# Learning Automata with Artificial Reflecting Barriers in Games with Limited Information

**Ismail Hassan**
Oslo Metropolitan University
Oslo, Norway
ismail@oslomet.no

**B. John Oommen**
Carleton University
Ottawa, Canada
oommen@scs.carleton.ca

**Anis Yazidi**
Oslo Metropolitan University
Oslo, Norway
anisy@oslomet.no

**Abstract**

This paper deals with the problem of solving stochastic games (which have numerous business and economic applications), using the interesting tools of Learning Automata (LA), the precursors to Reinforcement Learning (RL). Classical LA systems that possess properties of absorbing barriers, have been used as powerful tools in game theory to devise solutions that converge to the game's Nash equilibrium under *limited information* (Sastry, Phansalkar, and Thathachar 1994). Games with limited information are intrinsically hard because the player does not know the actions chosen of other players, neither their outcomes. The player might not be even aware of the fact that he/she is playing against an opponent.

With the state-of-the-art, the numerous works in LA applicable for solving game theoretical problems, can merely solve the case where the game possesses a saddle point in a pure strategy.

They are unable to reach mixed Nash equilibria when a saddle point is non-existent in pure strategies. Additionally, within the field of LA and RL in general, the theoretical and applied schemes of LA with artificial barriers are scarce, even though incorporating artificial barriers in LA has served as a powerful and yet under-explored concept, since its inception in the 1980's. More recently, the phenomenon of introducing artificial non-absorbing barriers was pioneered, and this renders the LA schemes to be resilient to being trapped in absorbing barriers. In this paper, we devise a LA with artificial barriers for solving a general form of stochastic bimatrix games. The problem's complexity has been augmented with the scenario that we consider games in which there is no saddle point in pure strategies. By resorting to the above-mentioned powerful concept of artificial reflecting barriers, we propose a LA that converges to an optimal mixed Nash equilibrium even though there may be no saddle point when a pure strategy is invoked.

## 1 Introduction

Narendra and Thathachar first presented the term Learning Automata (LA) in their 1974 survey paper (Narendra and Thathachar 1974; 2012). LA consists of an adaptive learning agent interacting with a stochastic Environment with incomplete information. Lacking prior knowledge, LA attempt to determine the optimal action to take by updating the action probabilities based on the Reward/Penalty input that the LA receives from the Environment. This process is repeated until the optimal action is, finally, hopefully achieved.

Research into LA over the past four decades is extensive, leading to the proposal of various types, throughout the years. LA are mainly characterized as being Fixed Structure Learning Automata (FSLA) or Variable Structure Learning Automata (VSLA). In FSLA, the probability of the transition from one state to another state is fixed, and the action probability of any action in any state is also fixed. Early research into LA centered around FSLA, and pioneers such as Tsetlin, Krylov, and Krinsky (Tsetlin and others 1973) proposed several examples of these. The research into LA moved gradually towards VSLA. Introduced by Varshavskii and Vorontsova in the early 1960's (Varshavskii and Vorontsova 1963), VSLA have transition and output functions that evolve as the learning process continues (Narendra and Thathachar 1974; 2012; John oommen 1986). The state transitions or the action probabilities are updated at every time step.

### 1.1 LA in Game Theory

Since the primary focus of this paper is on using LA to resolve games, it is prudent for us to give a brief overview of the results currently available. Studies on strategic games with LA have primarily focused on the traditional $L_{R-I}$ scheme, which is desirable because it can yield the Nash equilibrium in pure strategies (Sastry, Phansalkar, and Thathachar 1994) [1]. For instance, Vrancx et al. extend the latter work of Sastry to multi-state games while establishing the link between LA and replicator dynamics. However, seminal studies game theoretical LA such as (Bloembergen et al. 2015; Sastry, Phansalkar, and Thathachar 1994; Vrancx, Tuyls, and Westra 2008) only studied the traditional absorbing $L_{R-I}$ scheme.



[1] The abbreviations, $L_{R-I}$, as the Linear Reward-Inaction scheme, is a fundamental and popular machine, where the corresponding probability is increased linearly on receiving a "Reward", but is unchanged ("Inaction") upon receiving a penalty.

Although other ergodic schemes such as the $L_{R-P}$[2] were used in games with limited information (Viswanathan and Narendra 1974), they did not gain popularity, due to their inability to converge to the Nash equilibrium.

LA have found numerous applications in domains which can be perceived as being game-theoretic. These include sensor fusion without the knowledge of the ground truth (Yazidi et al. 2020), distributed power control in wireless networks and more particularly, NOMA (Rauniyar et al. 2020), optimization of cooperative tasks (Zhang, Wang, and Gao 2020), content placement in cooperative caching (Yang et al. 2020), congestion control in the Internet of Things (Gheisari and Tahavori 2019), QoS satisfaction in autonomous mobile edge computing (Apostolopoulos, Tsiropoulou, and Papavassiliou 2018), opportunistic spectrum access (Cao and Cai 2018), scheduling domestic shiftable loads in smart grids (Thapa et al. 2017), anti-jamming channel selection algorithm for interference mitigation (Jia et al. 2017), and relay selection in vehicular ad-hoc networks (Tian et al. 2017) etc.

We conclude this subsection by observing that all of these papers utilize what we shall refer to as the well-acclaimed families of LA (continuous, discretized, Pursuit etc.). However, none of them have ventured into the unexplored waters of considering LA which possess artificial *reflecting* barriers, as we have done in this paper.

### 1.2 Objectives and Contributions of this Paper

In this paper, we propose an algorithm addressing bimatrix games which is a more general version of the zero-sum game treated in (Lakshmivarahan and Narendra 1982; Yazidi, Silvestre, and Oommen 2021). We consider a stochastic game where the outcomes are either a Reward or a Penalty. The Reward probabilities are given by the corresponding payoff matrix of each player, and unknown to the LA. The games we work with possess *limited information*, where each Player only observes the outcome of his action in the form of a Reward or Penalty, without observing the action chosen by the other Player. The Player might not be even aware that he is playing against an opponent.

The novel contributions of this paper are the following:

1. While the numerous works in LA-based solutions for game-theoretic problems merely solve the case where the game possesses a saddle point in a pure strategy, they are unable to reach mixed Nash equilibria when there is no saddle point in pure strategies. This is our primary contribution.
2. In the new scheme, the artificial *non-absorbing* barriers render the schemes to be resilient to being trapped in absorbing barriers. However, by resorting to the above-mentioned phenomenon of artificial reflecting barriers, the LA converges to an optimal mixed Nash equilibrium even though there may be no saddle point when a pure strategy is invoked.
3. By virtue of the above, the well-known legacy $L_{R-I}$ scheme can be seen to be an instance of our proposed algorithm for a particular choice of the barrier.
4. All of the above results have been formally proven and experimentally verified.

All of the results mentioned above are novel, and to the best our knowledge, pioneering.

### 1.3 Organization of this Paper

The remainder of this article is organized as follows. In Section 2, we present the game model for both $P-$type and $S-$type Environments. In Section 3, we introduce our devised $L_{R-I}$ with artificial barriers for handling $P-$type Environments. The experimental results related to the $L_{R-I}$ are presented in Section 4. Section 5 concludes the paper.

## 2 The Game Model

In this section, we formalize the game model that is being investigated. Let $P(t) = [p_1(t) \quad p_2(t)]^\intercal$ denote the mixed strategy of Player $A$ at time instant $t$, where $p_1(t)$ accounts for the probability of adopting strategy 1 and, conversely, $p_2(t)$ stands for the probability of adopting strategy 2. Thus, $P(t)$ describes the distribution over the strategies of Player $A$. Similarly, we can define the mixed strategy of Player $B$ at time $t$ as $Q(t) = [q_1(t) \quad q_2(t)]^\intercal$. The extension to more than two actions per Player is straightforward following the method analogous to what was used by Papavassilopoulos (Papavassilopoulos 1989), which extended the work of Lakshmivarahan and Narendra (Lakshmivarahan and Narendra 1982).

Let $\alpha_A(t) \in \{1, 2\}$ be the action chosen by Player $A$ at time instant $t$ and $\alpha_B(t) \in \{1, 2\}$ be the one chosen by Player $B$, following the probability distributions $P(t)$ and $Q(t)$, respectively. The pair $(\alpha_A(t), \alpha_B(t))$ constitutes the joint action at time $t$, and are pure strategies. Specifically, if $(\alpha_A(t), \alpha_B(t)) = (i, j)$, the probability of Reward for Player $A$ is determined by $r_{ij}$ while that of Player $B$ is determined by $c_{ij}$. With regard to notation, in this case, Player $A$ is the row Player, while Player $B$ is the column Player.

When we are operating in the $P$-type mode, the game is defined by two payoff matrices, $R$ and $C$, describing the Reward probabilities of Players $A$ and $B$ respectively, where:

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}, \text{and} \quad (1) \quad C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}, \qquad (2)$$

where the entries of both matrices are probabilities.

In the case where the Environment is a $S$-model type, the latter two matrices are deterministic and describe the feedback as a scalar in the interval $[0, 1]$. For instance, if we operate in the $S$-type Environment, the feedback when both Players choose their respective first actions will be the scalar $c_{11}$ for Player $A$ and not a Bernoulli-distributed feedback as in the case of a $P$-type Environment. It is possible to also consider the $c_{ij}$'s as stochastic continuous variables whose means are $c_{ij}$, and with specific instantaneous realizations. However, for the sake of simplicity we consider $C$ and $R$ as being deterministic. The asymptotic convergence proofs for the $S-$type Environment are valid independent

[2]The abbreviations, $L_{R-P}$, as the Linear Reward-Penalty scheme, is a fundamental and popular machine, where the corresponding probability is increased linearly on receiving a "Reward", and decreasing linearly upon receiving a "Penalty".

of whether $C$ and $R$ are deterministic, or whether they are obtained from a distribution whose supports are in the interval $[0, 1]$, and whose means are defined by the matrices. The extension of our scheme to $S$-model type has been omitted from this paper due to number of page limitation.

We now resort to some fundamental principles of Strategic Game Theory. Indeed, independent of the type of the Environment (i.e., whether it is of a $P-$type or $S-$type), we can distinguish three cases when it concerns their equilibria:

- Case 1:

  If $(r_{11} - r_{21})(r_{12} - r_{22}) < 0$, $(c_{11} - c_{12})(c_{21} - c_{22}) < 0$ and $(r_{11} - r_{21})(c_{11} - c_{12}) < 0$, there is a single mixed equilibrium.

- Case 2:

  If $(r_{11} - r_{21})(r_{12} - r_{22}) > 0$ or $(c_{11} - c_{12})(c_{21} - c_{22}) > 0$, there is only a single pure equilibrium since there is at least a single Player who has a dominant strategy.

- Case 3:

  If $(r_{11} - r_{21})(r_{12} - r_{22}) < 0$, $(c_{11} - c_{12})(c_{21} - c_{22}) < 0$ and $(r_{11} - r_{21})(c_{11} - c_{12}) > 0$, there are two pure equilibria and one mixed equilibrium.

In strategic games, Nash equilibria are equivalently referred to as the game's "saddle points". Since the outcome for a given joint action is stochastic, the game is of stochastic genre.

# 3 Game Theoretical LA Scheme based on the $L_{R-I}$ with Artificial Barriers

In this section, we shall present our $L_{R-I}$ scheme that possesses artificial reflecting barriers, specifically devised for $P$-type Environments.

## 3.1 Non-Absorbing Artificial Barriers

What is unknown in the field of LA is a scheme which is originally absorbing, but which can be rendered ergodic. In many cases, this can be achieved by making the scheme behave according to the absorbing scheme's rule in the interior of the probability simplex, and by then forcing the probability back inside the simplex whenever the scheme approaches an absorbing barrier. In other words, the absorbing barrier is rendered to be "artificially reflecting". Such a scheme is novel in the field of LA and its advantage is that it avoids being absorbed in non-desirable absorbing barriers.

Interestingly, and apart from the above, by countering the absorbing barriers, the scheme can migrate stochastically towards a desirable *mixed* strategy. Also, as we will see later in the paper, even if the optimal strategy corresponds to an absorbing barrier, the scheme will approach it. Thus, the scheme converges to mixed strategies whenever they correspond to optimal strategies while approaching the absorbing states whenever they are the optimal strategies. Our newly-devised scheme that boasts these properties, is explained in the next section.

## 3.2 Non-Absorbing Game Playing

We now present the design of our proposed LA scheme together with some theoretical results demonstrating that it can converge to the saddle points of the game even if the saddle point is a *mixed* Nash equilibrium. Our solution represents a new variant of the $L_{R-I}$ scheme, which is made ergodic by modifying the update rule in a general form.

We introduce a quantity, $p_{max}$, which represents the highest values that any probability can take, as the artificial barrier. This is a real value close to 1. Similarly, we introduce a quantity, $p_{min} = 1 - p_{max}$ which corresponds to the lowest value any action probability can take. In order to enforce the constraint that the probability of any action for both Players remains within the interval $[p_{min}, p_{max}]$ one should start by choosing initial values of $p_1(0)$ and $q_1(0)$ in the same interval, and further resorting to strategically designed update rules that ensure that the modifications at each iteration, keep the probabilities *within* the same interval.

If the outcome from the Environment is a Reward at a time $t$ for action $i \in \{1, 2\}$, the update rule is given by:

$$\begin{aligned} p_i(t+1) &= p_i(t) + \theta(p_{max} - p_i(t)) \\ p_s(t+1) &= p_s(t) + \theta(p_{min} - p_s(t)) \quad \text{for} \quad s \neq i. \end{aligned} \tag{3}$$

where $\theta$ is a learning parameter. Following the Inaction Principle of the $L_{R-I}$, whenever the Player receives a Penalty, its action probabilities are kept unchanged, and thus:

$$\begin{aligned} p_i(t+1) &= p_i(t) \\ p_s(t+1) &= p_s(t) \quad \text{for} \quad s \neq i. \end{aligned} \tag{4}$$

The update rules for the mixed strategy $q(t+1)$ are defined in a similar fashion. We shall now move to the theoretical analysis of the convergence properties of our proposed algorithm for solving a strategic game. To do this, we assume that the optimal Nash equilibrium of the game, which is unknown to the LA, is the pair $(p_{\text{opt}}, q_{\text{opt}})$.

## 3.3 Proof Methodology

Before we proceed with a sketch of the formal proof, we shall describe the steps involved in the theoretical analysis. The proof invokes Norman's classical results (Norman 1972), which have been used to prove various convergence results for LA. However, the use of these results in LA-based game theory has been, to date, limited.

The first involves the *existence* and *location* of the positions where the Pure/Mixed equilibria could converge to. This is done in Theorem 1. It then involves a demonstration that the trajectory of the updating scheme does, indeed, move towards these equilibria. This is achieved by considering the "gradients", or rather the matrix of *derivatives* of the corresponding game matrices.

We shall first distinguish the details of the equilibrium condition, according to the entries in the payoff matrices $R$ and $C$ for Case 1, given below.

**Case 1: Only One Mixed Nash Equilibrium Case** The first case depicts the situation where there is no saddle point in pure strategies. In other words, the only Nash equilibrium

is a mixed one. Based on the fundamentals of Game Theory, the location of the optimal mixed strategy in the probability space can be shown to be the following:

$$p_{\text{opt}} = \frac{c_{22} - c_{21}}{L'}, \quad q_{\text{opt}} = \frac{r_{22} - r_{12}}{L},$$

where:

$L = (r_{11} + r_{22}) - (r_{12} + r_{21})$, and

$L' = (c_{11} + c_{22}) - (c_{12} + c_{21})$.

This can be further sub-divided into two sub-cases:

$$r_{11} > r_{21}, r_{12} < r_{22}; c_{11} < c_{12}, c_{21} > c_{22}, \text{and} \tag{5}$$

$$r_{11} < r_{21}, r_{12} > r_{22}; c_{11} > c_{12}, c_{21} < c_{22}. \tag{6}$$

Let the vector $X(t) = [p_1(t) \quad q_1(t)]^\intercal$. We resort to the notation $\Delta X(t) = X(t+1) - X(t)$, and for denoting the conditional expected value operator, we use the notation $\mathbb{E}[\cdot|\cdot]$. Using these notations, we prove the following theorem.

**Theorem 1.** *Consider a two-Player game with payoff matrices as in Eq.* (1) *and Eq.* (2)*, and a learning algorithm defined by Eq.* (3) *and Eq.* (4) *for both Players $A$ and $B$, with learning rate $\theta$. Then, $E[\Delta X(t)|X(t)] = \theta W(x)$ and for every $\epsilon > 0$, there exists a unique stationary point $X^* = [p_1^* \quad q_1^*]^\intercal$ satisfying:*

1. $W(X^*) = 0$;
2. $|X^* - X_{\text{opt}}| < \epsilon$.

The proofs of the theorems reported in this paper as well as some other additional theoretical results are omitted here due to space limitations and will be published in an extended version of the current article. A preprint of the extended version (Hassan, Oommen, and Yazidi 2022) is available on *arXiv*.

The next theorem states that the expected value of $\Delta X(t)$ has a negative definite gradient.

**Theorem 2.** *The matrix of partial derivatives, $\frac{\partial W(X^*)}{\partial x}$ is negative definite.*

**Theorem 3.** *We consider the update equations given by the $L_{R-I}$ scheme. For a sufficiently small $p_{min}$ approaching 0, and as $\theta \to 0$ and as time goes to infinity:*

$[E(p_1(t)) \quad E(q_1(t))] \to \left[p_{\text{opt}}^* \quad q_{\text{opt}}^*\right]$,

*where $\left[p_{\text{opt}}^* \quad q_{\text{opt}}^*\right]$ is the game's Nash equilibrium.*

## 4 Experimental results

To verify the theoretical properties of the proposed learning algorithm, we conducted several simulations. By using different instances of the payoff matrices $R$ and $C$, we can experimentally cover the three cases referred to in Section 3. Although the experiments were done for numerous game settings, in the interest of being concise and space, we report here only a few of these results for case 1 where the only Nash equilibrium that the game admits is a mixed one. Results for case 2 and case 3 are omitted due to page limitations. Experimental results which were not included in this paper show that the scheme possesses plausible convergence properties even in the case where there is a single saddle point in pure strategies, and that our proposed LA will approach the optimal pure equilibrium, which corresponds to case 2. In case 3, we observed that depending on the initial conditions, our LA converges to one of the two pure equilibria which is usually the one closest to the starting point.

| $p_{max}$ | $\theta = 0.001$ | $\theta = 0.0001$ |
|---|---|---|
| 0.990 | $1.77 \times 10^{-2}$ | $2.03 \times 10^{-2}$ |
| 0.991 | $1.71 \times 10^{-2}$ | $1.69 \times 10^{-2}$ |
| 0.992 | $1.33 \times 10^{-2}$ | $1.54 \times 10^{-2}$ |
| 0.993 | $1.32 \times 10^{-2}$ | $1.52 \times 10^{-2}$ |
| 0.994 | $1.18 \times 10^{-2}$ | $1.02 \times 10^{-2}$ |
| 0.995 | $1.17 \times 10^{-2}$ | $7.86 \times 10^{-3}$ |
| 0.996 | $8.50 \times 10^{-3}$ | $6.37 \times 10^{-3}$ |
| 0.997 | $5.57 \times 10^{-3}$ | $4.43 \times 10^{-3}$ |
| 0.998 | $5.27 \times 10^{-3}$ | $3.34 \times 10^{-3}$ |

Table 1: Error for different values of $\theta$ and $p_{max}$, when $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$ for the game specified by the $R$ and $C$ matrices given by Eq. ( 7) and Eq. ( 8) respectively.

### 4.1 Convergence in Case 1

We will now consider the case of the game which possesses only a single mixed Nash equilibrium, implying that there is no saddle point in pure strategies. The game matrices for which we report the results are $R$ and $C$ given by:

$$R = \begin{pmatrix} 0.2 & 0.6 \\ 0.4 & 0.5 \end{pmatrix}, \tag{7}$$

$$C = \begin{pmatrix} 0.4 & 0.25 \\ 0.3 & 0.6 \end{pmatrix}, \tag{8}$$

which admits $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$.

We ran our simulation for $5 \times 10^6$ iterations, and present the error in Table 1 for different values of $p_{max}$ and $\theta$, as the difference between $X_{\text{opt}}$ and the experimental mean over time of $X(t)$ after convergence[3]. The high value for the number of iterations was chosen in order to eliminate the Monte Carlo error. A significant observation is that the error monotonically decreases as $p_{max}$ goes towards 1 (i.e., when $p_{min} \to 0$). For instance, for $p_{max} = 0.998$ and $\theta = 0.001$, the proposed scheme yields an error of $5.27 \times 10^{-3}$, and further reducing $\theta$ to $0.0001$ leads to an error of $3.34 \times 10^{-3}$.

The time-based behavior of the scheme is illustrated in Figure 1. It displays the trajectory of the mixed strategies for both Players (given by $X(t)$) for an ensemble of 1,000 runs using $\theta = 0.01$ and $p_{max} = 0.99$. The trajectory of the ensemble enables us to notice the mean evolution of the mixed strategies. The spiral pattern results from one of the Players adjusting to the strategy used by the other before the former learns by readjusting its strategy. The process is repeated, thus leading to more minor corrections until the Players attain to the Nash equilibrium. The process can be visualized in Figure 2 presenting the time evolution of the strategies of both Players for a single experiment with $p_{max} = 0.99$

[3]The mean is taken over the last $10\%$ of the total number of iterations, which is a valid methodology since we are dealing with an ergodic Markov process, where the true time average is the same as the true ensemble average.

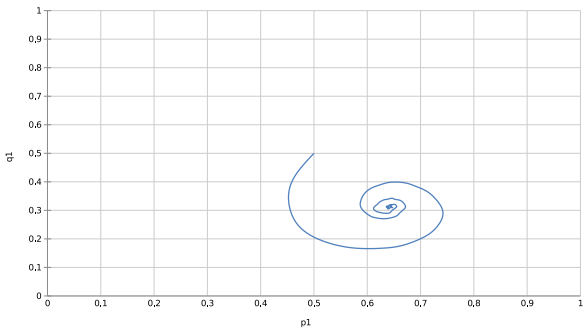


Figure 1: Trajectory of $[p_1(t), q_1(t)]^\intercal$ for the $R$ and $C$ matrices given by Eq. ( 7) and Eq. ( 8) respectively. Here $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$, and the parameters are $p_{max} = 0.99$ and $\theta = 0.01$.

and $\theta = 0.00001$ over $3 \times 10^7$ steps. We observe an oscillatory behavior that vanishes as the Players play for more iterations. It is worth noting that a larger value of $\theta$ will cause more steady state error (as specified in Theorem 1), but it will also disrupt this behavior as the Players take larger updates whenever they receive a Reward. Furthermore, decreasing $\theta$ results in a smaller convergence error, but also negatively affects the convergence speed, since more iterations are required to achieve convergence. Figure 3 depicts the trajectories of the probabilities $p_1$ and $q_1$ for the same settings as those in Figure 2.

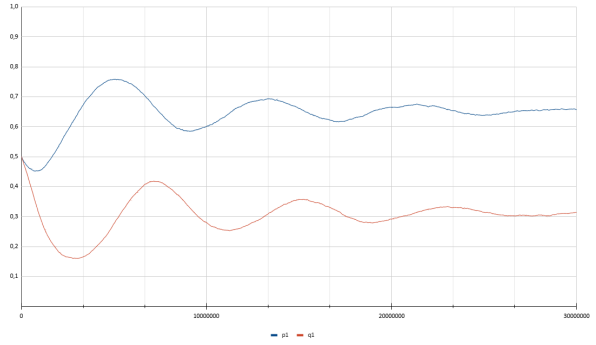

Figure 2: Time Evolution $X(t)$ for for the $R$ and $C$ matrices given by Eq. ( 7) and Eq. ( 8) respectively. Here $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$, and the parameters $p_{max} = 0.99$ and $\theta = 0.00001$.

Now, we turn our attention to the analysis of the deterministic Ordinary Differential Equation (ODE) corresponding to our LA with the reflecting barriers. The trajectory of the ODE can be seen to conform with our intuition, and the results of the LA run is given in Figure 3. Its plot is given in Figure 4. The two ODEs are given by:

$$\begin{aligned}\frac{dp_1}{dt} =& W_1(X)\\ =& p_1(p_{max} - p_1)D_1^A(q_1)\\ &+(1-p_1)(p_{min} - p_1)D_2^A(q_1), \quad \text{and}\end{aligned}$$

$$\begin{aligned}\frac{dq_1}{dt} =& W_1(X)\\ =& p_1(p_{max} - p_1)D_1^A(q_1)\\ &+(1-p_1)(p_{min} - p_1)D_2^A(q_1).\end{aligned}$$

and where $D_1^A$ and $D_2^A$ are:

$$D_1^A(q_1) = q_1 r_{11} + (1-q_1)r_{12}$$

$$D_2^A(q_1) = q_1 r_{21} + (1-q_1)r_{22}.$$

Observe that to obtain the ODE for a particular example, we need to merely replace the entries of $R$ and $C$ in the ODE by their specific values. Thus, we need to only know $R$ and $C$ to plot the trajectories of the ODE, and of course $p_{max}$.

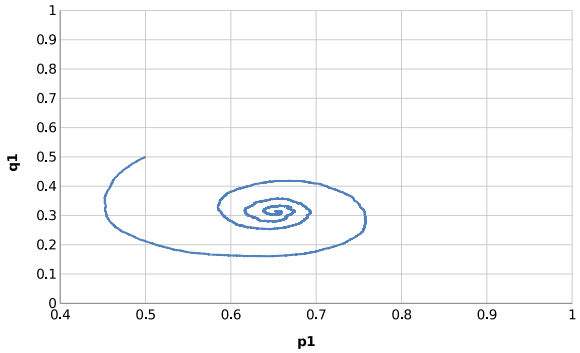


Figure 3: Trajectory of $X(t)$ for the example studied in Figure 2, where $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$, using $p_{max} = 0.99$ and $\theta = 0.00001$.

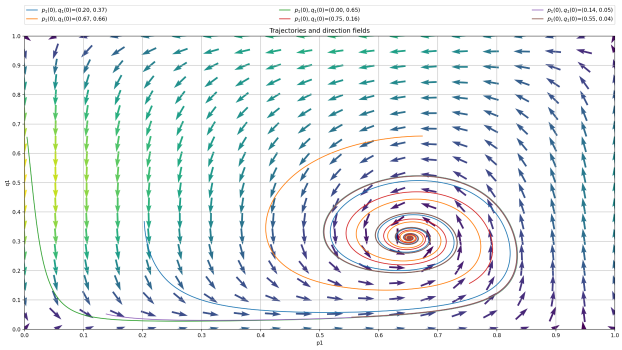


Figure 4: Trajectory of the ODE for Case 1 for $p_{max} = 0.99$.

## 5 Conclusion

In this paper, we propose a LA with artificially *non-absorbing* that is able to solve game theoretical problems. The scheme is able to converge to the game's Nash equilibrium under limited information that has clear advantages over the well-known LA solution for game theoretical due to Sastry et al. (Sastry, Phansalkar, and Thathachar 1994) and revisited by Vranx et al. (Vrancx, Tuyls, and Westra 2008).

Our scheme is an ergodic one and illustrates a design by which an inherently absorbing scheme, in our case, Linear Reward-Inaction ($L_{R-I}$), is rendered ergodic. Interestingly, while being able to solve the mixed Nash equilibrium case, our scheme maintains the plausible properties of the original $L_{R-I}$ as it is able to converge to a near-optimal to the pure strategies in the probability simplex whenever a saddle point exists for pure strategies. As a future work, we would like to extend our scheme to Stackelberg games which are often employed in security and that assume that the defender deploys a mixed strategy that can be fully observed by the attacker who will optimally reply to it. The extension would be interesting but far from being obvious. Furthermore, we aim to extend our scheme to a continuous strategy space by drawing inspiration from the Continuous Action LA (CALA) used by de Jong and Tuyls to solve the Ultimatum Game and the Nash Bargaining Game (de Jong and Tuyls 2011).

## References


Apostolopoulos, P. A.; Tsiropoulou, E. E.; and Papavassiliou, S. 2018. Game-theoretic learning-based qos satisfaction in autonomous mobile edge computing. In *2018 Global Information Infrastructure and Networking Symposium (GIIS)*, 1–5.

Bloembergen, D.; Tuyls, K.; Hennes, D.; and Kaisers, M. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research* 53:659–697.

Cao, H., and Cai, J. 2018. Distributed opportunistic spectrum access in an unknown and dynamic environment: A stochastic learning approach. *IEEE Transactions on Vehicular Technology* 67(5):4454–4465.

de Jong, S., and Tuyls, K. 2011. Human-inspired computational fairness. *Autonomous Agents and Multi-Agent Systems* 22(1):103–126.

Gheisari, S., and Tahavori, E. 2019. Cccla: A cognitive approach for congestion control in internet of things using a game of learning automata. *Computer Communications* 147:40–49.

Hassan, I.; Oommen, J.; and Yazidi, A. 2022. Adaptive learning with artificial barriers yielding nash equilibria in general games. *arXiv e-prints* arXiv:2203.15780.

Jia, L.; Xu, Y.; Zhang, Y.; Sun, Y.; Zhu, Y.; and Dai, X. 2017. A distributed anti-jamming channel selection algorithm for interference mitigation-based wireless networks. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, 151–155.

John oommen, B. 1986. Absorbing and ergodic discretized two-action learning automata. *IEEE transactions on systems, man, and cybernetics* 16(2):282–293.

Lakshmivarahan, S., and Narendra, K. S. 1982. Learning algorithms for two-person zero-sum stochastic games with incomplete information: A unified approach. *SIAM Journal on Control and Optimization* 20(4):541–552.

Narendra, K. S., and Thathachar, M. A. L. 1974. Learning automata - a survey. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-4(4):323–334.

Narendra, K. S., and Thathachar, M. A. 2012. *Learning automata: an introduction*. Courier corporation.

Norman, M. F. 1972. *Markov processes and learning models*, volume 84. Academic Press New York.

Papavassilopoulos, G. 1989. Learning algorithms for repeated bimatrix Nash games with incomplete information. *Journal of optimization theory and applications* 62(3):467–488.

Rauniyar, A.; Yazidi, A.; Engelstad, P.; and Østerbo, O. N. 2020. A reinforcement learning based game theoretic approach for distributed power control in downlink noma. In *2020 IEEE 19th International Symposium on Network Computing and Applications (NCA)*, 1–10.

Sastry, P.; Phansalkar, V.; and Thathachar, M. 1994. Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information. *IEEE Transactions on systems, man, and cybernetics* 24(5):769–777.

Thapa, R.; Jiao, L.; Oommen, B. J.; and Yazidi, A. 2017. A learning automaton-based scheme for scheduling domestic shiftable loads in smart grids. *IEEE Access* 6:5348–5361.

Tian, D.; Zhou, J.; Sheng, Z.; Chen, M.; Ni, Q.; and Leung, V. C. 2017. Self-organized relay selection for cooperative transmission in vehicular ad-hoc networks. *IEEE Transactions on Vehicular Technology* 66(10):9534–9549.

Tsetlin, M. L., et al. 1973. *Automaton theory and modeling of biological systems*. Academic Press New York.

Varshavskii, V. I., and Vorontsova, I. P. 1963. On the behavior of stochastic automata with a variable structure. *Automation and Remote Control* 24:327–333.

Viswanathan, R., and Narendra, K. S. 1974. Games of stochastic automata. *IEEE Transactions on Systems, Man, and Cybernetics* (1):131–135.

Vrancx, P.; Tuyls, K.; and Westra, R. 2008. Switching dynamics of multi-agent learning. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 307–313.

Yang, Z.; Liu, Y.; Chen, Y.; and Jiao, L. 2020. Learning automata based q-learning for content placement in cooperative caching. *IEEE Transactions on Communications* 68(6):3667–3680.

Yazidi, A.; Pinto-Orellana, M. A.; Hammer, H.; Mirtaheri, P.; and Herrera-Viedma, E. 2020. Solving sensor identification problem without knowledge of the ground truth using replicator dynamics. *IEEE Transactions on Cybernetics* 1–9.

Yazidi, A.; Silvestre, D.; and Oommen, B. J. 2021. Solving two-person zero-sum stochastic games with incomplete information using learning automata with artificial barriers. *IEEE Transactions on Neural Networks and Learning Systems* 1–12.

Zhang, Z.; Wang, D.; and Gao, J. 2020. Learning automata-based multiagent reinforcement learning for optimization of cooperative tasks. *IEEE Transactions on Neural Networks and Learning Systems* 1–14.