Cooperative Teaching and Learning of Actions *

Yangbo Song† Mofei Zhao‡

June 6, 2023

Abstract

This paper studies a novel game-theoretic setting: players may acquire new actions over time by observing the opponent's play. We model this scenario as finitely repeated games where players' action sets are private information and may endogenously expand over time. Three main implications emerge from this framework and its equilibria. First, players may target a payoff vector for the long run and voluntarily "teach" one another the actions needed in early periods. The action profile will be learned and sustained as long as each action is available to either player. Second, when no payoff target is prefixed, the players can always obtain or approximate strict ex-post efficiency via bilateral teaching and learning. Third, an alternative economic argument now exists for seemingly irrational cooperative behavior in games with finite horizon. For instance, fully rational players can play a cooperative equilibrium even if the stage game remains a Prisoner's Dilemma for everyone.

Keywords: acquisition of action, finitely repeated games, rational cooperation, strict efficiency

JEL Classification: C72, C73, D83

*We are grateful to Ichiro Obara, Moritz Meyer-ter-Vehn, Yi Chen, Peter Norman, Fei Li, Xi Weng, Jie Zheng and the seminar audience at the Hong Kong University of Science and Technology, the City University of Hong Kong, and Peking University for helpful suggestions. Yangbo Song gratefully acknowledges financial support from Natural Science Foundation of China (NSFC), Project number 72192805. Mofei Zhao gratefully acknowledges financial support from Natural Science Foundation of China (NSFC), Project number 72103016.

†School of Management and Economics, The Chinese University of Hong Kong (Shenzhen), 2001 Longxiang Road, Shenzhen, Guangdong, P.R. China 518172. Email: yangbosong@cuhk.edu.cn.

‡School of Economics and Management, Beihang University, New Main Building A901, 37 Xueyuan Road, Haidian District, Beijing, P.R. China 100191. Email: zhaomf@buaa.edu.cn. Corresponding author.

1 Introduction

The ability to observe and learn new skills and ideas is not only a human instinct, but also an important feature in various strategic scenarios. Consider, for instance, competing cellphone manufacturers who frequently bring forward "next-generation" models. Novel designs and functions that prove to be popular among consumers are often adopted by competitors in no time; examples include dual or triple camera, 3D Touch in user interaction, removal of headphone jack, and even the suffixes Pro, Max and Note in the model names. Similar phenomena prevail in other strategic environments such as financial markets and advertising, where participants may start with different sets of available strategies (on assembling hedge portfolios, creating brand images, etc.) but are capable of acquiring new techniques over time via observation. Conversely, a skilled participant may take this into account when contemplating strategies: they propagate certain techniques hoping for beneficial spillover effect, while conceal other abilities to prevent vicious competition.

The meaning of "learning" in these scenarios differs from that in most existing literature in two aspects. First, incomplete information exists not in payoffs but in availability of actions. For instance, before a chess tournament, an average contestant is typically aware of the existence of some secret tactic which entails her sure loss, but she remains uncertain about whether the opponent is capable of it. Similarly, a trader with no insider information still anticipates serious losses if a competitor practices malicious insider trading. Second and consequently, what a player may learn from her opponent is the means of executing certain actions, such as how to reach a Queen's Gambit or where to pry out information leaks, instead of their existence or the associated payoffs.

In game-theoretic language, the above examples call for a novel alternative setting in defining a game: instead of fixing every player's set of available actions from the beginning, or prescribing an exogenous rule for how it evolves, we may allow endogenous acquisition of available actions which makes the expansion of a player's action set an equilibrium outcome. Multiple incentive issues thus arise in this context. For instance, as emergence and acquisition of new actions are not automatic, it is worth investigating what type of action can be taught and learned in equilibrium, and under what qualification. Moreover, in many applications a player's action set is private information by nature. This uncertainty may promote mutually beneficial cooperation in gameplay when a player is unsure about what severe punishment her opponent is capable of, but may also obstruct cooperation since failure to play some action may result from either deviation or incapability. Finally, in terms of players' welfare, the set of sustainable payoffs requires careful characterization.

In this paper, we model endogenous acquisition of actions as a mechanism to enlarge players' action sets in repeated games with finite horizon. A player learns new actions

when she observes them from her opponent's gameplay. Our results convey a positive message on cooperative "teaching and learning" behavior: on the one hand, as long as an individually rational payoff vector is *feasible*, in the sense that the actions required for achieving it can be played by either player, it can emerge as a long-term *equilibrium* outcome. On the other hand, even if the players do not have a prefixed payoff target, they can always arrive at ex post efficiency in equilibrium. Interestingly, both results hinge on the existence of a destructive action that no player is certainly capable of initially. Apart from providing a framework to analyze the acquisition of new actions, this paper also contributes to the repeated-games literature by suggesting a novel mechanism to sustain cooperation under finite repetition, even with the stage game being a Prisoner's Dilemma, among fully rational agents with consistent payoff functions.

Our model considers two players playing a stage game repeatedly for finitely many periods. The initially available set of actions of each player is private information while the players have some common prior on the distribution of this set. The main digression of our setting from standard repeated games is that a player can learn new actions through observation. For example, suppose that a player is not able to play some action a^1 initially, but saw her opponent playing it in period 1. Then the player will be able to use a^1 starting from period 2. Therefore the term "learning" for a player in this paper should be understood as acquiring actions that she was unable to play before, which differs from most theoretical literature that interprets "learning" as updating information (e.g. Chassang (2010)) or maintaining the current action set (e.g. Joosten et al. (1995)). In this sense, the players are not strictly playing "repeated games", but a sequence of games where the action sets are monotonically expanding. The path of such expansion is endogenous.

To understand the difference made by enabling players to learn new actions, consider the following simple example. Suppose that a dominant action called a^D exists in the stage game but whether either player is able to play it is private information *ex ante*. If learning is infeasible, standard backward induction implies that a player capable of a^D will play it throughout the finitely repeated games, making cooperation impossible. However, if actions can be learned, whether to play a^D before the final period involves opposing incentives when outcome (a^D, a^D) is undesirable (as in a Prisoner's Dilemma). A player capable of a^D may be reluctant to play it, when she believes that the opponent will not, at the cost of leaving the subsequent games with (a^D, a^D) each period; at the same time, she would be less reluctant if she believes that the opponent will play a^D with a significant probability. Thus the uncertain existence of a^D creates a plausible reward-and-punishment scheme for cooperation, namely to have each player refrain from a^D until the last period on path but start playing a^D immediately off path.

Utilizing this observation, our first result, Theorem 1, characterizes the set of sustain-

able long-term payoffs, assuming the possible existence of an action a^D as above[1]. We find a payoff vector can serve as long-term average payoffs in equilibrium as long as it is individually rational and *ex post* feasible. The notion of feasibility becomes broader under acquisition of actions: for instance, suppose that some payoff vector requires player 1 to play a_1 and player 2 to play a_2. It is still feasible even if initially player 2 cannot play either action but player 1 can play both, because 2 could acquire a_2 from 1's gameplay. We construct a three-phase equilibrium for this result. In Phase I, players voluntarily reveal whether they can play the actions needed to achieve the target payoff vector, and thus make its feasibility (or infeasibility) common knowledge. In Phase II, they cooperate on sustaining the target payoffs if feasible. Phase III contains the above mentioned reward-and-punishment scheme: players start playing a^D early, if capable, if and only if prior deviation has been observed.

The downside of an equilibrium with fixed target payoffs is that players fail to cooperate when the target payoffs are actually infeasible. Our next result, Theorem 2, shows that in another type of equilibria, players need not prefix target payoffs but can always achieve ex-post efficiency. In such an equilibrium which also consists of three phases, Phase I now allows time for "trial and error" in order to establish common knowledge of what action profile is efficient among the feasible ones. In particular, the players will first try to achieve the action profile with the highest total payoff by playing out relevant actions, then if it turns out to be infeasible, they turn to the one with the second highest total payoff, and so on. Phases II and III are similar to before. This result also implies that construction of a cooperative equilibrium in our framework is straightforward yet robust: only strategies in Phase I need to be prescribed differently according to the nature of desired cooperation, while the subsequent phases require little alteration.

We have also considered a number of model variations and discussed how they may affect the main results. Interestingly, we find that when learning an action is probabilistic, i.e. observing an action once results in successful acquisition with probability $\mu \in (0, 1)$, an equilibrium may not exist given every μ close to 1 when the repeated games last for a sufficiently long time. This irregularity reveals contradicting incentives on some particular path of possibly off-equilibrium gameplay. Nevertheless, once the length of repetition becomes given while μ is set sufficiently close to 1, existence of equilibrium is restored.

Related literature. Our result proves the sustainability of mutually beneficial "teaching and learning" of actions, as well as subsequent long-term cooperation, under a framework with full rationality and consistent payoff matrices, which stands in stark contrast to the theoretical literature on finitely repeated games. The idea of equilibrium con-

[1]The dominance of a^D is not an essential assumption, but leads to the sharpest contrast with existing literature. See Section 3 for more discussion.

struction has common ground with Benoit and Krishna (1985) in the sense that reward or punishment after every possible history must survive sequential rationality with finite horizon; however, we do not require actual existence of multiple stage equilibria, as excluded by the dominance of action a^D. Another representative method in supporting cooperation in finitely repeated games is to introduce incomplete information on players' rationality, namely with a small probability a player faces a "crazy" opponent who adopts a predetermined strategy (e.g. Kreps et al. (1982), Fudenberg and Maskin (1986), Kreps and Wilson (1982), Milgrom and Roberts (1982)). A more recent paper by Weinstein and Yildiz (2016) also explains cooperation in finitely repeated Prisoner's Dilemma between fully rational players, by proving the existence of a set of stage game payoffs that rationalize the designated behavior of a "crazy" type (referred to as a "commitment" type in the paper). However, their model requires a considerable variation of the stage-game payoffs between different types, which means that the players are sometimes not playing Prisoner's Dilemma as cooperation is not dominated for a "crazy" player. Our theory with acquisition of actions, however, enforces cooperation even when the stage game always remains a Prisoner's Dilemma.

Apart from embedding the acquisition of actions in repeated games, our model can also be viewed as a variation of stochastic games (Dutta (1995), Fudenberg and Yamamoto (2011), Marlats (2015)), where the state is the actual game played and player can alter the state by playing some action the opponent was incapable of. The main difference between our framework and a typical stochastic game is two-fold. First, whether and how the state may transit from one to the other is determined by the players' types, and transition is irreversible. Second, in generic cases the state will be hidden throughout in equilibrium, and it is this incomplete information that enables cooperation even on a finite horizon.

Our theory also identifies an environment where finitely and infinitely repeated games produce similar cooperative equilibria. By adding our constructed play (Phase III) at the end of sufficiently long finitely repeated games, every long-run average payoff vector that is feasible in infinitely repeated games can be approximated in a finitely repeated version. This observation also brings forward the general question of whether an environment with learning of actions may produce new insights for other topics in finitely repeated games, such as monitoring (Mailath et al. (2002), Bhaskar and van Damme (2002), Miyahara and Sekiguchi (2013)) and evolution (Nachbar (1992), Cressman (1996)), and relates to applications of infinitely repeated games, such as market segmentation and collusion (Bos and Marini (2022)).

There is a large literature of behavioral and experimental economics in finitely repeated games, where the term "learning" is often mentioned but with different meanings from ours. In general, learning is usually related to the issue of bounded rationality in these works, and refers to identifying the opponent's possible non-equilibrium behavior

(Andreoni and Miller (1993), Mookherjee and Sopher (1994), Nagel (1995)), exploring superior ways to organize the same set of actions over time (Andreoni (1988), Crawford and Haller (1990), Camerer et al. (2002), Muller et al. (2008)), or changes in preferences induced by payment schemes (Chandrasekhar and Xandri (2023)). The experimental results vary for repeated Prisoner's Dilemma: some find the beginning of defection to occur earlier with experience (e.g. Selten and Stoecker (1986)) while others provide evidence that more experienced players cooperate for longer (e.g. Andreoni and Miller (1993)). Our model relates to the second interpretation above in that learning is essentially a way of expanding a player's payoff space, while we provide a theoretical justification for cooperation in a framework with full rationality.

Organization. The rest of the paper is organized as follows. Section 2 presents an illustrative example of our framework and main results. Section 3 introduces the model. Section 4 presents the results and discusses their robustness to model variations. Section 5 concludes. All proofs are in the appendices.

2 Illustrative Example

Consider the following stage game, played for a total of 4 periods. The whole history of action played is publicly observable.

	a^1	a^2	a^D
a^1	$0, 0$	$2, -0.5$	$-10, 0.1$
a^2	$-0.5, 2$	$1, 1$	$-20, 3$
a^D	$0.1, -10$	$3, -20$	$-9, -9$

Different from the usual repeated games, which action(s) are available to each player is private information. Suppose it is common knowledge that for each player *ex ante*, (1) a^1 is always available, (2) a^2 is available with probability 0.1, and (3) a^D is available with probability 0.6. If the available actions remain fixed regardless of the gameplay, the stage game would be a Prisoner's Dilemma as a^1 dominates a^2 and a^D dominates both a^2 and a^1. Thus there is a unique equilibrium which is independent of the prior beliefs: both players play a^D in every period if they can, and play a^1 otherwise.

Now suppose that a player can learn a previously unavailable action via observation, i.e. if the action is played by the opponent in some period, it becomes available to the player starting from the next period. To demonstrate how this setting works, consider the particular case where, at the beginning of the game, only a^2 and a^D are available to the row player while all actions are available to the column player, as shown in Figure 1. Suppose, for the sake of illustration, that in some equilibrium the row player plays a^2 in both periods 1 and 2, while the column player plays a^1 in period 1 and a^D in period

2. The availability and knowledge on each action then evolves as in Figures 2 and 3. In these figures, an action in black means it is available to the corresponding player and the availability has become common knowledge; an action in blue means it is available to the corresponding player but the availability is still private knowledge; an action in red means it is currently unavailable to the corresponding player. In other words, a^1 becomes available to both players, which also becomes common knowledge, after it is played by the column player in period 1. Meanwhile, as a^D has not been played by either player in period 1, its availability and each player's belief remains unchanged in period 2. Then with a^D played in period 2, it becomes common knowledge in period 3 that all actions are available to each player.

	a^1	a^2	a^D
a^1			
a^2			
a^D			

Figure 1: Period 1

	a^1	a^2	a^D
a^1			
a^2			
a^D			

Figure 2: Period 2, after (a^2, a^1) in period 1

	a^1	a^2	a^D
a^1			
a^2			
a^D			

Figure 3: Period 3, after (a^2, a^D) in period 2

Next, we propose below a particular equilibrium candidate of the 4-period repeated games. We leave it to interested readers to verify, by straightforward calculation, that this strategy profile indeed constitutes an equilibrium.

Period 1. Play a^2 if possible, and a^1 otherwise.

Period 2. If only a^2 or a^1 was played in period 1, play a^2 if a^2 was played and a^1 otherwise. Otherwise, i.e. if a^D was played in period 1, play a^D.

Period 3. If either player played a_2 in period 1, but some player did not play a^2 in period 2, play a^D if possible and a^1 otherwise. Otherwise, play a^1 if a^D has not been played, and a^D if it has been played before.

Period 4. Play a^D if possible. Otherwise play a^1.

In this strategy profile, players do not only refrain from playing the dominant strategy a^D earlier even if they can, but also manage to "teach" one another to cooperate from the beginning. Three economic forces enable it to be an equilibrium candidate. First, each agent is willing to play a^2 in period 1 if she can, because if the opponent can only play a^1, they will lose the opportunity of profitable coorperation in period 2. Second, once cooperation is enabled by some player playing a^2 in period 1, no agent wants to deviate in period 2 in fear of facing a^D as punishment in period 3. Finally, even an a^D-player

will not actually play a^D until the last period, because doing so earlier will trigger the undesirable outcome (a^D, a^D) in the subsequent games.

Note that both players are fully rational, thus the mechanism for sustaining cooperation stands in contrast to renowned prior works such as Kreps et al. (1982) and Fudenberg and Maskin (1986), which induce cooperation by assuming a positive probability of a player being irrational and playing a predetermined strategy, such as Tit-for-Tat. It is also different from Weinstein and Yildiz (2016) in not requiring variation of the game nature for different types of player, as the stage game remains a Prisoner's Dilemma *ex post* regardless of the players' available actions.

The example suggests further and more general implications. On the one hand, a stage game with more possible actions raises the question of how much can be taught. For players to achieve a payoff vector via teaching potential actions to one another, and to sustain cooperation in the subsequent periods, the design of an equilibrium goes beyond the standard Folk Theorem. Since players in our framework can always hide any potential action without fearing off-path punishment, incentivizing them to play designated potential actions can only be achieved through on-path strategy specification. On the other hand, although the initial beliefs need to satisfy certain conditions for desirable teaching to exist in equilibrium, longer repetition relaxes these conditions by increasing the opportunity cost of deviating players. In the subsequent sections, we formalize these implications by investigating a general stage game with arbitrary length of finite repetition.

3 Model

Players and actions. Consider two players 1 and 2, who play a simultaneous stage game $\mathcal{G}$ repeatedly for $T \in \mathbb{N}^+$ periods. Each player has a set of finite possibly available actions, denoted by $\bar{\mathcal{A}} := \{a^1, \cdots, a^n\}$.

Not all actions in $\bar{\mathcal{A}}$ are available to a player from the beginning; instead, what actions a player can choose initially follows a commonly-known probability distribution. For an arbitrary action a, we will use the phrase "player i can/is able to play a" for the event that i's set of available actions contains a. Let $\mathcal{A}_i$, $i = 1, 2$ denote the set of actual available actions for player i, and assume that for some $m < n$, $\forall k \in \{1, 2, ..., m\}$, the probability $\text{Prob}(a^k \in \mathcal{A}_i) := \lambda^k = 1$ for $i = 1, 2$. That is, the two players can both play $a^1, \cdots, a^m$ initially. We call them *endowed* actions and denote their collection by $\mathcal{A}^e$. For every $k \in \{m+1, m+2, ..., n\}$, $\text{Prob}(a^k \in \mathcal{A}_i) := \lambda^k \in [0, 1)$ for $i = 1, 2$. These actions are called a player's *potential* actions, and we denote their collection by $\mathcal{A}^p$. We assume that these probabilities are independent across actions and players.

Learning through observation. We allow mixed or correlated actions and assume that the mixing proportions are observable. We digress from the classical literature in repeated games in assuming that players can acquire new actions from past experience: once player i has observed some action $a \in \bar{\mathcal{A}}$ played by her opponent in period t, either in pure or mixed actions, she can play a if she has not been able to before.[2]

We can thus also use $\mathcal{A}_i$ to define a player's *type*. Note that in the context of learning actions via observation, $\mathcal{A}_i$ may vary according to past gameplay. Let $\Gamma := \{\mathcal{A}_i \subset \bar{\mathcal{A}} : \mathcal{A}_i \supset \mathcal{A}^e\}$ be the set of all possible types.

Payoffs. The players' payoffs in $\mathcal{G}$ is given by $g : \bar{\mathcal{A}} \times \bar{\mathcal{A}} \to \mathbb{R}^2$ where the first and the second arguments in the domain represent player 1's and 2's actions, and those in the range represent 1's and 2's payoffs respectively. A player's total payoff from the repeated games is the sum of her stage game payoffs. We assume no discounting. A player's *average* payoff from the repeated games is then her total payoff divided by T. When T is large, the average payoff in equilibrium measures a player's long-term sustainable payoff.

To highlight the difference made in equilibrium behavior by enabling acquisition of actions, we assume that one of the potential actions, say a^n without loss of generality, is a dominant action in $\bar{\mathcal{A}}$. We denote it in particular as a^D. Let $\mathcal{A}$ denote the set of possibly available actions except a^D. We shall call a player who can play a^D an "a^D-player" for expositional convenience. Clearly, the existence of a dominant action makes it impossible for an a^D-player to cooperate in a standard finitely-repeated-games framework.

Histories and belief updating. Let h_t, $t = 0, 1, 2, \cdots, T-1$, be a history of action played by period t in the repeated games with $h_0 = \varnothing$. Let $\mathcal{H}$ be the set of possible histories. The history of play is public knowledge to both players in every period. Hence, a player's strategy $s_i(\cdot, \cdot)$ is a mapping from $\mathcal{H} \times \Gamma$ to $\Delta(\bar{\mathcal{A}})$.

Define $\tilde{\lambda}_i^k : \mathcal{H} \to [0, 1]$ to be player i's belief updating rule: given history h_{t-1}, $\tilde{\lambda}_i^k$ gives i's believed probability that her opponent can play a^k (this formulation includes $k = D$, with a slight abuse of notation). Clearly, the belief updating process satisfies initial conditions $\tilde{\lambda}_i^k(\varnothing) = \lambda^k \ \forall k$. We assume that players follow the Bayes' rule in updating information.

Equilibrium. We use the perfect Bayesian equilibrium (PBE, or simply equilibrium, in the subsequent text) as the solution concept, and denote a PBE as $\sigma^* = \{s_i^*, \tilde{\lambda}_i^*\}_{i=1,2}$, where $\tilde{\lambda}_i^* := \{\tilde{\lambda}_i^{k*}\}_{k=1}^n$ is a series of belief updating rule.

It is clear that in every PBE, a player whose available actions include a^D by the last period will always play a^D in the last period. Also, if a^D is played once in some period, the

[2] In this setting, a player may learn multiple actions in a single period from the opponent's mixed action. We assume this for simplicity of exposition; our main results are unchanged under the alternative configuration where only one action per player can realize in every period. See Section 4.3 for a discussion.

availability of a^D to both players becomes common knowledge, and backward induction yields a unique equilibrium with both players playing a^D thereafter. We summarize these preliminary results as follows.

Proposition 1. *In every PBE σ^*:*

1. *$s_i^*(h_{T-1}, \mathcal{A}_i) = \delta(a^D)$ for $i = 1, 2$ and for all h_{T-1} and $\mathcal{A}_i$ such that $a^D \in \mathcal{A}_i$, where $\delta(\cdot)$ is the Dirac delta function.*

2. *$s_i^*(h_t, \mathcal{A}_i) = \delta(a^D)$ for $i = 1, 2$ and for all h_t such that a^D has been played for at least once.*

Specifications on parameters. Now we move on to specify some notations in the stage game payoffs and actions which will prove convenient in the subsequent analysis. We list them below. For conciseness of notations and without loss of generality, we assume that g is symmetric, i.e. $g_1(a, a') = g_2(a', a)$ for all $a, a' \in \bar{\mathcal{A}}$.

1. $g_{max} := \max_{a,a' \in \bar{\mathcal{A}}} g_1(a, a')$. This is the maximum payoff that a player can get from playing the stage game once.

2. $g_{br,a^D} := \max_{a \in \mathcal{A}} g_1(a, a^D)$. This is the maximum payoff that a player not able to play a^D can get from playing against another player using a^D. The subscript *br* stands for best response (without a^D).

3. $g_{min} := \min_{a,a' \in \mathcal{A}} g_1(a, a')$. This is the minimum payoff that a player not able to play a^D can get from playing against another such player in the stage game once.

4. Consider a stage game where $\lambda^k = 0 \ \forall k = m+1, \cdots, n-1$ while $\lambda^D \in [0, 1)$. That is, it is common knowledge that no player can play a potential action except a^D. By the Oddness Theorem, there exists a symmetric, possibly mixed-strategy Bayesian Nash equilibrium. We denote the corresponding action played by a player who cannot play a^D as $a(\lambda^D) \in \Delta(\mathcal{A}^e)$. We assume that for all $\lambda^D \in [0, 1)$, this BNE is also a BNE when $\lambda^k = 1 \ \forall k = m+1, \cdots, n-1$, i.e. when all potential actions except a^D are available to both players.[3] Without loss of generality, we normalize the payoffs such that $g_1(a(0), a(0)) = 0$.

[3]This assumption avoids certain technical details in equilibrium construction with little economics. We discuss our results' robustness without it in Section 4.3. As $\lambda^k < 1$ by the definition of potential actions, the case where $\lambda^k = 1 \ \forall k = m+1, \cdots, n-1$ cannot occur *a priori* and is only a hypothetical one for illustrating the assumption. Alternatively, one may think of this case occurring *ex interim* following a history where every a^k, $k = m+1, \cdots, n-1$, has been played at least once.

On important modeling choices. We have embedded two critical assumptions in the framework. First, we model the strategic environment as finitely repeated games and acquisition of actions as expanding the action set. In most applications such as sports games, the ability to acquire action is usually only important when newly acquired actions can be used repeatedly under familiar circumstances in the future. Hence repeated games provide a time-consistent environment for describing action acquisition and evaluate its consequences. Second, we assume that when a player is not able to play some action, she still knows its existence and relevant payoffs. This setting does exclude some interesting cases with possible total unawareness of certain actions, but keeps our analysis tractable and still encompasses important and realistic scenarios. For instance, an inexperienced fund manager may well know the existence of "rat trading" and its impact on the market, but has to actually observe or hear from an old hand to be able to implement the malpractice. As will be clear in the next section, our result readily extends to the scenario where the players may not know the exact payoffs following unavailable actions, but they do understand certain important relations between key payoff parameters.[4]

4 Analysis

In this section, we characterize and discuss long-term sustainable payoffs in the context of teaching and learning of actions. We begin with determining a range of such equilibrium payoffs.

4.1 Sustainable Payoffs under Strategic Teaching

Allowing for mixed actions, let V denote the convex hull of the set $\{g(a, a') : a, a' \in \mathcal{A}\}$, which is the set of feasible payoff vectors attainable in the stage game *without* playing the dominant action a^D. Notably, our definition of V is different from the usual definition of the feasible payoff set, which encompasses payoff vectors from all possible action profiles. This difference stems from the particular role a^D plays in our model. In the repeated games we consider, as soon as a^D is played by some player at some period t, it becomes available to both players and the mutual availability also becomes common knowledge; as a result, the consecutive repeated games must unravel in equilibrium as stage play becomes absorbed in (a^D, a^D) from t onward. Consequently, for every $a \in \mathcal{A}$, action profile (a, a^D) may appear for at most one period on any equilibrium path. Therefore to discuss long-term sustainable payoffs in the repeated games, in the sense that such payoffs may be supported by equilibrium actions played for many periods, it is natural to focus on feasible payoffs that does not require the play of a^D.

[4]See Assumption 1 and related discussions in Section 4.1 for details.

Let $V^*_{Nash} := V \cap (\mathbb{R}^+ \times \mathbb{R}^+)$. This set features a plausible lower bound for long-term average equilibrium payoffs, which is the stage-game BNE payoff between two players who cannot play a^D.[5] For every payoff vector $(v_1, v_2) \in V^*_{Nash}$, we say that it is *feasible* if it may realize from some action profile in $\Delta(\mathcal{A}_1 \cup \mathcal{A}_2 \times \mathcal{A}_1 \cup \mathcal{A}_2)$. This definition has a natural meaning: (v_1, v_2) must result from some action profile, and the minimal requirement for (v_1, v_2) to realize in equilibrium is each needed action can be played by either player initially, so that the players can feasibly teach one another. Note that we describe achievability on an ex-post basis, i.e. whether (v_1, v_2) is feasible is not initially common knowledge if it involves potential actions.

Our first result shows that if a payoff vector dominating the Nash equilibrium payoffs is feasible, it can be approximated in the long-run if the payoffs from a^D against a^D, and from the best response in $\mathcal{A}$ against a^D, are both undesirable. This qualification is formally described by the following condition, which we assume throughout this section.

Assumption 1. *The following inequalities are satisfied: (a)* $g_1(a^D, a^D) < \min\{g_{min} - \frac{g_{max}}{1-\lambda^D}, -\frac{g_{max}}{\lambda^D}\}$*; (b)* $g_{br,a^D} < \frac{1}{\lambda^D}(g_1(a^D, a^D) - (1-\lambda^D)(2g_{max} - g_{min}))$.

The formal proofs in Appendix A provide a detailed account of how the above two conditions work in the analysis. Heuristically, condition (a) means that the payoff from the stage-game dominant-strategy equilibrium (a^D, a^D) is sufficiently low, while condition (b) means that the best payoff of a player who cannot play a^D facing an opponent who can is even sufficiently lower. Together they imply that a^D, although a dominant action, is rather destructive to both players once played out: it ensures an undesirable equilibrium in the remaining periods, and does even greater harm instantly to whichever player not playing it.

Now we are ready to present the first theorem.

Theorem 1. *For* $i = 1, 2$*, every* $\epsilon > 0$ *and every* $(v_1, v_2) \in V^*_{Nash}$*, there exists an equilibrium such that, if* (v_1, v_2) *is feasible, then player* i*'s average payoff is within* ϵ *of* v_i *when* T *is sufficiently large.*

Equilibrium construction. We construct a strategy profile as candidate for a cooperative equilibrium where agents seek to achieve and sustain (v_1, v_2). We relegate the detailed description of strategies to Appendix A, while highlight their key features here and discuss how they facilitate teaching and cooperation. The equilibrium strategy profile has three phases, and we lay out below the on-path behavior and off-path punishment in turn.

[5] As indicated by the classical Folk Theorem, using BNE payoff as a lower bound is technically more convenient, but the bound can in fact be further reduced; in Appendix B, we present a parallel result that shows any payoff vector dominating the stage-game minimax payoff can also be sustained in equilibrium.

Phase I: teaching and learning. The first phase contains periods 1 and 2. In period 1, Player 1 tries her best to teach Player 2 the actions needed for achieving (v_1, v_2). Specifically, if she already can play all such actions, she plays an equal mixture of them;[6] otherwise, she plays as many of them as possible in another equal mixture. In period 2, Player 2 considers her potentially enlarged set of available actions, and if she can, plays an equal mixture of them that is sufficient for achieving (v_1, v_2). This mixture signals to Player 1 that profitable cooperation has become possible.

The main incentive issue here is that if achieving (v_1, v_2) requires some potential action that is not favorable in the stage game, a player may want to hide it and she can do so without facing off-path punishment. Thus we differentiate between subsequent on-path possibilities to incentivize teaching of such actions.

Phase II: cooperation. The meaning of cooperation is path-dependent. If either player has made it publicly in period 1 or 2, by playing a previously described mixture of actions, that (v_1, v_2) is feasible, they cooperate and obtain (v_1, v_2) every period in this phase. For instance, suppose that $g_1(a_1, a_2) = v_1$ while $g_2(a_1, a_2) = v_2$, and Player 1 played an equal mixture of a_1 and a_2 in period 1; then Player 1 will always play a_1 and Player 2 a_2 in this phase. Otherwise, if the achievability of (v_1, v_2) has not become common knowledge, the players coordinate on playing $(a(0), a(0))$, a stage-game Nash equilibrium among actions excluding a^D.

The above strategy resolves the incentive problem in Phase I when Phase II is sufficiently long. Consider Player 1 as an illustration. If she can play actions that achieve (v_1, v_2), say a_1 and a_2, but chooses not to play them out in Phase I, she faces the risk of her opponent not able to play them both initially, which will lead to the "worse cooperation" of $(a(0), a(0))$ (with payoff $(0, 0)$) instead of the "better cooperation" with payoff (v_1, v_2) during Phase II. When Phase II is sufficiently long, the opportunity cost stemming from the risk will eventually outweigh any one-period loss from playing the mixture of a_1 and a_2. However, as can be expected in all finitely repeated games, cooperation cannot last to the final period, and some off-path punishment is required for cooperation not to fail midway. These concerns call for a third and final phase.

Phase III: possible punishment with uncertainty. This phase contains a fixed number (at least 2) of periods up to period T. If the players cooperated successfully, in either form specified in Phase II, they will play $(a(0), a(0))$ up to period $T - 1$. In period T, each player play a^D if she can, and $a(\lambda^D)$ otherwise. However, if any detectable deviation was spotted in the previous phases, e.g. Player 1 played a mixture of a_1 and a_2 in period 1 but the players failed to play (a_1, a_2) in period 3, then they turn to punishment in every period of Phase III: play a^D if they can and $a(\lambda^D)$ otherwise.

The purpose of designating such an off-path strategy is to maintain player's incentives

[6]The mixture does not have to be equal, but must be commonly agreed upon in equilibrium. See Appendix A for more details.

to cooperate in Phase II. Note that once a^D is played in any period, the subsequent game unravels with the unique equilibrium (a^D, a^D) always. This is detrimental for both players because $g_1(a^D, a^D)$ is sufficiently small. Since each player has a positive probability *ex ante* to be able to play a^D, the players are willing to cooperate in Phase II to avoid the risk of triggering a^D prematurely. However when deviation has occurred, an a^D-player will play a^D instantly again because of uncertainty about a^D's availability to the opponent: better to play a^D now and trigger the dominant strategy equilibrium, than to refrain and possibly have payoff g_{br,a^D} which is sufficiently lower than $g_1(a^D, a^D)$ by assumption. Hence the punishment strategy is also incentive compatible by itself.

It may also be useful to mention a couple of other possible off-path behavior here. For instance, it is straightforward that no player will play a^D earlier if she will already refrain from it in Phase III. Also, when a player deviates from achieving (v_1, v_2) in Phase II, our strategy profile prescribes that they play $(a(0), a(0))$ for the rest of Phase II, so the timing of deviation is irrelevant as long as Phase III provides sufficient punishment. See Appendix A for a complete strategy specification and case-by-case discussion of incentives.

Remarks on Assumption 1. It is worth noting that for Assumption 1 to hold, λ^D must be bounded away from zero given the other payoff parameters. This stands in contrast to the usual assumption in reputation models, where the probability that a player is of commitment type may be arbitrarily close to zero if the game is repeated for an arbitrarily long time. Our cooperative equilibrium is built on a significantly positive λ^D because, unlike reputation models, punishment of deviation requires the play of a^D, after which the repeated games enter the dominant-strategy equilibrium with unfavorable payoffs. Should λ^D be close to zero, a player would lack incentives to punish deviations with a^D and unilaterally induce the undesired equilibrium in subsequent periods, as she would be faced with only a slight chance of a^D from the opponent.

We also briefly discuss here how realistic Assumption 1 is and what happens to equilibrium behavior if it is violated. First note that the existence of some dominant but detrimental behavior, i.e. a^D in our framework, is prevalent in many economic scenarios. In a prisoner's dilemma, it may refer to bribing the judge; in financial markets, insider trading; in industrial competitions, spying or malicious acquisition if feasible (note that a^D is different from cutthroat competition, since the latter is not likely to be a dominant action of the stage game, and most of the firms know how to conduct cutthroat competition); in global trade, political and/or military intervention; and so on.

When $g_1(a^D, a^D)$ is not sufficiently low, i.e. the dominant-action equilibrium (a^D, a^D) is not very undesirable, an a^D-player may find it worthwhile to play a^D earlier on equilibrium path, thus terminating cooperation. Conversely, when g_{br,a^D} is not sufficiently lower than $g_1(a^D, a^D)$, i.e. playing some non-a^D action against a^D is not much worse than playing a^D, an a^D-player may decide not to carry out the punishment off equilibrium

path, in the hope that the opponent may not be able to play a^D and the subsequent games need not unravel into (a^D, a^D). Therefore the two conditions in Assumption 1 properly bound an a^D player's incentives to play a^D, providing discouragement on path while encouragement off path.

There clearly exist parameter values violating Assumption 1. Respectively in applications of our framework, cooperation does not necessarily take place in every strategic scenario. Consider for instance an online game where all players know that bugs (or similarly, some uninteresting super-early-game rush) exist, but only a fraction of them know how to carry it out to steal victory. Using bugs means bad experience for every player, especially to those who could not exploit such bugs themselves; it means that consistent with our setting, $g_1(a^D, a^D)$ is low and g_{br,a^D} even lower. Now suppose that the two players join a multi-game series. Only when the bug is very detrimental (i.e. $g_1(a^D, a^D)$ is sufficiently low) will a player hesitate to use it early (because that teaches the opponent how to use the bug). Otherwise, it is natural that they take the advantage right away, which is actually not rare in real gameplay. Similarly, if g_{br,a^D} is not sufficiently low, then there is no sufficient threat against other deviations in the early rounds. It is also commonly seen in games that an expert deliberately surprises their opponent and wait for the countermeasure, because they are willing to bear one-period loss (g_{br,a^D}) for discovering whether the opponent has any secret tactic.

4.2 Ex-Post Efficiency

Theorem 1 has established cooperation for any feasible (v_1, v_2) that dominates the Nash equilibrium payoffs, but also exposes the players to the risk of living with the Nash equilibrium payoffs once (v_1, v_2) is not feasible ex post. In the next section, we introduce a result that enables the players to land a payoff vector that is uncertain *ex ante* but always efficient *ex post*.

Fix the players' initial sets of available actions, $\mathcal{A}_1$ and $\mathcal{A}_2$. A feasible payoff vector in $\Delta(\mathcal{A}_1 \cup \mathcal{A}_2 \times \mathcal{A}_1 \cup \mathcal{A}_2)$ is *ex post* efficient if it is not Pareto dominated by any other feasible payoff vector. In this section we focus on a particular such payoff vector which is also fair in distribution[7]: given $\mathcal{A}_1, \mathcal{A}_2$, let $v(\mathcal{A}_1 \cup \mathcal{A}_2)$ denote half of the maximized sum of payoffs from action space $\Delta(\mathcal{A}_1 \cup \mathcal{A}_2 \times \mathcal{A}_1 \cup \mathcal{A}_2)$. That is, we aim to sustain in equilibrium not only Pareto efficiency but also strict efficiency, in the sense of maximizing (and equally distributing) the players' total payoff.

The following result asserts that, although the players do not know *ex ante*, i.e. when they cannot directly observe the opponent's initially available actions, the largest set of available actions $\mathcal{A}_1 \cup \mathcal{A}_2$, they are able to approximate the arguably best possible

[7]We pick this particular payoff vector simply for selecting a concrete benchmark of efficiency. As can be seen from the proof, the distribution of payoffs, as long as it guarantees every player a larger payoff than from $(a(0), a(0))$, does not affect the results.

long-term payoffs $v(\mathcal{A}_1 \cup \mathcal{A}_2)$ in equilibrium.

Theorem 2. *For $i = 1, 2$ and every $\epsilon > 0$, when T is sufficiently large there exists an equilibrium where player i's average payoff is always within ϵ of $v(\mathcal{A}_1 \cup \mathcal{A}_2)$.*

To construct an equilibrium for Theorem 2, the key is to motivate players to honestly teach one another useful actions for the highest possible symmetric payoff, which by definition cannot be predetermined. For such cooperation to exist and persist in equilibrium, it is first necessary that no *ex post* efficient payoff vector requires any player to play a^D, after which the repeated games would simply unravel. This is guaranteed by Assumption 1, which implies that a player's highest possible payoff against a^D is still so low that the sum of payoffs can never be efficient. (See Appendix A.2 for a detailed argument.)

We thus propose a multi-stage teaching and learning phase from the beginning of the repeated games (as opposed to a one-stage, 2-period phase in the equilibrium for Theorem 1). Without loss of generality, suppose that *ex ante* there are L possible maximized total payoffs, $v^1, \cdots, v^L$, each corresponding to one or more pure action profiles in one or more possible configurations of $\mathcal{A}_1 \cup \mathcal{A}_2$. Assume that $v^1 > v^2 > \cdots > v^L > 0$. In our constructed equilibrium, the players will establish common knowledge about *ex post* efficiency by trial and error.

In period 1, Player 1 examines her set of available actions and see if she can play all the actions in some action profile that achieves $\frac{v^1}{2}$ for each player. If yes, she picks an action profile such that an equal mixture of the involved actions brings her the best payoff[8], and plays the mixture. The players can then start cooperating from period 2. Otherwise, Player 1 plays an equal mixture of as many actions in the union of such action profiles as she can. In this case, Player 2 in period 2 conducts an analogous evaluation on her now possibly enlarged set of available actions. That is, if achieving $(\frac{v^1}{2}, \frac{v^1}{2})$ is now feasible, she "announces" this to Player 1 by playing her best mixture of actions that can constitute an action profile to achieve $(\frac{v^1}{2}, \frac{v^1}{2})$, and the players start cooperating from period 3. If not, Player 2 will play another mixture of actions but aiming at achieving $\frac{v^2}{2}$ for each player. The process goes on by induction for at most L periods, by which time the players will have common knowledge of which v^l, $l = 1, \cdots, L$, to cooperate for. They then go through a sufficiently long cooperation phase and an ending phase similar to Phases II and III in Theorem 1, with similar off-path punishment.

The implication of Theorem 2 is two-fold. First, it proposes a different type of cooperation with learning of actions: although players may not accomplish a prefixed goal due to uncertainty in action availability, they can always cooperate on the best they can. While they do not necessarily agree initially, in terms of common knowledge, on the long-term payoff vector to be sustained, they do agree on each player's share $(\frac{1}{2})$ of

[8]This best payoff is calculated as against a pre-determined, always available action of Player 2. See Appendix A.2 for technical details.

the ultimate sum of long-term payoffs. This grants them incentives to make the pie as large as possible by playing out actions that help improve the payoff sum during the first phase of repeated gameplay, which only takes up an arbitrarily small fraction of time as the length of repetition extends towards infinity. Hence, given the subsequent strategy in the next phase that both players will coordinate on the total-payoff-maximizing action profile among those played so far, each player is self-motivated to teach the other as many useful action as possible, without even requiring a punishment scheme.

Second, we have identified an endgame device – Phase III with credible off-path threat to play a^D – in equilibrium design, as a robust method to prevent the finite repeated game from cooperation failure. The nature of finite repetition implies necessity of punishment on off-path behavior near the endgame, and our equilibrium construction for Theorems 1 and 2 shows that an identical punishment scheme can be applied to equilibria with different early-phase gameplay as well as different targeted long-term payoffs. One may consider an arbitrary target for cooperation, for instance the second highest possible payoff from $\Delta(\mathcal{A}_1 \cup \mathcal{A}_2 \times \mathcal{A}_1 \cup \mathcal{A}_2)$, to be sustained in some equilibrium, and now only needs to specify the first phase of gameplay where the target action profile emerges. Our construction of Phases II and III will resolve all other incentive issues.

The role of teaching and learning. To summarize, teaching and learning of actions have enabled different equilibrium behavior, as compared to standard finitely repeated games without such a feature, in two aspects. On the one hand, cooperation can be sustained in finitely repeated games in environments with non-common knowledge of the players' action sets, as long as a rather detrimental dominant action exists and can be learned through gameplay. This action is a dormant threat that will be played voluntarily and persistently once "activated" in any period by any player, and hence is a credible deterrence although never played in equilibrium before the endgame. Our theory thus offers an alternative explanation of cooperative behavior observed in practical scenarios where teaching and learning are clearly feasible. On the other hand, teaching and learning play an active role early in the repeated games in expanding the players' action set, as long as doing so changes the long-term payoff vector in each player's favor. Without common prior of the overall available actions, the players can identify the boundary for strict efficiency via gameplay and coordinate on staying on the boundary. Their economic incentives are guaranteed by the long cooperation phase thereafter, which as mentioned before hinges on effective and self-motivating punishment scheme featuring the dominant action a^D.

4.3 Discussion

We discuss here some further possible model variations and how they may or may not affect our results.

Imperfect learning. The first obvious case to examine is for learning of an action not to be certain and immediate upon the first occurrence of that action. Specifically, instead of learning with certainty after observing a previously unavailable action once, a player now learns the action with probability $\mu \in (0, 1)$ upon each observation, i.e. with probability $1 - \mu$ the player's action set remains unchanged. $\mu = 1$ coincides with our previous setting. In terms of learning, we do not distinguish between observation through a pure action or a mixed one, i.e. when a previously unavailable action is included in a mixed action, the probability of learning it via observation is also μ. Also for simplicity but without loss of generality, we assume here that the only potential action is a^D, i.e. $\mathcal{A}$ is a set of endowed actions to both players.

In Appendix B, we prove that when μ is close to 1, there exists no equilibrium when T gets sufficiently large. This nonexistence result stands in contrast to the general existence of equilibria in regular finite games with incomplete information. The main underlying reason is that, when a new action can be learned with high probability once it occurs even in a mixed action profile, a player's expected payoff is no longer continuous in her proportion of mixing actions, which renders the usual fixed-point theorems inapplicable. However, once we reverse the parametric setting, fixing T and let μ approach 1, existence of equilibria – in particular, existence of cooperative equilibria as in previous results – is restored.

Multi-period learning. We can relax the one-period perfect learning in the standard model in a different way from above: suppose that it takes observation of some action for $y > 1$ periods to learn it. As long as y is fixed and finite, this alternative setting makes little difference in the main results except for slight changes below in equilibrium design.

First, a longer Phase I is needed now for learning to occur. Second, Phase III lasts for at least $y + 1$ periods, and more specification is in place for an a^D-player in this phase. On path, an a^D player only plays a^D for the last y periods. Off path, she starts playing a^D from the beginning of Phase III: if the opponent does not respond by a^D, implying the opponent to be a non-a^D-player, she only plays a^D for the first $y-1$ periods of Phase III and save the last action a^D to period T; otherwise, she always plays a^D afterwards.

Non-observation of mixture. Another realistic modification of the model is to take out the assumption that the probabilities in mixed actions are observable and players can learn every action by seeing one mixture. Alternatively, assume that only one pure action will realize for each player per period. In this case, we can apply the idea of

"block" strategy in classical repeated games to achieve desired patterns of cooperation. For instance, for a player who is supposed to play an equal mixture of actions a^1 and a^2 in our standard model, she now plays a^1 and a^2 each for half of the time in a K-period block. As long as K can be sufficiently large, which is guaranteed by a sufficiently large T, we arrive at the same approximation of equilibrium payoffs as before.

Infinite repetition. Our results readily extend to repeated games with infinite horizon. In fact, the analysis with infinite repetition becomes much easier: there is even no need to assume a potential dominant action a^D, as the endgame effect disappears; a standard Folk Theorem thus follows. Mutually beneficial teaching and learning can occur in equilibrium, with the resulting payoffs sustained, as long as the payoffs dominate Nash equilibrium payoffs (or minimax payoffs) and players are sufficiently patient.

Stage-game BNE in potential actions. Suppose that when some potential action becomes available to both players, the symmetric action profile $a(\lambda^D)$ is no longer a BNE of the stage game for some λ^D. An immediate corollary under this configuration is that our constructed equilibrium still enables players to achieve and sustain feasible (v_1, v_2), as long as (v_1, v_2) dominates every possible stage-game BNE payoff vector. The only change to equilibrium strategy here is a possibly different stage-game BNE action profile for punishment, when a deviation has occurred but both player's ability to play a^D has not yet become common knowledge. When (v_1, v_2) is dominated by some stage-game BNE payoff vector, more technical specification is required for equilibrium design, but we choose not to include it in the current paper as it adds little new economics to the model.

5 Conclusion

In this paper, we have studied acquisition of actions in the context of finitely repeated games. Apart from a novel economic force, such acquisition makes a significant difference in theoretical prediction: mutually beneficial cooperation can be enforced on a finite horizon even when agents are fully rational and payoffs are consistent across different types. Moreover, the players need not both be capable of the actions needed for cooperation initially, but will voluntarily teach one another over time.

Our model provides one among many possible configurations for a theory of evolving game structure via players' observation of one another's gameplay. Broadly speaking, observing opponents' actions may change the underlying game in various other ways, including forming more accurate beliefs about its payoff, reducing possible associated cost, revealing potential subsequent games, etc. We believe that further investigation

from these perspectives leads to a rich set of potential research directions in game theory and related fields such as industrial organization.

A Proofs of Main Results

We first introduce some notations that will be useful for all subsequent proofs.

Let $\mathfrak{A}(v_1, v_2)$ denote the set of smallest sets of pure actions that can achieve (v_1, v_2). To be precise, for every $A \in \mathfrak{A}(v_1, v_2)$, $A \subset \mathcal{A}$, we have (1) (v_1, v_2) is feasible from $\Delta(A \times A)$, and (2) there exists no $A' \subsetneq A$ such that (v_1, v_2) is feasible from $\Delta(A' \times A')$. Let $\bar{A} = \cup_{A \in \mathfrak{A}(v_1, v_2)} A$ be the largest set of pure actions that can possibly be useful in achieving (v_1, v_2). Without loss of generality, we focus on the case where $\bar{A} \cap \mathcal{A}^e = \varnothing$. We use $a_{ep}(A)$ to denote the mixed action that places equal probability on each action in A, and use $a_v(A)$ to denote an action profile from $\Delta(A \times A)$ that achieves (v_1, v_2). Let $A_{i,t}$ denote the set of pure actions that have been played by player i before period t; and let $\lambda_{i,t}^k$ denote player i's believed probability, at period t, that her opponent can play a^k.

A.1 Proof of Theorem 1

The proof consists of the following two parts.

Part I: We propose the following strategy profile and associated belief updating rule, denoted as $\sigma_{Nash}^{(v_1,v_2)}$, where player i's average payoff is within ϵ of v_i when T is sufficiently large.

1. In period 1: if $2^{\mathcal{A}_1} \cap \mathfrak{A}(v_1, v_2) \neq \varnothing$, player 1 finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A), A \in 2^{\mathcal{A}_1} \cap \mathfrak{A}(v_1, v_2)$, and plays $a_{ep}(A^*)$; if $2^{\mathcal{A}_1} \cap \mathfrak{A}(v_1, v_2) = \varnothing$, player 1 plays $a_{ep}(\bar{A} \cap \mathcal{A}_1)$ if $\bar{A} \cap \mathcal{A}_1 \neq \varnothing$, and $a(0)$ otherwise. Player 2 plays $a(0)$.

2. In period $t > 1$, if either player has played a^D before, both players play a^D. Otherwise, follow the strategy described in 3-6.

3. In period 2:

   - a. If there was no publicly identified deviation (i.e. a mixture of actions such that its deviation from $\sigma_{Nash}^{(v_1,v_2)}$ is common knowledge between the players; same hereinafter) from 1, then (1) if player 1 played $a_{ep}(A)$ for some $A \in \mathfrak{A}(v_1, v_2)$ in period 1, the players play $a_v(A)$; (2) otherwise, then (2a) if $2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}(v_1, v_2) \neq \varnothing$, player 2 finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A), A \in 2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}(v_1, v_2)$, and plays $a_{ep}(A^*)$; (2b) player 2 plays $a(0)$ otherwise. Player 1 plays $a(0)$.
   - b. If there was a deviation from 1 which can be publicly identified (player 1 plays any other action than $a(0)$ and $a_{ep}(A')$, $A' \subset A$ for some $A \in \mathfrak{A}(v_1, v_2)$, or player 2 plays any other action than $a(0)$), the players play $(a(0), a(0))$.

4. Let $t^* \in \mathbb{N}^+$ be such that $T - t^* \geq 1$ and is a constant. In period $t \in \{3, \cdots, t^* - 1\}$:

- a. If there was no deviation from $1-3$, then (1) if player 2 played $a_{ep}(A)$ for some $A \in \mathfrak{A}(v_1, v_2)$ in period 2 and the players have played only $a_v(A)$ in periods $3, \cdots, t-1$, the players play $a_v(A)$; (2) otherwise, the players play $(a(0), a(0))$.
- b. If there was a deviation from $1-3$ which can be publicly identified, the players play $(a(0), a(0))$.

5. In period $t \in \{t^*, \cdots, T-1\}$: if there was no publicly identified deviation from $1-4$, and both players have played only $a_v(A)$ in periods $1, \cdots, t^*-1$ and $(a(0), a(0))$ in periods $t^*, \cdots, t-1$, then play $a(0)$. Otherwise, in period $t \in \{t^*, \cdots, T-1\}$: if $(\lambda^D_{i,t}, \lambda^D_{j,t}) = (\lambda^D, \lambda^D)$, play a^D if i is an a^D-player and $a(\lambda^D)$ if she is not. If $(\lambda^D_{i,t}, \lambda^D_{j,t}) = (0,0)$, play $a(0)$ regardless of i's type.

6. In period T: if i is an a^D-player, she plays a^D. If she is not an a^D-player, she plays $a(\lambda^D)$ if $(\lambda^D_{i,T}, \lambda^D_{j,T}) = (\lambda^D, \lambda^D)$, and play $a(0)$ if $(\lambda^D_{i,T}, \lambda^D_{j,T}) = (0,0)$.

7. The belief updating rule $\tilde{\lambda}^k_{i,t}$ for $k \geq m+1, t > 1, i = 1, 2$, is as follows:

- a. If either player has played a^k (with positive probability, same hereinafter) before, $\tilde{\lambda}^k_{1,t} = \tilde{\lambda}^k_{2,t} = 1$.
- b. Otherwise, if a^k should have been played by player $j \neq i$ with positive probability according to $1-5$ above, $\tilde{\lambda}^k_{i,t} = 0$.
- c. Otherwise, $\tilde{\lambda}^k_{i,t} = \lambda^k$.

To see that $\sigma^{(v_1,v_2)}_{Nash}$ is well-defined, it is necessary and sufficient to show that according to $\sigma^{(v_1,v_2)}_{Nash}$, $(\tilde{\lambda}^D_{i,t}, \tilde{\lambda}^D_{j,t})$ can only be (λ^D, λ^D), $(1,1)$ or $(0,0)$. Clearly, the only way to make $\tilde{\lambda}^D_{i,t}$ or $\tilde{\lambda}^D_{j,t}$ equal to 1 in $\sigma^{(a_1,a_2)}_{Nash}$ is for some player to play a^D with positive probability, which immediately enables the other player to learn and play a^D. Hence, once $\tilde{\lambda}^D_{i,t}$ or $\tilde{\lambda}^D_{j,t}$ becomes 1, the other player's belief must be 1 as well.

Next, notice that the only way to make $\tilde{\lambda}^D_{i,t}$ or $\tilde{\lambda}^D_{j,t}$ equal to 0 in $\sigma^{(v_1,v_2)}_{Nash}$ is to have someone play an action other than a^D in period $t \in \{t^*, \cdots, T-1\}$, following a history that digresses from $\sigma^{(v_1,v_2)}_{Nash}$ in which no player has played a^D. Consider $t \in \{t^*+1, \cdots, T\}$ such that $\tilde{\lambda}^D_{i,t-1} = \lambda$ while $\tilde{\lambda}^D_{i,t} = 0$. The assumption that $\tilde{\lambda}^D_{i,t} = 0$ implies that neither player has played a^D in period $t-1 \in \{t^*, \cdots, T-1\}$, which means that $\tilde{\lambda}^D_{j,t}$ must be 0. Hence, once $\tilde{\lambda}^D_{i,t}$ or $\tilde{\lambda}^D_{j,t}$ becomes 0, the other player's belief must be 0 as well. It then follows that if $\tilde{\lambda}^D_{i,t}$ or $\tilde{\lambda}^D_{j,t}$ equals λ, the other player's belief must also be λ.

Intuitively, what is required of the belief updating rule to make $\sigma^{(v_1,v_2)}_{Nash}$ an equilibrium is that an a^D-player is much more likely to play a^D after deviation occurs. As a matter

of fact, there are alternative beliefs that suffice for supporting our construction of equilibrium, but we choose the belief defined above for its tractability and clear implication.

The justification of the beliefs about the other actions is similar to and simpler than the above analysis and is thus omitted.

Part II: we categorize possible histories of play and identifies the conditions under which $\sigma_{Nash}^{(v_1,v_2)}$ is optimal for each player, and show that these conditions are captured by Assumption 1.

Step 1. We begin with histories where at least one deviation from $\sigma_{Nash}^{(v_1,v_2)}$ has occurred.

Step 1.1. $t \in \{2, \cdots, T\}$, a^D has occurred before. In this case, it is common knowledge that both players are a^D-players, they will play a^D from now on.

Step 1.2. $t \in \{t^*, \cdots, T\}$, a^D has not occurred before. Suppose that $\tilde{\lambda}_{i,t}^D = \tilde{\lambda}_{j,t}^D = 0$, note that this can only occur when $t \geq t^* + 1$ since no one is supposed to play a^D before t. If i is not an a^D-player, given that j will play $a(0)$ from now on according to $\sigma_{Nash}^{(v_1,v_2)}$, her best response is to also play $a(0)$ from now on. If i is an a^D-player, for $t = T$ it is clear that she will play a^D; for $t \in \{t^* + 1, \cdots, T-1\}$, it suffices for $\sigma_{Nash}^{(v_1,v_2)}$ to be optimal that she (weakly) prefers playing $a(0)$ now and a^D in the next period to playing a^D now. The corresponding condition is

$$\begin{aligned} 0 + g_1(a^D, a(0)) &\geq g_1(a^D, a(0)) + g_1(a^D, a^D) \\ 0 &\geq g_1(a^D, a^D). \end{aligned} \tag{1}$$

Next suppose that $\tilde{\lambda}_{i,t}^D = \tilde{\lambda}_{j,t}^D = \lambda^D$, since at least one deviation has occurred and since $t \geq t^*$, the beliefs jump to either 0 or 1 in the next period, conditional on the actions taken at t. If i is not an a^D-player, she knows that player j is going to play a^D in the next period if and only if j is an a^D-player, and information will inevitably become complete in the next period. Therefore, i should maximize her current-period payoff and play $a(\lambda^D)$. If i is an a^D-player, the one-step deviation principle gives a sufficient condition for a^D to be optimal:

$$\begin{aligned} &\lambda^D(T-t+1)g_1(a^D, a^D) + (1-\lambda^D)(g_1(a^D, a(\lambda^D)) + (T-t)g_1(a^D, a^D)) \\ \geq &\lambda^D(g_1(a(\lambda^D), a^D) + (T-t)g_1(a^D, a^D)) + (1-\lambda^D)(g_{max} + 0 \cdot (T-t-1) + g_{max}) \\ &\frac{\lambda^D}{1-\lambda^D}(g_1(a^D, a^D) - g_1(a(\lambda^D), a^D)) + (T-t)g_1(a^D, a^D) \geq 2g_{max} - g_1(a^D, a(\lambda^D)). \end{aligned}$$

In this condition, we relax i's payoff from playing an action other than a^D such that she can earn a strictly higher payoff g_{max} immediately and also g_{max} in period T when j cannot play a^D. By Assumption 1, $g_1(a^D, a^D) \leq 0$, then this condition

holds for every $t \in \{t^*, \cdots, T\}$ if it holds for $t = t^*$:

$$\frac{\lambda^D}{1-\lambda^D}(g_1(a^D, a^D) - g_1(a(\lambda^D), a^D)) + (T - t^*)g_1(a^D, a^D) \geq 2g_{max} - g_1(a^D, a(\lambda^D))$$
$$g_1(a(\lambda^D), a^D) \leq \frac{1}{\lambda^D}\{[(1-\lambda^D)(T-t^*) + \lambda^D]g_1(a^D, a^D) - (1-\lambda^D)(2g_{max} - g_1(a^D, a(\lambda^D)))\}. \tag{2}$$

Step 1.3. $t \in \{2, \cdots, t^* - 1\}$, a^D has not occurred before. The players' beliefs remain at $\lambda^D_{i,t} = \lambda^D_{j,t} = \lambda^D$. If i is not an a^D-player, following the one-step deviation principle it is clear that her optimal action is $a(0)$. If i is an a^D-player, she prefers $a(0)$ to a^D (and, of course, every other action in $\mathcal{A}$) if

$$\begin{aligned}
&0(t^* - t) + \lambda^D(T - t^* + 1)g_1(a^D, a^D) + (1-\lambda^D)(g_1(a^D, a(\lambda^D)) + (T-t^*)g_1(a^D, a^D)) \\
\geq &g_{max} + (T-t)g_1(a^D, a^D) \\
&(1-\lambda^D)g_1(a^D, a(\lambda^D)) \geq g_{max} + (t^* - t - \lambda^D)g_1(a^D, a^D).
\end{aligned}$$

Again, we relax the payoff from playing a^D so that i earns g_{max} immediately. As $g_1(a^D, a^D) \leq 0$, the above condition holds for every $t \in \{2, \cdots, t^* - 1\}$ if it holds for $t = t^* - 1$:

$$\begin{aligned}
(1-\lambda^D)g_1(a^D, a(\lambda^D)) &\geq g_{max} + (1-\lambda^D)g_1(a^D, a^D) \\
g_1(a^D, a^D) &\leq g_1(a^D, a(\lambda^D)) - \frac{1}{1-\lambda^D}g_{max}.
\end{aligned} \tag{3}$$

Step 2. Now we consider histories that have not deviated from $\sigma^{(a_1,a_2)}_{Nash}$.

Step 2.1. $t = T$. It is clear that an a^D-player will use a^D, and by the definition of $a(\lambda^D)$ a player who cannot play a^D will use $a(\lambda^D)$ against $a(\lambda^D)$.

Step 2.2. $t \in \{t^*, \cdots, T-1\}$. If i is not an a^D-player, she will not deviate from $a(0)$ if

$$\begin{aligned}
\lambda^D g_1(a(\lambda^D), a^D) + (1-\lambda^D)g_1(a(\lambda^D), a(\lambda^D)) \geq &0 + \lambda^D(g_1(a(\lambda^D), a^D) + (T-1-t)g_1(a^D, a^D)) \\
&+ (1-\lambda^D)(g_1(a(\lambda^D), a(\lambda^D)) + (T-1-t)\cdot 0),
\end{aligned}$$

where the first 0 on the right-hand side measures the upper bound of her current payoff from a deviation from $a(0)$. Since $g_1(a^D, a^D) \leq 0$, the above condition always holds for every $t \in \{t^*, \cdots, T-1\}$.

If i is an a^D-player, she will not deviate to any other action in $\mathcal{A}$ if

$$\begin{aligned}&\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\\ \geq&0+\lambda^D(T-t)g_1(a^D,a^D)+(1-\lambda^D)(g_1(a^D,a(\lambda^D))+(T-1-t)g_1(a^D,a^D)),\end{aligned}$$

where the first 0 on the right-hand side measures the upper bound of her current payoff from a deviation to any other action in $\mathcal{A}$. Since $g_1(a^D,a^D)\leq 0$, the above condition always holds for every $t\in\{t^*,\cdots,T-1\}$.

She will not deviate to a^D if

$$\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\geq g_1(a^D,a(0))+(T-t)g_1(a^D,a^D).$$

As $g_1(a^D,a^D)\leq 0$, the above condition holds for every $t\in\{t^*,\cdots,T-1\}$ if it holds for $t=T-1$:

$$\begin{aligned}(1-\lambda^D)(g_1(a^D,a(\lambda^D))-g_1(a^D,a^D))&\geq g_1(a^D,a(0))\\ g_1(a^D,a^D)&\leq g_1(a^D,a(\lambda^D)-\frac{1}{1-\lambda^D}g_1(a^D,a(0)).\end{aligned}$$

The above inequality is guaranteed by (3).

Step 2.3. $t\in\{3,\cdots,t^*-1\}$. Suppose that there exists a_1, a_2 such that all the pure actions that constitute a_1 and a_2 are played with positive weights in the first two periods and $g_1(a_1,a_2)=v_1$, $g_2(a_1,a_2)=v_2$. As no publicly identified deviation occurred, since period 3, the players must have played (a_1,a_2), which is an element of $a_v(A)$. If i is not an a^D-player, she will not deviate from playing a_i if

$$\begin{aligned}&(t^*-t)v_i+0\cdot(T-t^*)+\lambda^D g_1(a(\lambda^D),a^D)+(1-\lambda^D)g_1(a(\lambda^D),a(\lambda^D))\\ \geq&g_{max}+0(t^*-t-1)+\lambda^D(g_1(a(\lambda^D),a^D)+(T-t^*)g_1(a^D,a^D))\\ &+(1-\lambda^D)(g_1(a(\lambda^D),a(\lambda^D))+0(T-t^*)),\end{aligned}$$

where g_{max} measures the upper bound of her current payoff from a deviation from a_i. As $v_i\geq 0$, $g_1(a^D,a^D)\leq 0$, the above condition holds for every $t\in\{1,\cdots,t^*-1\}$ if it holds for $t=t^*-1$, $v_i=0$:

$$\begin{aligned}0&\geq g_{max}+\lambda^D(T-t^*)g_1(a^D,a^D)\\ g_1(a^D,a^D)&\leq-\frac{1}{\lambda^D(T-t^*)}g_{max}.\end{aligned}\tag{4}$$

If i is an a^D-player, she will not deviate to any other action in $\mathcal{A}$ if

$$\begin{aligned}&(t^*-t)v_i+0\cdot(T-t^*)+\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\\ \geq& g_{max}+0(t^*-t-1)+\lambda^D(T+1-t^*)g_1(a^D,a^D)\\ &+(1-\lambda^D)(g_1(M,a(\lambda^D))+(T-t^*)g_1(a^D,a^D)),\end{aligned}$$

where g_{max} measures the upper bound of her current payoff from a deviation to any other action in $\mathcal{A}$. The above condition holds for every $t\in\{1,\cdots,t^*-1\}$ if it holds for $t=t^*-1$, $v_i=0$:

$$\begin{aligned}0&\geq g_{max}+(T-t^*)g_1(a^D,a^D)\\ g_1(a^D,a^D)&\leq -\frac{1}{T-t^*}g_{max}.\end{aligned}\tag{5}$$

The above condition is guaranteed by (4).

She will not deviate to a^D if

$$\begin{aligned}&(t^*-t)v_i+0(T-t^*)+\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\\ \geq& g_1(a^D,a(0))+(T-t)g_1(a^D,a^D).\end{aligned}$$

The above condition holds for every $t\in\{1,\cdots,t^*-1\}$ if it holds for $t=t^*-1$, $v_i=0$:

$$\begin{aligned}\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))&\geq g_1(a^D,a(0))+(T+1-t^*)g_1(a^D,a^D)\\ g_1(a^D,a^D)&\leq\frac{1}{T+1-t^*-\lambda^D}[(1-\lambda^D)g_1(a^D,a(\lambda^D))-g_1(a^D,a(0))].\end{aligned}\tag{6}$$

Next suppose that such a_1, a_2 does not exist. Following $\sigma_{Nash}^{(v_1,v_2)}$, the players play $(a(0),a(0))$. If i is not an a^D-player, she will not deviate from playing $a(0)$ if

$$\begin{aligned}&(t^*-t)0+0\cdot(T-t^*)+\lambda^D g_1(a(\lambda^D),a^D)+(1-\lambda^D)g_1(a(\lambda^D),a(\lambda^D))\\ \geq& g_{max}+0(t^*-t-1)+\lambda^D(g_1(a(\lambda^D),a^D)+(T-t^*)g_1(a^D,a^D))\\ &+(1-\lambda^D)(g_1(a(\lambda^D),a(\lambda^D))+0(T-t^*)),\end{aligned}$$

which is exactly (4). An a^D-player, on the other hand, will not deviate to any other

action in $\mathcal{A}$ if

$$\begin{aligned}&(t^*-t)0+0\cdot(T-t^*)+\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\\ \geq& g_{max}+0(t^*-t-1)+\lambda^D(T+1-t^*)g_1(a^D,a^D)\\ &+(1-\lambda^D)(g_1(a^D,a(\lambda^D))+(T-t^*)g_1(a^D,a^D)),\end{aligned}$$

which is exactly (5) and is therefore guaranteed by (4). She will not deviate to a^D if

$$\begin{aligned}&(t^*-t)v_i+0(T-t^*)+\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))\\ \geq& g_1(a^D,a(0))+(T-t)g_1(a^D,a^D),\end{aligned}$$

which is exactly (6).

Step 2.4. In period 2:

- a. Suppose that player 1 played $a_{ep}(A)$ for some $A\in\mathfrak{A}(v_1,v_2)$ in period 1. For $i=1,2$, if player i follows $\sigma_{Nash}^{(v_1,v_2)}$, she gets v_i in periods $2,\cdots,t^*-1$ and 0 in periods $t^*,\cdots,T-1$, and an expected payoff E in period T. The value of E is $\lambda^D g_1(a^D,a^D)+(1-\lambda^D)g_1(a^D,a(\lambda^D))$ if player i is an a^D-player, and $\lambda^D g_1(a(\lambda^D),a^D)+(1-\lambda^D)g_1(a(\lambda^D),a(\lambda^D))$ if she is not. If player i takes a one-step deviation, she gets at most g_{max} in period 2, 0 in periods $3,\cdots,t^*-1$, E in period t^*, and a negative expected payoff afterwards. When T (and hence t^*) is sufficiently large, it is optimal to follow $\sigma_{Nash}^{(v_1,v_2)}$.
- b. Suppose that player 1 did not play $a_{ep}(A)$ for any $A\in\mathfrak{A}(v_1,v_2)$ in period 1, and $2^{\mathcal{A}_2\cup A_{1,2}}\cap\mathfrak{A}(v_1,v_2)\neq\varnothing$. For player 2, it is clear that playing $a_{ep}(A)$ for some $A\in\mathfrak{A}(v_1,v_2)$ now dominates any other action when T is sufficiently large; among the set of possible $a_{ep}(A)$ that player 2 can choose from, the one that currently best responds to $a(0)$ is optimal since they generate identical continuation payoffs from the next period onwards. For player 1, every one-step deviation from $\sigma_{Nash}^{(v_1,v_2)}$ yields at most g_{max} currently, 0 in periods $3,\cdots,t^*-1$, E in period t^*, and a negative expected payoff afterwards; however, following $\sigma_{Nash}^{(v_1,v_2)}$ bears a positive probability of yielding v_1 in periods $3,\cdots,t^*-1$. When T (and hence t^*) is sufficiently large, it is optimal to follow $\sigma_{Nash}^{(v_1,v_2)}$.
- c. Suppose that player 1 did not play $a_{ep}(A)$ for any $A\in\mathfrak{A}(v_1,v_2)$ in period 1, and $2^{\mathcal{A}_2\cup A_{1,2}}\cap\mathfrak{A}(v_1,v_2)=\varnothing$. Player 1 has identical incentives to step 2.4.b, and the corresponding proof applies. For player 2, every one-step deviation from $\sigma_{Nash}^{(v_1,v_2)}$ yields a lower continuation payoff from the next period onwards, and $a(0)$ is a best response to $a(0)$ which will be played by player 1 in the current period. Hence, it is optimal to follow $\sigma_{Nash}^{(v_1,v_2)}$.

Step 2.5. In period 1:

- a. Player 2 has incentives similar to those of player 1 in step 2.4.b, and the corresponding proof applies.
- b. Suppose that $2^{\mathcal{A}_1} \cap \mathfrak{A}(v_1, v_2) \neq \varnothing$. For player 1, playing $a_{ep}(A)$ for some $A \in \mathfrak{A}(v_1, v_2)$ now dominates any other action when T is sufficiently large, since cooperation can thus be sustained for sure instead of only with a probability strictly smaller than 1. Among the set of possible $a_{ep}(A)$ that player 1 can choose from, the one that currently best responds to $a(0)$ is optimal since they generate identical continuation payoffs from the next period onwards.
- c. Suppose that $2^{\mathcal{A}_1} \cap \mathfrak{A}(v_1, v_2) = \varnothing$ and $\bar{A} \cap \mathcal{A}_1 \neq \varnothing$. The current gain for player 1, from deviating from $\sigma_{Nash}^{(v_1,v_2)}$, is at most $g_{max} - g_{min}$. However, every additional action in $\bar{A}$ that player 1 includes in her current period's play increases the probability that cooperation will start from period 3. Hence, when T is sufficiently large, playing $a_{ep}(\bar{A} \cap \mathcal{A}_1)$ is optimal.
- d. Suppose that $\bar{A} \cap \mathcal{A}_1 = \varnothing$. Now player 1 has incentives similar to those of player 2 in step 2.4.c, and the corresponding proof applies.

Note that $g_1(a^D, a(\lambda^D)), g_1(a^D, a(0)) \in [g_{min}, g_{max}]$. Hence we can conclude that (i) given $T - t^*$, inequalities (1) and (3)-(6) must hold when $g_1(a^D, a^D)$ is sufficiently small; (ii) since $g_1(a(\lambda^D), a^D) \leq g_{br,a^D}$, given $g_1(a^D, a^D)$ and $T - t^*$, inequality (2) must hold when g_{br,a^D} is sufficiently smaller than $g_1(a^D, a^D)$. Finally, we let $t^* = T - 1$ in particular, and it is straightforward to verify by direct calculation that Assumption 1(a) implies the condition for (i) and Assumption 1(b) implies the condition for (ii).[9] Hence, $\sigma_{Nash}^{(a_1,a_2)}$ constitutes an equilibrium for all T.

Finally, given $g_1(a^D, a^D)$ and g_{br,a^D}, a player's sum of payoffs in periods $t^*, \cdots, T$ is a type-dependent constant which is bounded; the upper bound is g_{max}, and we denote the lower bound as K. Thus, fixing $T - t^*$ and for every ϵ, there must exist T^* sufficiently large such that for every $T \geq T^*$ and for $i = 1, 2$, $[\frac{(T-t^*)(K-v_i)}{T}, \frac{(T-t^*)(g_{max}-v_i)}{T}] \subseteq [-\epsilon, \epsilon]$, i.e. player i's average payoff is within ϵ of v_i. This completes the proof.

A.2 Proof of Theorem 2

We first prove that no efficient payoff vector can be implemented with either player playing a^D. Suppose an efficient payoff vector is attained from the action profile (a^D, a) for some

[9] The verification for conditions (1)-(5) is simple. For (6), when $t^* = T - 1$, (6) becomes $g_1(a^D, a^D) \leq \frac{1-\lambda^D}{2-\lambda^D}[g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D}]$. Note that $g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D} \geq g_{min} - \frac{g_{max}}{1-\lambda^D}$, if $g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D} \leq 0$, then $\frac{1-\lambda^D}{2-\lambda^D}[g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D}] \geq g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D} \geq g_{min} - \frac{g_{max}}{1-\lambda^D}$; if $g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D} > 0$, then $\frac{1-\lambda^D}{2-\lambda^D}[g_1(a^D, a(\lambda^D)) - \frac{g_1(a^D,a(0))}{1-\lambda^D}] > 0 > g_{min} - \frac{g_{max}}{1-\lambda^D}$. In conclusion, Assumption 1(a) implies a sufficient condition for (6).

$a \in \mathcal{A}$ (since (a^D, a^D) yields less payoff than $(a(0), a(0))$ and is clearly not efficient). The player who takes a^D obtains at most g_{max}. The player who does not take a^D obtains at most g_{br,a^D}. However, Assumption 1(a) implies $-g_1(a^D, a^D) > \frac{g_{max}}{\lambda^D} \geq \lambda^D g_{max}$, and hence Assumption 1(b) implies

$$\begin{aligned} g_{max} + g_{br,a^D} &< \frac{1}{\lambda^D}[g_1(a^D, a^D) - (1-\lambda^D)(2g_{max} - g_{min})] + g_{max} \\ &= \frac{1}{\lambda^D}[g_1(a^D, a^D) - (1-\lambda^D)(2g_{max} - g_{min}) + \lambda^D g_{max}] \\ &\leq \frac{1}{\lambda^D}[g_1(a^D, a^D) - (1-\lambda^D)(2g_{max} - g_{min}) - g_1(a^D, a^D)] \\ &= -\frac{1-\lambda^D}{\lambda^D}(2g_{max} - g_{min}) \\ &< 0 \\ &= g(a(0), a(0)). \end{aligned}$$

Since $a(0)$ only involves the actions that are always feasible, a^D can never appear in an action profile that results in efficient payoffs.

Next, suppose that there are L possible ex-post efficient total payoffs, $v^1, \cdots, v^L$, each corresponding to one or more pure action profiles in one or more possible configurations of $\mathcal{A}_1 \cup \mathcal{A}_2$. Without loss of generality, assume that $v^1 > v^2 > \cdots > v^L > 0$. Let $\mathfrak{A}^l$ denote the set of smallest sets of pure actions that can achieve v^l. For each $A \in \mathfrak{A}^l$ (each A consists of at most two pure actions), we use $a^l(A)$ to denote an action profile in $\Delta(A \times A)$ to achieve $\frac{v^l}{2}$ for each player. Let $\bar{A}^l = \cup_{A \in \mathfrak{A}^l} A$.

Consider the following strategy profile, denoted σ^e:

1. The players follow the strategies below:
   - a. In period 1: if $2^{\mathcal{A}_1} \cap \mathfrak{A}^1 \neq \varnothing$, player 1 finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A)$, $A \in 2^{\mathcal{A}_1} \cap \mathfrak{A}^1$, and plays $a_{ep}(A^*)$; if $2^{\mathcal{A}_1} \cap \mathfrak{A}^1 = \varnothing$, player 1 plays $a_{ep}(\bar{A}^1 \cap \mathcal{A}_1)$ if $\bar{A}^1 \cap \mathcal{A}_1 \neq \varnothing$, and $a(0)$ otherwise. Player 2 plays $a(0)$.
   - b. In period $t > 1$, if either player has played a^D before, both players play a^D. Otherwise, follow the strategy described below.
2. In period 2:
   - a. If there was no deviation from 1, then (1) if player 1 played $a_{ep}(A)$ for some $A \in \mathfrak{A}^1$ in period 1, the players play $a^1(A)$; (2) otherwise, then (2a) if $2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^1 \neq \varnothing$, player 2 finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A)$, $A \in 2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^1$, and plays $a_{ep}(A^*)$; (2b) if $2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^1 = \varnothing$ and $2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^2 \neq \varnothing$, player 2 finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A)$, $A \in 2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^2$, and plays $a_{ep}(A^*)$; (2c) if $2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^1 =$

$2^{\mathcal{A}_2 \cup A_{1,2}} \cap \mathfrak{A}^2 = \varnothing$ and $\bar{A}^2 \cap (\mathcal{A}_2 \cup A_{1,2}) \neq \varnothing$, player 2 plays $a_{ep}(\bar{A}^2 \cap (\mathcal{A}_2 \cup A_{1,2}))$; (2d) player 2 plays $a(0)$ otherwise. Player 1 plays $a(0)$.

- b. If there was a deviation from 1, the players play $(a(0), a(0))$.

...

t ($t \leq L$). In period t: let $i = 1$ if t is an odd number and 2 if t is an even number.

- a. If there was no deviation from 1 to $t-1$, then (1) if player $j \neq i$ played $a_{ep}(A)$ for some $A \in \mathfrak{A}^l$ for some $l \leq t-1$ in period $t-1$, the players play $a^l(A)$; (2) otherwise, then (2a) if $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} \neq \varnothing$, player i finds A^* such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A), A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1}$, and plays $a_{ep}(A^*)$; (2b) if $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} = \varnothing$ and $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t \neq \varnothing$, player i finds $a_{ep}(A^*)$ such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A), A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t$, and plays $a_{ep}(A^*)$; (2c) if $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} = 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t = \varnothing$ and $\bar{A}^t \cap (\mathcal{A}_i \cup A_{j,t}) \neq \varnothing$, player i plays $a_{ep}(\bar{A}^t \cap (\mathcal{A}_i \cup A_{j,t}))$; (2d) player i plays $a(0)$ otherwise. Player j plays $a(0)$.
- b. If there was a deviation from 1 to $t-1$, the players play $(a(0), a(0))$.

...

$L+1$. In period $L+1$: Let i be 1 if L is an even number and 2 if L is an odd number.

- a. If there was no deviation from 1 to L, then (1) if player $j \neq i$ played $a_{ep}(A)$ for some $A \in \mathfrak{A}^l$ for some l in period L, the players play $a^l(A)$; (2) otherwise, then (2a) if $2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L \neq \varnothing$, player i finds $a_{ep}(A^*)$ such that $a_{ep}(A^*)$ is a best response to $a(0)$ among $a_{ep}(A), A \in 2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L$, and plays $a_{ep}(A^*)$; (2b) player i plays $a(0)$ otherwise. Player j plays $a(0)$.
- b. If there was a deviation from 1 to L, the players play $(a(0), a(0))$.

$L+2$. Let $t^* \in \mathbb{N}^+$ be such that $T - t^* \geq 1$ and is a constant. In period $t \in \{L+2, \cdots, t^*-1\}$:

- a. If there was no deviation from 1 to $L+1$, then (1) if player 2 played $a_{ep}(A)$ for some $A \in \mathfrak{A}^l$ for some l in period $L+1$, the players play $a^l(A)$; (2) if player 2 played $a(0)$ in period $2L$, the players play $(a(0), a(0))$.
- b. If there was a deviation from 1 to $L+1$, the players play $(a(0), a(0))$.

$L+3$. In period $t \in \{t^*, \cdots, T-1\}$, if there was no deviation from 1 to $L+2$, the players play $(a(0), a(0))$. Otherwise, for $i = 1, 2$, play a^D if i is an a^D-player and $a(\lambda^D)$ if i is not an a^D-player.

$L+4$. In period T, for $i=1,2$, if i is an a^D-player, play a^D; if i is not an a^D-player, play $a(\lambda^D)$.

$L+5$. The belief updating rule $\tilde{\lambda}_{i,t}^k$ is the same as the proof of Theorem 1.

To verify that σ^e is an equilibrium when T is sufficiently large, given the proof of Theorem 1, we only need to prove optimality of the above strategies on path (in the sense that no deviation has been publicly identified) from period 1 to $L+1$, for agent 1 in odd-number periods and agent 2 in even-number periods. Our argument consists of the following steps.

Step 1. In period $L+1$:

- a. If player j played $a_{ep}(A)$ for some $A \in \mathfrak{A}^l$ for some l in period L, by definition of $a^l(A)$ we know that each player's per-period payoff in periods $L+2, \cdots, t^*-1$ is $\frac{v^l}{2} > 0$, while the current gain from any one-step deviation is at most $g_{max} - g_{min}$ at the cost of lowering the subsequent payoffs in periods $L+2, \cdots, t^*-1$ to 0. When T (and hence t^*) is sufficiently large, following σ^e is optimal. Otherwise:
- b. If $2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L \neq \varnothing$, it is first strictly better for player i to play $a_{ep}(A)$ for some $A \in 2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L$ when T is sufficiently large. It yields the maximum possible continuation payoff, $\frac{v^L}{2}$ in periods $L+2$ to t^*-1 with certainty, while the payoff from any one-step deviation is upper-bounded by $g_{max} - g_{min}$ in the current period and 0 in periods $t+1$ to t^*-1. Therefore, the optimal action for player i is a best response to $a(0)$ among $a_{ep}(A)$ for $A \in 2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L$.
- c. If $2^{\mathcal{A}_i \cup A_{j,L+1}} \cap \mathfrak{A}^L = \varnothing$, playing any other action than $a(0)$ triggers M from a pro opponent in periods $\{t^*, \cdots, T-1\}$. Hence given the conditions on the payoff parameters, $a(0)$ is optimal.

Step 2. In period t, $t \leq L$:

- a. If player j played $a_{ep}(A)$ for some $A \in \mathfrak{A}^l$ for some $l \leq t-1$ in period $t-1$, the proof follows step 1a. Otherwise:
- b. If $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} \neq \varnothing$, it is strictly better for player i to play $a_{ep}(A)$ for some $A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1}$ when T is sufficiently large. It yields the maximum possible continuation payoff, $\frac{v^{t-1}}{2}$ in periods $t+1$ to t^*-1 with certainty, while the payoff from any one-step deviation is upper-bounded by $g_{max} - g_{min}$ in the current period and at most $\frac{v^t}{2}$ in periods $t+1$ to t^*-1. Therefore, the optimal action for player 1 is a best response to $a(0)$ among $a_{ep}(A)$ for $A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1}$.

– c. If $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} = \varnothing$ and $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t \neq \varnothing$, it is strictly better for player i to play $a_{ep}(A)$ for some $A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t$ when T is sufficiently large. It yields the maximum possible continuation payoff, $\frac{v^t}{2}$ in periods $t+1$ to t^*-1 with certainty, while the payoff from any one-step deviation is upper-bounded by $g_{max} - g_{min}$ in the current period and at most $\frac{v^t}{2}$ only with a < 1 probability in periods $t+1$ to t^*-1. Therefore, the optimal action for player 1 is a best response to $a(0)$ among $a_{ep}(A)$ for $A \in 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t$.

– d. If $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} = 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t = \varnothing$ and $\bar{A}^t \cap (\mathcal{A}_i \cup A_{j,t}) \neq \varnothing$, every additional action in $\bar{A}^t$ that player 1 includes in her current period's play increases the probability that cooperation with payoff $(\frac{v^t}{2}, \frac{v^t}{2})$ will start from period $t+1$. Hence, when T is sufficiently large, player i's best response is to play $a(\bar{A}^t \cap (\mathcal{A}_i \cup A_{j,t}))$.

– e. If $2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^{t-1} = 2^{\mathcal{A}_i \cup A_{j,t}} \cap \mathfrak{A}^t = \bar{A}^t \cap (\mathcal{A}_i \cup A_{j,t}) = \varnothing$, playing any other action than $a(0)$ triggers a^D from a pro opponent in periods $\{t^*, \cdots, T-1\}$, as well as prevents any possible cooperation in periods $2q+1, \cdots, t^*-1$. Hence when T is sufficiently large, $a(0)$ is optimal.

In conclusion, when T is sufficiently large, σ^e is an equilibrium. Moreover, we know from Theorem 1 that i's payoff in σ^e is within ϵ of $v(\mathcal{A}_1 \cup \mathcal{A}_2)$ given sufficiently large T. This completes the proof.

B Additional Results

B.1 On Imperfect Learning

For $\mu \in (0,1)$, a prominent difference from "perfect learning" ($\mu = 1$) is that, now a player having used a^D is still unsure about her opponent's type unless the opponent has also used a^D, and her belief may evolve further afterwards. Specifically, suppose that player i has played a^D in period t, her belief at the beginning of period $t+1$ becomes

$$\tilde{\lambda}_i^D(h_t) = \tilde{\lambda}_i^D(h_{t-1}) + (1 - \tilde{\lambda}_i^D(h_{t-1}))\mu.$$

If her opponent plays a^D in period $t+1$, $\tilde{\lambda}_i$ jumps to 1 and persists afterwards. If the opponent plays an action in $\mathcal{A}$, i's belief depends on the opponent's strategy profile, which i takes as given in an equilibrium.

For the simplicity of notations, here we once again normalize the payoffs such that $g_{min} = 0$. In addition, we assume that g_{br,a^D} is sufficiently small. In particular:

$$g_{br,a^D} < -g_{max}. \tag{7}$$

We will present two results in this section. First, the existence of equilibrium as well as the supportable payoff space is discontinuous at $\mu = 1$, as for every μ sufficiently close to 1 we can find a sufficiently large T such that the repeated games have no equilibrium.

Proposition 2. *There exists $\mu^* \in (0,1)$ such that for every $\mu > \mu^*$, there exists no equilibrium when T is sufficiently large.*

Proof. Suppose that the game has some equilibrium σ^*. Consider the following (possibly off-equilibrium path) history: player i can play a^D and plays a^D at $t = 1$, while player j plays an action in $\mathcal{A}$. Bayes' updating of beliefs indicates that for every $t > 1$, $\lambda^D_{j,t} = 1$. We prove the nonexistence of equilibrium in five steps.

Step 1. We first introduce some definitions that will be useful for the rest of the proof.

Definition 1. *If player j's action for the current period is (1) a^D if available, and (2) an action (or a combination of actions) in $\mathcal{A}$ if a^D is not available, we say that she is using a separating strategy (Se) at this period. If j's action for the current period is invariant across types, we say that she is using a pooling strategy (Po).*

Definition 2. *Suppose that in equilibrium, starting from the current period, both player i and j only play actions other than a^D until period $T-1$ and $a^D \times Se$ at period T. We call this path a "cooperative path." Suppose that in equilibrium, starting from the current period, player i and j play $a^D \times Se$ up to the last period. We call this path a "noncooperative path."*

In the remaining steps 2-5, we will show that for every μ sufficiently close to 1, there cannot be any equilibrium play given the above history when T is sufficiently large.

Step 2. We prove that there exists $\bar{\mu}$ such that for every $\mu > \bar{\mu}$, if $\lambda^D_{i,T-1} = 0$, a cooperative path ensues at $t = T-1$ in every equilibrium.

At $t = T$, it is clear that in every equilibrium player i plays a^D and player j plays Se.

At $t = T-1$, in every equilibrium, player j must play Se since her opponent will surely play a^D in the last period. Let $\lambda^D_{i,T-1} = 0$. If player i's equilibrium play at $T-1$ is a^D, player j's equilibrium play must be $a(1)$, and, for every $a \in \mathcal{A}$

$$\begin{aligned} g_1(a^D, a(1)) + \mu g_1(a^D, a^D) + (1-\mu) g_1(a^D, a(1)) &\geq g_1(a, a(1)) + g_1(a^D, a(1)) \\ \mu g_1(a^D, a^D) + (1-\mu) g_1(a^D, a(1)) &\geq g_1(a, a(1)) \\ \mu &\leq \frac{g_1(a, a(1)) - g_1(a^D, a(1))}{g_1(a^D, a^D) - g_1(a^D, a(1))}. \end{aligned} \tag{8}$$

Let $\bar{\mu} = \max_{a \in \mathcal{A} \setminus \{a^D\}} \frac{g_1(a,a(1)) - g_1(a^D,a(1))}{g_1(a^D,a^D) - g_1(a^D,a(1))} < \frac{-g_{max}}{g_1(a^D,a^D) - g_{max}} \in (0,1)$ and $\mu > \bar{\mu}$, then (8) is violated, which means that in no equilibrium will player i play a^D at $T-1$ if his belief is 0.

Step 3. We prove that there exists $\hat{\mu}$ sufficiently close to 1 such that for every $\mu > \hat{\mu}$, if $\lambda^D_{i,T-1} \geq \mu$, a noncooperative path ensues at $t = T-1$ in every equilibrium.

At $t = T$, it is clear that in every equilibrium player i plays a^D and player j plays Se.

At $t = T-1$, in every equilibrium, player j must play Se since her opponent will surely play a^D in the last period. If player i's equilibrium play at $T-1$ is some $a \in \mathcal{A}$ and player j's equilibrium play is some $a' \in \mathcal{A}$ when a^D is not available, it is necessary that

$$\begin{aligned}
&\lambda^D_{i,T-1}(g_1(a,a^D)+g_1(a^D,a^D)) + (1-\lambda^D_{i,T-1})(g_1(a,a')+g_1(a^D,a(1))) \\
\geq &\lambda^D_{i,T-1}(g_1(a^D,a^D)+g_1(a^D,a^D)) + (1-\lambda^D_{i,T-1})(g_1(a^D,a') \\
&+\mu g_1(a^D,a^D) + (1-\mu)g_1(a^D,a(1))) \\
\lambda^D_{i,T-1} \leq &\frac{g_1(a^D,a') + \mu g_1(a^D,a^D) + (1-\mu)g_1(a^D,a(1)))}{g_1(a^D,a') + g_1(a,a^D) - (1-\mu)g_1(a^D,a^D) - \mu g_1(a^D,a(1)) - g_1(a,a')}. \qquad (9)
\end{aligned}$$

Note that given (7), the denominator in (9) is negative and the value of the right-hand side of (9) is strictly less than 1 when $\mu = 1$. Take

$$\hat{\mu} = \max_{a,a' \in \mathcal{A}\setminus\{a^D\}} \min_{\mu}\{\mu : \frac{g_1(a^D,a') + \mu' g_1(a^D,a^D) + (1-\mu')g_1(a^D,a(1)))}{g_1(a^D,a') + g_1(a,a^D) - (1-\mu')g_1(a^D,a^D) - \mu' g_1(a^D,a(1)) - g_1(a,a')} \leq \mu' \quad \forall \mu' \geq \mu\},$$

and we know that $\hat{\mu}$ is upper-bounded by

$$\min_{\mu}\{\mu : \frac{\mu' g_1(a^D,a^D) + (1-\mu')g_{max}}{g_{br,a^D} - (1-\mu')g_1(a^D,a^D) - \mu' g_{max}} \leq \mu' \ \forall \mu' \geq \mu\}.$$

For every $\mu > \hat{\mu}$, if $\lambda^D_{i,T-1} \geq \mu$, (9) is violated. Hence, in every equilibrium player i always plays a^D at both $t = T$ and $t = T-1$.

Step 4. We prove that there exists $\mu^* \in (0,1)$ such that for every $\mu > \mu^*$, if for some integer $k \geq 2$ and every $t \in (T-k, T)$, ever equilibrium has a cooperative path when $\lambda^D_{i,t} = 0$ and a noncooperative path when $\lambda^D_{i,t} \geq \mu$, then at $t = T-k-1$, every equilibrium has a cooperative path when $\lambda^D_{i,t} = 0$ and a noncooperative path when $\lambda^D_{i,t} \geq \mu$.

Let $\mu' = \frac{g_{max}}{g_{max} - g_1(a^D,a^D)}$, and let $\mu^* = \max\{\bar{\mu}, \hat{\mu}, \mu'\}$.

Consider $t = T-k-1$ and $\lambda^D_{i,t} = 0$. If player i's equilibrium play at t is a^D (note that her equilibrium play must be constituted by either pure a^D or a possibly mixed action without a^D, as a mixed action with a positive proportion of a^D generates the same learning probability for the opponent), by Step 2 we know that a noncooperative path ensues at period $T-k$. Hence, since $\mu > \mu^* \geq \mu'$, player i's continuation payoff is bounded above by $g_1(a^D,a(1)) + \mu g_1(a^D,a^D) + (1-\mu)g_1(a^D,a(1))$. If she switches to playing $\arg\max_{a \in \mathcal{A}} g_1(a,a(1))$, her continuation payoff is bounded below by $\max_{a \in \mathcal{A}} g_1(a,a(1)) + g_1(a^D,a(1))$. From (8) we know that $g_1(a^D,a(1)) + \mu g_1(a^D,a^D) + (1-\mu)g_1(a^D,a(1)) <$

$\max_{a\in\mathcal{A}} g_1(a, a(1)) + g_1(a^D, a(1))$ when $\mu > \mu^* \geq \bar{\mu}$, and hence i's equilibrium play can never be a^D. From Step 2, every equilibrium consists a cooperative path.

Consider $t = T - k - 1$ and $\lambda_{i,t}^D \geq \mu$. Suppose that player i's equilibrium play at t is not a^D. If player j's equilibrium play is Po, by Step 2 we know that a noncooperative path must ensue at period $T - k$, which means that Po is never optimal for player j, a contradiction. On the other hand, if player j's equilibrium play is Se and if j can play a^D, a noncooperative path must ensue at period $T - k$. However, now j has a profitable deviation of mimicking j's equilibrium action when he is not able to play a^D in period t and deferring a^D to period $T - k$, yielding a net benefit of at least $-g_1(a^D, a^D)$, again, this is a contradiction. Hence, player i's equilibrium play at t must be a^D. Since a noncooperative path must ensues at period $T - k$, player j's equilibrium play at t must be Se. Therefore, every equilibrium has a noncooperative path.

Combining Steps 2, 3 and 4, we know that for every $\mu > \mu^*$ at every period $t \leq T-1$, every equilibrium consists a cooperative path when $\lambda_{i,t}^D = 0$ and a noncooperative path when $\lambda_{i,t}^D \geq \mu$.

Step 5. We prove that for every $\mu > \mu^*$, there exists $T(\mu) \in \mathbb{N}^+$ such that for every $T > T(\mu)$, player i never plays a^D in period 2 in any equilibrium .

We know that $\lambda_{i,2}^D$ is bounded above by $\mu'' = \mu + \mu(1 - \mu) \in (0, 1)$. Suppose that player i's equilibrium play at $t = 2$ is a^D. By Step 4, it implies that a noncooperative path ensues at period 2. Let $p \in (0, 1)$ be a probability such that

$$pg_1(a^D, a^D) + (1 - p)g_{max} \leq (\mu'' + \mu''(1 - \mu''))g_1(a^D, a^D).$$

On a noncooperative path, we know that there exists $T'(\mu)$ such that the probability that j can play a^D at $t = T'(\mu)$, evaluated at the beginning of $t = 2$, is at least p. In other words, i's expected payoff in every period after $t = T'(\mu)$, evaluated at the beginning of $t = 2$, is at most $(\mu'' + \mu''(1 - \mu''))g_1(a^D, a^D)$. Let $T''(\mu) = \min\{T'(\mu)\}$. Player i's continuation payoff at the beginning of $t = 2$, assuming that her equilibrium play is a^D, is then bounded above by

$$(T''(\mu) - 1)g_{max} + (T - T''(\mu))(\mu'' + \mu''(1 - \mu''))g_1(a^D, a^D). \tag{10}$$

If player i deviate by playing $a(1)$, with probability μ'' her opponent can play a^D and she gets $g_1(a^D, a^D)$ from $t = 3$ onwards, with probability $1 - \mu''$ her opponent cannot play a^D and a cooperative path ensues at $t = 3$. Player i's continuation payoff at the beginning of $t = 2$, assuming her deviating to $a(1)$, is then bounded below by

$$\mu''(g_{br,a^D} + (T - 2)g_1(a^D, a^D)). \tag{11}$$

Taking the difference (11)-(10), the net benefit from deviation is bounded below by

$$\mu'' g_{br,a^D} - (T''(\mu) - 1) g_{max} - \mu''(2 + (T - T''(\mu))(1 - \mu'')) g_1(a^D, a^D). \tag{12}$$

Since $g_1(a^D, a^D) < 0$, there must exist $T(\mu)$ such that when $T > T(\mu)$, (12) > 0. Hence, player i never plays a^D in period 2. However, this is a contradiction to our assertion in Step 4 that a noncooperative path must occur when $\lambda_{i,2}^D \geq \mu$, which implies that there exists no equilibrium. This completes the proof.

□

Proposition 2 highlights contradicting incentives when μ is large in the case where one player, say i, has revealed herself to be an a^D-player early in the repeated games. When time t is close to the end, i takes opposite actions in every equilibrium (if any) according to her belief about the opponent j's type: if she believes with at least probability μ in facing another a^D-player, she will use a^D since it only affects the belief marginally, and j will also very likely respond by a^D; however, if she believes with certainty that j still cannot play a^D, she will not use a^D as it brings a probability of at least μ that she will face a^D later.

Now consider a history where i believes that j is an a^D-player with at least probability μ at the beginning of period $t-1$. It is straight forward that such a history, possibly off-path, can always occur. We can thus deduce that, i must now play a^D in every equilibrium: otherwise, j will not play a type-dependent action because j would rather pretend to be unable to play a^D, in exchange for not starting (a^D, a^D) prematurely; j will not play the same action between types either, because thus i will play a^D from the next period onwards and j, if an a^D-player, would rather start a^D now. Therefore, the only possible equilibrium play in period $t-1$ is for i to play a^D and j to play a^D as well if she can. This argument then unravels backwards to the very period right after i revealed her type. However, it is clear that she should refrain from a^D now in order to observe j's type, a contradiction.

Nevertheless, Proposition 2 does not exclude the existence of equilibrium for large μ once and for all. Reversing the parametric setting, i.e. when a sufficiently large T is fixed and μ approaches 1, the existence of a (cooperative) equilibrium is preserved. We state the result below; its proof follows that of Theorem 1.

Proposition 3. *For $i = 1, 2$, every $\epsilon > 0$ and every $(v_1, v_2) \in V^*_{Nash}$, there exists $T(\epsilon) \in \mathbb{N}^+$ and $\mu(T(\epsilon)) \in (0, 1)$ such that for all $T > T(\epsilon)$, $\mu > \mu(T(\epsilon))$, there exists an equilibrium where player i's average payoff is within ϵ of v_i.*

B.2 On Sustaining Minimax Payoffs

Theorem 1 and 2 shows the sustainability of payoffs that dominate Nash equilibrium payoffs. In this section, we propose another equilibrium to enlarge the set of sustainable payoffs from $\Delta(\mathcal{A}\times\mathcal{A})$ to their lower bound – the minimax payoffs. To begin with, we introduce some additional notations below.

Let $a_{minimax} := \arg\min_{a_i\in\mathcal{A}^e}\max_{a_j\in\mathcal{A}^e} g_j(a_i,a_j)$ denote a player's minimax strategy against her opponent in $\mathcal{A}^e\times\mathcal{A}^e$. We normalize the payoffs such that $\min_{a_i}\max_{a_j} g_j(a_i,a_j) = 0$, and let $g_{minimax} := g_1(a_{minimax}, a_{minimax})$. Note that $g_{minimax}\leq 0$ and $g_1(a(0),a(0))\geq 0$ with this normalization. If $g_{minimax}=0$ or $g_1(a(0),a(0))=0$, we can directly apply Theorem 1 and 2; hence we assume here that $g_{minimax}<0$ and $g_1(a(0),a(0))>0$. Let V^e denote the convex hull of the set $\{g(a,a') : a,a'\in\mathcal{A}^e\}$. Let $V^*_{minimax} := int(V^e\cap(\mathbb{R}^+\times\mathbb{R}^+)\cup V^*_{Nash})$. For every payoff vector $(v_1,v_2)\in V^*_{minimax}$, let (a_1,a_2) denote a (possibly mixed) action profile that achieves (v_1,v_2). To incorporate the more prohibitive punishment of minimax payoffs and thus the larger set of sustainable equilibrium payoffs, we rewrite Assumption 1 as the following more general version.

Assumption 1'. *Fix $\{\lambda^k\}_{k=m+1,\cdots,n}, \lambda^D, g_{max}$, and g_{min}. There exists a sufficiently small number $\hat{g}\in\mathbb{R}^-$ and a sufficiently large number $\tilde{g}\in\mathbb{R}^+$ such that $g_1(a^D,a^D)<\hat{g}$ and $g_{br,a^D}<g_1(a^D,a^D)-\tilde{g}$.*

Our next result, Theorem 3, is a direct extension of Theorem 1. It enlarges the range of (approximately within ϵ) supportable average payoffs to $V^*_{minimax}$. The key to sustain payoffs lower than $g_1(a(0),a(0))$ is to introduce a new phase after Phase II (cooperation) to allow the players to end any carried out punishment in a fixed number of periods, prior to the Phase III (possible punishment with uncertainty). Theorem 2 can be readily extended using the same approach.

Theorem 3. *For $i=1,2$, every $\epsilon>0$ and every $(v_1,v_2)\in V^*_{minimax}$, there exists an equilibrium such that, if (v_1,v_2) is feasible, then player i's average payoff is within ϵ of v_i when T is sufficiently large.*

Proof. Fix $T-t^*\in\mathbb{N}^+$. Find integer k such that for any $(v_1,v_2)\in V^*_{minimax}$, there exists $(v_1',v_2')\in V^*_{minimax}$ such that (v_1',v_2') is (strictly) within ϵ of (v_1,v_2) and that $\min\{v_1',v_2'\}\geq\frac{\epsilon}{k}$. For any $(v_1,v_2)\in V^*_{minimax}$, let (a_1,a_2) be an action profile such that $g_i(a_1,a_2)=v_i'$ for $i=1,2$.

Next, similar to the proof of Theorem 1, we propose the following strategy profile and associated belief updating rule, denoted as $\sigma^{(v_1,v_2)}_{minimax}$, and prove that they form an equilibrium where player i's average payoff is within ϵ of v_i.

1. In period 1 and 2, both players' strategy are the same as $\sigma^{(v_1,v_2)}_{Nash}$.

2. In period $t > 2$, if either player has played a^D before, play a^D. Otherwise, follow the strategy described in 4-7 below.

3. In period $t \in \{2, \cdots, t^* - \hat{t} - 1\}$ for some integer $t^* \in (1, T)$ and some integer $\hat{t} \in (0, t^* - 1)$, if there was no deviation from 1, then (1) if player 2 played $a(A)$ for some $A \in \mathfrak{A}(v_1, v_2)$ in period 2 and the players have played only $a_v(A)$ in periods $t - \hat{t}, \cdots, t - 1$, the players play $a_v(A)$; (2) if player 2 played $a(0)$ in period 2, the players play $(a(0), a(0))$. Otherwise, if there was a deviation from 1 which can be publicly identified, the players play:

- a. When a deviation from (a_1, a_2) occurs, enter the *punishment phase* from the next period: play $a_{minimax}$ for $\hat{t}$ periods.
- b. If no player deviated in the punishment phase, switch back to playing a_i (or $a(0)$ if a_i is not available) when the punishment phase is over.
- c. Restart the punishment phase if any player deviates from $a_{minimax}$.

4. In period $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$: if no punishment phase has started or restarted within $\hat{t} - 1$ periods, then play $a(0)$. Otherwise:

- a. If no player has deviated from $a_{minimax}$ in periods $t^* - \hat{t}, \cdots, t - 1$, then play $a_{minimax}$.
- b. Otherwise, play $a(0)$.

5. In period $\{t^*, \cdots, T - 1\}$: if no punishment phase has started or restarted at or after period $t^* - 2\hat{t} + 1$, or no player has deviated from $a_{minimax}$ within a punishment phase in periods $t^* - \hat{t}, \cdots, t^* - 1$, then play $a(0)$. Otherwise, if $(\lambda_{i,t}^D, \lambda_{j,t}^D) = (\lambda, \lambda)$, play a^D if available and $a(\lambda)$ if not; if $(\lambda_{i,t}^D, \lambda_{j,t}^D) = (0, 0)$, play $a(0)$.

6. In period T: player a^D if available, otherwise, play $a(\lambda)$ if $(\lambda_{i,T}^D, \lambda_{j,T}^D) = (\lambda, \lambda)$, and play $a(0)$ if $(\lambda_{i,T}^D, \lambda_{j,T}^D) = (0, 0)$.

7. The belief updating rule is analogous to that in Section 3.1.

Again, we categorize possible histories of play and identifies the conditions under which $\sigma_{minimax}^{(v_1,v_2)}$ is optimal for each player, we show that these conditions are captured by Assumption 1.

Step 1. We begin with histories where at least one deviation from $\sigma_{minimax}^{(v_1,v_2)}$ has occurred.

Step 1.1. For $t \in \{2, \cdots, T\}$, a^D has occurred before. Following step 1.1 in the proof of Theorem 1, playing a^D is optimal for each player.

Step 1.2. For $t \in \{t^*, \cdots, T\}$, a^D has not occurred before. Following step 1.2 in the proof of Theorem 1, $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player.

Step 1.3. For $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$, a^D has not occurred before. The players' beliefs remain at $\lambda^D_{i,t} = \lambda^D_{j,t} = \lambda$. There are three possible scenarios:

- a. Some player has deviated from a punishment phase in periods $t^* - \hat{t}, \cdots, t - 1$. For player i, playing $a(0)$ clearly dominates every other action in $\mathcal{A}\backslash\{a^D\}$ regardless of her type. If i can play a^D, playing $a(0)$ dominates playing a^D if

$$\begin{aligned}&(t^* - t)g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1 - \lambda)g_1(a^D, a(\lambda)) + (T - t^*)g_1(a^D, a^D)\\ \geq& g_1(a^D, a(0)) + (T - t)g_1(a^D, a^D).\end{aligned}$$

  The above condition holds for every $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$ if

$$g_1(a(0), a(0)) + (1 - \lambda)g_1(a^D, a(\lambda)) \geq g_1(a^D, a(0)) + (1 - \lambda)g_1(a^D, a^D). \quad (13)$$

- b. No player has deviated from a punishment phase in periods $t^* - \hat{t}, \cdots, t - 1$, and the game is in a punishment phase according to $\sigma^{(a_1,a_2)}_{minimax}$. Suppose that the punishment phase has $t' \leq t^* - t$ periods left, including period t. For player i, playing $a_{minimax}$ dominates every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$\begin{aligned}&t' g_{minimax} + (T - t - t')g_1(a(0), a(0)) + \lambda g_1(a(\lambda), a^D) + (1 - \lambda)g_1(a(\lambda), a(\lambda))\\ \geq& 0 + (t^* - 1 - t)g_1(a(0), a(0)) + \lambda g_1(a(\lambda), a^D)\\ &+ (1 - \lambda)g_1(a(\lambda), a(\lambda)) + (T - t^*)g_1(a^D, a^D).\end{aligned}$$

  The above condition holds for every $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$ if

$$\hat{t} g_{minimax} + (T - t^*)g_1(a(0), a(0)) \geq 0 + (\hat{t} - 1)g_1(a(0), a(0)) + (T - t^*)g_1(a^D, a^D).$$

  A further sufficient condition is

$$\hat{t} g_{minimax} + g_1(a(0), a(0)) \geq 0 + (\hat{t} - 1)g_1(a(0), a(0)) + g_1(a^D, a^D). \quad (14)$$

  If i can play a^D, playing $a_{minimax}$ also dominates a^D if

$$\begin{aligned}&t' g_{minimax} + (T - t - t')g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1 - \lambda)g_1(a^D, a(\lambda))\\ \geq& g_{max} + (T - t)g_1(a^D, a^D).\end{aligned}$$

  Here we relax the payoff of playing a^D against $a_{minimax}$ to g_{max}. The above

condition holds for every $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$ if

$$\hat{t} g_{minimax} + g_1(a(0), a(0)) \geq g_{max} + \hat{t} g_1(a^D, a^D). \tag{15}$$

– c. No player has deviated from a punishment phase in periods $t^* - \hat{t}, \cdots, t-1$, and the game is not in a punishment phase according to $\sigma^{(a_1,a_2)}_{minimax}$. For player i, playing $a(0)$ clearly dominates every other action in $\mathcal{A}\backslash\{a^D\}$ regardless of her type. If i can play a^D, $a(0)$ also dominates a^D if

$$\begin{aligned}&(T-t) g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1-\lambda) g_1(a^D, a(\lambda)) \\ \geq & g_1(a^D, a(0)) + (T-t) g_1(a^D, a^D).\end{aligned}$$

The above condition holds for every $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$ if

$$(T - t^* + 1) g_1(a(0), a(0)) \geq g_1(a^D, a(0)) + (T - t^*) g_1(a^D, a^D).$$

A further sufficient condition is

$$2 g_1(a(0), a(0)) \geq g_1(M, a(0)) + g_1(M, M). \tag{16}$$

Step 1.4. For $t \in \{2, \cdots, t^* - \hat{t} - 1\}$, a^D has not occurred before. The players' beliefs remain at $\lambda^D_{i,t} = \lambda^D_{j,t} = \lambda$. There are two possible scenarios:

– a. The game is not in a punishment phase. If $t \leq t^* - 2\hat{t} - 1$, i weakly prefers playing a_i to every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$\begin{aligned}&(t^* - \hat{t} - t) v_i' + (T - t^* + \hat{t}) g_1(a(0), a(0)) \\ \geq & g_{max} + \hat{t} g_{minimax} + (t^* - 2\hat{t} - t - 1) v_i' + (T - t^* + \hat{t}) g_1(a(0), a(0)).\end{aligned}$$

Here we relax the payoff from deviation in the current period to g_{max}. The above condition can be simplified as

$$(\hat{t} + 1) v_i' \geq g_{max} + \hat{t} g_{minimax}. \tag{17}$$

If i can play a^D, she weakly prefers playing $a_{minimax}$ to playing a^D if

$$\begin{aligned}&(t^* - \hat{t} - t) v_i' + (T - t^* + \hat{t}) g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1-\lambda) g_1(a^D, a(\lambda)) \\ \geq & g_{max} + (T-t) g_1(a^D, a^D).\end{aligned}$$

The above condition holds for every $t \leq t^* - 2\hat{t} - 1$ if

$$v_i' + (T - t^* + \hat{t})g_1(a(0), a(0)) \geq g_{max} + (T - t^* + \hat{t})g_1(a^D, a^D). \tag{18}$$

If $t > t^* - 2\hat{t} - 1$, i weakly prefers playing a_i to every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$\begin{aligned}&(t^* - \hat{t} - t)v_i' + (T - t^* + \hat{t})g_1(a(0), a(0)) \\ \geq &g_{max} + \hat{t}g_{minimax} + (T - t - \hat{t} - 1)g_1(a(0), a(0)).\end{aligned}$$

The above condition holds for every $t > t^* - 2\hat{t} - 1$ if

$$g_1(a(0), a(0)) \geq g_{max} + \hat{t}g_{minimax}. \tag{19}$$

If i can play a^D, she weakly prefers playing a_i to playing a^D if

$$\begin{aligned}&(t^* - \hat{t} - t)v_i' + (T - t^* + \hat{t})g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1 - \lambda)g_1(a^D, a(\lambda)) \\ \geq &g_{max} + (T - t)g_1(a^D, a^D).\end{aligned}$$

The above condition holds for every $t > t^* - 2\hat{t} - 1$ given (18).

– b. The game is in a punishment phase. Let $t' \leq \hat{t}$ denote the number of remaining periods (including period t) of the punishment phase. If $t \leq t^* - 2\hat{t} - 1$, which implies that $t' < t^* - \hat{t} - t$, i weakly prefers playing $a_{minimax}$ to every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$\begin{aligned}&t'g_{minimax} + (t^* - \hat{t} - t - t')v_i' + (T - t^* + \hat{t})g_1(a(0), a(0)) \\ \geq &0 + \hat{t}g_{minimax} + (t^* - 2\hat{t} - t - 1)v_i + (T - t^* + \hat{t})g_1(a(0), a(0)).\end{aligned}$$

The above condition can be simplified as

$$(\hat{t} - t' + 1)v_i' \geq (\hat{t} - t')g_{minimax},$$

which holds for every $t \leq t^* - 2\hat{t} - 1$ and $t' \leq \hat{t}$. If i can play a^D, she weakly prefers playing $a_{minimax}$ to playing a^D if

$$\begin{aligned}&t'g_{minimax} + (t^* - \hat{t} - t - t')v_i' + (T - t^* + \hat{t})g_1(a(0), a(0)) \\ &+ \lambda g_1(a^D, a^D) + (1 - \lambda)g_1(a^D, a(\lambda)) \\ \geq &g_{max} + (T - t)g_1(a^D, a^D).\end{aligned}$$

Here we relax the payoff from deviation in the current period to be g_{max}. The

above condition holds for every $t \leq t^* - 2\hat{t} - 1$ and $t' \leq \hat{t}$ if

$$\hat{t} g_{minimax} \geq g_{max} + (T - t^* + \hat{t}) g_1(a^D, a^D). \tag{20}$$

If $t^* - 2\hat{t} - 1 < t \leq t^* - \hat{t} - t'$, i weakly prefers playing $a_{minimax}$ to every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$\begin{aligned} &t' g_{minimax} + (t^* - \hat{t} - t - t') v_i' + (T - t^* + \hat{t}) g_1(a(0), a(0)) \\ \geq &0 + \hat{t} g_{minimax} + (T - t - \hat{t} - 1) g_1(a(0), a(0)). \end{aligned}$$

The above condition can be simplified as

$$(t^* - \hat{t} - t - t') v_i' + (2\hat{t} + t + 1 - t^*) g_1(a(0), a(0)) \geq (\hat{t} - t') g_{minimax},$$

which holds for every $t \in (t^* - 2\hat{t} - 1, t^* - \hat{t} - t')$ and $t' \leq \hat{t}$. If i can play a^D, she weakly prefers playing $a_{minimax}$ to playing a^D if (20) holds.
If $t > t^* - \hat{t} - t'$, i weakly prefers playing $a_{minimax}$ to every other action in $\mathcal{A}\backslash\{a^D\}$ if

$$t' g_{minimax} + (T - t - t') g_1(a(0), a(0)) \geq 0 + \hat{t} g_{minimax} + (T - t - \hat{t} - 1) g_1(a(0), a(0)).$$

The above condition can be simplified as

$$(\hat{t} + 1 - t') g_1(a(0), a(0)) \geq (\hat{t} - t') g_{minimax},$$

which holds for every $t > t^* - \hat{t} - t'$ and $t' \leq \hat{t}$. If i can play a^D, she weakly prefers playing $a_{minimax}$ to playing a^D if

$$\begin{aligned} &t' g_{minimax} + (T - t - t') g_1(a(0), a(0)) + \lambda g_1(a^D, a^D) + (1 - \lambda) g_1(a^D, a(\lambda)) \\ \geq &g_{max} + (T - t) g_1(a^D, a^D). \end{aligned}$$

The above condition holds for every $t > t^* - \hat{t} - t'$ and $t' \leq \hat{t}$ if (20) holds.

Next, we consider histories that have not deviated from $\sigma^{(a_1,a_2)}_{minimax}$.

Step 2.1. $t = T$. Following step 2.1 in the proof of Theorem 1, $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player.

Step 2.2. $t \in \{t^*, \cdots, T\}$. Following step 2.2 in the proof of Theorem 1, $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player.

Step 2.3. $t \in \{t^* - \hat{t}, \cdots, t^* - 1\}$. Apply step 1.3(c), and $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player given the corresponding conditions.

Step 2.4. $t \in \{3, \cdots, t^* - \hat{t} - 1\}$. Apply step 1.4(a), and $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player given the corresponding conditions.

Step 2.5. $t \in \{1, 2\}$. Following step 2.4 and 2.5 in the proof of Theorem 1, $\sigma^{(a_1,a_2)}_{minimax}$ is optimal for each player.

The set of sufficient conditions for $\sigma^{(a_1,a_2)}_{minimax}$ to be an equilibrium is (13)-(20). Since $v_i' \geq \frac{\epsilon}{k}$ by construction and $g_{minimax}$ is fixed and negative, we can now find $\hat{t} \in \mathbb{N}^+$ so that (17) and (19) are satisfied. Fix one such $\hat{t}$, it is clear that (13)-(16), (18) and (20) are satisfied when $g_1(a^D, a^D)$ is sufficiently small. Hence similar to the proof of Theorem 1, we can find a sufficiently small number $\hat{g} \in \mathbb{R}^-$ and a sufficiently large number $\tilde{g} \in \mathbb{R}^+$ such that $\sigma^{(a_1,a_2)}_{minimax}$ is an equilibrium. This completes the proof.

□

References

Andreoni, J. (1988). Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics 37*(3), 291–304.

Andreoni, J. and J. H. Miller (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal 103*(418), 570–585.

Benoit, J.-P. and V. Krishna (1985). Finitely repeated games. *Econometrica 53*(4), 905–922.

Bhaskar, V. and E. van Damme (2002). Private strategies in finitely repeated games with imperfect public monitoring. *Journal of Economic Theory 102*(1), 16–39.

Bos, I. and M. A. Marini (2022). Collusion in quality-segmented markets. *Journal of Public Economic Theory 24*(2), 293–323.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory 104*(1), 137–188.

Chandrasekhar, A. G. and J. P. Xandri (2023). A note on payments in the lab for infinite horizon dynamic games with discounting. *Economic Theory 75*, 389–426.

Chassang, S. (2010). Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts. *American Economic Review 100*(1), 448–465.

Crawford, V. P. and H. Haller (1990). Learning how to cooperate: Optimal play in repeated coordination games. *Econometrica 58*(3), 571–595.

Cressman, R. (1996). Evolutionary stability in the finitely repeated prisoner 's dilemma game. *Journal of Economic Theory 68*(1), 234–248.

Dutta, P. K. (1995). A folk theorem for stochastic games. *Journal of Economic Theory 66*(1), 1–32.

Fudenberg, D. and E. Maskin (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica 54*(3), 533–554.

Fudenberg, D. and Y. Yamamoto (2011). The folk theorem for irreducible stochastic games with imperfect public monitoring. *Journal of Economic Theory 146*(4), 1664–1683.

Joosten, R., H. Peters, and F. Thuijsman (1995). Unlearning by not doing: Repeated games with vanishing actions. *Games and Economic Behavior 9*(1), 1–7.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory 27*(2), 245–252.

Kreps, D. M. and R. Wilson (1982). Reputation and imperfect information. *Journal of Economic Theory 27*(2), 253–279.

Mailath, G. J., S. A. Matthews, and T. Sekiguchi (2002). Private strategies in finitely repeated games with imperfect public monitoring. *The B.E. Journal of Theoretical Economics 2*(1), 1–23.

Marlats, C. (2015). A folk theorem for stochastic games with finite horizon. *Economic Theory 58*(3), 485–507.

Milgrom, P. and J. Roberts (1982). Predation, reputation, and entry deterrence. *Journal of Economic Theory 27*(2), 280–312.

Miyahara, Y. and T. Sekiguchi (2013). Finitely repeated games with monitoring options. *Journal of Economic Theory 148*(5), 1929–1952.

Mookherjee, D. and B. Sopher (1994). Learning behavior in an experimental matching pennies game. *Games and Economic Behavior 7*(1), 62–91.

Muller, L., M. Sefton, R. Steinberg, and L. Vesterlund (2008). Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior and Organization 67*(3–4), 782–793.

Nachbar, J. H. (1992). Evolution in the finitely repeated prisoner's dilemma. *Journal of Economic Behavior and Organization 19*(3), 307–326.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review 85*(5), 1313–1326.

Selten, R. and R. Stoecker (1986). End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach. *Journal of Economic Behavior and Organization 7*(1), 47–70.

Weinstein, J. and M. Yildiz (2016). Reputation without commitment in finitely repeated games. *Theoretical Economics 11*(1), 157–185.