On Learning Stable Cooperation in the Iterated Prisoner's Dilemma with Paid Incentives

Xiyue Sun
LMU Munich

Fabian R. Pieroth
Technical University of Munich

Kyrill Schmid
LMU Munich

Martin Wirsing
LMU Munich

Lenz Belzner
Technische Hochschule Ingolstadt

***Abstract*—An essential step towards collective intelligence in systems comprised of multiple independent and autonomous agents is that individual decision-makers are capable of acting cooperatively. Cooperation is especially challenging in environments where collective and individual rationality diverge, like in the Prisoner's Dilemma (PD), which is often used to test whether algorithms are capable of circumventing the single non-optimal Nash equilibrium. In this paper, we extend the approach "Learning to Incentivize other Learning Agents" in two ways: 1. We analyze the impact of the payoff matrices on incentive updates, as different payoff matrices could accelerate or decelerate the growth of incentives. 2. We adapt the concept of the market from "Action Markets in Deep Multi-Agent Reinforcement Learning" to iterated PD games as to trade incentives, i.e., the final revenue of the agent is the game revenue minus the incentive it provided, and propose (sufficient) conditions for reaching stable two-way cooperation under specific assumptions.**

***Index Terms*—Prisoner's Dilemma, Cooperation, Incentivizing, Reinforcement Learning**

I. Introduction

Numerous applications critically depend on the decision-makers' ability to behave cooperatively. Examples are autonomous vehicles [1], smart clouds, and more generally, ensembles [2] and collective adaptive systems [3]. Such systems comprise heterogeneous decision-makers and require the ability that individuals can act in the general interest to assure that their decisions do not lead to dysfunctional or even catastrophic system behavior. Despite that, there is evidence that independently trained agents tend to act defectively or develop overly greedy strategies, particularly when shared resources are scarce [4]. Such behavior might have fatal consequences in scenarios with depletable resources where greediness gives rise to the tragedy of the commons and eventually might lead to the total exhaustion of resources [5]. This raises the question of how independently trained agents with individual goals and objectives can be incentivized to make collectively desirable decisions.

To address this question, in this work, we apply methods from the field of reinforcement learning (RL) to model (boundedly) rational decision-making of individual agents. After reinforcement learning for individual agents has been extensively studied, more and more researchers are focusing on how cooperation can be maintained for maximum welfare in multi-agent environments [6]. Game-theory and social dilemma are inevitable topics in this discussion. Humans have difficulty deciding in a social dilemma situation, as individuals need to choose between increasing their benefits at the cost of the overall good and giving up some of their individual benefits to maximize the overall payoffs. In a social dilemma like the Prisoner's Dilemma (PD), the simple equilibrium is continuous mutual betrayal. Bó and Fréchette [7] conducted 18 experimental sessions with 266 participants, where the simple PD game is played repeatedly with a given probability of continuation. These experiments showed that the evolution of cooperation is independent of experience gained by the subjects, and cooperation may not prevail even when it is a possible equilibrium.

Inspired by humans incentivizing others to influence their behavior, "learning to incentivize others" (LIO) via inter-agent incentivization was proposed by Yang et al. [8]. LIO allows agents to give rewards directly to others in a multi-agent environment. Agents also learn their incentive functions by considering the recipients' reactions. The emergence of stable cooperation in this setup was analyzed in an iterated PD (IPD) with a particular payoff matrix. Agents have the memory of the last iteration, including probabilities of cooperation as well as incentives provided by the opponent. Based on this observation, each agent's policy and incentive function are updated using gradient ascent. They proved that two LIO agents converge to mutual cooperation.

Due to the particular payoff matrix used in [8], incentives increase or decrease at a fixed rate regardless of agents' current willingness to cooperate. Therefore, some terms related to the payoff matrix in the update equations are eliminated. In addition, agents are not required to "pay" incentives in the LIO formulation. That means agents can create new value by incentivization rather than transferring utility in the sense of a market. With these "unpaid" incentives, the incentives provided could be greater than the expected benefits, and because of the loose restrictions on the scope of incentives, the continuously increasing incentives for cooperation will make mutual cooperation (CC) a global Nash equilibrium after a certain time.

In this work, we tackle these shortcomings of the LIO setup by using generic payoff matrices for the IPD in our analysis and requiring agents to "pay" for any incentives they provide to others. We make the following contributions:

- We provide a formal description of the IPD with paid incentives with parametrized payoff matrices.
- We analyze convergence to stable cooperation in IPD with paid incentives.

- Our main theorem provides sufficient conditions required for stable two-way cooperation, or the lack thereof, under specific assumptions.
- We support our theoretical findings with empirical results from numerical experiments.

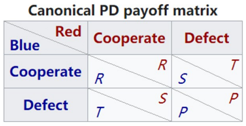
The paper is structured as follows: Section II introduces related work on learning to collaborate. In Section III, we formalize the IPD with paid incentives and prove our main theorem. Section IV contains empirical results in support of our theory. Finally, Section V concludes and points to directions of further work.

II. Related Work

The question of emerging cooperation between independent decision-makers has been historically addressed in game-theory. Cooperative AI [6] complements this field by utilizing artificial intelligence that allows to analyze complex scenarios for which game-theoretic models are hardly applicable. One line of cooperative AI uses machine learning methods to analyze the dynamics in complex multi-agent systems. Perolat et al. [9] apply reinforcement learning to estimate equilibria in common-pool resource domains where temporal and spatial aspects make theoretic models inapplicable. Also in this line of work, Leibo et al. [4] extend social dilemmas to bring them closer towards real-world situations and find that independently trained agents tend to become less cooperative the scarcer shared resources become. Other approaches aim at developing cooperative algorithms, e.g., that resemble proven game-theoretic strategies. Lerer and Peysakhovich [10] introduce a method that learns the tit-for-tat strategy. This strategy is considered cooperative but also has other intriguing features like forgiving without being exploitable [11].

More recently, different peer incentivization mechanisms have been presented that rely on the direct exchange of rewards to promote cooperation among independent agents. One approach called *action trading* lets agents exchange environmental rewards against their actions, so agents can incentivize each other to specific behaviors [12]. A related approach called *gifting* allows agents to give rewards to their peers unconditionally to let them collaborate more effectively by sharing their rewards [13].

The starting point of our work is the paper *learning to incentivize other learning agents (LIO)* [8]. In LIO an incentive function is defined that outputs the amount of reward the opponent agent receives as an incentive to become more cooperative.This extension demonstrates that a policy gradient algorithm applied to a variant of the Prisoner's dilemma converges to a fully cooperative policy. In this work, we further investigate the incentive approach in the following way. Here, agents can only incentivize others by paying them with their environmental reward, so reward cannot be created out of anything. In that sense, reward defines a locally-conserved quantity that might prevent pathological behaviors from emerging [14]. Also, in this work, instead of analyzing a single variant of the Prisoner's Dilemma, we focus on the generalized form of the Prisoner's Dilemma, which is defined by the set of inequalities given in section III.



Canonical PD payoff matrix

Red Blue	Cooperate	Defect
Cooperate	R R	T S
Defect	S T	P P

Fig. 1. Payoff matrices in PD game.

III. Iterated Prisoner's Dilemma with Incentives

The Prisoner's Dilemma reflects that the best choice of individuals does not result in the best choice for groups. It is also a classic example of game-theory in which multiple participants react to each other and influence the gains of others [15]. This section focuses on the effect of each variable on the emergence of cooperation under different revenue settings in IPD games with agents who are able to incentivize others.

Using the common notations for the PD game, R is the payoff both agents could gain in a mutual cooperative (CC) situation. In contrast, two defectors could get a profit of P individually (DD). A player can gain the highest payoff T by choosing defection while the opponent is cooperating. In this case, the cooperator only gets the minimum profit S (CD/DC). A parametrized payoff matrix for a PD is shown in Fig. 1.

To be a PD game in a strong sense, the following inequalities must hold:

$$T > R > P > S \tag{1}$$

$$2R > T + S, \tag{2}$$

which restrict the payoffs in PD games. The PD reflects a social dilemma, because of the condition (2) alternative exploitation could not be better for both players than mutual cooperation, since patient players can improve by defecting alternately [16].

If the same players repeatedly face off in the PD game, the resulting interaction is called an iterated Prisoner's Dilemma (IPD) [17]. Yang et al. [18] study this interaction under the addition that agents can incentivize each other. In their interaction, RL agents in a shared multi-agent environment can learn an incentive function to reward other agents by explicitly accounting for the impact of incentives on their own performance through recipients' learning.

A PD with incentives gives the agents an additional choice by providing an incentive for cooperation or defection. The given incentive influences the received payoff of the agents.

Definition 1. *A PD game with incentives can be described by the Tuple* $(\mathcal{N}, \mathcal{A}, \mathcal{H}, r)$.

- $\mathcal{N} = \{1, 2\}$ *is the set of players.*
- $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ *is the joint action space with* $\mathcal{A}_i = \{C, D\}$.
- $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ *with* $\mathcal{H}_i \subset (\mathbb{R}_0^+)^{|\mathcal{A}_i|}$ *is a set of feasible incentives provided for the individual actions.*

- *$r = (r^1, r^2)$ is the tuple of reward functions with $r^i : \mathcal{A} \times \mathcal{H} \to \mathbb{R}$.*

If there are no incentives, the dominant strategy is to betray in every round, as mutual betrayal is the unique dominant strategy in PD games. On the contrary, if the incentive for cooperation reaches a certain value, the dynamics change and can lead to mutual cooperation becoming the dominant strategy.

Deducing an iterated game from a PD with incentives results in an IPD game with incentives. For each iteration t, the agent with index i observes the probability θ_t^i of cooperating and the incentive η_t^i provided by the opponent. The rewards r^i for each agent can depend on the agents actions and incentives. Furthermore, future rewards are discounted by a factor γ. Using this information, agents update their policy parameter to θ_{t+1}^i and their incentives to η_{t+1}^i. More specifically, we have the following definition.

Definition 2. *An IPD with incentives is an iterated extension of a PD game with incentives $(\mathcal{N}, \mathcal{A}, \mathcal{H}, r)$ and consists of the following elements for each round t:*

- *Agent i chooses its action a_t^i to be cooperation (C) with probability θ_t^i and defection (D) with probability $1 - \theta_t^i$. Furthermore, it provides two kinds of incentives to the opposing agent. The incentive $\eta_t^i = (\eta_{C,t}^i, \eta_{D,t}^i) \in \mathcal{H}_i$ consists of the incentive for cooperation $\eta_{C,t}^i$ and the incentive for defection $\eta_{D,t}^i$.*
- *Agent i's reward is given by $r_t^i = \mathbb{E}[r^i(a_t, \eta_t)|\theta_t]$.*
- *Agent i observes $\theta_t = (\theta_t^1, \theta_t^2)$ and η_t and updates its probabilities for cooperation and incentives to θ_{t+1}^i and η_{t+1}^i respectively.*
- *Agent i's objective is to maximize its discounted long-term return $J^i(\theta_t, \eta_t) = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\left[r_k^i(a_k^1, a_k^2, \eta_t)|\theta_t\right]$.*

In [18], the policy is decoupled from incentivization, taking regular actions and giving incentives are two fundamentally different behaviors. That means agent i is not penalized for the behavior of its incentive, i.e., r^i is independent of η_t^i. This paper considers whether incentivization can be seen as a trading behavior. The agents exchange profits through incentivizing, thus achieving the effect of influencing each other's behavior.

We expand the result of [18] by applying the concept of markets [12], [19] and explore the conditions under which mutual cooperation can be expected when players need to pay for their own incentives. We consider the following dynamics of an IPD game with incentives. At the beginning of each iteration t, agents observe probabilities of cooperation θ_t, and the probabilities for four possible situations [CC, CD, DC, DD] are calculated and stored in the set p_t with

$$p_t = \left[\theta_t^1 \theta_t^2, \theta_t^1 \left(1 - \theta_t^2\right), \left(1 - \theta_t^1\right) \theta_t^2, \left(1 - \theta_t^1\right)\left(1 - \theta_t^2\right)\right]^T. \tag{3}$$

The rewards for players are the sum of fixed payoff in the payoff matrix and incentives from the opponent, minus the incentives they provide to the opponent. The set of feasible incentives $\mathcal{H}$ is restricted to $[0, T - S]^4$, where the maximum value corresponds to the maximum achievable benefit of a defector in the case of (CD) or (DC).

The achievable reward vectors for each agent are:

$$\begin{aligned} r_t^1 =& [R + \eta_{C,t}^2 - \eta_{C,t}^1, S + \eta_{C,t}^2 - \eta_{D,t}^1, \\ & T + \eta_{D,t}^2 - \eta_{C,t}^1, P + \eta_{D,t}^2 - \eta_{D,t}^1]^T \end{aligned} \tag{4}$$

$$\begin{aligned} r_t^2 =& [R + \eta_{C,t}^1 - \eta_{C,t}^2, T + \eta_{D,t}^1 - \eta_{C,t}^2, \\ & S + \eta_{C,t}^1 - \eta_{D,t}^2, P + \eta_{D,t}^1 - \eta_{D,t}^2]^T \end{aligned} \tag{5}$$

The value function for each agent is defined the same as in [18] and coincides with the expected long-term return:

$$V^i\left(\theta_t, \eta_t\right) = \sum_{k=0}^{\infty} \gamma^k p_t^T r_t^i = \frac{1}{1-\gamma} p_t^T r_t^i. \tag{6}$$

Define a projection on an interval by

$$\Gamma_{[b,c]}(x) = \begin{cases} b, & \text{for } x < b \\ x, & \text{for } x \in [b, c] \\ c, & \text{for } x > c \end{cases} \tag{7}$$

Agent 2 updates its policy by calculating the gradient of the value function from (6) with learning rate α:

$$\begin{aligned} \theta_{t+1}^2 =& \Gamma_{[0,1]}\Big(\theta_t^2 + \alpha \nabla_{\theta_t^2} V^2\left(\theta_t, \eta_t\right)\Big) \\ =& \Gamma_{[0,1]}\Big(\theta_t^2 + \frac{\alpha}{1-\gamma} \nabla_{\theta_t^2}\big(\theta_t^1 \theta_t^2 \left(R + \eta_{C,t}^1 - \eta_{C,t}^2\right) \\ & + \theta_t^1 \left(1 - \theta_t^2\right)\left(T + \eta_{D,t}^1 - \eta_{C,t}^2\right) \\ & + \left(1 - \theta_t^1\right) \theta_t^2 \left(S + \eta_{C,t}^1 - \eta_{D,t}^2\right) \\ & + \left(1 - \theta_t^1\right)\left(1 - \theta_t^2\right)\left(P + \eta_{D,t}^1 - \eta_{D,t}^2\right)\big)\Big) \\ =& \Gamma_{[0,1]}\left(\theta_t^2 + \Delta_t^2\right), \end{aligned} \tag{8}$$

where its policy update Δ_t^2 is

$$\Delta_t^2 := \frac{\alpha}{1-\gamma}[(R + P - T - S)\,\theta_t^1 + \eta_{C,t}^1 - \eta_{D,t}^1 + S - P]. \tag{9}$$

Similarly, for agent 1:

$$\theta_{t+1}^1 = \Gamma_{[0,1]}\left(\theta_t^1 + \Delta_t^1\right) \tag{10}$$

$$\Delta_t^1 := \frac{\alpha}{1-\gamma}[(R + P - T - S)\,\theta_t^2 + \eta_{C,t}^2 - \eta_{D,t}^2 + S - P]. \tag{11}$$

The update of incentives with learning rate β provided by agent 1 based on the new rewards is:

$$
\begin{aligned}
\eta^1_{t+1} =& \Gamma_{[0,T-S]}\Big(\eta^1_t + \beta\nabla_{\eta^1_t}\frac{1}{1-\gamma}p^T_{t+1}r^1_t\Big)\\
=& \Gamma_{[0,T-S]}\Big(\eta^1_t + \frac{\beta}{1-\gamma}\nabla_{\eta^1_t}\left[\theta^1_{t+1}\left(\theta^2_t+\Delta^2_t\right)\right.\\
&\times\left(R+\eta^2_{C,t}-\eta^1_{C,t}\right)\\
&+\theta^1_{t+1}\left(1-\theta^2_t-\Delta^2_t\right)\left(S+\eta^2_{C,t}-\eta^1_{D,t}\right)\\
&+\left(1-\theta^1_{t+1}\right)\left(\theta^2_t+\Delta^2_t\right)\left(T+\eta^2_{D,t}-\eta^1_{C,t}\right)\\
&\left.+\left(1-\theta^1_{t+1}\right)\left(1-\theta^2_t-\Delta^2_t\right)\left(P+\eta^2_{D,t}-\eta^1_{D,t}\right)\right]\Big)\\
=& \Gamma_{[0,T-S]}\Big(\eta^1_t + \frac{\beta}{1-\gamma}\nabla_{\eta^1_t}\left[\theta^1_{t+1}\Delta^2_t(R+P-T-S)\right.\\
&+\Delta^2_t(T-P)+\theta^2_t(\eta^1_{D,t}-\eta^1_{C,t})\\
&\left.+\Delta^2_t(\eta^1_{D,t}-\eta^1_{C,t})-\eta^1_{D,t}\right]\Big)\\
=& \Gamma_{[0,T-S]}\Big(\eta^1_t + \frac{\beta}{1-\gamma}\nabla_{\eta^1_t}\left[\theta^1_{t+1}\Delta^2_t(R+P-T-S)\right.\\
&\left.+\Delta^2_t(T-P)+\theta^2_{t+1}(\eta^1_{D,t}-\eta^1_{C,t})-\eta^1_{D,t}\right]\Big)\\
=& \Gamma_{[0,T-S]}\Big(\eta^1_t + \frac{\beta}{1-\gamma}\left[\begin{array}{c} r^1_{C,t}\\ r^1_{D,t}\end{array}\right]\Big),
\end{aligned}
\tag{12}
$$

where $r^i_{C,t}$ and $r^i_{D,t}$ are given by

$$
r^1_{C,t} = \frac{\alpha}{1-\gamma}[\theta^1_{t+1}(R+P-T-S)+(T-P)]-\theta^2_{t+1} \tag{13}
$$

and

$$
\begin{aligned}
r^1_{D,t} =& -\frac{\alpha}{1-\gamma}[\theta^1_{t+1}(R+P-T-S)+(T-P)]\\
&+\theta^2_{t+1}-1\\
=& -r^1_{C,t}-1.
\end{aligned}
\tag{14}
$$

Likewise, the incentive update for agent 2 is

$$
\begin{aligned}
\eta^2_{t+1} =& \Gamma_{[0,T-S]}\Big(\eta^2_t + \beta\nabla_{\eta^2_t}\frac{1}{1-\gamma}p^T_t r^2_t\Big)\\
=& \Gamma_{[0,T-S]}\Big(\eta^2_t + \frac{\beta}{1-\gamma}\left[\begin{array}{c} r^2_{C,t}\\ r^2_{D,t}\end{array}\right]\Big),
\end{aligned}
\tag{15}
$$

where

$$
r^2_{C,t} = \frac{\alpha}{1-\gamma}[\theta^2_{t+1}(R+P-T-S)+(T-P)]-\theta^1_{t+1}, \tag{16}
$$

and

$$
\begin{aligned}
r^2_{D,t} =& -\frac{\alpha}{1-\gamma}[\theta^2_{t+1}(R+P-T-S)+(T-P)]\\
&+\theta^1_{t+1}-1\\
=& -r^2_{C,t}-1.
\end{aligned}
\tag{17}
$$

We are now ready to define the IPD with paid incentives with these definitions in place.

Definition 3. *An IPD game with* paid incentives *is an IPD game with incentives satisfying the following specifications:*

- *The theoretical profits an agent could receive is the sum of the predefined payoff in the PD game (Fig.1) and the incentives provided by its opponent, minus the incentives provided by itself for corresponding actions in each iteration.*
- *The set of feasible incentives for agent i is $\mathcal{H}_i = [0, T-S]^2 \subset \mathbb{R}^2$ for $i \in \{1,2\}$.*[1]
- *In each iteration, agents update the probability of cooperation using (8) and (10), updating the incentives using (12) and (15).*

Algorithm 1 implements an IPD with paid incentives.

```
Algorithm 1 IPD game with paid incentives
Inputs:
    $S, R, P, T$ satisfying (1) and (2)
    discount factor: $\gamma$; learning rates: $\alpha, \beta$
    maximum time-step: $t_{max}$
Initialize:
    $t \leftarrow 0$
    Assign random values $\theta^i_t \in [0,1]$ and
    $\eta^i_t \in [0, T-S]^2$, $i = 1, 2$
for $t < t_{max}$ do
    Generate the set $p_t$ following (3)
    Generate the sets $r^i_t$ of rewards for two agents using (4) and (5)
    Compute the value function according to (6)
    Update $\theta^i_{t+1}$ using (8) and (10)
    Update $\eta^i_{t+1}$ using (12) and (15)
    $\theta^i_t \leftarrow \theta^i_{t+1}$, $\eta^i_t \leftarrow \eta^i_{t+1}$, $t \leftarrow t+1$
end for
```

We now want to establish sufficient conditions for stable cooperation, or lack thereof, in the IPD with paid incentives. In Lemma 1 we prove that incentives for betrayal always decrease to zero regardless of their initial values.

Lemma 1. *In an IPD with paid incentives, the incentives for defection are guaranteed to reach 0 after at most $K_1 := \lceil\frac{(T-S)(1-\gamma)^2}{\alpha\beta\cdot k_{min}}\rceil$ steps.*

Proof. We show the statement by deriving an upper bound for the agents' incentive updates for defection $r^i_{D,t}$ for $i \in \{1,2\}$. This upper bound is negative, ensuring that the incentive for defection reaches 0 in a finite amount of steps. Define the function $f : [0,1] \to \mathbb{R}, x \mapsto (R+P-T-S)x + T - P$. As f is linear on a compact interval, we have

$$
\begin{aligned}
\min_{x\in[0,1]} f(x) &= \min\{f(0), f(1)\}\\
&= \min\{T-P, R-S\} =: k_{\min}.
\end{aligned}
$$

Note that $k_{\min} > 0$ due to (1). Furthermore, define $g : [0,1]^2 \to \mathbb{R}, (x,y) \mapsto -\frac{\alpha}{1-\gamma}f(x) + y - 1$. Then it holds that $g(\theta^1_{t+1}, \theta^2_{t+1}) = r^2_{D,t}$ and $g(\theta^2_{t+1}, \theta^1_{t+1}) = r^1_{D,t}$. As g is linear in each argument, we see that an upper bound is given by

$$
r^i_{D,t} \leq \max_{(x,y)\in[0,1]^2} g(x,y) \leq -\frac{\alpha}{1-\gamma}k_{\min},
$$

[1] Without considering incentives, a defector could get the maximum fixed benefit in the case of (DC). The maximum value of the incentive provided is set not to exceed the difference between T and S.

for $i \in \{1, 2\}$ and all $t \geq 0$. Therefore, player i's incentive for defection $\eta^i_{D,t} \in [0, T-S]$ decreases by at least $\frac{\alpha\beta}{(1-\gamma)^2}k_{\min}$ in every step. That results in $\eta^i_{D,t} = 0$ for $t \geq \lceil \frac{(T-S)(1-\gamma)^2}{\alpha\beta \cdot k_{\min}} \rceil$. □

Compared to the monotonically decreasing $\eta^i_{D,t}$, $\eta^i_{C,t}$ may increase or decrease under different conditions. When the opponent has shown a high willingness to cooperate, an appropriate reduction in the cooperative incentive value can be expected, which could help to maximize the agent's benefit. The conditions under which stable two-way cooperation can be achieved are discussed below. To simplify the analysis, we make the following assumption.

Assumption 1. *Both players have the same initial values for the probability of cooperation ($\theta^1_0 = \theta^2_0$) and value of incentives ($\eta^1_0 = \eta^2_0$).*

We denote an IPD where Assumption 1 holds as *symmetric* IPD. According to Lemma 1, the defective incentive $\eta^i_{D,t}$ decreases monotonically to zero after finite time steps and only affects the time when cooperation occurs, thus it can be regarded as zero in the following analysis.

Theorem 2. *Consider a symmetric IPD with paid incentives. Then, sustained mutual cooperation is guaranteed to occur if* $\frac{\alpha}{1-\gamma}(R-S) - 1 \geq 0$. *Furthermore, if* $\frac{\alpha}{1-\gamma}(R-S) - 1 < 0$, *the probability of cooperation does not converge towards* 1.

Proof. Due to Assumption 1, the incentives and strategies of both players are identical. Therefore, it holds that $\eta^1_t = \eta^2_t =: \tilde{\eta}_t$ and $\theta^1_t = \theta^2_t =: \tilde{\theta}_t$ for every $t \geq 0$. Define the update for the incentive to cooperate $r^1_{C,t} = r^2_{C,t}$ as function

$$h(x) = \left(\frac{\alpha}{1-\gamma}(R+P-T-S) - 1\right)x + \frac{\alpha}{1-\gamma}(T-P). \tag{18}$$

The function h is linear, connecting the two points

$$\{h(0), h(1)\} = \left\{\frac{\alpha}{1-\gamma}(T-P), \frac{\alpha}{1-\gamma}(R-S) - 1\right\}.$$

a) Case 1: $\frac{\alpha}{1-\gamma}(R-S) - 1 \geq 0$*:* In this case, it holds that $h(x) \geq 0$ for all $x \in [0, 1]$. Therefore, the sequence $\{\tilde{\eta}_{C,t}\}_{t \geq 0}$ is monotonically increasing. We conduct a proof by contradiction. Suppose the sequence $\{\tilde{\theta}_t\}_{t\geq 0}$ does not converge to 1. Then, there exists an $\epsilon_0 > 0$ such that for any $k \in \mathbb{N}$, there exists a $\tilde{t} \geq k$ such that $\tilde{\theta}_{\tilde{t}} < 1 - \epsilon_0$. For any such $\tilde{t}$, $\tilde{\eta}_{C,\tilde{t}}$ increases by at least $h(1-\epsilon_0) > 0$. Therefore, $\tilde{\eta}_{C,t} \to T - S$. However, that means there exists an $\epsilon_1 > 0$ and a $k_2 \in \mathbb{N}$ such that $\tilde{\theta}_t$ increases by at least

$$\Delta^i_t \geq \frac{\alpha}{1-\gamma}\min\{T-P, R-S\} - \epsilon_1 > 0$$

for every $t \geq k_2$ and $i \in \{1, 2\}$. This results in $\tilde{\theta}_t \to 1$, a contradiction. Therefore, $\tilde{\theta}_t \to 1$.

b) Case 2: $\frac{\alpha}{1-\gamma}(R-S) - 1 < 0$*:* We show that $\tilde{\theta}_t \nrightarrow 1$. It is sufficient to show that there exists an $\epsilon_2 > 0$ such that for any t with $\tilde{\theta}_t \in (1-\epsilon_2, 1]$, there exists a $k_{\epsilon_2} \in \mathbb{N}$ such that $\tilde{\theta}_{t+k_{\epsilon_2}} < 1 - \epsilon_2$. As $\min_{x\in[0,1]} h(x) = h(1) = \frac{\alpha}{1-\gamma}(R - S) - 1 < 0$, there exists an $\epsilon_3 > 0$ such that $h(1-\epsilon_3) < 0$. We choose $\epsilon_2 := \min\{\epsilon_3, |\max\{R-T, S-P\}|\}$. If $\tilde{\theta}_{t+l}$ is smaller than $1 - \epsilon_2$ for $1 \leq l \leq \lceil (T-S)/h(1-\epsilon_2) \rceil$, we are done. Otherwise, the incentive for cooperation decreases by at least $h(1-\epsilon_2)$ in every step. Leading to $\tilde{\eta}_{C,t+l} = 0$ for some $l \leq \lceil (T-S)/h(1-\epsilon_2) \rceil$. However, that means $\Delta(\tilde{\theta}_{t+l}) \leq \max\{R-T, S-P\} < -\epsilon_2$. By the choice of ϵ_2, there exists a $k_{\epsilon_2} \leq l+1$ such that $\tilde{\theta}_{t+k_{\epsilon_2}} < 1 - \epsilon_2$, which gives us the statement. □

IV. Experimental Results

In order to demonstrate this conclusion empirically, Fig. 2 compares the impact of different learning rates on the results using a numerical example, wherein $\theta^1_0 = \theta^2_0 = 0.5, R = -1, T = 0, S = -3, P = -2, \eta^1_0 = \eta^2_0 = [0, 0]^T$ to simplify the analysis and comparison. Since both players have the same initial probability of cooperation, i.e. the game is symmetric. If we set $\gamma = 0.99$, α needs to be greater than or equal to 0.005 to fulfill the condition for mutual cooperation in Theorem 2.

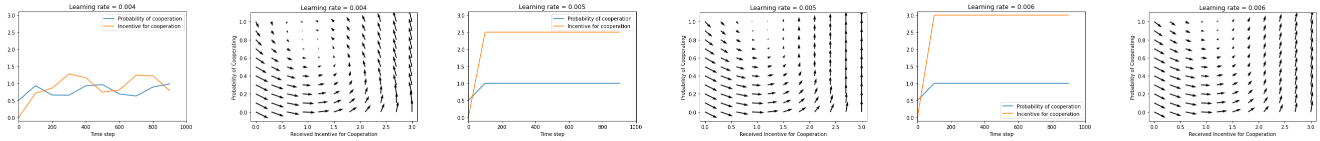
V. Summary and Further Work

On the basis of the incentivizing concept of [18], we give players the ability to trade incentives in the above-defined IPD. Furthermore, the conditions that need to be satisfied to achieve stable two-way cooperation when players need to pay incentives to their opponents are discussed in Theorem 2. The restriction on the factors in Theorem 2 is mainly to balance the negative effect of the opponent's cooperation probability on the incentive increase. Incentive reduction should be avoided when a player just demonstrates a very low probability of cooperation, otherwise, it will further cause the player's willingness to cooperate to decrease.

Specialized assumptions are made to simplify the analysis in the system, such as the symmetry of two players and zero incentive for betrayal. As the results above suggest that the incentive for betrayal in IPD games is continuously falling, it only affects the time it takes for players to switch to a cooperative strategy but does not affect the emergence and development of cooperative trends. The payoff matrices can influence the rate of incentive renewal, further altering the timing of the emergence of stable two-way cooperation.

In an IPD game, when the value of S is very small, it is risky for the players to choose to cooperate in each round. Matthias Blonski and Giancarlo Spagnolo suggest that the less risky cooperation is, the more tolerant and forgiving agents could be in the face of betrayers [16]. But in an incentivized environment, the smaller the value of S, the faster the increase in the incentives. The high risk that one player's choice to betray poses to another player makes them inclined to offer more incentives for avoiding being defected.

This paper only considers changes to the strategies in an ideal theoretical situation, but in general, mutations cannot be



(a) $\alpha = 0.004$ (b) $\alpha = 0.004$ (c) $\alpha = 0.005$ (d) $\alpha = 0.005$ (e) $\alpha = 0.006$ (f) $\alpha = 0.006$

Fig. 2. Three scenarios of earning behaviors and vector fields in a symmetric IPD with paid incentives. The scenarios correspond to the three different cases $\frac{\alpha}{1-\gamma}(R-S)-1(<),(=),(>)0$ following the condition for convergence in Theorem 2. We set $\theta_0^1 = \theta_0^2 = 0.5, R = -1, T = 0, S = -3, P = -2, \eta_0^1 = \eta_0^2 = [0,0]^T$, and $\gamma = 0.99$ for all cases.
Case (> 0) with $\alpha = 0.004$: Fig. 2a, does not show convergence to cooperation as predicted by Theorem 2. Instead, the value for cooperation rises with a high incentive for cooperation. However, the vector field in Fig. 2b shows that the incentive to cooperate sinks for a high probability of cooperation, leading to an oscillating dynamic for both values.
Cases $(= 0)$ with $\alpha = 0.005$: Sustained cooperation arises quickly. The incentive to cooperate remains at 2.5 (see Fig. 2c), as the probability of cooperating is already at 1 and the incentive to cooperate is no longer increasing (see Fig. 2d).
Case (> 0) with $\alpha = 0.006$: Sustained cooperation arises quickly. The incentive to cooperate rises to the maximum value $T - S = 3$ (see Fig. 2e). That is due to the incentive to cooperate rising for any value of θ, which can be seen in the vector field of Fig. 2f.

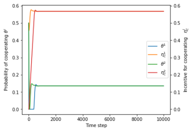


Fig. 3. $\alpha = 0.007, \beta = 0.001, \gamma = 0.9, \theta_0^1 = \theta_0^2 = 0.5, R = -1, T = 0, S = -2.5, P = -2, \eta_0^1 = [0.5, 0]^T, \eta_0^2 = [0,0]^T$. The probability of cooperation is maintained at a fixed value after around 300 time-steps.

neglected because collectively optimal strategies are fragile and unstable [20]. In cases of IPD with incentives, a selfish mutant could have a short-term impact on the process but no consequential change to the final state.

Further Work Referring to the update function of incentives, in some particular where Theorem 2 does not hold, it can be observed that the probability of the agents cooperating is stable at a value other than 1 (Fig. 3), rather than fluctuating all the time. It makes sense to consider the circumstances under which two players can achieve a stable state, except for 100% cooperation, and under what conditions the probability of their cooperation will fluctuate in what range. In addition, this paper sets the initial cooperation probability values and the incentive values to be the same for both players, enabling them to behave consistently throughout the process. If the individual variables of the two players do not correspond equally, the analysis will be more complex and may yield different results. If the maximum value of incentives is restricted differently, the results obtained may also be different from here. If the incentivizing approach is applied to other matrix games, the betrayal incentive may also fluctuate rather than be monotonically decreasing. Moreover, we also make the proof that the probability of cooperation converges towards 1 in finitely many steps in case of $\frac{\alpha}{1-\gamma}(R-S)-1 > 0$, which is not included in this paper due to space limitations.

Acknowledgements We thank the anonymous reviewers whose comments helped to improve this paper.

References

[1] S. Mariani and F. Zambonelli, "Degrees of autonomy in coordinating collectives of self-driving vehicles," in *ISOLA 2020*, T. Margaria and B. Steffen, Eds. Springer, 2020, pp. 189–204.

[2] M. Wirsing, J.-P. Banâtre, M. M. Hölzl, and A. Rauschmayer, Eds., *Software-Intensive Systems and New Computing Paradigms*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 5380.

[3] S. Kernbach, T. Schmickl, and J. Timmis, "Collective adaptive systems: challenges beyond evolvability," *CoRR abs/1108.5643*, 2011.

[4] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-Agent Reinforcement Learning in Sequential Social Dilemmas," in *AAMAS*, 2017, pp. 464–473.

[5] E. Ostrom, "Tragedy of the commons," *The new palgrave dictionary of economics*, vol. 2, 2008.

[6] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, "Open problems in cooperative ai," *arXiv preprint arXiv:2012.08630*, 2020.

[7] P. Dal Bó and G. R. Fréchette, "The evolution of cooperation in infinitely repeated games: Experimental evidence," *American Economic Review*, vol. 101, no. 1, pp. 411–29, 2011.

[8] J. Yang, A. Li, M. Farajtabar, P. Sunehag, E. Hughes, and H. Zha, "Learning to incentivize other learning agents," in *NeurIPS 2020*, 2020.

[9] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel, "A multi-agent reinforcement learning model of common-pool resource appropriation," in *Advances in Neural Information Processing Systems*, 2017, pp. 3643–3652.

[10] A. Lerer and A. Peysakhovich, "Maintaining cooperation in complex social dilemmas using deep reinforcement learning," *arXiv preprint arXiv:1707.01068*, 2017.

[11] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[12] K. Schmid, L. Belzner, T. Gabor, and T. Phan, "Action markets in deep multi-agent reinforcement learning," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 240–249.

[13] A. Lupu and D. Precup, "Gifting in multi-agent reinforcement learning," in *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, 2020, pp. 789–797.

[14] M. S. Miller and K. E. Drexler, "Comparative ecology: A computational perspective," *The Ecology of Computation. North-Holland*, 1988.

[15] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, "Algorithmic game theory, 2007," *Book available for free online*, 2007.

[16] M. Blonski and G. Spagnolo, "Prisoners' other dilemma," *International Journal of Game Theory*, vol. 44, no. 1, pp. 61–81, 2015.

[17] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[18] J. Yang, A. Li, M. Farajtabar, P. Sunehag, E. Hughes, and H. Zha, "Learning to incentivize other learning agents," *arXiv preprint arXiv:2006.06051*, 2020.

[19] L. Belzner, K. Schmid, T. Phan, T. Gabor, and M. Wirsing, "The sharer's dilemma in collective adaptive systems of self-interested agents," in *International Symposium on Leveraging Applications of Formal Methods*. Springer, 2018, pp. 241–256.

[20] F.-X. Dechaume-Moncharmont, "Evolutionarily stable strategies," *Encyclopedia of Animal Cognition and Behavior*, pp. 1–6, 2018.