SIAM J. CONTROL OPTIM.
Vol. 48, No. 1, pp. 373–396

PAYOFF-BASED DYNAMICS FOR MULTIPLAYER WEAKLY ACYCLIC GAMES[*]

JASON R. MARDEN[†], H. PEYTON YOUNG[‡], GÜRDAL ARSLAN[§], AND JEFF S. SHAMMA[¶]

Abstract. We consider repeated multiplayer games in which players repeatedly and simultaneously choose strategies from a finite set of available strategies according to some strategy adjustment process. We focus on the specific class of weakly acyclic games, which is particularly relevant for multiagent cooperative control problems. A strategy adjustment process determines how players select their strategies at any stage as a function of the information gathered over previous stages. Of particular interest are "payoff-based" processes in which, at any stage, players know only their own actions and (noise corrupted) payoffs from previous stages. In particular, players do not know the actions taken by other players and do not know the structural form of payoff functions. We introduce three different payoff-based processes for increasingly general scenarios and prove that, after a sufficiently large number of stages, player actions constitute a Nash equilibrium at any stage with arbitrarily high probability. We also show how to modify player utility functions through tolls and incentives in so-called congestion games, a special class of weakly acyclic games, to guarantee that a centralized objective can be realized as a Nash equilibrium. We illustrate the methods with a simulation of distributed routing over a network.

Key words. game theory, cooperative control, learning in games

AMS subject classifications. 91A10, 91A80, 68W15

DOI. 10.1137/070680199

1. Introduction.

The objective in distributed cooperative control for multiagent systems is to enable a collection of "self-interested" agents to achieve a desirable "collective" objective. There are two overriding challenges to achieving this objective. The first is complexity. Finding an optimal solution by a centralized algorithm may be prohibitively difficult when there are large numbers of interacting agents. This motivates the use of adaptive methods that enable agents to "self-organize" into suitable, if not optimal, collective solutions.

The second challenge is limited information. Agents may have limited knowledge about the status of other agents, except perhaps for a small subset of "neighboring" agents. An example is collective motion control for mobile sensor platforms (see, e.g., [7]). In these problems, mobile sensors seek to position themselves to achieve various collective objectives such as rendezvous or area coverage. Sensors can communicate with neighboring sensors, but otherwise they do not have global knowledge of the domain of operation or the status and locations of nonneighboring sensors.

A typical assumption is that agents are endowed with a reward or utility function

[*] Received by the editors January 15, 2007; accepted for publication (in revised form) October 14, 2008; published electronically February 11, 2009. This research was supported by NSF grant ECS-0501394, ARO grant W911NF–04–1–0316, and AFOSR grant FA9550-05-1-0239.
http://www.siam.org/journals/sicon/48-1/68019.html

[†] Information Science and Technology, California Institute of Technology, Pasadena, CA 91125 (marden@caltech.edu).

[‡] Department of Economics, University of Oxford and the Brookings Institute, Oxford OX1 3UQ, UK (pyoung@brookings.edu).

[§] Department of Electrical Engineering, University of Hawaii, Honolulu, HI 96822 (gurdal@hawaii.edu).

[¶] Corresponding author. School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (shamma@gatech.edu).

that depends on their own strategies and the strategies of other agents. In motion coordination problems, for example, an agent's utility function typically depends on its position relative to other agents or environmental targets, and knowledge of this function guides local motion adjustments.

In other situations, agents may know nothing about the structure of their utility functions and how their own utility depends on the actions of other agents (whether local or far away). In this case, the only course of action is to observe rewards based on experience and "optimize" on a trial and error basis. The situation is further complicated because all agents are trying simultaneously to optimize their own strategies. Therefore, even in the absence of noise, an agent trying the same strategy twice may see different results because of the nonstationary nature of the strategies of other agents.

There are several examples of multiagent systems that illustrate this situation. In distributed routing for ad hoc data networks (see, e.g., [2]), routing nodes seek to route packets to neighboring nodes based on packet destinations without knowledge of the overall network structure. The objective is to minimize the delay of packets to their destinations. This delay must be realized through trial and error, since the functional dependence of delay on routing strategies is not known. A similar problem is automotive traffic routing, in which drivers seek to minimize the congestion experienced to reach a desired destination. Drivers can experience the congestion on selected routes as a function of the routes selected by other drivers, but drivers do not know the structure of the congestion function. Finally, in a multiagent approach to designing manufacturing systems (see, e.g., [9]), it may not be known in advance how performance measures (such as throughput) depend on manufacturing policy. Rather, performance can only be measured once a policy is implemented.

Our interest in this paper is to develop algorithms that enable coordination in multiagent systems for precisely this "payoff-based" scenario, in which agents only have access to (possibly noisy) measurements of the rewards received through repeated interactions with other agents. We adopt the framework of "learning in games." (See [5, 10, 25, 26] for an extensive overview. See also the recent special issue containing [22] or survey article [18] for perspectives from machine learning.) Unlike most of the learning rules in this literature, which assume that agents adjust their behavior based on the observed behavior of other agents, we shall assume that agents know only their own past actions and the payoffs that resulted. It is far from obvious that Nash equilibrium can be achieved under such a restriction, but in fact it has recently been shown that such "payoff-based" learning rules can be constructed that work in any game [4, 8].

In this paper we show that there are simpler and more intuitive adjustment rules that achieve this objective for a large class of multiplayer games known as "weakly acyclic" games. This class captures many problems of interest in cooperative control [13, 14]. It includes the very special case of "identical interest" games, where each agent receives the same reward. However, weakly acyclic games (and the related concept of potential games) capture other scenarios such as congestion games [19] and similar problems such as distributed routing in networks, weapon target assignment, consensus, and area coverage. See [15, 1] and references therein for a discussion of a learning in games approach to cooperative control problems, but under less stringent assumptions on informational constraints than considered in this paper.

For many multiagent problems, operation at a pure Nash equilibrium may reflect optimization of a collective objective.[1] We will derive payoff-based dynamics that

[1] Nonetheless, there are varied viewpoints on the role of Nash equilibrium as a solution concept for multiagent systems. See [22] and [12].

guarantee asymptotically that agent strategies will constitute a pure Nash equilibrium with arbitrarily high probability. It need not always be the case that at least one Nash equilibrium optimizes a collective objective. Motivated by this consideration, we also discuss the introduction of incentives or tolls in a player's payoff function to assure that there is at least one Nash equilibrium that optimizes a collective objective. Even in this case, however, there may still be suboptimal Nash equilibria.

The remainder of this paper is organized as follows. Section 2 provides background on finite strategic-form games and repeated games. This is followed by three types of payoff-based dynamics in section 3 for increasingly general problems. Subsection 3.1 presents "safe experimentation dynamics" which is restricted to identical interest games. Subsection 3.2 presents "simple experimentation dynamics" for the more general class of weakly acyclic games but with noise-free payoff measurements. Subsection 3.3 presents "sample experimentation dynamics" for weakly acyclic games with noisy payoff measurements. Section 4 discusses how to introduce tolls and incentives in payoffs so that a Nash equilibrium optimizes a collective objective. Section 5 presents an illustrative example of a traffic congestion game. Finally, section 6 contains some concluding remarks. An important analytical tool throughout is the method of resistance trees for perturbed Markov chains [24], which is reviewed in an appendix.

2. Background.

In this section, we will present a brief background of the game theoretic concepts used in the paper. We refer the readers to [6, 25, 26] for a more comprehensive review.

2.1. Finite strategic-form games.

Consider a finite strategic-form game with n-player set $\mathcal{P} := \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ where each player $\mathcal{P}_i \in \mathcal{P}$ has a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We will sometimes use a single symbol, e.g., G, to represent the entire game, i.e., the player set, $\mathcal{P}$, action sets, $\mathcal{A}_i$, and utility functions U_i.

For an action profile $a = (a_1, a_2, \ldots, a_n) \in \mathcal{A}$, let a_{-i} denote the profile of player actions *other than* player $\mathcal{P}_i$, i.e.,

$$a_{-i} = \{a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\}.$$

With this notation, we will sometimes write a profile a of actions as (a_i, a_{-i}). Similarly, we may write $U_i(a)$ as $U_i(a_i, a_{-i})$.

An action profile $a^* \in \mathcal{A}$ is called a *pure Nash equilibrium* if for all players $\mathcal{P}_i \in \mathcal{P}$,

$$U_i(a_i^*, a_{-i}^*) = \max_{a_i \in \mathcal{A}_i} U_i(a_i, a_{-i}^*). \tag{2.1}$$

Furthermore, if the above condition is satisfied with a unique maximizer for every player $\mathcal{P}_i \in \mathcal{P}$, then a^* is called a *strict (Nash) equilibrium*.

In this paper we will consider three classes of games: identical interest games, potential games, and weakly acyclic games. Each class of games has a connection to general cooperative control problems and multiagent systems for which there is some global utility or potential function $\phi : \mathcal{A} \to \mathbb{R}$ that a global planner seeks to maximize [13].

2.1.1. Identical interest games.

The most restrictive class of games that we will review in this paper is identical interest games. In such a game, the players' utility functions $\{U_i\}_{i=1}^n$ are chosen to be the same. That is, for some function $\phi : \mathcal{A} \to \mathbb{R}$,

$$U_i(a) = \phi(a)$$

for every $\mathcal{P}_i \in \mathcal{P}$ and for every $a \in \mathcal{A}$. It is easy to verify that all identical interest games have at least one pure Nash equilibrium, namely, any action profile a that maximizes $\phi(a)$.

2.1.2. Potential games. A significant generalization of an identical interest game is a potential game. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the global utility. Specifically, there is a function $\phi : \mathcal{A} \to \mathbb{R}$ such that for every player $\mathcal{P}_i \in \mathcal{P}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}).$$

When this condition is satisfied, the game is called an exact potential game with the potential function ϕ.[2] It is easy to see that, in potential games, any action profile maximizing the potential function is a pure Nash equilibrium, and hence every potential game possesses at least one such equilibrium. An example of an exact potential game is illustrated in Figure 1.

	L	R
U	$0, 0$	$-1, 1$
D	$1, -1$	$0, 0$

Payoffs

	L	R
U	0	1
D	1	2

Potential

FIG. 1. *An example of a two player exact potential game.*

2.1.3. Weakly acyclic games. Consider any finite game G with a set $\mathcal{A}$ of action profiles. A *better reply path* is a sequence of action profiles $a^1, a^2, \ldots, a^L$ such that for each successive pair a^j, a^{j+1} there is exactly one player such that $a_i^j \neq a_i^{j+1}$ and for that player $U_i(a^{j+1}) > U_i(a^j)$. In other words, one player moves at a time, and each time a player moves he increases his own utility.

Suppose now that G is a potential game with potential function ϕ. Starting from an arbitrary action profile $a \in \mathcal{A}$, construct a better reply path $a = a^1, a^2, \ldots, a^L$ until it can no longer be extended. Note first that such a path cannot cycle back on itself, because ϕ is strictly increasing along the path. Since $\mathcal{A}$ is finite, the path cannot be extended indefinitely. Hence, the last element in a maximal better reply path from any joint action, a, must be a Nash equilibrium of G.

This idea may be generalized as follows. The game G is *weakly acyclic* if for any $a \in \mathcal{A}$, there exists a better reply path starting at a and ending at some pure Nash equilibrium of G [25, 26]. Potential games are special cases of weakly acyclic games. An example of a two player weakly acyclic game is illustrated in Figure 2. Notice that the illustrated game is not a potential game.

2.2. Repeated games. In a repeated game, at each time $t \in \{0, 1, 2, \ldots\}$, each player $\mathcal{P}_i \in \mathcal{P}$ simultaneously chooses an action $a_i(t) \in \mathcal{A}_i$ and receives the utility $U_i(a(t))$, where $a(t) := (a_1(t), \ldots, a_n(t))$. Each player $\mathcal{P}_i \in \mathcal{P}$ chooses action $a_i(t)$ at time t according to a probability distribution $p_i(t)$, which we will refer to as the

[2] There are weaker notions of potential games such as ordinal or weighted potential games. Rather than discuss each variation specifically, we will discuss a more general framework, weakly acyclic games, in the ensuing section. Any potential game, whether exact, ordinal, or weighted, is a weakly acyclic game.

	L	C	R
U	$0,0$	$0.1,0$	$1,1$
M	$1,0$	$0,1$	$0,0$
D	$0,1$	$1,0$	$0,0$

Fig. 2. *An example of a two player weakly acyclic game.*

strategy of player $\mathcal{P}_i$ at time t. A player's strategy at time t can rely only on observations from times $\{0, 1, 2, \ldots, t-1\}$. Different learning algorithms are specified by both the assumptions on available information and the mechanism by which the strategies are updated as information is gathered. For example, if a player knows the functional form of his utility function and is capable of observing the actions of all other players at every time step, then the strategy adjustment mechanism of player $\mathcal{P}_i$ can be written in the general form

$$p_i(t) = F_i\big(a(0), \ldots, a(t-1); U_i\big).$$

An example of a learning algorithm, or strategy adjustment mechanism, of this form is the well-known fictitious play [16]. For a detailed review of learning in games, we direct the reader to [5, 25, 26, 11, 23, 20].

In this paper we deal with the issue of whether players can learn to play a pure Nash equilibrium through repeated interactions under the most restrictive observational conditions; players *only* have access to (i) the action they played and (ii) the utility (possibly noisy) they received. In this setting, the strategy adjustment mechanism of player $\mathcal{P}_i$ takes on the form

$$(2.2)\quad p_i(t) = F_i\big(\{a_i(0), U_i(a(0)) + \nu_i(0)\}, \ldots, \{a_i(t-1), U_i(a(t-1)) + \nu_i(t-1)\}\big),$$

where the $\nu_i(t)$ are zero mean independent and identically distributed (i.i.d.) random variables.

3. Payoff-based learning algorithms.

In this section, we will introduce three simple payoff-based learning algorithms. The first, called *safe experimentation*, guarantees convergence to a pure optimal Nash equilibrium in any identical interest game. Such an equilibrium is optimal because each player's utility is maximized. The second learning algorithm, called *simple experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game. The third learning algorithm, called *sample experimentation*, guarantees convergence to a pure Nash equilibrium in any weakly acyclic game even when utility measurements are corrupted with noise.

3.1. Safe experimentation dynamics for identical interest games.

3.1.1. Constant exploration rates.

Before introducing the learning dynamics, we introduce the following function. Let

$$U_i^{\max}(t) := \max_{0 \le \tau \le t-1} U_i(a(\tau))$$

be the maximum utility that player $\mathcal{P}_i$ has received up to time $t-1$.

We will now introduce the safe experimentation dynamics for identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time $t = 1$ and is denoted by $a_i^b(1) = a_i(0)$.

2. **Action selection:** At each subsequent time step, each player selects his baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ, i.e.,
   - $a_i(t) = a_i^b(t)$ with probability $(1-\epsilon)$;
   - $a_i(t)$ is chosen randomly (uniformly) over $\mathcal{A}_i$ with probability ϵ.

   The variable ϵ will be referred to as the player's *exploration rate*.
3. **Baseline strategy update:** Each player compares the actual utility received, $U_i(a(t))$, with the maximum received utility $U_i^{\max}(t)$ and updates the baseline action as follows:

$$a_i^b(t+1) = \begin{cases} a_i(t), & U_i(a(t)) > U_i^{\max}(t), \\ a_i^b(t), & U_i(a(t)) \leq U_i^{\max}(t). \end{cases}$$

   Each player updates the maximum received utility regardless of whether or not step 2 involved exploration.
4. Return to step 2 and repeat.

The reason that this learning algorithm is called "safe" experimentation is that the utility evaluated at the baseline action, $U(a^b(t))$, is nondecreasing with respect to time.

THEOREM 3.1. *Let G be a finite n-player identical interest game in which all players use the safe experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is an optimal Nash equilibrium of G with at least probability p.*

Proof. Since G is an identical interest game, let the utility of each player be expressed as $U : \mathcal{A} \to \mathbb{R}$, and let $\mathcal{A}^*$ be the set of "optimal" Nash equilibria of G, i.e.,

$$\mathcal{A}^* = \left\{ a^* \in \mathcal{A} : U(a^*) = \max_{a \in \mathcal{A}} U(a) \right\}.$$

For any joint action, $a(t)$, the ensuing joint action will constitute an optimal Nash equilibrium with at least probability

$$\left(\frac{\epsilon}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon}{|\mathcal{A}_2|}\right) \cdots \left(\frac{\epsilon}{|\mathcal{A}_n|}\right),$$

where $|\mathcal{A}_i|$ denotes the cardinality of the action set of player $\mathcal{P}_i$. Therefore, an optimal Nash equilibrium will eventually be played with probability 1 for any $\epsilon > 0$.

Suppose an optimal Nash equilibrium is first played at time t^*, i.e., $a(t^*) \in \mathcal{A}^*$ and $a(t^*-1) \notin \mathcal{A}^*$. Then the baseline joint action must remain constant from that time onwards, i.e., $a^b(t) = a(t^*)$ for all $t > t^*$. An optimal Nash equilibrium will then be played at any time $t > t^*$ with at least probability $(1-\epsilon)^n$. Since $\epsilon > 0$ can be chosen arbitrarily small, and in particular such that $(1-\epsilon)^n > p$, this completes the proof. □

3.1.2. Diminishing exploration rates. In the safe experimentation dynamics, the exploration rate ϵ was defined as a constant. Alternatively, one could let the exploration rate vary to induce desirable behavior. One example would be to let the exploration rate decay, such as $\epsilon_t = (1/t)^{1/n}$. This would induce exploration at early stages and reduce exploration at later stages of the game. The theorem and proof hold under the following conditions for the exploration rate:

$$\lim_{t\to\infty} \epsilon_t = 0,$$

$$\lim_{t\to\infty} \prod_{\tau=1}^{t} \left[1 - \left(\frac{\epsilon_\tau}{|\mathcal{A}_1|}\right)\left(\frac{\epsilon_\tau}{|\mathcal{A}_2|}\right) \cdots \left(\frac{\epsilon_\tau}{|\mathcal{A}_n|}\right) \right] = 0.$$

3.2. Simple experimentation dynamics for weakly acyclic games. We will now introduce the simple experimentation dynamics for weakly acyclic games. These dynamics will allow us to relax the assumption of identical interest games.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0)$. This action will be initially set as the player's *baseline action* at time 1, i.e., $a_i^b(1) = a_i(0)$. Likewise, the player's *baseline utility* at time 1 is initialized as $u_i^b(1) = U_i(a(0))$.
2. **Action selection:** At each subsequent time step, each player selects a baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ, i.e.,
   - $a_i(t) = a_i^b(t)$ with probability $(1-\epsilon)$;
   - $a_i(t)$ is chosen randomly (uniformly) over $\mathcal{A}_i$ with probability ϵ.

   The variable ϵ will be referred to as the player's *exploration rate.* Whenever $a_i(t) \neq a_i^b(t)$, we will say that player $\mathcal{P}_i$ *experimented.*
3. **Baseline action and baseline utility update:** Each player compares the utility received, $U_i(a(t))$, with his baseline utility, $u_i^b(t)$, and updates his baseline action and utility as follows:
   - If player $\mathcal{P}_i$ *experimented* (i.e., $a_i(t) \neq a_i^b(t)$) and if $U_i(a(t)) > u_i^b(t)$, then
     $a_i^b(t+1) = a_i(t)$,
     $u_i^b(t+1) = U_i(a(t))$.
   - If player $\mathcal{P}_i$ *experimented* and if $U_i(a(t)) \leq u_i^b(t)$, then
     $a_i^b(t+1) = a_i^b(t)$,
     $u_i^b(t+1) = u_i^b(t)$.
   - If player $\mathcal{P}_i$ *did not experiment* (i.e., $a_i(t) = a_i^b(t)$), then
     $a_i^b(t+1) = a_i^b(t)$,
     $u_i^b(t+1) = U_i(a(t))$.
4. Return to step 2 and repeat.

As before, these dynamics require only utility measurements and hence almost no information regarding the structure of the game.

THEOREM 3.2. *Let G be a finite n-player weakly acyclic game in which all players use the simple experimentation dynamics. Given any probability $p < 1$, if the exploration rate $\epsilon > 0$ is sufficiently small, then for all sufficiently large times t, $a(t)$ is a Nash equilibrium of G with at least probability p.*

The remainder of this subsection is devoted to the proof of Theorem 3.2. The proof relies on the theory of resistance trees for perturbed Markov chains (see the appendix for a brief review).

Define the *state* of the dynamics to be the pair $[a, u]$, where a is the baseline joint action and u is the baseline utility vector. We will omit the superscript b to avoid cumbersome notation.

Partition the state space into the following three sets. First, let X be the set of states $[a, u]$ such that $u_i \neq U_i(a)$ for at least one player $\mathcal{P}_i$. Let E be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a Nash equilibrium. Let D be the set of states $[a, u]$ such that $u_i = U_i(a)$ for all players $\mathcal{P}_i$ and a is a disequilibrium (not a Nash equilibrium). These are all the states.

Claim 3.1.

(a) *Any state $[a,u] \in X$ transitions to a state in $E \cup D$ in one period with probability $O(1)$.*

(b) *Any state $[a,u] \in E \cup D$ transitions to a different state $[a',u']$ with probability at most $O(\varepsilon)$.*

Proof. For any $[a,u'] \in X$, there exists at least one player $\mathcal{P}_i$ such that $u_i' \neq U_i(a)$. If all players repeat their part of the joint action profile a, which occurs with probability $(1-\epsilon)^n$, then $[a,u']$ transitions to $[a,u]$, where $u_i = U_i(a)$ for all players $\mathcal{P}_i$. Thus the process moves to $[a,u] \in E \cup D$ with prob $O(1)$. This proves statement (a). As for statement (b), any state in $E \cup D$ transitions back to itself whenever no player experiments, which occurs with probability at least $O(1)$. □

Claim 3.2. *For any state $[a,u] \in D$, there is a finite sequence of transitions to a state $[a^*,u^*] \in E$, where the transitions have the form*[3]

$$[a,u] \underset{O(\epsilon)}{\rightarrow} [a^1,u^1] \underset{O(\epsilon)}{\rightarrow} \cdots \underset{O(\epsilon)}{\rightarrow} [a^*,u^*],$$

where $u_i^k = U_i(a^k)$ for all i and for all $k > 0$, and each transition occurs with probability $O(\epsilon)$.

Proof. Such a sequence is guaranteed by weak acyclicity. Since a is not an equilibrium, there is a better reply path from a to some equilibrium a^*, say $a, a^1, a^2, \ldots, a^*$.

At $[a,u]$ the appropriate player $\mathcal{P}_i$ experiments with probability ϵ and chooses the appropriate better reply with probability $1/|\mathcal{A}_i|$, and no one else experiments. Thus the process moves to $[a^1,u^1]$, where $u_i^1 = U_i(a^1)$ for all players $\mathcal{P}_i$ with probability $O(\epsilon)$ (more precisely, $O(\epsilon(1-\epsilon)^{n-1})$). Notice that for the deviator $\mathcal{P}_i$, $U_i(a^1) > U_i(a)$, and therefore $u_i^1 = U_i(a^1)$. For the nondeviator, say, player $\mathcal{P}_j$, $u_j^1 = U_j(a^1)$ since $a_j^1 = a_j$. Thus $[a^1,u^1] \in D \cup E$. In the next period, the appropriate player deviates, and so forth. □

Claim 3.3. *For any equilibrium $[a^*,u^*] \in E$, any path from $[a^*,u^*]$ to another state $[a,u] \in E \cup D$, $a \neq a^*$, that does not loop back to $[a^*,u^*]$ must be one of the following two forms:*

(1) $[a^*,u^*] \underset{O(\epsilon)}{\rightarrow} [a^*,u'] \underset{O(\epsilon^k)}{\rightarrow} [a',u''] \rightarrow \cdots \rightarrow [a,u]$, *where* $k \geq 1$;

(2) $[a^*,u^*] \underset{O(\epsilon^k)}{\rightarrow} [a',u''] \rightarrow \cdots \rightarrow [a,u]$, *where* $k \geq 2$.

Proof. The path must begin by either one player experimenting or more that one player experimenting. Case (2) results if more than one player experiments. Case (1) results if exactly one agent, say, agent $\mathcal{P}_i$, experiments with an action $a_i' \neq a_i^*$ and all other players continue to play their part of a^*. This happens with probability $(\epsilon/|\mathcal{A}_i|)(1-\epsilon)^{n-1}$. In this situation, player $\mathcal{P}_i$ cannot be better off, meaning that $U_i(a_i', a_{-i}^*) \leq U_i(a^*)$, since by assumption a^* is an equilibrium. Hence the baseline action next period remains a^* for all players, though their baseline utilities may change. Denote the next state by $[a^*,u']$. If in the subsequent period all players continue to play their part of the action a^*, which occurs with probability $(1-\epsilon)^n$, then the state reverts back to $[a^*,u^*]$ and we have a loop. Hence, the only way the path can continue without a loop is for one or more players to experiment in the next stage, which has probability $O(\epsilon^k)$, $k \geq 1$. This is exactly what case (1) alleges. □

Proof of Theorem 3.2. This is a finite aperiodic Markov process on the state space $\mathcal{A} \times \bar{U}_1 \times \cdots \times \bar{U}_n$, where $\bar{U}_i$ denotes the (finite) range of $U_i(\cdot)$. Furthermore, from

[3] We will use the notation $z \rightarrow z'$ to denote the transition from state z to state z'. We use $z \underset{O(\epsilon)}{\rightarrow} z'$ to emphasize that this transition occurs with probability of order ϵ.

every state there exists a positive probability path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will now show that within any recurrent class, the trees (see the appendix) rooted at the Nash equilibrium will have the lowest resistance. Therefore, according to Theorem A.1, the a priori probability that the state will be a Nash equilibrium can be made arbitrarily close to 1.

In order to apply Theorem A.1, we will construct minimum resistance trees with vertices consisting of every possible state (within a recurrence class). Each edge will have resistance $0, 1, 2, \ldots$ associated with the transition probabilities $O(1), O(\epsilon)$, $O(\epsilon^2), \ldots$, respectively.

Our analysis will deviate slightly from the presentation in the appendix. In the discussion in the appendix, the vertices of minimum resistance trees are recurrence classes of an associated unperturbed Markov chain. In this case, the unperturbed Markov chain corresponds to simple experimentation dynamics with $\epsilon = 0$, and so the recurrence classes are all states in $E \cup D$. Nonetheless, we will construct resistance trees with the vertices being all possible states, i.e., $E \cup D \cup X$. The resulting conclusions remain the same (see Lemma 1 in [24]). Since the states in X are transient with probability $O(1)$, the resistance to leave a node corresponding to a state in X is 0. Therefore, the presence of such states does not affect the conclusions determining which states are stochastically stable.

Suppose a minimum resistance tree T is rooted at a vertex v that is not in E. If $v \in X$, it is easy to construct a new tree that has lower resistance. Namely, by Claim 3.1(a), there is a zero-resistance one-hop path P from v to some state $[a, u] \in E \cup D$. Add the edge of P to T and subtract the edge in T that exits from the vertex $[a, u]$. This results in a $[a, u]$-tree T'. It has lower resistance than T because the added edge has zero resistance, while the subtracted edge has resistance greater than or equal to 1 because of Claim 3.1(b). This argument is illustrated in Figure 3, where the edge of strictly positive resistance ($R \geq 1$) is removed and replaced with the edge of zero resistance ($R = 0$).

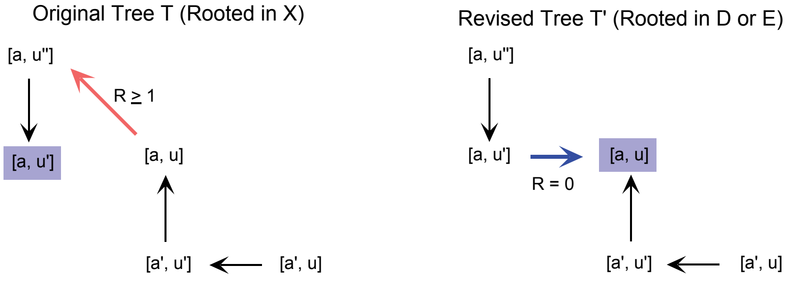

FIG. 3. *Construction of alternative to tree rooted in X.*

Suppose next that $v = [a, u] \in D$ but not in E. Construct a path P as in Claim 3.2 from $[a, u]$ to some state $[a^*, u^*] \in E$. As above, construct a new tree T' rooted at $[a^*, u^*]$ by adding the edges of P to T and taking out the redundant edges (the edges in T that exit from the vertices in P). The nature of the path P guarantees that the edges taken out have total resistance at least as high as the resistances of the edges put in. This is because the entire path P lies in $E \cup D$, each transition on the path has resistance 1, and, from Claim 3.2(b), the resistance to leave any state in $E \cup D$ is at least 1.

To construct a new tree that has strictly lower resistance, we will inspect the effect of removing the exiting edge from $[a^*, u^*]$ in T. Note that this edge must fit either case (1) or (2) of Claim 3.3.

In case (2), the resistance of the exiting edge is at least 2, which is larger than any edge in P. Hence the new tree has strictly lower resistance than T, which is a contradiction. This argument is illustrated in Figure 4. A new path is created from the original root $[a, u] \in D$ to the equilibrium $[a^*, u^*] \in E$ ($R = 1$ edges). Redundant ($R \geq 1$, $R \geq 2$) edges emanating from the new path are removed. In case (2), the redundant edge emanating from $[a^*, u^*]$ has a resistance of at least 2.

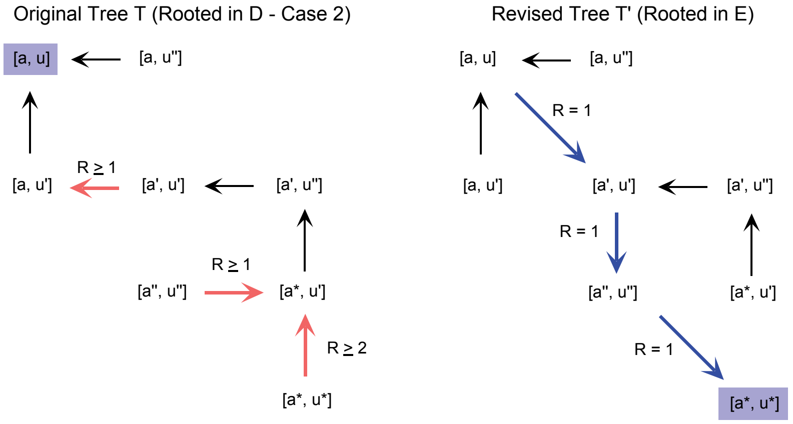

FIG. 4. *Construction of alternative to tree rooted in D for case* (2).

In case (1), the exiting edge has the form $[a^*, u^*] \rightarrow [a^*, u']$ which has resistance 1 where $u^* \neq u'$. The next edge in T, say, $[a^*, u'] \rightarrow [a', u'']$, also has at least resistance 1. Remove the edge $[a^*, u'] \rightarrow [a', u'']$ from T, and put in the edge $[a^*, u'] \rightarrow [a^*, u^*]$. The latter has resistance 0 since $[a^*, u'] \in X$. This results in a tree T'' that is rooted at $[a^*, u^*]$ and has strictly lower resistance than does T, which is a contradiction. This argument is illustrated in Figure 5. As in Figure 4, a new ($R = 1$, $R = 0$) path is constructed and redundant ($R \geq 1$, $R = 1$) edges are removed. The difference is that the edge $[a^*, u'] \rightarrow [a', u'']$ is removed and replaced with $[a^*, u'] \rightarrow [a^*, u^*]$.

To recap, a minimum resistance tree cannot be rooted at any state in X or D, but rather only at a state in in E. Therefore, when ϵ is sufficiently small, the long-run probability on E can be made arbitrarily close to 1, and in particular, larger than any specified probability p. $\square$

3.3. Sample experimentation dynamics for weakly acyclic games with noisy utility measurements.

3.3.1. Noise-free utility measurements. In this section we will focus on developing payoff-based dynamics for which the limiting behavior exhibits that of a pure Nash equilibrium with arbitrarily high probability in any finite weakly acyclic game *even in the presence of utility noise*. We will show that a variant of the so-called regret testing algorithm [4] accomplishes this objective for weakly acyclic games with noisy utility measurements.

We now introduce sample experimentation dynamics.

1. **Initialization:** At time $t = 0$, each player randomly selects and plays any action, $a_i(0) \in \mathcal{A}_i$. This action will be initially set as each player's *baseline action*, $a_i^b(1) = a_i(0)$.

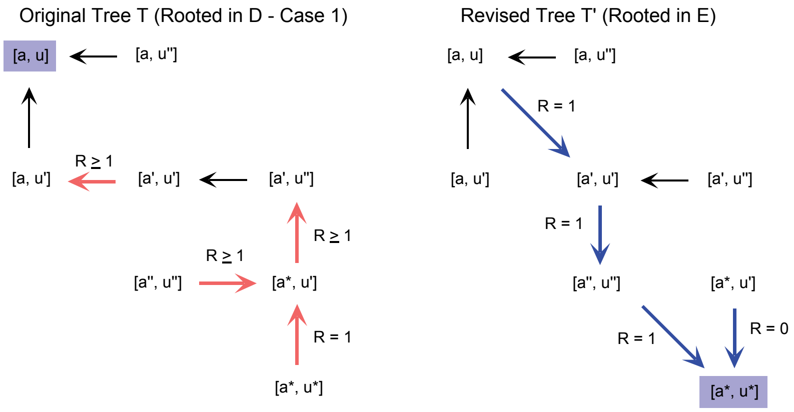

FIG. 5. *Construction of alternative to tree rooted in D for case* (1).

2. **Exploration phase:** After the baseline action is set, each player engages in an *exploration phase* over the next m periods. The exploration phases need not be synchronized or of the same length for each player, but we will assume that they are for the proof. For convenience, we will double index the time of the actions played as

$$\check{a}(t_1, t_2) = a(m\, t_1 + t_2),$$

where t_1 indexes the number of the exploration phase and t_2 indexes the actions played in that exploration phase. We will refer to t_1 as the *exploration phase time* and to t_2 as the *exploration action time*. By construction, the exploration phase time and exploration action time satisfy $t_1 \geq 1$ and $m \geq t_2 \geq 1$, respectively. The baseline action will be updated only at the end of the exploration phase and will therefore be indexed only by the exploration phase time.

During the exploration phase, each player selects a baseline action with probability $(1-\epsilon)$ or experiments with a new random action with probability ϵ. That is, for any exploration phase time $t_1 \geq 1$ and for any exploration action time satisfying $m \geq t_2 \geq 1$,

- $\check{a}_i(t_1, t_2) = a_i^b(t_1)$ with probability $(1-\epsilon)$,
- $\check{a}_i(t_1, t_2)$ is chosen randomly (uniformly) over $(\mathcal{A}_i \setminus a_i^b(t_1))$ with probability ϵ.

Again, the variable ϵ will be referred to as the player's *exploration rate.*

3. **Action assessment:** After the exploration phase, each player evaluates the average utility received when playing each of his actions during the exploration phase. Let $n_i^{a_i}(t_1)$ be the number of times that player $\mathcal{P}_i$ played action a_i during the exploration phase at time t_1. The average utility for action a_i during the exploration phase at time t_1 is

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1, t_2)\} U_i(\check{a}(t_1, t_2)), & n_i^{a_i}(t_1) > 0, \\ U_{\min}, & n_i^{a_i}(t_1) = 0, \end{cases}$$

where $I\{\cdot\}$ is the usual indicator function and $U_{\min}$ satisfies

$$U_{\min} < \min_i \min_{a \in \mathcal{A}} U_i(a).$$

In other words, $U_{\min}$ is less than the smallest payoff any agent can receive.

4. **Evaluation of better response set:** Each player compares the average utility received when playing a baseline action, $\hat{V}_i^{a_i^b(t)}(t_1)$, with the average utility received for each of the other actions, $\hat{V}_i^{a_i}(t_1)$, and finds all played actions which performed δ better than the baseline action. The term δ will be referred to as the players' *tolerance level*. Define $\mathcal{A}_i^*(t_1)$ to be the set of actions that outperformed the baseline action as follows:

$$\mathcal{A}_i^*(t_1) = \left\{ a_i \in \mathcal{A}_i : \hat{V}_i^{a_i}(t_1) \geq \hat{V}_i^{a_i^b(t_1)}(t_1) + \delta \right\}. \tag{3.1}$$

5. **Baseline strategy update:** Each player updates a baseline action as follows:
   - If $\mathcal{A}_i^*(t_1) = \emptyset$, then $a_i^b(t_1+1) = a_i^b(t_1)$.
   - If $\mathcal{A}_i^*(t_1) \neq \emptyset$, then
     - with probability ω, set $a_i^b(t_1+1) = a_i^b(t_1)$. (We will refer to ω as the player's inertia.)
     - with probability $1-\omega$, randomly select $a_i^b(t_1+1) \in \mathcal{A}_i^*(t_1)$ with uniform probability.
6. Return to step 2 and repeat.

For simplicity, we will first state and prove the desired convergence properties using noiseless utility measurements. The setup for the noisy utility measurements will be stated afterwards.

Before stating the following theorem, we define the constant $\alpha > 0$ as follows. If $U_i(a^1) \neq U_i(a^2)$ for any joint actions $a^1, a^2 \in \mathcal{A}$ and any player $\mathcal{P}_i \in \mathcal{P}$, then $|U_i(a^1) - U_i(a^2)| > \alpha$. In other words, if any two joint actions result in different utilities at all, then the difference would be at least α.

THEOREM 3.3. *Let G be a finite n-player weakly acyclic game in which all players use the sample experimentation dynamics. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$,*
- *inertia $\omega \in (0,1)$, and*
- *exploration rate ϵ satisfying $\min\{(\alpha-\delta)/4, \delta/4, 1-p\} > (1-(1-\epsilon)^n) > 0$,*

if the exploration phase length m is sufficiently large, then for all sufficiently large times $t > 0$, $a(t)$ is a Nash equilibrium of G with at least probability p.

The remainder of this subsection is devoted to the proof of Theorem 3.3.

We will assume for simplicity that utilities are between $-1/2$ and $1/2$, i.e., $|U_i(a)| \leq 1/2$ for any player $\mathcal{P}_i \in \mathcal{P}$ and any joint action $a \in \mathcal{A}$.

We begin with a series of useful claims. The first claim states that for any player $\mathcal{P}_i$ the average utility for an action $a_i \in \mathcal{A}_i$ during the exploration phase can be made arbitrarily close (with high probability) to the actual utility the player would have received provided that all other players never experimented. This can be accomplished if the experimentation rate is sufficiently small and the exploration phase length is sufficiently large.

CLAIM 3.4. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For*

- *any probability $p < 1$,*
- *any $\delta^* > 0$, and*
- *any exploration rate $\epsilon > 0$ satisfying $\delta^*/2 \geq (1-(1-\epsilon)^{n-1}) > 0$,*

if the exploration phase length m is sufficiently large, then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] < 1 - p.$$

Proof. Let $n_i(a_i)$ represent the number of times player $\mathcal{P}_i$ played action a_i during the exploration phase. In the following discussion, *all probabilities and expectations are conditioned on* $n_i(a_i) > 0$. We omit making this explicit for the sake of notational simplicity. The event $n_i(a_i) = 0$ has diminishing probability as the exploration phase length m increases, and so this case will not affect the desired conclusions for increasing phase lengths.

For an arbitrary $\delta^* > 0$,

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad\le \mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| + \left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*\right] \\
&\quad\le \underbrace{\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \delta^*/2\right]}_{(*)} + \underbrace{\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right]}_{(**)}.
\end{aligned}$$

First, let us focus on $(**)$. We have

$$E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b) = \left[1 - (1-\epsilon)^{n-1}\right]\left[E\{U_i(a_i, a_{-i}(t)) \mid a_{-i}(t) \neq a_{-i}^b\} - U_i(a_i, a^b)\right],$$

which approaches 0 as $\epsilon \downarrow 0$. Therefore, for any exploration rate ϵ satisfying $\delta^*/2 > (1 - (1-\epsilon)^{n-1}) > 0$, we know that

$$\mathbf{Pr}\left[\left|E\{\hat{V}_i^{a_i}\} - U_i(a_i, a_{-i}^b)\right| > \delta^*/2\right] = 0.$$

Now we will focus on $(*)$. By the weak law of large numbers, $(*)$ approaches 0 as $n_i(a_i) \uparrow \infty$. This implies that for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$, there exists a sample size $n_i^*(a_i)$ such that if $n_i(a_i) > n_i^*(a_i)$, then

$$\mathbf{Pr}\left[\left|\hat{V}_i^{a_i} - E\{\hat{V}_i^{a_i}\}\right| > \rho/2\right] < 1 - \bar{p}.$$

Lastly, for any probability $\bar{p} < 1$ and any fixed exploration rate, there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$,

$$\mathbf{Pr}\left[n_i(a_i) \ge n_i^*(a_i)\right] \ge \bar{p}.$$

In summary, for any fixed exploration rate ϵ satisfying $\delta^*/2 \ge (1 - (1-\epsilon)^{n-1}) > 0$, $(*) + (**)$ can be made arbitrarily close to 0, provided that the exploration length m is sufficiently large. □

CLAIM 3.5. *Let a^b be the joint baseline action at the start of an exploration phase of length m. For any*

- *probability $p < 1$,*
- *tolerance level $\delta \in (0, \alpha)$, and*
- *exploration rate $\epsilon > 0$ satisfying* $\min\{(\alpha - \delta)/4, \delta/4\} \ge (1 - (1-\epsilon)^{n-1}) > 0$,

if the exploration length m is sufficiently large, then each player's better response set $\mathcal{A}_i^*$ will contain only and all *actions that are a better response to the joint baseline action, i.e.,*

$$a_i^* \in \mathcal{A}_i^* \Leftrightarrow U_i(a_i^*, a_{-i}^b) > U_i(a^b)$$

with at least probability p.

Proof. Suppose a^b is not a Nash equilibrium. For some player $\mathcal{P}_i \in \mathcal{P}$, let a_i^* be a strict better reply to the baseline joint action, i.e., $U_i(a_i^*, a_{-i}^b) > U_i(a^b)$, and let a_i^w be a nonbetter reply to the baseline joint action, i.e., $U_i(a_i^w, a_{-i}^b) \leq U_i(a^b)$.

Using Claim 3.4, for any probability $\bar{p} < 1$ and any exploration rate $\epsilon > 0$ satisfying $\min\{(\alpha - \delta)/4, \delta/4\} \geq (1 - (1 - \epsilon)^{n-1}) > 0$ there exists a minimum exploration length $\underline{m} > 0$ such that for any exploration length $m > \underline{m}$ the following expressions are true:

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.2}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.3}$$

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] \geq \bar{p}, \tag{3.4}$$

where $\delta^* = \min\{(\alpha - \delta)/2, \delta/2\}$. Rewriting (3.2), we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) < (\alpha - \delta)/2\right],$$

and rewriting (3.3), we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^*, a_{-i}^b) > -(\alpha - \delta)/2\right] \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - (U_i(a_i^b, a_{-i}^b) + \alpha) > -(\alpha - \delta)/2\right] \\ &= \mathbf{Pr}\left[\hat{V}_i^{a_i^*} - U_i(a_i^b, a_{-i}^b) > (\alpha + \delta)/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^* \in \mathcal{A}_i^*\right] \geq \bar{p}^2.$$

Similarly, rewriting (3.2), we obtain

$$\mathbf{Pr}\left[|\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b)| < \delta^*\right] \leq \mathbf{Pr}\left[\hat{V}_i^{a_i^b} - U_i(a_i^b, a_{-i}^b) > -\delta/2\right],$$

and rewriting (3.4), we obtain

$$\begin{aligned}\mathbf{Pr}\left[|\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b)| < \delta^*\right] &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^w, a_{-i}^b) < \delta/2\right] \\ &\leq \mathbf{Pr}\left[\hat{V}_i^{a_i^w} - U_i(a_i^b, a_{-i}^b) < \delta/2\right],\end{aligned}$$

meaning that

$$\mathbf{Pr}\left[a_i^w \notin \mathcal{A}_i^*\right] \geq \bar{p}^2.$$

Since $\bar{p}$ can be chosen arbitrarily close to 1, the proof is complete. □

Proof of Theorem 3.3. The evolution of the baseline actions from phase to phase is a finite aperiodic Markov process on the state space of joint actions, $\mathcal{A}$. Furthermore, since G is weakly acyclic, from every state there exists a better reply path to a Nash equilibrium. Hence, every recurrent class has at least one Nash equilibrium. We will show that these dynamics can be viewed as a perturbation of a certain Markov chain whose recurrent classes are restricted to Nash equilibria. We will then appeal to Theorem A.1 to derive the desired result.

We begin by defining an "unperturbed" process on baseline actions. For any $a^b \in \mathcal{A}$, define the *true* better reply set as

$$\bar{\mathcal{A}}_i^*(a^b) = \left\{a_i : U_i(a_i, a_{-i}^b) > U_i(a^b)\right\}.$$

Now define the transition process from $a^b(t_1)$ to $a^b(t_1+1)$ as follows:

- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) = \emptyset$, then $a_i^b(t_1+1) = a_i^b(t_1)$.
- If $\bar{\mathcal{A}}_i^*(a^b(t_1)) \neq \emptyset$, then
  - with probability ω, set $a_i^b(t_1+1) = a_i^b(t_1)$.
  - with probability $1-\omega$, randomly select $a_i^b(t_1+1) \in \bar{\mathcal{A}}_i^*(t_1)$ with uniform probability.

This is a special case of a so-called "better reply process with finite memory and inertia." From [26, Theorem 6.2], the joint actions of this process converge to a Nash equilibrium with probability 1 in any weakly acyclic game. Therefore, the recurrence classes of this unperturbed are precisely the set of pure Nash equilibria.

The above unperturbed process closely resembles the baseline strategy update process described in step 5 of sample experimentation dynamics. The difference is that the above process uses the true better reply set, whereas step 5 uses a better reply set constructed from experimentation over a phase. However, by Claim 3.5, for any probability $\bar{p} < 1$, acceptable tolerance level δ, and acceptable exploration rate ϵ, there exists a minimum exploration phase length $\underline{m}$ such that for any exploration phase length $m > \underline{m}$, each player's better response set will contain only and all actions that are a strict better response with at least probability $\bar{p}$.

With parameters selected according to Claim 3.5, the transitions of the baseline joint actions in sample experimentation dynamics follow that of the above unperturbed better reply process with probability $\bar{p}$ arbitrarily close to 1. Since the recurrence classes of the unperturbed process are only Nash equilibria, we can conclude from Theorem A.1 that as $\bar{p}$ approaches 1, the probability that the baseline action for sufficiently large t_1 will be a (pure) Nash equilibrium can be made arbitrarily close to 1. By selecting the exploration probability ϵ sufficiently small, we can also conclude that the joint action during exploration phases, i.e., $a(mt_1 + t_2)$, will also be a Nash equilibrium with probability arbitrarily close to 1. □

3.3.2. Noisy utility measurements. Suppose that each player receives a noisy measurement of his true utility, i.e.,

$$\tilde{U}_i(a_i, a_{-i}) = U_i(a_i, a_{-i}) + \nu_i,$$

where n_i is an i.i.d. random variable with zero mean. In the regret testing algorithm with noisy utility measurements, the average utility for action a_i during the exploration phase at time t_1 is now

$$\hat{V}_i^{a_i}(t_1) = \begin{cases} \frac{1}{n_i^{a_i}(t_1)} \sum_{t_2=1}^{m} I\{a_i = \check{a}_i(t_1, t_2)\} \tilde{U}_i(\check{a}(t_1, t_2)), & n_i^{a_i}(t_1) > 0, \\ U_{\min}, & n_i^{a_i}(t_1) = 0. \end{cases}$$

A straightforward modification of the proof of Theorem 3.3 leads to the following theorem.

THEOREM 3.4. *Let G be a finite n-player weakly acyclic game where players' utilities are corrupted with a zero mean noise process. If all players use the sample experimentation dynamics, then for any*

- *probability* $p < 1$,
- *tolerance level* $\delta \in (0, \alpha)$,
- *inertia* $\omega \in (0, 1)$, *and*
- *exploration rate* ϵ *satisfying* $\min\{(\alpha - \delta)/4, \delta/4, 1 - p\} > (1 - (1 - \epsilon)^n) > 0$,

if the exploration phase length m *is sufficiently large, then for all sufficiently large times* $t > 0$, $a(t)$ *is a Nash equilibrium of* G *with at least probability* p.

3.3.3. Comment on length and synchronization of players' exploration phases. In the proof of Theorem 3.3, we assumed that all players' exploration phases were synchronized and of the same length. This assumption was used to ensure that when a player assessed the performance of a particular action, the baseline action of the other players remained constant. Because of the players' inertia this assumption is unnecessary. The general idea is as follows: a player will repeat a baseline action regardless of the better response set with positive probability because of the inertia. Therefore, if all players repeat their baseline action a sufficient number of times, which happens with positive probability, then the joint baseline action would remain constant long enough for any player to evaluate an accurate better response set for that particular joint baseline action.

4. Influencing Nash equilibria in resource allocation problems.

In this section we will derive an approach for influencing the Nash equilibria of a resource allocation problem using the idea of marginal cost pricing. We will illustrate the setup and our approach on a congestion game which is an example of a resource allocation problem.

4.1. Congestion game setup. We consider a transportation network with a finite set R of road segments (or resources) that needs to be shared by a set of selfish drivers labeled as $D := \{d_1, \dots, d_n\}$. Each driver has a fixed origin/destination pair connected through multiple routes. The set of all routes available to driver d_i is denoted by $\mathcal{A}_i$. A route $a_i \in \mathcal{A}_i$ consists of multiple road segments, therefore, $a_i \subset R$. Player $\mathcal{P}_i$ taking route a_i incurs a cost c_r for each road segment $r \in a_i$. The utility of driver d_i taking route a_i is defined as the negative of the total cost incurred, i.e., $U_i = -\sum_{r \in a_i} c_r$. Of course, the utility of each driver will depend on the routes chosen by other drivers.

If we assume that the cost incurred in a road segment depends *only* on the total number of drivers sharing that road, then drivers are anonymous, and this leads to a *congestion game* [19]. The utility of driver d_i is now stated more precisely as

$$U_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)),$$

where $a := (a_1, \dots, a_n)$ is the profile of routes chosen by all drivers and $\sigma_r(a)$ is the total number of drivers using the road segment r.

It is known that a congestion game admits the following potential function:

$$\hat{\phi}(a) = \sum_{r \in R} \sum_{k=1}^{\sigma_r(a)} c_r(k).$$

Unfortunately, this potential function lacks practical significance for measuring the effectiveness of a routing strategy in terms of the overall congestion.

4.2. Congestion game with tolls setup. One approach for equilibrium manipulation is to influence drivers' utilities with tolls [21]. In a congestion game with tolls, a driver's utility takes on the form

$$U_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)) + t_r(\sigma_r(a)),$$

where $t_r(k)$ is the toll imposed on route r if there are k users.

Suppose that the global planner is interested in minimizing the total congestion experienced by all drivers on the network, which can be evaluated as

$$T_c(a) := \sum_{r \in R} \sigma_r(a) c_r(\sigma_r(a)).$$

It has been shown that there exists a set of tolls such that the potential function associated with the congestion game with tolls is aligned with the total congestion experienced by all drivers on the network (see [15, Proposition 4.1]).

PROPOSITION 4.1. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (k-1)[c_r(k) - c_r(k-1)] \quad \forall k \geq 1,$$

then the total negative congestion experienced by all drivers, $\phi_c(a) = -T_c(a)$, *is a potential function for the congestion game with tolls.*

This tolling scheme results in drivers' local utility functions being aligned with the global objective of minimal total congestion.

Now suppose that the global planner is interested in minimizing a *more general* measure,[4]

$$\phi(a) := \sum_{r \in R} f_r(\sigma_r(a)) c_r(\sigma_r(a)), \tag{4.1}$$

where $f_r : \{0, 1, 2, \ldots\} \to \mathbb{R}$ is any arbitrary function. An example of an objective function that fits within this framework and may be practical for general resource allocation problems is

$$\phi(a) = \sum_{r \in R} c_r(\sigma_r(a)).$$

We will now show that there exists a set of tolls, $t_r(\cdot)$, such that the potential function associated with the congestion game with tolls will be aligned with the global planner's objective function of the form given in (4.1).

PROPOSITION 4.2. *Consider a congestion game of any network topology. If the imposed tolls are set as*

$$t_r(k) = (f_r(k) - 1)c_r(k) - f_r(k-1)c_r(k-1) \quad \forall k \geq 1,$$

then the global planners objective, $\phi_c(a) = -\phi(a)$, *is a potential function for the congestion game with tolls.*

[4] In fact, if $c_r(\sigma_r(a)) \neq 0$ for all a, then (4.1) is equivalent to $\sum_{r \in R} f_r(\sigma_r(a))$.

Proof. Let $a^1 = \{a_i^1, a_{-i}\}$ and $a^2 = \{a_i^2, a_{-i}\}$. We will use the shorthand notation $\sigma_r^{a^1}$ to represent $\sigma_r(a^1)$. The change in utility incurred by driver d_i in changing from route a_i^2 to route a_i^1 is

$$\begin{aligned} U_i(a^1) - U_i(a^2) &= -\sum_{r \in a_i^1} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big) \\ &= -\sum_{r \in a_i^1 \setminus a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2 \setminus a_i^1} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big). \end{aligned}$$

The change in the total negative congestion from the joint action a^2 to a^1 is

$$\phi_c(a^1) - \phi_c(a^2) = -\sum_{r \in (a_i^1 \cup a_i^2)} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big).$$

Since

$$\sum_{r \in (a_i^1 \cap a_i^2)} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) = 0,$$

the change in the total negative congestion is

$$\begin{aligned} &\phi_c(a^1) - \phi_c(a^2) \\ &\quad = -\sum_{r \in a_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) \\ &\qquad - \sum_{r \in a_i^2 \setminus a_i^1} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big). \end{aligned}$$

Expanding the first term, we obtain

$$\begin{aligned} &\sum_{r \in a_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - f_r(\sigma_r^{a^2}) c_r(\sigma_r^{a^2})\big) \\ &\qquad = \sum_{r \in a_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - (f_r(\sigma_r^{a^1} - 1)) c_r(\sigma_r^{a^1} - 1)\big) \\ &\qquad = \sum_{r \in a_i^1 \setminus a_i^2} \big(f_r(\sigma_r^{a^1}) c_r(\sigma_r^{a^1}) - ((f_r(\sigma_r^{a^1}) - 1) c_r(\sigma_r^{a^1}) - t_r(\sigma_r^{a^1}))\big) \\ &\qquad = \sum_{r \in a_i^1 \setminus a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big). \end{aligned}$$

Therefore,

$$\begin{aligned} \phi_c(a^1) - \phi_c(a^2) &= -\sum_{r \in a_i^1 \setminus a_i^2} \big(c_r(\sigma_r^{a^1}) + t_r(\sigma_r^{a^1})\big) + \sum_{r \in a_i^2 \setminus a_i^1} \big(c_r(\sigma_r^{a^2}) + t_r(\sigma_r^{a^2})\big) \\ &= U_i(a^1) - U_i(a^2). \qquad \square \end{aligned}$$

By implementing the tolling scheme set forth in Proposition 4.2, we guarantee that all action profiles that minimize the global planner's objective are equilibrium of the congestion game with tolls.

In the special case that $f_r(\sigma_r(a)) = \sigma_r(a)$, Proposition 4.2 produces the same tolls as Proposition 4.1.

5. Illustrative example—congestion game. We will consider a discrete representation of the congestion game setup considered in Braess' paradox [3]. In our setting, there are 1000 vehicles that need to traverse through the network. The network topology and associated congestion functions are illustrated in Figure 6. Each vehicle can select one of the four possible paths to traverse across the network.

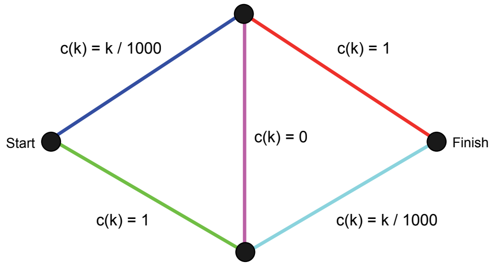

FIG. 6. *Congestion game setup.*

The reason for using this setup as an illustration of the learning algorithms and equilibrium manipulation approach developed in this paper is that the Nash equilibrium of this particular congestion game is easily identifiable. The unique Nash equilibrium is when all vehicles take the route as highlighted in Figure 7. At this Nash equilibrium each vehicle has a utility of 2 and the total congestion is 2000.

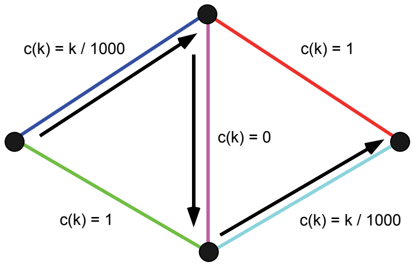

FIG. 7. *Illustration of Nash equilibrium in proposed congestion game.*

Since a potential game is weakly acyclic, the payoff-based learning dynamics in this paper are applicable learning algorithms for this congestion game. In a congestion game, a payoff-based learning algorithm means that drivers have access *only* to the actual congestion experienced. Drivers are unaware of the congestion level on any alternative routes. Figure 8 shows the evolution of drivers on routes when using the simple experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. One can observe that the vehicles' collective behavior does indeed approach that of the Nash equilibrium.

In this congestion game, it is also easy to verify that this vehicle distribution does not minimize the total congestion experience by all drivers over the network. The distribution that minimizes the total congestion over the network is when half the

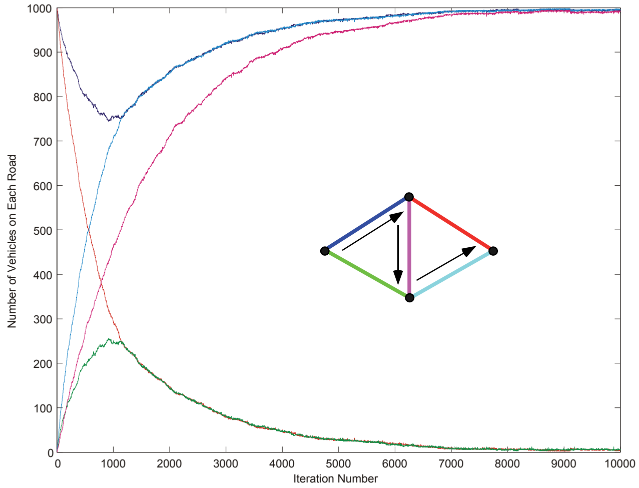

FIG. 8. *Evolution of number of vehicles on each road using simple experimentation dynamics: the number of vehicles on the roads highlighted by arrows approaches* 1000 *while the number of vehicles on all remaining roads approaches* 0.

vehicles occupy the top two roads and the other half occupy the bottom two roads. The middle road is irrelevant.

One can employ the tolling scheme developed in the previous section to locally influence vehicle behavior to achieve this objective. In this setting, the new cost functions, i.e., congestion plus tolls, are illustrated in Figure 9.

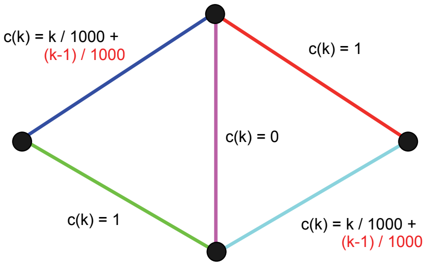

FIG. 9. *Congestion game setup with tolls to minimize total congestion.*

Figure 10 shows the evolution of drivers on routes when using the simple experimentation dynamics. This simulation used an experimentation rate of $\epsilon = 0.25\%$. When using this tolling scheme, the vehicles' collective behavior approaches the new Nash equilibrium which now minimizes the total congestion experienced on the network. The total congestion experienced on the network is now approximately 1500.

There are other tolling schemes that would have resulted in the desired allocation. One approach is to assign an infinite cost to the middle road, which is equivalent to removing it from the network. Under this scenario, the unique Nash equilibrium is for half the vehicles to occupy the top route and the other half to occupy the bottom,

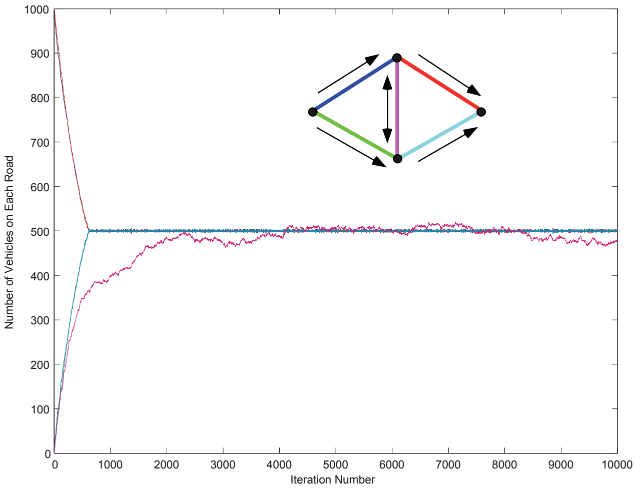

FIG. 10. *Evolution of number of vehicles on each road using simple experimentation dynamics with optimal tolls: the number of vehicles on the middle road fluctuates around* 500 *while the number of vehicles on all remaining roads stabilizes to around* 500*.*

which would minimize the total congestion on the network. Therefore, the existence of this extra road, even though it has zero cost, resulted in the unique Nash equilibrium having a higher total congestion. This is Braess' paradox [3].

The advantage of the tolling scheme set forth in this paper is that it gives a systematic method for influencing the Nash equilibria of any congestion game. We would like to highlight that this tolling scheme guarantees only that the action profiles that maximize the desired objective function are Nash equilibria of the new congestion game with tolls. However, it does not guarantee the lack of suboptimal Nash equilibria.

In many applications, players may not have access to their true utility, but do have access to a noisy measurement of their utility. For example, in the traffic setting, this noisy measurement could be the result of accidents or weather conditions. We will revisit the original congestion game (without tolls) as illustrated in Figure 6. We will now assume that a driver's utility measurement takes on the form

$$\tilde{U}_i(a) = -\sum_{r \in a_i} c_r(\sigma_r(a)) + \nu_i,$$

where ν_i is a random variable with zero mean and variance of 0.1. We will assume that the noise is driver specific rather than road specific.

Figure 11 shows a comparison of the evolution of drivers on routes when using the simple and sample experimentation dynamics. The simple experimentation dynamics simulation used an experimentation rate $\epsilon = 0.25\%$. The sample experimentation dynamics simulation used an exploration rate $\epsilon = 0.25\%$, a tolerance level $\delta = 0.002$, an exploration phase length $m = 500000$, and inertia $\omega = 0.85$. As expected, the noisy utility measurements influenced vehicle behavior more in the simple experimentation dynamics than the sample experimentation dynamics.

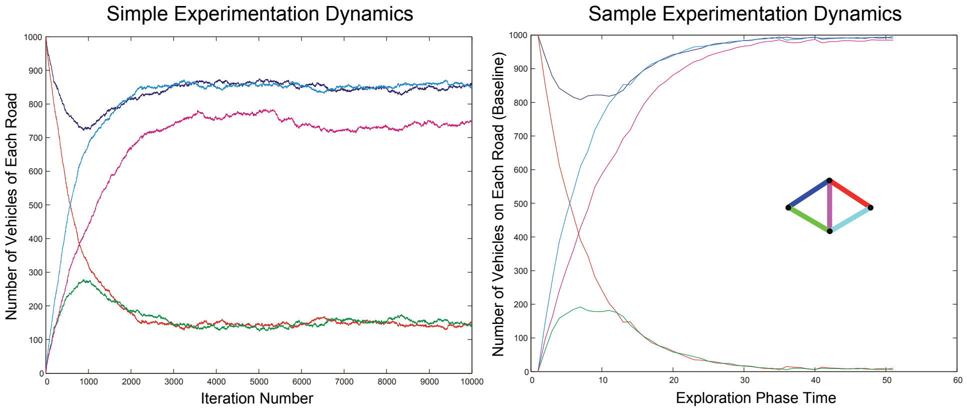

FIG. 11. *Comparison of evolution of number of vehicles on each road using simple experimentation dynamics and sample experimentation dynamics (baseline) with noisy utility measurements: the number of vehicles on the route (upper left, middle, lower right) dominates the number of vehicles on all remaining roads in both settings.*

6. Concluding remarks. We have introduced safe experimentation dynamics for identical interest games, simple experimentation dynamics for weakly acyclic games with noise-free utility measurements, and sample experimentation dynamics for weakly acyclic games with noisy utility measurements. For all three settings, we have shown that for sufficiently large times, the joint action taken by all players will constitute a Nash equilibrium. Furthermore, we have shown how to guarantee that a collective objective in a congestion game is a (nonunique) Nash equilibrium. An important, but unaddressed, topic in this work is characterizing resulting convergence rates. It is likely that tools regarding mixing times of Markov chains [17] will be relevant.

Our motivation has been that in many engineered systems, the functional forms of utility functions are not available, and so players must adjust their strategies through an adaptive process using only payoff measurements. In the dynamic processes defined here, there is no explicit cooperation or communication between players. On the one hand, this lack of explicit coordination offers an element of robustness to a variety of uncertainties in the strategy adjustment processes. Nonetheless, on the other hand, an interesting future direction would be to investigate to what degree explicit coordination through limited communications could be beneficial.

Appendix. Background on resistance trees. For a detailed review of the theory of resistance trees, please see [24].

Let P^0 denote the probability transition matrix for a finite state Markov chain over the state space Z. Consider a "perturbed" process such that the size of the perturbations can be indexed by a scalar $\epsilon > 0$, and let P^ϵ be the associated transition probability matrix. The process P^ϵ is called a *regular perturbed Markov process* if P^ϵ is ergodic for all sufficiently small $\epsilon > 0$ and P^ϵ approaches P^0 at an exponentially smooth rate [24]. Specifically, the latter condition means that for all $z, z' \in Z$,

$$\lim_{\epsilon \to 0^+} P^\epsilon_{zz'} = P^0_{zz'},$$

and

$$P^\epsilon_{zz'} > 0 \text{ for some } \epsilon > 0 \;\Rightarrow\; 0 < \lim_{\epsilon \to 0^+} \frac{P^\epsilon_{zz'}}{\epsilon^{r(z \to z')}} < \infty$$

for some nonnegative real number $r(z \to z')$, which is called the *resistance* of the transition $z \to z'$. (Note in particular that if $P^0_{zz'} > 0$, then $r(z \to z') = 0$.)

Let the recurrence classes of P^0 be denoted by $E_1, E_2, \ldots, E_N$. For each pair of distinct recurrence classes E_i and E_j, $i \neq j$, an ij-path is defined to be a sequence of distinct states $\zeta = (z_1 \to z_2 \to \cdots \to z_n)$ such that $z_1 \in E_i$ and $z_n \in E_j$. The resistance of this path is the sum of the resistances of its edges, that is, $r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + \cdots + r(z_{n-1} \to z_n)$. Let $\rho_{ij} = \min r(\zeta)$ be the least resistance over all ij-paths ζ. Note that ρ_{ij} must be positive for all distinct i and j, because there exists no path of zero resistance between distinct recurrence classes.

Now construct a complete directed graph with N vertices, one for each recurrence class. The vertex corresponding to class E_j will be called j. The weight on the directed edge $i \to j$ is ρ_{ij}. A tree, T, rooted at vertex j, also called a j-tree, is a set of $N-1$ directed edges such that, from every vertex different from j, there is a unique directed path in the tree to j. The resistance of a rooted tree, T, is the sum of the resistances ρ_{ij} on the $N-1$ edges that compose it. The *stochastic potential*, γ_j, of the recurrence class E_j is defined to be the minimum resistance over all trees rooted at j. The following theorem gives a simple criterion for determining the stochastically stable states (see [24, Theorem 4]).

THEOREM A.1. *Let P^ϵ be a regular perturbed Markov process, and for each $\epsilon > 0$ let μ^ϵ be the unique stationary distribution of P^ϵ. Then $\lim_{\epsilon\to 0} \mu^\epsilon$ exists and the limiting distribution μ^0 is a stationary distribution of P^0. The stochastically stable states (i.e., the support of μ^0) are precisely those states contained in the recurrence classes with minimum stochastic potential.*

REFERENCES

[1] G. ARSLAN, J. R. MARDEN, AND J. S. SHAMMA, *Autonomous vehicle-target assignment: A game theoretical formulation*, ASME J. Dynam. Systems Measurement and Control, 129 (2007), pp. 584–596.

[2] V. S. BORKAR AND P. R. KUMAR, *Dynamic Cesaro-Wardrop equilibration in networks*, IEEE Trans. Automat. Control, 48 (2003), pp. 382–396.

[3] D. BRAESS, *Uber ein paradoxen der verkehrsplanning*, Unternehmensforschung, 12 (1968), pp. 258–268.

[4] D. P. FOSTER AND H. P. YOUNG, *Regret testing: Learning to play Nash equilibrium without knowing you have an opponent*, Theoret. Econom., 1 (2006), pp. 341–367.

[5] D. FUDENBERG AND D. K. LEVINE, *The Theory of Learning in Games*, MIT Press, Cambridge, MA, 1998.

[6] D. FUDENBERG AND J. TIROLE, *Game Theory*, MIT Press, Cambridge, MA, 1991.

[7] A. GANGULI, S. SUSCA, S. MARTINEZ, F. BULLO, AND J. CORTES, *On collective motion in sensor networks: Sample problems and distributed algorithms*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 4239–4244.

[8] F. GERMANO AND G. LUGOSI, *Global Nash convergence of Foster and Young's regret testing*, Games Econom. Behavior, 60 (2007), pp. 135–154.

[9] S. B. GERSHWIN, *Manufacturing Systems Engineering*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[10] S. HART, *Adaptive heuristics*, Econometrica, 73 (2005), pp. 1401–1430.

[11] J. HOFBAUER AND K. SIGMUND, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[12] S. MANNOR AND J. S. SHAMMA, *Multi-agent learning for engineers*, Artificial Intelligence, 171 (2007), pp. 417–422.

[13] J. R. MARDEN, G. ARSLAN, AND J. S. SHAMMA, *Connections between cooperative control and potential games illustrated on the consensus problem*, in Proceedings of the 2007 European Control Conference (ECC '07), Kos, Greece, 2007, pp. 4604–4611.

[14] J. R. MARDEN, G. ARSLAN, AND J. S. SHAMMA, *Regret based dynamics: Convergence in weakly acyclic games*, in Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Honolulu, HI, 2007, article 42.

[15] J. R. Marden, G. Arslan, and J. S. Shamma, *Joint strategy fictitious play with inertia for potential games*, IEEE Trans. Automat. Control, to appear.

[16] D. Monderer and L. S. Shapley, *Fictitious play property for games with identical interests*, J. Econom. Theory, 68 (1996), pp. 258–265.

[17] R. Montenegro and P. Tetali, *Mathematical Aspects of Mixing Times in Markov Chains*, Now Publishers, Hanover, MA, 2006.

[18] L. Panait and S. Luke, *Cooperative multi-agent learning: The state of the art*, Autonomous Agents and Multi-Agent Systems, 11 (2005), pp. 387–434.

[19] R. W. Rosenthal, *A class of games possessing pure-strategy Nash equilibria*, Internat. J. Game Theory, 2 (1973), pp. 65–67.

[20] L. Samuelson, *Evolutionary Games and Equilibrium Selection*, MIT Press, Cambridge, MA, 1997.

[21] W. Sandholm, *Evolutionary implementation and congestion pricing*, Rev. Econom. Stud., 69 (2002), pp. 667–689.

[22] Y. Shoham, R. Powers, and T. Grenager, *If multi-agent learning is the answer, what is the question?*, Artificial Intelligence, 171 (2007), pp. 365–377.

[23] J. W. Weibull, *Evolutionary Game Theory*, MIT Press, Cambridge, MA, 1995.

[24] H. P. Young, *The evolution of conventions*, Econometrica, 61 (1993), pp. 57–84.

[25] H. P. Young, *Individual Strategy and Social Structure*, Princeton University Press, Princeton, NJ, 1998.

[26] H. P. Young, *Strategic Learning and Its Limits*, Oxford University Press, Oxford, UK, 2005.