Games and Economic Behavior 75 (2012) 882–897

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Learning efficient Nash equilibria in distributed systems

Bary S.R. Pradelski, H. Peyton Young *

Department of Economics, University of Oxford, Manor Road, Oxford OX1 3UQ, United Kingdom

ARTICLE INFO

Article history:
Received 3 September 2010
Available online 8 March 2012

JEL classification:
C72
C73

Keywords:
Stochastic stability
Completely uncoupled learning
Equilibrium selection
Distributed control

ABSTRACT

An individual's learning rule is *completely uncoupled* if it does not depend directly on the actions or payoffs of anyone else. We propose a variant of log linear learning that is completely uncoupled and that selects an efficient (welfare-maximizing) pure Nash equilibrium in all generic n-person games that possess at least one pure Nash equilibrium. In games that do not have such an equilibrium, there is a simple formula that expresses the long-run probability of the various disequilibrium states in terms of two factors: (i) the sum of payoffs over *all* agents, and (ii) the maximum payoff gain that results from a unilateral deviation by *some* agent. This *welfare/stability trade-off criterion* provides a novel framework for analyzing the selection of disequilibrium as well as equilibrium states in n-person games.

1. Learning equilibrium in complex interactive systems

Game theory has traditionally focused on situations that involve a small number of players. In these environments it makes sense to assume that players know the structure of the game and can predict the strategic behavior of their opponents. But there are many situations involving huge numbers of players where these assumptions are not particularly persuasive. Commuters in city traffic are engaged in a game because each person's choice of route affects the driving time of many other drivers, yet it is doubtful that anyone 'knows the game' or fully takes into account the strategies of the other players as is usually posited in game theory. Other examples include decentralized procedures for routing data on the internet, and the design of information sharing protocols for distributed sensors that are attempting to locate a target.

These types of games pose novel and challenging questions. Can such systems equilibrate even though agents are unaware of the strategies and behaviors of most (or perhaps all) of the other agents? What kinds of adaptive learning rules make sense in such environments? How long does it take to reach equilibrium assuming it can be reached at all? And what can be said about the welfare properties of the equilibria that result from particular learning rules?

In the last few years the study of these issues has been developing rapidly among computer scientists and distributed control theorists (Papadimitriou, 2001; Roughgarden, 2005; Mannor and Shamma, 2007; Marden and Shamma, 2008, Marden et al. 2009a, 2009b; Asadpour and Saberi, 2009; Shah and Shin, 2010). Concurrently game theorists have been investigating the question of whether decentralized rules can be devised that converge to Nash equilibrium (or correlated equilibrium) in general n-person games (Hart and Mas-Colell 2003, 2006; Foster and Young 2003, 2006; Young, 2009; Hart and Mansour, 2010). A related question is whether decentralized learning procedures can be devised that optimize some overall measure of performance or welfare without necessarily inducing equilibrium (Arieli and Babichenko, 2011; Marden et al., 2011). This is particularly relevant to problems of distributed control, where measures of system performance are given (e.g., the total

* Corresponding author. Fax: +44 1865 271094.
E-mail address: peyton.young@economics.ox.ac.uk (H.P. Young).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2012.02.017

power generated by a windfarm, the speed of data transmission in a communications network), and the aim is to design local response functions for the components that achieve maximum overall performance.

Much of the recent research on these topics has focused on potential games, which arise frequently in applications (Marden and Shamma, 2008; Marden et al., 2009a, 2009b). For this class of games there exist extremely simple and intuitively appealing learning procedures that cause the system to equilibrate from any initial conditions. A notable example is logit learning, in which an agent chooses actions with log probabilities that are a linear function of their payoffs. In this case equilibrium occurs at a local or global maximum of the potential function. However, the potential function need not measure the overall welfare of the agents, hence the equilibrium selected may be quite inefficient. This is a well-known problem in congestion games for example. The problem of inefficient equilibrium selection can be overcome by a congestion pricing scheme, but this requires some type of centralized (or at least not fully decentralized) mechanism for determining the price to charge on each route (Sandholm, 2002).

The contribution of this paper is to demonstrate a simple learning rule that incorporates log linear learning as one component, and that selects an efficient equilibrium in any game with generic payoffs that possesses at least one pure Nash equilibrium. (An equilibrium is *efficient* if there is no other equilibrium in which someone is better off and no one is worse off.) By 'select' we mean that, starting from arbitrary initial conditions, the process is in an efficient equilibrium in a high proportion of all time periods. Our learning rule is *completely uncoupled*, that is, the updating procedure does not depend on the actions or payoffs of anyone else. Thus it can be implemented even in environments where players know nothing about the game, or even whether they are in a game. All they do is react to the pattern of recent payoffs, much as in reinforcement learning (though the rule differs in certain key respects from reinforcement learning).

Our notion of selection – in equilibrium a high proportion of the time – is crucial for this result. It is not true that the process converges to equilibrium or even that it converges to equilibrium with high probability. Indeed it can be shown that, for general n-person games, there exist no completely uncoupled rules with finite memory that select a Nash equilibrium in this stronger sense (Babichenko, 2010a; see also Hart and Mas-Colell, 2003, 2006).

The learning rule that we propose has a similar architecture to the trial and error learning procedure of Young (2009), and is more distantly related to the 'learning by sampling' procedure of Foster and Young (2006) and Germano and Lugosi (2007).[1] An essential feature of these rules is that players have two different search modes: (i) deliberate experimentation, which occurs with low probability and leads to a change of strategy only if it results in a higher payoff than the current aspiration level; (ii) random search, which leads to a change of strategy that may or may not have a higher payoff. Young (2009) demonstrates a procedure of this type that selects pure Nash equilibria in games where such equilibria exist and payoffs are generic. However this approach does not provide a basis for discriminating between pure equilibria, nor does it characterize the states that are selected when such equilibria do not exist.

In contrast to these earlier papers, the learning rule described here permits a sharp characterization of the equilibrium and disequilibrium states that are favored in the long run. This results from several key features that distinguish our approach from previous ones, including Young (2009). First, we do not assume that agents invariably accept the outcome of an experiment even when it results in a strict payoff improvement: acceptance is probabilistic and is merely increasing in the size of the improvement. Second, players accept the outcome of a random search with a probability that is increasing in its realized level of payoff rather than the gain in payoff. Third, the acceptance functions are assumed to have a log linear format as in Blume (1993, 1995). These assumptions define a learning process that selects efficient pure Nash equilibria whenever pure Nash equilibria exist. Moreover when such equilibria do not exist we obtain a precise characterization of the *disequilibrium* states that have high probability in the long run. These states represent a trade-off between welfare and stability: the most likely disequilibrium states are those that maximize a linear combination of: (i) the total welfare (sum of payoffs) across *all* agents and (ii) the payoff gain that would result from a deviation by *some* agent, where the first is weighted positively and the second negatively.

2. The learning model

We shall first describe the learning rule informally in order to highlight some of its qualitative features. At any given point in time an agent may be searching in one of two ways depending on his internal state or 'mood'. In the *content* state an agent occasionally experiments with new strategies, and adopts the new strategy with a probability that increases with the associated gain in payoff. (This is the conventional exploration/exploitation form of search.) In the *discontent* state an agent flails around, trying out randomly chosen strategies every period. The search ends when he spontaneously accepts the strategy he is currently using, where the probability of acceptance is an increasing function of its realized payoff. The key differences between these modes of search are: (i) the rate of search (slow for a content agent, fast for a discontent agent); and (ii) the probability of accepting the outcome of the search. In the content state the probability of acceptance is determined by the *change* in payoff, whereas in the discontent state the probability of acceptance is determined by the *level* of payoff. The rationale for the latter assumption is that a discontent agent will typically try out many different strategies

[1] Another distant relative is the aspiration-based learning model of Karandikar et al. (1998). In this procedure each player has an endogenously generated aspiration level that is based on a smoothed average of his prior payoffs. He changes strategy with positive probability if his current payoff falls below his current aspirations. Unlike the present method, this procedure does not necessarily lead to Nash equilibrium behavior even in 2×2 games.