Generalizing Better Response Paths and Weakly Acyclic Games

Bora Yongacoglu, Gürdal Arslan, Lacra Pavel, and Serdar Yüksel

***Abstract*— Weakly acyclic games generalize potential games and are fundamental to the study of game theoretic control. In this paper, we present a generalization of weakly acyclic games, and we observe its importance in multi-agent learning when agents employ experimental strategy updates in periods where they fail to best respond. While weak acyclicity is defined in terms of path connectivity properties of a game's better response graph, our generalization is defined using a generalized better response graph. We provide sufficient conditions for this notion of generalized weak acyclicity in both two-player games and n-player games. To demonstrate that our generalization is not trivial, we provide examples of games admitting a pure Nash equilibrium that are not generalized weakly acyclic. The generalization presented in this work is closely related to the recent theory of satisficing paths, and the counterexamples presented here constitute the first negative results in that theory.**

I. Introduction

Algorithms for seeking Nash equilibrium play an important role in the game theoretic approach to distributed control [1]–[4]. In Nash-seeking algorithms, players in a game iteratively adjust their strategies over time, and the goal of the algorithm designer is to guide the collective strategy profile to a Nash equilibrium of the game, where play should stabilize. In this paper, we identify and study mathematical properties of games with the aim of informing the design of Nash-seeking algorithms. We do not, however, present an explicit algorithm of our own and we abstract away from learning theoretic considerations, focusing instead on the structure of the underlying game.

This article presents a new class of games relevant to game theoretic learning and Nash-seeking algorithms. We refer to this class of games as *generalized weakly acyclic games* (GenWAGs), since they constitute a meaningful generalization of weakly acyclic games and are defined in an analogous manner [5]–[9]. In turn, weakly acyclic games generalize potential games [10], a class of games used to model cooperative and distributed control [11]–[13].

Whereas weakly acyclic games are defined in terms graph theoretic properties of a game's better response graph, GenWAGs are defined using a game's *satisficing graph*, introduced in this paper, which contains the game's better response graph as a subgraph. The definition we propose for a game's satisficing graph is based on the concept of *satisficing paths*, first presented in the context of multi-state Markov games in [14]. In that work, a satisficing path is any sequence of strategy profiles for which the strategy of an optimizing agent (that is, an agent whose strategy is a best response to that of its counterparts at a given time) is not altered in the next period. Despite thematic similarities, there are salient differences between the graphs studied here and the paths studied in [14]. In the latter, paths are defined on the set of randomized/mixed strategies, and continuity arguments play an important role in the analysis. By contrast, satisficing graphs are defined here on the set of pure strategies, our analysis centres on discrete objects, and our arguments do not hinge on continuity. Moreover, while [14] provided some sufficient conditions for existence of satisficing paths to equilibrium, necessary conditions were not provided and it was left open whether pathological counterexamples exist.

There are several factors motivating the study of generalized weakly acyclic games. As we show in Theorem 1, GenWAGs arise naturally in the analysis of certain learning-relevant stochastic processes on the set of strategy profiles in a game, with important special cases such as randomized variants of inertial better/best response dynamics. A key insight of this result is that Nash convergence can be guaranteed in a wider class of games by incorporating experimental (possibly suboptimal) strategy revision when failing to best respond, rather than rigidly requiring players to revise their strategies to better or best responses. A second motivation for studying generalized weakly acyclic games is the relative simplicity of verifying sufficient conditions. For instance, verifying the existence of a strict pure Nash equilibrium in a two-player game (Theorem 2) or symmetry conditions in an n-player game ([14, Theorem 3.6]) is often more practical than verifying path connectivity properties of the game's better response graph.

Contributions: This paper presents Generalized Weakly Acyclic Games (GenWAGs). In Theorem 1 we show that GenWAGs coincide exactly with the class of games for which a specific Markov chain converges to a pure Nash equilibrium. In Theorems 2 and 3, we provide sufficient conditions for guaranteeing that a normal-form game is a GenWAG. We also provide the first negative results in the theory of satisficing and demonstrate that our generalization is non-trivial: we provide an example of a game that is not weakly acyclic but is a GenWAG, showing that the generalization is strict, and we provide examples of games that are not GenWAGs but nevertheless admit a pure Nash equilibrium, showing that our class of GenWAGs is a strict subset of the set of games admitting a pure Nash equilibrium. Finally, we identify an open question on sufficient conditions for games with three or more players.

B. Yongacoglu and L. Pavel are with the Department of Electrical and Computer Engineering at the University of Toronto. G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa. S. Yüksel is with the Department of Mathematics and Statistics, Queen's University. Correspondence to bora.yongacoglu@utoronto.ca.

II. Normal-Form Games

Our setting is that of finite n-player normal-form games. An n-player game Γ is described by a triple

$$\Gamma = (n, \mathbf{A}, \mathbf{r}),$$

where n is the number of players, $\mathbf{A} = \mathbb{A}^1 \times \cdots \times \mathbb{A}^n$ is a finite set of strategy profiles (also called action profiles, joint actions, or pure strategies), and $\mathbf{r} = \{r^i\}_{i=1}^n$ is a collection of reward functions, with $r^i : \mathbf{A} \to \mathbb{R}$ being player i's reward function. The i^{th} component of $\mathbf{A}$ is player i's set of actions/pure strategies $\mathbb{A}^i$.

Notation. We use $[n] := \{1, \ldots, n\}$ to denote the set of players. For an element $\mathbf{a} \in \mathbf{A}$, we write $\mathbf{a} = (a^i)_{i \in [n]}$. To isolate the role of player i we write $\mathbf{a} = (a^i, \mathbf{a}^{-i})$, so that $\mathbf{a}^{-i}$ is interpreted as $(a^1, \ldots, a^{i-1}, a^{i+1}, \ldots, a^n)$. In a slight abuse of notation, we write $\mathbf{A} = \mathbb{A}^i \times \mathbf{A}^{-i}$. For a given player $i \in [n]$, we refer to the remaining players in $[n] \setminus \{i\}$ as i's counterplayers or counterparts.

Description of play. Each player $i \in [n]$ selects its own action $a^i \in \mathbb{A}^i$, resulting in an action profile $\mathbf{a} = (a^i)_{i=1}^n$. Once this action profile has been selected, each player $i \in [n]$ receives a reward $r^i(\mathbf{a}) = r^i(a^i, \mathbf{a}^{-i})$. Player i's objective is to maximize its reward $r^i(a^i, \mathbf{a}^{-i})$ by optimizing over its action choice $a^i \in \mathbb{A}^i$. Since player i's objective function depends on the action selections of its counterplayers, we have the following definitions of better and best responding.

Definition 1: For player $i \in [n]$ and an action profile $(a^i, \mathbf{a}^{-i}) \in \mathbf{A}$, an action $a^i_\star \in \mathbb{A}^i$ is called a *better response to* $(a^i, \mathbf{a}^{-i})$ if $r^i(a^i_\star, \mathbf{a}^{-i}) \geq r^i(a^i, \mathbf{a}^{-i})$.

If $r^i(a^i_\star, \mathbf{a}^{-i}) \geq r^i(\bar{a}^i, \mathbf{a}^{-i})$ for any $\bar{a}^i \in \mathbb{A}^i$, then the action $a^i_\star$ is called a *best response to* $\mathbf{a}^{-i}$.

We let $\mathrm{Better}^i(\mathbf{a}) \subseteq \mathbb{A}^i$ denote the subset of player i's pure actions that are better responses to $\mathbf{a} = (a^i, \mathbf{a}^{-i})$, and we let $\mathrm{Best}^i(\mathbf{a}^{-i}) \subseteq \mathbb{A}^i$ denote the subset of player i's pure actions that are best responses to $\mathbf{a}^{-i}$.

For an action profile $\mathbf{a}$ and a player $i \in [n]$, we say that player i is *satisfied* at $\mathbf{a}$ if $a^i \in \mathrm{Best}^i(\mathbf{a}^{-i})$, and otherwise we say that player i is unsatisfied at $\mathbf{a}$. We let $\mathrm{Sat}(\mathbf{a}) \subseteq [n]$ denote the subset of players who are satisfied at an action profile $\mathbf{a}$, and let $\mathrm{UnSat}(\mathbf{a}) \subseteq [n]$ denote the subset of players who are unsatisfied at $\mathbf{a}$.

The solution concept of interest to this paper is the celebrated Nash equilibrium, which captures a situation in which all players are simultaneously best responding to one another.

Definition 2: A strategy profile $\mathbf{a}_\star \in \mathbf{A}$ is called a *(pure) Nash equilibrium* if $a^i_\star \in \mathrm{Best}(\mathbf{a}^{-i}_\star)$ for all $\forall i \in [n]$.

Remark. In this paper, we study only the finite game Γ and *not* its mixed extension. That is, we do not allow any player $i \in [n]$ to select its action a^i randomly according to a mixed strategy (a probability distribution over $\mathbb{A}^i$). Consequently, pure Nash equilibrium need not exist in a general finite normal-form game, even though Nash equilibrium in mixed strategies always exists.

III. Graph Theoretic Structure in Games

A. Best Response Paths and Graphs

To formalize the concepts of better and best response paths mentioned in the introduction, we now introduce the best response graph and better response graph of the game Γ. In what follows, all graphs are directed, and our notational conventions are such: $D = (V, E)$ represents a directed graph, where V represents the finite set of vertices of D and $E \subseteq V \times V$ represents a collection of directed edges, with $(v_1, v_2) \in E$ meaning there is a directed edge from v_1 to v_2 in D.

Definition 3: The (multi-agent) *best response graph* of the game Γ is a directed graph $\mathcal{D}_{\mathrm{Best}}(\Gamma) = (\mathbf{A}, \mathcal{E}_{\mathrm{Best}})$, where, for any $(\mathbf{a}_1, \mathbf{a}_2) \in \mathbf{A} \times \mathbf{A}$, we have $(\mathbf{a}_1, \mathbf{a}_2) \in \mathcal{E}_{\mathrm{Best}}$ if and only if the following conditions hold for each player $i \in [n]$:

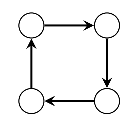
1. $a^i_1 \in \mathrm{Best}^i(\mathbf{a}^{-i}_1) \Rightarrow a^i_2 = a^i_1$, and
2. $a^i_2 \neq a^i_1 \Rightarrow a^i_2 \in \mathrm{Best}^i(\mathbf{a}^{-i}_1)$.

Intuitively, a directed edge from action profile $\mathbf{a}_1$ to $\mathbf{a}_2$ exists when $\mathbf{a}_2$ is obtained by switching the strategies of (at most) players in some subset of $\mathcal{C}_1 \subseteq \mathrm{UnSat}(\mathbf{a}_1)$, and for each such player $i \in \mathcal{C}_1$, a^i_2 belongs to $\mathrm{Best}^i(\mathbf{a}^{-i}_1)$. Note that this construction allows for the actions of several players to be changed simultaneously.

Consider the 2×2 discoordination game in Figure 1a. In this game, Player 1 selects the row, Player 2 selects the column. Player 1 is paid the first quantity in the chosen cell, and Player 2 receives the second quantity. Player 1's best response is to copy the action of Player 2, and Player 2's best response is to mismatch the action of Player 1.

	a	b
A	1, 0	0, 1
B	0, 1	1, 0

(a) Discoordination game



(b) Best response graph

Fig. 1: A discoordination game and its best response graph, with self-loops omitted.

The best response graph of this game is displayed in Figure 1b, with node labels (e.g. (A, a), etc.) and self-loops omitted for visual clarity. For example, there is a directed edge $(B, a) \to (A, a)$, because action $A \in \mathrm{Best}^1(a)$. On the other hand, there is no directed edge from $(B, a) \to (A, b)$, because $a \in \mathrm{Best}^2(B)$ and thus player 2 is satisfied at (B, a).

A (multi-agent) *best response path* in the game Γ is defined as any path in the directed graph $\mathcal{D}_{\mathrm{Best}}(\Gamma)$.

Next, we define the better response graph for the game Γ. This construction is similar to the best response graph, but allows for suboptimal strategy revision when a player is not satisfied.

Definition 4: The (multi-agent) *better response graph* of the game Γ is a directed graph $\mathcal{D}_{\mathrm{Better}}(\Gamma) = (\mathbf{A}, \mathcal{E}_{\mathrm{Better}})$, where $\mathcal{E}_{\mathrm{Better}} \subseteq \mathbf{A} \times \mathbf{A}$ is characterized as follows: for a pair $(\mathbf{a}_1, \mathbf{a}_2) \in \mathbf{A} \times \mathbf{A}$, one has $(\mathbf{a}_1, \mathbf{a}_2) \in \mathcal{E}_{\text{Better}}$ if and only if both of the following hold for each player $i \in [n]$,

1. $a_1^i \in \text{Best}^i(\mathbf{a}_1^{-i}) \Rightarrow a_2^i = a_1^i$, and
2. $a_2^i \neq a_1^i \Rightarrow a_2^i \in \text{Better}^i(\mathbf{a}_1)$.

A multi-agent better response path is defined as a path in the directed graph $\mathcal{D}_{\text{Better}}(\Gamma)$. We note that in the discoordination game of Figure 1a, the better and best response graphs coincide ($\mathcal{D}_{\text{Best}}(\Gamma) = \mathcal{D}_{\text{Better}}(\Gamma)$) but this is not generally the case.

With the preceding definitions in hand, we are now ready to present the definition of weakly acyclic games, which have been defined in several related forms of differing generality [5]–[9].

Definition 5: A game Γ is called *weakly acyclic* if, for any action profile $\mathbf{a} \in \mathbf{A}$, there exists a multi-agent better response path beginning at $\mathbf{a}$ and ending at a pure Nash equilibrium.

In graph theoretic terms, this definition has two main parts. First, the better response graph $\mathcal{D}_{\text{Better}}(\Gamma)$ must possess at least one sink (a node with no outgoing vertices), which corresponds to the existence of pure Nash equilibrium. Second, there must exist a directed path in $\mathcal{D}_{\text{Better}}(\Gamma)$ from any non-sink node to some sink node.

We observe that the discoordination game of Figure 1a is not a weakly acyclic game, since it possesses no pure Nash equilibrium and thus fails the first part of the definition. There are also examples of games that admit pure Nash equilibrium but are nevertheless not weakly acyclic because they fail the second condition on the existence of paths to pure equilibrium. For one such example, consider the game in Figure 2. This game admits a unique pure Nash equilibrium, (T, L), but is not weakly acylic because there are no multi-agent better response paths from (for instance) the initial action profile (M, C) to (T, L).

	L	C	R
T	9, 9	0, 0	0, 0
M	0, 0	2,1	1,2
B	0, 0	1,2	2,1

Fig. 2: A game with a pure Nash equilibrium that is not weakly acyclic

Weakly acyclic games appear in many studies on multi-agent game theoretic learning with distributed and/or decentralized information, e.g. [15]–[17]. One reason for their practical relevance is that weakly acyclic games are the largest class of games for which randomized inertial better response dynamics is guaranteed to converge to Nash equilibrium. That is, suppose an initial action profile $\mathbf{a}_1 \in \mathbf{A}$ is selected arbitrarily, and then for each time $t \geq 1$, the strategy of every player $i \in [n]$ is set according to the following randomized strategy update rule:

$$a_{t+1}^i = \begin{cases} a_t^i, \text{ if } a_t^i \in \text{Best}^i(\mathbf{a}_t^{-i}), \\ a^i \sim \text{Uniform}(\text{Better}^i(\mathbf{a}_t) \cup \{a_t^i\}), \text{ else.} \end{cases}$$

(We have included a_t^i in the latter case as a substitute for a random inertia condition, as discussed in [16], [17].)

Although randomized, distributed inertial better response dynamics converge in weakly acyclic games and have other desirable qualities, such as being individually rational, one obvious shortcoming is that such algorithms do not lead to Nash equilibrium strategies in games lacking the weakly acyclic structure. This remains true even when some strategy profiles are Pareto optimal, such as the Nash equilibrium (T, L) in the game of Figure 2. Thus, beyond weakly acyclic games, one must rely on different graph theoretic structure when designing Nash-seeking algorithms. This deficiency of the weakly acyclic structure is addressed in the next sections.

B. Satisficing Paths and Graphs

Having established the need to identify structure beyond that of weakly acyclic games, we now present the notion of the *satisficing graph* of a game. The contents of this section are thematically related to the notion of satisficing paths presented in [14], but here we consider only pure strategies and define explicit graphs for the first time.

Definition 6: The (multi-agent) *satisficing graph* of the game Γ is a directed graph $\mathcal{D}_{\text{Sat}}(\Gamma) = (\mathbf{A}, \mathcal{E}_{\text{Sat}})$, where, for any $(\mathbf{a}_1, \mathbf{a}_2) \in \mathbf{A} \times \mathbf{A}$, we have $(\mathbf{a}_1, \mathbf{a}_2) \in \mathcal{E}_{\text{Sat}}$ if and only if the following conditions hold for each player $i \in [n]$:

1. $a_1^i \in \text{Best}^i(\mathbf{a}_1^{-i}) \Rightarrow a_2^i = a_1^i$, and
2. $a_2^i \neq a_1^i \Rightarrow a_2^i \in \mathbb{A}^i$.

The second condition in the definition above is redundant, but was included to elucidate the successive generalization from the best response graph (Definition 3) to the better response graph (Definition 4) and then from the better response graph to the satisficing graph (Definition 6) of a game: since $\text{Best}^i(\mathbf{a}_1^{-i}) \subseteq \text{Better}^i(\mathbf{a}_1) \subseteq \mathbb{A}^i$, we immediately have

$$\mathcal{E}_{\text{Best}} \subseteq \mathcal{E}_{\text{Better}} \subseteq \mathcal{E}_{\text{Sat}}. \tag{1}$$

Intuitively, a directed edge from action profile $\mathbf{a}_1$ to action profile $\mathbf{a}_2$ exists when $\mathbf{a}_2$ is obtained by switching the strategies of (at most) players in some subset $\bar{\mathcal{C}}_1 \subseteq \text{UnSat}(\mathbf{a}_1)$, and for each player $i \in \bar{\mathcal{C}}_1$, the action a_2^i can take any value. In particular, a_2^i need not be a better or best response to $\mathbf{a}_2^{-i}$.

A *(pure) satisficing path* for the game Γ is defined as any directed path in $\mathcal{D}_{\text{Sat}}$. From (1), one has that any best response path (or any better response path) is also a satisficing path, but the reverse is not generally true.

Definition 7: A game Γ is called a *generalized weakly acyclic game* (GenWAG) if, for any action profile $\mathbf{a} \in \mathbf{A}$, there exists a satisficing path beginning at $\mathbf{a}$ and ending at a pure Nash equilibrium.

We now state some simple but useful results, which show that GenWAGs are a meaningful generalization of weakly acyclic games. First, in Lemma 1, we observe that all weakly acyclic games are generalized weakly acyclic, but the converse does not typically hold. Then, we provide an example of a game that admits pure Nash equilibrium but is not generalized weakly acyclic.

Lemma 1: If a game Γ is weakly acyclic, then it is also generalized weakly acyclic. Moreover, there are games that are generalized weakly acyclic but not weakly acyclic.

Proof. The first part follows from the fact that any multi-agent better response path is automatically a pure satisficing path. For the second part, we observe that the game presented in Figure 2 is not weakly acyclic but is indeed generalized weakly acyclic. $\diamond$

Since $\mathcal{E}_{\mathrm{Sat}}$ is defined by a rather weak constraint, it is natural to ask whether *all* games with pure Nash equilibrium are generalized weakly acyclic. We now present an example to show that the class of generalized weakly acyclic games is not trivial. That is, we show that existence of pure Nash equilibrium is not sufficient for a game to be generalized weakly acyclic. Consider the game in Figure 3, below.

	L	C	R
T	1, 1	0, 1	0, 1
M	1, 0	1,0	0,1
B	1, 0	0,1	1,0

Fig. 3: A game that is not generalized weakly acyclic

In this game, the unique pure Nash equilibrium is (T, L), but the game does not admit pure satisficing paths to (T, L) from any of the strategy profiles (M, C), (M, R), (B, C), or (B, R).

The qualitative difference here between the games of Figure 2 and Figure 3 has to do with indifference. In the game of Figure 3, beginning at one of the strategy profiles (M, C), (M, R), (B, C), or (B, R), reaching pure Nash equilibrium (T, L) requires switching the actions of both players. However, in each of these action profiles, there is exactly one satisfied player and one unsatisfied player. Upon switching the action of the unsatisfied player, one of two situations arises: either (1) the satisfied player remains satisfied, and is not compelled to change its behavior, but the unsatisfied player remains unsatisfied; or (2) the unsatisfied player becomes satisfied at the new action profile while the previously satisfied player becomes unsatisfied. In any case, there are no directed edges with tail in $\mathbf{A} \setminus \{(T, L)\}$ and head (T, L).

IV. GenWAGs and Satisficing Markov Chains

In this section, we define *satisficing Markov chains* and we characterize generalized weakly acyclic games as being exactly those games for which a satisficing Markov chain eventually converges to pure Nash equilibrium.

Definition 8: For a game Γ and an action profile $\mathbf{a} \in \mathbf{A}$, a *satisficing Markov chain* beginning at $\mathbf{a}$ is a Markov chain $\{\mathbf{a}_t\}_{t=1}^{\infty}$ on $\mathbf{A}$ with $\mathbb{P}(\mathbf{a}_1 = \mathbf{a}) = 1$ and satisfying the following evolution rule for each $t \geq 1$:

$$a^i_{t+1} = \begin{cases} a^i_t, & \text{if } a^i_t \in \mathrm{Best}^i(\mathbf{a}^{-i}_t) \\ a^i \sim \mathrm{Uniform}(\mathbb{A}^i), & \text{else,} \end{cases} \tag{2}$$

where the collection $\{a^i_{t+1}\}_{i \in [n]}$ is jointly conditionally independent given $\mathbf{a}_t$.

Definition 8 is motivated by a family of closely related game theoretic learning algorithms [14], [15], [18]–[26]. In these learning algorithms, each player i has a baseline action a^i_t that it periodically revises to a^i_{t+1}.[1] Learning phases occur between baseline action revision times, and in a learning phase player i may experiment with non-baseline actions and use the resulting reward observations to evaluate the performance of its various alternative actions, $a \in \mathbb{A}^i$. At the end of such a learning phase, each player i revises its baseline action from a^i_t to a^i_{t+1}. Although the particular revision mechanism may vary, it is typically characterized by a "win stay, lose shift" condition similar to that of (2), whereby unsatisfied agents consider randomized experimental action revision and satisfied agents continue using their previous baseline action.[2] A number of Nash-seeking algorithms and multi-agent learning algorithms from this family, including [14], [18], [20] and [27], have been studied by first analyzing the convergence properties of a Markov chain $\{\mathbf{a}_t\}_{t=1}^{\infty}$ and then using $\{\mathbf{a}_t\}_{t=1}^{\infty}$ to approximate a sequence of learned strategy iterates $\{\widehat{\mathbf{a}}_t\}_{t=1}^{\infty}$.

We now review some Markov chain terminology as it relates to Definition 8. For action profiles $\mathbf{a}, \tilde{\mathbf{a}} \in \mathbf{A}$, $\tilde{\mathbf{a}}$ is said to be *accessible* from $\mathbf{a}$ if there exists a positive integer $m = m(\mathbf{a}, \tilde{\mathbf{a}})$ such that $\mathbb{P}(\mathbf{a}_{m+1} = \tilde{\mathbf{a}} | \mathbf{a}_1 = \mathbf{a}) > 0$. That is, the satisficing Markov chain beginning at $\mathbf{a}$ transits to $\tilde{\mathbf{a}}$ in finitely many steps with non-zero probability. The action profiles $\mathbf{a}$ and $\tilde{\mathbf{a}}$ are said to *communicate* if $\mathbf{a}$ is accessible from $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{a}}$ is accessible from $\mathbf{a}$. It is easy to see that communication defined this way allows one to partition $\mathbf{A}$ into equivalence classes called *communicating classes*, where $\mathbf{a}$ and $\tilde{\mathbf{a}}$ belong to the same communicating class if and only if $\mathbf{a}$ and $\tilde{\mathbf{a}}$ communicate. A communicating class $\mathbf{A}' \subseteq \mathbf{A}$ is called *absorbing* if the probability of transiting out of $\mathbf{A}'$ is zero when the Markov chain starts in $\mathbf{A}'$. That is, $\mathbf{A}'$ is absorbing if

$$\mathbb{P}\left(\bigcup_{t \geq 1} \{\mathbf{a}_t \notin \mathbf{A}'\} \middle| \mathbf{a}_1 \in \mathbf{A}' \right) = 0.$$

An action profile $\mathbf{a}_\dagger \in \mathbf{A}$ is said to be absorbing if the singleton $\{\mathbf{a}_\dagger\}$ is an absorbing communicating class. Before presenting the main result of this section, we prove that an action profile $\mathbf{a}_\dagger$ is absorbing for the satisficing Markov chain if and only if it is a pure Nash equilibrium of the game.

Lemma 2: For a game Γ and an action profile $\mathbf{a}_\dagger$, we have that $\mathbf{a}_\dagger$ is absorbing for the satisficing Markov chain $\{\mathbf{a}_t\}_{t=1}^{\infty}$ if and only if $\mathbf{a}_\dagger$ is a Nash equilibrium of Γ.

[1] The subscripts t and $t+1$ here denote successive revision times, which usually do not correspond to successive periods of actual play of the game.

[2] The definition of "winning" is handled differently by different algorithms. In some cases, such as [23] or [24], players maintain a *mood variable* to guide their revisions. In others, such as [22] and [26], players set a scalar *aspiration level* and winning is defined as exceeding this aspiration level. In still others, such as [14] or [18], the best response condition of (2) is used as the winning condition.

Proof. Suppose $\mathbf{a}_\dagger \in \mathbf{A}$ is a Nash equilibrium of Γ. By the best response condition of (2), one has

$$\mathbb{P}(a_2^i = a_\dagger^i | \mathbf{a}_1 = \mathbf{a}_\dagger) = 1, \, \forall i \in [n].$$

Thus, $\mathbb{P}(\mathbf{a}_2 = \mathbf{a}_\dagger | \mathbf{a}_1 = \mathbf{a}_\dagger) = 1^n = 1$, which shows that $\{\mathbf{a}_\dagger\}$ is absorbing.

For the reverse direction, suppose $\{\mathbf{a}_\dagger\}$ is absorbing. Then, by (2), the best response condition holds for each player. (Otherwise there exists a player whose action is switched with non-zero probability, which would contradict the fact that $\mathbf{a}_\dagger$ is absorbing.) Since all players are best responding, $\mathbf{a}_\dagger$ is a Nash equilibrium. $\diamond$

We now present the first of our main results, which says that GenWAGs are the largest class of games for which satisficing Markov chains are guaranteed to converge to Nash equilibrium.

Theorem 1: A game Γ is generalized weakly acyclic if and only if for every $\mathbf{a} \in \mathbf{A}$, the satisficing Markov chain beginning at $\mathbf{a}$ converges to a pure Nash equilibrium almost surely.

Proof. ($\Rightarrow$) To prove the forward direction, let $\mathbf{a} \in \mathbf{A}$ be arbitrary, suppose $\{\mathbf{a}_t\}_{t=1}^\infty$ is a satisficing Markov chain beginning at $\mathbf{a}$, and suppose Γ is generalized weakly acyclic.

Since Γ is a GenWAG, there exists a satisficing path connecting any initial point $\tilde{\mathbf{a}} \in \mathbf{A}$ to a Nash equilibrium. Let $T(\tilde{\mathbf{a}}) \in \mathbb{N}$ be the length of a shortest such path beginning at $\tilde{\mathbf{a}}$, and let $\rho(\tilde{\mathbf{a}}) > 0$ be the probability that the satisficing Markov chain follows such a path when initialized at $\tilde{\mathbf{a}}$. Since $\mathbf{A}$ is finite, we have

$$\max_{\tilde{\mathbf{a}} \in \mathbf{A}} T(\tilde{\mathbf{a}}) =: \tau < \infty, \text{ and } \min_{\tilde{\mathbf{a}} \in \mathbf{A}} \rho(\tilde{\mathbf{a}}) =: \rho > 0$$

By Lemma 2, any pure Nash equilibrium is a sink in the satisficing graph and thus corresponds to an absorbing state for the satisficing Markov chain.

From the preceding discussion, one obtains the following inequality:

$$\mathbb{P}\left(\mathbf{a}_{t+\tau} \notin \text{Nash} | \mathbf{a}_t = \tilde{\mathbf{a}}\right) \leq (1 - \rho),$$

for any $\tilde{\mathbf{a}} \in \mathbf{A}$ and any $t \geq 1$, where Nash denotes the set of pure Nash equilibria of Γ. Recursively, one then obtains

$$\mathbb{P}\left(\mathbf{a}_{t+m\tau} \notin \text{Nash} | \mathbf{a}_t = \tilde{\mathbf{a}}\right) \leq (1 - \rho)^m,$$

for any $t, m \geq 1$ and $\tilde{\mathbf{a}} \in \mathbf{A}$. Taking $m \to \infty$, one sees that $\{\mathbf{a}_t\}_{t=1}^\infty$ converges to some pure Nash equilibrium almost surely for any initial condition $\mathbf{a}_1 = \mathbf{a}$.

($\Leftarrow$) We argue the backward direction by contrapositive. That is, we will argue that if Γ is not a GenWAG, then there exists some initial condition $\mathbf{a}_1 = \mathbf{a} \in \mathbf{A}$ such that the satisficing Markov chain will fail to converge almost surely to a Nash equilibrium.

Indeed, if Γ is not a GenWAG, then there is some action profile $\mathbf{a} \in \mathbf{A}$ for which there are no satisficing paths beginning at $\mathbf{a}$ and terminating at a pure Nash equilibrium of the game Γ. This implies that for the satisficing Markov chain, the communicating class of $\mathbf{a}$ contains no Nash equilibrium strategy profiles, and the satisficing Markov chain beginning at $\mathbf{a}_1 = \mathbf{a}$ will avoid Nash equilibrium strategy profiles at all times. $\diamond$

V. A Sufficient Condition, a Counterexample, and a Conjecture

We previously encountered an example of a game, shown in Figure 3 of Section III, that possessed pure Nash equilibrium but was not generalized weakly acyclic. This showed that existence of pure Nash equilibrium alone is not sufficient for a game to be generalized weakly acyclic, and constitutes the first negative example in the theory on satisficing paths and graphs. In this section, we present sufficient conditions to complement that observation. We also identify a conjecture as an open question for future research.

A. A Sufficient Condition for Two-Player Games

Definition 9: For a player i and an action profile $\mathbf{a}^{-i}$, an action $a_\star^i \in \mathbb{A}^i$ is a *strict best response* to $\mathbf{a}^{-i}$ if

$$r^i(a_\star^i, \mathbf{a}^{-i}) > r^i(a^i, \mathbf{a}^{-i}), \quad \forall a^i \neq a_\star^i.$$

Definition 10: A pure Nash equilibrium $\mathbf{a}_\star$ is *strict* if for each $i \in [n]$, $a_\star^i$ is a strict best response to $\mathbf{a}_\star^{-i}$.

Theorem 2: Let Γ be a two-player game and suppose Γ admits a strict pure Nash equilibrium $\mathbf{a}_\star$. Then, Γ is generalized weakly acyclic.

Proof. Let $\mathbf{a}_1 = \mathbf{a} \in \mathbf{A}$ be our initial strategy profile. We proceed in cases.

Case 1: If $\text{UnSat}(\mathbf{a}) = \varnothing$, then $\mathbf{a}$ is a pure Nash equilibrium itself and so the path $(\mathbf{a})$ is a satisficing path beginning at $\mathbf{a}$ and terminating at pure Nash equilibrium.

Case 2: If $\text{UnSat}(\mathbf{a}) = \{1, 2\}$, then neither player is satisfied at $\mathbf{a}$, and so $(\mathbf{a}, \mathbf{a}') \in \mathcal{E}_{\text{Sat}}$ is a valid satisficing path for any $\mathbf{a}' \in \mathbf{A}$, since (vacuously) it does not change the action of any player that was previously best responding. Selecting $\mathbf{a}' = \mathbf{a}_\star$, one obtains a satisficing path from $\mathbf{a}$ to a pure Nash equilibrium of Γ.

Case 3: $\text{UnSat}(\mathbf{a}) = \{i\}$ for some $i \in \{1, 2\}$. That is, exactly one player, i, is unsatisfied at $\mathbf{a}$, while the other player, $j \neq i$, is satisfied at $\mathbf{a}$. We put $\mathbf{a}_2 = (a_\star^i, a^j)$ and note $(\mathbf{a}_1, \mathbf{a}_2) \in \mathcal{E}_{\text{Sat}}$.

If $a^j = a_\star^j$, then we have constructed our path and there is nothing left to show, since $\mathbf{a}_2 = \mathbf{a}_\star$ is a pure Nash equilibrium. Otherwise, if $a^j \neq a_\star^j$, then since $\mathbf{a}_\star$ is a strict Nash equilibrium, one has that $a^j \notin \text{Best}^j(a_\star^i)$. Then, we put $\mathbf{a}_3 = (a_\star^i, a_\star^j) = \mathbf{a}_\star$, and note that $(\mathbf{a}_2, \mathbf{a}_3) \in \mathcal{E}_{\text{Sat}}$, completing the proof. $\diamond$

B. A Sufficient Condition for n-Player Games

To state the results of this section, we require some additional notation. For a (possibly empty) player subset $\mathcal{N} \subseteq [n]$, we denote a *partial* action profile $\mathbf{a}^{\mathcal{N}} = (a^i)_{i \in \mathcal{N}}$ as the action profile $\mathbf{a}$ with components for players in the subset $\mathcal{N}$. We write $\mathbf{a}^{-\mathcal{N}}$ to denote $\mathbf{a}^{[n] \setminus \mathcal{N}}$, and we then have $\mathbf{a} = (\mathbf{a}^{\mathcal{N}}, \mathbf{a}^{-\mathcal{N}})$. We let $\mathbf{A}^{\mathcal{N}} = \times_{i \in \mathcal{N}} \mathbb{A}^i$.

Definition 11: Let $\Gamma = (n, \mathbf{A}, \mathbf{r})$ be an n-player game. Let $\mathcal{N} \subseteq [n]$ be an m-player subset, with $0 \leq m \leq n$, and let $\mathbf{a}^{-\mathcal{N}}$ be a partial action profile. The *subgame induced by* $\mathbf{a}^{-\mathcal{N}}$ is an m-player game $\Gamma_\dagger$,

$$\Gamma_\dagger = \left(m,\ \mathbf{A}_\dagger = \mathbf{A}^{\mathcal{N}},\ \mathbf{r}_\dagger\right),$$

with player set $\mathcal{N}$, action sets $\mathbb{A}^i_\dagger = \mathbb{A}^i$ for $i \in \mathcal{N}$, and the following reward functions for each $i \in \mathcal{N}$:

$$r^i_\dagger\left(\mathbf{a}^{\mathcal{N}}\right) = r^i\left(\mathbf{a}^{\mathcal{N}}, \mathbf{a}^{-\mathcal{N}}\right), \quad \forall \mathbf{a}^{\mathcal{N}}.$$

Intuitively, the subgame induced by $\mathbf{a}^{-\mathcal{N}}$ is the game that one obtains when the actions of players in $[n]\setminus\mathcal{N}$ are fixed at $\mathbf{a}^{-\mathcal{N}}$. The players of $[n]\setminus\mathcal{N}$ become fixed aspects of the environment, and the remaining players in $\mathcal{N}$ play a smaller game. We say that $\Gamma_\dagger$ is an induced subgame of Γ if there exists a player subset $\mathcal{N}$ and a partial action profile $\mathbf{a}^{-\mathcal{N}}$ for which $\Gamma_\dagger$ is the subgame induced by $\mathbf{a}^{-\mathcal{N}}$.

Note: In the definitions above, we allow $\mathcal{N}$ to be empty and we also allow $\mathcal{N} = [n]$. Thus, a game Γ is always an induced subgame of itself.

Theorem 3: Let Γ be an n-player game. Suppose that for any induced subgame $\Gamma_\dagger$ of Γ, the induced subgame $\Gamma_\dagger$ admits a unique pure Nash equilibrium and this Nash equilibrium is strict. Then, Γ is generalized weakly acyclic.

For brevity, we say that a game has *the induced subgame property* (ISP) if it satisfies the condition appearing Theorem 3. That is, a game has the ISP if any induced subgame admits a unique pure Nash equilibrium and this Nash equilibrium is strict. We also remark that if a given game has the induced subgame property, then any induced subgame also has the ISP.

Proof. We prove this theorem by induction on the number of players. The base case of $m = 2$ players is a special case of Theorem 2. Our induction hypothesis is such: for some $m \geq 2$, if $\Gamma_\dagger$ is an m-player game that has the ISP, then $\Gamma_\dagger$ is generalized weakly acyclic.

Now let Γ be an n-player game that has the ISP, and suppose $n = m + 1$. We will argue that our induction hypothesis implies that Γ is generalized weakly acyclic.

Let $\mathbf{a}_1 = \mathbf{a} \in \mathbf{A}$ be an arbitrary initial action profile for the game Γ. We prove the result by showing that there exists a satisficing path from $\mathbf{a}_1$ to the unique Nash equilibrium of Γ, which is strict and which we denote by $\mathbf{a}_\star \in \mathbf{A}$.

We begin by ruling out trivial cases. If $\mathrm{UnSat}(\mathbf{a}_1) = [n]$, then for any $\mathbf{a}'$ we have that $(\mathbf{a}_1, \mathbf{a}')$ is a valid edge in $\mathcal{E}_{\mathrm{Sat}}(\Gamma)$. We may then take $\mathbf{a}' = \mathbf{a}_\star$ to be the unique Nash equilibrium of Γ, and the proof is complete in this case. Thus, we focus on the case where $\mathrm{Sat}(\mathbf{a}_1) \neq \varnothing$, and there exists some player i who is satisfied at $\mathbf{a}_1$: $a^i_1 \in \mathrm{Best}^i(\mathbf{a}^{-i}_1)$.

Recalling that $m = n - 1$, consider the $(n-1)$-player subgame $\Gamma_\dagger$ induced by fixing the action of player i to be a^i_1. Since Γ has the ISP and $\Gamma_\dagger$ is an induced subgame of Γ, it follows that $\Gamma_\dagger$ also has the ISP and, by the induction hypothesis, that $\Gamma_\dagger$ is generalized weakly acyclic. Thus, for any $(n-1)$-player strategy profile $\tilde{\mathbf{a}}^{-i}_1$ of the game $\Gamma_\dagger$, there exists a directed path in the graph $\mathcal{D}_{\mathrm{Sat}}(\Gamma_\dagger)$ from $\tilde{\mathbf{a}}^{-i}_1$ to the unique Nash equilibrium of the $(n-1)$-player game $\Gamma_\dagger$. Denote such a path by $\tilde{\mathbf{a}}^{-i}_1, \cdots, \tilde{\mathbf{a}}^{-i}_k$, where each $\tilde{\mathbf{a}}^{-i}_t \in \mathbf{A}^{-i}$, and note that $\tilde{\mathbf{a}}^{-i}_k$ is a strict Nash equilibrium for the subgame $\Gamma_\dagger$.

Using the $(n-1)$-player joint actions $\tilde{\mathbf{a}}^{-i}_1, \ldots, \tilde{\mathbf{a}}^{-i}_k$, we construct n-player strategy profiles for Γ by fixing player i's component at a^i_1:

$$\mathbf{a}_t = (a^i_1, \tilde{\mathbf{a}}^{-i}_t), \quad \forall\, t \leq k.$$

Since player i's action is fixed at $a^i_t = a^i_1$ for $t \leq k$, the sequence $(\mathbf{a}_t)^k_{t=1}$ is a valid satisficing path of the original game Γ (i.e., in the directed graph $\mathcal{D}_{\mathrm{Sat}}(\Gamma)$). Depending on whether or not $\mathbf{a}_k$ is a Nash equilibrium for the game Γ, we again proceed in cases to complete the proof.

Since $\tilde{\mathbf{a}}^{-i}_k$ was selected to be the Nash equilibrium of the induced subgame $\Gamma_\dagger$, for each player $j \neq i$, it holds that $a^j_k \in \mathrm{Best}^j(\mathbf{a}^{-j}_k)$. That is, player j is satisfied at $\mathbf{a}^{-i}_k$. If player i is also satisfied at $\mathbf{a}_k$, then $\mathbf{a}_k$ is the Nash equilibrium of Γ and the proof is complete.

If, on the other hand, player i is not satisfied at $\mathbf{a}_k$ (that is, $a^i_1 \notin \mathrm{Best}^i(\mathbf{a}^{-i}_k)$), then we define

$$\mathbf{a}_{k+1} := (a^i_\star, \mathbf{a}^{-i}_k).$$

In other words, we change the action of player i from a^i_1 to $a^i_\star$, its component of the unique Nash equilibrium of Γ.

Let $\Gamma^\star_\dagger$ denote the $(n-1)$-player subgame of Γ induced by fixing player i's action at $a^i_\star$. Since Γ has the ISP, $\Gamma^\star_\dagger$ admits a unique pure Nash equilibrium, which is readily verified as being $\mathbf{a}^{-i}_\star = (a^j_\star)_{j\neq i}$, the action profile in which players $j \neq i$ play their components of $\mathbf{a}_\star$, the unique Nash equilibrium of Γ. Moreover, by our induction hypothesis, $\Gamma^\star_\dagger$ is generalized weakly acyclic. Thus, from any initial action profile $\ddot{\mathbf{a}}^{-i}_1$ in the $(n-1)$-player game $\Gamma^\star_\dagger$, there exists a satisficing path $\ddot{\mathbf{a}}^{-i}_1, \ldots, \ddot{\mathbf{a}}^{-i}_L$ in the directed graph $\mathcal{D}_{\mathrm{Sat}}(\Gamma^\star_\dagger)$ beginning at $\ddot{\mathbf{a}}^{-i}_1$ and terminating at $\ddot{\mathbf{a}}^{-i}_L := \mathbf{a}^{-i}_\star$, the unique pure Nash equilibrium of $\Gamma^\star_\dagger$.

Using $\ddot{\mathbf{a}}^{-i}_1 := \mathbf{a}^{-i}_{k+1} = \mathbf{a}^{-i}_k$ in the discussion above, we define n-player strategy profiles $(\mathbf{a}_{k+t})^L_{t=1}$ as

$$\mathbf{a}_{k+t} = \left(a^i_\star, \ddot{\mathbf{a}}^{-i}_t\right), \quad t = 1, 2, \ldots, L.$$

Since the component of player i is fixed at $a^i_{k+t} = a^i_{k+1} = a^i_\star$, the sequence $(\mathbf{a}_{k+t})^L_{t=1}$ is a satisficing path of the original game Γ. Furthermore, the extended sequence $(\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+L})$ is also a satisficing path of the game Γ. This concludes the proof, since $\mathbf{a}_\star = \mathbf{a}_{k+L}$ is the unique pure Nash equilibrium of Γ. $\diamond$

C. An Open Question for Future Work

In the preceding sections, we encountered an example of a two-player game (Figure 3) in which indifference posed a problem for generalized weak acyclicity. In that example, a (unique) non-strict pure Nash equilibrium existed, but this equilibrium was not accessible from all starting action profiles because players had non-unique best responses and could not be compelled to switch actions.

While indifference may threaten generalized weak acyclicity, we also encountered sufficient conditions for generalized weak acyclicity based on non-indifference. The strictness hypotheses of Theorems 2 and 3 assume the existence of equilibrium action profiles in which players are not indifferent between alternative actions but instead strictly prefer one alternative to the rest.

In the case of two-player games, existence of a strict pure Nash equilibrium was sufficient for generalized weak acyclicity. Of note, uniqueness of Nash equilibrium was not assumed in the two-player case of Theorem 2. On the other hand, our multi-player result, Theorem 3, assumes the existence and uniqueness of a pure Nash equilibrium and further that such an equilibrium is strict. It is natural to ask whether this uniqueness assumption can be relaxed while still preserving generalized weak acyclicity, or whether a pathological counterexample exists. We conclude the technical part of this paper by offering a conjecture as an open question for future research.

Conjecture. *Let Γ be an n-player game. Suppose that for any induced subgame $\Gamma_\dagger$ of Γ we have that $\Gamma_\dagger$ admits a strict pure Nash equilibrium. Then, the game Γ is generalized weakly acyclic.*

VI. Conclusion and Future Directions

We have presented GenWAGs, a generalization of the class of weakly acyclic games. We have demonstrated by example that this generalization is non-trivial, in the sense that there are examples of games that are GenWAGs but not weakly acyclic, and furthermore the class of GenWAGs does not include all games with pure Nash equilibrium. We have argued that the class of GenWAGs is practically relevant for exploratory multi-agent learning applications because of its connection to the satisficing Markov chain, and we have provided multiple sufficient conditions for a game to be generalized weakly acyclic.

As an open question for future work, we ask whether the existence of a pure Nash equilibrium that is strict is a sufficient condition for a game to be generalized weakly acyclic when there are at least three players.

References

[1] P. Frihauf, M. Krstic, and T. Basar, "Nash equilibrium seeking in noncooperative games," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1192–1207, 2011.

[2] F. Salehisadaghiani and L. Pavel, "Distributed Nash equilibrium seeking: A gossip-based algorithm," *Automatica*, vol. 72, pp. 209–216, 2016.

[3] M. Ye and G. Hu, "Distributed Nash equilibrium seeking by a consensus based approach," *IEEE Transactions on Automatic Control*, vol. 62, no. 9, pp. 4811–4818, 2017.

[4] D. Gadjov and L. Pavel, "A passivity-based approach to Nash equilibrium seeking over networks," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1077–1092, 2018.

[5] H. P. Young, "The evolution of conventions," *Econometrica: Journal of the Econometric Society*, pp. 57–84, 1993.

[6] I. Milchtaich, "Congestion games with player-specific payoff functions," *Games and economic behavior*, vol. 13, no. 1, pp. 111–124, 1996.

[7] H. P. Young, *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press, 1998.

[8] H. P. Young, *Strategic learning and its limits*. OUP Oxford, 2004.

[9] A. Fabrikant, A. D. Jaggard, and M. Schapira, "On the structure of weakly acyclic games," *Theory of Computing Systems*, vol. 53, no. 1, pp. 107–122, 2013.

[10] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[11] J. R. Marden, G. Arslan, and J. S. Shamma, "Cooperative control and potential games," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 6, pp. 1393–1407, 2009.

[12] N. Li and J. R. Marden, "Designing games for distributed optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 230–242, 2013.

[13] R. Gopalakrishnan, J. R. Marden, and A. Wierman, "Potential games are necessary to ensure pure Nash equilibria in cost sharing games," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1252–1296, 2014.

[14] B. Yongacoglu, G. Arslan, and S. Yüksel, "Satisficing paths and independent multiagent reinforcement learning in stochastic games," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 3, pp. 745–773, 2023.

[15] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.

[16] G. Arslan and S. Yüksel, "Decentralized Q-learning for stochastic teams and games," *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2016.

[17] B. Swenson, C. Eksin, S. Kar, and A. Ribeiro, "Distributed inertial best-response dynamics," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4294–4300, 2018.

[18] D. Foster and H. P. Young, "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, vol. 1, no. 3, pp. 341–367, 2006.

[19] F. Germano and G. Lugosi, "Global Nash convergence of Foster and Young's regret testing," *Games and Economic Behavior*, vol. 60, no. 1, pp. 135–154, 2007.

[20] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, no. 2, pp. 626–643, 2009.

[21] B. S. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," *Games and Economic behavior*, vol. 75, no. 2, pp. 882–897, 2012.

[22] G. C. Chasparis, A. Arapostathis, and J. S. Shamma, "Aspiration learning in coordination games," *SIAM Journal on Control and Optimization*, vol. 51, no. 1, pp. 465–490, 2013.

[23] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.

[24] J. R. Marden, "Selecting efficient correlated equilibria through distributed learning," *Games and Economic Behavior*, vol. 106, pp. 114–133, 2017.

[25] Z. Hu, M. Zhu, P. Chen, and P. Liu, "On convergence rates of game theoretic reinforcement learning algorithms," *Automatica*, vol. 104, pp. 90–101, 2019.

[26] B. Yongacoglu, G. Arslan, and S. Yüksel, "Decentralized learning for optimality in stochastic dynamic teams and games with local control and global state information," *IEEE Transactions on Automatic Control*, vol. 67, no. 10, pp. 5230–5245, 2022.

[27] J. Gaveau, C. J. Le Martret, and M. Assaad, "Performance analysis of trial and error algorithms," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 6, pp. 1343–1356, 2020.