This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Solving Two-Person Zero-Sum Stochastic Games With Incomplete Information Using Learning Automata With Artificial Barriers

Anis Yazidi, *Senior Member, IEEE*, Daniel Silvestre, and B. John Oommen, *Life Fellow, IEEE*

***Abstract*— Learning automata (LA) with artificially absorbing barriers was a completely new horizon of research in the 1980s (Oommen, 1986). These new machines yielded properties that were previously unknown. More recently, absorbing barriers have been introduced in continuous estimator algorithms so that the proofs could follow a martingale property, as opposed to monotonicity (Zhang *et al.*, 2014), (Zhang *et al.*, 2015). However, the applications of LA with artificial barriers are almost nonexistent. In that regard, this article is pioneering in that it provides effective and accurate solutions to an extremely complex application domain, namely that of solving two-person zero-sum stochastic games that are provided with incomplete information. LA have been previously used (Sastry *et al.*, 1994) to design algorithms capable of converging to the game's Nash equilibrium under limited information. Those algorithms have focused on the case where the saddle point of the game exists in a pure strategy. However, the majority of the LA algorithms used for games are absorbing in the probability simplex space, and thus, they converge to an exclusive choice of a single action. These LA are thus unable to converge to other mixed Nash equilibria when the game possesses no saddle point for a pure strategy. The pioneering contribution of this article is that we propose an LA solution that is able to converge to an optimal mixed Nash equilibrium even though there may be no saddle point when a pure strategy is invoked. The scheme, being of the linear reward-inaction (L_{R-I}) paradigm, is in and of itself, absorbing. However, by incorporating artificial barriers, we prevent it from being "stuck" or getting absorbed in pure strategies. Unlike the linear reward-ϵpenalty ($L_{R-\epsilon P}$) scheme proposed by Lakshmivarahan and Narendra almost four decades ago, our new scheme achieves the same goal with much less parameter tuning and in a more elegant manner. This article includes the nontrial proofs of the theoretical results characterizing our scheme and also contains experimental verification that confirms our theoretical findings.**

***Index Terms*— Games with incomplete information, learning automata (LA), LA with artificial barriers.**

Manuscript received October 2, 2020; revised February 16, 2021; accepted May 18, 2021. The work of Daniel Silvestre was supported in part by the Portuguese Fundação para a Ciência e a Tecnologia (FCT) through the Institute for Systems and Robotics (ISR), Laboratory for Robotics and Engineering Systems (LARSyS), under Project UIDB/50009/2020, through FirePuma Project under Grant PCIF/MPG/0156/2019, and through COPELABS, University Lusófona, under Project UIDB/04111/2020. The work of B. John Oommen was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by the Natural Sciences and Engineering Council of Canada. *(Corresponding author: B. John Oommen.)*

Anis Yazidi is with the Department of Computer Science, Oslo Metropolitan University, 0190 Oslo, Norway (e-mail: anis.yazidi@oslomet.no).

Daniel Silvestre is with Dynamical Systems and Ocean Robotics Laboratory (DSOR), Department of Electrical and Computer Engineering, Lusófona University, 1749-024 Lisbon, Portugal, and also with the Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal (e-mail: dsilvestre@isr.tecnico.ulisboa.pt).

B. John Oommen is with the School of Computer Science, Carleton University, Ottawa, ON K1S 5B6, Canada, and also with the Department of IKT, University of Agder, 4879 Grimstad, Norway (e-mail: oommen@scs.carleton.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3099095.

Digital Object Identifier 10.1109/TNNLS.2021.3099095

I. Introduction

THE term learning automata (LA) denotes a whole subfield of research within adaptive systems with several books being dedicated to its study [2], [5], [6], [12], [14]. The work on LA dates to the Soviet Union in the 1960s when the mathematical giant Tsetlin *et al.* [15] devised the so-called Tsetlin machine that is a learning mechanism with finite memory. Tsetlin's learning machines were demonstrated to give birth to self-organizing behavior through collective learning. In his work, Tsetlin pioneered the Goore game, which is a distributed coordination game with limited feedback that has many practical applications, as shown by Tung and Kleinrock [16]. The early works in the field of LA, such as the Tsetlin machine, fall under the family of fixed structure LA. The mainstream of current LA research concerns the family of variable structure LA (VSLA) which, loosely speaking, differs from fixed structure LA in the fact that they operate with a probability vector that is updated dynamically over time. In fixed structure LA, the choice is governed by a transition matrix whose transitions do not depend on time and that describes how the internal states of the LA are updated based on the environment's feedback. The term LA was coined for the first time by Narendra and Thathachar [6].

Markovian Representations of LA: LA can also be characterized by their Markovian representations. They thus fall into one of two families, being either ergodic or those that possess absorbing barriers [8]. Such a characterization is crucial to the tenets of this article. Absorbing automata have underlying Markov chains that get absorbed or locked into a barrier state. Sometimes, this can occur even after a relatively small, finite number of iterations. The classic references [2], [5], [6], [12], [14] report numerous LA families that contain such absorbing barriers. On the other hand, as these same references explain, the literature has also reported scores of ergodic automata, which converge in distribution. In these cases, the asymptotic distribution of the action probability vector converges to a value that is independent of its initial vector. Absorbing LA are usually designed to operate in stationary environments. As opposed to these, ergodic LA are preferred for nonstationary environments, namely those that possess time-dependent reward probabilities. These characterizations, and their corresponding implications for game playing, will be explained presently.

Continuous or Discretized VSLA: VSLA can also be characterized as being continuous or discretized. This depends on the values that the action probabilities can take. Continuous LA allow the action probabilities to assume any value in the interval [0, 1]. Such algorithms have a relatively slow rate of convergence. The problem with continuous LA is that they approach a goal but never reach there. This was mitigated in the 1980s by introducing the concept of discretization, where if an action probability was close enough to zero or unity, it could jump to that endpoint in a single step. This also rendered the LA to have a faster convergence because one could increase their speeds of convergence by incorporating this phenomenon [3], [4], [9]. This is implemented by constraining the action selection probability to be one of a finite number of values in the interval $[0, 1]$. By incorporating discretization, almost all of the reported VSLA of the continuous type have also been discretized [9], [10], [19].

LA With Artificially Absorbing Barriers: LA with artificially introduced absorbing barriers were a novelty in the 1980s. These yielded machines, which had properties that were previously unknown. This was due to the fact that a discretized machine, even though it was ergodic, could be rendered absorbing by forcing the machine to stay at one of the absorbing barriers [8]. Ironically, this simple step introduced families of new LA, with properties that were previously unknown. For example, ADL$_{R-P}$ and ADL$_{I-P}$ are absorbing versions of their corresponding ergodic counterparts but have been proven to be ϵ-optimal in all random environments. This phenomenon, of including artificially absorbing barriers, has been recently applied to the family of pursuit LA [18].

Estimator LA With Artificial Barriers: The concept of introducing absorbing barriers is also central to the proofs of estimator algorithms. For three decades, these pursuit algorithms were "proven" to be ϵ-optimal by virtue of the monotonicity property. However, recently, these proofs have been shown to be flawed. To remedy this, absorbing barriers have been introduced in continuous estimator algorithms so that the proofs could follow a martingale property, as opposed to monotonicity. Consequently, Zhang *et al.* [18]–[20] have shown that one can invoke this weaker property, namely, the martingale property, by artificially providing such an absorbing barrier. Thus, whenever an action probability is close enough to unity, the LA is forced to jump to this absorbing barrier.

Applications of LA: LA have boasted scores of applications. These include theoretical problems, such as the graph partitioning problem. They have been used in controlling intelligent vehicles. When it concerns neural networks and hidden Markov models, Meybodi *et al.* have used them in adapting the former, and others have applied them in training the latter. Network call admission, traffic control, and quality-of-service routing have been resolved using LA, while others have also found applications in tackling problems involving network and communications issues. Apart from these, the entire field of LA and stochastic learning has had a myriad of applications listed in the reference books [2], [5], [6], [14]. In the interest of the page-limit constraints, the citations to these applications are not included. However, they can be easily found by executing a simple search, and many are included in the above benchmark references.

Game Playing With LA: While artificially introduced barriers have been shown to have powerful theoretical and design implications, the applications of them are few. This is where this article finds its place—it presents one such application. LA have also been used to resolve stochastic games with incomplete information. This article pioneers a merge of the above two issues. First of all, we present a mechanism by which LA can be augmented with artificial barriers, but unlike the state of the art, these barriers are nonabsorbing. We then proceed to use these to play zero-sum games with incomplete information. Games of this type were studied four decades ago for scenarios when the game matrix had a saddle point using traditional L_{R-I} and L_{R-P} LA [13]. Our results generalize those when the game does not possess the Nash equilibrium. Rather, we propose the nontrivial use of LA with artificial nonabsorbing barriers to resolve such games. This article contains the theoretical results and those from simulations using the corresponding benchmark games.

Landscape of Our Present Work: In this article, we propose an algorithm addressing zero-sum games, which can be generalized to nonzero-sum games in a manner similar to the principle by which the method in [1] was generalized in [17]. In the latter, Xing and Chandramouli [17] proved that the linear reward$-\epsilon$penalty ($L_{R-\epsilon P}$) algorithm, devised in [1], is able to work in nonzero-sum games. Thus, without further elaborating on this,[1] our results are generalizable to nonzero-sum games.

Since the game is zero-sum, the outcomes are either a loss for player A, with reward -1, and the corresponding win for player B with value $+1$, or the converse for the case of a win for player A. We emphasize that this is a limited information game where each player is unaware of both the mixed strategy and the selected action of the other player. The available information to each player is whether its action resulted in a win or a loss. The reader should note that either/both players might not even be aware of the existence of another player and be working with the assumption that he is playing against nature, as in the classical multiarmed bandit algorithms. However, if both players learn using our algorithm based on the assumption that they are operating in an adversarial environment, we show that they will both converge to the desired equilibrium. Our proposed scheme has players adjusting their strategy whenever it obtains a "win" for that round. This conforms to the linear reward-inaction (L_{R-I}) paradigm, described in detail, presently. It is thus, unarguably, radically different from the mechanism proposed by Lakshmivarahan and Narendra [1], where the probability updates are performed upon receiving both reward and penalty responses, which thus renders changes to occur at every time instant.

[1] Some preliminary unpublished work is being conducted for extending this work to nonzero-sum games.

Objective and Contribution of This Article: Based on the above discussion, one can summarize the objective of this article as to study the behavior of a nonabsorbing barrier-based L_{R-I} mechanism in a stochastic zero-sum game played by two players A and B, with two actions each, as earlier done in [1]. Each player uses an LA to decide his strategy, where the only received feedback from the environment is the reward of the joint actions of both players. The game is played iteratively and the players are able to revise their mixed strategies.

Applications of the Proposed Method: Learning within the context of games has a natural application in the realm of game theory. However, in the context of multiagent systems (MASs), this has been shown to be suitable for the cooperative control of robotic systems [21]. In such a design, it is assumed that the mission can be fully described as a potential game, where the utility function measures how well the nodes in the network are complying with the objectives. Nevertheless, having robots converging to pure strategies means that the network designer is favoring exploitation and disregarding exploration. If the environment changes and causes a different payoff matrix, the agent would be locked into repeatedly playing the same strategy. Moreover, this assumes that the utility must be known and deterministic. Therefore, instead of designing application-specific algorithms, the proposed learning algorithm can be used to address problems in cooperative control such as the so-called "rendezvous" problem for a fleet of robots [26], [27], the desynchronization of the use of a shared medium [22], [23], and a consensus algorithm to have the agents agree on a common value [24] or to solve distributed computation such as the PageRank [25], by only considering the current stochastic payoff.

It is also pertinent to mention that the mechanism that we propose here can be used by the agents to learn how to act if the payoff corresponds to how successful they are in following the objectives of the "mission." Much can be said about this, but we terminate these discussions here in the interest of brevity and due to space limitations. However, with respect to future research, it is wise to mention that the question of whether they can be applied to synchronization, as in the analysis of the family of so-called "Firefly" algorithms, is yet open.

A. Notation Used

Most of the notations that we use are well established from the theory of matrices and in the field of LA [2], [6], and stating them would trivialize this article. However, we mention that apart from the well-established notations used in these areas, we will use the notation that the conditional expectation of some variable v with respect to w is written as $E[v|w]$ and the partial derivative of a variable $v(t)$ with respect to time t is denoted by $(\partial v(t)/\partial t)$.

II. Game Model

To initiate discussions, we formalize the game model that is being investigated. Let $P(t) = [p_1(t)\ p_2(t)]^{\mathsf{T}}$ 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 $Q(t) = [q_1(t)\ q_2(t)]^{\mathsf{T}}$. The extension to more than two actions per player is straightforward following the method analogous to what was used by Papavassilopoulos [11], which extended the work of Lakshmivarahan and Narendra [1].

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 is pure strategy. Specifically, if $(\alpha_A(t), \alpha_B(t)) = (i, j)$, the probability of gain for player A is determined by d_{ij}, as formalized in [1]. We thus construct a matrix with the set of probabilities $D = [d_{ij}]$, $1 \le i \le 2$, which is the so-called payoff matrix associated with the game.

The matrix D is given by

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \tag{1}$$

where all the entries are probabilities.

Clearly, the actual game matrix G is given by $g_{ij} = 2\,d_{ij} - 1$, with entries in the interval $[-1, 1]$. Without loss of generality, player A corresponds to the row player, whereas B is the column player. Furthermore, when referring to a "gain," we are seeing this from the perspective of player A.

In zero-sum 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 the stochastic form of a zero-sum game. The "zero-sum" property implies that at any time t, there is only one winning player.[2]

In the interest of completeness, we present the original scheme proposed in [1] based on the $L_{R-\epsilon P}$ rule. It uses two parameters θ_R and θ_P as the learning rates associated with the reward and penalty responses, respectively. When player A gains at time instant t by playing action i, he updates his mixed strategy as

$$\begin{aligned} p_i(t+1) &= p_i(t) + \theta_R(1 - p_i(t)) \\ p_s(t+1) &= p_s(t) - \theta_P\, p_s(t) \quad \text{for} \quad s \ne i. \end{aligned}$$

However, if player A loses after using action i, his mixed strategy is updated by the following:

$$\begin{aligned} p_i(t+1) &= p_i(t) - \theta_P\, p_i(t) \\ p_s(t+1) &= p_s(t) + \theta_R(1 - p_s(t)) \quad \text{for} \quad s \ne i. \end{aligned}$$

The exact update mechanism for player B is obtained by replacing the corresponding $p(t)$ by $q(t)$ and by recalling that a gain for A maps onto a loss scenario for player B. We now introduce our novel solution that is proposed to learn a new mixed strategy.

[2]The results inferred from this article can be extended to nonzero-sum games. However, for the sake of simplicity, we only consider the case of zero-sum games.

III. LA Algorithm Based on L_{R-I} With Artificial Barriers

A. Nonabsorbing Artificial Barriers

We have earlier seen that an ergodic LA can be made absorbing by artificially rendering the end states to become absorbing. This was briefly addressed above. However, what has not been discussed in the literature is a strategy by which a scheme which is, in and of itself, absorbing, can be rendered to be ergodic. In other words, the LA are allowed to move within the probability simplex by utilizing an absorbing scheme. However, when it enters an absorbing barrier, the scheme is forced to go back into the simplex in order to render it to be ergodic. No such scheme has ever been reported in the literature, and the advantage of having such a scheme is that one does not get locked into a suboptimal absorbing barrier. Rather, we can permit it to move around so that it can migrate stochastically toward an optimal mixed strategy. This is, precisely, what we shall do.

B. Nonabsorbing Game Playing

We now present our strategic LA-based game algorithm together with a formal analysis that demonstrates the convergence to the saddle points of the game even if the saddle point corresponds to a mixed Nash equilibrium. Our LA solution is based on the L_{R-I} scheme, but as alluded to earlier, it has been modified in order to nontrivially provide nonabsorbing barriers. The proof of convergence is based on Norman's theory for learning processes characterized by small learning steps [6], [7].

Considering that $p_{\max}$ denotes an artificial barrier, we use the notation that $p_{\min} = 1 - p_{\max}$. We further constrain the probability for each action by restricting it, by design, to belong to the interval $[p_{\min}, p_{\max}]$ if $p_1(0)$ and $q_1(0)$ are initially chosen to belong to the same interval. If the outcome from the environment is a gain at a time t for action $i \in \{1, 2\}$, the update rule is given by

$$\begin{aligned} p_i(t+1) &= p_i(t) + \theta(p_{\max} - p_i(t)) \\ p_s(t+1) &= p_s(t) + \theta(p_{\min} - p_s(t)) \quad \text{for} \quad s \neq i. \end{aligned} \tag{2}$$

The reader will observe that this update mechanism is identical to the well-established linear schemes, except that $p_{\min}$ and $p_{\max}$ replace the values zero and unity, respectively. When the player receives a loss, the probabilities are not updated, which translates into

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

The update rules for the mixed strategy $q(t+1)$ are defined in a similar fashion by recalling the dichotomy that whenever player A gains, it corresponds to a loss for player B and vice versa. Analogous to the L_{R-I} paradigm, mixed strategies are not changed in the case of a loss.

We now proceed to analyze the convergence properties of the proposed algorithm. To aid in the analysis, we identify the Nash equilibrium of the game by the pair $(p_{\text{opt}}, q_{\text{opt}})$. To render the presentation to be less cumbersome, we divide the analysis into two cases.

1) Case 1 [Only One Mixed Nash Equilibrium Case (No Saddle Point in Pure Strategies)]: 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 easily shown to be the following:

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

where $L = (d_{11}+d_{22}) - (d_{12}+d_{21})$. Without loss of generality, we assume that

$$d_{11} > \max\{d_{12}, d_{21}\} \qquad \text{and } d_{22} > \max\ \{d_{12}, d_{21}\}. \tag{4}$$

Notice that the above inequalities are not restrictive, as games not satisfying them can be mapped in a symmetric manner by reindexing the actions of the players and/or the indices of the players.

2) Case 2 (There Is a Saddle Point in Pure Strategies): The case where the game matrix has saddle points in pure strategies corresponds to either: 1) $d_{11} > d_{12}$, $d_{12} < d_{21}$, $d_{21} > d_{22}$, and $d_{22} < d_{11}$ or 2) in the symmetric case, where $d_{11} < d_{12}$, $d_{12} > d_{21}$, $d_{21} < d_{22}$, and $d_{22} > d_{11}$.

Since the other cases can be proven in identical manners, in the interest of brevity, we consider only the case where

$$d_{21} < d_{11} < d_{12}. \tag{5}$$

In this case, $p_{\text{opt}} = 1$ and $q_{\text{opt}} = 1$. The other subcases within Case 2 can be obtained by reindexing the actions of the players and/or the indices of the players, as in Case 1.

Let the vector $X(t) = \left[p_1(t)\ q_1(t)\right]^\mathsf{T}$. We introduce the notation that $\Delta X(t) = X(t+1) - X(t)$. We also represent the conditional expected value operator by $\mathbb{E}[\cdot|\cdot]$. Using these, we claim the next theorem.

Theorem 1: Consider a zero-sum game with a payoff matrix as in (1) and a learning algorithm defined by (2) and (3) 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 $X^* = \left[p_1^*\ q_1^*\right]^\mathsf{T}$ satisfying the following conditions.

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

Proof: Let us first compute the conditional expected value[3] 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 (2).

[3]Computing the "expected value of the increment" is a standard procedure in the theory of LA. This is because the increment, in and of itself, is a random variable, which is sometimes positive and sometimes negative. Quantifying the latter is not possible due to the randomness of the updating rule. However, the conditional expected value of the increment can be determined, whence (by invoking the "Law of the Unconscious Statistician"), one can determine the expected value of the increment itself.

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

$$\begin{aligned}
W_1(X) &\\
&= p_1 q_1 d_{11}(p_{\max} - p_1) + p_1(1-q_1)d_{12}(p_{\max} - p_1)\\
&\quad +(1-p_1)q_1 d_{21}(p_{\min} - p_1)\\
&\quad +(1-p_1)(1-q_1)d_{22}(p_{\min} - p_1)\\
&= p_1[q_1 d_{11} + (1-q_1)d_{12}](p_{\max} - p_1)\\
&\quad +(1-p_1)[q_1 d_{21} + (1-q_1)d_{22}](p_{\min} - p_1)\\
&= p_1(p_{\max} - p_1)D_1^A(q_1) + (1-p_1)(p_{\min} - p_1)D_2^A(q_1)
\end{aligned} \tag{6}$$

where

$$D_1^A(q_1) = q_1 d_{11} + (1-q_1)d_{12} \tag{7}$$
$$D_2^A(q_1) = q_1 d_{21} + (1-q_1)d_{22}. \tag{8}$$

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

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

Similarly, we can get

$$\begin{aligned}
W_2(X) &\\
&= q_1 p_1(1-d_{11})(p_{\max} - q_1)\\
&\quad +q_1(1-p_1)(1-d_{12})(p_{\max} - q_1)\\
&\quad +(1-q_1)p_1(1-d_{21})(p_{\min} - q_1)\\
&\quad +(1-q_1)(1-p_1)(1-d_{22})(p_{\min} - q_1)\\
&= q_1[p_1(1-d_{11}) + (1-p_1)(1-d_{12})](p_{\max} - q_1)\\
&\quad +(1-q_1)[p_1(1-d_{21}) + (1-p_1)(1-d_{22})](p_{\min} - q_1)\\
&= q_1(p_{\max} - q_1)\big[1 - D_1^B(p_1)\big]\\
&\quad +(1-q_1)(p_{\min} - q_1)\big[1 - D_2^B(p_1)\big]
\end{aligned} \tag{9}$$

where

$$D_1^B(p_1) = p_1 d_{11} + (1-p_1)d_{21} \tag{10}$$
$$D_2^B(p_1) = p_1 d_{12} + (1-p_1)d_{22}. \tag{11}$$

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

$$\begin{aligned}
W_2(X) &\\
&= q_1(1-q_1)\big(1 - D_1^B(p_1)\big) - q_1 p_{\min}\big(1 - D_1^B(p_1)\\
&\quad +(1-q_1)p_{\min}\big(1 - D_2^B(p_1) - q_1(1-q_1)\big(1 - D_2^B(p_1)\big)\\
&= -q_1(1-q_1)\big[D_1^B(p_1) - D_2^B(p_1)\big]\\
&\quad +p_{\min}\big[-q_1\big(1 - D_1^B(p_1)\big) + (1-q_1)\big(1 - D_2^B(p_1)\big)\big]\\
&= -q_1(1-q_1)\big[D_1^B(p_1) - D_2^B(p_1)\big]\\
&\quad +p_{\min}\big[q_1 D_1^B(p_1) - (1-q_1)D_2^B(p_1) + (1-2q_1)\big].
\end{aligned} \tag{12}$$

We need to address the two identified cases. Consider Case 1), where there is only a single mixed equilibrium. According to (4), we get

$$\begin{aligned}
D_{12}^A(q_1) &= D_1^A(q_1) - D_2^A(q_1)\\
&= (d_{12} - d_{22}) + Lq_1.
\end{aligned} \tag{13}$$

Given that $L > 0$, since $d_{11} > d_{12}$ and $d_{22} > d_{21}$, $D_{12}^A(q_1)$ is an increasing function of q_1 and

$$\begin{cases}
D_{12}^A(q_1) < 0, & \text{if } q_1 < q_{\text{opt}}\\
D_{12}^A(q_1) = 0, & \text{if } q_1 = q_{\text{opt}}\\
D_{12}^A(q_1) > 0, & \text{if } q_1 > q_{\text{opt}}.
\end{cases} \tag{14}$$

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

$$\begin{aligned}
W_1\left(\begin{bmatrix}0\\ q_1\end{bmatrix}\right) &= p_{\min}D_2^A(q_1) > 0\\
W_1\left(\begin{bmatrix}1\\ q_1\end{bmatrix}\right) &= -p_{\min}D_1^A(q_1) < 0.
\end{aligned} \tag{15}$$

Since $W_1(X)$ is quadratic with a negative second derivative with respect to p_1 and the inequalities in (15) 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

$$\begin{cases}
p_1 < \dfrac{1}{2}, & \text{if } q_1 < q_{\text{opt}}\\[2mm]
p_1 = \dfrac{1}{2}, & \text{if } q_1 = q_{\text{opt}}\\[2mm]
p_1 > \dfrac{1}{2}, & \text{if } q_1 > q_{\text{opt}}.
\end{cases} \tag{16}$$

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

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

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

In the proof of Case 1), we have utilized the fact that for small enough $p_{\min}$, the learning algorithm admits 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) 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, $((\partial W(X^*))/\partial x)$, is negative definite.

Proof: We start the proof by writing the explicit format for

$$\frac{\partial W(X)}{\partial X} = \begin{bmatrix} \dfrac{\partial W_1(X)}{\partial p_1} & \dfrac{\partial W_1(X)}{\partial q_1}\\[2mm] \dfrac{\partial W_2(X)}{\partial p_1} & \dfrac{\partial W_2(X)}{\partial q_1}\end{bmatrix}$$

and then computing each of the entries as follows:

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

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 $((\partial W(X^*))/\partial X) \approx ((\partial W(X_{\text{opt}}))/\partial X)$. We now subdivide the analysis in the two cases identified as above, which are equivalent to the following.

1) *Case 1:* No saddle point in pure strategies.
2) *Case 2:* There is a saddle point in pure strategies.

3) Case 1 (No Saddle Point in Pure Strategies): In this case, we have

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

which makes

$$\frac{\partial W_1\big(X_{\text{opt}}\big)}{\partial p_1} = -2p_{\min} D_1^A\big(q_{\text{opt}}\big). \tag{17}$$

Similarly, we can compute

$$\frac{\partial W_1\big(X_{\text{opt}}\big)}{\partial q_1} = (1-2p_{\min})p_{\text{opt}}\big(1-p_{\text{opt}}\big)L. \tag{18}$$

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

$$\frac{\partial W_2\big(X_{\text{opt}}\big)}{\partial p_1} = -(1-2p_{\min})q_{\text{opt}}\big(1-q_{\text{opt}}\big)L \tag{19}$$

and

$$\frac{\partial W_2\big(X_{\text{opt}}\big)}{\partial q_1} = -2p_{\min}\big(1-D_1^B\big(p_{\text{opt}}\big)\big) \tag{20}$$

resulting in (21), as shown at the bottom of the page.

The matrix given in (21) satisfies

$$\det\left(\frac{\partial W\big(X_{\text{opt}}\big)}{\partial x}\right) > 0, \ \text{trace}\left(\frac{\partial W\big(X_{\text{opt}}\big)}{\partial x}\right) < 0 \tag{22}$$

which implies that the 2×2 matrix is negative definite.

4) Case 2 (There Is a Saddle Point in Pure Strategies): In Theorem 1, Case 2 reduces to considering $q_{\text{opt}} = 1$ and $p_{\text{opt}} = 1$.

Computing the entries of the matrix for this case yields

$$\frac{\partial W_1\big(X_{\text{opt}}\big)}{\partial p_1} = -(d_{11}-d_{21}) - p_{\min}(d_{11}+d_{21}) \tag{23}$$

and

$$\frac{\partial W_1\big(X_{\text{opt}}\big)}{\partial q_1} = -p_{\min}(d_{11}-d_{12}). \tag{24}$$

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

$$\frac{\partial W_2\big(X_{\text{opt}}\big)}{\partial p_1} = p_{\min}(d_{11}-d_{21}) \tag{25}$$

and

$$\frac{\partial W_2\big(X_{\text{opt}}\big)}{\partial q_1} = (d_{11}-d_{12}) - p_{\min}(2-d_{11}-d_{12}) \tag{26}$$

resulting in (27), as shown at the bottom of the next page.

The matrix in (27) satisfies

$$\det\left(\frac{\partial W\big(X_{\text{opt}}\big)}{\partial X}\right) > 0, \ \text{trace}\left(\frac{\partial W\big(X_{\text{opt}}\big)}{\partial X}\right) < 0 \tag{28}$$

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

Theorem 3: Let V be the von Neumann value of the game given by matrix D. Let $\mathbf{p}(t) = [p_1, p_2]$ and $\mathbf{q}(t) = [q_1, q_2]$. For a sufficiently small $p_{\min}$ approaching 0, $\eta(t)$ converges to V as $\theta \to 0$ where

$$\eta(t) \triangleq E\big[\mathbf{p}(t)\big]DE\big[\mathbf{q}^T(t)\big]. \tag{29}$$

Proof: The proof of these results requires a classic result due to Norman [7], given in the Appendix, in the interest of completeness.

The convergence of $\big[E(p_1(t)) \ E(q_1(t))\big]$ to $\big[p_{\text{opt}}^* \ q_{\text{opt}}^*\big]$ is a consequence of this theorem. Interestingly enough, this theorem is a classical fundamental result that has been used to prove many of the convergence results in LA. It has, for example, been used by the seminal paper by Lakshmivarahan and Narendra [1] to derive similar convergence properties of the $L_{R-\epsilon P}$, applicable for the same game settings as ours. Indeed, it is easy to verify that Assumptions (1)–(6) required for Norman's result are satisfied. Thus, by further invoking Theorem 1 and Theorem 2, the result follows. □

We conclude this section by mentioning that like all LA algorithms, the computational complexity of our scheme is linear in the size of the action probability vector. This is because, at the most, all the action probabilities are updated at every time instant.

For the benefit of future researchers, we believe that it will be profitable to record the hurdles we encountered in

$$\frac{\partial W\big(X_{\text{opt}}\big)}{\partial X} = \begin{bmatrix} -2p_{\min}D_1^A\big(q_{\text{opt}}\big) & (1-2p_{\min})p_{\text{opt}}\big(1-p_{\text{opt}}\big)L \\ -(1-2p_{\min})q_{\text{opt}}\big(1-q_{\text{opt}}\big)L & -2p_{\min}\big(1-D_1^B\big(p_{\text{opt}}\big)\big)\end{bmatrix}. \tag{21}$$

this research. The breakthrough came when we were able to devise/design LA systems that possessed no-absorbing barriers. In other words, it involved the concept of forcing the LA back into the probability space when it was close enough to the absorbing barriers. This was a phenomenon that we had not earlier seen in the literature. The consequent problem was the analysis. The underlying Markov process could not be easily analyzed using the properties of absorbing chains. Neither could it be trivially modeled as an ergodic chain converging to an equilibrium distribution. The analysis that we presented here came as a "brain wave," and once the building blocks were established, everything naturally seemed to fall in place. These few sentences, requested by an anonymous referee, should clarify the difficulties encountered in this research, in order to show that the present research is pioneering and this is not a trivial extension of existing methodologies.

IV. Simulations

In this section, we present simulations to confirm the abovementioned theoretical properties of the proposed learning algorithm. In the interest of maintaining benchmarks, we adopt the same examples as those reported in [1]. Also, by using different instances of the payoff matrix D, we are able to experimentally cover the two cases referred to in Section III. Again, we refer to those cases as Cases 1 and 2, as done in [1].

A. Convergence in Case 1

We consider an instance of the game where only one mixed Nash equilibrium exists, i.e., there is no saddle point in pure strategies. We adopt the same game matrix D as in [1] given by

$$D = \begin{pmatrix} 0.6 & 0.2 \\ 0.35 & 0.9 \end{pmatrix} \tag{30}$$

which admits $p_{\text{opt}} = 0.5789$ and $q_{\text{opt}} = 0.7368$.

In order to eliminate the Monte Carlo error, we ran our scheme for 5×10^6 iterations and report the error in Table I for different values of $p_{\max}$ and θ as the difference between X_{opt} and the mean over time of $X(t)$ after convergence.[4] An important remark is that the error decreases as $p_{\max}$ approaches 1 (i.e., when $p_{\min} \to 0$). Please observe that in this case, we have particularly chosen to not let $p_{\max}$ be unity. If we allow it to be precisely unity, it would mean that we would not require an artificial barrier close to unity (for example, between 0.990 and 0.999 as in Table I). In fact, for $p_{\max} = 0.999$ and $\theta = 0.001$, the method achieves an error of 2.1621×10^{-3}, and further reducing $\theta = 0.0001$ leads to an error of 1.6820×10^{-3}.

To better visualize the scheme, Fig. 1 shows the evolution over time of the mixed strategies for both players (given by

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

TABLE I

Error for Different Values of θ and p_{MAX} When $p_{\text{OPT}} = 0.5789$ and $q_{\text{OPT}} = 0.7368$ for the Game Specified by the D Matrix Given by (30). The Point That You Have Raised Is Pertinent

p_{max}	$\theta = 0.001$	$\theta = 0.0001$
0.999	2.1621×10^{-3}	1.6820×10^{-3}
0.998	2.5456×10^{-3}	1.7059×10^{-3}
0.997	3.7380×10^{-3}	2.2332×10^{-3}
0.996	3.4007×10^{-3}	2.0155×10^{-3}
0.995	5.4371×10^{-3}	3.7888×10^{-3}
0.994	5.5962×10^{-3}	4.2018×10^{-3}
0.993	7.3416×10^{-3}	5.4064×10^{-3}
0.992	7.7319×10^{-3}	7.8230×10^{-3}
0.991	9.6127×10^{-3}	6.7476×10^{-3}
0.990	9.3467×10^{-3}	9.6713×10^{-3}

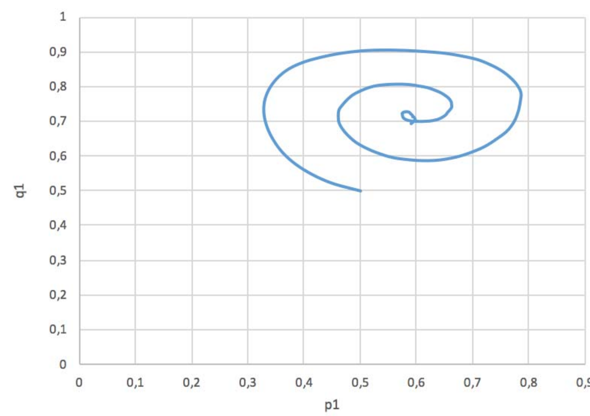


Fig. 1. Time evolution of $[p_1(t), q_1(t)]^\top$ for the same settings as in Fig. 2.

$X(t)$) for an ensemble of 1000 runs using $\theta = 0.01$ and $p_{\max} = 0.999$.

The trajectory of the ensemble allows us to perceive the mean evolution of the mixed strategies. The spiral pattern is caused by one of the players adapting to the strategy being used by the other before the former learns by overcorrecting its strategy. The procedure is continued leading to smaller corrections until the players reach the Nash equilibrium.

The abovementioned behavior can also be visualized in Fig. 2 that presents the trajectory for a single experiment with $p_{\max} = 0.99$ and $\theta = 0.00001$ over 3×10^7 steps. The described oscillatory behavior is attenuated as the players play for more iterations. The reader should particularly observe that a larger value of θ will cause more steady-state error (as specified in Theorem 1), but it will also perturb this behavior as the nodes take larger updates whenever they win. On the other hand, further decreasing θ results in a smaller error of the stationary point of the method but also decreases the convergence speed. This well-established inherent tradeoff between the steady-state error and rate of convergence can be better visualized by comparing Fig. 1 with $\theta = 0.001$ against Fig. 3 for a smaller value of $\theta = 10^{-5}$.

$$\frac{\partial W(X_{\text{opt}})}{\partial X} = \begin{bmatrix} -(d_{11} - d_{21}) - p_{\min}(d_{11} + d_{21}) & -p_{\min}(d_{11} - d_{12}) \\ p_{\min}(d_{11} - d_{21}) & (d_{11} - d_{12}) - p_{\min}(2 - d_{11} - d_{12}) \end{bmatrix}. \tag{27}$$

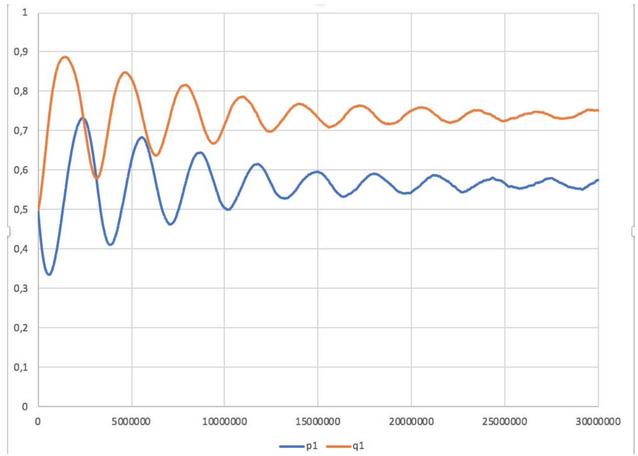


Fig. 2. Trajectory of $X(t)$ for the case of the D matrix given by (30) with $p_{\text{opt}} = 0.5789$ and $q_{\text{opt}} = 0.7368$ and using $p_{\max} = 0.99$ and $\theta = 0.00001$.

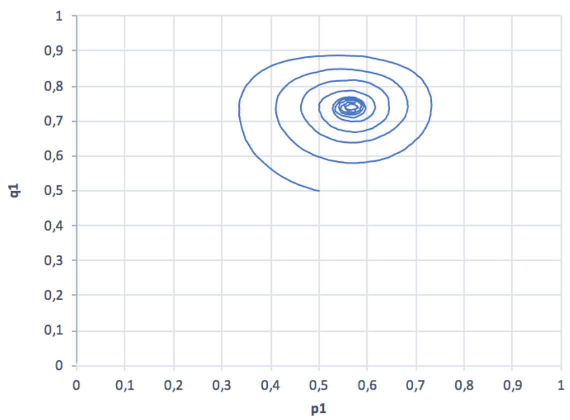


Fig. 3. Time evolution of $X(t)$ where $p_{\text{opt}} = 0.5789$ and $q_{\text{opt}} = 0.7368$ using $p_{\max} = 0.99$ and $\theta = 0.00001$.

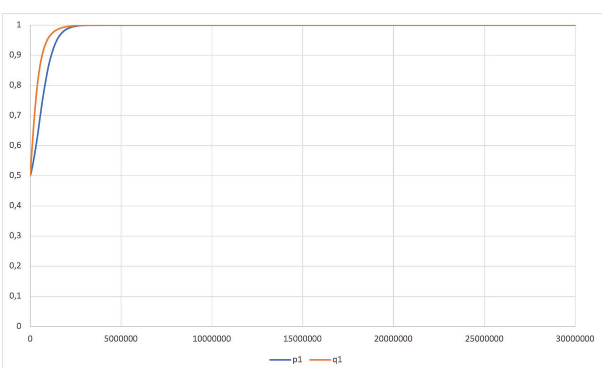


Fig. 4. Trajectory of $X(t)$ for the case of the D matrix given by (30) and using an absorbing barrier $p_{\max} = 1$ and $\theta = 0.00001$.

Furthermore, in order to clearly emphasize the necessity of using an artificial barrier, we have specifically repeated the same experiment except that we have included an absorbing barrier instead, i.e., set $p_{\max} = 1$. The result is shown in Fig. 4. In this case, we expect that the scheme enters an absorbing barrier. Since it is impossible for the human eye to detect whether or not we entered an absorbing barrier by merely examining the graph, we also manually checked the log of

TABLE II

ERROR FOR DIFFERENT VALUES OF θ AND p_{MAX} FOR D_1

p_{max}	$\theta = 0.001$	$\theta = 0.0001$
0.999	2.1073×10^{-3}	2.0971×10^{-3}
0.998	4.2753×10^{-3}	4.4573×10^{-3}
0.997	6.6147×10^{-3}	6.9025×10^{-3}
0.996	8.7588×10^{-3}	8.9192×10^{-3}
0.995	1.0815×10^{-2}	1.1044×10^{-2}
0.994	1.3424×10^{-2}	1.2894×10^{-2}
0.993	1.5005×10^{-2}	1.5415×10^{-2}
0.992	1.7347×10^{-2}	1.7805×10^{-2}
0.990	1.9772×10^{-2}	1.9670×10^{-2}
0.99	2.2516×10^{-2}	2.2548×10^{-2}

TABLE III

ERROR FOR DIFFERENT VALUES OF θ AND p_{MAX} FOR D_2

P_{max}	$\theta = 0.001$	$\theta = 0.0001$
0.999	6.2439×10^{-3}	4.8359×10^{-3}
0.998	9.8350×10^{-3}	9.9923×10^{-3}
0.997	1.5470×10^{-2}	1.3583×10^{-2}
0.996	1.8769×10^{-2}	2.1704×10^{-2}
0.995	2.5573×10^{-2}	2.4587×10^{-2}
0.994	3.1426×10^{-2}	2.8006×10^{-2}
0.993	3.6112×10^{-2}	3.5181×10^{-2}
0.992	3.7508×10^{-2}	4.0789×10^{-2}
0.991	4.6255×10^{-2}	4.2545×10^{-2}
0.99	4.8299×10^{-2}	4.5069×10^{-2}

the experiment and verified that the probabilities became exactly unity after around 6 798 000 iterations. Although the theoretical convergence should have occurred in the limit and not after a finite number of iterations, the machine limited accuracy rounded the probabilities to unity after this juncture.

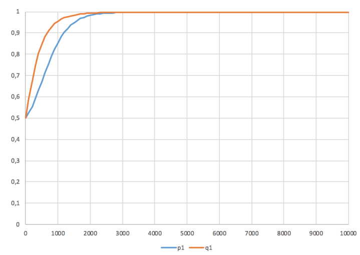
B. Pure Equilibrium

In order to assess the performance of the proposed learning algorithm on cases with a pure equilibrium, we consider two instances of games falling in the category of Case 2 with $p_{\text{opt}} = 1$ and $q_{\text{opt}} = 1$. The payoff matrices D_1 and D_2 for the two games are given by

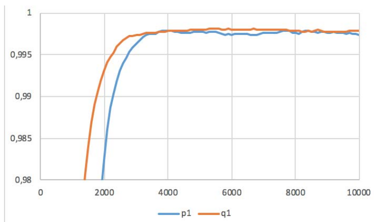
$$D_1 = \begin{pmatrix} 0.6 & 0.8 \\ 0.35 & 0.9 \end{pmatrix}$$
$$D_2 = \begin{pmatrix} 0.7 & 0.9 \\ 0.6 & 0.8 \end{pmatrix}.$$

We first show the convergence errors of our method for both games D_1 and D_2 in Tables II and III, respectively. As in the previous simulation for Case 1, the errors are on the order to 10^{-3} for larger values of $p_{\max}$. However, given that our algorithm uses artificial barriers to prevent absorbing states, the error is lower bounded by $p_{\min}$. A similar issue is present in game D_2. We have also included this simulation since it is a more challenging game to learn with our method for a larger steady-state error, even for very small values of θ.

In Fig. 5, 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

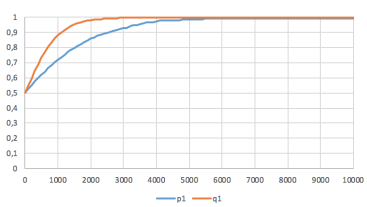


(a)

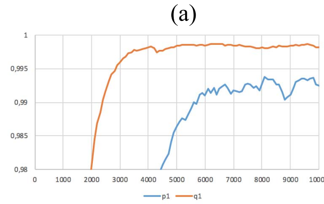


(b)

Fig. 5. (a) Evolution over time of $X(t)$ for $\theta = 0.01$ and $P_{\max} = 0.999$ when applied to game with payoffs D_1. (b) Zoomed version around the steady-state value.



(a)



(b)

Fig. 6. (a) Evolution over time of $X(t)$ for $\theta = 0.01$ and $P_{\max} = 0.999$ when applied to game with payoffs D_2. (b) Zoomed version around the steady-state value.

reinforces the strategy. However, Fig. 5(a) could lead make one believe that the LA method has converged to a pure strategy. Fig. 5(b) zooms around the point where the strategies have converged to showcase that their maximum value is limited by $p_{\max}$, as per the design of our updating rule. This mechanism is particularly favorable to prevent players from converging to absorbing states for games with time-varying payoff matrices. However, the study of such a scenario is left for future research, namely that of determining how to design $p_{\max}$ and θ that represent a good tradeoff between learning the game and adapting to a change in the payoffs.

Game D_2 presents a harder challenge for our method as we can see from its larger steady-state error. Fig. 6 shows the time evolution of the probabilities for each player when the algorithm is applied to D_2 with $\theta = 0.01$, $P_{\max} = 0.999$ and for an ensemble with 1000 runs.

The main remark regarding the results presented in Fig. 6(a) is that the convergence is much slower when compared to game D_1. This behavior is governed by the fact that the entries in matrix D_2 are closer to each other, unlike in D_1 where there is a clear disadvantage for player A when selecting action 2. There will, thus, be much fewer updates for player A where it increases the probability of action 2 in game D_1—which is not pertinent in game D_2. Fig. 6(b) further emphasizes this remark by displaying a zoom, depicting a sharper change in the probabilities in comparison with the smooth behavior in game D_1.

C. Comparisons With Related Works

Now that we have explained our new techniques and established its theoretical basis, we continue this discussion with a brief comparison with some of the prior art.[5]

First of all, one possible alternative when the payoff matrix is known can be to consider the problem as that of designing a local controller for each of the agents. One alternative is to explore the results in [28] and further investigated in [29]. However, this is only possible when D is known, which is not the scenario that we have assumed in this article.

It is not out of place to review some of the relevant works in game theory that are not necessarily solved using LA. However, in the interest of space and brevity, we will not aim at submitting an extensive review of the field of game theory. Rather, we shall cite some pertinent works inasmuch as our main contribution in this article centers on advancing the field of LA-based solutions and, more specifically, those dealing with the special case of games with "incomplete information."

There are different variants of zero-sum stochastic games in the literature. Flesch *et al.* [30] have proven the general result that every positive zero-sum stochastic game with countable state and action spaces admits a value if at least one player has a finite action space. A similar value-existence result was obtained for a zero-sum stochastic game [31] with a continuous-time Markov chain, where the players have also the possibility of stopping the game. Ziliotto [32] considered weighted-average stochastic games, that is, stochastic games where Player 1 maximizes (in expectation) a fixed weighted average of the sequence of rewards. A so-called pumping algorithm was proposed in [33] for two-person zero-sum undiscounted stochastic games. Other approaches map the game onto a dynamic programming problem and solve it based on Bellman's optimal principle using concepts from the theory of optimal control [34].

The research on game theoretical learning with incomplete information [13] is scarce in the literature. Incomplete information is a taxonomy used within the field of LA games to denote the case where the players do not observe the action of the opponent players and where each player does not know his own payoff function but only observes outcomes in the form of a reward or a penalty. The informed reader would observe that the games we deal with in this article fall under this class of games characterized by such incomplete information.

The case of incomplete information is not usually treated by the mainstream of literature in game theory. Indeed,

[5]We are thankful to the anonymous referee who requested this comprehensive section. It significantly adds to the quality of this article.

the main game learning algorithms available in the literature, such as fictitious play [35], best response dynamics, and gradient-based learning approaches, deal with the complete knowledge case, where the players know their own payoff function and observe the history of the choices of other players. Fictitious play is one of the few algorithms that can converge to a mixed strategy equilibrium by maintaining various frequency-based beliefs over the action of the opponent players, and using those beliefs, for deciding the next action to be played. However, the fictitious play algorithm cannot solve our settings of incomplete information.

When it comes to games with incomplete information, different algorithms have been suggested, which are based on the Bush–Mosteller learning paradigm. Notable examples include the ones reported in [36]–[39]. All those algorithms share a similar structure to our proposed LA, in particular, and to VSLA in general, in the sense, that the action probabilities are updated iteratively based on feedback and using some learning parameter. In this context, one should note that many LA models can be seen as extensions of Bush–Mosteller learning. However, the difference with our work is the fact that all the aforementioned algorithms have absorbing barriers. The theoretical analyses of the convergence to pure equilibria for this family of algorithms rely usually on the theory replicator dynamics.

Another family of methods that can operate with limited information includes the Erev–Roth algorithm [40] and the Arthur algorithm [41], which in turn can be seen as a variant of the Erev–Roth algorithm. The Erev–Roth algorithm is alternatively called the Erev–Roth payoff matching algorithm and relies on updating the so-called "propensity" for each action, which is, loosely speaking, the cumulative payoff for that action. Thereafter, each action is played in a manner proportional to its corresponding relative propensity. The Erev–Roth algorithm is one of the few examples of limited-information game learning approaches that converge to unique mixed strategy equilibria. However, the Erev–Roth algorithm requires storing the entire history of rewards and penalties for each action. Furthermore, we have not been able to locate any research study that reports the analysis of the Erev–Roth algorithm for the case of our stochastic zero-sum game. We therefore opted to implement it for our game. Experimental results (not reported here, in the interest of not distracting from the main contribution of this article) show that it neither converges to the desired equilibrium nor does it possess consistent convergence results.

D. Real-Life Application Scenarios

One referee had requested a brief explanation of a complex environment, or different scenarios in a game, by which we could utilize our newly proposed solution. We agree that providing an insightful discussion could be insightful for interested readers and active researchers. This, of course, can be open-ended, but to satisfy the referee, we present the following brief example.

Our learning algorithm admits potential applications in many security games as well as in communication problems. The intersection between game theory and security is an emerging field of research. Algorithms that can converge to mixed equilibria are of great interest to the security community because mixed equilibria are usually preferred over pure ones. In fact, randomization gives less predictive ability to the attacker to guess the deployed strategy of the defender [42]. For instance, let us take a repetitive game involving a jammer and a transmitter, which, in turn, constitute our players [43]. The jammer aims to disturb and block communication between a transmitter and its associated receiver. The transmitter can choose the channel over which his message is communicated, while the jammer chooses a channel to attack. We suppose that the outcome is stochastic depending on the choice of the attacker (jammer) and defender (transmitter) and the stochastic characteristics of the channel. Both the jammer and transmitter can observe whether the attack was successful or not, and for instance, this common observation can be due to the receiver acknowledging the correct reception of the message over a wireless channel that both the attacker and jammer can overhear. Thus, the game is stochastic zero-sum.

V. Conclusion

The theoretical applications LA with artificially absorbing barriers have been reported since the 1980s [8] and, more recently, in Estimator LA [18]–[20]. This article pioneers the study of LA with artificial nonabsorbing barriers. LA have been previously used [13] to design algorithms capable of converging to the game's Nash equilibrium under limited information. The majority of the LA algorithms used for games are absorbing in the probability simplex space, and they converge to an exclusive choice of a single action. These LA are, thus, unable to converge to other mixed Nash equilibria when the game possesses no saddle point for a pure strategy. As opposed to these, we propose an LA solution that is able to converge to an optimal mixed Nash equilibrium even though there may be no saddle point when a pure strategy is invoked. The scheme is inherently of the absorbing L_{R-I} paradigm. However, by introducing reflecting barriers, we prevent it from being "stuck" or getting absorbed in pure strategies. Unlike the linear reward-ϵpenalty ($L_{R-\epsilon P}$) scheme proposed in [1], our new scheme achieves the same goal with much less parameter tuning and in a more elegant manner.

As far as know, our method is only the second reported algorithm in the literature capable of finding mixed strategies whenever no saddle point exists for pure strategies. If a saddle point exists for pure strategies, the scheme converges to a near-optimal solution close to the pure strategies in the probability simplex. This article includes the nontrial proofs of the theoretical results characterizing the convergence and stability of the algorithm. These are presented and illustrated through simulations for benchmark games presented in the literature.

With regard to future work, we believe that it will be useful in real-life applications that can be modeled using such game-like behavior.

Appendix
Norman Theorem

Theorem 4: 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) $E\big[\Delta X(t)^3|X(t) = y\big] = O(\theta^3)$. where $sup_{y\in I}\big(O(\theta^k)/\theta^k\big) < \infty$ for $K = 2, 3$ and $sup_{y\in I}\big(o(\theta^2)/\theta^2\big) \to 0$ 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 $(dw/dy)\Big|_{y=y^*} \leq 0$; then, the following conditions hold.

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 $(dy(\tau)/d\tau) = w(y(t))$ 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) $\big((X(t) - y(t\theta))/\sqrt{\theta}\big)$ has a normal distribution with zero mean and finite variance as $\theta \to 0$ and $t\theta \to \infty$.

Acknowledgment

We are very grateful for the feedback from the anonymous Referees of the original submission. Their input significantly improved the quality of this final version.

References

[1] S. Lakshmivarahan and K. S. Narendra, "Learning algorithms for two-person zero-sum stochastic games with incomplete information: A unified approach," *SIAM J. Control Optim.*, vol. 20, no. 4, pp. 541–552, 1982.

[2] S. Lakshmivarahan, *Learning Algorithms Theory and Applications: Theory and Applications*. New York, NY, USA: Springer, 2012.

[3] J. K. Lanctot and B. J. Oommen, "On discretizing estimator-based learning algorithms," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 2, pp. 1417–1422, 1991.

[4] J. K. Lanctot and B. J. Oommen, "Discretized estimator learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. 22, no. 6, pp. 1473–1483, Nov./Dec. 1992.

[5] K. Najim and A. S. Poznyak, *Learning Automata: Theory and Applications*. Amsterdam, The Netherlands: Elsevier, 2014.

[6] K. S. Narendra and M. A. Thathachar, *Learning Automata: An Introduction*. North Chelmsford, MA, USA: Courier Corporation, 2012.

[7] M. F. Norman, *Markov Processes and Learning Models*, vol. 84. New York, NY, USA: Academic, 1972.

[8] B. Johnoommen, "Absorbing and ergodic discretized two-action learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. 16, no. 2, pp. 282–293, Mar. 1986.

[9] B. J. Oommen and J. K. Lanctot, "Discretized pursuit learning automata," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 4, pp. 931–938, Jul./Aug. 1990.

[10] B. J. Oommen and M. Agache, "Continuous and discretized pursuit learning schemes: Various algorithms and their comparison," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 31, no. 3, pp. 277–287, Jun. 2001.

[11] G. P. Papavassilopoulos, "Learning algorithms for repeated bimatrix Nash games with incomplete information," *J. Optim. Theory Appl.*, vol. 62, no. 3, pp. 467–488, 1989.

[12] A. Rezvanian, A. M. Saghiri, S. M. Vahidipour, M. Esnaashari, and M. R. Meybodi, *Recent Advances in Learning Automata*, vol. 754. New York, NY, USA: Springer, 2018.

[13] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst., Man Cybern.*, vol. 24, no. 5, pp. 769–777, May 1994.

[14] M. A. L. Thathachar and P. S. Sastry, *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Boston, MA, USA: Kluwer, 2003.

[15] M. L. Tsetlin *et al.*, *Automaton Theory and Modeling of Biological Systems*. New York, NY, USA: Academic, 1973.

[16] B. Tung and L. Kleinrock, "Using finite state automata to produce self-optimization and self-control," *IEEE Trans. Parallel Distrib. Syst.*, vol. 7, no. 4, pp. 439–448, Apr. 1996.

[17] Y. Xing and R. Chandramouli, "Stochastic learning solution for distributed discrete power control game in wireless data networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 932–944, Aug. 2008.

[18] X. Zhang, O. C. Granmo, B. J. Oommen, and L. Jiao, "A formal proof of the ε-optimality of absorbing continuous pursuit algorithms using the theory of regular functions," *Appl. Intell.*, vol. 41, pp. 974–985, May 2014.

[19] X. Zhang, B. J. Oommen, O. C. Granmo, and L. Jiao, "A formal proof of the ε-optimality of discretized pursuit algorithms," *Appl. Intell.*, vol. 44, pp. 282–294, 2016, doi: 10.1007/s10489-015-0670-1.

[20] X. Zhang, B. J. Oommen, and O. C. Granmo, "The design of absorbing Bayesian pursuit algorithms and the formal analyses of their ε-optimality," *Pattern Anal. Appl.*, vol. 20, no. 3, pp. 797–808, 2015.

[21] J. R. Marden, G. Arslan, and J. S. Shamma, "Cooperative control and potential games," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 39, no. 6, pp. 1393–1407, Dec. 2009, doi: 10.1109/TSMCB.2009.2017273.

[22] D. Silvestre, J. P. Hespanha, and C. Silvestre, "Resilient desynchronization for decentralized medium access control," *IEEE Control Syst. Lett.*, vol. 5, no. 3, pp. 803–808, Jul. 2021, doi: 10.1109/LCSYS.2020.3005819.

[23] D. Silvestre, J. Hespanha, and C. Silvestre, "Desynchronization for decentralized medium access control based on gauss-seidel iterations," in *Proc. Amer. Control Conf. (ACC)*, Jul. 2019, pp. 4049–4054, doi: 10.23919/ACC.2019.8814471.

[24] D. Silvestre, J. P. Hespanha, and C. Silvestre, "Broadcast and gossip stochastic average consensus algorithms in directed topologies," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 2, pp. 474–486, Jun. 2019, doi: 10.1109/TCNS.2018.2839341.

[25] D. Silvestre, J. Hespanha, and C. Silvestre, "A PageRank algorithm based on asynchronous gauss-seidel iterations," in *Proc. Annu. Amer. Control Conf. (ACC)*, Milwaukee, WI, USA, Jun. 2018, pp. 484–489, doi: 10.23919/ACC.2018.8431212.

[26] R. Ribeiro, D. Silvestre, and C. Silvestre, "A rendezvous algorithm for multi-agent systems in disconnected network topologies," in *Proc. 28th Medit. Conf. Control Autom. (MED)*, Sep. 2020, pp. 592–597, doi: 10.1109/MED48518.2020.9183093.

[27] R. Ribeiro, D. Silvestre, and C. Silvestre, "Decentralized control for multi-agent missions based on flocking rules," *CONTROLO* (Lecture Notes in Electrical Engineering), vol. 695, J. A. Gonçalves, M. Braz-César, and J. P. Coelho, Eds. Cham, Switzerland: Springer, 2021, pp. 445–454, doi: 10.1007/978-3-030-58653-9_43.

[28] Y. Wu and R. Lu, "Output synchronization and L_2-gain analysis for network systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2105–2114, Dec. 2018, doi: 10.1109/TSMC.2017.2754544.

[29] Y. Wu, A. Isidori, R. Lu, and H. K. Khalil, "Performance recovery of dynamic feedback-linearization methods for multivariable nonlinear systems," *IEEE Trans. Autom. Control*, vol. 65, no. 4, pp. 1365–1380, Apr. 2020, doi: 10.1109/TAC.2019.2924176.

[30] J. Flesch, A. Predtetchinski, and W. Sudderth, "Positive zero-sum stochastic games with countable state and action spaces," *Appl. Math. Optim.*, vol. 82, pp. 499–516, Nov. 2018.

[31] C. Pal and S. Saha, "Continuous-time zero-sum stochastic game with stopping and control," *Operations Res. Lett.*, vol. 48, no. 6, pp. 715–719, Nov. 2020.

[32] B. Ziliotto, "A tauberian theorem for nonexpansive operators and applications to zero-sum stochastic games," *Math. Operations Res.*, vol. 41, no. 4, pp. 1522–1534, Nov. 2016.

[33] E. Boros, K. Elbassioni, V. Gurvich, and K. Makino, "A potential reduction algorithm for two-person zero-sum mean payoff stochastic games," *Dyn. Games Appl.*, vol. 8, no. 1, pp. 22–41, Mar. 2018.

[34] K. Du, R. Song, Q. Wei, and B. Zhao, "A solution of two-person zero sum differential games with incomplete state information," in *Advances in Neural Networks—ISNN 2019* (Lecture Notes in Computer Science), vol. 11554, H. Lu, H. Tang, and Z. Wang, Eds. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-22796-8_46.

[35] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, Nov. 2002.

[36] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, vol. 77, no. 1, pp. 1–14, Nov. 1997.
[37] L. R. Izquierdo, S. S. Izquierdo, N. M. Gotts, and J. G. Polhill, "Transient and asymptotic dynamics of reinforcement learning in games," *Games Econ. Behav.*, vol. 61, no. 2, pp. 259–276, Nov. 2007.
[38] A. S. Poznyak and K. Najim, "Bush-Mosteller learning for a zero-sum repeated game with random pay-offs," *Int. J. Syst. Sci.*, vol. 32, no. 10, pp. 1251–1260, 2001.
[39] Q. Zhu, H. Tembine, and T. Basar, "Heterogeneous learning in zero-sum stochastic games with incomplete information," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Dec. 2010, pp. 219–224.
[40] I. Erev and A. E. Roth, "Multi-agent learning and the descriptive value of simple models," *Artif. Intell.*, vol. 171, no. 7, pp. 423–428, May 2007.
[41] W. B. Arthur, "On designing economic agents that behave like human agents," *J. Evol. Econ.*, vol. 3, no. 1, pp. 1–22, Mar. 1993.
[42] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Comput. Surv.*, vol. 45, no. 3, pp. 1–39, Jun. 2013.
[43] V. Vadori, M. Scalabrin, A. V. Guglielmi, and L. Badia, "Jamming in underwater sensor networks as a Bayesian zero-sum game with position uncertainty," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 1–6.

Anis Yazidi (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Agder, Grimstad, Norway, in 2008 and 2012, respectively.

He was a Researcher with Teknova AS, Grimstad. From 2014 to 2019, he was an Associate Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway. He is currently a Full Professor with the Department of Computer Science, where he is leading the research group in applied artificial intelligence. He is also a Professor II with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. His current research interests include machine learning, learning automata, stochastic optimization, and autonomous computing.

Daniel Silvestre received the B.Sc. degree in computer networks from the Instituto Superior Técnico (IST), Lisbon, Portugal, in 2008, and the M.Sc. degree in advanced computing and the Ph.D. degree (Hons.) in electrical and computer engineering from Imperial College London, London, U.K., in 2009 and 2017, respectively.

He was a Visiting Scholar at the University of California at Santa Barbara, Santa Barbara, CA, USA. He is currently with the Institute for Systems and Robotics, IST. He holds a research assistant appointment with the University of Macau, Macau. His research interests span the fields of fault detection and isolation, distributed systems, network control systems, computer networks, set-valued estimation and control methods, and randomized algorithms.

B. John Oommen (Life Fellow, IEEE) was born in India, in 1953. He received the Bachelor of Technology degree in electrical engineering from IIT Madras, Chennai, India, in 1975, the Master of Engineering degree from the Indian Institute of Science, Bengaluru, India, in 1977, and the Master of Science degree and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1979 and 1982, respectively.

He has been teaching at the School of Computer Science, Carleton University, Ottawa, ON, Canada, since 1981. He was elevated to be a Chancellor's Professor at Carleton University in 2006. He has published more than 485 refereed publications, many of which have been award-winning.

Dr. Oommen was nominated as a fellow of the Institute of Electrical and Electronic Engineers (IEEE) for research in a subfield of artificial intelligence, namely in learning automata in 2003. He was also nominated as a fellow of the International Association of Pattern Recognition (IAPR) in August 2006 for contributions to fundamental and applied problems in syntactic and statistical pattern recognition. He received the Carleton University's Research Achievement Award four times, in 1995, 2001, 2007, and 2015. At IIT Madras and the Indian Institute of Science, he received the medal for being the Best Graduating Student. He has served on the Editorial Board of the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS and *Pattern Recognition*.