The Knowledge Engineering Review (2023), **38**, e10, pp. 1–24
doi:10.1017/S0269888923000103

The Knowledge Engineering Review

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Adaptive learning with artificial barriers yielding Nash equilibria in general games[1]

Ismail Hassan[1], B. John Oommen[2] and Anis Yazidi[1]

[1]OsloMet – Oslo Metropolitan University, Oslo, Norway
[2]Carleton University, Ottowa, Canada
Corresponding author: Ismail Hassan; Email: ismail@oslomet.no

Received: 22 February 2023; **Revised:** 22 October 2023; **Accepted:** 23 October 2023

Abstract

Artificial barriers in Learning Automata (LA) is a powerful and yet under-explored concept although it was first proposed in the 1980s. Introducing artificial non-absorbing barriers makes the LA schemes resilient to being trapped in absorbing barriers, a phenomenon which is often referred to as lock in probability leading to an exclusive choice of one action after convergence. Within the field of LA and reinforcement learning in general, there is a sacristy of theoretical works and applications of schemes with artificial barriers. In this paper, we devise a LA with artificial barriers for solving a general form of stochastic bimatrix game. Classical LA systems possess properties of absorbing barriers and they are a powerful tool in game theory and were shown to converge to game's of Nash equilibrium under limited information. However, the stream of works in LA for solving game theoretical problems can merely solve the case where the Saddle Point of the game exists in a pure strategy and fail to reach mixed Nash equilibrium when no Saddle Point exists for a pure strategy.

Furthermore, we provide experimental results that are in line with our theoretical findings.

1. Introduction

Narendra and Thathachar (1974) first presented the term Learning Automata (LA) in their 1974 survey. LA consists of an adaptive learning agent interacting with a stochastic environment with incomplete information. Lacking prior knowledge, LA attempts to determine the optimal action to take by first choosing an action randomly and then 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, achieved. The LA update process can be described by the learning loop shown in Figure 1.

The feedback from the LA is a scalar that falls in the interval [0,1]. If the feedback is binary, meaning 0 or 1, then the Environment is called P-type. Whenever the feedback is a discrete values, we call the environment $Q-$type. In the third case where the feedback is any real number in the interval [0,1], we call the environment as $S-$type.

Depending on their Markovian properties, LA can be classified as either ergodic or equipped with characteristics of absorbing barriers. In an ergodic LA system, the final steady state does not depend on the initial state. In contrast, LA with absorbing barriers, the steady state depends on the initial state and when the LA converges, it gets locked into an absorbing state.

[1]A preliminary version of this paper appeared in the 35th International FLAIRS Conference, Florida, USA, May 15–18, 2022. Also, during the course of this work, the Second Author was an Adjunct Professor with University of Agder, in Grimstad, Norway.

Cite this article: I. Hassan, B. J. Oommen and A. Yazidi. Adaptive learning with artificial barriers yielding Nash equilibria in general games. *The Knowledge Engineering Review* 38(e10): 1-24. https://doi.org/10.1017/S0269888923000103

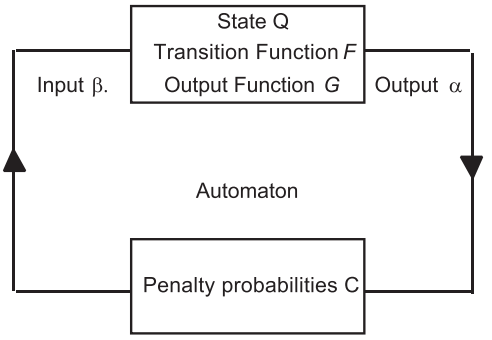


Figure 1. *LA interacting with the environment.*

Absorbing barrier LA are preferred in static environments, while ergodic LA are suitable for dynamic environments.

LA with artificially absorbing barrier were introduced in the 1980s. In the context of LA, artificial barriers refer to additional constraints or obstacles intentionally imposed on the agent's learning environment. These barriers can be designed to shape the agent's behavior, steer its exploration, or promote the discovery of optimal solutions that might not lie in the corners of the simplex.

John Oommen (1986), turned a discretized ergodic scheme into an absorbing one by introducing an artificially absorbing barrier that forces the scheme to converge to one of the absorbing barriers. Such a modification led to the advent of new LA families with previously unknown behavior.

In this paper, we devise a LA with artificial barriers for solving a general form of stochastic bimatrix game. Our proposed algorithm addresses bimatrix games which is a more general version of the zero-sum game treated in Lakshmivarahan and Narendra (1982).

Reward-ϵPenalty ($L_{R-\epsilon P}$) scheme proposed by Lakshmivarahan and Narendra (1982) almost four decades ago, is the only LA scheme that was shown to converge to the optimal mixed Nash equilibrium when no Saddle Point exists in pure strategy, and the proofs were limited to only zero-sum games.

By resorting to the powerful concept of artificial 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. Our deployed scheme is of Linear Reward-Inaction (L_{R-I}) flavor which is originally an absorbing LA scheme, however, we render it non-absorbing by introducing artificial barriers in an elegant and natural manner, in the sense that the well-known legacy L_{R-I} scheme can be seen as an instance of our proposed algorithm for a particular choice of the barrier. Furthermore, we present an S-Learning version of our LA with absorbing barriers that is able to handle S-Learning environment in which the feedback is continuous and not binary as in the case of the L_{R-I}. For a generalized analysis of reinforcement learning in game theory, including the L_{R-I}, we refer the reader to Bloembergen *et al.* (2015).

The contributions of this article can be summarized as follows:

- We introduce a stochastic game with binary outcomes, specifically a reward or a penalty. We extend our consideration to a game where the probabilities of receiving a reward are determined by the corresponding payoff matrix of each player. Furthermore, we propose a limited information framework, a variant often examined in LA. In this game, each player only observes the outcome of his action, either as a reward or penalty, without knowledge of the other player's choice. The player might not be even aware that he is playing against an opponent player.
- We introduce a design principle for our scheme, where players adapt their strategies upon receiving a reward during each round of the repetitive game, yet retain their strategies when faced with a penalty. This approach is in line with the Linear Reward-Inaction, L_{R-I} paradigm. We further extend our discussion by noting a stark contrast to the paradigm presented by Lakshmivarahan and Narendra (1982). In their approach, players consistently revise their

strategies every round, with the magnitude of probability adjustments solely based on the receipt of a reward or penalty at each time instance.
- Furthermore, we provide an extension of our scheme to handle *S*-learning environment where the feedback is not binary but rather continuous. The informed reader will notice that our main focus is on the case of *P*-type environment, while we give enough exposure and attention related to the *S*-type environment.

2. Related work

Studies on strategic games with LA were focused mainly on traditional L_{R-I} which is desirable to use as it can yield Nash equilibrium in pure strategies (Sastry *et al.,* 1994). Although other ergodic schemes such as L_{R-P} were used in games (Viswanathan & Narendra, 1974) with limited information, they did not gain popularity at least when it comes to applications due to their inability to converge to Nash equilibrium. LA has found numerous applications in game theoretical applications such as sensor fusion without knowledge of the ground truth (Yazidi *et al.,* 2022), for distributed power control in wireless networks and more particularly NOMA (Rauniyar *et al.,* 2020), optimization of cooperative tasks (Zhang *et al.,* 2020), for content placement in cooperative caching (Yang *et al.,* 2020), congestion control in Internet of Things (Gheisari & Tahavori, 2019), reaching agreement in Ultimatum games utilizing a continuous space strategy rather than working within a discrete actions space (De Jong *et al.,* 2008), QoS satisfaction in autonomous mobile edge computing (Apostolopoulos *et al.,* 2018), opportunistic spectrum access (Cao & 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), relay selection in vehicular ad-hoc networks (Tian et *al.,* 2017), load balancing by invoking the feedback from a purely local agent (Schaerf *et al.,* 1994) etc.

The application of game theory in cybersecurity is also a promising research area attracting lots of attention (Do *et al.,* 2017; Fielder, 2020; Sokri, 2020). Our LA-based solution is well suited for that purpose. In cybersecurity, algorithms that can converge to mixed equilibria are preferred over those that get locked into pure ones since randomization reduces an attacker's predictive capability to guess the implemented strategy of the defender. For example, let us consider a repetitive two-person security game comprising of a hacker and network administrator. The hacker intends to disrupt the network by launching a Distributed Denial of Service attack (DDOS) of varying magnitudes that could be classified as high or low. The administrator can use varying levels of security measures to protect the assets. We can assume that the adoption of a strong defense strategy by the defender has an extra cost compared to a low one. Similarly, the usage of a high magnitude attack strategy by the attacker has a higher cost compared to a low magnitude attack strategy. Another example of a security game is the jammer and transmitter game (Vadori *et al.,* 2015) where a jammer is trying to guess the communication channel of the transmitter to interfere and block the communication. The transmitter chooses probabilistically a channel to transmit over and the jammer chooses probabilistically a channel to attack. Clearly converging to pure strategies is neither desirable by the jammer nor by the transmitter as it will give a predictive advantage to the opponent. A pertinent example of an application of our proposed scheme is distributed discrete power control problem (Xing & Chandramouli, 2008). Indeed, in Xing and Chandramouli (2008), these authors used the counter-part scheme due to Lakshmivarahan and Narendra (1982) in order to decide the power level of each terminal in a non-cooperative game setting. The latter authors adopt a utility function due to Saraydar et al. which is expressed as 'the number of information bits received successfully per Joule of energy expended' (Saraydar *et al.,* 2002). This utility function depends on a certain number of parameters, such as the interference exercised by the other terminals. More specifically, each terminal is only able to observe local information, namely its utility, without having a knowledge of the power level used by the other terminals. The authors of Xing and Chandramouli (2008) proved that whenever the power levels of the terminals are discrete, there might be cases where there are multiple Nash equilibria or there may also be mixed ones. However, for continuous power level choices, Saraydar *et al.* (2002) have shown that the Nash equilibrium is unique.

3. The game model

In this section, we begin by presenting the formal definition of LA in detail, ensuring a clear understanding of its foundational concepts. Following that, we delve into formalizing the game model that is being investigated.

Formally, a LA is defined by the mean of a quintuple $\langle A, B, Q, F(.,.), G(.)\rangle$, where the elements of the quintuple are defined term by term as:

1. $A = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ gives the set of actions available to the LA, while $\alpha(t)$ is the action selected at time instant t by the LA. Note that the LA selects one action at a time, and the selection is sequential.
2. $B = \{\beta_1, \beta_2, \ldots, \beta_m\}$ denotes the set of possible input values that the LA can receive. $\beta(t)$ denotes the input at time instant t which is a form of feedback.
3. $Q = \{q_1, q_2, \ldots, q_s\}$ represents the states of the LA where $Q(t)$ is the state at time instant t.
4. $F(.,.) : Q \times B \mapsto Q$ is the transition function at time t, such that, $q(t+1) = F[q(t), \beta(t)]$. In simple terms, $F(.,.)$ returns the next state of the LA at time instant $t+1$ given the current state and the input from the environment both at time t using either a deterministic or a stochastic mapping.
5. $G(.)$ defines *output function*, it represents a mapping $G{:}Q \mapsto A$ which determines the action of the LA as a function of the state.

The Environment, E is characterized by :

- $C = \{c_1, c_2, \ldots, c_r\}$ is a set of penalty probabilities, where $c_i \in C$ corresponds to the penalty of action α_i.

Let $P(t) = \left[p_1(t)\ p_2(t)\right]^{\intercal}$ 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 $Q(t) = \left[q_1(t)\ q_2(t)\right]^{\intercal}$. The extension to more than two actions per player is straightforward following the method analogous to what was used by Papavassilopoulos (1989), which extended the work of 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}. Player A is in this case 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 player A and player B respectively:

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}, \tag{3.1}$$

and the matrix C

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}, \tag{3.2}$$

where, as aforementioned, all 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 Bernoulli feedback such in the case of P-type environment. It is possible also to consider c_{11} as stochastic continuous variable with mean c_{11} and which realization in c_{11}, however, for the sake of simplicity we consider c_{11}, and consequently C and R as deterministic. The asymptotic convergence proofs

for the $S-$type environment will remain valid independently of whether C and R are deterministic or whether they are obtained from a distribution with support in the interval [0,1] and with their means defined by the matrices.

Independently of the environment type, whether it is $P-$type or $S-$type environments, we have three cases to be distinguished for 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 just one mixed equilibrium. The first case depicts the situation where no Saddle Point exists in pure strategies. In other words, the only Nash equilibrium is a mixed one. Based on the fundamentals of Game Theory, the optimal mixed strategies 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 $L' = (c_{11} + c_{22}) - (c_{12} + c_{21})$.
- Case 2: if $(r_{11} - r_{21})(r_{12} - r_{22}) > 0$ or $(c_{11} - c_{12})(c_{21} - c_{22}) > 0$, then there is just one pure equilibrium since there is one player at least 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 called the 'Saddle Points' for the game. Since the outcome for a given joint action is stochastic, the game is of stochastic genre.

4. Game theoretical LA algorithm based on the L_{R-I} with artificial barriers

In this section, we shall present our L_{R-I} with artificial barriers that is devised specially for the P-type environments.

4.1. Non-absorbing artificial barriers

As we have seen above from surveying the literature, an originally ergodic LA can be rendered absorbing by operating a change in its end states. However, what is unknown in the literature is a scheme which is originally absorbing can be rendered ergodic. In many cases, this can be achieved by making the scheme behave according to the absorbing scheme rule over the probability simplex and pushing the probability back inside the simplex whenever the scheme approaches absorbing barriers. Such a scheme is novel in the field of LA and its advantage is that the strategies avoid being absorbed in non-desirable absorbing barriers. Further, and interestingly, by countering the absorbing barriers, the scheme can migrate stochastically towards a desirable mixed strategy. Interestingly, 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. We shall give the details of our devised scheme in the next section which enjoys the above mentioned properties.

4.2. Non-absorbing Gagme playing

At this juncture, we shall 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 presents a new variant of the L_{R-I} scheme, which is made rather ergodic by modifying the update rule in a general form which makes the original L_{R-I} with absorbing barriers corresponding to the corners of the simplex an instance of the latter general scheme for a particular choice of parameters of the scheme. The proof of convergence is based on Norman's theory for learning processes characterized by small learning steps (Norman, 1972; Narendra & Thathachar, 2012).

We introduce p_{max} as the artificial barrier which is a real value close to 1. Similarly, we introduce $p_{min} = 1 - p_{max}$ 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 updates rules that ensure that each update keeps 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:

$$\begin{aligned} p_i(t+1) &= p_i(t) + \theta(p_{max} - p_i(t)) \\ p_s(t+1) &= p_s(t) + \theta(p_{min} - p_s(t)) \quad \text{for} \quad s \neq i. \end{aligned} \tag{4.1}$$

where θ is a learning parameter. The informed reader observes that the update rules coincide with the classical L_{R-I} except that p_{max} replaces unity for updating $p_i(t+1)$ and that p_{min} replaces zero for updating $p_s(t+1)$.

Following the Inaction principle of the L_{R-I}, whenever the player receives a penalty, its action probabilities are kept unchanged which is formally given by:

$$\begin{aligned} p_i(t+1) &= p_i(t) \\ p_s(t+1) &= p_s(t) \quad \text{for} \quad s \neq i. \end{aligned} \tag{4.2}$$

The update rules for the mixed strategy $q(t+1)$ are defined in a similar fashion. We shall now move to a theoretical analysis of the convergence properties of our proposed algorithm for solving a strategic game. In order to denote the optimal Nash equilibrium of the game we use the pair $(p_{\text{opt}}, q_{\text{opt}})$.

Let the vector $X(t) = \left[p_1(t)\ q_1(t)\right]^{\intercal}$. We resort to the notation $\Delta X(t) = X(t+1) - X(t)$. For denoting the conditional expected value operator we use the nomenclature $\mathbb{E}[\cdot|\cdot]$. Using those notations, we introduce the next theorem of the article.

Theorem 1. *Consider a two-player game with a payoff matrices as in –Equations (3.1) and (3.2), and a learning algorithm defined by Equations (4.1) and (4.2) for both players A and B, with learning rate θ. Then, $E[\Delta X(t)|X(t)] = \theta W(x)$ and for every $\epsilon > 0$, there exists a unique stationary point $X^* = \left[p_1^*\ q_1^*\right]^{\intercal}$ satisfying:*

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

Proof We start by first computing the conditional expected value of the increment $\Delta X(t)$:

$$\begin{aligned} E[\Delta X(t)|X(t)] &= E[X(t+1) - X(t)|X(t)] \\ &= \begin{bmatrix} E[p_1(t+1) - p_1(t)|X(t)] \\ E[q_1(t+1) - q_1(t)|X(t)]) \end{bmatrix} \\ &= \theta \begin{bmatrix} W_1(X(t)) \\ W_2(X(t)) \end{bmatrix} \\ &= \theta W(X(t)), \end{aligned}$$

where the above format is possible since all possible updates share the form $\Delta X(t) = \theta W(t)$, for some $W(t)$, as given in Equation (4.1). For ease of notation, we drop the dependence on t with the implicit assumption that all occurrences of X, p_1 and q_1 represent $X(t)$, $p_1(t)$ and $q_1(t)$ respectively. $W_1(x)$ is then:

$$\begin{aligned} W_1(X) &= p_1 q_1 r_{11}(p_{max} - p_1) + p_1(1-q_1) r_{12}(p_{max} - p_1) \\ &\quad + (1-p_1) q_1 r_{21}(p_{min} - p_1) \\ &\quad + (1-p_1)(1-q_1) r_{22}(p_{min} - p_1) \\ &= p_1 \left[q_1 r_{11} + (1-q_1) r_{12} \right] (p_{max} - p_1) \\ &\quad + (1-p_1) \left[q_1 r_{21} + (1-q_1) r_{22} \right] (p_{min} - p_1) \\ &= p_1 (p_{max} - p_1) D_1^A(q_1) + (1-p_1)(p_{min} - p_1) D_2^A(q_1), \end{aligned} \tag{4.3}$$

where,

$$D_1^A(q_1) = q_1 r_{11} + (1-q_1) r_{12} \tag{4.4}$$

$$D_2^A(q_1) = q_1 r_{21} + (1-q_1) r_{22}. \tag{4.5}$$

By replacing $p_{max} = 1 - p_{min}$ and rearranging the expression we get:

$$\begin{aligned} W_1(X) &= p_1(1-p_1) D_1^A(q_1) - p_1 p_{min} D_1^A(q_1) \\ &\quad + (1-p_1) p_{min} D_2^A(q_1) - p_1(1-p_1) D_2^A(q_1) \\ &= p_1(1-p_1) \left[D_1^A(q_1) - D_2^A(q_1) \right] \\ &\quad - p_{min} \left[p_1 D_1^A(q_1) - (1-p_1) D_2^A(q_1) \right]. \end{aligned}$$

Similarly, we can get

$$\begin{aligned} W_2(X) &= q_1 p_1 c_{11}(p_{max} - q_1) + q_1(1-p_1) c_{21}(p_{max} - q_1) \\ &\quad + (1-q_1) p_1 c_{12}(p_{min} - q_1) + (1-q_1)(1-p_1) c_{22}(p_{min} - q_1) \\ &= q_1 \left[p_1 c_{11} + (1-p_1) c_{21} \right] (p_{max} - q_1) \\ &\quad + (1-q_1) \left[p_1 c_{12} + (1-p_1) c_{22} \right] (p_{min} - q_1) \\ &= q_1 (p_{max} - q_1) D_1^B(p_1) + (1-q_1)(p_{min} - q_1) D_2^B(p_1) \end{aligned} \tag{4.6}$$

where

$$D_1^B(p_1) = p_1 c_{11} + (1-p_1) c_{21} \tag{4.7}$$

$$D_2^B(p_1) = p_1 c_{12} + (1-p_1) c_{22}. \tag{4.8}$$

By replacing $p_{max} = 1 - p_{min}$ and rearranging the expression we get:

$$\begin{aligned} W_2(X) &= q_1(1-q_1)(1 - D_1^B(p_1)) - q_1 p_{min} D_1^B(p_1) \\ &\quad + (1-q_1) p_{min} D_2^B(p_1) - q_1(1-q_1) D_2^B(p_1) \\ &= q_1(1-q_1) \left[D_1^B(p_1) - D_2^B(p_1) \right] \\ &\quad - p_{min} \left[q_1 D_1^B(p_1) - (1-q_1) D_2^B(p_1) \right]. \end{aligned} \tag{4.9}$$

We need to address the three identified cases.

Consider Case 1: Only One Mixed Equilibrium Case, where there is only a single mixed equilibrium. We get

$$\begin{aligned} D_{12}^A(q_1) &= D_1^A(q_1) - D_2^A(q_1) \\ &= (r_{12} - r_{22}) + L q_1. \end{aligned} \tag{4.10}$$

We also should distinguish details of the equilibrium according to the entries in the payoff matrices R and C for Case 1. This case can be divided into two sub-cases. The first sub-case is given by:

$$r_{11} > r_{21}, r_{12} < r_{22}; c_{11} < c_{12}, c_{21} > c_{22}, \tag{4.11}$$

The second sub-case is given by:

$$r_{11} < r_{21}, r_{12} > r_{22}; c_{11} > c_{12}, c_{21} < c_{22}, \tag{4.12}$$

For the sake of brevity, we consider the first sub-case given by condition Equation 4.11. We have $L > 0$, since $r_{11} > r_{12}$ and $r_{22} > r_{21}$. Therefore $D^A_{12}(q_1)$ is an increasing function of q_1 and

$$\begin{cases} D^A_{12}(q_1) < 0, \text{ if } q_1 < q_{\text{opt}}, \\ D^A_{12}(q_1) = 0, \text{ if } q_1 = q_{\text{opt}}, \\ D^A_{12}(q_1) > 0, \text{ if } q_1 > q_{\text{opt}}. \end{cases} \tag{4.13}$$

For a given q_1, $W_1(X)$ is quadratic in p_1. Also, we have:

$$\begin{aligned} W_1\left(\begin{bmatrix} 0 \\ q_1 \end{bmatrix}\right) &= p_{min} D^A_2(q_1) > 0 \\ W_1\left(\begin{bmatrix} 1 \\ q_1 \end{bmatrix}\right) &= -p_{min} D^A_1(q_1) < 0. \end{aligned} \tag{4.14}$$

Since $W_1(X)$ is quadratic with a negative second derivative with respect to p_1, and since the inequalities in Equation (4.14) are strict, it admits a single root p_1 for $p_1 \in [0, 1]$. Moreover, we have $W_1(X) = 0$ for some p_1 such that:

$$\begin{cases} p_1 < \dfrac{1}{2}, \text{ if } q_1 < q_{\text{opt}}, \\ p_1 = \dfrac{1}{2}, \text{ if } q_1 = q_{\text{opt}}, \\ p_1 > \dfrac{1}{2}, \text{ if } q_1 > q_{\text{opt}}. \end{cases} \tag{4.15}$$

Using a similar argument, we can see that there exists a single solution for each p_1, and as $p_{min} \to 0$, we conclude that $W_1(X) = 0$ whenever $p_1 \in \{0, p_{\text{opt}}, 1\}$. Arguing in a similar manner we see that $W_2(X) = 0$ when:

$$X \in \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} p_{\text{opt}} \\ q_{\text{opt}} \end{bmatrix} \right\}.$$

Thus, there exists a small enough value for p_{min} such that $X^* = [p^*, q^*]^\intercal$ satisfies $W_2(X^*) = 0$, proving Case 1).

In the proof of Case 1), we take advantage of the fact that for small enough p_{min}, the learning algorithm enters a stationary point, and also identified the corresponding possible values for this point. It is thus always possible to select a small enough $p_{min} > 0$ such that X^* approaches X_{opt}, concluding the proof for Case 1).

Case 2) and Case 3) can be derived in a similar manner, and the details are omitted to avoid repetition. □

In the next theorem, we show 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.*

Proof. We start the proof by writing the explicit format for $\frac{\partial W(X)}{\partial X} = \begin{bmatrix} \frac{\partial W_1(X)}{\partial p_1} & \frac{\partial W_1(X)}{\partial q_1} \\ \frac{\partial W_2(X)}{\partial p_1} & \frac{\partial W_2(X)}{\partial q_1}, \end{bmatrix}$ and then computing each of the entries as below:

$$\begin{aligned}
\frac{\partial W_1(X)}{\partial p_1} &= (1-2p_1)\left(D_1^A(q_1) - D_2^A(q_1)\right) \\
&\quad - p_{min}\left(D_1^A(q_1) + D_2^A(q_1)\right) \\
&= (1-2p_1)D_{12}^A(q_1) - p_{min}\left(D_1^A(q_1) + D_2^A(q_1)\right). \\
\frac{\partial W_1(X)}{\partial q_1} &= p_1(1-p_1)L - p_{min}(p_1(r_{11}-r_{12}) + (1-p_1)(r_{22}-r_{21})). \\
\frac{\partial W_2(X)}{\partial p_1} &= q_1(1-q_1)L' - p_{min}(q_1(c_{11}-c_{12}) + (1-q_1)(c_{22}-c_{21})). \\
\frac{\partial W_2(X)}{\partial q_1} &= (1-2q_1)D_{12}^B(p_1) - p_{min}\left(D_1^B(p_1) + D_2^B(p_1)\right).
\end{aligned}$$

As seen in Theorem 1, for a small enough value for p_{min}, we can ignore the terms that are weighted by p_{min}, and we will thus have $\frac{\partial W(X^*)}{\partial X} \approx \frac{\partial W(X_{\text{opt}})}{\partial X}$. We now subdivide the analysis into the three cases.

Case 1: No Saddle Point in pure strategies. In this case, we have:

$$D_1^A(q_{\text{opt}}) = D_2^A(q_{\text{opt}}) \quad \text{and } D_1^B(p_{\text{opt}}) = D_2^B(p_{\text{opt}})$$

which makes

$$\frac{\partial W_1(X_{\text{opt}})}{\partial p_1} = -2p_{min}D_1^A(q_{\text{opt}}). \tag{4.16}$$

Similarly, we can compute

$$\frac{\partial W_1(X_{\text{opt}})}{\partial q_1} = (1-2p_{min})p_{\text{opt}}(1-p_{\text{opt}})L. \tag{4.17}$$

The entry $\frac{\partial W_2(X_{\text{opt}})}{\partial p_1}$ can be simplified to:

$$\frac{\partial W_2(X_{\text{opt}})}{\partial p_1} = (1-2p_{min})q_{\text{opt}}(1-q_{\text{opt}})L' \tag{4.18}$$

and

$$\frac{\partial W_2(X_{\text{opt}})}{\partial q_1} = 2p_{min}D_1^B(p_{\text{opt}}) \tag{4.19}$$

resulting in:

$$\frac{\partial W(X_{\text{opt}})}{\partial X} = \begin{bmatrix} -2p_{min}D_1^A(q_{\text{opt}}) & (1-2p_{min})p_{\text{opt}}(1-p_{\text{opt}})L \\ (1-2p_{min})q_{\text{opt}}(1-q_{\text{opt}})L' & -2p_{min}D_1^B(p_{\text{opt}}) \end{bmatrix}. \tag{4.20}$$

We know that this case can be divided into two sub-cases. Let us consider the first sub-case given by:

$$r_{11} > r_{21}, r_{12} < r_{22}; c_{11} < c_{12}, c_{21} > c_{22}, \tag{4.21}$$

Thus, $L > 0$ and $L' < 0$ as a consequence of Equation (4.21)

Thus, the matrix given in Equation (4.20) satisfies:

$$\det\left(\frac{\partial W(X_{\text{opt}})}{\partial x}\right) > 0\,,\ \text{trace}\left(\frac{\partial W(X_{\text{opt}})}{\partial x}\right) < 0, \tag{4.22}$$

which implies the 2×2 matrix is negative definite.

Case 2: Only one single pure equilibrium. According to this case: $(r_{11} - r_{21})(r_{12} - r_{22}) > 0$ or $(c_{11} - c_{12})(c_{21} - c_{22}) > 0$.

The condition for only one pure equilibrium can be divided into four different sub-cases.

Without loss of generality, we can consider a particular sub-case where $q_{\text{opt}} = 1$ and $p_{\text{opt}} = 1$. This reduces to $r_{11} - r_{21} > 0$ and $c_{11} - c_{12} > 0$.

Computing the entries of the matrix for this case yields:

$$\frac{\partial W_1(X_{\text{opt}})}{\partial p_1} = -(r_{11} - r_{21}) - p_{min}(r_{11} + r_{21}), \tag{4.23}$$

and

$$\frac{\partial W_1(X_{\text{opt}})}{\partial q_1} = -p_{min}(r_{11} - r_{12}). \tag{4.24}$$

The entry $\dfrac{\partial W_2(X_{\text{opt}})}{\partial p_1}$ can be simplified to:

$$\frac{\partial W_2(X_{\text{opt}})}{\partial p_1} = -p_{min}(c_{11} - c_{12}) \tag{4.25}$$

and

$$\frac{\partial W_2(X_{\text{opt}})}{\partial q_1} = -(c_{11} - c_{12}) - p_{min}(c_{11} + c_{12}) \tag{4.26}$$

resulting in:

$$\frac{\partial W(X_{\text{opt}})}{\partial X} = \begin{bmatrix} -(r_{11} - r_{21}) - p_{min}(r_{11} + r_{21}) & -p_{min}(r_{11} - r_{12}) \\ -p_{min}(c_{11} - c_{12}) & -(c_{11} - c_{12}) - p_{min}(c_{11} + c_{12}) \end{bmatrix}. \tag{4.27}$$

The matrix in (4.34) satisfies:

$$\det\left(\frac{\partial W(X_{\text{opt}})}{\partial X}\right) > 0\,,\ \text{trace}\left(\frac{\partial W(X_{\text{opt}})}{\partial X}\right) < 0 \tag{4.28}$$

for a sufficiently small value of p_{min}, which again implies that the 2×2 matrix is negative definite.

Case 3: Two pure equilibrium and one mixed equilibrium. In this case, $(r_{11} - r_{21})(r_{12} - r_{22}) < 0$, $(c_{11} - c_{12})(c_{21} - c_{22}) < 0$ and $(r_{11} - r_{21})(c_{11} - c_{12}) > 0$.

Without loss of generality, we suppose that $(p_{\text{opt}}, q_{\text{opt}}) = (1, 1)$ and $(p_{\text{opt}}, q_{\text{opt}}) = (0, 0)$ are the two pure Nash equilibria. This corresponds to a sub-case where:

$$r_{11} - r_{21} > 0, c_{11} - c_{12} > 0, r_{22} - r_{12} > 0, c_{22} - c_{21} > 0, \tag{4.29}$$

$r_{11} - r_{21} > 0$ and $c_{11} - c_{12} > 0$ because of the Nash equilibrium $(p_{\text{opt}}, q_{\text{opt}}) = (1, 1)$. Similarly, $r_{22} - r_{12} > 0$ and $c_{22} - c_{21} > 0$ because of the Nash equilibrium $(p_{\text{opt}}, q_{\text{opt}}) = (1, 1)$.

Whenever $(p_{\text{opt}}, q_{\text{opt}}) = (1, 1)$, we obtain stability of the fixed point as demonstrated in the previous case, case 2.

Now, let us consider the stability for $(p_{\text{opt}}, q_{\text{opt}}) = (0, 0)$.

Computing the entries of the matrix for this case yields:

$$\frac{\partial W_1(X_{\text{opt}})}{\partial p_1} = (r_{12} - r_{22}) - p_{min}(r_{12} + r_{22}), \tag{4.30}$$

and

$$\frac{\partial W_1(X_{\text{opt}})}{\partial q_1} = -p_{min}(r_{22} - r_{21}). \tag{4.31}$$

The entry $\frac{\partial W_2(X_{\text{opt}})}{\partial p_1}$ can be simplified to:

$$\frac{\partial W_2(X_{\text{opt}})}{\partial p_1} = -p_{min}(c_{22} - c_{12}) \tag{4.32}$$

and

$$\frac{\partial W_2(X_{\text{opt}})}{\partial q_1} = (c_{21} - c_{22}) - p_{min}(c_{21} + c_{22}) \tag{4.33}$$

resulting in:

$$\frac{\partial W(X_{\text{opt}})}{\partial X} = \begin{bmatrix} (r_{12} - r_{22}) - p_{min}(r_{12} + r_{22}) & -p_{min}(r_{22} - r_{21}) \\ -p_{min}(c_{22} - c_{12}) & (c_{21} - c_{22}) - p_{min}(c_{21} + c_{22}) \end{bmatrix}. \tag{4.34}$$

The matrix in (4.34) satisfies:

$$\det\left(\frac{\partial W(X_{\text{opt}})}{\partial X}\right) > 0\,,\ \text{trace}\left(\frac{\partial W(X_{\text{opt}})}{\partial X}\right) < 0 \tag{4.35}$$

for a sufficiently small value of p_{min}, which again implies that the 2×2 matrix is negative definite.

Now, what remains to be shown is that the mixed Nash equilibrium in this case is unstable.

$$\frac{\partial W(X_{\text{opt}})}{\partial X} = \begin{bmatrix} -2p_{min}D_1^A(q_{\text{opt}}) & (1 - 2p_{min})p_{\text{opt}}(1 - p_{\text{opt}})L \\ (1 - 2p_{min})q_{\text{opt}}(1 - q_{\text{opt}})L' & -2p_{min}D_1^B(p_{\text{opt}}) \end{bmatrix}. \tag{4.36}$$

Using Equation 4.29, we can see that $L > 0$ and $L' > 0$ and thus:

$$\det\left(\frac{\partial W(X_{\text{opt}})}{\partial X}\right) < 0 \tag{4.37}$$

□

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:*

$\left[E(p_1(t))\ E(q_1(t))\right] \to \left[p^*_{\text{opt}}\ q^*_{\text{opt}}\right]$

where $\left[p^*_{\text{opt}}\ q^*_{\text{opt}}\right]$ corresponds to a Nash equilibrium of the game.

Proof The proof of the result is obtained by virtue of applying a classical result due to Norman (1972), given in the Appendix A, in the interest of completeness.

Norman theorem has been traditionally used to prove considerable amount of the results in the field of LA. In the context of game theoretical LA schemes, Norman theorem has been adapted by Lakshmivarahan and Narendra to derive similar convergence properties of the $L_{R-\epsilon P}$ (Lakshmivarahan & Narendra, 1982) for the zero-sum game. It is straightforward to verify that Assumptions (1)-(6) as required for Norman's result in the appendix are satisfied.

Thus, by further invoking Theorem 1 and Theorem 2, the result follows.

Indeed, the convergence proof follows the same line as the proofs in Lakshmivarahan and Narendra (1982) and Xing and Chndramouli (2008) which builds upon the Norman theorem.

We can write

$$E\{[\Delta X(t) - X(t)]^{\intercal}[\Delta X(t) - X(t)]|X(t)\} = \theta^2 s(X(t)),$$

where $s(X(t)) = a(X(t)) - W(X(t))^{\intercal}W(X(t))$ and $a(X(t)) = E[\Delta X(t)^{\intercal}\Delta X(t)|X(t)]$.

The elements of the matrix $a(X(t))$ can be easily computed. All the states are non-absorbing, it follows that $s(X(t))$ is positive definite.

Furthermore,

$$E[\,|\Delta X(t)|\,|^3 X(t)] = O(\theta^3),$$

where $|.|$ is the norm function. $W(X(t))$ has a bounded Lipschitz derivative. In addition, $s(X(t))$ is Lipschitz.

It follows that the process $X(t)$ satisfies all conditions of the Norman theorem.

Hence,

$$E[\Delta X(t)|X(0) = X] = \theta y(t\theta) + O(\theta), \tag{4.38}$$

where

$$y\prime(t) = W(y(t))$$

where $y(0) = X(0) = X$,

For properties of $y\prime(t)$ it follows that Equation (4.38) is uniformly asymptotically stable, that is $y(t)$ converges to X^* as $t \to \infty$.

This implies that $\dfrac{X(t) - y(t\theta)}{\sqrt{\theta}}$ converges in distribution, which implies that $E[X(t)]$ converges.

That is:

$\lim_{t\to\infty} E[X(t)]$ exists.

For small enough p_{min}, X^* approximates X_{opt}.

It follows that for any $\delta > 0$ there exits $0 < \theta^* < 1$, such that $\lim_{t\to\infty} \left|E[X(t)) - X_{\text{opt}}\right| < \delta$ □

5. Game theoretical LA algorithm based on the S-learning with artificial barriers

In this section, we give the update equations for the LA when the environment is of $S-$ type.

In the case of $S-$ type, the game is defined by two payoff matrices, R and C describing a deterministic feedback of player A and player B respectively.

All the entries of both matrices are deterministic numbers like in classical game theory settings.

The environment returns $u_i^A(t)$: the payoff defined by the matrix R for player A at time t whenever player A question chooses an action $i \in \{1, 2\}$.

The update rule for the player A that takes into account $u_i^A(t)$ is given by:

$$\begin{aligned} p_i(t+1) &= p_i(t) + \theta u_i^A(p_{max} - p_i(t)) \\ p_s(t+1) &= p_s(t) + \theta u_i^A(p_{min} - p_s(t)) \quad \text{for} \quad s \neq i. \end{aligned} \tag{5.1}$$

where θ is a learning parameter.

Note u_i^A is the feedback for action i of the player A which is one entry in the $i'h$ row of the matrix R, depending on the action of the player B.

Similarly we can define $u_i^B(t)$ the payoff defined by the matrix C for player B at time t whenever player B question chooses an action $i \in \{1, 2\}$.

For instance, if at time t, player A takes action 1 and player B takes action 2, then $u_1^A(t) = r_{12}$ and $u_2^B(t) = c_{21}$.

The update rules for player B can be obtained by analogy to those given for player A.

Table 1. *Error for different values of θ and p_{max}, when $p_{opt} = 0.6667$ and $q_{opt} = 0.3333$ for the game specified by the R matrix given by Equation (6.1) and the C matrix given by Equation (6.2)*

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

Theorem 4. *We consider the update equations given by the S− Learning scheme given above in this Section. For a sufficiently small p_{min} approaching 0, and as $\theta \to 0$ and as time goes to infinity:*
$\left[E(p_1(t))\ E(q_1(t))\right] \to \left[p^*_{\text{opt}}\ q^*_{\text{opt}}\right]$
where $\left[p^*_{\text{opt}}\ q^*_{\text{opt}}\right]$ corresponds to a Nash equilibrium of the game.

Proof The proofs of this theorem follows the same lines as the proofs given in Section 4 and are omitted here for the sake of brevity. □

6. Experimental results

In this Section, we focus on providing thorough experiments for L_{R-I} scheme. Some experiments of $S-$ LA for handling $S-$ type environments are given in the Appendix 7 that mainly aim to verify our theoretical findings.

To verify the theoretical properties of the proposed learning algorithm, we conducted several simulations that will be presented in this section. By using different instances of the payoff matrices R and C, we can experimentally cover the three cases referred to in Section 4.

6.1. Convergence in Case 1

We examine a case of the game where only one mixed Nash equilibrium exists meaning that there is no Saddle Point in pure strategies. The game matrices R and C are given by:

$$R = \begin{pmatrix} 0.2 & 0.6 \\ 0.4 & 0.5 \end{pmatrix}, \tag{6.1}$$

$$C = \begin{pmatrix} 0.4 & 0.25 \\ 0.3 & 0.6 \end{pmatrix}, \tag{6.2}$$

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

We ran our simulation for 5×10^6 iterations, and present the error in Table 1 for different values of p_{max} and θ as the difference between X_{opt} and the mean over time of $X(t)$ after convergence[2]. 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

[2]The mean is taken over the last 10% of the total number of iterations.

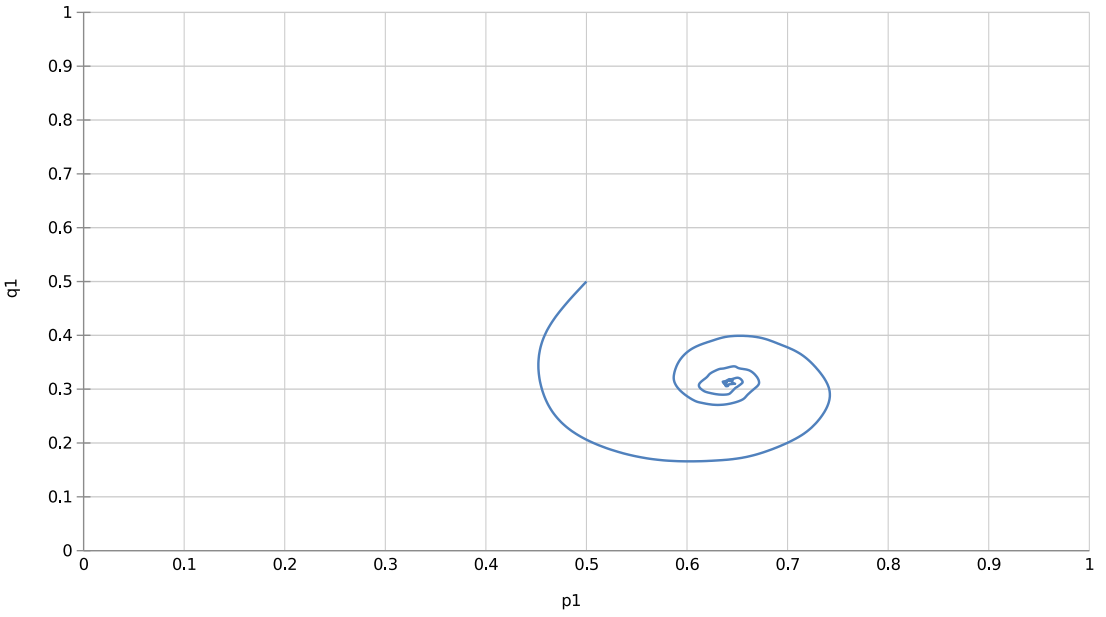


Figure 2. *Trajectory of* $[p_1(t), q_1(t)]^\intercal$ *for the case of the R matrix given by Equation (6.1) and the C matrix given by Equation (6.2) with* $p_{opt} = 0.6667$ *and* $q_{opt} = 0.3333$*, and using* $p_{max} = 0.99$ *and* $\theta = 0.01$.

instance, for $p_{max} = 0.998$ and $\theta = 0.001$, the proposed scheme yields an error of 5.27×10^{-3}, and further reducing $\theta = 0.0001$ leads to an error of 3.34×10^{-3}.

The behavior scheme is illustrated in Figure 2 showing the trajectory of the mixed strategies for both players (given by $X(t)$) for an ensemble of 1000 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 reach the Nash equilibrium.

The process can be visualized in Figure 3 presenting the time evolution of the strategies of both players for a single experiment with $p_{max} = 0.99$ and $\theta = 0.00001$ over 3×10^7 steps. We observe an oscillatory behavior which vanishes as the players play for more iterations. It is worth noting that a larger value of θ 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 θ results in a smaller convergence error, but also affects negatively the convergence speed as more iterations are required to achieve convergence. Figure 4 depicts the trajectories of the probabilities p_1 and q_1 for the same settings as those in Figure 3.

Now, we turn our attention to the analysis of the deterministic Ordinary Differential Equation (ODE) corresponding to our LA with barriers and plot it in Figure 5. The trajectory of the ODE is conform with our intuition and the results of the LA run in Figure 4. The two ODE are given by:

$$\frac{dp_1}{dt} = W_1(X) = p_1(p_{max} - p_1)D_1^A(q_1) + (1 - p_1)(p_{min} - p_1)D_2^A(q_1), \tag{6.3}$$

and,

$$\frac{dq_1}{dt} = W_1(X) = p_1(p_{max} - p_1)D_1^A(q_1) + (1 - p_1)(p_{min} - p_1)D_2^A(q_1), \tag{6.4}$$

To obtain the ODE for a particular example, we need just to replace the entries of R and C in the ODE by their values. In this sense to plot the ODE trajectories we only need to know R and C and of course p_{max}.

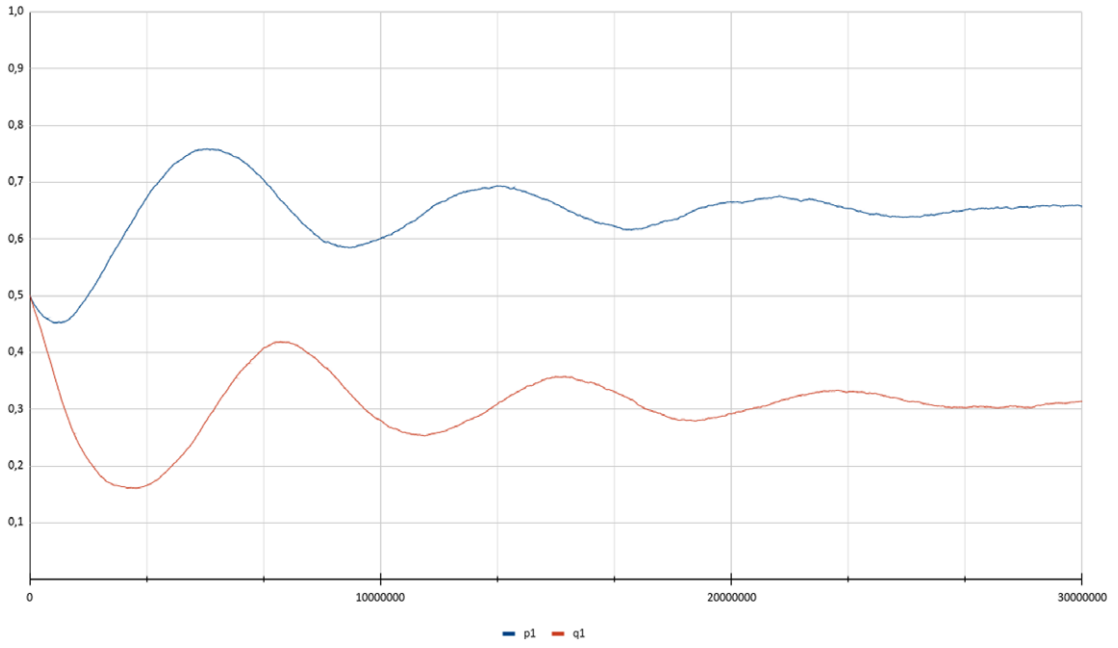


Figure 3. *Time Evolution X(t) for the case of the R matrix given by Equation (6.1) and the C matrix given by Equation (6.2) with* $p_{opt} = 0.6667$ *and* $q_{opt} = 0.3333$*, and using* $p_{max} = 0.99$ *and* $\theta = 0.00001$.

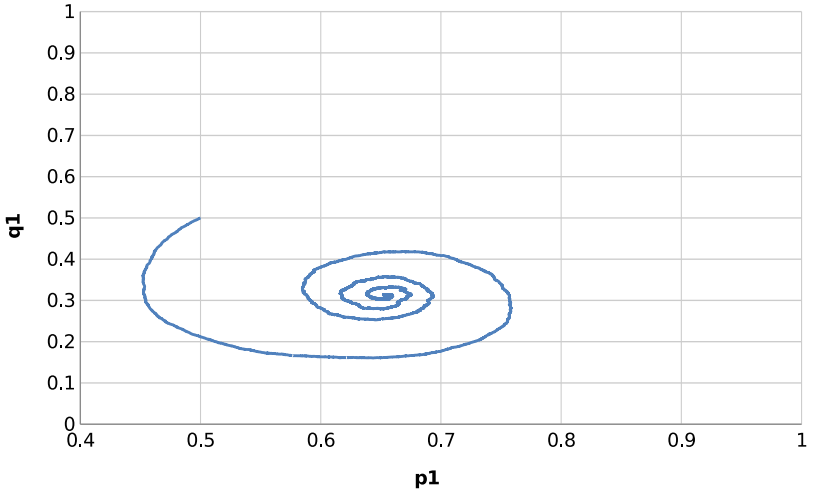


Figure 4. *Trajectory of X(t) where* $p_{opt} = 0.6667$ *and* $q_{opt} = 0.3333$*, using* $p_{max} = 0.99$ *and* $\theta = 0.00001$.

6.2. Case 2*: One pure equilibrium*

At this juncture, we shall experimentally show that the scheme possess still plausible convergence properties even in case where there is a single saddle point in pure strategies and that our proposed LA will approach the optimal pure equilibria. For this sake, we consider a case of the game where there is a single pure equilibrium which falls in the category of Case 2 with $p_{opt} = 1$ and $q_{opt} = 0$. The payoff matrices R and C for the games are given by:

Table 2. *Error for different values of θ and p_{max} for the game specified by the R matrix and the C matrix given by Equations (6.5) and (6.6)*

p_{max}	$\theta = 0.0001$	$\theta = 0.00001$
0.990	6.57×10^{-2}	6.51×10^{-2}
0.991	5.88×10^{-2}	5.82×10^{-2}
0.992	5.30×10^{-2}	5.21×10^{-2}
0.993	4.67×10^{-2}	4.64×10^{-2}
0.994	4.00×10^{-2}	4.02×10^{-2}
0.995	3.36×10^{-2}	3.38×10^{-2}
0.996	2.68×10^{-2}	2.64×10^{-2}
0.997	2.04×10^{-2}	2.08×10^{-2}
0.998	1.40×10^{-2}	1.37×10^{-2}

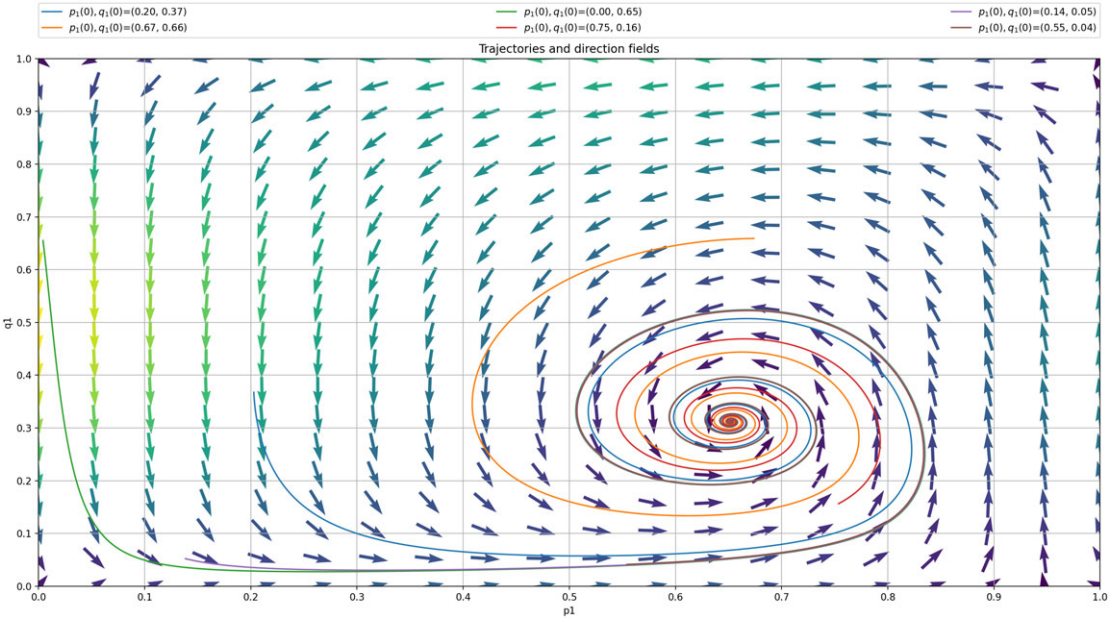


Figure 5. *Trajectory of ODE using $p_{max} = 0.99$ for case 1.*

$$R = \begin{pmatrix} 0.7 & 0.9 \\ 0.6 & 0.8 \end{pmatrix}, \tag{6.5}$$

$$C = \begin{pmatrix} 0.6 & 0.8 \\ 0.8 & 0.9 \end{pmatrix}, \tag{6.6}$$

We first show the convergence errors of our method in Table 2. As in the previous simulation for Case 1, the errors are on the order to 10^{-2} for larger values of p_{max}. We also observe that steady state error is slightly higher compared to the previous case of mixed Nash described by Equations (6.1) and (6.2) which is treated in the previous section.

We then plot the ODE for $p_{max} = 0.99$ as shown in Figure 6. According to the ODE in Figure 6, we are expecting that the LA will converge towards the attractor of the ODE which corresponds to

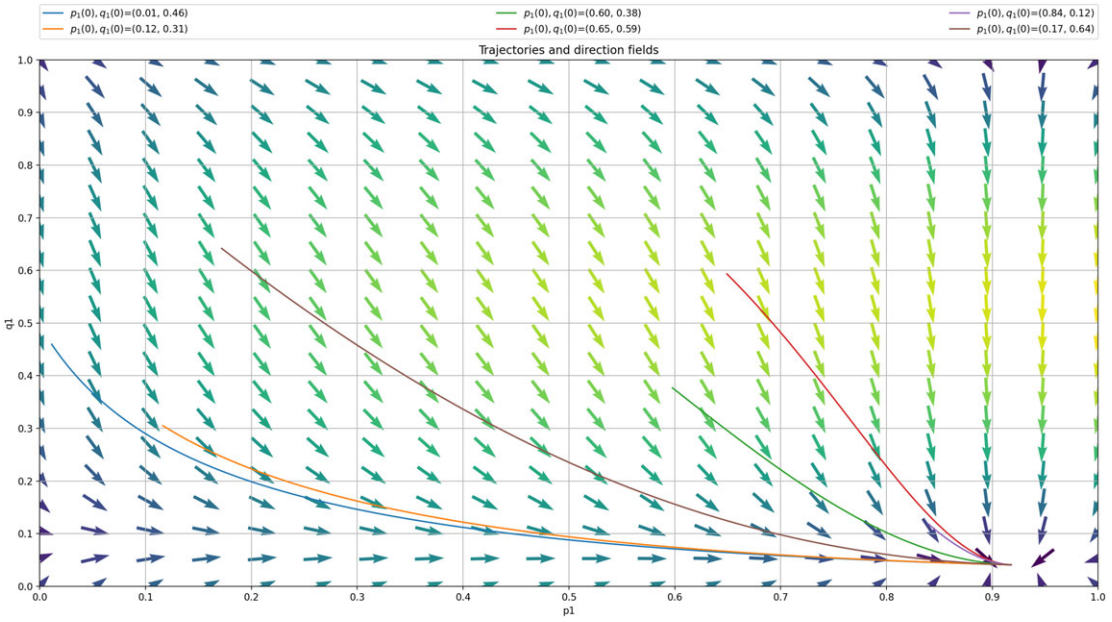


Figure 6. *Trajectory of the deterministic ODE using $p_{max} = 0.99$ for case 2.*

$(p^*, q^*) = 0.917, 0.040)$ as θ goes to zero. We see that $(p^*, q^*) = (0.917, 0.040)$ approaches $(p_{opt}, q_{opt}) = (1, 0)$ but there is still a gap between them. This is also illustrated in Figure 7 where we also consistently observe that the LA converges towards $(p^*, q^*) = (0.916, 0.041)$ after running our LA for 30 000 iterations with an ensemble of 1000 experiments. The main limitation of introducing artificial barriers whenever the optimal equilibrium is pure is choosing p_{max} even slightly away from 1 (such as 0.99), gives a large deviation in the convergence $(p^*, q^*) = (0.916, 0.041)$ which is somehow far from the optimal value $(p_{opt}, q_{opt}) = (1, 0)$. It appears as if, in the case of pure equilibrium, for slightly large p_{min}, we can get a large deviation from (p_{opt}, q_{opt}). In order to mitigate this issue, one can introduce an extra mechanism that would detect that convergence should take place into the corners of the simplex, implying a pure Nash equilibrium, and reducing p_{min} over time to ensure such convergence.

Observing the small dispersancy between $(p^*, q^*) = (0.917, 0.040)$ and $(p_{opt}, q_{opt}) = (1, 0)$ from the ODE and from the LA trajectory as shown in Figures 6 and 7 motivates us to choose even a larger value of p_{max}. Thus, we increase p_{max} from 0.99 to 0.999 and observe the expected convergence results from the ODE in Figure 8. We observe a single attraction point close of the ODE close to the pure Nash equilibrium. We can read from the ODE trajectory that $(p^*, q^*) = (0.991, 0.004)$ which is closer $(p_{opt}, q_{opt}) = (1, 0)$ than the previous case with a smaller p_{max}.

In Figure 9a, we depict the time evolution of the two components of the vector $X(t)$ using the proposed algorithm for an ensemble of 1000 runs. In the case of having a Pure Nash equilibrium, there is no oscillatory behavior as when a player assigns more probability to an action, since the other player reinforces the strategy. However, Figure 9a could mislead the reader to believe that the LA method has converged to a pure strategy for both players. In order to clarify that we are not converging to an absorbing state for the player A, we provide Figure 9b which zooms on Figure 9a around the region where the strategy of player A has converged in order to visualize that its maximum first action probability is limited by p_{max}, as per the design of our updating rule. Similarly, we zoom on the evolution of the first action probability of player B in Figure 9c. We observe that the first action instead to converging to zero as it would be if we did not have absorbing barriers, its rather converges to a small probability limited by p_{min} which approaches zero. Such propriety of evading lock in probability even for pure optimal strategies and which emanates from the ergodicity of our L_{R-I} scheme with absorbing barriers is a desirable

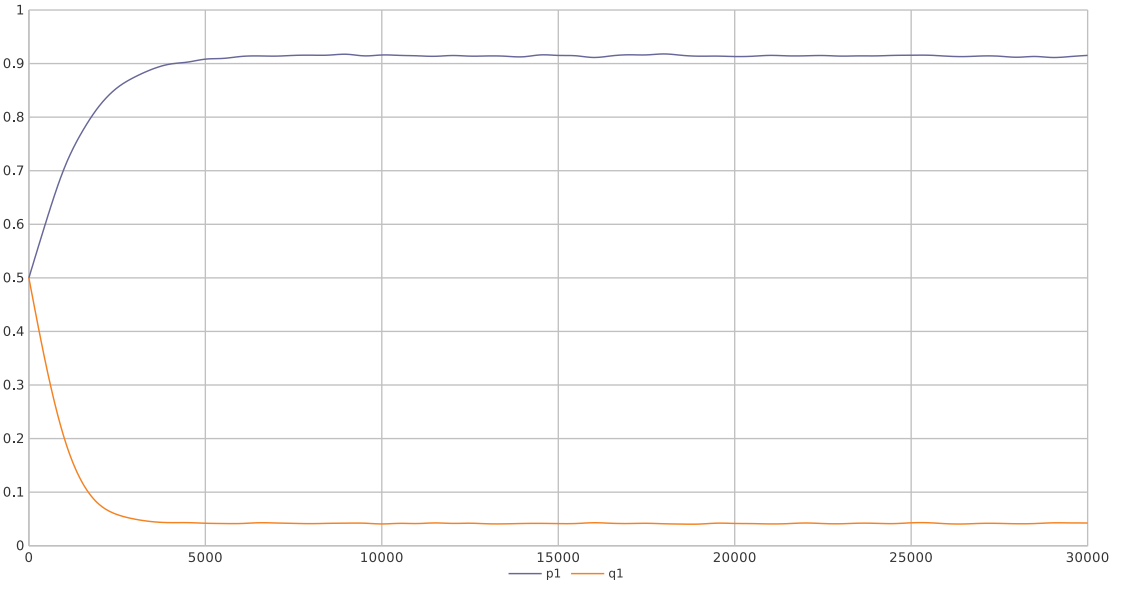


Figure 7. *Time evolution over time of X(t) for* $\theta = 0.01$ *and* $p_{max} = 0.99$ *for the case of the R matrix given Equation (6.5) and for the C matrix given by Equation (6.6).*

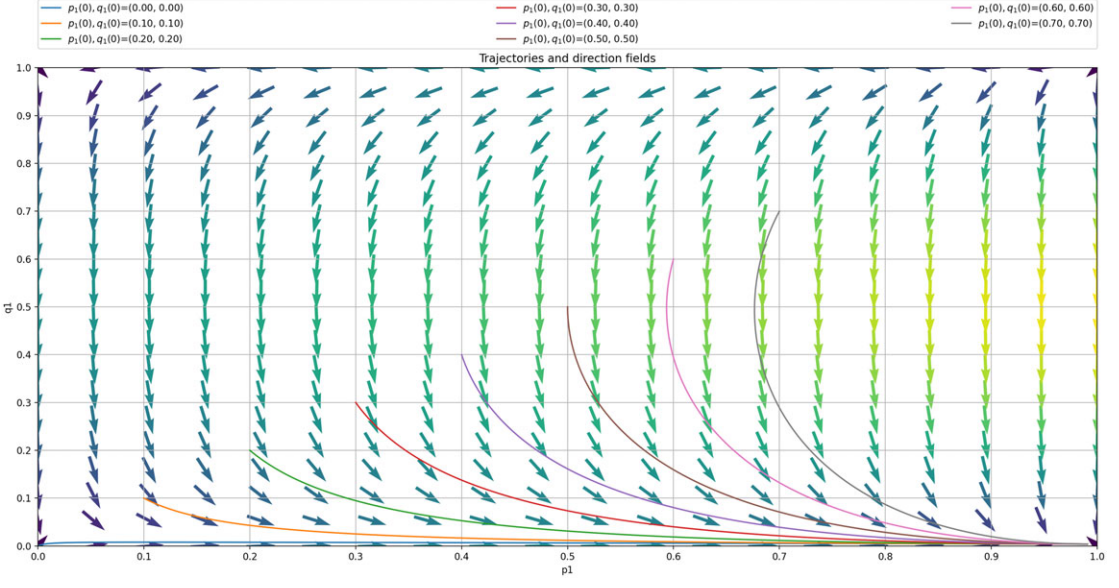
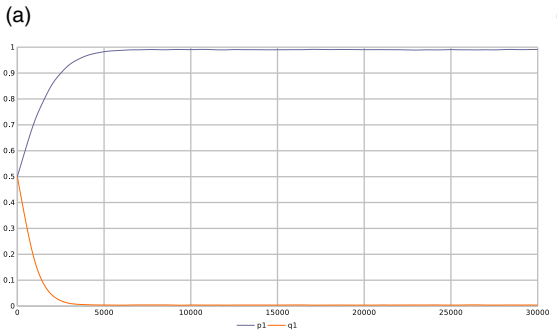


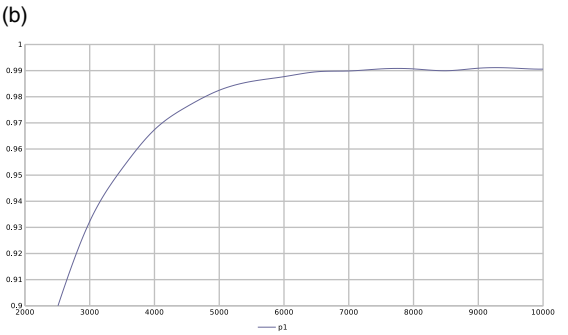
Figure 8. *Trajectory of ODE using* $p_{max} = 0.999$ *for case 2.*

property specially when the payoff matrices are time-varying and thus the optimal equilibrium point might change over time. Such a case deserves a separate study to better understand the behavior of the scheme and to also understand the effect of the tuning parameters and how to control and vary them in this case to yield a compromise between learning and forgetting stale information.

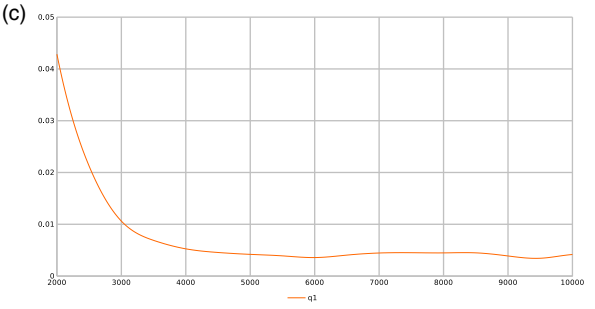
Figure 9 depicts the time evolution of the probabilities for each player, with $\theta = 0.01$, $p_{max} = 0.999$ and for an ensemble with 1000 runs.



Evolution over time of $X(t)$.



Zoomed version for player A strategy.



Zoomed version for player B strategy.

Figure 9. *The figure shows (a) the evolution over time of X(t) for* $\theta = 0.01$ *and* $p_{max} = 0.999$ *when applied to game with payoffs specified by the R matrix and the C matrix given by Equation (6.5) and Equation (6.6), and (b) is a zoomed version around player A strategy c) and is a zoomed version around player B strategy.*

6.3. Case 3: 2 Pure equilibria and 1 mixed

Now, we shall consider the last case 3.

As an instance of case 3, we consider the payoff matrices R and C given by:

$$R = \begin{pmatrix} 0.3 & 0.1 \\ 0.2 & 0.3 \end{pmatrix}, \tag{6.7}$$

$$C = \begin{pmatrix} 0.3 & 0.2 \\ 0.1 & 0.2 \end{pmatrix}, \tag{6.8}$$

In Figure 10, we plot 9 trajectories for the LA for a number of iterations is 1 000 000. We observe that depending on the initial conditions, our LA converges to one of the two pure equilibria which is usually the closest to the starting point. We have also performed extensive simulations with initial values $(0.5, 0.5)$ of the probabilities and we found that almost 50% of the time the LA converges to the Nash equilibrium close to (1,1) and 50% close to (0,0). As a future work, we would like to explore how to push the LA to favor one of the two equilibria as there is usually an equilibrium that is superior to the other, and thus it is more desirable for both players to converge to the superior Nash equilibrium.

We plot the ODE corresponding to our LA for case 3 in Figure 11. We can see two attractions points which approach the two Nash equilibria.

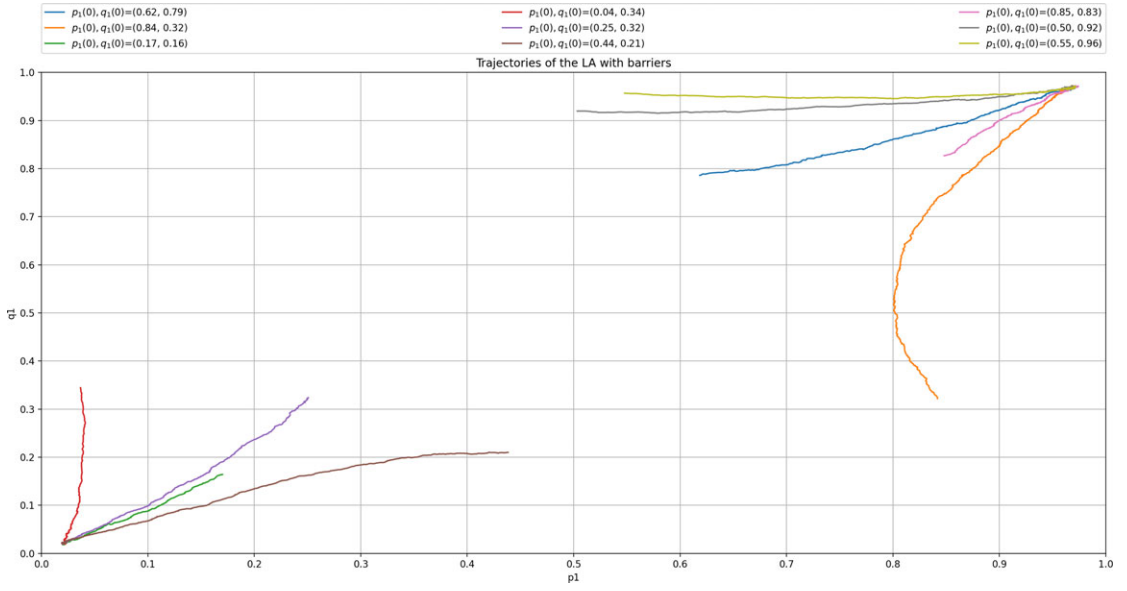


Figure 10. *9 Trajectories of the LA with barriers starting from random initial point with* $p_{max} = 0.99$ *and* $\theta = 0.0001$.

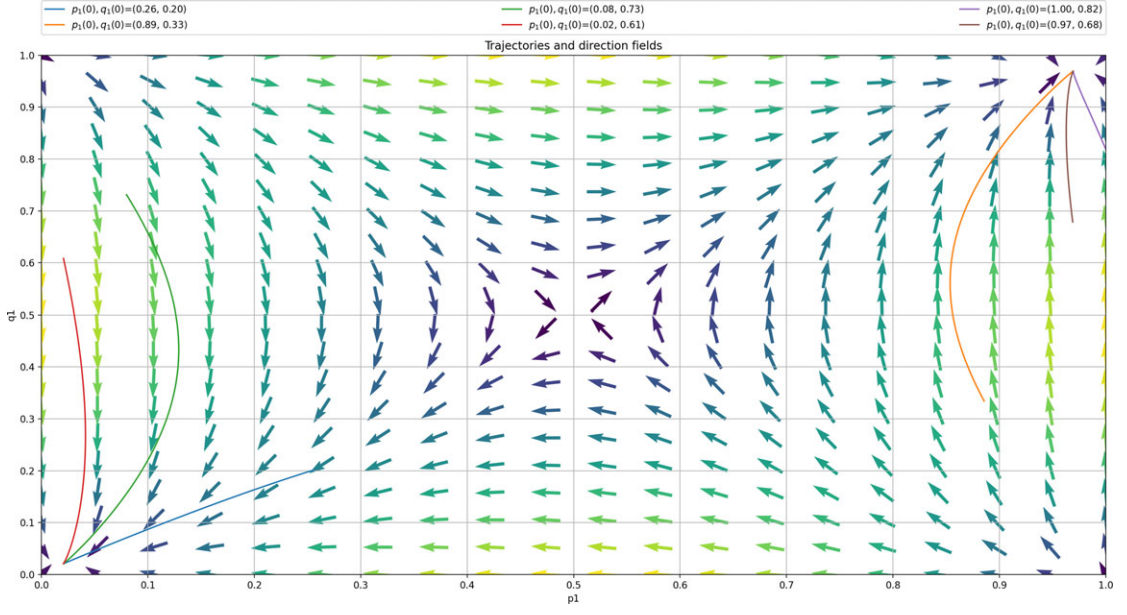


Figure 11. *Trajectory of ODE using* $p_{max} = 0.99$ *for case 3.*

Although for $p_{max} \neq 1$, w our scheme is in theory ergodic and not absorbing, this is not the case in practice as shown in the simulation reported in Table 3. In fact for $\theta = 0.0001$ and as p_{max} becomes larger or equal to 0.995, we observe that the error is zero meaning that the LA has converged already to an absorbing state! This lock in probability phenomenon is due to the limited accuracy of the machine and limitations of the random number generator. For smaller $\theta = 0.00001$, we expect that the LA will approximate better the ODE. Indeed, this is the case the absorption this time does not happen for

Table 3. *Error for different values of θ and p_{max} for the game specified by the R matrix and the C matrix given by Equations (6.7) and (6.8)*

p_{max}	$\theta = 0.0001$	$\theta = 0.00001$
0.990	3.08×10^{-2}	2.06×10^{-2}
0.991	$2,76 \times 10^{-2}$	2.76×10^{-2}
0.992	1.64×10^{-2}	2.43×10^{-2}
0.993	1.42×10^{-2}	2.12×10^{-2}
0.994	1.85×10^{-2}	1.21×10^{-2}
0.995	0.0	1.53×10^{-2}
0.996	0.0	1.21×10^{-2}
0.997	0.0	0.0
0.998	0.0	0.0
0.999	0.0	0.0

$p_{max} = 0.995$ and $p_{max} = 0.996$ as in the previous case, but happen for only p_{max} larger or equal to $p_{max} = 0.997$.

Solving the ODE for $p_{max} = 0.999$, gives two solutions, namely, $(p^*, q^*) = (0.99699397, 0.99699397)$ and $(0.00200603, 0.00200603)$ which approach $(p_{\text{opt}}, q_{\text{opt}}) = (1, 1)$ and $(p_{\text{opt}}, q_{\text{opt}}) = (0, 0)$ respectively.

Solving the ODE for $p_{max} = 0.998$, gives two solutions $(p^*, q^*) = (0.99397576, 0.99397576)$ and $(0.00402424, 0.00402424)$.

While solving the ODE for $p_{max} = 0.997$, gives $(p^*, q^*) = (0.99094517, 0.99094517)$ and $(0.00605483, 0.00605483)$.

7. Experimental results for *S*-type environments

In this section, we present the results of the experiments for the S-type learning game. We conducted several simulations similar to those presented in Section 6. The same instances of the payoff matrices R and C were used, covering the cases referred to in Section 4.

For all the experiments conducted for the S-LA, 9 trajectories were plotted for 2 000 000 iteration, with $p_{max} = 0.99$ and $\theta = 0.0001$. A general observation that we noticed when performing that the experiments is that the S-LA converges slower than the L_{R-I}. Therefore, we have doubled the number of iterations to allow the S-LA to converge in our experiments.

7.1. Case 1: Only one mixed Nash equilibrium exists

Figure 12 depicts the situation where the only Nash equilibrium that exists is a mixed one. We can easily observe that the S-LA approaches the trajectories of the ODE given in Figure 5. Please note that the ODE regardless of the LA type, whether it is L_{R-I} or $S-$LA.

7.2. Case 2: One pure equilibrium

We also examined the case where the game has a single pure equilibrium. The exhibited behavior is comparable to those reported in Section 6. The trajectory of the LA depicted in Figure 13 tightly follows the trajectories of the ODE depicted in Figure 6. As θ goes to zero, the trajectories of the LA and those of the ODE will be indistinguishable (Sastry *et al.,* 1994).

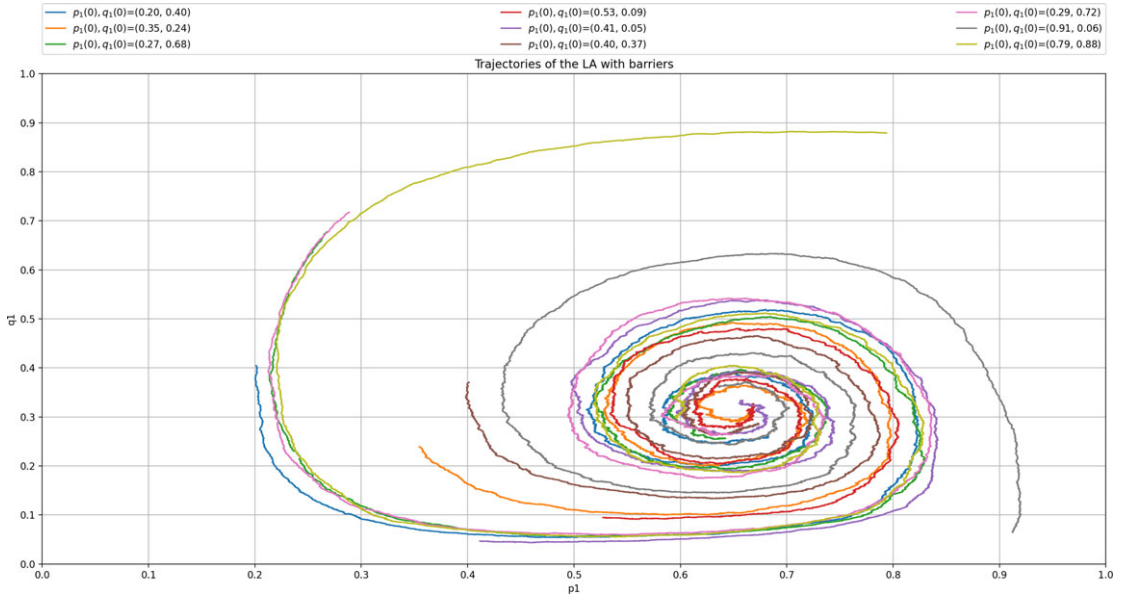


Figure 12. *Trajectory of* S-*LA using* $p_{max} = 0.99$ *and* $\theta = 0.0001$ *for case* 1.

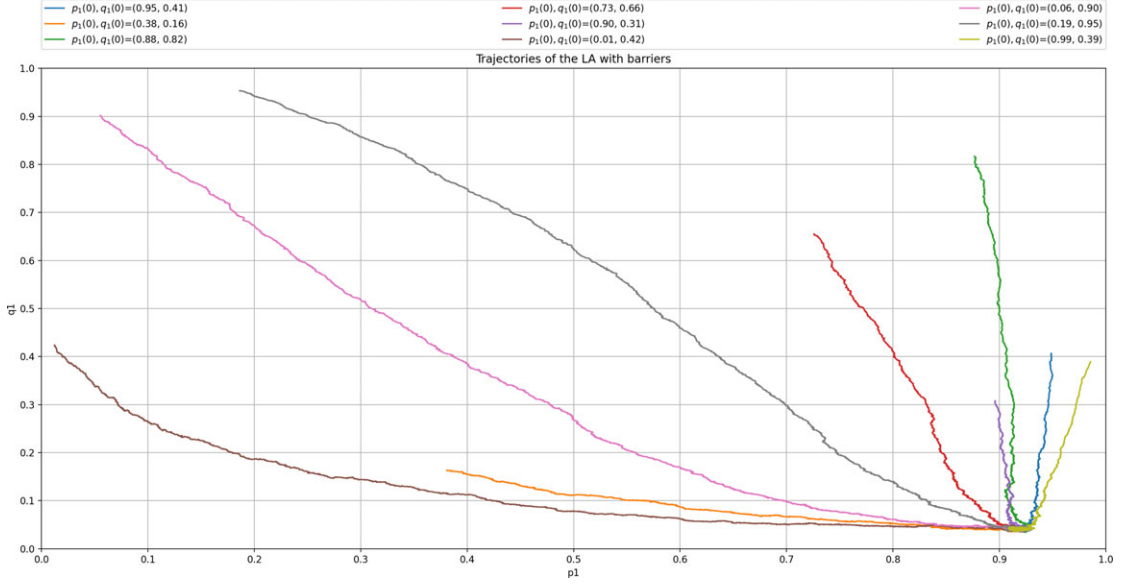


Figure 13. *Trajectory of* S-*LA using* $p_{max} = 0.99$ *and* $\theta = 0.0001$ *for case 2.*

7.3. Case 3: Two pure equilibria and one mixed

Figure 14 shows the situation where there are two pure equilibria and one mixed.

We observe that the LA converges to one of the two pure equilibria that is closest to the starting point. The *S*-LA behaves much similar to the L_{R-I} LA as shown in Figure 10. We also observe that the *S*-LA respectively converged to the Nash equilibrium close to (1, 1) and close to (0,0) approximately 50% of the time.

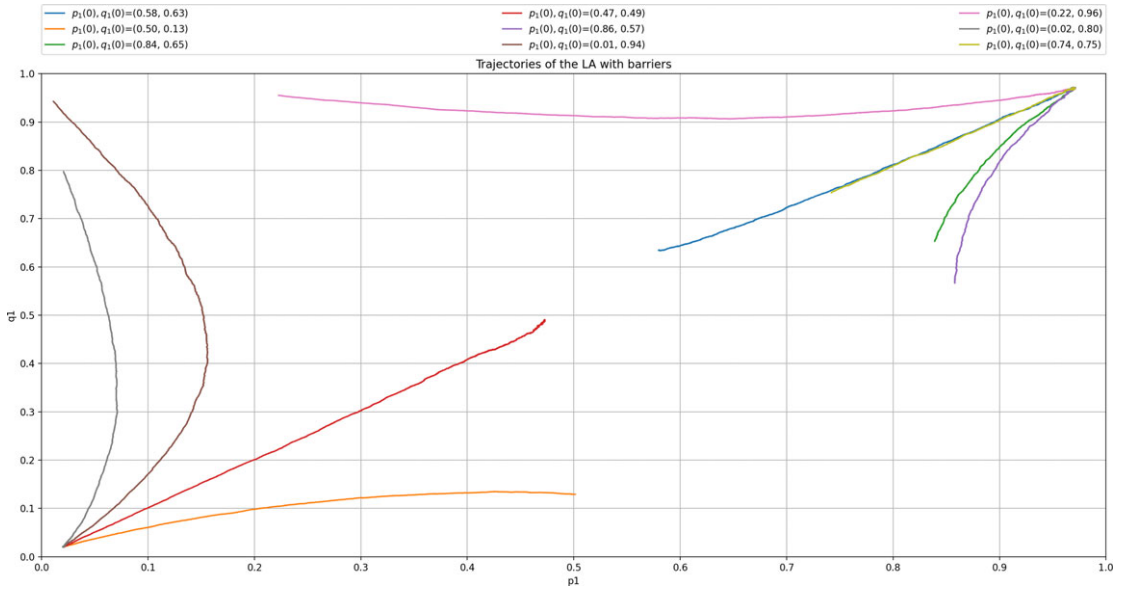


Figure 14. *Trajectory of* S-*LA using* $p_{max} = 0.99$ *and* $\theta = 0.0001$ *for case 3.*

References

Apostolopoulos, P. A., Tsiropoulou, E. E. & 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. https://doi.org/10.1109/GIIS.2018.8635770.

Bloembergen, D. et al. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research* **53**, 659–697.

Cao, H. & 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., Uyttendaele, S. & Tuyls, K. 2008. Learning to reach agreement in a continuous ultimatum game. *Journal of Artificial Intelligence Research* **33**, 551–574.

Do, C. T., et al. 2017. Game theory for cyber security and privacy. *ACM Computing Surveys (CSUR)* **50**(2), 1–37.

Fielder, A. 2020. Modelling the impact of threat intelligence on advanced persistent threat using games. In *From Lambda Calculus to Cybersecurity Through Program Analysis*. Springer, 216–232.

Gheisari, S. & 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.

Jia, L., et al. 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. https://doi.org/10.1109/ICCT.2017.8359621.

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 zerosum stochastic games with incomplete information: A unified approach. *SIAM Journal on Control and Optimization* **20**(4), 541–552.

Narendra, K. S. & Thathachar, M. A. L. 2012. *Learning Automata: An Introduction*. Courier Corporation.

Narendra, K. S. & Thathachar, M. A. L. 1974. Learning automata – A survey. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-4**(4), 323–334. https://doi.org/10.1109/TSMC.1974.5408453.

Norman, M. F. 1972. *Markov Processes and Learning Models*. Vol. 84. Academic Press.

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., et al. 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. https://doi.org/10.1109/NCA51143.2020. 9306737.

Saraydar, C. U., Mandayam, N. B. & Goodman, D. J. 2002. Efficient power control via pricing in wireless data networks. *IEEE Transactions on Communications* **50**(2), 291–303.

Sastry, P. S., Phansalkar, V. V. & Thathachar, M. A. L. 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.

Schaerf, A., Shoham, Y. & Tennenholtz, M. 1994. Adaptive load balancing: A study in multi-agent learning. *Journal of Artificial Intelligence Research* **2**, 475–500.

Sokri, A. 2020. Game theory and cyber defense. *In Games in Management Science* **280**, 335–352.
Thapa, R., et al. 2017. A learning automaton-based scheme for scheduling domestic shiftable loads in smart grids. *IEEE Access* **6**, 5348–5361.
Tian, D., et al. 2017. Self-organized relay selection for cooperative transmission in vehicular ad-hoc networks. *IEEE Transactions on Vehicular Technology* **66**(10), 9534–9549.
Vadori, V., et al. 2015. Jamming in underwater sensor networks as a Bayesian zerosum game with position uncertainty. In *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.
Viswanathan, R. & Narendra, K. S. 1974. Games of stochastic automata. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-4**(1), 131–135. https://doi.org/10.1109/TSMC.1974.5408539.
Xing, Y. & Chandramouli, R. 2008. Stochastic learning solution for distributed discrete power control game in wireless data networks. *IEEE/ACM Transactions on Networking* **16**(4), 932–944.
Yang, Z., et al. 2020. Learning automata based Q-learning for content placement in cooperative caching. *IEEE Transactions on Communications* **68**(6), 3667–3680.
Yazidi, A., et al. 2022. Solving sensor identification problem without knowledge of the ground truth using replicator dynamics. *IEEE Transactions on Cybernetics* **52**(1), 16–24. https://doi.org/10.1109/TCYB.2019.2958627.
Zhang, Z., Wang, D. & Gao, J. 2020. Learning automata-based multiagent reinforcement learning for optimization of cooperative tasks. *IEEE Transactions on Neural Networks and Learning Systems* **32**(10), 4639–4652.

Appendices

Appendix A. Norman theorem

Theorem 5. *Let $X(t)$ be a stationary Markov process dependent on a constant parameter $\theta \in [0, 1]$. Each $X(t) \in I$, where I is a subset of the real line. Let $\Delta X(t) = X(t+1) - X(t)$. The following are assumed to hold:*

1. *I is compact.*
2. *$E[\Delta X(t)|X(t) = y] = \theta w(y) + O(\theta^2)$*
3. *$Var[\Delta X(t)|X(t) = y] = \theta^2 s(y) + o(\theta^2)$*
4. *$E[\Delta X(t)^3|X(t) = y] = O(\theta^3)$ where $sup_{y\in I} \frac{O(\theta^k)}{\theta^k} < \infty$ for $K = 2, 3$ and $sup_{y\in I} \frac{o(\theta^2)}{\theta^2} \to 0$ as $\theta \to 0$.*
5. *$w(y)$ has a Lipschitz derivative in I.*
6. *$s(y)$ is Lipschitz I.*

If Assumptions (1)-(6) hold, $w(y)$ has a unique root y^* in I and $\left.\frac{dw}{dy}\right|_{y=y^*} \leq 0$ then

1. *$var[\Delta X(t)|X(0) = x] = 0(\theta)$ uniformly for all $x \in I$ and $t \geq 0$. For any $x \in I$, the differential equation $\frac{dy(\tau)}{d\tau} = w(y(t))$ has a unique solution $y(\tau) = y(\tau, x)$ with $y(0) = x$ and $E[\delta X(t)|X(0) = x] = y(t\theta) + O(\theta)$ uniformly for all $x \in I$ and $t \geq 0$.*
2. *$\frac{X(t)-y(t\theta)}{\sqrt{\theta}}$ has a normal distribution with zero mean and finite variance as $\theta \to 0$ and $t\theta \to \infty$.*