Neurocomputing 127 (2014) 258–265

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# An actor–critic algorithm for multi-agent learning in queue-based stochastic games

D. Krishna Sundar [a], K. Ravikumar [b,*]

[a] Indian Institute of Management Bangalore, Bangalore-560076, India
[b] D-103, Adarsh Palm Retreat, Outer Ring Road, Bangalore-560103, India

ARTICLE INFO

*Article history:*
Received 12 December 2012
Received in revised form
30 May 2013
Accepted 29 July 2013
Communicated by N.T. Nguyen
Available online 11 September 2013

*Keywords:*
Service markets
Queues
Dynamic pricing
Stochastic games
Learning in games
Reinforcement learning

ABSTRACT

We consider state-dependent pricing in a two-player service market stochastic game where state of the game and its transition dynamics are modeled using a semi-Markovian queue. We propose a multi-time scale actor–critic based reinforcement algorithm for multi-agent learning under self-play and provide experimental results on Nash convergence.



## 1. Introduction

Queues emerge in markets in two distinct settings – in the stochastic setting when demand and supply are subject to random variations, and dynamic adjustment of capacity to cope with these variations is costly, and in the nonstochastic setting when the price is set above or below the market clearing price due to costs or constraints on prices [11]. In the former case, queues create *negative network externalities*: the utility to each consumer decreases with an increase in the total number of customers in the queue. The full price to the customer is the money price plus the value of his/her time. Naor [16] studied a system in which potential customers can observe the state of the queue before joining. Customers will join if queuelength observed is below a critical size, depending on the value of service to them and upon their cost of waiting. It was shown that the critical queue size for welfare maximization should in general be lower than the *self-optimizing* equilibrium size and admission therefore be restricted. In situations, where it is not feasible or is costly to inspect queue size at arrival instants, it is more practical and efficient to impose uniform fee on all customers regardless of the state of the system. That is, instead of controlling the rate of admission, the arrival rate itself is reduced. For studies on analysis of pricing in queues, refer [15,17].

In a commodity market, competition in prices drives them downward resulting in zero profits [20]. In contrast, in queue-based service markets, the downward pressures on prices is countered to some extent by the congestion or *dis-utilities* consumers incur from waiting costs. Competing parties might choose to differentiate themselves by offering different prices, and thus different qualities of service in terms of congestion. This type of *queue-driven discrimination* can be viewed as a special sort of third degree price discrimination [21] that is widely practiced in commodity markets – different discounts to different age groups, for example. In a simple two-firm scenario, one firm will offer a higher price than the other, appealing only to the most congestion sensitive customers while the majority of customers will use the service of the other, less expensive firm. In this paper, we build a model of service market with two players (or firms) and make the above intuition precise by developing a multi-agent learning game model for dynamic price setting, and experimentally demonstrate convergent behavior of rational learners.

Nash equilibrium is a natural equilibrium concept in competitive games and is a point in the joint policy space where no agent has incentive to deviate unilaterally. If all agents of the game follow the same *rational learning* algorithm (self-play) that tries to learn best response to opponents' actions, and if it *converges*, then the agents will be *locked in* a Nash equilibrium. However, it is now well established that *simultaneous* disequilibrium strategy adjustments

* Corresponding author. Tel.: +91 997 258 9102.
*E-mail addresses:* diatha@iimb.ernet.in (D. Krishna Sundar), rkkaruman@gmail.com (K. Ravikumar).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.07.020

by competing agents may not converge (to an equilibrium) in general. In view of the fact, we mildly perturb the simultaneous adjustment condition and allow players to run at different time scales while updating their beliefs or strategies. This feature coupled with the assumption that state of the game is known to all the players led us to experimentally demonstrate convergent dynamics. To be precise, we propose an actor–critic-type of reinforcement learning scheme; that is, we model the two players as two actor–critic learners, and the actors (policies) update their *strategies* on different time scales. We propose this scheme with the intuition that if two actors run on different time scales, the slower player sees the other player as "equilibrated" and the faster player sees the other player as quasi-static and hence, both the learners might converge (to a Nash equilibrium). Though no claims can be made on convergence of the algorithm in general, our experimental analysis can be treated as the first step in that direction.

### 1.1. Related literature

Reinforcement learning as a paradigm for multi-agent learning in stochastic games has been studied by Littman [12] in zero-sum games and Patek and Bertsekas [19] in zero-sum stochastic – path games using minmax-Q learning that is shown to converge. Nash-Q learning for general-sum games of Hu and Wellman [8] imposes many restrictive assumptions for convergence whereas more general and convergent Friend or Foe Q-learning of Littman [13] requires information with regard to opponent: friend or foe and uses Nash-Q or minimax-Q accordingly. Busoniu et al. [6] gives a comprehensive survey of reinforcement learning in games.

Even in the iterated game cases, no algorithms with guaranteed convergence are known to exist. In the complete information iterated two-action bi-matrix games, Singh et al. [18], develop a gradient ascent algorithm with constant steps and show that either the agents converge to a Nash equilibrium or their average pay-offs will converge to the pay-offs corresponding to a Nash equilibrium. Bowling and Veloso [5] modify the above algorithm to include learning rates that vary with time depending on whether a player is Winning or Losing (WoLF) and show convergence with variable learning rates. However, in this scheme, each player is expected to know about Nash equilibirum policy and pay-offs. In [1], a Weighted Policy Learner scheme is proposed which modifies the WoLF algorithm to weigh the learning rates in proportional to the current probability of choosing actions. The procedure is shown to converge to mixed strategy equilibria in some benchmark iterated general-sum games.

More recently, Akchurina [2] reported an algorithm for general-sum stochastic games where evolution of the game is modeled through differential equations, and under certain assumptions on the initial condition, convergence to a $\varepsilon$-Nash equilibrium is established. Young [22] tries to identify the boundary between the possible and impossible in multi-agent learning. The boundary between the possible and the impossible depends on differences in assumptions about the amount of information that agents have, the extent to which they optimize, the desired form of convergence.

#### 1.1.1. Contributions of the present work

In this paper, we present a two-player stochastic game whose transition dynamics is governed by semi-Markovian queues associated with the players. Subsequently, we develop a reinforcement learning algorithm for agent learning and numerically establish its convergent behavior in many example scenarios. The model and the associated learning dynamics presented here is an attempt to comprehend dynamic pricing behavior in service markets or markets with congestible resources, in general. To the best of our knowledge, we are not aware of any work that models and analyzes learning dynamics in stochastic games to the level of generality contained herein. Further, our semi-Markovian game model allows for modeling transition intervals between states of the game, which in fact affects the pay-offs received by the players.

As for reinforcement learning, our proposed algorithm differs from the works cited above in the following ways: first, a vast majority of the algorithms follow the philosophy of value iteration scheme of Markov decision processes (or more generally, Markovian games). At every step of learning such schemes involve solving Linear Program (in the case of Zero sum games or Foe learning) or a quadratic program (in Nash Q learning) to identify the policy for next step of learning. In Nash-Q learning one needs to maintain estimates of Q-values of the opponent. Since Q-learning can only learn deterministic policies [23], learning mixed strategy equilibria is not possible through such Q-learning based schemes. In this paper we give an actor–critic type of learner (Barto et al. [3] and Konda and Borkar [9]), a derivative of policy-iteration scheme, that maintains values as well as policy and updates these in a coupled fashion but on different time scales. Further, it does not entail maintaining estimates of opponent's pay-offs as in Nash-Q learning. The actor–critic algorithm described here is along the lines of Borkar [4] but with a difference. As opposed to actors described therein that operate on same time scale resulting in convergence to $\varepsilon$-Nash equilibria. To obviate this difficulty, we allow the actors in our algorithm to operate on different time scales. This models a situation in any competitive game where players differ in their information acquisition capabilities. The work by Leslie and Collins [10] which reports a multi-time scale reinforcement learning algorithm is closest in spirit to our work. However, their algorithm builds on Q-value updates as well as policy updates to learn mixed strategy equilibria. Maintaining Q-values increases the dimensionality of the problem, and requires through exploration for convergence. In contrast, in our approach we address the scale problem by employing policy iteration scheme. Further, in [10] experimental results on convergence are reported only in the case of normal form games. Majority of the existing works report convergence only in iterated games whereas in this paper, we report convergence results in a very general stochastic game.

To validate convergence of the algorithm to Nash equilibria in general sum games, we consider a few iterated game cases discussed in the literature. In particular, we consider a fairly complex six-action general-sum bi-matrix game (an example from Mangasarian and Stone [14]) with no apparent special structure. The algorithm converges exactly to the unique Nash equilibrium mentioned therein. We also provide results on a constant-sum game and an iterated bi-matrix game example case presented in Bowling and Veloso [5] (that exposes some difficulties involved with expected pay-off based rules in identifying "winning" position in their WoLF algorithm).

The rest of the paper is organized as follows. In Section 2, we develop a semi-Markov game model pertinent to queues. In Section 3 we present our multi-time scale actor–critic algorithm interspersing intuition at various places to motivate the algorithm. Section 4 gives results of our computational experiments.

## 2. Queueing model

We consider a simple model of a service market with two service providers shown in Fig. 1. Buyers approaching the market are classified into two categories: A Type 1 buyer randomly chooses a service provider and receives a quote on price to render requested service and the expected delay to initiate processing his request. In contrast, Type 2 buyers (comparison shoppers) search the market to learn *posted price* quote and also the *posted expected*