913–6918, 2013.

[Hilbe *et al.*, 2014] Christian Hilbe, Bin Wu, Arne Traulsen, and Martin A Nowak. Cooperation and control in multiplayer social dilemmas. *Proceedings of the National Academy of Sciences*, 111(46):16425–16430, 2014.

[Hilbe *et al.*, 2015] Christian Hilbe, Arne Traulsen, and Karl Sigmund. Partners or rivals? strategies for the iterated prisoner's dilemma. *Games and Economic Behavior*, 92:41–52, 2015.

[Kandori, 2002] Michihiro Kandori. Introduction to repeated games with private monitoring. *Journal of Economic Theory*, 102(1):1–15, 2002.

[Li and Kendall, 2014] Jiawei Li and Graham Kendall. The effect of memory size on the evolutionary stability of strategies in iterated prisoner's dilemma. *IEEE Transactions on Evolutionary Computation*, 18(6):819–826, 2014.

[Mailath and Samuelson, 2006] George J Mailath and Larry Samuelson. *Repeated Games and Reputations: Long-Run Relationships*. Oxford university press, 2006.

[Mathieu and Delahaye, 2015] Philippe Mathieu and Jean-Paul Delahaye. New winning strategies for the iterated prisoner's dilemma. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1665–1666, 2015.

[Nowak *et al.*, 1995] Martin A Nowak, Karl Sigmund, and Esam El-Sedy. Automata, repeated games and noise. *Journal of Mathematical Biology*, 33(7):703–722, 1995.

[Pan *et al.*, 2015] Liming Pan, Dong Hao, Zhihai Rong, and Tao Zhou. Zero-determinant strategies in iterated public goods game. *Scientific reports*, 5:13096, 2015.

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

[Stewart and Plotkin, 2012] Alexander J Stewart and Joshua B Plotkin. Extortion and cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences*, 109(26):10134–10135, 2012.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press Cambridge, 1998.