Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Economic Theory 122 (2005) 1–36

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

On the convergence of reinforcement learning

A.W. Beggs

Wadham College, Oxford, OX1 3PN, UK

Received 8 April 2002; final version received 30 March 2004
Available online 15 September 2004

Abstract

This paper examines the convergence of payoffs and strategies in Erev and Roth's model of reinforcement learning. When all players use this rule it eliminates iteratively dominated strategies and in two-person constant-sum games average payoffs converge to the value of the game. Strategies converge in constant-sum games with unique equilibria if they are pure or if they are mixed and the game is 2 × 2. The long-run behaviour of the learning rule is governed by equations related to Maynard Smith's version of the replicator dynamic. Properties of the learning rule against general opponents are also studied.

JEL classification: C72; D83

Keywords: Reinforcement learning; Games

1. Introduction

This paper studies the convergence properties of a class of naïve reinforcement learning models in games. These were originally proposed by Roth and Erev [37] and Erev and Roth [14] as a means of modelling the observed behaviour of subjects in experiments on games. They argue that their behaviour can be well approximated by a simple model in which players tend to put more weight on strategies that have enjoyed past success, as measured by the cumulated payoffs they have achieved. Harley [20] proposed a similar model in a biological context.

E-mail address: alan.beggs@economics.ox.ac.uk.

0022-0531/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2004.03.008

This model has considerable attraction as a simple model of boundedly rational players. The amount of information players are assumed to gather is small. Players need only observe their realised payoffs and may not be aware they are even playing a game, let alone the payoff matrix of their opponent or even their actions. It also builds in a certain amount of inertia, in that players are slow to switch from actions that have performed well in the past, which seems a plausible feature of learning. Despite this, little is known about the analytical properties of the model.

This paper aims to reduce this gap. It studies the behaviour of players' payoffs and strategies when other players use the same rule and when they do not.

It shows that when all players use this rule dominated strategies are iteratively deleted. In addition in two-person constant-sum games, players' average payoffs converge to the value of the game. It also shows that their strategies converge to equilibrium in a constant-sum game if it has a unique pure strategy equilibrium or it has a unique-mixed strategy equilibrium and is 2×2.

In the course of the analysis it is shown that the long-run behaviour of the players' strategies is governed by equations related to Maynard Smith's [33] version of the replicator dynamic. This may be of independent interest since in 2×2 constant-sum games mixed equilibria are stable under it, while the ordinary replicator dynamic cycles around it. There have, however, been few derivations of the Maynard Smith dynamic from primitive assumptions [6,25], which justify versions from models of imitation, seem the only examples. The current dynamic is similar to the Maynard Smith dynamic and shares its convergence properties.

When opponents do not use the Erev and Roth rule, it is shown that a player using it learns not to play dominated strategies. It is also shown that a player's long-run average payoff cannot be forced permanently below his minmax payoff. That is if a player uses the ER rule, the lim sup of his average payoffs will be at least this. More generally it is shown that his long-run average payoff cannot be forced below any payoff he can guarantee himself on average by playing a fixed action. A clever player can, however, exploit the inertia in the Erev and Roth scheme and force the lim inf of the player's average payoffs below his minmax value, so that his average payoff is below this infinitely often.

There are of course many other models of learning. Much attention recently has focussed on fictitious play and variations of it. For example Benaïm and Hirsch [4] study convergence of strategies in games with randomly perturbed payoffs. Although fictitious play itself has poor optimality properties, smoothed versions have quite good properties—stronger than those mentioned above for the procedure studied here. Fudenberg and Levine [16] provide a good summary of this work. Hart and Mas-Collel [21–23] study procedures based on 'regrets', which in some cases share these properties. On the other hand, fictitious play and regret-based strategies require greater knowledge of the game and sophistication. Auer et al. [2] and Hart and Mas-Collel [22] study versions which do not require knowledge of the game, but still require some sophistication. The feature of inertia which these procedures lack, but is shared by Hart and Mas-Collel [22] where it also can be exploited by clever players, also seems an appealing feature of a model of learning.

In any case it is not argued that this is the only plausible model of learning, only that it is of enough interest to be make it worth further study. Camerer and Ho [9] suggest that both it and fictitious play have features which match the data and present a synthesis.