Revisiting Log-Linear Learning: Asynchrony, Completeness and Payoff-Based Implementation*

Jason R. Marden† Jeff S. Shamma‡§

September 16, 2008
January 5, 2011 (revised)

Abstract

Log-linear learning is a learning algorithm with equilibrium selection properties. Log-linear learning provides guarantees on the percentage of time that the joint action profile will be at a potential maximizer in potential games. The traditional analysis of log-linear learning has centered around explicitly computing the stationary distribution. This analysis relied on a highly structured setting: i) players' utility functions constitute a potential game, ii) players update their strategies one at a time, which we refer to as *asynchrony*, iii) at any stage, a player can select any action in the action set, which we refer to as *completeness*, and iv) each player is endowed with the ability to assess the utility he would have received for any alternative action provided that the actions of all other players remain fixed. Since the appeal of log-linear learning is not solely the explicit form of the stationary distribution, we seek to address to what degree one can relax the structural assumptions while maintaining that only potential function maximizers are the stochastically stable action profiles. In this paper, we introduce slight variants of log-linear learning to include both synchronous updates and incomplete action sets. In both settings, we prove that only potential function maximizers are stochastically stable. Furthermore, we introduce a payoff-based version of log-linear learning, in which players are only aware of the utility they received and the action that they played. Note that log-linear learning in its original form is not a payoff-based learning algorithm. In payoff-based log-linear learning, we also prove that only potential maximizers are stochastically stable. The key enabler for these results is to change the focus of the analysis away from deriving the explicit form of the stationary distribution of the learning process towards characterizing the stochastically stable states. The resulting analysis uses the theory of resistance trees for regular perturbed Markov decision processes, thereby allowing a relaxation of the aforementioned structural assumptions.

1 Introduction

The theory of learning in games has sought to understand how and why equilibria emerge in non-cooperative games. Traditionally, social science literature develops *descriptive* game theoretic models for players, analyzes the limiting behavior, and generalizes the results for larger classes of games. Recently, there has been

*Research supported by the Social and Information Sciences Laboratory at the California Institute of Technology, AFOSR grants #FA9550-08-1-0375 and #FA9550-05-1-0321, and NSF grant #ECS-0501394.

†J. R. Marden is with the Department of Electrical, Computer and Energy Engineering, UCB 425, Boulder, Colorado 80309-0425, jason.marden@colorado.edu.

‡J. S. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Dr NW, Atlanta, GA 30332-0250, shamma@gatech.edu.

§Corresponding author.

a significant amount of research seeking to understand these behavioral models not from a descriptive point of view, but rather from a *prescriptive* point of view [2,14,15,27]. The goal is to use these behavioral models as a prescriptive control approach in distributed multi-agent systems where the guaranteed limiting behavior would represent a desirable operating condition.

A game theoretic approach to distributed control of multi-agent systems involves designing the interactions of the agents within a non-cooperative game framework. It turns out that a design of such distributed systems is strongly related to a class of non-cooperative games known as *potential games*, or more generally *weakly acyclic games* [15]. The reason for this stems from the fact that in distributed engineering systems each agent's utility function should be appropriately aligned to the global objective and this class of games captures this notion of alignment. This connection is important for two main reasons. First, potential games provide a paradigm for designing, analyzing and controlling multi-agent systems. In fact, there are existing methodologies for designing local agent utility functions that guarantee that the resulting game is a potential game [19, 30]. Furthermore, these methodologies also guarantee that the action profiles that maximize the global objective of the multi-agent system coincides with the potential function maximizers in the resulting potential game (see Section 2.2). Second, potential games have been studied extensively in the game theory literature and several established learning algorithms with guaranteed asymptotic results could be implemented to control these multi-agent systems.

Most of the learning algorithms for potential games guarantee convergence to a Nash equilibrium. A representative sampling of these algorithms includes Fictitious Play [21], Joint Strategy Fictitious Play [18], Adaptive Play [31], and many others [16, 20, 32, 33]. However, in potential games a pure Nash equilibrium may be inefficient with regards to the global objective. In certain settings, such as a congestion game with linear cost functions, it may be possible to bound the efficiency, but in general, such a bound need not be possible [24, 28]. Therefore, from a design perspective, having a learning algorithm that guarantees convergence to the most efficient Nash equilibrium is desirable. This is especially true when utility functions have been designed to ensure that the action profiles that maximize the global objective of the multi-agent system coincides with the potential function maximizers.

Log-linear learning, originally introduced in [8], is one of the few learning dynamics that embodies this

notion of equilibrium selection. In potential games, log-linear learning guarantees that only the joint action profiles that maximize the potential function are stochastically stable. The enabler for these results is the introduction of *noise* into the decision making process [8–10, 31]. In log-linear learning, this noise permits players to occasionally make mistakes, where mistakes represent the selection of suboptimal actions. The structure of the noise in log-linear learning is such that the probability of selecting a suboptimal action is associated with the magnitude of the payoff difference associated with a best response and the suboptimal action. As the noise vanishes, the probability that a player selects a suboptimal action goes to zero.

The traditional analysis of log-linear learning has centered around explicitly computing the stationary distribution. This analysis relies on a highly structured setting:

(i) Players' utility functions constitute a potential game.

(ii) Players update their strategies one at a time, which we refer to as *asynchrony*.[1]

(iii) At any stage, a player can select any action in the action set, which we refer to as *completeness*.

(iv) Each player is endowed with the ability to assess the utility he would have received for any alternative action provided that the actions of all other players remain fixed.

Nonetheless, log-linear learning has received significant research attention [1, 2, 6, 9, 10, 15, 32]. These results range from analyzing convergence rates [6, 25] to the necessity of the structural requirements [1]. In particular, [1] demonstrates that if the structural requirements of (i) and (ii) are relaxed arbitrarily, then the equilibrium selection properties of log-linear learning are no longer guaranteed.

Since the appeal of log-linear learning is not solely the explicit form of the stationary distribution, we seek to address to what degree one can relax the structural assumptions while maintaining that only potential function maximizers are stochastically stable. One motivation for this relaxation is that the structured setting of log-linear learning may inhibit its applicability as a distributed control mechanism in many distributed systems. However, these results are of broader interest in the setting of game theoretic learning.

Our main contribution in this paper is demonstrating that the structural assumption of log linear learning can be relaxed while maintaining that only potential function maximizers are stochastically stable. We

[1]It is worth noting that the usage of asynchrony in this paper is inconsistent with the standard usage of the term in computer science where the term references each agent's common knowledge of the clock.

introduce slight variants of log-linear learning to include both synchronous updates and incomplete action sets. In both settings, we prove that only potential function maximizers are stochastically stable. Furthermore, we introduce a payoff-based version of log-linear learning, in which players are only aware of the utility they received and the action that they played. Note that log-linear learning in its original form is not a payoff-based learning algorithm. In the payoff-based log-linear learning, we also prove that only potential maximizers are stochastically stable. This result follows a string of research analyzing payoff-based dynamics, also referred to as completely uncoupled dynamics, in games [4, 7, 11–13, 20, 34]. In general, the existing literature focuses on convergence to Nash equilibrium or ϵ-Nash equilibrium in different classes of games using only payoff-based information. In contrast to these results, our proposed algorithm provides convergence to the best Nash equilibrium, i.e., the potential function maximizer, using only payoff-based information.

The key enabler for these results is to change the focus of the analysis away from deriving the explicit form of the stationary distribution of the learning process towards characterizing the stochastically stable states. The resulting analysis uses the theory of resistance trees for regular perturbed Markov decision processes [31]. One important issue that is not addressed in this paper is convergence rates for the proposed algorithms. Recent results have proposed slight variations of log-linear learning that result in desirable convergence rates for a class of congestion games [3, 25] and social networks [23]. It remains an open and important research question to characterize convergence rates for variations of log-linear learning that are more suitable for distributed control.

The outline of the paper is as follows. In Section 2 we review the game theoretic concepts that we will use in this paper. In Section 3 we prove that log-linear learning in its original form is a regular perturbed Markov process. Utilizing this connection, we reprove the equilibrium selection properties of log-linear learning using the theory of resistance trees for regular perturbed Markov processes. Furthermore, we analyze the convergence properties of log-linear learning outside the realm of potential games. In Section 4, we analyze a variant of log-linear learning that includes synchronous updates. In Section 5, we show that if action sets are not complete, then potential function maximizers may not be stochastically stable in potential games. Furthermore, we derive a variant of the traditional log-linear learning that rectifies this problem. In

Section 6, we introduce a payoff-based version of log-linear learning that maintains the equilibrium selection properties of log-linear learning. Lastly, in Section 7 we provide some concluding remarks.

2 Setup

We consider a finite strategic-form game with n-player set $\mathcal{I} := \{1, ..., n\}$ where each player i has a finite action set $\mathcal{A}_i$ and a utility function $U_i : \mathcal{A} \to \mathbb{R}$ where $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$.

For an action profile $a = (a_1, a_2, ..., a_n) \in \mathcal{A}$, let a_{-i} denote the profile of player 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 will sometimes write a profile a of actions as (a_i, a_{-i}). Similarly, we may write $U_i(a)$ as $U_i(a_i, a_{-i})$. Let $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ denote the set of possible collective actions of all players other than player i. We define player i's *best response set* for an action profile $a_{-i} \in \mathcal{A}_{-i}$ as

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

2.1 Classes of Games

In this paper, we focus on three classes of games: identical interest, potential, and weakly acyclic. In each of these classes of games, a pure Nash equilibrium is guaranteed to exist. Each class of games imposes a restriction on the admissible utility functions.

2.1.1 Identical Interest Games

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

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

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

2.1.2 Potential Games

A significant generalization of an identical interest game is a potential game [22]. In a potential game, the change in a player's utility that results from a unilateral change in strategy equals the change in the global utility. Specifically, there is a function $\phi : \mathcal{A} \to \mathbb{R}$ such that for every player $i \in \mathcal{I}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) = \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}). \tag{1}$$

When this condition is satisfied, the game is called an (exact) potential game with the potential function ϕ. In potential games, any action profile maximizing the potential function is a pure Nash equilibrium, hence every potential game possesses at least one such equilibrium.

We will also consider a more general class of potential games known as *ordinal potential games*. In ordinal potential games there is a global function $\phi : \mathcal{A} \to \mathbb{R}$ such that for every player $i \in \mathcal{I}$, for every $a_{-i} \in \mathcal{A}_{-i}$, and for every $a_i', a_i'' \in \mathcal{A}_i$,

$$U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i}) > 0 \;\Leftrightarrow\; \phi(a_i', a_{-i}) - \phi(a_i'', a_{-i}) > 0.$$

2.1.3 Weakly Acyclic Games

Consider any game with a set $\mathcal{A}$ of action profiles. A *better reply path* is a sequence of action profiles $a^1, a^2, ..., a^L$ such that, for every $1 \leq \ell \leq L-1$, there is exactly one player i_ℓ such that i) $a_{i_\ell}^\ell \neq a_{i_\ell}^{\ell+1}$, ii) $a_{-i_\ell}^\ell = a_{-i_\ell}^{\ell+1}$, and iii) $U_{i_\ell}(a^\ell) < U_{i_\ell}(a^{\ell+1})$. In other words, one player moves at a time and that player increases his own utility. A *best reply path* is a better reply path with the additional requirement that each unilateral deviation is the result of a best response. More specifically, a best reply path is a sequence of action profiles $a^1, a^2, ..., a^L$ such that, for every $1 \leq \ell \leq L-1$, there is exactly one player i_ℓ such that i) $a_{i_\ell}^\ell \neq a_{i_\ell}^{\ell+1}$, ii) $a_{-i_\ell}^\ell = a_{-i_\ell}^{\ell+1}$, and iii) $a_i^{\ell+1} \in B_i(a_{-i}^\ell)$.

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

This idea may be generalized as follows. A game is *weakly acyclic* if for any $a \in \mathcal{A}$, there exists a better reply path starting at a and ending at some pure Nash equilibrium [32, 33]. A game is *weakly acyclic under best replies* if for any $a \in \mathcal{A}$, there exists a best reply path starting at a and ending at some pure Nash equilibrium [32, 33]. Potential games are special cases of weakly acyclic games.

An equivalent definition of weakly acyclic games utilizing potential functions is given in [15]. A game is weakly acyclic if and only if there exists a potential function $\phi : \mathcal{A} \to \mathbb{R}$ such that for any action $a \in \mathcal{A}$ that is *not* a Nash equilibrium, there exists a player $i \in \mathcal{I}$ with an action $a_i' \in \mathcal{A}_i$ such that $U_i(a_i', a_{-i}) > U_i(a_i, a_{-i})$ and $\phi(a_i', a_{-i}) > \phi(a_i, a_{-i})$.

2.2 Utility Design

Utilizing game theoretic tools for distributed control of multi-agent systems requires defining a local utility function for each agent. Designing these utility functions is nontrivial as there are several pertinent issues that need to be considered including scalability, locality, tractability, and efficiency of the resulting stable solutions [19]. The starting point of utility design for multi-agent systems is a global objective function of the form $G : \mathcal{A} \to R$ which captures the behavior that the system designer would like to achieve. The utility design question is how to distribute this global objective function to meet a variety of design objectives. One approach is the *wonderful life utility* [30], or marginal contribution utility, which takes on the form

$$U_i(a_i, a_{-i}) = G(a_i, a_{-i}) - G(a_i^0, a_{-i})$$

where $a_i^0 \in \mathcal{A}_i$ is fixed and referred to as the null action of player i. It is straightforward to verify that the resulting game is a potential game with potential function G. Therefore the potential function maximizers of the resulting game coincide with the optimal system behavior.

The classes of games consider in Section 2.1 provides a paradigm for designing these local utility functions where weakly acyclic games provides the most flexibility with respect to utility design. Research in utility design for multi-agent systems has sought to identify how to utilize this flexibility to meet various design objectives.[2]

[2]The results in this paper focus on all three classes of games as they are all relevant for multi-agent systems. For example, in [17] the authors showed that for the problem of rendezvous it is impossible to design utility functions within the framework of potential games such that all resulting Nash equilibrium satisfy a given coupled constraint performance criterion. However, using the extra flexibility of weakly acyclic games, utility functions could be designed to meet the desired performance criterions.

3 Log Linear Learning

In a repeated game, at each time $t \in \{0, 1, 2, \dots\}$, each player $i \in \mathcal{I}$ simultaneously chooses an action $a_i(t) \in \mathcal{A}_i$ and receives the utility $U_i(a(t))$ where $a(t) := (a_1(t), \dots, a_n(t))$. The action of player i is chosen at time t according to a probability distribution $p_i(t) \in \Delta(\mathcal{A}_i)$ where $\Delta(\mathcal{A}_i)$ denotes the set of probability distributions over the finite set $\mathcal{A}_i$. Let $p_i^{a_i}(t)$ denotes the probability that player i will select action a_i. We refer to $p_i(t)$ as the *strategy* of player i at time t.

The following learning algorithm is known as log-linear learning [8]. At each time $t > 0$, one player $i \in \mathcal{I}$ is randomly chosen and allowed to alter his current action. All other players must repeat their current action at the ensuing time step, i.e. $a_{-i}(t) = a_{-i}(t-1)$. At time t, player i employs the strategy $p_i(t) \in \Delta(\mathcal{A}_i)$ where

$$p_i^{a_i}(t) = \frac{e^{\frac{1}{\tau}U_i(a_i, a_{-i}(t-1))}}{\sum_{\bar{a}_i \in \mathcal{A}_i} e^{\frac{1}{\tau}U_i(\bar{a}_i, a_{-i}(t-1))}}, \tag{2}$$

for any action $a_i \in \mathcal{A}_i$ and temperature $\tau > 0$. The temperature τ determines how likely player i is to select a suboptimal action. As $\tau \to \infty$, player i will select any action $a_i \in \mathcal{A}_i$ with equal probability. As $\tau \to 0$, player i will select a best response to the action profile $a_{-i}(t-1)$, i.e., $a_i(t) \in B_i(a_{-i}(t-1))$ with arbitrarily high probability. In the case of a non-unique best response, player i will select a best response at random (uniformly).

Consider any potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$. In the repeated potential game in which all players adhere to log-linear learning, the stationary distribution of the joint action profiles is $\mu \in \Delta(\mathcal{A})$ where [8]

$$\mu(a) = \frac{e^{\frac{1}{\tau}\phi(a)}}{\sum_{\bar{a} \in \mathcal{A}} e^{\frac{1}{\tau}\phi(\bar{a})}}. \tag{3}$$

One can interpret the stationary distribution μ as follows. For sufficiently large times $t > 0$, $\mu(a)$ equals the probability that $a(t) = a$. As one decreases the temperature, $\tau \to 0$, all the weight of the stationary distribution μ is on the joint actions that maximize the potential function.

The above analysis characterizes the precise stationary distribution as a function of the temperature τ. The importance of the result is not solely the explicit form of the stationary distribution, but rather the recog-

nition that as $\tau \to 0$, the only stochastically stable states of the process are the joint actions that maximize the potential function. From this point on, rather than looking for the explicit stationary distribution, we focus on analyzing the stochastically stable states of the process using the theory of resistance trees for regular perturbed Markov decision processes. This relaxation will allow us to modify the traditional log-linear learning to include synchronous updates, incomplete action sets, and a payoff based implementation.

3.1 Background on Resistance Trees

The following is a very brief summary of the detailed review presented in [31]. Let P^0 denote the probability transition matrix for a finite state Markov chain over the state space Z. We refer to P^0 as the "unperturbed" process. Let $|Z|$ denote the number of states. 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 [31]. Specifically, the latter condition means that $\forall z, z' \in Z$,

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

and

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

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 $P^0_{z \to z'} > 0$ then $R(z \to z') = 0$.)

Construct a complete directed graph with $|Z|$ vertices, one for each state. The vertex corresponding to state z_j will be called j. The weight on the directed edge $i \to j$ is denoted as $\rho_{ij} = R(z_i \to z_j)$. A tree, T, rooted at vertex j, or j-tree, is a set of $|Z| - 1$ 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 $|Z| - 1$ edges that compose it. The *stochastic potential*, γ_j, of state z_j is defined to be the minimum resistance over all trees rooted at j. The following theorem gives a simple criterion for determining the stochastically stable states ([31], Lemma 1).

Theorem 3.1 *Let P^ϵ be a regular perturbed Markov process, and for each $\epsilon > 0$ let μ^ϵ be the unique stationary distribution of P^ϵ. Then $\lim_{\epsilon \to 0} \mu^\epsilon$ exists and the limiting distribution μ^0 is a stationary distribution of P^0. The stochastically stable states (i.e., the support of μ^0) are precisely those states with minimum stochastic potential. Furthermore, if a state is stochastically stable then the state must be in a recurrent class of the unperturbed process P^0.*

3.2 Proof for Log-Linear Learning Using Theory of Resistance Trees

Before discussing variants of log-linear learning, we will reprove the original result using the theory of resistance trees. The proof approach presented in this section can also be found in [5, 9]. We present the proof in detail for two reasons: i) the proof approach is similar to the forthcoming arguments for the variations of log-linear learning presented in the subsequent sections and ii) the structure of the proof can be exploited to analyze the limiting behavior of log-linear learning in games outside the realm of potential games.

Before proving that the only stochastically stable states are the potential function maximizers, we will prove two lemmas. The first lemma establishes that log-linear learning induces a regular perturbed Markov decision process.

Lemma 3.1 *Log-linear learning induces a regular perturbed Markov process where the unperturbed Markov process is an asynchronous best reply process and the resistance of any feasible transition $a^0 \to a^1 = (a_i^1, a_{-i}^0)$ is*

$$R(a^0 \to a^1) = \max_{a_i^* \in \mathcal{A}_i} U_i(a_i^*, a_{-i}^0) - U_i(a^1). \tag{4}$$

Proof: The unperturbed process, P^0, is the following: At each time $t > 0$, one player $i \in \mathcal{I}$ is randomly chosen and allowed to alter his current action. All other players must repeat their current action at the ensuing time step, i.e. $a_{-i}(t) = a_{-i}(t-1)$. At time t, player i selects a best response to the action profile $a_{-i}(t-1)$, i.e., $a_i(t) \in B_i(a_{-i}(t-1))$. In the case of multiple best responses, player i selects one at random (uniformly). We refer to these dynamics as an *asynchronous best reply process*.

Log-linear learning induces a finite, aperiodic, irreducible process over the state space $\mathcal{A}$. We will analyze this process with respect to $\epsilon := e^{-\frac{1}{\tau}}$ rather than the temperature τ. Let P^ϵ denote the associated

transition probability matrix. The probability of transitioning from a^0 to $a^1 := (a_i^1, a_{-i}^0)$ is

$$P^{\epsilon}_{a^0 \to a^1} = \frac{1}{n} \frac{\epsilon^{-U_i(a_i^1, a_{-i}^0)}}{\sum\limits_{a_i \in \mathcal{A}_i} \epsilon^{-U_i(a_i, a_{-i}^0)}}. \tag{5}$$

We assume for convenience that the updating player is selected randomly according to a uniform distribution.[3] Define the maximum utility of player i for any action profile $a_{-i} \in \mathcal{A}_{-i}$ as

$$V_i(a_{-i}) := \max_{a_i \in \mathcal{A}_i} U_i(a_i, a_{-i}).$$

Multiplying the numerator and denominator of (5) by $\epsilon^{V_i(a_{-i}^0)}$, we obtain

$$P^{\epsilon}_{a^0 \to a^1} = \frac{1}{n} \frac{\epsilon^{V_i(a_{-i}^0) - U_i(a_i^1, a_{-i}^0)}}{\sum\limits_{a_i \in \mathcal{A}_i} \epsilon^{V_i(a_{-i}^0) - U_i(a_i, a_{-i}^0)}}.$$

Accordingly,

$$\lim_{\epsilon \to 0^+} \frac{P^{\epsilon}_{a^0 \to a^1}}{\epsilon^{V_i(a_{-i}^0) - U_i(a_i^1, a_{-i}^0)}} = \frac{1}{n|B_i(a_{-i}^0)|}.$$

where $|B_i(a_{-i}^0)|$ denotes the size, or number of actions, in player i's best response set. This implies that the process P^{ϵ} is a regular perturbed Markov process where the resistance of the transition $a^0 \to a^1$ is

$$R(a^0 \to a^1) = V_i(a_{-i}^0) - U_i(a_i^1, a_{-i}^0). \tag{6}$$

Notice that $R(a^0 \to a^1) \geq 0$.

□

Before stating the second lemma we introduce some notation. A feasible action path $\mathcal{P}$ is a sequence of joint actions

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

that are the result of unilateral deviations, i.e., for each $k \in \{1, 2, ..., m\}$, $a^k = (a_i, a_{-i}^{k-1})$ for some player i and action $a_i \in \mathcal{A}_i$. The resistance of a path $\mathcal{P}$ is the sum of the resistance of each edge

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

[3]The player selection process could be relaxed to any probability distribution where the probability of selecting a given player is bounded away from 0.

Lemma 3.2 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$ where all players adhere to log-linear learning. For any feasible action path*

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

and reverse path

$$\mathcal{P}^R = \{a^m \to a^{m-1} \to ... \to a^0\},$$

the difference in the total resistance across the paths is

$$R(\mathcal{P}) - R(\mathcal{P}^R) = \phi(a^0) - \phi(a^m).$$

Proof: We will start by analyzing any edge in the feasible path $a^k \to a^{k+1}$. Suppose that $a^{k+1} = (a_i', a_{-i}^k)$ for some player i. The resistance across this edge is

$$R(a^k \to a^{k+1}) = V_i(a_{-i}^k) - U_i(a^{k+1}).$$

The resistance across the reverse edge $a^{k+1} \to a^k$ is

$$\begin{aligned} R(a^{k+1} \to a^k) &= V_i(a_{-i}^{k+1}) - U_i(a^k), \\ &= V_i(a_{-i}^k) - U_i(a^k), \end{aligned}$$

where the second equality is because $a_{-i}^k = a_{-i}^{k+1}$. The difference in the resistances across this edge is

$$\begin{aligned} R(a^k \to a^{k+1}) - R(a^{k+1} \to a^k) &= U_i(a^k) - U_i(a^{k+1}), \\ &= \phi(a^k) - \phi(a^{k+1}). \end{aligned}$$

Using the above equality, the difference in the resistances across the two paths is

$$\begin{aligned} R(\mathcal{P}) - R(\mathcal{P}^R) &= \sum_{k=0}^{m-1} R(a^k \to a^{k+1}) - R(a^{k+1} \to a^k), \\ &= \sum_{k=0}^{m-1} \left(\phi(a^k) - \phi(a^{k+1})\right), \\ &= \phi(a^0) - \phi(a^m). \end{aligned}$$

□

Before stating the following theorem, we introduce the notation of a resistance tree in the context of log-linear learning. A tree, T, rooted at an action profile a, is a set of $|\mathcal{A}| - 1$ directed edges such that, from every action profile a', there is a unique directed path in the tree to a. The resistance of a rooted tree, T, is the sum of the resistances on the edges

$$R(T) = \sum_{a' \to a'' \in T} R(a' \to a'').$$

Let $\mathcal{T}(a)$ be defined as the set of trees rooted at the action profile a. The stochastic potential of the action profile a is defined as

$$\gamma(a) = \min_{T \in \mathcal{T}(a)} R(T).$$

We refer to a minimum resistance tree as any tree that has minimum stochastic potential, that is, any tree T that satisfies

$$R(T) = \min_{a \in \mathcal{A}} \gamma(a).$$

Proposition 3.1 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$ where all players adhere to log-linear learning. The stochastically stable states are the set of potential maximizers, i.e., $\{a \in \mathcal{A} : \phi(a) = \max_{a^* \in \mathcal{A}} \phi(a^*)\}$.*

Proof: As mentioned earlier, the proposition follows from the known form of the stationary distribution. We now present a proof based on minimum resistance trees in preparation for the forthcoming analysis on variations of log-linear learning. By Lemma 3.1, we know that log-linear learning induces a regular perturbed Markov process. Therefore, an action profile $a \in \mathcal{A}$ is stochastically stable if and only if there exists a minimum resistance tree rooted at a.

Suppose that a minimum resistance tree, T, is rooted at an action profile a that does not maximize the potential function. Let a^* be any action profile that maximizes the potential function. Since T is a rooted tree, there exists a path $\mathcal{P}$ from a^* to a of the form

$$\mathcal{P} = \{a^* \to a^1 \to ... \to a^m \to a\}.$$

Notice that $\mathcal{P}$ is a feasible action path. Consider the reverse path $\mathcal{P}^R$ that goes from a to a^*,

$$\mathcal{P}^R = \{a \to a^m \to ... \to a^1 \to a^*\}.$$

Construct a new tree T' rooted at a^* by adding the edges of $\mathcal{P}^R$ to T and removing the redundant edges $\mathcal{P}$. The new tree will have the following resistance

$$R(T') = R(T) + R(\mathcal{P}^R) - R(\mathcal{P}). \tag{7}$$

By Lemma 3.2, we know that

$$\begin{aligned} R(T') &= R(T) + \phi(a) - \phi(a^*), \\ &< R(T). \end{aligned}$$

We constructed a new tree T' rooted at a^* with strictly less resistance than T. Therefore T cannot be a minimum resistance tree.

The above analysis can be repeated to show that all action profiles that maximize the potential function have the same stochastic potential; hence, all potential function maximizers are stochastically stable.

□

3.3 Additional Comments for Log-Linear Learning

The insight that log-linear learning is a regular perturbed Markov decision process enables us to analyze the stochastically stable states for any game structure (i.e., not only potential games) using Lemma 2 and Theorem 4 in [31].

Corollary 3.1 *Consider any finite n-player where all players adhere to log-linear learning. If an action profile is stochastically stable then the action profile must be contained in a best reply cycle, i.e., a best reply path of the form $a^* \rightarrow a^1 \rightarrow ... \rightarrow a^m \rightarrow a^*$.*

Proof: In any regular perturbed Markov process, the stochastically stable states must be contained in the recurrent classes of the unperturbed process. In the case of log-linear learning, the unperturbed process is an asynchronous best reply process. Therefore, regardless of the game structure, if an action profile is stochastically stable, it must be contained in a best reply cycle.

□

Consider any game that is (i) weakly acyclic under best replies and (ii) all Nash equilibrium are strict. For this class of games, the recurrent class of the unperturbed process are the strict Nash equilibrium, hence we obtain the following sharper characterization of the limiting behavior.

Corollary 3.2 *Consider any finite n-player weakly acyclic game under best replies where all players adhere to log-linear learning. If all Nash equilibria are strict, then the set of stochastically stable states must be contained in the set of Nash equilibrium.*

If a game is weakly acyclic under best replies then from any joint action there exists a best reply path leading to a pure Nash equilibrium. Since a strict Nash equilibrium is a recurrent class of our unperturbed process this gives us Corollary 3.2. In the case of non-strict Nash equilibrium, we fall back on Corollary 3.1 which gives us that the stochastically stable states must be contained in a best reply cycle.

The above analysis also holds true for games that are "close" to being potential games as stated in the following corollary.

Corollary 3.3 *Consider any finite n-player game with potential function $\phi : \mathcal{A} \rightarrow \mathbb{R}$ where all players adhere to log-linear learning. Suppose for any player $i \in \mathcal{I}$, actions $a_i', a_i'' \in \mathcal{A}_i$, and joint action $a_{-i} \in \mathcal{A}_{-i}$, players' utility satisfies*

$$\Big|\big(U_i(a_i', a_{-i}) - U_i(a_i'', a_{-i})\big) - \big(\phi(a_i', a_{-i}) - \phi(a_i'', a_{-i})\big)\Big| \leq \delta,$$

for some $\delta > 0$. If δ is sufficiently small, then the stochastically stable states are contained in the set of potential maximizers.

We omit the proof of Corollary 3.3 as it is identical to the resistance tree proof for log-linear learning in the case of an exact potential function. The only difference is the implication of Lemma 3.2 where equality is replaced by a bound on the difference of the resistances across two paths. This situation could arise when utility functions are corrupted with a slight degree of noise.

The main point of this section is that utilizing the theory of resistance trees to calculate the limiting behavior allows for a degree of relaxation in the learning process without significantly impacting the limiting behavior (or for that matter requiring new proofs). The price that we pay for this relaxation is that we

forego characterizing the precise stationary distribution in favor of characterizing the support of the limiting distribution.

4 Revisiting Asynchrony

In this section we explore whether asynchrony is necessary to guarantee the optimality of the stochastically stable states for log-linear learning in potential games. Do the properties of the stochastically stable states change if a limited number of players are allowed to update each period? Or are we bound by asynchrony?

Suppose at each time $t > 0$, a group of players $G \subseteq \mathcal{I}$ is randomly chosen according to a fixed probability distribution $q \in \Delta(2^{\mathcal{I}})$ where $2^{\mathcal{I}}$ denotes the set of subsets of $\mathcal{I}$ and q^G denotes the probability that group G will be chosen. We will refer to q as the *revision process*. At time t, each player $i \in G$ plays a strategy $p_i(t) \in \Delta(\mathcal{A}_i)$ where

$$p_i^{a_i}(t) = \frac{e^{\frac{1}{\tau}U_i(a_i, a_{-i}(t-1))}}{\sum\limits_{\bar{a}_i \in \mathcal{A}_i} e^{\frac{1}{\tau}U_i(\bar{a}_i, a_{-i}(t-1))}},$$

for any action $a_i \in \mathcal{A}_i$. All players not in G must repeat their previous action, i.e., $a_j(t) = a_j(t-1)$ for all $j \notin G$. We will refer to this learning algorithm as *synchronous log-linear learning* with revision process q. This setting is the focus of [1].

4.1 A Counterexample

Synchronous log-linear learning not only alters the resulting stationary distribution but also affects the set of stochastically stable states. The following example illustrates this phenomenon.

Consider a three player identical interest game with the following payoffs:

a_3

	a_2	b_2
a_1	0	0
b_1	0	1

b_3

	a_2	b_2
a_1	0	1
b_1	1	1

In this example, there are several Nash equilibria, namely all joint action profiles that yield a utility of 1 in addition to the action profile (a_1, a_2, a_3) that yields a utility of 0. Suppose the revision process q has full

support on the player sets $\{1, 2, 3, (1, 2, 3)\}$. An example of such a revision process is

$$\begin{aligned} q^1 &= 1/4, \\ q^2 &= 1/4, \\ q^3 &= 1/4, \\ q^{(1,2,3)} &= 1/4. \end{aligned}$$

We will see that *all* action profiles are stochastically stable.

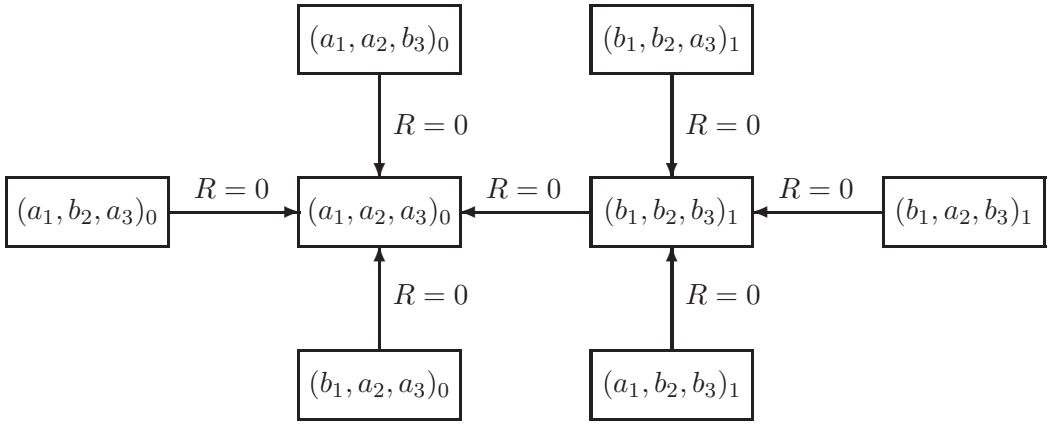
Before proceeding with the discussion, we first introduce the following notation. For any action profiles $a, a' \in \mathcal{A}$, the set of groups where the transition $a \to a'$ is feasible given the revision process q is defined as

$$\bar{G}(a, a', q) := \{G \subseteq \mathcal{I} : \{i \in \mathcal{I} : a_i \neq a'_i\} \subseteq G,\ q^G > 0\}.$$

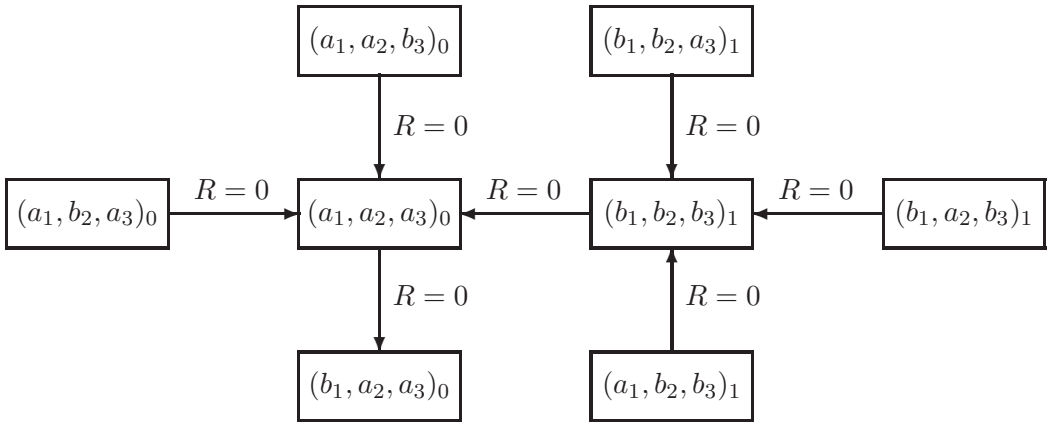
Synchronous log-linear learning with revision process q induces a regular perturbed Markov process, with the unperturbed process being synchronous learning under best replies. In synchronous log-linear learning, a transition $a^0 \to a^1$ is feasible if and only if $\bar{G}(a, a', q) \neq \emptyset$. It is straightforward to show that the resistance of a feasible transition $a^0 \to a^1$ is

$$R(a^0 \to a^1) = \min_{G \in \bar{G}(a^0, a^1, q)} \sum_{i \in G} \left(\max_{a_i^* \in \mathcal{A}_i} U_i(a_i^*, a_{-i}^0) - U_i(a_i^1, a_{-i}^0) \right). \tag{8}$$

Since synchronous log-linear learning is a regular perturbed Markov process we know that a minimum resistance tree must be rooted at any stochastically stable state. It is straightforward to show that any action profile that yields a utility of 1 has a rooted tree of resistance 0; hence all such action profiles are stochastically stable. Furthermore, there also exists a 0-resistance tree rooted at (a_1, a_2, a_3) of the form



Therefore (a_1, a_2, a_3) is also stochastically stable. Building off the same tree, it is straightforward to construct trees rooted at (b_1, a_2, a_3), (a_1, b_2, a_3), and (a_1, a_2, b_3) that also have 0 resistance, e.g.,



Therefore, all action profiles are stochastically stable.

4.2 Regular Revision Processes

Characterizing the stochastically stable states of a synchronous log-linear learning process is highly dependent on the structure of the revision process q in addition to the structure of the game. We start by defining a *regular revision process* (introduced in [1]).

Definition 4.1 (Regular Revision Process) *A regular revision process is a probability distribution $q \in \Delta(2^{\mathcal{I}})$ where for each player $i \in \mathcal{I}$, the probability assigned to the group consisting of only player i is positive, i.e., $q^i > 0$.*

The previous example demonstrates that any action profile may be stochastically stable even if we restrict our attention to regular revision processes and identical interest games.[4] However, one can easily characterize the stochastically stable states if the game embodies a special structure. This leads to the following theorem, which extends Theorem 2 in [1] to a larger class of games.

Theorem 4.1 *Consider any finite n-player weakly acyclic game under best replies where all Nash equilibria are strict. If all players adhere to synchronous log-linear learning with regular revision process q, then the*

[4]See [1] for alternative examples of games and revision processes outside the realm of regular revision processes that exhibit similar behavior.

stochastically stable states are contained in the set of Nash equilibria.[5]

Proof: Synchronous log-linear learning with regular revision process q is a regular perturbed Markov process; therefore, the stochastically stable states must be contained in the recurrent class of the unperturbed process. The unperturbed process is as follows: at each time $t > 0$, a group of players $G \subseteq \mathcal{I}$ is randomly chosen according to q. At time t, each player $i \in G$ selects an action from his best reply set, i.e., $a_i(t) \in B_i(a_{-i}(t-1))$. All players not in G must repeat their previous action, i.e., $a_j(t) = a_j(t-1)$ for all $j \notin G$.

For any game that is weakly acyclic game under best replies, if all Nash equilibria are strict, then the recurrent classes of the unperturbed process are precisely the set of Nash equilibria.

□

4.3 Independent Revision Process

In general, guaranteeing optimality of the stochastically stable states in potential games for an arbitrary revision process and utility interdependence structure is unattainable. In this section, we focus on particular class of revision processes that we refer to as *independent revision processes*. In such setting, each player independently decides whether to update his strategy by the log-linear learning rule with some probability $\omega > 0$. More precisely, at each time t, each player $i \in \mathcal{I}$ simultaneously plays the following strategy: (i) with probability $(1-\omega)$, player i selects his previous action, i.e., $a_i(t) = a_i(t-1)$, or (ii) with probability ω player i plays a strategy $p_i(t) \in \Delta(\mathcal{A}_i)$ where for any action $a_i \in \mathcal{A}_i$

$$p_i^{a_i}(t) = \frac{e^{\frac{1}{\tau}U_i(a_i, a_{-i}(t-1))}}{\sum\limits_{\bar{a}_i \in \mathcal{A}_i} e^{\frac{1}{\tau}U_i(\bar{a}_i, a_{-i}(t-1))}}.$$

Theorem 4.2 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$ where all players adhere to synchronous log-linear learning with an independent revision process with update parameter $\omega > 0$. Set $\omega = (e^{-\frac{1}{\tau}})^m$. For sufficiently large m, the stochastically stable states are the set of potential maximizers.*

[5]Any best response potential game is weakly acyclic game under best replies [29]. However, the converse is not true. The strictness condition could be relaxed as stated in the footnote of Corollary 3.2.

Note that Theorem 4.2 considers stochastic stability in terms of the *combined* perturbation of log-linear (vs. maximizing) action selection through τ and the group selection through ω. The perturbations are related by $\omega = (e^{-1/\tau})^m$, with $\tau \downarrow 0^+$ as usual.

Proof: The probability of transitioning from $a \to a'$ is

$$\sum_{S \subseteq \mathcal{I}: G \subseteq S} \epsilon^{m|S|}(1-\epsilon^m)^{|\mathcal{I}\setminus S|} \prod_{i \in S} \frac{\epsilon^{U_i(a_i', a_{-i})}}{\sum_{a_i'' \in \mathcal{A}_i} \epsilon^{U_i(a_i'', a_{-i})}}$$

where $G = \{i \in \mathcal{I} : a_i \neq a_i'\}$ and $\epsilon = e^{-\frac{1}{\tau}}$. It is straightforward to verify that synchronous log-linear learning with an independent revision process is a regular perturbed Markov process where the resistance of any transition $(a \to a')$ is

$$R(a \to a') = m|G| + \sum_{i \in G} \left(\max_{a_i^* \in \mathcal{A}_i} U_i(a_i^*, a_{-i}) - U_i(a_i', a_{-i}) \right).$$

The unperturbed process corresponds to players never experimenting, i.e., $\omega = 0$.

Without loss of generality, we will assume that $\phi(a)$ is bounded between 0 and 1 for all $a \in \mathcal{A}$. Utilizing this normalization, the resistance of the transition $(a \to a')$ satisfies the following inequality

$$m|G| + n \;\geq\; R(a \to a') \;\geq\; m|G|. \tag{9}$$

Let T be a minimum resistance tree rooted at a^*. If for each edge $[a \to \tilde{a}] \in T$ there is a single deviator, i.e., $|\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\}| = 1$, then arguments from the proof of Proposition 3.1 establish that a^* must be a potential function maximizer.

Suppose there exists an edge $[a \to \tilde{a}] \in T$ with multiple deviators, i.e., $|\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\}| \geq 2$. Let $G = \{i \in \mathcal{I} : a_i \neq \tilde{a}_i\}$. The resistance of this transition is at least

$$R(a \to \tilde{a}) \geq m|G|.$$

Consider any path $\mathcal{P} = \{a = a^0 \to a^1 \to ... \to a^{|G|} = \tilde{a}\}$ where each transition $a^k \to a^{k+1}$, $k \in \{0, .., |G|-1\}$ reflects a unilateral deviation by some player $i \in G$. According to (9), the resistance of each edge along this path is at most

$$R(a^{k-1} \to a^k) \leq m + n.$$

Therefore the resistance of the path $\mathcal{P}$ is at most

$$R(\mathcal{P}) \ \leq \ |G|(m+n) \ \leq \ m|G| + n^2.$$

Construct a new tree T' rooted at a^* by adding the edges of $\mathcal{P}$ to T and removing the redundant edges $\mathcal{P}^R$. The redundant edges are the set of edges leaving the action profiles $\{a^0, ..., a^{|G|-1}\}$ in the original tree T. The redundant edges $\mathcal{P}^R$ include the edge $[a \to \tilde{a}]$ in addition to $(|G|-1)$ other edges each of which has resistance at least m. The total resistance of the redundant edges is at least

$$R(\mathcal{P}^R) \geq m|G| + m(|G|-1).$$

The new tree T' has resistance

$$\begin{aligned} R(T') &= R(T) + R(\mathcal{P}) - R(\mathcal{P}^R), \\ &\leq R(T) + m|G| + n^2 - m|G| - m(|G|-1), \\ &= R(T) + n^2 - m(|G|-1). \end{aligned}$$

Since $|G| \geq 2$, if $m \geq n^2$ then $R(T') < R(T)$. This implies that any edge in a minimum resistance tree consists of only a single deviator which in turn implies that only potential function maximizers are stochastically stable.

□

4.4 Games Played on Graphs

The independent revision process discussed in the previous section guarantees that the only stochastically stable states are potential function maximizers irrespective of the structure of the utility interdependence. In this section, we focus on a particular structure of utility interdependence by considering games played on graphs where there are potentially a large number of players and each player's utility is only effected by the actions of a subset of other players. In such settings, we demonstrate that one can exploit this structure to expand the class of revision processes that guarantee optimality of the stochastically stable states.

In games played on graphs, each player's utility is influenced by a subset of other players. Let $N_i \subseteq \mathcal{I}$ represent the set of players that impact player i's utility. We will refer to N_i as the neighbors of player

i. In such a setting, player i's utility is of the form $U_i : \mathcal{A}^{N_i} \to \mathbb{R}$ where $\mathcal{A}^{N_i} := \prod_{j \in N_j} \mathcal{A}_j$. To avoid unnecessary notation, we will still express the utility of player i given the action profile a as $U_i(a)$. If player i's utility depends on the complete action profile, then $N_i = \mathcal{I}$.

We make the following assumption on the revision process q.

Assumption 4.1 *For any player $i \in \mathcal{I}$ there exists a group $G \subseteq \mathcal{I}$ such that $i \in G$ and $q^G > 0$.*

Proposition 4.1 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$. Let all players adhere to synchronous log-linear learning with a revision process satisfying Assumptions 4.1. Assume further that the revision process is such that any group of players G with $q^G > 0$ is conflict free, i.e.,*

$$i, j \in G,\ i \neq j \ \Rightarrow\ i \notin N_j.$$

Then the stochastically stable states are the set of potential maximizers.

Proof: For a given revision process q, a transition $a^0 \to a^1$ is possible if $\bar{G}(a^0, a^1, q) \neq \emptyset$. According to (8), it is straightforward to exploit the conflict free assumption to verify that

$$R(a^0 \to a^1) - R(a^1 \to a^0) = \phi(a^0) - \phi(a^1).$$

The remainder of the proof is identical to the proof for log-linear learning in Section 3.2. We note that this result also could have been proved by analyzing the precise stationary distribution using the detailed balance condition as set forth in [32].

□

This result is unsatisfying in that it requires a special structure on the revision process. Consider the following relaxation of the revision process. Roughly speaking, suppose that the revision process q usually selects conflict free groups but occasionally selects conflicted groups. To formalize this intuition, we make the following assumption on the revision process.

Assumption 4.2 *Let the revision process $q(\alpha)$ be continuously parameterized by $\alpha \geq 0$, where $q(\alpha)$ satisfies Assumption 4.1 for all α.*

- *For any G, $q^G(\alpha) > 0$ for some $\alpha > 0$ implies that there exists a $\rho \geq 0$ such that*

$$0 < \lim_{\alpha \to 0} \frac{q^G(\alpha)}{\alpha^\rho} < \infty;$$

- *For any conflicted (i.e., not conflict free) group C such that $q^C(\alpha) > 0$ and any player $i \in C$, there exists a conflict free group $F \subseteq \mathcal{I}$ such that*

  - $i \in F$;
  - $q^F(0) > 0$;
  - $q^F(\alpha) \geq \left(\frac{\kappa}{\alpha}\right) q^C(\alpha)$ *for some $\kappa > 0$ independent of F and C.*

Assumption 4.2 is of the presented form to ensure that the resulting process is a regular perturbed process, i.e., transitional probabilities decay at an exponentially smooth rate. One could write a simpler representation of this assumption but the resulting analysis would require relaxing the conditions for regular perturbed processes and the results in [31]. Such developments are unnecessary for this paper.

Theorem 4.3 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$ where all players adhere to synchronous log-linear learning with a revision process $q(\alpha)$ that satisfies Assumption 4.1–4.2. Set $\alpha = (e^{-\frac{1}{\tau}})^m$. For sufficiently large m, the stochastically stable states are the set of potential function maximizers.*

As with Theorem 4.2, Theorem 4.3 considers stochastic stability in terms of the *combined* perturbation of log-linear (vs. maximizing) action selection through τ and conflicted (vs. conflict free) group selection through α. The perturbations are related by $\alpha = (e^{-\frac{1}{\tau}})^m$, with $\tau \downarrow 0^+$ as usual.

Proof: Without loss of generality, we assume that $\phi(a)$ is bounded between 0 and 1 for all $a \in \mathcal{A}$. Synchronous log-linear learning with a revision process $q(\alpha)$ induces a regular perturbed Markov decision process. The unperturbed process consists of conflict free group selection with unilateral best replies.

Let T be a minimum resistance tree rooted at a^*. If for each edge $[a \to \tilde{a}] \in T$ there exists a conflict free group G such that $q^G > 0$ and $\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\} \subseteq G$ then arguments from the proof of Proposition 4.1 establish that a^* must be a potential function maximizer.

Suppose there exists an edge $[a \to \tilde{a}]$ where there does not exist a conflict free group G such that $q^G > 0$ and $\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\} \subseteq G$. From Assumption 4.2, the probability of this transition is at most order $\epsilon^{m\rho}$ for some $\rho \geq 1$. Accordingly, the associated resistance satisfies $R(a \to \tilde{a}) \geq m\rho$.

By Assumption 4.2, there exists a path $\mathcal{P} = \{a = a^0 \to a^1 \to ... \to a^l = \tilde{a}\}$ of at most length $|\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\}|$, where each transition $a^{k-1} \to a^k,\ k \in \{1, ..., l\}$ reflects a unilateral deviation that can be accomplished by a conflict free group. Assumption 4.2 further implies that the resistance of each transition $a^{k-1} \to a^k$ along the path $\mathcal{P}$ is is at most

$$R(a^{k-1} \to a^k) \quad \leq \quad n.$$

The main idea behind the above inequality is as follows. Resistance computation is based on the product of the probability of a group being selected and the probability of an action being selected. There exists a conflict free group with nonvanishing probability that can accomplish the desired transition. The selection of this group contributes zero to the resistance value. The action selection contributes to the resistance by at most n (cf., equation (8) and the assumed normalized bounds of the potential function.) The resistance along the path $\mathcal{P}$ is bounded by

$$R(\mathcal{P}) \ \leq \ |\{i \in \mathcal{I} : a_i \neq \tilde{a}_i\}|n \ \leq \ n^2.$$

We can conclude that for $m\rho > n^2$,

$$R(\mathcal{P}) < R(a \to \tilde{a}).$$

Construct a new tree T' still rooted at a^* by adding the edges of $\mathcal{P}$ to T and removing the redundant edges. The new tree T' has strictly less resistance than T, contradicting the assumption that T was a minimum resistance tree. Hence, all minimum resistance trees must consist of only conflict free transitions. This implies that only potential function maximizers are stochastically stable.

□

5 Revisiting Completeness

One role for learning algorithms in distributed control is to guide the decisions of players in real time. That is, the iterations of a learning algorithm correspond to the sequential decisions of players. In such a setting,

players may not have the ability to select any particular action in their action set at any given time. One example is multi-vehicle motion control, where an agent's action set represents discrete spatial locations. Mobility limitations restrict the ability to traverse from one location to another in a given time period.

Standard log-linear learning assumes an agent can access any action at each iteration, which we refer to as "completeness". Furthermore, in order to implement log-linear learning agents must have access to information, i.e., hypothetical payoffs, for all possible actions. Even in non-control theoretic applications, having this degree of information when each player's action set is quite large is demanding. Rather, having an algorithm that allows each agent to select the next action using only *local* information is desirable. These considerations motivate the introduction of constraints between iterations of a learning algorithm.

Let $a(t-1)$ be the joint action at time $t-1$. With constrained action sets[6], the set of actions available to player i at time t is a function of his action at time $t-1$ and will be denoted as $C_i(a_i(t-1)) \subseteq \mathcal{A}_i$. We will adopt the convention that $a_i \in C_i(a_i)$ for any action $a_i \in \mathcal{A}_i$, i.e., a player is always allowed to stay with his previous action. We will say that a player's action set is *complete* if $C_i(a_i) = \mathcal{A}_i$ for all actions $a_i \in \mathcal{A}_i$.[7]

We make the following two assumptions on constrained action sets.

Assumption 5.1 *For any player $i \in \mathcal{I}$ and any action pair $a_i^0, a_i^m \in \mathcal{A}_i$, there exists a sequence of actions $a_i^0 \to a_i^1 \to \cdots \to a_i^m$ satisfying $a_i^k \in C_i(a_i^{k-1})$ for all $k \in \{1, 2, \ldots, m\}$.*

Assumption 5.2 *For any player $i \in \mathcal{I}$ and any action pair $a_i^1, a_i^2 \in \mathcal{A}_i$,*

$$a_i^2 \in C_i(a_i^1) \Leftrightarrow a_i^1 \in C_i(a_i^2).$$

5.1 A Counterexample

Log-linear learning, in its original form, requires completeness of players' action sets. Suppose log-linear learning is employed in a situation with constrained action sets. At time t, player i plays a strategy $p_i(t) \in$

[6]One could view constrained action sets as either a constraint on available moves for distributed engineering systems or a constraint on information in game theoretic learning.

[7]We note that this scenario could have been formulated as a stochastic game [26] where the state is defined as the previous action profile and the state dependent action sets are defined according to the constrained action sets. We avoid formally defining the game as a stochastic game in favor of a more direct treatment.

$\Delta(\mathcal{A}_i)$ where

$$p_i^{a_i}(t) = \begin{cases} \frac{e^{\frac{1}{\tau}U_i(a_i,a_{-i}(t-1))}}{\sum_{\bar{a}_i \in C_i(a_i)} e^{\frac{1}{\tau}U_i(\bar{a}_i,a_{-i}(t-1))}} & \text{for any action } a_i \in C_i(a_i), \\ 0 & \text{for any action } a_i \notin C_i(a_i). \end{cases} \tag{10}$$

It is easy to see that log-linear learning in this setting induces a regular perturbed Markov process. The resistance of a feasible transition $a^0 \rightarrow a^1$, i.e., a^1 is of the form (a_i^1, a_{-i}^0) for some player i and action $a_i^1 \in C_i(a_i^0)$, is now

$$R(a^0 \rightarrow a^1) = \max_{a_i^* \in C_i(a_i^0)} U_i(a_i^*, a_{-i}^0) - U_i(a^1).$$

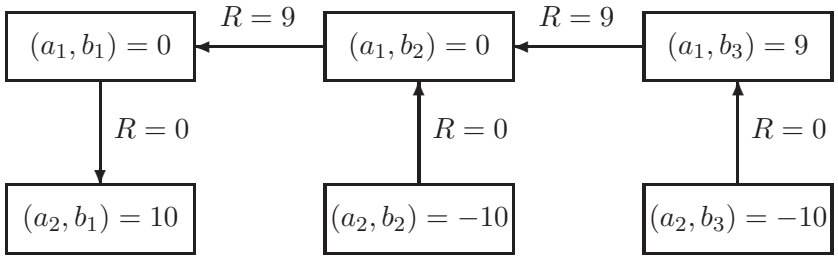
The following example demonstrates that log-linear learning applied to constrained action sets need not result in stochastically stable joint actions that are potential function maximizer. Consider a two player identical interest game with payoffs

		Player 2		
		b_1	b_2	b_3
Player 1	a_1	0	0	9
	a_2	10	-10	-10

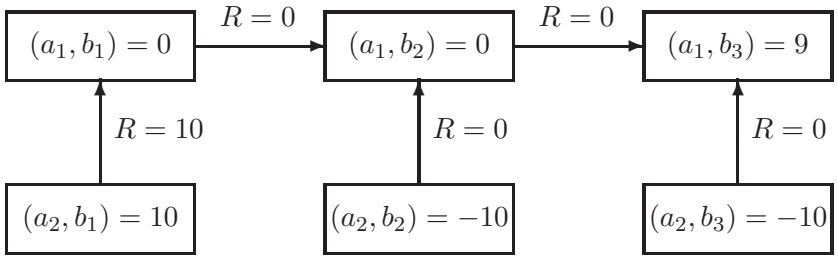
In this example, there are two Nash equilibria (a_2, b_1) and (a_1, b_3), but only one potential maximizer (a_2, b_1). Let the constrained action sets satisfy

$$\begin{aligned} C_1(a_1) &= \{a_1, a_2\}, \\ C_1(a_2) &= \{a_1, a_2\}, \\ C_2(b_1) &= \{b_1, b_2\}, \\ C_2(b_2) &= \{b_1, b_2, b_3\}, \\ C_2(b_3) &= \{b_2, b_3\}. \end{aligned}$$

A state is stochastically stable if and only if the state has minimum stochastic potential. The stochastic potential of state (a_2, b_1) is 18 highlighted by the following resistance tree



The stochastic potential of state (a_1, b_3) is 9 highlighted by the following resistance tree



Hence, (a_2, b_1) is not stochastically stable. In fact, one can construct an alternative example which illustrate that the stochastically stable states need not be contained in the set of Nash equilibria for potential games when action sets are incomplete.

The key insight for this example lies in the best reply graph. In log-linear learning with complete action sets, the recurrent classes of the unperturbed process are the Nash equilibria. Constraining action sets changes the structure of the best reply graph. In particular, these changes may induce new *local* Nash equilibria, i.e., action profiles where no player can unilaterally improve his utility by selecting an action in his constrained action set. These local Nash equilibria are now recurrent classes of the unperturbed process and are candidates for the stochastically stable states of the perturbed process.

Furthermore, this example illustrates that Lemma 3.2 no longer holds in the constrained setting because the difference in the resistances across two path is no longer a function of the endpoints.

5.2 Binary Log-Linear Learning

In this subsection, we introduce a variant of log-linear learning that will rectify the problems caused by constrained action sets.

Consider the following learning algorithm, which we refer to as *binary log-linear learning* originally

proposed in [2, 15]. At each time $t > 0$, one player $i \in \mathcal{I}$ is randomly chosen (uniformly) and allowed to alter his current action. All other players must repeat their current action at the ensuing time step, i.e. $a_{-i}(t) = a_{-i}(t-1)$. At time t, player i selects one trial action $\hat{a}_i$ (uniformly) from his constrained action set $C_i(a_i(t-1)) \subset \mathcal{A}_i$. Player i plays a strategy $p_i(t) \in \Delta(\mathcal{A}_i)$ where

$$\begin{aligned} p_i^{a_i(t-1)}(t) &= \frac{e^{\frac{1}{\tau}U_i(a(t-1))}}{e^{\frac{1}{\tau}U_i(a(t-1))} + e^{\frac{1}{\tau}U_i(\hat{a}_i, a_{-i}(t-1))}}, \\ p_i^{\hat{a}_i}(t) &= \frac{e^{\frac{1}{\tau}U_i(\hat{a}_i, a_{-i}(t-1))}}{e^{\frac{1}{\tau}U_i(a(t-1))} + e^{\frac{1}{\tau}U_i(\hat{a}_i, a_{-i}(t-1))}}, \\ p_i^{a_i}(t) &= 0, \ \ \forall\, a_i \neq a_i(t-1), \hat{a}_i \end{aligned}$$

If $a_i(t-1) = \hat{a}_i$, then $p_i^{a_i(t-1)} = 1$.[8]

Reference [15] analyzes the stationary distribution of binary log-linear learning when action sets are constrained. In particular, if the probability of selecting a trial action satisfies a certain specified condition then the stationary distribution of (3) is preserved. This condition involves accommodating for the time-varying cardinality $|C_i(a_i(t-i))|$. We now show that this issue can be resolved by changing the focus of the analysis from the stationary distribution to the stochastically stable states.

Theorem 5.1 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$ and constrained action sets satisfying Assumptions 5.1 and 5.2. If all players adhere to binary log-linear learning, then the stochastically stable states are the set of potential maximizers.*

Before proving that the only stochastically stable states are the potential maximizers, we introduce two lemmas. The first lemma establishes that binary log-linear learning is a regular perturbed Markov decision process.

Lemma 5.1 *Binary log-linear learning induces a regular perturbed Markov process where the resistance of any feasible transition, $a^0 \to a^1 = (a_i^1, a_{-i}^0)$ where $a_i^1 \in C_i(a_i^0)$, is*

$$R(a^0 \to a^1) = \max_{a_i^* \in \{a_i^0, a_i^1\}} U_i(a_i^*, a_{-i}^0) - U_i(a^1).$$

[8]The results presented in this section still hold even if the probability distributions are not uniform. In the case of the player selection process, any probability distribution with full support on the player set $\{1, 2, ..., n\}$ would work. The probability of selecting a trial action can be relaxed in the same fashion.

Proof: The unperturbed process employs a maximizing strategy as opposed to log-linear strategy. Binary log-linear learning with constrained action sets induces a finite, aperiodic, irreducible process over the state space $\mathcal{A}$. Irreducibility is a consequence of Assumption 5.1. Let $\epsilon := e^{-\frac{1}{\tau}}$. Let P^ϵ denote the associated probability transition matrix. The probability of transitioning from a^0 to $a^1 := (a_i^1, a_{-i}^0)$ where $a_i^1 \in C_i(a_i^0)$ is now of the form

$$P^\epsilon_{a^0 \to a^1} = \left(\frac{1}{n|C_i(a_i^0)|}\right)\left(\frac{\epsilon^{-U_i(a^1)}}{\epsilon^{-U_i(a^0)} + \epsilon^{-U_i(a^1)}}\right). \tag{11}$$

Redefine the functions:

$$\begin{aligned} V_i(a^0, a^1) &:= \max\{U_i(a^0), U_i(a^1)\}, \\ B_i(a^0, a^1) &:= \{a \in \{a^0, a^1\} : U_i(a) = V_i(a^0, a^1)\}. \end{aligned}$$

Accordingly,

$$\lim_{\epsilon \to 0^+} \frac{P^\epsilon_{a^0 \to a^1}}{\epsilon^{V_i(a^0,a^1) - U_i(a^1)}} = \frac{1}{n|C_i(a_i^0)||B_i(a^0, a^1)|}.$$

This implies that the process P^ϵ is a regular perturbed Markov process where the resistance of the transition $a^0 \to a^1$ is

$$R(a^0 \to a^1) = V_i(a^0, a^1) - U_i(a^1) \geq 0.$$

□

We state the following lemma without proof as it is almost identical to the proof of Lemma 3.2.

Lemma 5.2 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$, where all players adhere to binary log-linear learning. For any feasible action path*

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

and reverse path

$$\mathcal{P}^R = \{a^m \to a^{m-1} \to ... \to a^0\},$$

the difference in the total resistance across the paths is

$$R(\mathcal{P}) - R(\mathcal{P}^R) = \phi(a^0) - \phi(a^m).$$

Proof of Theorem 5.1: By Lemma 5.1, we know that binary log-linear learning induces a regular perturbed Markov process. Therefore, an action profile $a \in \mathcal{A}$ is stochastically stable if and only if there exists a minimum resistance tree rooted at a.

Suppose that a minimum resistance tree, T, was rooted at an action profile a that does not maximize the potential function. Let a^* be any action profile that maximizes the potential function. Since T is a rooted tree, there exists a path $\mathcal{P}$ from a^* to a of the form

$$\mathcal{P} = \{a^* \rightarrow a^1 \rightarrow ... \rightarrow a^m \rightarrow a\}.$$

Notice that $\mathcal{P}$ is a feasible action path. Consider the reverse path $\mathcal{P}^R$ that goes from a to a^*,

$$\mathcal{P}^R = \{a \rightarrow a^m \rightarrow ... \rightarrow a^1 \rightarrow a^*\}.$$

Such a path exists because of Assumption 5.2. Construct a new tree T' rooted at a^* by adding the edges of $\mathcal{P}^R$ to T and removing the redundant edges $\mathcal{P}$. The new tree will have the following resistance

$$R(T') = R(T) + R(\mathcal{P}^R) - R(\mathcal{P})$$

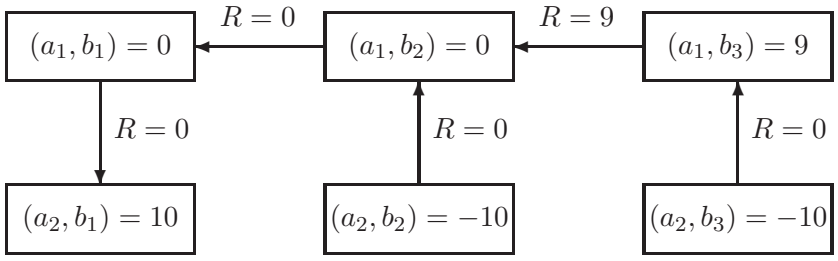
By Lemma 3.2, we know that

$$\begin{aligned} R(T') &= R(T) + \phi(a) - \phi(a^*), \\ &< R(T). \end{aligned}$$

Therefore, we constructed a new tree T' rooted at a^* with strictly less resistance than T. Accordingly, T is cannot be a minimum resistance tree.

The above analysis can be repeated to show that all action profiles that maximize the potential function have the same stochastic potential; hence, all potential function maximizers are stochastically stable. □

When revisiting the example in Section 5.1, the stochastic potential of state (a_2, b_1) is now 9 highlighted by the following resistance tree



The potential maximizer (a_2, b_1) is the only stochastically stable state.

6 Payoff Based Implementation

In any of version of log-linear learning, each player needs to be endowed with the ability to assess the utility that the player would have received had the player selected an alternative action. That is, given any action profile a, each player i is able to compute $U_i(a_i', a_{-i})$ for all $a_i' \in \mathcal{A}_i$. In this section, we seek to understand what happens if players do not have access to this information. Rather, players have access to *only* (i) the action they played and (ii) the utility they received.

Payoff-based learning algorithms have received significant attention recently [4, 11, 12, 20, 34]. While most of this research focusing on introducing dynamics that show that pure Nash equilibria are stochastically stable, in this section we introduce a payoff-based learning algorithm such that only potential function maximizers are stochastically stable. In fact, we combine two of the algorithms perviously studied, synchronous log-linear learning with an independent revision process in Section 4.3 and binary log-linear learning in Section 5.2, to develop a payoff based version of log-linear learning with the desired properties.

We introduce the following learning algorithm called *payoff based log-linear learning*. For any time $t \in \{1, 2, ...\}$, let $a(t-1)$ and $a(t)$ be the action profile at time $t-1$ and t, respectively. Define $x_i(t) \in \{0, 1\}$ to be a binary flag that indicates whether player i experimented in time t. At time $t+1$ each player i simultaneously selects an action $a_i \in \mathcal{A}_i$ according to the following rule:

- If player i did not experiment in period t, i.e., $x_i(t) = 0$, then
  - With probability ω,
    - $a_i(t+1)$ is chosen randomly (uniformly) over $\mathcal{A}_i$ with probability ω,

$*$ $x_i(t+1) = 1$,

– With probability $(1-\omega)$,

$*$ $a_i(t+1) = a_i(t)$,

$*$ $x_i(t+1) = 0$.

where $\omega > 0$ is the exploration rate.

- If player i experimented in period t, i.e., $x_i(t) = 1$, then

$$a_i(t+1) = \begin{cases} a_i(t-1) & \text{with probability } \frac{e^{\frac{1}{\tau}U_i(a(t-1))}}{e^{\frac{1}{\tau}U_i(a(t-1))}+e^{\frac{1}{\tau}U_i(a(t))}} \\ a_i(t) & \text{with probability } \frac{e^{\frac{1}{\tau}U_i(a(t))}}{e^{\frac{1}{\tau}U_i(a(t-1))}+e^{\frac{1}{\tau}U_i(a(t))}} \end{cases} \tag{12}$$

and

$$x_i(t+1) = 0.$$

These dynamics can be described as follows. Occasionally players experiment with a new action. In the event that a player does experiment, then the player switches to the new action with a log-linear strategy over the utility received in the previous two time periods.

The above dynamics require that a player knows only his own utility received and action employed in the last two time steps and whether he experimented or not in the previous time step. These dynamics do not require players to have any knowledge regarding the actions, strategies, or utilities of the other players.

Theorem 6.1 *Consider any finite n-player potential game with potential function $\phi : \mathcal{A} \to \mathbb{R}$. where all players adhere to payoff-based log-linear learning. Set $\omega = (e^{-\frac{1}{\tau}})^m$. For sufficiently large m, the stochastically stable states are the set of potential maximizers.*

In a similar manner to Theorem 4.3, Theorem 6.1 considers stochastic stability in terms of the *combined* perturbation of log-linear action selection through τ and experimentation through ω. The perturbations are related by $\omega = (e^{-\frac{1}{\tau}})^m$.

The remainder of this subsection is devoted to the proof of Theorem 6.1. Without loss of generality, we will assume that $\phi(a)$ is bounded between 0 and 1 for all $a \in \mathcal{A}$.

First, define the *state* of the dynamics at time t to be the tuple $z(t) := [a(t-1), a(t), x(t)] \in \mathcal{A} \times \mathcal{A} \times \{0,1\}^n$.

Claim 6.1 *Payoff-based log-linear learning induces a regular perturbed Markov decision process. Furthermore, the resistance of a transition from state $z^1 = [a^0, a^1, x^1]$ to state $z^2 = [a^1, a^2, x^2]$ is*

$$R(z^1 \to z^2) = \underbrace{ms(x^2)}_{*} + \underbrace{\sum_{i:x_i^1=1, a_i^2=a_i^0} \big(V_i(a^0,a^1) - U_i(a^0)\big) + \sum_{i:x_i^1=1, a_i^2=a_i^1} \big(V_i(a^0,a^1) - U_i(a^1)\big)}_{**} \quad (13)$$

where

$$\begin{aligned} V_i(a^0,a^1) &:= \max\{U_i(a^0), U_i(a^1)\}, \\ s(x) &:= \sum_i x_i. \end{aligned}$$

Proof: The unperturbed process corresponds to players never experimenting, i.e., $\omega = 0$. The recurrent classes of the unperturbed process are states of the form $[a, a, \mathbf{0}]$, where boldface $\mathbf{0}$ denotes a zero vector of appropriate dimension.

Valid transitions of the perturbed process are of the following form:

$$[a^0, a^1, x^1] \to [a^1, a^2, x^2],$$

where

$$x_i^1 = 0 \quad \Rightarrow \quad \begin{cases} x_i^2 = 0,\ a_i^2 = a_i^1, \text{ or} \\ x_i^2 = 1,\ a_i^2 \in \mathcal{A}_i, \end{cases}$$

$$x_i^1 = 1 \quad \Rightarrow \quad x_i^2 = 0,\ a_i^2 \in \{a_i^0, a_i^1\}.$$

Let $z^1 := [a^0, a^1, x^1]$, $z^2 := [a^1, a^2, x^2]$, and $\epsilon := e^{-\frac{1}{\tau}}$. The probability of the transition from $z^1 \to z^2$ is

$$\begin{aligned} P^{\epsilon}_{z^1 \to z^2} = & \\ & \left(\prod_{i:x_i^1=0, x_i^2=0} (1-\omega) \right) \left(\prod_{i:x_i^1=0, x_i^2=1} \frac{\omega}{|\mathcal{A}_i|} \right) \\ & \left(\prod_{i:x_i^1=1, a_i^2=a_i^0} \frac{\epsilon^{-U_i(a^0)}}{\epsilon^{-U_i(a^0)} + \epsilon^{-U_i(a^1)}} \right) \left(\prod_{i:x_i^1=1, a_i^2=a_i^1} \frac{\epsilon^{-U_i(a^1)}}{\epsilon^{-U_i(a^0)} + \epsilon^{-U_i(a^1)}} \right) \end{aligned}$$

It is straightforward to verify that

$$0 < \lim_{\epsilon \to 0^+} \frac{P^{\epsilon}_{z^1 \to z^2}}{\epsilon^{R(z^1 \to z^2)}} < \infty,$$

with $R(z^1 \to z^2)$ as stated in the claim.

□

Note that the resistance of a particular transition is expressed as two terms. The first term in equation (13), denoted $(*)$, captures the number of synchronous deviators. The second term, denoted $(**)$, captures the log-linear response of players.

Partition the state space into the following five sets: X_1, X_2, D, E, and E^*:

- X_1 is the set of states $[a, a', x]$ such that $x_i = 1$ for at least one player i.
- X_2 is the set of states $[a, a', \mathbf{0}]$ such that $a \neq a'$.
- D is the set of states $[a, a, \mathbf{0}]$ such that a is not a pure Nash equilibrium.
- E is the set of states $[a, a, \mathbf{0}]$ such that a is a pure Nash equilibrium, but is not a maximizer of ϕ.
- E^* is the set of states $[a, a, \mathbf{0}]$ such that a is a pure Nash equilibrium and is a maximizer of ϕ.

These account for all possible states.

The only candidates for the stochastically stable states are the recurrent classes of the unperturbed process. This eliminates the possibility that states in X_1 and X_2 are stochastically stable. Therefore, from now on we restrict our attention only to trees rooted at states of D, E, and E^*. For simplicity, we use shorthand notation to write $[a, a, \mathbf{0}]$ as simply a.

Claim 6.2 *All edges $a^0 \to a^1$ have a resistance $R(a^0 \to a^1) \geq m$.*

Proof: Any transition requires at least one player to experiment which happens with at most probability on the order of ϵ^m.

□

Note that a single edge between states in D, E, or E^*, will involve transitions between multiple states. Accordingly, the expression $R(a^0 \to a^1)$ refers to the resistance of the *path* between a^0 and a^1.

Claim 6.3 *All edges $a^0 \to a^1$ in a minimum resistance tree must consist of a unilateral deviator, i.e.,* $|\{i \in \mathcal{I} : a_i^0 \neq a_i^1\}| = 1$.

Proof: We will construct resistance trees over vertices in $D \cup E \cup E^*$. Let T be a minimum resistance tree. Suppose there exists an edge $a^0 \to a^1$ where multiple players switched actions, i.e., there exists a group of players $G \subseteq \mathcal{I}$, $|G| > 1$, such that $a_i^0 \neq a_i^1$ if and only if $i \in G$. We can express the resistance of this transition as

$$R(a^0 \to a^1) = m|G| + r(a^0 \to a^1),$$

where $r(a \to a')$ captures the effect of $(**)$ in equation (13). Because of the assumed normalized potential function,

$$0 \leq r(a^0 \to a^1) \leq |G|.$$

Construct a new path $\mathcal{P}$ of unilateral deviations

$$\mathcal{P} = \{a^0 = \bar{a}^0 \to \bar{a}^1 \to ... \to \bar{a}^{|G|} = a^1\},$$

where for all $k \in \{0, 1, ..., |G| - 1\}$, $\bar{a}^{k+1} = (a_i^1, \bar{a}_{-i}^k)$ for some player $i \in G$. From (13), the resistance of the path $\mathcal{P}$ satisfies

$$m|G| \leq R(\mathcal{P}) \leq m|G| + |G|.$$

Construct a new tree T' by adding the edges of $\mathcal{P}$ to T and removing the redundant edges. The removed resistance of the edge emanating from the originating node is at least $m|G|$. Furthermore, the removed resistance from each of the remaining edges is at least m (by Claim 6.2). Therefore, the overall removed resistance is at least $m|G|+m(|G|-1)$. The added path $\mathcal{P}$ has resistance of at most $(m+1)|G|$. Accordingly, the new tree has strictly less resistance as long as $|G| \geq 2$. This contradicts the original T being a minimum resistance tree.

□

Claim 6.4 *A minimum resistance tree must be rooted at an action profile that maximizes the potential function.*

Proof: Let T be a minimum resistance tree rooted at $[a^0] \in D$. There exists a better reply path $\mathcal{P}$ from a^0 to a Nash equilibrium, a^*,

$$\mathcal{P} = \{a^0 \rightarrow a^1 \rightarrow ... \rightarrow a^k = a^*\}.$$

Since $\mathcal{P}$ is a better reply path, the resistance of the path is precisely $R(\mathcal{P}) = km$. Construct a new tree T' rooted at a^m by adding the edges of $\mathcal{P}$ to T and removing the redundant edges. This results in $R(T) > R(T')$, because the removed edge exiting a^* has resistance $R(a^* \rightarrow a') > m$, which is a contradiction. Accordingly, a minimum resistance tree cannot be rooted at a state in D.

Suppose that a minimum resistance tree, T, was rooted at a Nash equilibrium, a^0, that did not maximize the potential function, i.e., $a^0 \in E$. Let a^* be any action profile that maximizes the potential function. Since T is a rooted tree, there exists a path $\mathcal{P}$ from a^* to a^0 of the form

$$\mathcal{P} = \{a^* = a^k \rightarrow a^{k-1} \rightarrow ... \rightarrow a^0\},$$

where each edge $a^{l+1} \rightarrow a^l$ is the result of unilateral deviation. Consider the reverse path $\mathcal{P}^R$

$$\mathcal{P}^R = \{a^0 \rightarrow a^1 \rightarrow ... \rightarrow a^k = a^*\}.$$

Exploiting the similarities between payoff-based log-linear learning with unilateral updates and binary log-linear learning, we can use Lemma 5.2 to show that the difference in the total resistance across the paths is

$$R(\mathcal{P}) - R(\mathcal{P}^R) = \phi(a^0) - \phi(a^*).$$

Construct a new tree T' rooted at a^* by adding the edges of $\mathcal{P}^R$ to T and removing the redundant edges $\mathcal{P}$. The new tree will have resistance

$$\begin{aligned} R(T') &= R(T) + R(\mathcal{P}^R) - R(\mathcal{P}), \\ &= R(T) + \phi(a^0) - \phi(a^*), \\ &< R(T). \end{aligned}$$

Since the new tree T' rooted at a^* has strictly less resistance than T, T cannot be a minimum resistance tree.

□

The above analysis can be repeated to show that all action profiles that maximize the potential function have the same stochastic potential; hence, all potential function maximizers are stochastically stable. □

7 Conclusion

The central issue explored in this paper is whether the structural requirements of log-linear learning are necessary to guarantee convergence to the potential function maximizer in potential games. In particular, we have shown that the requirements of asynchrony and completeness are not necessary. The key enabler for these results is to change the focus of the analysis away from deriving the explicit form of the stationary distribution of the learning process towards characterizing the stochastically stable states. Establishing that log-linear learning constitutes a regular perturbed Markov process allows the utilization of theory of resistance trees, as in [31], to analyze the limiting properties of variations on log-linear learning.

References

[1] C. Alos-Ferrer and N. Netzer. The logit-response dynamics. *Games and Economic Behavior*, 68:413–427, 2010.

[2] 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*, 129:584–596, September 2007.

[3] A. Asadpour and A. Saberi. On the inefficiency ratio of stable equilibria in congestion games. In *Proceedings of the 5th International Workshop on Internet and Network Economics*, pages 545–552, 2009.

[4] Y Babichenko. Completely uncoupled dynamics and nash equilibria. working paper, 2010.

[5] A. Beggs. Waiting times and equilibrium selection. *Economic Theory*, 25:599–628, 2005.

[6] M. Benaim and W. H. Sandholm. Logit evolution in potential games: Reversibility, rates of convergence, large deviations, and equilibrium selection. working paper, 2007.

[7] A. Blum, E. Even-Dar, and K. Ligett. Routing without regret: On convergence to nash equilibria of regret-minimizing algorithms in routing games. In *Symposium on Principles of Distributed Computing (PODC)*, pages 45–52, 2006.

[8] L. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5:387–424, 1993.

[9] L. Blume. Population games. In B. Arthur, S. Durlauf, and D. Lane, editors, *The Economy as an evolving complex system II*, pages 425–460. Addison-Wesley, Reading, MA, 1997.

[10] L. Blume. How noise matters. *Games and Economic Behavior*, 44:251–271, 2003.

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

[12] F. Germano and G. Lugosi. Global Nash convergence of Foster and Young's regret testing. *Games and Economic Behavior*, **60**:135–154, July 2007.

[13] S. Hart and A. Mas-Colell. A reinforcement learning procedure leading to correlated equilibrium. In G. Debreu, W. Neuefeind, and W. Trockel, editors, *Economic Essays*, pages 181–200. Springer, 2001.

[14] S. Mannor and J.S. Shamma. Multi-agent learning for engineers. *Artificial Intelligence*, pages 417–422, may 2007. special issue on Foundations of Multi-Agent Learning.

[15] J. R. Marden, G. Arslan, and J. S. Shamma. Connections between cooperative control and potential games illustrated on the consensus problem. In *Proceedings of the 2007 European Control Conference (ECC '07)*, July 2007.

[16] J. R. Marden, G. Arslan, and J. S. Shamma. Regret based dynamics: Convergence in weakly acyclic games. In *Proceedings of the 2007 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Honolulu, Hawaii, May 2007.

[17] 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*, 39:1393–1407, December 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*, 54:208–220, February 2009.

[19] J. R. Marden and A. Wierman. Distributed welfare games. under submission, 2008.

[20] 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*, 48:373–396, February 2009.

[21] D. Monderer and L. Shapley. Fictitious play property for games with identical interests. *Games and Economic Theory*, 68:258–265, 1996.

[22] D. Monderer and L.S. Shapley. Potential games. *Games and Economic Behavior*, **14**:124–143, 1996.

[23] A. Montanari and A. Saberi. The spread of innovations in social networks. In *Proceedings of the National Academy of Sciences*, 2010.

[24] T. Roughgarden. *Selfish Routing and the Price of Anarchy*. MIT Press, Cambridge, MA, USA, 2005.

[25] D. Shah and J. Shin. Dynamics in congestion games. In *ACM SIGMETRICS*, 2010.

[26] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095–1100, 1953.

[27] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? forthcoming special issue in *Artificial Intelligence*, 2007.

[28] A. Vetta. Nash equilibria in competitive societies with applications to facility location, traffic routing, and auctions. In *Proc. of Symp. on Fdns. of Comp. Sci.*, pages 416–425, 2002.

[29] M. Voorneveld. Best-response potential games. *Economic Letters*, 66:289–295, 2000.

[30] D. Wolpert and K. Tumor. An overview of collective intelligence. In J. M. Bradshaw, editor, *Handbook of Agent Technology*. AAAI Press/MIT Press, 1999.

[31] H. P. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, January 1993.

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

[33] H. P. Young. *Strategic Learning and its Limits*. Oxford University Press, 2005.

[34] H. P. Young. Learning by trial and error. *Games and economic behavior*, 65:626–643, 2009.