2013 American Control Conference (ACC)
Washington, DC, USA, June 17-19, 2013

Convergence Guarantees for a Decentralized Algorithm Achieving Pareto Optimality

Anup Menon and John S. Baras

Abstract— **We consider N agents, each picking actions from a finite set and receiving a payoff according to its individual utility function that may depend on the actions picked by others. An agent has no knowledge about the functional form of its utility and can only measure its instantaneous value. It is assumed that all agents pick actions and receive payoffs synchronously. For this setting, a fully decentralized iterative algorithm for achieving Pareto optimality i.e. picking actions that maximize the sum of all utilities was proposed by Marden et. al. in [1] that lacks convergence guarantees. By scheduling a certain noise parameter to go to zero along iterations of this algorithm, conditions that guarantee convergence in probability are derived in this paper.**

I. INTRODUCTION

The paradigm of Game Theoretic Control is a promising direction of research in control and optimization in the context of muti-agent systems. It comprises of: i) designing individual utility or payoff functions with special structure such that solution concepts like Nash equilibira (NE) etc. of the resulting game correspond to desired system-wide outcome (for instance, NE in potential games correspond to extrema of the potential function); and ii) devising learning rules for the agents to discover such equilibria [2]. Both the utilities and the learning rule must conform with the informational constraints of the problem at hand. Examples of such utility design for specific applications range from distributed optimization [3] to coverage problems in sensor networks [4] and power control in wireless networks [5].

Several learning rules (or, interchangeably, algorithms) have been proposed in the evolutionary games literature that help agents learn NE in games with special structure like potential, weakly-acyclic, congestion games, etc. [6], [7]. Thus, designing utilities with such special structure facilitates direct use of these algorithms. Another desirable feature of some of these learning rules is payoff-based implementation i.e. no knowledge of the payoff structure is needed and an agent adjusts its play based on observed payoffs alone.

However, there are situations where this paradigm of designing utilities with special structure is too restrictive. To illustrate this point, consider the problem of maximizing the total power production of a wind farm [8]. Aerodynamic interactions between different wind turbines are not well understood and there are no good models to predict the effects of one turbine's actions on the power production of other turbines downstream. The information available to each turbine is its own power output and a decentralized algorithm that maximizes the total power production of the farm is sought. Since there are no good models for the interactions, there is little hope to design utilities with special structure that are functions of such individual power measurements. This points towards the need for algorithms that are applicable when there is little structural information about the utilities (for instance, a turbine can be assigned its individual power as its utility which, in turn, can depend on the actions taken by others in complex ways).

To summarize, we require a decentralized, payoff-based algorithm that

- requires little assumptions on the structure of the utilities; and
- helps agents learn a solution concept that corresponds to desirable system-wide behavior.

A fully decentralized learning rule which addresses exactly these concerns has been recently proposed in [1] with the objective of making the agents learn to play *efficient actions* that maximize the sum of the individual utilities i.e. the welfare function. Roughly speaking, this algorithm prescribes certain probability distributions for the agents to pick actions from; the distributions depend on the measured payoffs and a certain noise parameter ε. It is proved in [1] that for a sufficiently small ε, the realized actions of the agents in the limit are drawn from a distribution close to one with support over efficient actions. These results are based on the theory of perturbed Markov chains that was developed by Young [9] to explain equilibrium section in evolutionary games. While this learning rule and related results in [1] are encouraging, they have the following shortcomings:

1) Viewed as an algorithm, an adequate notion in which the individual actions converge to the efficient ones is absent.
2) There are results regarding perturbed Markov chains (see, for instance, [10]) that suggest that the expected waiting time before efficient actions are picked associated with a small ε can be too long.

The contribution of this paper is analogous to that of [11] for proving convergence of Simulated Annealing. We modify the learning rule of [1] by allowing the parameter ε to decrease to zero along the iterations of the algorithm ("annealing") and derive conditions on the rate of decrease

Research partially supported by the US Air Force Office of Scientific Research MURI grant FA9550-09-1-0538 and by the National Science Foundation (NSF) grant CNS-1035655.

The authors are with the Institute for Systems Research and the Department of Electrical and Computer Engineering at the University of Maryland, College Park, MD 20742, USA amenon@umd.edu, baras@umd.edu

978-1-4799-0176-0/$31.00 ©2013 AACC

of ε that guarantee convergence of the resulting algorithm w.p. 1. A sufficient condition for ergodicity of perturbed Markov chains with certain time decreasing perturbations is also derived in the process. While this directly addresses the first of the two concerns raised above, in view of recent results [12], it is also a step towards the second.

The remainder of the paper is organized as follows. We state the problem and introduce the learning rule of [1] in section II. In section III we introduce perturbed Markov chains and derive conditions for their ergodicity with certain time-varying perturbations. Section IV applies results of section III to derive the convergence guarantees in Theorem 5. Some illustrative numerical simulations are presented and directions for future work are discussed in the last section.

NOTATION

The paper deals exclusively with discrete-time, finite state space Markov chains. A Markov chain with Q as its 1-step transition probability matrix means that the i^{th} row and j^{th} column entry $Q_{i,j} = \mathbb{P}(\mathbf{X}_{t+1} = j | \mathbf{X}_t = i)$, where $\mathbf{X}_t$ denotes the state of the chain at time t. More generally, if $Q(t)$ denotes the 1-step transition probability matrix of a nonhomogeneous Markov chain at time t, then for all $m > n$, $\mathbb{P}(\mathbf{X}_m = j | \mathbf{X}_n = i) = Q_{i,j}^{(n,m)}$, where the matrix $Q^{(n,m)} = Q(n) \cdot Q(n+1) \cdots Q(m-1)$. The time indices of all Markov chains take consecutive values from the set of natural numbers $\mathbb{N}$. A Markov chain should be understood to be homogeneous unless stated otherwise. Given a vector x, its i^{th} component is denoted by x_i; and that of x_t by $(x_t)_i$.

II. PROBLEM FORMULATION AND ALGORITHM

To motivate the formulation, recall the wind farm example introduced earlier. The amount of energy a turbine extracts from the wind can be controlled by adjusting its *axial induction factor* and we shall call this the action variable of the turbine [8]. Thus, a change in the action of a turbine clearly affects the power production of turbines downstream from it. However, the relationship between the power produced by a turbine as a function of the actions picked by neighboring turbines is not accurately modeled. Also, a turbine's controller does not know what actions other turbines have picked. A turbine controller can only measure the instantaneous power produced by the turbine (the payoff) which is the consequence of such interaction. Thus, we seek an on-line, decentralized algorithm for individual turbine controllers to implement on the basis of their own past actions and measured payoffs to maximize the total power produced by the wind farm.

A. Problem Statement

Consider N agents indexed by i. Each agent can pick actions from the set A_i, $2 \leq |A_i| < \infty$, and the collective action set $\prod_{i=1}^{N} A_i$ is denoted by $\mathscr{A}$. There is a payoff function $u_i : \mathscr{A} \to [0,1)$ corresponding to each agent i and the sum of the payoffs is denoted by $W(a) = \sum_{i=1}^{N} u_i(a)$. Let U_i denote the range of $u_i(a)$ for $a \in \mathscr{A}$. At every time step t, agent i measures or receives the payoff $(u_t^{mes})_i = u_i(a_t)$, where $a_t \in \mathscr{A}$ is the joint action picked by the agents at time t and is not known to agent i. The objective is to design a mechanism for choosing the action of agent i at time t, $(a_t)_i$, on the basis of $\{(a_{t-1})_i, (u_{t-1}^{mes})_i, ..., (a_0)_i, (u_0^{mes})_i\}$ such that the joint action a_t is eventually picked from

$$\mathscr{A}^* = \{\arg\max_{a \in \mathscr{A}} W(a)\}.$$

B. The Learning Rule of [1]

Endow agent i with a state $x_i = [a_i, \overline{u}_i, m_i]$. The attribute $a_i \in A_i$ corresponds to the action picked, $\overline{u}_i$ to the payoff received and m_i is the $\{C,D\}$-valued 'mood' of agent i. When the mood variable equals C we call the agent "content" else "discontent". The collective state of all agents is denoted by $x = (a, \overline{u}, m)$ where $a \in \mathscr{A}$, $\overline{u} \in \prod_{i=1}^{N} U_i$ and $m \in \{C,D\}^N$.

At $t = 0$, agent i picks an arbitrary $(a_0)_i \in A_i$, records $(\overline{u}_0)_i = (u_0^{mes})_i$ and initializes $(m_0)_i = D$. For a fixed $\varepsilon > 0$ and $c > N$, agent i performs the following sequentially at every ensuing time instant $t \in \mathbb{N}$.

Start

Step 1: Pick $(\mathbf{a}_t)_i$ as follows.

1) If $(m_{t-1})_i = C$, pick $(\mathbf{a}_t)_i$ from A_i according to the p.m.f.

$$p(b) = \begin{cases} 1 - \varepsilon^c & \text{if } b = (a_{t-1})_i \\ \frac{\varepsilon^c}{|A_i| - 1} & \text{otherwise.} \end{cases} \tag{1}$$

2) Else, if $(m_{t-1})_i = D$, pick $(\mathbf{a}_t)_i$ according to the uniform distribution on A_i:

$$p(b) = \frac{1}{|A_i|} \text{ for all } b \in A_i. \tag{2}$$

Step 2: Receive payoff $(u_t^{mes})_i$ $(= u_i(a_t))$.

Step 3: Update $(\overline{\mathbf{u}}_t)_i$ and $(\mathbf{m}_t)_i$ as follows.

1) If $((a_t)_i, (u_t^{mes})_i) = ((a_{t-1})_i, (\overline{u}_{t-1})_i)$ and $(m_{t-1})_i = C$, then do nothing i.e. set $(x_t)_i = (x_{t-1})_i$.
2) Else, update $(\overline{\mathbf{u}}_t)_i \leftarrow (u_t^{mes})_i$. Set

$$(\mathbf{m}_t)_i = \begin{cases} D & \text{w.p.} \quad 1 - \varepsilon^{1-(u_t^{mes})_i} \\ C & \text{w.p.} \quad \varepsilon^{1-(u_t^{mes})_i}. \end{cases} \tag{3}$$

Stop

Heuristically, the learning rule says that a discontent agent experiments far more often than a content one and prescribes certain transition rules from one to the other.

Since $(\overline{u}_t)_i$ is set equal to $(u_t^{mes})_i$ in both steps 3.1 and 3.2, the following holds.

Lemma 2.1: The set of states realized by the algorithm $S \subset \prod_{i=1}^{N} (A_i \times U_i \times \{C,D\})$, satisfy $x \in S$, $x = [a, \overline{u}, m] \Rightarrow \overline{u}_i = u_i(a)$ for all i.

Next, we will make an assumption on the structure of the payoffs.

Assumption 1: For every $a \in \mathscr{A}$ and every proper subset of agents $J \subset \{1,...,N\}$, there exists an agent $i \notin J$ and a choice of actions $a'_J \in \prod_{j \in J} A_j$ such that $u_i(a'_J, a_{-J}) \neq u_i(a)$.[1]

It is easy to see that the transitions described by the algorithm define a Markov chain on S. Let us denote the corresponding 1-step transition matrix by $P(\varepsilon)$. Further, it can be verified that, under Assumption 1, $P(\varepsilon)$ is irreducible and aperiodic and thus has a unique stationary distribution $\mu(\varepsilon)$ (see [1]). It turns out that, as $\varepsilon \to 0$, $\mu(\varepsilon) \to \mu(0)$, for a certain density $\mu(0)$ over S. The following result characterizes the support of $\mu(0)$.

Theorem 1 ([1],Theorem 1): Let $P(\varepsilon)$ denote the 1-step transition probability matrix of the Markov chain defined by the algorithm and let $\mu(\varepsilon)$ denote its unique stationary distribution. Then $\mu(\varepsilon) \to \mu(0)$ as $\varepsilon \to 0$ and if Assumption 1 holds, for a state $x \in S$, $x = [a, \overline{u}, m]$, $\mu_x(0) > 0$ if and only if $a \in \mathscr{A}^*$ and $m_i = C$ for all $i \in \{1,...,N\}$.

The support of $\mu(0)$ is also called the *stochastically stable set*. One can now make the argument that by picking a sufficiently small ε, the realized states of the algorithm in the limit are drawn from a distribution close to one which has support over states where the joint action is from the set $\mathscr{A}^*$. This is a somewhat unsatisfactory argument from the point of view of convergence to $\mathscr{A}^*$ since there are no quantitative relations to guide the choice of such an ε. It is in this sense that convergence guarantees are lacking for this algorithm.

We will modify the algorithm by picking successively smaller values of ε along the iterates and derive conditions that guarantee convergence of the resulting algorithm w.p. 1 (see Theorem 5). The analysis of the algorithm is based on the theory of perturbed Markov chains that will be introduced in the next section. Detailed explanation about the specific structure of the algorithm is beyond the scope of this paper and we refer the interested reader to [1] or [13].

III. THEORY OF PERTURBED MARKOV CHAINS

The theory of perturbed Markov chains was developed by Young [9] to explain selection of some equilibria over others in finite player evolutionary games. Its mathematical description involves a Markov chain $P(0)$ with possibly several stationary distributions and one wishes to "choose one" among these. In order to do so, individual elements of $P(0)$ are perturbed by functions of a 'noise parameter' ε to obtain a perturbed chain $P(\varepsilon)$ ($\varepsilon \to 0 \Rightarrow P(\varepsilon) \to P(0)$) with a unique stationary distribution $\mu(\varepsilon)$. As $\varepsilon \to 0$, $\mu(\varepsilon) \to \mu(0)$, where $\mu(0)$ is a stationary distribution of $P(0)$ and the support of $\mu(0)$ can be characterized in terms of the rate at which components of $P(\varepsilon)$ converge. In this sense one can choose amongst the stationary distributions of $P(0)$.

[1]We borrow notation from the game theory literature: a_J denotes the actions taken by the agents in subset J from the collective action a and the actions of the rest is denoted by a_{-J}.

We will not discuss these results here and refer the interested reader to Theorem 4 in [9]. We consider reducing ε to zero along the evolution of $P(\varepsilon)$ (rendering it nonhomogeneous) and derive conditions that ensure ergodicity of the resulting chain with $\mu(0)$ as its limiting distribution. This is the content of the main result of this section, Theorem 4, and we begin by building the background and notation to state and prove this result. Results similar to Theorem 4 have been derived in the economics and game theory literature (see [14],[15]) but are inadequate for our purposes. Also, we deliberately use the same notation (like $P(\varepsilon)$, S etc.) for both, the Markov chain induced by the algorithm of section II and the general perturbed Markov chains as we wish to view the former as a special case of the latter and use the results of this section to analyze the algorithm.

A. Perturbed Markov Chains [9]

Let $P(0)$ be the transition probability matrix of a Markov chain on a finite state space S. We refer to this chain as the *unperturbed chain*. A *regular perturbation* of $P(0)$ consists of a stochastic matrix valued function $P(\varepsilon)$ on some interval $(0, a]$ that satisfies, for all $x, y \in S$,

1) $P(\varepsilon)$ is irreducible and aperiodic for each $\varepsilon \in (0, a]$ ($\Rightarrow \exists$ unique $\mu(\varepsilon)$ s.t. $\mu(\varepsilon) = \mu(\varepsilon)P(\varepsilon)$),
2) $\lim_{\varepsilon \to 0} P_{x,y}(\varepsilon) = P_{x,y}(0)$ and
3) if $P_{x,y}(\varepsilon) > 0$ for some ε, then $\exists\, r(x,y) \geq 0$ such that $0 < \lim_{\varepsilon \to 0} \varepsilon^{-r(x,y)} P_{x,y}(\varepsilon) < \infty$.

It follows that for a sufficiently small ε^*, $\exists\, 0 < \underline{\alpha}(x,y) < \overline{\alpha}(x,y) < \infty$, such that

$$\underline{\alpha}(x,y) < \varepsilon^{-r(x,y)} P_{x,y}(\varepsilon) < \overline{\alpha}(x,y),\ \forall\, \varepsilon < \varepsilon^*.$$

By denoting $\min_{x,y \in S} \underline{\alpha}(x,y) = \underline{\alpha}$ and $\max_{x,y \in S} \overline{\alpha}(x,y) = \overline{\alpha}$, we have

$$\underline{\alpha}\, \varepsilon^{r(x,y)} < P_{x,y}(\varepsilon) < \overline{\alpha}\, \varepsilon^{r(x,y)},\ \forall\, \varepsilon < \varepsilon^*. \tag{4}$$

Let $\mathfrak{L} = \{ f \in \mathfrak{C}^\infty \mid f(\varepsilon) = \sum_{i=1}^{L} a_i \varepsilon^{b_i}$ for some $a_i \in \mathbb{R}, b_i \geq 0 \}$ for some large enough but fixed $L \in \mathbb{N}$. The following assumption will be invoked later.

Assumption 2: For all $x, y \in S$, $P_{x,y}(\varepsilon) \in \mathfrak{L}$.

The following is an immediate consequence of the second requirement in the definition of a regular perturbation.

Lemma 3.1 (see [9], Lemma 1): The stationary distribution $\mu(\varepsilon)$ of $P(\varepsilon)$ satisfies $\lim_{\varepsilon \to 0} \mu(\varepsilon) = \mu(0)$, where $\mu(0)$ is a stationary distribution of the unperturbed chain $P(0)$.

We now develop some notation[2]:

1) A *path* from $a \in S$ to $b \in S$, $h(a \to b)$, is an ordered set $\{a = x_1, x_2, \ldots, x_n = b\} \subseteq S$ such that each transition $x_k \to x_{k+1}$ has positive 1-step probability in $P(\varepsilon)$; the resistance of a path is $r(h) = \sum_{k=1}^{n-1} r(x_k, x_{k+1})$ (where $r(\cdot,\cdot)$ is as in (4)).

[2]These definitions are adopted from relevant literature [1], [14], [15].

2) The *resistance* from x to y is given by $\rho(x,y) = \min\{r(h)|\ h(x \to y) \text{ is a path}\}$.
3) Given a subset $A \subset S$, its *co-radius* is given by $CR(A) = \max_{x \in S \setminus A} \min_{y \in A} \rho(x,y)$.
4) A *recurrence class* of a Markov chain is a non-empty subset of S s.t. once the state of the chain enters the set it remains in the set for all future times and there is a positive probability of transitioning from any point in the set to any other.

Thus, $\rho(x,y)$ is the resistance of the path with least resistance among all possible paths starting at state x and ending at state y and the co-radius of a set specifies the maximum resistance that must be overcome to enter it from outside. Since $P(\varepsilon)$ is irreducible for $\varepsilon > 0$, $\rho(x,y)$ is well defined for all $x,y \in S$. Also, notice that $r(x,y) = 0$ only for the one step transitions $x \to y$ allowed under $P(0)$.

With reference to the algorithm of section II-B, we would like to point out that the specific structure of the p.m.f. in (1), (2) and (3) are precisely designed to obtain appropriate values of $r(\cdot,\cdot)$. For instance, a content agent can only change its action by a transition with $r = c$ according to (1) while a discontent one can do so with $r = 0$ according to (2).

B. Ergodicity of nonhomogeneous Markov chains

We briefly recall results on ergodicity of a nonhomogeneous Markov chain on a finite state space S, with $Q(t)$ being the 1-step transition probability matrix at time t.

Definition 3.1 (Ergodicity): The chain is

- weakly ergodic (*WE*) if for all $t' \in \mathbb{N}$ and all $x,y,z \in S$,
$$\lim_{t \to \infty} |Q_{x,z}^{(t',t)} - Q_{y,z}^{(t',t)}| = 0.$$
- strongly ergodic (*SE*) if there exists a probability distribution π on S such that for any initial distribution η_0 on S and any $t' \in \mathbb{N}$,
$$\lim_{t \to \infty} \eta_0 Q^{(t',t)} = \pi.$$
We call π the limiting distribution of the chain.

Definition 3.2 (Ergodic Coefficient): Given a row stochastic matrix $Q \in \mathbb{R}^{|S| \times |S|}$, its ergodic coefficient is given by
$$\delta(Q) = 1 - \min_{x,y \in S} \sum_{z \in S} \min\{Q_{x,z}, Q_{y,z}\}.$$
The following result due to Doeblin provides a characterization for *WE* based on the ergodic coefficient.

Theorem 2 (see [16], Theorem 8.2): The chain is weakly ergodic if and only if there exists a strictly increasing sequence of positive integers $\{t_n\}_{n \in \mathbb{N}}$ such that
$$\sum_{n \in \mathbb{N}} (1 - \delta(Q^{(t_n, t_{n+1})})) = \infty. \quad (5)$$
The next Theorem provides a sufficiency condition for *SE*.

Theorem 3 (see [16], Theorem 8.3): Suppose the chain is weakly ergodic and at all t, there exists π_t such that $\pi_t Q(t) = \pi_t$ and
$$\sum_{t \in \mathbb{N}} \|\pi_{t+1} - \pi_t\|_1 < \infty, \quad (6)$$
then the chain is strongly ergodic. Furthermore, the limiting distribution π is the same as the limit of the sequence $\{\pi_t\}_{t \in \mathbb{N}}$.

C. Time-decreasing Noise and Ergodicity

Consider the nonhomogeneous Markov chain resulting from picking ε along the evolution of the perturbed chain $P(\varepsilon)$ as the sequence $\{\hat{\varepsilon}_t\}_{t \in \mathbb{N}}$; i.e. at each t, $\varepsilon = \hat{\varepsilon}_t$. When $\hat{\varepsilon}_t \to 0$ as $t \to \infty$, we refer to such a sequence as an *annealing schedule*. Denote the resulting nonhomogeneous chain by the bold-font $\mathbf{P}$, i.e. $\mathbf{P}(t) = P(\hat{\varepsilon}_t)$. The following technical lemma is easily proved.

Lemma 3.2 (see [13], Lemma 3.1): Let $\sum_{n \in \mathbb{N}} a(n) = \infty$ and $a(n) \geq a(n+1)\ \forall\, n$. Then for any $n', l \in \mathbb{N}$, $\sum_{n \in \mathbb{N}} a(n'+l+n) = \infty$.

Let us denote the recurrence classes of the unperturbed chain $P(0)$ as $E_1, ..., E_M$. Define
$$\kappa = \min_{E \in \{E_i\}} CR(E). \quad (7)$$

Theorem 4: Let the recurrence classes of an unperturbed chain $P(0)$ be aperiodic and the parameter ε in its regular perturbation $P(\varepsilon)$ be scheduled according to the monotone decreasing sequence $\{\hat{\varepsilon}_t\}_{t \in \mathbb{N}}$, with $\hat{\varepsilon}_t \to 0$ as $t \to \infty$, as described above. Then, a sufficient condition for weak ergodicity of the resulting nonhomogeneous Markov chain is
$$\sum_{t \in \mathbb{N}} \hat{\varepsilon}_t^{\kappa} = \infty.$$
Furthermore, if the chain is weakly ergodic and Assumption 2 holds, then it is strongly ergodic with the limiting distribution being $\mu(0)$ as described in Lemma 3.1.

Proof: (Weak Ergodicity) Let E^* be a recurrent class of $P(0)$ such that $CR(E^*) = \kappa$. Since E^* is aperiodic according to $P(0)$, there exists an $l_1 \in \mathbb{N}$ such that for all $m \geq l_1$ and $x,y \in E^*$, $P_{x,y}^m(0) > 0$ (see [16], Theorem 4.3, pp. 75). Since any path under $P(0)$ has zero resistance, once the chain enters a state in E^*, it can remain there with zero resistance via a path of any length greater than l_1.

Let $e^* \in E^*$ be such that $\exists\, x' \in S \setminus E^*$ such that $\rho(x', e^*) = \kappa$ i.e. the transition $x' \to e^*$ has the most resistance among all $x \to e^*$, $x \in S$. For all $x \in S$, consider the shortest paths $h(x \to e^*)$ such that $r(h(x \to e^*)) = \rho(x, e^*)$ and denote the length of such paths by $l(x, e^*)$. Let $l_2 = \max_{x \in S} l(x, e^*)$. So, by waiting for l_2 transitions, there is a path to e^* from all states $x \in S$ with resistance $\rho(x, e^*)$. Thus, by allowing more than

$l = l_1 + l_2$ transitions, we have for any $x \in S$ and a sufficiently small ε^*,

$$P^m_{x,e^*}(\varepsilon) > \underline{\alpha}^m \varepsilon^\kappa, \quad \forall\, \varepsilon < \varepsilon^*,\, m \geq l.$$

From (4), since $\hat{\varepsilon}_t \to 0$, for sufficiently large t,

$$\underline{\alpha}\, \hat{\varepsilon}_t^{r(x,y)} < \mathbf{P}_{x,y}(t) < \overline{\alpha}\, \hat{\varepsilon}_t^{r(x,y)}.$$

Consequently, by choosing a subsequence such that $t_{n+1} - t_n = l$, for sufficiently large n,

$$\mathbf{P}^{(t_n,t_{n+1})}_{x,e^*} > \underline{\alpha}^l \hat{\varepsilon}^\kappa_{t_{n+1}},\ \forall x \in S.$$

Then, for sufficiently large n, we can bound

$$\sum_{z\in S} \min\{\mathbf{P}^{(t_n,t_{n+1})}_{x,z}, \mathbf{P}^{(t_n,t_{n+1})}_{y,z}\} > \underline{\alpha}^l \hat{\varepsilon}^\kappa_{t_{n+1}},\ \forall x,y \in S.$$

Taking minimum over x,y, for sufficiently large n,

$$\min_{x,y\in S} \sum_{z\in S} \min\{\mathbf{P}^{(t_n,t_{n+1})}_{x,z}, \mathbf{P}^{(t_n,t_{n+1})}_{y,z}\} > \underline{\alpha}^l \hat{\varepsilon}^\kappa_{t_{n+1}}. \tag{8}$$

Since $\{t_n\}_{n\in\mathbb{N}}$ is an equally spaced subsequence, from Lemma 3.2 and the hypothesis of this Theorem, $\sum_{n\in\mathbb{N}} \hat{\varepsilon}^\kappa_{t_{n+1}} = \infty$. In view of this and (8), *WE* follows by noting that (5) is verified with $Q = \mathbf{P}$. The proof for *SE* involves invoking some known facts about the structure of $\mu(\varepsilon)$ and is omitted due to space constraints. It can be found in [13], Theorem 5. ■

IV. THE MODIFIED ALGORITHM

It is clear that the Markov chain $P(\varepsilon)$ defined by the algorithm of section II-B is a regular perturbation of the chain defined by the algorithm with $\varepsilon = 0$. Now, at time t, let each agent pick ε as $\hat{\varepsilon}_t$ for a given annealing schedule $\{\hat{\varepsilon}_t\}_{t\in\mathbb{N}}$. We refer to this new algorithm as the *Modified Algorithm* and denote the nonhomogeneous Markov chain defined by it by $\mathbf{P}(t)$. In the main result of this section, Theorem 5, we use Theorem 4 to obtain conditions for $\mathbf{P}(t)$ to retain $\mu(0)$ (as in Theorem 1) as its limiting distribution. We begin with some lemmas but exclude their proofs due to space limitations.

Define

$$C^0 = \{x \in S | x = [a,\overline{u},m], m_i = C,\ \forall\, i = 1,...,N\} \text{ and}$$
$$D^0 = \{x \in S | x = [a,\overline{u},m], m_i = D,\ \forall\, i = 1,...,N\}.$$

Lemma 4.1 ([1], Theorem 1): The recurrence classes of the unperturbed chain $P(0)$ are D^0 and the singletons $z \in C^0$.

Lemma 4.2 ([13], Lemma 4.5): For the Markov chain defined on S by the Modified Algorithm, κ as defined in (7) equals c.

Theorem 5 (Convergence Guarantee): Under Assumption 1, the nonhomogeneous Markov chain $\mathbf{P}(t)$ defined by the Modified Algorithm is strongly ergodic if

$$\sum_{t=1}^{\infty} \hat{\varepsilon}_t^c \quad = \quad \infty. \tag{9}$$

Furthermore, if (9) holds and $\mathbf{X}_t = [\mathbf{a}_t, \overline{\mathbf{u}}_t, \mathbf{m}_t]$ denotes the state of the chain at time t, then

$$\lim_{t\to\infty} \mathbb{P}[\mathbf{a}_t \in \mathbb{A}^*] = 1.$$

Proof: All transition probabilities in the algorithm of section II-B belong to $\mathfrak{L}$; thus Assumption 2 holds. For any $y \in D^0$ and $z \in C^0$, $P_{y,y}(0) > 0$ and $P_{z,z}(0) > 0$. Hence the recurrence classes of $P(0)$ are aperiodic and the first part of the theorem follows from Theorem 4 and Lemma 4.2.

Next, for any initial distribution η_0 on S and any subset $\tilde{S} \subset S$, $\mathbb{P}(\mathbf{X}_t \in \tilde{S}) = \sum_{j\in\tilde{S}} (\eta_0 \mathbf{P}^{(1,t)})_j$. Since (9) implies *SE* with limiting distribution $\mu(0)$ as in Theorem 1 and from the definition of *SE*, $\lim_{t\to\infty} \mathbb{P}(\mathbf{X}_t \in \tilde{S}) = \sum_{j\in\tilde{S}} \mu_j(0)$. Let $\tilde{S} = \{x \in C^0 | x = [a,\overline{u},m], W(a) = W^*,\ m_i = C\ \forall\, i\}$, then

$$\lim_{t\to\infty} \mathbb{P}[\mathbf{a}_t \in \mathscr{A}^*] = 1.$$

■

V. CONCLUSIONS

By restricting the rate of decrease in the annealing schedule, we derive the convergence guarantee we set out for in Theorem 5. From a practical viewpoint, while an annealing scheme seems natural given Theorem 1, the value of Theorem 5 lies in giving the system designer the freedom to choose amongst annealing schedules without worrying about adverse effects on convergence of the algorithm. Along the way, we also derive conditions for ergodicity of perturbed Markov chains with time-decreasing noise, which can be used to derive similar conditions for other algorithms based on the theory of perturbed Markov chains.

To illustrate the use of the proposed annealing scheme, consider a system with two agents A and B, and the payoff structure given by

$$\begin{bmatrix} (\frac{1}{2},\frac{1}{2}) & (0,\frac{2}{3}) \\ (\frac{2}{3},0) & (\frac{1}{3},\frac{1}{3}) \end{bmatrix},$$

where the first (second) number in the parenthesis of the (i,j) entry refers to the payoff received by agent A (B) when agents A and B play i and j respectively. This payoff structure corresponds to the classical game Prisoner's dilemma. The welfare maximizing joint action is $(1,1)$ with a welfare of 1. Let the agents implement the algorithm of section II-B with ε held constant and that of section IV with an annealing schedule $\hat{\varepsilon}_t = \frac{1}{\sqrt{t}}$. In Figure 1 we plot the resulting welfare profile and an indicator function which is non-zero whenever one of the agents is discontent against iterates. The first three plots correspond to the fixed ε runs with successively lower values of ε while the last one corresponds to the annealing schedule. In the first plot with $\varepsilon = 10^{-2}$, the welfare profile does not seem to converge to the optimal value 1 suggesting that $||\mu(\varepsilon) - \mu(0)||$ is relatively large. It is observed that as ε is reduced to 10^{-3} and 10^{-4} (second and third plots resp.), the welfare profiles do converges to 1. However, learning seems to take longer with smaller values of ε as suggested by the increasing duration for which

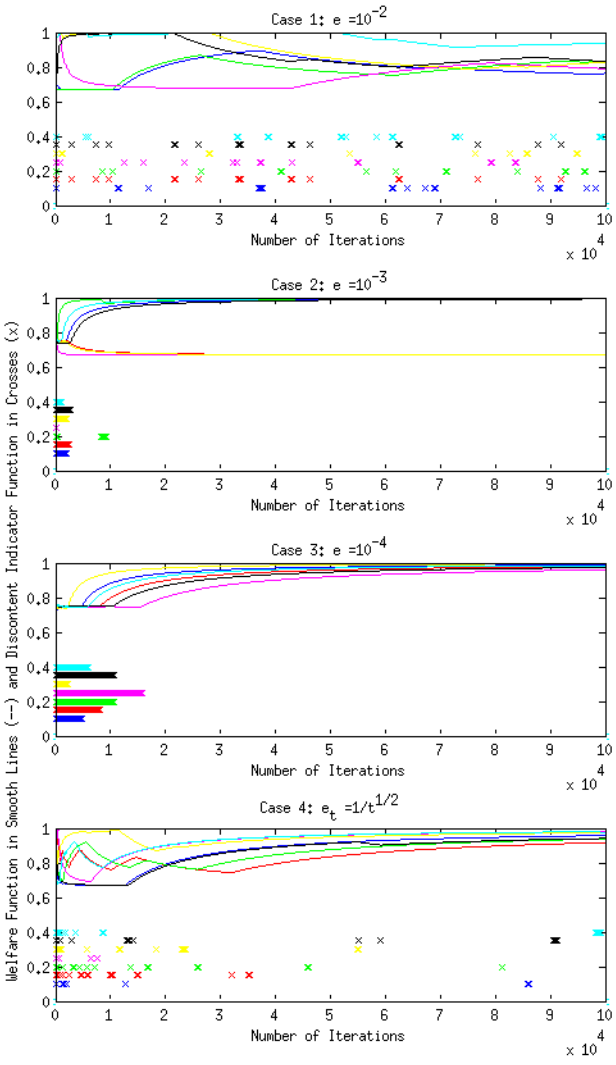

Fig. 1. Plots of welfare and discontent-indicator (horizontal '×'s) against iterates. Different colors represent different runs. Cases 1, 2 and 3 correspond to runs with ε fixed at 10^{-2}, 10^{-3} and 10^{-4} resp. while case 4 corresponds to $\hat{\varepsilon}_t = \frac{1}{\sqrt{t}}$.

the agents remain discontent. In the fourth plot where the agents implement the modified algorithm, we observe that by sweeping across a larger range of ε over time, the agents initially spend time *exploring* with relatively large $\hat{\varepsilon}_t$ and subsequently, with lower values of $\hat{\varepsilon}_t$, move to an *exploitation* phase after learning the welfare maximizing actions.

A concern that limits applicability of the algorithm remains: Assumption 1 must be satisfied by the utilities in the application. One way to mitigate this concern is to design proxy utilities as suggested in [8]: consider a undirected connected graph on the set of agents and assign proxy utility $\tilde{u}_i = u_i + \sum_{j \in \mathcal{N}(i)} u_j$, where $\mathcal{N}(i)$ is the adjacency list of i. Assumption 1 is satisfied by the proxies and, under appropriate conditions, the extrema of the welfare are the same for the utilities and the proxies. Intuitively, a condition like Assumption 1 ensures that each agent can be influenced to change its action; without such a condition one can always construct a payoff-structure where an agent which cannot be influenced picks actions that maximize its payoff but result in suboptimal joint action w.r.t. welfare. A conceivable approach then is to affect such influence by means of explicit inter-agent communication and will be pursued in future work [13].

An important question is determining the rate of convergence of the algorithm. One way to answer this question is to calculate the rate of convergence of $||\eta_t - \mu(0)||$ as $t \to \infty$, where η_t is the density of the state $\mathbf{X}_t$ of the modified algorithm. Such a calculation will also help design appropriate annealing schedules. We will address these issues in future work.

References

[1] J. R. Marden, H. P. Young, , and L. Y. Pao, "Achieving pareto optimality through distributed learning," *Discussion Paper Series, University of Oxford*, 2011.

[2] 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.

[3] N. Li and J. R. Marden, "Designing games for distributed optimization," in *Proc. of 50th IEEE Conference on Decision and Control (CDC-ECC), 2011*, pp. 2434–2440, 2011.

[4] M. Zhu and S. Martnez, "Distributed coverage games for energy-aware mobile sensor networks," *SIAM Journal on Control and Optimization*, vol. 51, no. 1, pp. 1–27, 2013.

[5] E. Altman and Z. Altman, "S-modular games and power control in wireless networks," *IEEE Transactions on Automatic Control*, vol. 48, no. 5, pp. 839–842, 2003.

[6] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games and Economic Behavior*, vol. 75, no. 5, pp. 788–808, 2012.

[7] 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, Feb 2009.

[8] J. R. Marden, S. D. Ruben, and L. Y. Pao, "A model-free approach to wind farm control using game theoretic methods," 2012. submitted for journal publication.

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

[10] G. Ellison, "Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution," *Review of Economic Studies*, vol. 67, no. 1, pp. 17–45, 2000.

[11] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 721–741, 1984.

[12] K. Aiba, "Waiting times in evolutionary dynamics with time-decreasing noise." preprint, 2011.

[13] A. Menon and J. S. Baras, "A distributed learning algorithm with bit-valued communications for multi-agent welfare optimization." Institute of Systems Research Technical Report. Available online at http://hdl.handle.net/1903/13702, 2013.

[14] J. Robles, "Evolution with changing mutation rates," *Journal of Economic Theory*, vol. 79, no. 2, pp. 207–223, 1998.

[15] M. Pak, "Stochastic stability and time-dependent mutations," *Games and Economic Behavior*, vol. 64, no. 2, pp. 650–665, 2008.

[16] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, vol. 31. Springer-Verlag, 1999.