Games and Economic Behavior 70 (2010) 71–83

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

A payoff-based learning procedure and its application to traffic games

Roberto Cominetti[a,1], Emerson Melo[b,*], Sylvain Sorin[c,d]

[a] Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático, Universidad de Chile, Chile
[b] Departamento de Economía, Universidad de Chile and Banco Central de Chile, Chile
[c] Equipe Combinatoire et Optimisation, Faculté de Mathématiques, Université P. et M. Curie – Paris 6, 175 rue du Chevaleret, 75013 Paris, France
[d] Laboratoire d'Econométrie – Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France

ARTICLE INFO

Article history:
Received 6 July 2007
Available online 25 December 2008

Keywords:
Games
Learning
Adaptive dynamics
Stochastic algorithms
Congestion games

ABSTRACT

A stochastic process that describes a payoff-based learning procedure and the associated adaptive behavior of players in a repeated game is considered. The process is shown to converge almost surely towards a stationary state which is characterized as an equilibrium for a related game. The analysis is based on techniques borrowed from the theory of stochastic algorithms and proceeds by studying an associated continuous dynamical system which represents the evolution of the players' evaluations. An application to the case of finitely many users in a congested traffic network with parallel links is considered. Alternative descriptions for the dynamics and the corresponding rest points are discussed, including a Lagrangian representation.

1. Introduction

This paper belongs to the growing literature on the dynamics of repeated games in which players make decisions day after day based on partial information derived from prior experience. The accumulated experience is summarized in a *state variable* that determines the strategic behavior of players through a certain stationary rule. This is the framework of learning and adaptation in games, an area intensively explored in the last decades (Fudenberg and Levine, 1998; Young, 2004). Instead of studying the dynamics at an aggregate level in which the informational and strategic aspects are unspecified, we consider the question from the perspective of the individual player's strategy. At this level the most prominent adaptive procedure is *fictitious play*, early studied by Brown (1951) and Robinson (1951), which assumes that at each stage players choose a best reply to the observed empirical distribution of past moves of their opponents. A recent account of the convergence of this procedure for games with perturbed payoffs can be found in Benaim and Hirsch (1999) introducing tools that we will use here. This variant, called *smooth fictitious play*, is closely tied with Logit random choice and is analyzed in Fudenberg and Levine (1998) and Hofbauer and Sandholm (2002).

The assumption that players are able to record the past moves of their opponents is very stringent for games involving many players with limited observation capacity and bounded rationality. A milder assumption is that each player observes only the outcome vector, namely, the payoff obtained at every stage and the payoff that would have resulted if a different move had been played. Several procedures such as exponential weight (Freund and Schapire, 1999), calibration (Foster and Vohra, 1997), and no-regret procedures *à la Hannan* (Hannan, 1957; Hart, 2005), deal with such limited information contexts: players build statistics of their past performance and infer what the outcome would have been if a different strategy had been played. Eventually, adaptation leads to configurations where no player regrets the choices he makes.

[*] Corresponding author at: California Institute of Technology, Division of the Humanities and Social Sciences, MC 228-77, Pasadena, CA, USA.
E-mail addresses: rcominet@dim.uchile.cl (R. Cominetti), emelos@hss.caltech.ed (E. Melo), sorin@math.jussieu.fr (S. Sorin).

[1] Supported by FONDAP grant in Applied Mathematics, CONICYT-Chile.

0899-8256/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.geb.2008.11.012

Although these procedures are flexible and robust, the underlying rationality may still be considered as too demanding in the context of games where players are boundedly rational and less informed. This is the case for traffic in congested networks where a multitude of small players make routing decisions with little or no information about the strategies of the other drivers nor on the actual congestion in the network. The situation is often described as a game where traffic equilibrium is seen as a sort of steady state that emerges from an underlying adaptive mechanism. Wardrop (1952) considered a non-atomic framework ignoring individual drivers and using continuous variables to represent aggregate flows, while Rosenthal (1973) studied the case in which drivers are taken as individual players. Traffic has also been described using random utility models for route choice, leading to the notion of *stochastic user equilibrium* (Daganzo and Sheffi, 1977; Dial, 1971). All these models assume implicitly the existence of a hidden mechanism in travel behavior that leads to equilibrium. Empirical support for this has been given in Avineri and Prashker (2006), Horowitz (1984), Selten et al. (2007) based on laboratory experiments and simulations of different discrete time adaptive dynamics, though it has been observed that the steady states attained may differ from all the standard equilibria and may even depend on the initial conditions. Additional empirical evidence to support the use of discrete choice models in the context of games is presented in McKelvey and Palfrey (1995). From an analytical point of view, the convergence of a class of finite-lag adjustment procedures was established in Cantarella and Cascetta (1995), Cascetta (1989), Davis and Nihan (1993). On a different direction, several continuous time dynamics describing plausible adaptive mechanisms that converge to Wardrop equilibrium were studied in Friesz et al. (1994), Sandholm (2002), Smith (1984), though these models are of an aggregate nature and are not directly linked to the behavior of individual players.

A simpler idea is considered in this paper. We assume that each player has a prior perception or estimate of the payoff performance for each possible move and makes a decision based on this rough information using a random choice rule such as Logit. The payoff of the chosen alternative is then observed and is used to update the perception for that particular move. This procedure is repeated day after day, generating a discrete time stochastic process which we call the *learning process*. The basic ingredients are therefore: a state parameter; a decision rule from states to actions; an updating rule on the state space. This structure is common to many procedures in which the incremental information leads to a change in a state parameter that determines the current behavior through a given stationary map. The specificity here is that the updating rule depends uniquely on the realized payoffs: although players observe only their own payoffs, these values are affected by everybody else's choices revealing information on the game as a whole. The question is whether a simple learning mechanism based on such a minimal piece of information may be sufficient to induce coordination and make the system stabilize to an equilibrium.

It is worth noting that several learning procedures, initially conceived for the case when the complete outcome vector is available, have been adapted to deal with the case where only the realized payoff is known: see Fudenberg and Levine (1998, §4.8) for smooth fictitious play, Auer et al. (2002) for exponential weight, Foster and Vohra (1998) for calibration, and Hart and Mas-Collel (2001) for non-regret. The idea of this kind of approaches is to use the observed payoffs to build an unbiased estimator of the outcome vector, to which the initial version of the procedure is applied. More explicitly, a *pseudo-outcome* vector is defined by the observed payoff divided by the probability with which the actual move was played, on the component corresponding to that move, and completed by zeroes on the other components. Alternatively, the pseudo-outcome vector is built as the empirical average of payoffs obtained on random *exploration stages* having a positive density. The resulting update rules depend not only on the observed payoffs, but also on the probability according to which a move was played as well as on the nature of the stage (exploitation or exploration).

Our process is much simpler in that it relies only on the past sequence of realized moves and payoffs. The idea is closer to the so-called *reinforcement dynamics* in which the only information of a player is her daily payoff (Arthur, 1993; Beggs, 2005; Borgers and Sarin, 1997; Erev and Roth, 1998; Laslier et al., 2001; Posch, 1997), though it differs in the way the state variable is updated as well as in the choice of the decision rule as a function of the state. Usually, reinforcement models use a cumulative rule on a *propensity vector* in which the current payoff is added to the component played, while the remaining components are kept unchanged. A stage-by-stage normalization of the propensity vector leads to a mechanism which is related to the replicator dynamics (Posch, 1997). In this context, convergence has been established for the case of an i.i.d. environment, for zero-sum games, and also for some games with unique equilibria (Laslier et al., 2001; Beggs, 2005). We should also mention here the mechanism proposed in Borgers and Sarin (1997) which uses an averaging rule with payoff dependent weights. Our updating rule uses instead a time average criteria that induces a specific dynamics on perceptions and strategies which appears to be structurally different from the previously studied ones, while preserving the qualitative features of *probabilistic choice* and *sluggish adaptation* (Young, 2004, §2.1).

The paper is organized as follows. Section 2 describes the learning process in the general setting of repeated games, providing sufficient conditions for this process to converge almost surely towards a stationary state which is characterized as an equilibrium for a related game. The analysis relies on techniques borrowed from stochastic algorithms (see e.g. Benaim, 1999; Kushner and Yin, 1997), and proceeds by studying an associated continuous deterministic dynamical system which we call the *adaptive dynamics*. Under suitable assumptions the latter has a unique rest point which is a global attractor, from which the convergence of the learning process follows. In Section 3 we apply the general convergence result to a simple traffic game on a network with parallel links. In this restricted setting the convergence results are established in terms of a "*viscosity parameter*" which represents the amount of noise in players' choices, namely, if noise is large enough the learning process and the associated adaptive dynamics have a unique global attractor. Besides, we obtain a potential function that yields an equivalent Lagrangian description of the dynamics together with alternative characterizations of the rest points.