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 game with generic payoffs. 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 generally understand. The method has potential application 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.

I. Introduction

The field of game theory is gaining popularity as a paradigm for 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 “players” embedded in an interactive, game-theoretic environment. This step involves defining a set of choices and a local objective function for each player. Second, the system designer must specify the players’ 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 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 primarily focuses on distributed learning algorithms that are suitable for implementation in large scale engineering systems [2], [3], [10]–[13]. Accordingly, most

This research was supported by AFOSR grant #FA9550-09-1-0538 and by ONR grant #N00014-09-1-0751.

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.

of the results focus on particular classes of games, notably potential games, that are pertinent to distributed engineering systems. The motivation for this work stems from the fact that the interaction framework for a distributed engineering system can frequently be represented as a potential game. Consequently, these distributed learning algorithms can be utilized as distributed control algorithms that provide strong asymptotic guarantees on the emergent global behavior [5]–[7], [14], [15]. This approach provides a hierarchical decomposition in the design (*game design*) and control (*learning rule*) of a multiagent system where the intermediate layer is constrained by the potential game structure [5].

There are two limitations to this framework however. First, most results in this domain focus on convergence to Nash equilibrium, which may be very inefficient with regard to the system level objective. Characterizing this inefficiency is a highly active research area in algorithmic game theory [16]. 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. This stems from the fact that a given engineering system possesses 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 parts of the system. Furthermore, in many complex systems the relationship between the behavior of the components and the overall system performance is not well characterized.

One example of a system that exhibits these challenges is the control of a wind farm to maximize total power production. 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 energy production [17]. 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 approach to handle this large-scale coordination problem is to model the interactions of the turbines in a game theoretic environment. The space of admissible utility functions for the individual turbines is limited because of the following informational limitations:

(i) Each turbine does not have access to the actions[1] of other turbines due to the lack of a

[1] A turbine's action is called an *axial induction factor*. The axial induction factor captures the amount of energy the turbine extracts from the wind.

suitable communication system.

(ii) No turbine has access to the functional relationship between the total power generated and the action of the other turbines. This is because the aerodynamic interaction between the turbines is poorly understood.

These limitations restrict the ability to represent the interaction framework as a potential game. For example, one of the common utility design approaches is to assign each turbine an objective function which measures the turbine's marginal contribution to the power production of the wind farm, i.e., the difference between the total power produced when that turbine is active and the total power produced when that turbine is inactive [6], [14]. Calculating this difference is not possible due to a lack of knowledge about the aerodynamic interactions. Essentially, the interaction framework is constrained to being the case where each turbine responds to it's individual power production. It is not known whether this interaction framework can be represented by a potential game, or even whether the game is question possesses a pure strategy Nash equilibrium. The existing results in the literature do not provide suitable control algorithms for this type of 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 agent's utility functions. In a wind farm this sum is precisely equal to the total power generated. The main result of this work is a simple payoff-based learning algorithm that guarantees convergence to this Pareto efficient action profile when 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 [13] who show how to achieve constrained efficiency *within* the set of Nash equilibrium outcomes.

II. Background

We consider finite strategic-form games with n agents denoted by the set $N := \{1, ..., n\}$. 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 refer to a finite strategic-form game as "a game," and we sometimes use a single symbol, e.g., G, to represent the entire game, i.e., the player set, N, action sets, $\mathcal{A}_i$, and utility functions U_i. For an action profile $a = (a_1, a_2, ..., a_n) \in \mathcal{A}$, let a_{-i}

denote the profile of agent actions *other than* player i, i.e., $a_{-i} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$. 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. Define the *welfare* of an action profile $a \in \mathcal{A}$ as $W(a) = \sum_{i \in N} U_i(a)$. An action profile that optimizes the welfare will be denoted by $a^{\text{opt}} \in \arg\max_{a \in \mathcal{A}} W(a)$.

A. Repeated Games

In a repeated game, at each time $t \in \{0, 1, 2, \ldots\}$, each player $i \in N$ 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 $i \in N$ chooses the action $a_i(t)$ at time t according to a probability distribution $p_i(t) \in \Delta(\mathcal{A}_i)$, which we will refer to as the *strategy* of player i at time t where $\Delta(\mathcal{A}_i)$ is defined as the simplex over the set $\mathcal{A}_i$. We adopt the convention that $p_i^{a_i}(t)$ is the probability that player i selects action a_i at time t according to the strategy $p_i(t)$. A player's strategy at time t can rely only on observations from times $\{0, 1, 2, ..., t-1\}$. 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 his own utility function and is capable of observing the actions of all other players at every time step but does not know their utility functions, then the strategy adjustment mechanism of player i can be written in the general form

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

Such an algorithm is said to be *uncoupled* [18], [19].

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

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

Such a learning rule is said to be *completely uncoupled* or *payoff-based* [20]. Recent work in [21] has shown that for generic two-player games there are completely uncoupled learning rules that lead to Pareto optimal behavior. In this paper we exhibit a different class of learning procedures that lead to Pareto optimal outcomes in any finite n-person game with generic payoffs.[2]

III. A Payoff Based Algorithm for Maximizing Welfare

In this section we introduce a payoff-based algorithm that converges to the Pareto efficient action profile in any finite n-person game with generic payoffs. The proposed algorithm is a variant of the approach in [13], where each player possesses an internal state variable which impacts the player's behavior rule. The core difference between our proposed algorithm and the one in [13] is the asymptotic guarantees. In particular, [13] guarantees convergence to the Pareto efficient Nash equilibrium while our proposed algorithm converges to the 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 design in [13].

At each point in time a player'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 at each iteration $k = 1, 2, ...$, is conditioned on agent i's underlying benchmark payoff $\bar{u}_i(k)$, benchmark action $\bar{a}_i(k)$, and mood $m_i(k) \in \{C, D\}$. We divide the dynamics into the following two parts: the player dynamics and the state dynamics. Without loss of generality we focus on the case where player utility functions are strictly bounded between 0 and 1, i.e., for any player $i \in N$ 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 $n > W(a) \geq 0$.

Player Dynamics: Fix an experimentation rate $\epsilon > 0$. Let $[\bar{a}_i, \bar{u}_i, m_i]$ be the current state of agent i.

[2] Such a result might seem reminiscent of the Folk Theorem, which specifies conditions under which a Pareto efficient action profile can be implemented as an equilibrium of a repeated game. See among others [22], [23]. In the present context, however, we are interested in whether players can learn to play a Pareto efficient action profile without having any information about the game as a whole or even what the other players are doing.

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

$$p_i^{a_i} = \begin{cases} \frac{\epsilon^c}{|\mathcal{A}_i|-1} & \text{for } a_i \neq \bar{a}_i \\ 1-\epsilon^c & \text{for } a_i = \bar{a}_i \end{cases} \tag{2}$$

where $c \geq n$ is a constant.

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

$$p_i^{a_i} = \frac{1}{|\mathcal{A}_i|} \quad \text{for every} \quad a_i \in \mathcal{A}_i \tag{3}$$

Note that the benchmark action and utility play no role with regards to the player dynamics when the player is discontent.

State Dynamics: Once the player 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 players other than player 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

$$[\bar{a}_i, \bar{u}_i, C] \xrightarrow{[\bar{a}_i, \bar{u}_i]} [\bar{a}_i, \bar{u}_i, C] \tag{4}$$

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

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

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

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

Ensuring that the dynamics converge to the Pareto efficient action profile requires the following level of interdependence in the game structure.

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 players $J \subset N$, there exists a player $i \notin J$ 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 is relatively weak and states that it is not possible to divide the players into two distinct subsets that do not mutually interact with one another.

The above 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 denote the transition probability matrix by P^ϵ 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 focus on characterizing the support of the limiting stationary distribution which is referred to as the *stochastically stable states*. 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^ϵ for a fixed $\epsilon > 0$. We now provide the following characterization of the stochastically stable states.

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

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

(ii) The benchmark actions and payoffs are aligned, i.e., $u_i = U_i(a)$.

(iii) The mood of each player is content, i.e., $m_i = C$.

IV. Proof of Theorem 1

In this section we provide the proof of Theorem 1. We rely on the theory of resistance trees for regular perturbed Markov decision processes to prove that an action profile is stochastically stable if and only if it is Pareto efficient. We first provide a brief background on the theory of resistance tree.

A. 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 $\forall z, z' \in Z$,

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

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 $E_1, E_2, ..., E_N$. 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 $\zeta = (z_1 \to z_2 \to ... \to z_n)$ 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, $r(\zeta) = r(z_1 \to z_2) + r(z_2 \to z_3) + ... + r(z_{n-1} \to z_n)$. 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, or 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 result provides a simple criterion for determining the stochastically stable states ([24], Theorem 4).

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.

B. Proof of Theorem 1

We will prove Theorem 1 by the following sequence of lemmas. First, we introduce the following notation by dividing up the state space Z. Let C^0 be the subset of states in which each player is content and the benchmark action and utility are aligned. That is, if $[a, u, m] \in C^0$ then $u_i = U_i(a)$ and $m_i = C$ for each player $i \in N$. Let D^0 represent the set of states in which everyone is discontent. That is, if $[a, u, m] \in D^0$ then $u_i = U_i(a)$ and $m_i = D$ for each player $i \in N$.

The process described above is clearly a regular perturbed Markov decision process. The unperturbed process, denoted as P^0, is the Markov decision process where $\epsilon = 0$. The first lemma provides a characterization of the recurrent classes of the unperturbed process.

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

Proof: First, the states D^0 clearly represent a single recurrent class of the unperturbed process since the probability of transitioning between any two states $z_1, z_2 \in D^0$ is $O(1)$. Next, suppose that a proper subset players $S \subset N$ is discontent and the benchmark action and benchmark utility of all other players are a_{-S} and u_{-S} respectively. By our interdependence condition there exists a player $j \notin S$ such that $u_j \neq U_j(a'_S, a_{-S})$ for some action $a'_S \in \prod_{i \in S} \mathcal{A}_i$. Hence, the player set S will eventually play action a'_S with probability 1 thereby causing player j to become discontent. Hence, this cannot be a recurrent class of the unperturbed process. This process can be repeated to show that all players will become discontent; hence any state that consists of a partial collection of discontent players $S \subset N$ is not a recurrent class of the unperturbed process. Lastly, consider a state $[a, u, C]$ where all players are content but there exists at least one player i whose benchmark action and benchmark utility are not aligned, i.e., $u_i \neq U_i(a)$. For the unperturbed process, at the ensuing time step the action profile a will be played and player i will become discontent since $u_i \neq U_i(a)$. Since one player is discontent, all players will become discontent as highlighted above. This completes the proof. ■

We know from [24] 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 recurrent class. 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 $[a, u]$ and drop the extra notation highlighting that the players 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 $r(w \rightarrow z)$ denote the total resistance of the least-resistance path that starts in w and end in z. We call $w \rightarrow z$ an edge and $r(w \rightarrow z)$ the resistance of the edge.*

Let $z = [a, u]$ and $z' = [a', u']$ be any two distinct states in C^0. We point out the following observations regarding the resistance of transitions between the states z, z', and D.

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

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

This is true since one experimentation can cause all players to become discontent.

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

$$r(D \rightarrow z) = \sum_{i \in N} (1 - u_i) = n - W(a).$$

This is true since each player i needs to accept the benchmark payoff u_i which has a resistance $(1 - u_i)$.

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

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

This is true since by definition of edge resistance we have that $r(z \rightarrow z') \leq r(z \rightarrow D) + r(D \rightarrow z')$. Therefore, each transition of minimum resistance includes at most one player that experiments.

Before stating the next lemma we introduce the notion of a path or sequence of edges. A path $\mathcal{P}$ over the states $D \cup C^0$ is a sequence transitions of the form

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

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

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

Lemma 3. *The stochastic potential associated with any state $z = [a, u]$ in C^0 is*

$$\gamma(z) = c\left(\left|C^0\right| - 1\right) + \sum_{i \in N} (1 - u_i)\,. \tag{5}$$

Proof: We first prove that (5) 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 $z' \rightarrow D$. The total resistance associated with these edges is $c\left(|C^0| - 1\right)$.

P-2: The edge existing the state D is of the form $D \rightarrow z$. The resistance associated with this edge is $\sum_{i \in N}(1 - u_i)$.

The constructed tree T is clearly rooted at z and has a total resistance $c\left(|C^0| - 1\right) + \sum_{i \in N}(1 - u_i)$. Therefore we know that $\gamma(z) \leq c\left(|C^0| - 1\right) + \sum_{i \in N}(1 - u_i)$.

We now prove that (5) is also a lower bound for the stochastic potential by contradiction. Suppose there exists a tree T rooted at z with resistance $R(T) < c\left(|C^0| - 1\right) + \sum_{i \in N}(1 - u_i)$. Since the tree T is rooted at z we know that there exists a path $\mathcal{P}$ from D to z of the form

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

where $z^k \in C^0$ for each $k \in \{1, ..., m\}$. The resistance associated with this path of $m + 1$ transition satisfies

$$R(\mathcal{P}) \geq mc + \sum_{i \in N}(1 - u_i)$$

where mc comes from invoking observation (iii) at the last m transitions in the path $\mathcal{P}$ and $\sum_{i \in N}(1 - u_i)$ comes from the fact that each player needs to accept 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:

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

The new tree T' has a total resistance that satisfies $R(T') \leq R(T)$. Note that if the path $\mathcal{P}$ was of the form $D \rightarrow z$ then this augmentation did not alter the tree structure.

Now, suppose there exists an edge $z' \rightarrow z''$ in the tree T' for some states $z', z'' \in C^0$. By observation (iii) the resistance of this edge satisfies $r(z' \rightarrow z'') \geq c$. Construct a new tree T'' by removing the edge $z' \rightarrow z''$ and adding the edge $z' \rightarrow D$ which has a resistance c. Note that this new tree T'' is rooted at z. The resistance associated with the tree T'' 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 $R(T^*) = c\left(|C^0| - 1\right) + \sum_{i \in N} \left(1 - u_i\right)$. Since by construction $R(T^*) \leq R(T)$ we have a contradiction which completes the proof. ■

We will now complete the proof by analyzing the minimum resistance trees using the above lemmas. We first show that the state D is not stochastically by contradiction. Suppose there exists a minimum resistance tree, denoted as 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 the tree T and adding the edge $D \to z$ which has at most a resistance of $n < c$. Therefore, we have that

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

Hence, T is not a minimum resistance tree. Consequently, the state D is not stochastically stable.

Therefore, we know that the stochastically stable stable must be contained in the set C^0. From Lemma 3 we know that a state $z = [a, u]$ in C^0 is stochastically stable if and only if

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

or equivalently

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

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

V. Relaxing Interdependence

In this section we focus on whether our interdependence condition can be relaxed while ensuring that the only stochastically stable states remain the Pareto efficient action profiles. Interdependence is a relatively weak condition which states that it is not possible to partition the players into two distinct groups S and $N \setminus S$ that do not mutually interact with one another. One trivially relaxation of this condition is the situation where neither group interacts with the other group, i.e., for any action profile a, any player $j \in S$ (or $j \in N \setminus S$), and any action $a'_S \in \prod_{j \in S} \mathcal{A}_j$ we have $U_j(a) = U_j(a'_S, a_{-S})$. Roughly speaking, this condition states that the game can be broken into separate games that can be analyzed independently. Therefore, the proposed algorithm ensures that in each game the only stochastically stable states are the Pareto efficient action profiles. Hence, Theorem 1 directly carries over to this setting.

Unfortunately, relaxing this condition further does not yield similar results. For example, consider the following game:

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

Here, the row player impacts the column player but the reverse is not true. Consequently, the recurrent states of the unperturbed process are now $\{AA, AB, BA, BB, A\emptyset, B\emptyset, \emptyset\emptyset\}$ where $A\emptyset$ indicates the state where player 1 is content with action profile A and player 2 is discontent. Alternatively, $\emptyset\emptyset$ represents the state where both players are discontent. Figure V proves that the action profile (A, A), which is not Pareto efficient, is stochastically stable by evaluating the minimum resistance tree rooted at (A, A). In the figure we set $c = n = 2$. The illustrated resistance tree is of minimum stochastic potential because each edge in the given tree is of minimum resistance. That is for every edge $z \rightarrow z'$ in the highlighted tree $r(z \rightarrow z') = \min_{z''} r(z \rightarrow z'')$. Consequently, in this new domain the unique Nash equilibrium is stochastically stable as opposed to the Pareto efficient action profile.

This example demonstrates that in games with such a structure the proposed dynamics do not converge to the Pareto efficient action profile. Rather, the proposed dynamics converge to a Nash equilibrium of a new game where the interconnected player sets are viewed as a single player with a payoff equal to the sum of the players' payoffs. We avoid including a rigorous analysis of the resulting behavior for such scenarios.

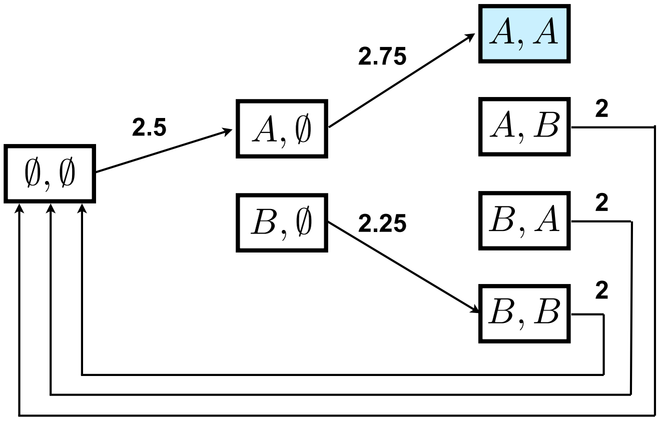

Fig. 1. Illustration of the minimum resistance tree rooted at the action profile (A, A).

VI. Illustrative examples

In this section we highlight the contribution of this paper on two problem. First, we focus on the game of prisoner's dilemma to illustrate our algorithms through a simple example. Next, we demonstrate how the proposed algorithm can be used to optimize energy production in a wind farm.

A. Prisoner's dilemma

Consider the following prisoner's dilemma game where all player utilities are scaled between 0 and 1:

	C	D
C	$1/2, 1/2$	$0, 2/3$
D	$2/3, 0$	$1/3, 1/3$

The prisoner's dilemma game is generic and hence satisfies our interdependence condition. Consequently, the presented algorithm guarantees that the action profile (C, C) is the only stochastically stable state. We will now verify this by computing the resistances for of each of the transitions. The recurrent classes of the unperturbed process are $(CC, CD, DC, DD, \emptyset)$ where the players are content for the given action profiles and $\emptyset$ corresponds to the scenario where both players are discontent. Here, we omit explicitly highlighting the baseline utility for

	CC	CD	DC	DD	$\emptyset$
CC	$\cdot$	$2+(1-2/3)+(1-0)=10/3$	$2+(1-2/3)+(1-0)=10/3$	$2+2(1-1/3)=10/3$	2
CD	$2+2(1-1/2)=3$	$\cdot$	$2+(1-2/3)+(1-0)=10/3$	$2+2(1-1/3)=10/3$	2
DC	$2+2(1-1/2)=3$	$2+(1-2/3)+(1-0)=10/3$	$\cdot$	$2+2(1-1/3)=10/3$	2
DD	$2+2(1-1/2)=3$	$2+(1-2/3)+(1-0)=10/3$	$2+(1-2/3)+(1-0)=10/3$	$\cdot$	2
$\emptyset$	$2(1-1/2)=1$	$(1-2/3)+(1-0)=4/3$	$(1-2/3)+(1-0)=4/3$	$2(1-1/3)=4/3$	$\cdot$

TABLE I

EVALUATION OF RESISTANCES FOR PRISONER'S DILEMMA GAME.

each of the 4 joint action profiles to avoid redundancy. Lets initially focus on one particular transition $CC \to DD$. The resistance of this transition is

$$r(CC \to DD) = c + (1-1/3) + (1-1/3) = c + 4/3$$

where c comes from the fact that we have only one experimenter and the $2(1-1/3)$ results from the fact that both players 1 and 2 need to accept the new benchmark payoff of $1/3$ in this transition. Let $c = n = 2$ for the remaining portion of this section. The resistances of all possible transitions are highlighted in Table VI-A. Each entry in this table represents the resistance going from the state highlighted by the row on the left column to the state highlighted by the column on the the top row. First, note that the resistances associated with any two state $z, z' \in C^0$ to a new state $z'' \in C^0$ is the same. This is because a single experimentation leads to all players becoming discontent. Hence, all players need to accept the new benchmarks in z'' which gives us the equivalence.

The stochastic potential of each of the 5 states can be evaluated by analyzing the family of trees rooted at each state. The minimum resistance tree rooted at each of the 5 states is illustrated in Figure VI-A. Note that each of the minimum resistance trees has a very simple structure as identified in Lemma 3. Hence, this analysis verifies that (C, C) is of minimum stochastic potential and hence is the unique stochastically stable state.

B. *Wind farms*

In this section we focus on the control of a wind farm where the goal is to generate as much power as possible. The ingredients of the problem are the following:

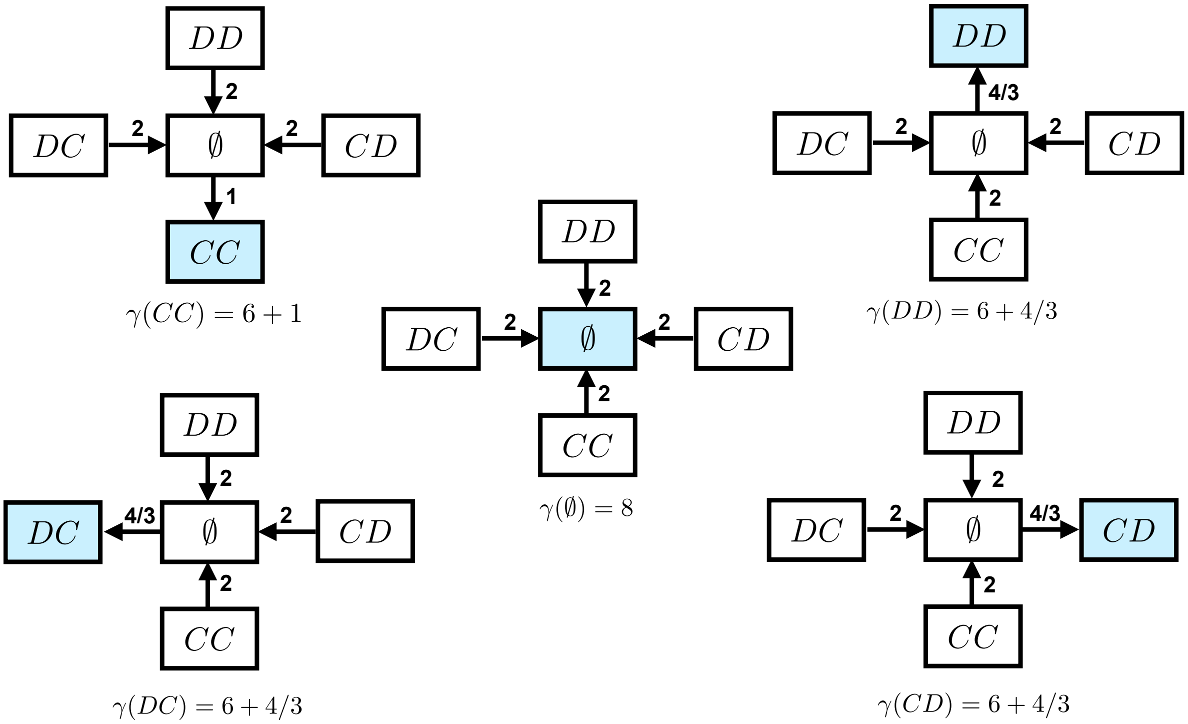

Fig. 2. Stochastic potential for each state in the prisoner's dilemma game.

- **Agents**: Individual wind turbines denoted by the set N.
- **Decisions**: The action set for turbine i includes the orientation and the axial induction factor. The set of such choices is denoted by $\mathcal{A}_i$. The axial induction factor relates to how much energy the turbine extracts from the wind given the current wind conditions.
- **Power production**: The power produced by turbine i is a function of the current wind conditions and the actions of all turbines. The power generated by turbine i given the decision of all turbines $a = (a_1, a_2, ..., a_n) = (a_i, a_{-i})$ is given by $P_i(a_i, a_{-i})$. We assume throughout that the exogenous wind conditions are fixed so we omit this in the power expression for each turbine.
- **System level objective:** The goal is to optimize the total energy production in the wind farm, i.e.,

$$P(a) = \sum_{i \in N} P_i(a)$$

Most of the existing research on the control of wind turbines focuses on the single turbine setting [25]. Controlling an array of turbines in a wind farm is fundamentally more challenging than controlling a single turbine stemming from the aerodynamic interaction between the turbines. These aerodynamic interactions render most of the single turbine control algorithms *highly inefficient* for optimizing wind farm productivity [17], [26] by introducing a degree of interconnectivity between the objective (or power) functions of the individual turbines. More specifically, the power generated by one turbine is dependent on the exogenous wind conditions coupled with the axial induction factors of other turbines. Lastly, these aerodynamic interactions are poorly characterized hence the precise structural form of the power generated by the wind farm $P(a_1, ..., a_n)$ as a function of the axial induction factors in not characterized.

The results in this paper provide a model-free approach to energy optimization in a wind farm. Our proposed payoff-based algorithm guarantees convergence to the action profile which maximizes the total power generated in the wind farm provided that the underlying game satisfies the interdependence condition. This condition seems plausible in the context of wind farms since the turbines mutually interact with one another through the aerodynamic interactions, i.e., the action by one turbine impacts the local wind conditions seen by neighboring turbines. The key point of this exposition is that we do not need to characterize the aerodynamic interactions between the turbines in order to design algorithms that optimize energy production.

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 at Boulder.

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

[4] J. R. Marden, “State based potential games,” 2011, discussion paper, Department of ECEE, University of Colorado at Boulder.

[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] 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.

[11] 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.

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

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

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

[15] J. R. Marden and A. Wierman, "Distributed welfare games," 2008, discussion paper, Department of ECEE, University of Colorado at Boulder.

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

[17] 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.

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

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

[20] 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.

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

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

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

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

[25] L. Pao and K. Johnson, "Control of wind turbines: Approaches, challenges, and recent developments," *Control Systems, IEEE*, vol. 31, no. 2, pp. 44 –62, 2011.

[26] R. J. Barthelmie and L. E. Jensen, "Evaluation of wind farm efficiency and wind turbine wakes at the nysted offshore wind farm," *Wind Energy*, vol. 13, no. 6, pp. 573–586, 2010.