# Achieving Pareto Optimality Through Distributed Learning

Jason R. Marden, H. Peyton Young, and Lucy Y. Pao

**Abstract**

We propose a simple payoff-based learning rule that is completely decentralized, and that leads to an efficient configuration of actions in any $n$-person finite strategic-form game with generic payoffs. The algorithm follows the theme of exploration versus exploitation and is hence stochastic in nature. We prove that if all agents adhere to this algorithm, then the agents will select the action profile that maximizes the sum of the agents' payoffs a high percentage of time. The algorithm requires no communication. Agents respond solely to changes in their own realized payoffs, which are affected by the actions of other agents in the system in ways that they do not necessarily understand. The method can be applied to the optimization of complex systems with many distributed components, such as the routing of information in networks and the design and control of wind farms. The proof of the proposed learning algorithm relies on the theory of large deviations for perturbed Markov chains.

## I. Introduction

Game theory has important applications to the design and control of multiagent systems [1]–[9]. This design choice requires two steps. First, the system designer must model the system components as "agents" embedded in an interactive, game-theoretic environment. This step involves defining a set of choices and a local objective function for each agent. Second, the system designer must specify the agents' behavioral rules, i.e., the way in which they react to local conditions and information. The goal is to complete both steps in such a way that the

This research was supported by AFOSR grants #FA9550-09-1-0538 and #FA9550-12-1-0359, ONR grants #N00014-09-1-0751 and #N00014-12-1-0643, and the Center for Research and Education in Wind.

J. R. Marden is with the Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO 80309, jason.marden@colorado.edu. Corresponding author.

H. Peyton Young is with the Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom, peyton.young@nuffield.ox.ac.uk.

Lucy Y. Pao is with the Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO 80309, pao@colorado.edu.

agents' behavior leads to desirable system wide behavior even though the agents themselves do not have access to the information needed to determine the state of the system.

The existing literature focuses primarily on distributed learning algorithms that are suitable for implementation in engineering systems. Accordingly, most of the results in distributed learning are concerned with the attainment of (pure) Nash equilibria for particular classes of games that arise in distributed engineering systems [6], [7], [10]–[20]. For example, [10]–[12] establish learning algorithms that converge to Nash equilibria for games that are relevant to mobile sensor networks, while [6], [7], [17], [18] establish learning algorithms that converge to Nash equilibria for *potential games* [21].

There are, however, two limitations to this body of work. First, as highlighted above, most results in this domain focus on convergence to Nash equilibrium, which may be very inefficient in achieving the system level objective. Characterizing this inefficiency is a highly active research area in algorithmic game theory [22]. The second limitation of this framework is that it is frequently impossible to represent the interaction framework of a given system as a potential game, because engineering systems possess inherent constraints on the types of objective functions that can be assigned to the agents. These constraints are a byproduct of the information available to different components of the system. Furthermore, in many complex systems, the relationship between the behavior of the components and the overall system performance is not known with any precision.

One example of a system that exhibits these challenges is the control of a wind farm to maximize total power production [23]. Controlling an array of turbines in a wind farm is fundamentally more challenging than controlling a single turbine. The reason is the aerodynamic interactions amongst the turbines, which render many of the single turbine control algorithms *highly inefficient* for optimizing total power production [24]. Here, the goal is to establish a *distributed* control algorithm that enables the individual turbines to adjust their behavior based on local conditions, so as to maximize total system performance. One way to handle this large-scale coordination problem is to model the interactions of the turbines in a game theoretic environment. However, the space of admissible utility functions for the individual turbines is limited because of the following informational constraints:

(i) No turbine has access to the actions[1] of other turbines, due to the lack of a suitable communication system;

(ii) No turbine has access to the functional relationship between the total power generated and the action of the other turbines. The reason is that the aerodynamic interaction between the turbines is poorly understood from an engineering standpoint.

These limitations restrict the ability of the designer to represent the interaction framework as a potential game. For example, one of the common design approaches is to assign each turbine an objective function that measures the turbine's marginal contribution to the power production of the wind farm, that is, the difference between the total power produced when the turbine is active and the total power produced when the turbine is inactive [6], [25]. This assignment ensures that the resulting interaction framework is a potential game and that the action profile which optimizes the potential function also optimizes the total power production of the wind farm. Calculating this marginal contribution may not be possible due to lack of knowledge about the aerodynamic interactions, hence the existing literature does not provide suitable control algorithms for this situation.

The contribution of this paper is to demonstrate the existence of simple, completely decentralized learning algorithms that lead to efficient system-wide behavior *irrespective* of the game structure. We measure the efficiency of an action profile by the sum of the agents' utility functions. In a wind farm, this sum is precisely equal to the total power generated. Our main result is the development of a simple payoff-based learning algorithm that guarantees convergence to an efficient action profile whenever the underlying game has generic payoffs. This result holds whether or not this efficient action profile is a Nash equilibrium. It therefore differs from the approach of [20], which shows how to achieve constrained efficiency *within* the set of Nash equilibrium outcomes.

In the prior literature, the principal approaches to distributed or networked optimization are subgradient methods [26]–[31], consensus based methods [32]–[34], and two-step consensus based methods [35], [36]. A key difference between our proposed algorithm and these earlier methods is that the latter depend heavily on the particular structure of the objective function

[1]In our related work [23], a turbine's action is called an *axial induction factor*. The axial induction factor indicates the fractional amount of power the turbine extracts from the wind.

and also on the form of the underlying network. For example, a common requirement is that the objective function be convex and the communication graph must be connected. This stands in sharp contrast to our approach where there is no communication between the agents and the objective function is completely general.

## II. Background

Let $G$ be a finite strategic-form game with $n$ agents. The set of agents is denoted by $N := \{1, ..., n\}$. Each agent $i \in N$ has a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the joint action set. We shall henceforth refer to a finite strategic-form game simply as "a game." Given an action profile $a = (a_1, a_2, ..., a_n) \in \mathcal{A}$, let $a_{-i}$ denote the profile of agent actions *other than* agent $i$, that is, $a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$. With this notation, we shall sometimes denote a profile $a$ of actions by $(a_i, a_{-i})$ and $U_i(a)$ by $U_i(a_i, a_{-i})$. We shall also let $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ denote the set of possible collective actions of all agents other than agent $i$. The *welfare* of an action profile $a \in \mathcal{A}$ is defined as

$$W(a) = \sum_{i \in N} U_i(a).$$

An action profile $a$ is said to be *efficient* if the action profile $a$ optimizes the welfare, i.e, $a \in \arg\max_{a' \in \mathcal{A}} W(a')$.

### A. Repeated Games

We shall assume that a given game $G$ is repeated once each period $t \in \{0, 1, 2, \ldots\}$. In period $t$, the agents simultaneously choose actions $a(t) = (a_1(t), ..., a_n(t))$ and receive payoffs $U_i(a(t))$. Agent $i \in N$ chooses the action $a_i(t)$ according to a probability distribution $p_i(t) \in \Delta(\mathcal{A}_i)$, which is the simplex of probability distributions over $\mathcal{A}_i$. We shall refer to $p_i(t)$ as the *strategy* of agent $i$ at time $t$. We adopt the convention that $p_i^{a_i}(t)$ is the probability that agent $i$ selects action $a_i$ at time $t$ according to the strategy $p_i(t)$. An agent's strategy at time $t$ relies only on observations from times $\{0, 1, 2, ..., t-1\}$.

Different learning algorithms are specified by the agents' information and the mechanism by which their strategies are updated as information is gathered. Suppose, for example, that an agent knows his own utility function and is capable of observing the actions of all other agents at every

time step but does not know their utility functions. Then the strategy adjustment mechanism of a given agent $i$ can be written in the form

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

Such an algorithm is said to be *uncoupled* [37], [38].

In this paper, we ask whether agents can learn to play the welfare maximizing action profile under even more restrictive observational conditions. In particular, we shall assume that agents *only* have access to: (i) the action they played and (ii) the payoff they received. In this setting, the strategy adjustment mechanism of agent $i$ takes the form

$$p_i(t) = F_i\left(\{a_i(\tau), U_i(a(\tau))\}_{\tau=0,...,t-1}\right). \tag{2}$$

Such a learning rule is said to be *completely uncoupled* or *payoff-based* [39]. Recent work has shown that for finite games with generic payoffs, there exist completely uncoupled learning rules that lead to Pareto optimal Nash equilibria [20]; see also [19], [40], [41]. Here we exhibit a different class of learning procedures that lead to Pareto optimal outcomes whether or not they are Nash equilibria.[2]

## III. A Payoff-Based Algorithm for Maximizing Welfare

Our proposed algorithm, like many others, exploits the tradeoff between exploration versus exploitation. Each agent possesses a baseline action that he expects to play and a baseline utility that he expects to receive. Each agent also possesses an internal state variable, which we refer to as a *mood*, which determines the agent's underlying behavior. There are two distinct moods: "content" and "discontent". When an agent is content he selects his baseline action with high probability. When an agent is discontent, there is a high probability that he selects an action that differs from the baseline action. Upon selecting an action and receiving a payoff, an agent updates his mood by comparing the action played and the payoff received with his baseline

[2]Such a result might seem reminiscent of the Folk Theorem, which specifies conditions under which an efficient action profile can be implemented as an equilibrium of a repeated game (see among others [41], [42]). In the present context, however, we are interested in whether agents can learn to play an efficient action profile without having any information about the game as a whole or what the other agents are doing. Hence, they cannot condition their behavior on the observed behavior of others, which is a key requirement of most repeated game equilibria.

action and baseline payoff. By defining this updating rule appropriately, we demonstrate that the learning process leads to a stochastically stable action profile that is in fact efficient.

Our proposed algorithm is a variant of the approach in [20], where each agent also possesses an internal state variable which impacts the agent's behavior rule. The key difference between our algorithm and the one in [20] is the asymptotic guarantees. In particular, [20] guarantees convergence to a Pareto Nash equilibrium, whereas our proposed algorithm converges to a Pareto (efficient) action profile irrespective of whether or not this action profile is a Nash equilibrium. Furthermore, our algorithm uses fewer state variables than the method in [20].

At each point in time an agent's *state* can be represented as a triple $[\bar{a}_i, \bar{u}_i, m_i]$, where

- The ***benchmark action*** is $\bar{a}_i \in \mathcal{A}_i$.
- The ***benchmark payoff*** is $\bar{u}_i$, which is in the range of $U_i(\cdot)$.
- The ***mood*** is $m_i$, which can take on two values: *content* (C) and *discontent* (D).

The learning algorithm produces a sequence of action profiles $a(1)$, ..., $a(t)$, where the behavior of an agent $i$ in each period $t = 1, 2, ...,$ is conditioned on agent $i$'s underlying benchmark payoff $\bar{u}_i(t)$, benchmark action $\bar{a}_i(t)$, and mood $m_i(t) \in \{C, D\}$.

We divide the dynamics into the following two parts: the agent dynamics and the state dynamics. Without loss of generality, we focus on the case where agent utility functions are strictly bounded between 0 and 1, i.e., for any agent $i \in N$ and action profile $a \in \mathcal{A}$ we have $1 > U_i(a) \geq 0$. Consequently, for any action profile $a \in \mathcal{A}$, the welfare function satisfies $n > W(a) \geq 0$.

***Agent Dynamics:*** Fix an experimentation rate $1 > \epsilon > 0$ and constant $c > n$. Let $[\bar{a}_i, \bar{u}_i, m_i]$ be the current state of agent $i$.

- **Content** $(m_i = C)$: In this state, the agent chooses an action $a_i$ according to the following probability distribution

$$p_i^{a_i} = \begin{cases} \frac{\epsilon^c}{|\mathcal{A}_i|-1} & \text{for } a_i \neq \bar{a}_i \\ 1 - \epsilon^c & \text{for } a_i = \bar{a}_i \end{cases} \tag{3}$$

  where $|\mathcal{A}_i|$ represents the cardinality of the set $\mathcal{A}_i$.

- **Discontent** $(m_i = D)$: In this state, the agent chooses an action $a_i$ according to the following probability distribution:

$$p_i^{a_i} = \frac{1}{|\mathcal{A}_i|} \quad \text{for every} \quad a_i \in \mathcal{A}_i \tag{4}$$

Note that the benchmark action and utility play no role in the agent dynamics when the agent is discontent.

***State Dynamics:*** Once the agent selects an action $a_i \in \mathcal{A}_i$ and receives the payoff $u_i = U_i(a_i, a_{-i})$, where $a_{-i}$ is the action selected by all agents other than agent $i$, the state is updated as follows:

- **Content** $(m_i = C)$: If $[a_i, u_i] = [\bar{a}_i, \bar{u}_i]$, the new state is determined by the transition

$$[\bar{a}_i, \bar{u}_i, C] \overset{[\bar{a}_i, \bar{u}_i]}{\longrightarrow} [\bar{a}_i, \bar{u}_i, C]. \tag{5}$$

  If $[a_i, u_i] \neq [\bar{a}_i, \bar{u}_i]$, the new state is determined by the transition

$$[\bar{a}_i, \bar{u}_i, C] \overset{[a_i, u_i]}{\longrightarrow} \begin{cases} [a_i, u_i, C] & \text{with prob } \epsilon^{1-u_i} \\ [a_i, u_i, D] & \text{with prob } 1 - \epsilon^{1-u_i}. \end{cases}$$

- **Discontent** $(m_i = D)$: If the selected action and received payoff are $[a_i, u_i]$, the new state is determined by the transition

$$[\bar{a}_i, \bar{u}_i, D] \overset{[a_i, u_i]}{\longrightarrow} \begin{cases} [a_i, u_i, C] & \text{with prob } \epsilon^{1-u_i} \\ [a_i, u_i, D] & \text{with prob } 1 - \epsilon^{1-u_i}. \end{cases}$$

To ensure that the dynamics converge to an efficient action profile, we require the following notion of interdependence in the game structure [19].

**Definition 1** (Interdependence)**.** *An $n$-person game $G$ on the finite action space $\mathcal{A}$ is interdependent if, for every $a \in \mathcal{A}$ and every proper subset of agents $J \subset N$, there exists an agent $i \notin J$ and a choice of actions $a'_J \in \prod_{j \in J} \mathcal{A}_j$ such that $U_i(a'_J, a_{-J}) \neq U_i(a_J, a_{-J})$.*

Roughly speaking, the interdependence condition states that it is not possible to divide the agents into two distinct subsets that do not mutually interact with one another.

These dynamics induce a Markov process over the finite state space $Z = \prod_{i \in N} (\mathcal{A}_i \times \mathcal{U}_i \times M)$, where $\mathcal{U}_i$ denotes the finite range of $U_i(a)$ over all $a \in \mathcal{A}$ and $M = \{C, D\}$ is the set of moods. We shall denote the transition probability matrix by $P^\epsilon$ for each $\epsilon > 0$. Computing the stationary distribution of this process is challenging because of the large number of states and the fact that the underlying process is not reversible. Accordingly, we shall focus on characterizing the *support* of the limiting stationary distribution, whose elements are referred to as the *stochastically stable*

*states* [43]. More precisely, a state $z \in Z$ is stochastically stable if and only if $\lim_{\epsilon \to 0^+} \mu(z, \epsilon) > 0$ where $\mu(z, \epsilon)$ is a stationary distribution of the process $P^\epsilon$ for a fixed $\epsilon > 0$.

**Theorem 1.** *Let $G$ be an interdependent $n$-person game on a finite joint action space $\mathcal{A}$. Under the dynamics defined above, a state $z = [\bar{a}, \bar{u}, m] \in Z$ is stochastically stable if and only if the following conditions are satisfied:*

*(i) The action profile $\bar{a}$ optimizes $W(\bar{a}) = \sum_{i \in N} U_i(\bar{a})$.*

*(ii) The benchmark actions and payoffs are aligned, i.e., $\bar{u}_i = U_i(\bar{a})$ for all $i$.*

*(iii) The mood of each agent is content, i.e., $m_i = C$ for all $i$.*

## IV. Proof of Theorem 1

The proof relies on the theory of resistance trees for regular perturbed Markov decision processes [44], which we briefly review here. Let $P^0$ denote the probability transition matrix of a finite state Markov chain on the state space $Z$. Consider a "perturbed" process $P^\epsilon$ on $Z$ where the "size" of the perturbations can be indexed by a scalar $\epsilon > 0$. The process $P^\epsilon$ is called a *regular perturbed Markov process* if $P^\epsilon$ is ergodic for all sufficiently small $\epsilon > 0$ and $P^\epsilon$ approaches $P^0$ at an exponentially smooth rate, that is,

$$\forall z, z' \in Z, \quad \lim_{\epsilon \to 0^+} P^\epsilon_{zz'} = P^0_{zz'},$$

and

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

where $r(z \to z')$ is a nonnegative real number 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 $E_1, E_2, ..., E_M$. 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 $\zeta = (z_1 \to z_2 \to ... \to z_m)$ such that $z_k \in Z$ for all $k \in \{1, \ldots, m\}$, $z_1 \in E_i$, and $z_m \in E_j$. The *resistance* of this path is the sum of the resistances of its edges, that is,

$$r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + ... + r(z_{m-1} \to z_m).$$

Let $\rho_{ij} = \min r(\zeta)$ be the least resistance over all $ij$-paths $\zeta$. Note that $\rho_{ij}$ 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 $M$ 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 $\rho_{ij}$. A *$j$-tree* $T$ is a set of $M-1$ directed edges such that, from every vertex different from $j$, there is a unique directed path in the tree to $j$. The resistance of such a tree is the sum of the resistances on the $M-1$ edges that compose it. The *stochastic potential*, $\gamma_j$, of the recurrence class $E_j$ is the minimum resistance over all trees rooted at $j$. The following result provides a simple criterion for determining the stochastically stable states ([44], Theorem 4).

*Let $P^\epsilon$ be a regular perturbed Markov process, and for each $\epsilon > 0$ let $\mu^\epsilon$ be the unique stationary distribution of $P^\epsilon$. Then $\lim_{\epsilon \to 0} \mu^\epsilon$ exists and the limiting distribution $\mu^0$ is a stationary distribution of $P^0$. The stochastically stable states (i.e., the support of $\mu^0$) are precisely those states contained in the recurrence classes with minimum stochastic potential.*[3]

It can be verified that the dynamics introduced above define a regular perturbed Markov process. The proof of Theorem 1 proceeds by a series of lemmas. Let $C^0$ be the subset of states in which each agent is content and the benchmark action and utility are aligned. That is, if $[\bar{a}, \bar{u}, m] \in C^0$, then $\bar{u}_i = U_i(\bar{a})$ and $m_i = C$ for each agent $i \in N$. Let $D^0$ represent the set of states in which everyone is discontent. That is, if $[\bar{a}, \bar{u}, m] \in D^0$, then $\bar{u}_i = U_i(\bar{a})$ and $m_i = D$ for each agent $i \in N$. Accordingly, for any state in $D^0$, each agent's benchmark action and utility are aligned.

The first lemma provides a characterization of the recurrence classes of the unperturbed process $P^0$.

**Lemma 2.** *The recurrence classes of the unperturbed process $P^0$ are $D^0$ and all singletons $z \in C^0$.*

*Proof:* The set of states $D^0$ represents a single recurrence class of the unperturbed process since the probability of transitioning between any two states $z_1, z_2 \in D^0$ is $O(1)$ and when $\epsilon = 0$ there is no possibility of exiting from $D^0$.[4] Any state $[\bar{a}, \bar{u}, C] \in C^0$ is a recurrent class of the

[3] In Section VI-A, we illustrate how to compute the resistances and stochastic potential of each state in several concrete examples.

[4] The notation $O(1)$ refers to transition probabilities that are bounded away from 0. For the situation highlighted above, the probability of the transition $z_1 \to z_2$ is $1/|\mathcal{A}|$. The notation $O(\epsilon)$ refers to transition probabilities that are on the order of $\epsilon$.

unperturbed process, because all agents will continue to play their baseline action at all future times.

We will now show that the states $D^0$ and all singletons $z \in C^0$ represent the *only* recurrent states. Suppose that a proper subset of agents $S \subset N$ is discontent, and the benchmark actions and benchmark utilities of all other agents are $\bar{a}_{-S}$ and $\bar{u}_{-S}$, respectively. By interdependence, there exists an agent $j \notin S$ and an action tuple $a'_S \in \prod_{i \in S} \mathcal{A}_i$ such that $u_j \neq U_j(a'_S, \bar{a}_{-S})$. This situation cannot be a recurrence class of the unperturbed process because the agent set $S$ will eventually play action $a'_S$ with probability 1, thereby causing agent $j$ to become discontent. Agent set $S$ will eventually play action $a'_S$ with probability 1, because each agent $i \in S$ is discontent and hence selects actions uniformly for the action set $\mathcal{A}_i$. Consequently, at each subsequent period, the action $a'_S$ will be played with probability $1/|\mathcal{A}_S|$. Note that once the agent set $S$ selects action $a'_S$, the payoff of agent $j$ will be different from agent $j$'s baseline utility, i.e., $\bar{u}_j \neq U_j(a'_S, \bar{a}_{-S})$, thereby causing agent $j$ to become discontent. This process can be repeated to show that all agents will eventually become discontent with probability $O(1)$; hence any state that consists of a partial collection of discontent agents $S \subset N$ is not a recurrence class of the unperturbed process.

Lastly, consider a state $[\bar{a}, \bar{u}, C]$ where all agents are content, but there exists at least one agent $i$ whose benchmark action and benchmark utility are not aligned, i.e., $\bar{u}_i \neq U_i(\bar{a})$. For the unperturbed process, at the ensuing time step the action profile $\bar{a}$ will be played and agent $i$ will become discontent since $\bar{u}_i \neq U_i(\bar{a})$. Since one agent is discontent, all agents will eventually become discontent. This completes the proof of Lemma 2. ■

We know from [44] that the computation of the stochastically stable states can be reduced to an analysis of rooted trees on the vertex set consisting solely of the recurrence classes. We denote the collection of states $D^0$ by a single variable $D$ to represent this single recurrence class, since the exit probabilities are the same for all states in $D^0$. By Lemma 2, the set of recurrence classes consists of the singleton states in $C^0$ and also the singleton state $D$. Accordingly, we represent a state $z \in C^0$ by just $[\bar{a}, \bar{u}]$ and drop the extra notation highlighting that the agents are content. We now reiterate the definition of edge resistance.

**Definition 2** (Edge resistance)**.** *For every pair of distinct recurrence classes $w$ and $z$, let $r(w \to z)$ denote the total resistance of the least-resistance path that starts in $w$ and ends in $z$. We call*

$w \to z$ *an edge and* $r(w \to z)$ *the resistance of the edge.*

Let $z = [\bar{a}, \bar{u}]$ and $z' = [\bar{a}', \bar{u}']$ be any two distinct states in $C^0$. The following observations will be useful.

(i) The resistance of the transition $z \to D$ satisfies

$$r(z \to D) = c.$$

To see this, consider any state $z \in C^0$. In order to transition out of the state $z$, at least one agent needs to experiment, which happens with a probability $O(\epsilon^c)$. This experimenting agent will become discontent at the ensuing step with probability $O(1)$. Given this event, Lemma 2 implies that all agents will become discontent with probability $O(1)$. Hence, the resistance of the transition $z \to D$ equals $c$.

(ii) The resistance of the transition $D \to z$ satisfies

$$r(D \to z) = \sum_{i \in N} (1 - \bar{u}_i) = n - W(\bar{a}).$$

According to the state dynamics, transitioning from discontent to content requires that each agent must accept the benchmark payoff $\bar{u}_i$, which has a resistance $(1 - \bar{u}_i)$. Consequently, the resistance associated with this transition is $\sum_{i \in N} (1 - \bar{u}_i) = n - W(\bar{a})$.

(iii) The resistance of the transition $z \to z'$ satisfies

$$c \leq r(z \to z') < 2c.$$

This follows directly from the definition of edge resistance, which requires that $r(z \to z') \leq r(z \to D) + r(D \to z')$. Therefore, each transition of minimum resistance includes at most one agent who experiments.

The following lemma characterizes the stochastic potential of the states in $C^0$. Before stating this lemma, we define a *path* $\mathcal{P}$ over the states $D \cup C^0$ to be a sequence of edges of the form

$$\mathcal{P} = \{z^0 \to z^1 \to ... \to z^m\},$$

where each $z^k$ for $k \in \{0, 1, ..., m\}$ is in $D \cup C^0$. The *resistance* of a path $\mathcal{P}$ is the sum of the resistance of each edge in the path, i.e.,

$$R(\mathcal{P}) = \sum_{k=1}^{m} r(z^{k-1} \to z^k).$$

**Lemma 3.** *The stochastic potential of any state $z = [\bar{a}, \bar{u}]$ in $C^0$ is*

$$\gamma(z) = c\left(\left|C^0\right| - 1\right) + \sum_{i \in N} (1 - \bar{u}_i). \tag{6}$$

*Proof:* We first prove that (6) is an upper bound for the stochastic potential of $z$ by constructing a tree rooted at $z$ with the prescribed resistance. To that end, consider the tree $T$ with the following properties:

**P-1**: The edge exiting each state $z' \in C^0 \setminus \{z\}$ is of the form $z' \to D$. The total resistance associated with these edges is $c\left(|C^0| - 1\right)$.

**P-2**: The edge exiting the state $D$ is of the form $D \to z$. The resistance associated with this edge is $\sum_{i \in N} (1 - \bar{u}_i)$.

The tree $T$ is rooted at $z$ and has total resistance $c\left(|C^0| - 1\right) + \sum_{i \in N} (1 - \bar{u}_i)$. It follows that $\gamma(z) \leq c\left(|C^0| - 1\right) + \sum_{i \in N} (1 - \bar{u}_i)$, hence (6) holds as an inequality. It remains to be shown that the right-hand side of (6) is also a lower bound for the stochastic potential.

We argue this by contradiction. Suppose there exists a tree $T$ rooted at $z$ with resistance $R(T) < c\left(|C^0| - 1\right) + \sum_{i \in N} (1 - \bar{u}_i)$. Since the tree $T$ is rooted at $z$ we know that there exists a path $\mathcal{P}$ from $D$ to $z$ of the form

$$\mathcal{P} = \{D \to z^1 \to z^2 \to ... \to z^m \to z\},$$

where $z^k \in C^0$ for each $k \in \{1, ..., m\}$. We claim that the resistance associated with this path of $m + 1$ transitions satisfies

$$R(\mathcal{P}) \geq mc + \sum_{i \in N} (1 - \bar{u}_i).$$

The term $mc$ comes from applying observation (iii) to the last $m$ transitions on the path $\mathcal{P}$. The term $\sum_{i \in N} (1 - \bar{u}_i)$ comes from the fact that each agent needs to accept $\bar{u}_i$ as the benchmark payoff at some point during the transitions.

Construct a new tree $T'$ still rooted at $z$ by removing the edges in $\mathcal{P}$ and adding the following edges:

- $D \to z$ which has resistance $\sum_{i \in N} (1 - \bar{u}_i)$.
- $z^k \to D$ for each $k \in \{1, ..., m\}$ which has total resistance $mc$.

The new tree $T'$ is still rooted at $z$ and has a total resistance that satisfies $R(T') \leq R(T)$. Note that if the path $\mathcal{P}$ was of the form $D \to z$, then this augmentation does not alter the tree structure.

Now suppose that there exists an edge $z' \to z''$ in the tree $T'$ for some states $z', z'' \in C^0$. By observation (iii), the resistance of this edge satisfies $r(z' \to z'') \geq c$. Construct a new tree $T''$ by removing the edge $z' \to z''$ and adding the edge $z' \to D$, which has a resistance $c$. This new tree $T''$ is rooted at $z$, and its resistance satisfies

$$\begin{aligned} R(T'') &= R(T') + r(z' \to D) - r(z' \to z'') \\ &\leq R(T') \\ &\leq R(T). \end{aligned}$$

Repeat this process until we have constructed a tree $T^*$ for which no such edges exist. Note that the tree $T^*$ satisfies properties P-1 and P-2 and consequently has a total resistance $R(T^*) = c\left(|C^0| - 1\right) + \sum_{i \in N}\left(1 - \bar{u}_i\right)$. Since by construction $R(T^*) \leq R(T)$ we have a contradiction. This completes the proof of Lemma 3. ■

We will now prove Theorem 1 by analyzing the minimum resistance trees using the above lemmas. We first show that the state $D$ is not stochastically stable. Suppose, by way of contradiction, that there exists a minimum resistance tree $T$ rooted at the state $D$. Then there exists an edge in the tree $T$ of the form $z \to D$ for some state $z \in C^0$ and the resistance of this edge is $c$. Create a new tree $T'$ rooted at $z$ by removing the edge $z \to D$ from $T$ and adding the edge $D \to z$. The latter has resistance at most $n < c$. Therefore

$$\begin{aligned} R(T') &= R(T) + r(D \to z) - r(z \to D) \\ &\leq R(T) + n - c \\ &< R(T). \end{aligned}$$

Hence $T$ is not a minimum resistance tree. This contradiction shows that the state $D$ is not stochastically stable. It follows that all the stochastically stable states are contained in the set $C^0$.

From Lemma 3, we know that a state $z = [\bar{a}, \bar{u}]$ in $C^0$ is stochastically stable if and only if

$$\bar{a} \in \arg\min_{a^* \in \mathcal{A}} \left\{ c\left(\left|C^0\right| - 1\right) + \sum_{i \in N}\left(1 - U_i(a^*)\right) \right\},$$

equivalently

$$\bar{a} \in \arg\max_{a^* \in \mathcal{A}} \left\{ \sum_{i \in N} U_i(a^*) \right\}.$$

Therefore, a state is stochastically stable if and only if the action profile is efficient. This completes the proof of Theorem 1. □

## V. The Importance of Interdependence

In this section, we focus on whether the interdependence condition in Definition 1 can be relaxed while ensuring that the stochastically stable states remain efficient. Recall that a game is interdependent if it is not possible to partition the agents into two distinct groups $S$ and $N \setminus S$ that do not mutually interact with one another. One way that this condition can fail is that the game can be broken into two completely separate sub-games that can be analyzed independently. In this case, our algorithm ensures that in each sub-game the only stochastically stable states are the efficient action profiles. Hence, this remains true in the full game.

In general, however, some version of interdependence is needed. To see why, consider the following two-player game:

| | $A$ | $B$ |
|---|---|---|
| $A$ | $1/2, 1/4$ | $1/2, 0$ |
| $B$ | $1/4, 0$ | $1/4, 3/4$ |

Here, the row agent affects the column agent, but the reverse is not true. Consequently, the recurrence states of the unperturbed process are $\{AA, AB, BA, BB, A\emptyset, B\emptyset, \emptyset\emptyset\}$ where: $A\emptyset$ is the state where agent $1$ is content with action profile $A$ and agent $2$ is discontent; $\emptyset\emptyset$ is the state where both agents are discontent. We claim that the action profile $(A, A)$, which is not efficient, is stochastically stable. This can be deduced from Figure 1 (here we choose $c = n = 2$). The illustrated resistance tree has minimum stochastic potential because each edge in the given tree has minimum resistance among the edges exiting from that vertex. Consequently, this inefficient action profile $AA$ is stochastically stable.

At first glance, this example merely demonstrates that our proposed algorithm does not guarantee convergence to the efficient action profile for all finite strategic-form games. However, it turns out that this example also establishes that there does not exist a distributed learning algorithm that guarantees convergence to the efficient action profile. The following proposition makes this precise.

**Proposition 4.** *There exists no uncoupled learning algorithm that leads to an efficient action*

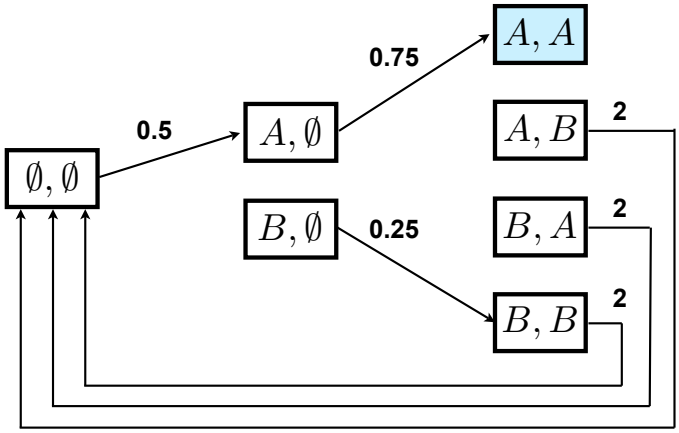



Fig. 1. Illustration of the minimum resistance tree rooted at the action profile $(A, A)$.

*profile for all finite strategic-form games.*[5]

*Proof:* We will prove this proposition by contradiction. Suppose that there exists a coupled learning algorithm of the form (1) that leads to an efficient action profile for all finite strategic-form games. When considering the two-player game highlighted above, this algorithm will lead behavior to the action profile $(B, B)$. However, from player 1's perspective, this game is equivalent to a simple one-player game with payoffs.

| | |
|---|---|
| $A$ | $1/2$ |
| $B$ | $1/4$ |

This learning algorithm leads to the behavior $A$ in this one-player game. Therefore, we have a contradiction since the same learning algorithm is not able to ensure that behavior leads to $(A)$ for the one-player setting and $(B, B)$ for the two-player setting. ■

## VI. Illustrations

In this section, we provide two simulations that illustrate the mechanics of our proposed algorithm. In the first subsection, we apply our results to a prisoner's dilemma game and provide

[5]Here, we use the term "lead to" to mean either convergence, almost sure convergence, or convergence in the sense of stochastic stability. The authors would like to acknowledge conversations with Yakov Babichenko which led to this result.

a detailed analysis of the stochastically stable states. In the second subsection, we simulate our algorithm on a three-player game that exhibits many of the same challenges associated with the prisoner's dilemma game.

### A. Prisoner's Dilemma

Consider the following prisoner's dilemma game where all players' utilities are scaled between $0$ and $1$. It is easy to verify that these payoffs satisfy the interdependence condition.

| | $A$ | $B$ |
|---|---|---|
| $A$ | $3/4, 3/4$ | $0, 5/4$ |
| $B$ | $4/5, 0$ | $1/3, 1/3$ |

Fig. 2. A two-player strategic-form game in which both player 1 (row player) and player 2 (column player) choose either $A$ or $B$. $(B, B)$ is the unique pure Nash equilibrium.

Consequently, our algorithm guarantees that the action profile $(A, A)$ is the only stochastically stable state. We will now verify this by computing the resistances for each of the transitions. The recurrence classes of the unperturbed process are $(AA, AB, BA, BB, \emptyset)$, where the agents are content for the four listed action profiles and $\emptyset$ corresponds to the scenario where both agents are discontent. (For notational simplicity, we omit the baseline utilities for each of the four action profiles.)

Consider the transition $AA \to BB$. Its resistance is

$$r(AA \to BB) = c + (1 - 1/3) + (1 - 1/3) = c + 4/3.$$

The term $c$ comes from the fact that we have only one experimenter. The term $2(1-1/3)$ results from the fact that both agents $1$ and $2$ need to accept the new benchmark payoff of $1/3$ to make this transition. For the sake of concreteness, let $c = n = 2$ for the remainder of this section. The resistances of all possible transitions are shown in Table I. Each entry in the table represents the resistance going from the row-state to the column-state. The stochastic potential of each of the five states can be evaluated by analyzing the trees rooted at each state. These are shown in Figure 3. Note that each of the minimum resistance trees has the very simple structure identified

| | $AA$ | $AB$ | $BA$ | $BB$ | $\emptyset$ |
|---|---|---|---|---|---|
| $AA$ | $\cdot$ | $2+(1-4/5)+(1-0)=16/5$ | $2+(1-4/5)+(1-0)=16/5$ | $2+2(1-1/3)=10/3$ | 2 |
| $AB$ | $2+2(1-3/4)=5/2$ | $\cdot$ | $2+(1-4/5)+(1-0)=16/5$ | $2+2(1-1/3)=10/3$ | 2 |
| $BA$ | $2+2(1-3/4)=5/2$ | $2+(1-4/5)+(1-0)=16/5$ | $\cdot$ | $2+2(1-1/3)=10/3$ | 2 |
| $BB$ | $2+2(1-3/4)=5/2$ | $2+(1-4/5)+(1-0)=16/5$ | $2+(1-4/5)+(1-0)=16/5$ | $\cdot$ | 2 |
| $\emptyset$ | $2(1-3/4)=1/2$ | $(1-4/5)+(1-0)=6/5$ | $(1-4/5)+(1-0)=6/5$ | $2(1-1/3)=4/3$ | $\cdot$ |

TABLE I

EVALUATION OF RESISTANCES FOR PRISONER'S DILEMMA GAME.

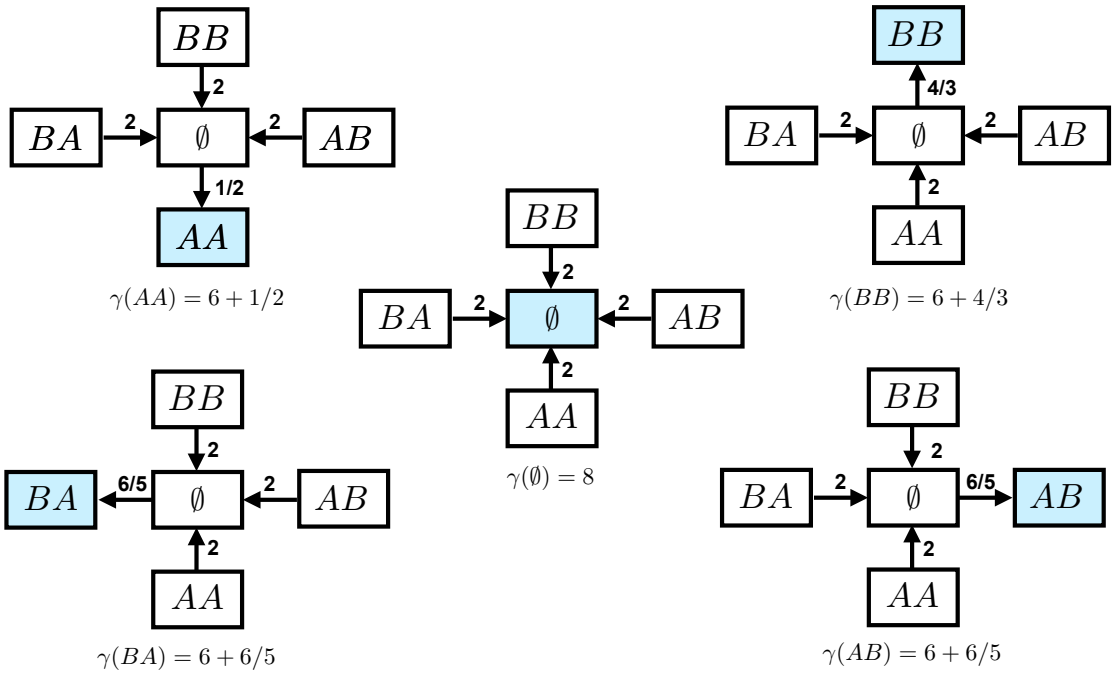



Fig. 3. Stochastic potential for each state in the prisoner's dilemma game.

in Lemma 3. It is evident that $AA$ has minimum stochastic potential, hence it is the unique stochastically stable state.

Figure 5(a) illustrates a simulation of our learning algorithm on the prisoner's dilemma game defined above. The experimentation parameter was set at $\epsilon = 0.001$. Furthermore, the initial state of each agent $i \in \{1, 2\}$ was set at $[\bar{a}_i = B, \bar{u}_i = 1/3, m_i = C]$. Observe that the empirical frequency of the joint actions reaches a stage where most of the weight is placed on the joint

action $(A, A)$, which maximizes the sum of the agents' payoffs. It is important to note that either of the algorithms presented in [19], [20] would converge in a stochastic stability sense to the inefficient action profile $(B, B)$. The reason for these different asymptotic behaviors stems from the structural differences between the algorithms. In particular, [20] and [19] make use of four mood variables for the agents (content, discontent, hopeful, and watchful) instead of the two mood variables as in our approach. The transitional moods (hopeful and watchful) have the effect of making pure Nash equilibria more attractive than alternative action profiles, including efficient non-equilibrium profiles.

### *B. A Three-Player Version of the Prisoner's Dilemma*

In this section, we consider the following three-player version of the prisoner's dilemma game where all players' utilities are scaled between $0$ and $1$ as in Figure 4. Figure 5(b) illustrates a simulation of our learning algorithm for this game. The experimentation parameter was set at $\epsilon = 0.001$, and the initial state of each agent $i \in \{1, 2, 3\}$ was set at $[\bar{a}_i = B, \bar{u}_i = 0.2, m_i = C]$. Observe that the empirical frequency of the joint actions reaches a stage where most of the weight is placed on the joint action $(A, A, A)$ which maximizes the sum of the agents' payoffs.

$A$

| | $A$ | $B$ |
|---|---|---|
| $A$ | $0.95, 0.95, 0.95$ | $0, 1, 0$ |
| $B$ | $1, 0, 0$ | $0.1, 0, 0$ |

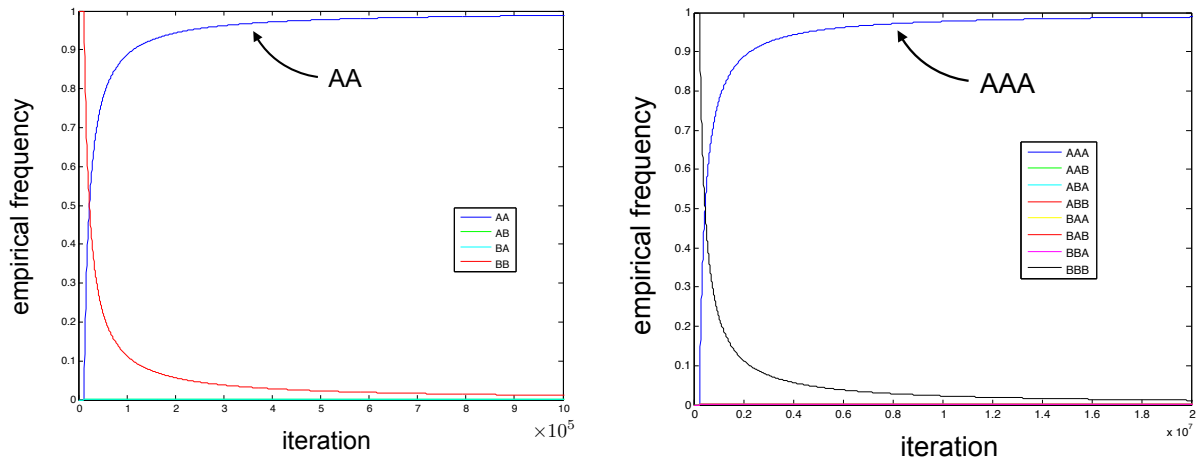
$B$

| | $A$ | $B$ |
|---|---|---|
| $A$ | $0, 0, 1$ | $0, 0, 0.1$ |
| $B$ | $0, 0.1, 0$ | $0.2, 0.2, 0.2$ |

Fig. 4. A three-player strategic-form game in which player 1 (row player), player 2 (column player), and player 3 (box player, i.e., the player that chooses if we are in the payoff matrix on the left or right) choose either $A$ or $B$. Observe that $(B, B, B)$ is the unique pure Nash equilibrium.

## VII. Conclusion

Most of the distributed learning algorithms in the prior literature focus on reaching a Nash equilibrium for particular game structures, such as potential games. However, there are many engineering applications that cannot be represented by potential games, in which case these




(a) Two-player prisoner's dilemma game.

(b) Three-player prisoner's dilemma game.

Fig. 5. Simulation results for both the two-player and three-player prisoner's dilemma games with experimentation rate $\epsilon = 0.001$. Observe that the empirical frequency of the joint actions reaches a stage where most of the weight is placed on the joint action $(A, A)$ or $(A, A, A)$ which is expected. Note that for both settings, the agents began content in the inefficient action profile $(B, B)$ or $(B, B, B)$.

algorithms are inadequate. This paper establishes the existence of simple learning algorithms that lead to efficiency for arbitrary game structures. The methodology raises a number of important theoretical and practical issues. How long does it take, in expectation, to reach the efficient outcome? Would providing agents with more information improve the transient behavior? If so, what information should be communicated to the agents? We suspect that the answers will depend importantly on the payoff structure of the game under consideration, and hence we leave this issue for further investigation.

## REFERENCES

[1] G. Chasparis and J. Shamma, "Distributed dynamic reinforcement of efficient outcomes in multiagent coordination and network formation," 2011, discussion paper, Department of Electrical Engineering, Georgia Tech.

[2] N. Li and J. R. Marden, "Decoupling coupled constraints through utility design," 2011, discussion paper, Department of ECEE, University of Colorado, Boulder.

[3] ——, "Designing games for distributed optimization," 2011, discussion paper, Department of ECEE, University of Colorado, Boulder.

[4] J. R. Marden, "State based potential games," *Automatica*, vol. 48, pp. 3075–3088, 2012.

[5] R. Gopalakrishnan, J. R. Marden, and A. Wierman, "An architectural view of game theoretic control," *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 3, pp. 31–36, 2011.

[6] J. R. Marden, G. Arslan, and J. S. Shamma, "Connections between cooperative control and potential games," *IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics*, vol. 39, pp. 1393–1407, December 2009.

[7] G. Arslan, J. R. Marden, and J. S. Shamma, "Autonomous vehicle-target assignment: a game theoretical formulation," *ASME Journal of Dynamic Systems, Measurement and Control*, vol. 129, pp. 584–596, September 2007.

[8] R. Johari, "The price of anarchy and the design of scalable resource allocation mechanisms," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, Eds. Cambridge University Press, 2007.

[9] R. S. Komali and A. B. MacKenzie, "Distributed topology control in ad-hoc networks: A game theoretic perspective," in *Proceedings of IEEE Consumer Communication and Network Conference*, 2007.

[10] H. B. Durr, M. S. Stankovic, and K. H. Johansson, "Distributed positioning of autonomous mobile sensors with application to coverage control," in *Proceedings of the American Control Conference*, 2011.

[11] M. S. Stankovic, K. H. Johansson, and D. M. Stipanovic, "Distributed seeking of Nash equilibria in mobile sensor networks," in *IEEE Conference on Decision and Control*, 2010.

[12] M. Zhu and S. Martinez, "Distributed coverage games for mobile visual sensor networks," *SIAM Journal on Control and Optimization*, 2010, under submission.

[13] P. Frihauf, M. Krstic, and T. Basar, "Nash equilibrium seeking for games with non-quadratic payoffs," in *IEEE Conference on Decision and Control*, 2010.

[14] ——, "Nash equilibrium seeking with infinitely-many players," in *American Control Conference*, 2011.

[15] M. Krstic, P. Frihauf, J. Krieger, and T. Basar, "Nash equilibrium seeking with finitely- and infinitely-many players," in *8th IFAC Symposium on Nonlinear Control Systems*, 2010.

[16] Q. Zhu, H. Tembine, and T. Basar, "Heterogeneous learning in zero-sum stochastic games with incomplete information," in *IEEE Conference on Decision and Control*, 2010.

[17] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff based dynamics for multi-player weakly acyclic games," *SIAM Journal on Control and Optimization*, vol. 48, pp. 373–396, February 2009.

[18] J. R. Marden, G. Arslan, and J. S. Shamma, "Joint strategy fictitious play with inertia for potential games," *IEEE Transactions on Automatic Control*, vol. 54, pp. 208–220, February 2009.

[19] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, pp. 626–643, 2009.

[20] B. R. Pradelski and H. P. Young, "Learning efficient Nash equilibria in distributed systems," 2010, discussion paper, Department of Economics, University of Oxford.

[21] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. **14**, pp. 124–143, 1996.

[22] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. New York, NY, USA: Cambridge University Press, 2007.

[23] J. R. Marden, S. Ruben, and L. Pao, "A model-free approach to wind farm control using game theoretic methods," 2011, discussion paper, Department of ECEE, University of Colorado, Boulder. [Online]. Available: http://ecee.colorado.edu/marden/files/WF_Optimization.pdf

[24] K. E. Johnson and N. Thomas, "Wind farm control: Addressing the aerodynamic interaction among wind turbines," in *Proceedings of the 2009 American Control Conference*, 2009.

[25] D. Wolpert and K. Tumor, "An overview of collective intelligence," in *Handbook of Agent Technology*, J. M. Bradshaw, Ed. AAAI Press/MIT Press, 1999.

[26] M. V. Solodov, "Incremental gradient algorithms with stepsizes bounded away from zero," *Computational Optimization and Applications*, vol. 11, no. 1, pp. 23–35, 1998.

[27] D. Blatt, A. Hero, and H. Gauchman, "A convergent incremental gradient method with a constant step size," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 29–51, 2008.

[28] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[29] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[30] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Transactions on Automatic Control*, no. 99, pp. 1–1, 2010.

[31] M. Zhu and S. Martinez, "On distributed convex optimization under inequality and equality constraints via primal-dual subgradient methods," *Arxiv preprint arXiv:1001.2612*, 2010.

[32] J. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 42–50, 1984.

[33] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[34] Y. Hatano and M. Mesbahi, "Agreement over random networks," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1867–1872, 2005.

[35] I. Androulakis and G. Reklaitis, "Approaches to asynchronous decentralized decision making," *Computers and Chemical Engineering*, vol. 23, no. 3, pp. 339–354, 1999.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundation and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[37] S. Hart and A. Mas-Colell, "Stochastic uncoupled dynamics and Nash equilibrium," *Games and Economic Behavior*, vol. 57, no. 2, pp. 286–303, 2006.

[38] ——, "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, vol. 93, no. 5, pp. 1830–1836, 2003.

[39] D. Foster and H. Young, "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, vol. **1**, pp. 341–367, 2006.

[40] I. Arieli and Y. Babichenko, "Average testing and the efficient boundary," 2011, discussion paper, Department of Economics, University of Oxford and Hebrew University.

[41] D. Fudenberg and E. Maskin, "The folk theorem in repeated games with discounting or with incomplete information," *Econometrica*, vol. **54**, pp. 533–554, 1986.

[42] M. Osborne and A. Rubinstein, *A Course in Game Theory*. Cambridge, MA: MIT Press, 1994.

[43] D. Foster and H. Young, "Stochastic evolutionary games dynamics," *Journal of Theoretical Population Biology*, vol. **38**, pp. 219–232, 1990.

[44] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, January 1993.