The Design of Absorbing Bayesian Pursuit Algorithms and the Formal Analyses of their ε-Optimality

Xuan Zhang[*], B. John Oommen[†], and Ole-Christoffer Granmo[‡]

Abstract

The fundamental phenomenon that has been used to enhance the convergence speed of Learning Automata (LA) is that of incorporating the running Maximum Likelihood (ML) estimates of the action reward probabilities into the probability updating rules for selecting the actions. The frontiers of this field have been recently expanded by replacing the ML estimates with their corresponding Bayesian counterparts that incorporate the properties of the conjugate priors [1–3]. These constitute the Bayesian Pursuit Algorithm (BPA) [1], and the Discretized Bayesian Pursuit Algorithm (DBPA) [2, 3]. Although these algorithms have been designed and efficiently implemented[1], and are, arguably, the fastest and most accurate LA reported in the literature[2], the proofs of their ε-optimal convergence has been unsolved. This is precisely the intent of this paper. In this paper, we present a *single unifying analysis* by which the proofs of both the continuous and discretized schemes are proven. We emphasize that unlike the ML-based pursuit schemes, the Bayesian schemes have to not only consider the estimates themselves but also the *distributional forms* of their conjugate posteriors and their higher order moments – all of which render the proofs to be particularly challenging. As far as we know, apart from the results themselves, the *methodologies* of this proof have been unreported in the literature - they are both pioneering and novel.

Keywords : *Bayesian Pursuit Algorithms (BPA), Discretized Bayesian Pursuit Algorithms (DBPA), ε-optimality of LA, Beta Distribution.*

[*]This author can be contacted at: Department of ICT, University of Agder, Grimstad, Norway. E-mail address: xuan.zhang@uia.no.

[†]*Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. This author can be contacted at: School of Computer Science, Carleton University, Ottawa, Canada : K1S 5B6. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway. E-mail address: oommen@scs.carleton.ca.

[‡]This author can be contacted at: Department of ICT, University of Agder, Grimstad, Norway. E-mail address: ole.granmo@uia.no.

[1]The version of the BPA presented here, namely the *Absorbing* Bayesian Pursuit Algorithm (ABPA), is distinct from the version presented in [1]. The reason for proposing this newer *absorbing* version will be explained in the body of the paper.

[2]The families of BPA are faster and more accurate than their counterparts that invoke the ML estimates because unlike the former which use the information in the mean, the BPA families utilize the information at a higher-end quantile (95%th precentile) of the posterior Bayesian distribution. This is the rationale for the claim that they are probably, the fastest and most accurate reported LA.

1 Introduction

This paper deals with the formal analysis of the convergence properties of the most-recently proposed family of Learning Automata (LA) which (a) are estimator-based, (b) are of a pursuit nature, and (c) utilize the family of Bayesian estimates in their updating rules. We shall clarify all of these salient features of these schemes and the consequent proof that we embark on, in the forthcoming paragraphs. To guide the reader through the various aspects at stake, we shall briefly visit the crucial issues individually.

What is a Learning Automaton (LA): An LA is an adaptive decision-making unit that learns the optimal action from among a set of actions offered by the Environment it operates in. At each iteration, the LA selects one action, the Environment takes this action as its input and gives back to the LA a response based on the action chosen. The response can be either a reward or a penalty, given the stochastic property of the Environment. In this paper, we work with the so-called *P*-model of LA where the stochastic property is characterized by the Bernoulli distributed reward probabilities, and the action associated with the greatest reward probability is uniquely defined as the optimal action. Based on the response and the knowledge acquired in the past iterations, the LA adjusts its action selection strategy to make a "wiser" decision in the next iteration. In this way, the LA, even though it lacks a complete knowledge about the Environment, is able to learn through repeated interactions with the Environment, and adapts itself to the optimal decision. The field of LA has been studied for more than four decades and well-documented surveys of the field are given in [4–6].

Applications of LA: LA have found applications in a variety of fields, including game playing [7], parameter optimization [8], channel selecting for secondary users in cognitive radio networks [9,10], solving knapsack problems [11], optimizing the web polling problem [12, 13], stochastically optimally allocating limited resources [11,14,15], service selection in stochastic environments [16], vehicle path control [17], and assigning capacities in prioritized networks [18]. LA have also been used in natural language processing, string taxonomy [19], graph patitioning [20], and map learning [21].

Structural Development of Field of LA: The development of LA has gone through four stages. Initial LA were Fixed Structure Stochastic Automata (FSSA), with the state update and decision functions being time invariant. Tsetlin, Krylov and Krinsky automata [5] are the most notable examples of this type. Later, Variable Structure Stochastic Automata (VSSA) were developed, which are characterized by functions that update the probability of selecting the various actions. Typical examples of traditional VSSA includes the Linear Reward-Penalty (L_{R-P}) scheme and the Linear Reward-Inaction (L_{R-I}) scheme [5]. The entire field of LA was raised by a quantum level by the discovery and invention of the family of so-called Estimator Algorithms (EAs) explained below.

Estimator Algorithms (EAs): EAs augment an action probability updating scheme with the use of estimates of the reward probabilities of the respective actions. The design of EAs was pioneered by the study of Pursuit Algorithms (PAs) [22]. The first PA was designed to operate by updating the action probabilities

based on the L_{R-I} paradigm. Later, Oommen and Lanctot [23] presented the Discretized Pursuit Algorithm (DPA) by discretizing the action probability space. The DPA was shown to be superior to its continuous counterpart. In order to highlight the distinct characteristics of the DPA and the PA, in this paper, the latter is referred to as the Continuous Pursuit Algorithm (CPA). By the same token, being an EA in its own right, every PA maintains running Maximum Likelihood (ML) reward probability estimates, to determine the current "Best" action for the present iteration. More generally, the PA then pursues the current best action by linearly increasing *its* action probability in a either a L_{R-I}, L_{R-P}, DL_{R-I} or DL_{R-P} paradigm [6]. As the PA considers both the *short-term* responses of the Environment and the *long-term* reward probability estimates in formulating the action probability updating rules, it outperforms traditional VSSA schemes in terms of its accuracy and its rate of convergence.

The Estimates used in EAs: Prior to the work that we have reported recently, *all* of the reported Pursuit and Estimator Algorithms exclusively incorporated the running ML estimates of the action reward probabilities into the probability updating rule for selecting the actions. These estimates have, indeed, been used in [6, 23–26].

The Use of Bayesian Estimates in PAs: As opposed to invoking ML estimates, more recently, the authors of this present paper have proposed the introduction of the family of Bayesian estimates in the LA. This has led to various Bayesian Pursuit Algorithms (BPAs), including the Continuous Bayesian Pursuit Algorithm (CBPA)[3] [1], and the Discretized Bayesian Pursuit Algorithm (DBPA) [2, 3]. Although the BPAs follow the same "pursuit" paradigm of learning, by virtue of the fact that one can invoke the properties of their posterior distributions, the Bayesian estimates provide more accurate estimation and are, consequently, superior to their counterparts which invoke the ML paradigm. Indeed, the families of BPAs are, arguably, the fastest and most accurate LA reported in the literature.

The State-of-the-Art of BPAs: Both the above-mentioned Bayesian-based algorithms have been designed and efficiently implemented[4], and their effectiveness in solving the learning problem have been clearly demonstrated. However, the issue that remains unsolved is the proofs of their ε-optimal convergence. We thus present in this paper a *single unifying analysis* by which the proofs of both the schemes are proven.

Difficulty of the Proofs of EAs: The most difficult part in the design and analysis of LA consists of the formal proofs of their convergence accuracies. The mathematical techniques used for the various families (FSSA, VSSA, Discretized etc.) are quite distinct. The proof methodology for the family of FSSA is the simplest: it quite simply involves formulating the Markov chain for the LA, computing its equilibrium (or steady state) probabilities, and then computing the asymptotic action selection probabilities. The proofs of convergence for traditional VSSA are more complex and involve the theory of small-step Markov processes, distance diminishing operators, and the theory of Regular functions. The proofs for Discretized LA involve the asymptotic analysis of the Markov chain that represents the LA in the discretized space, whence the *total*

[3] In the interest of compactness, unless otherwise stated, we will refer to the CBPA as the BPA.

[4] As mentioned in the Abstract, the version of the BPA presented here, namely the *Absorbing* Bayesian Pursuit Algorithm (ABPA), is distinct from the version presented in [1]. The reason for this is explained presently.

probability of convergence to the various actions is evaluated. However, understandably, the most difficult proofs involve the family of EAs. This is because the convergence involves two intertwined phenomena, namely the convergence of the reward estimates *and* the convergence of the action probabilities themselves. Ironically, the combination of these vectors in the updating rule is what renders the EA fast. However, if the accuracy of the estimates are poor because of inadequate estimation (i.e., if the sub-optimal actions are not sampled "enough number of times"), the convergence accuracy can be diminished. Hence the dilemma!

Proofs of PAs: The proofs for the convergence of PAs have been studied and reported for decades in [6], [23], [24], [25] [26], which, unfortunately, all have a common flaw that has been recently discovered by the authors of [27]. Further, the authors of [27] submitted a new proof for the convergence of the CPA which adequately rectified the flawed proofs. The new proof, focusing only on the CPA, was based on the monotonicity property of the probability of selecting the optimal action, and required the introduction of an additional assumption that forced the learning parameter to be continuously decreasing over time. In order to provide a unifying analysis that is applicable to both the CPA and the DPA, the authors of [28–31] presented a completely different proof methodology based on the submartingale property and the theory of Regular functions. The latter proof, indeed, requires the PAs to have absorbing states [32, 33], and thus when it is applied to the CPA, the CPA needs to be artificially rendered absorbing, i.e., by constraining the learning process to jump to an absorbing barrier in a single step when any of the action probabilities is greater than or equal to a user-defined threshold that is close to unity. The artificially rendered absorbing CPA was called Absorbing CPA (ACPA) to highlight the difference.

Intent of this Paper: In this paper, we intend to prove the convergence accuracies of the latest EAs, namely, the BPA and the DBPA. To ensure that the algorithms have absorbing barriers, as in the case of the CPA mentioned above, the BPA is rendered artificially absorbing, and is thus referred to as the Absorbing BPA (ABPA). The discretized BPA, however, is truly absorbing, and it is thus unmodified. The unified proof that we present is thus valid for both the ABPA and the DBPA. From the perspective of Martingale convergence theory and the theory of Regular functions, in principle, the present proof is related to the foundational concepts of the proofs in [28–31]. However, the difference is non-trivial and is not just cosmetic. Indeed, we emphasize that unlike the ML-based pursuit schemes, the Bayesian schemes have to not only consider the estimates themselves but also the *distributional forms of the conjugate posteriors* – all of which render the present proofs particularly challenging. Thus, as far as we know, apart from the results themselves, the *methodologies* of this proof have been unreported in the literature - they are both pioneering and novel.

The rest of the paper is organized as follows: Section 2 reviews the notations and algorithmic processes followed by both the ABPA and the DBPA. Section 3 proves, in detail, the convergence of the ABPA and the DBPA, i.e., that they are both ε-optimal in all stationary environments. We conclude the paper in Section 4.

2 Overview of the ABPA and the DBPA

We first submit a brief survey of the ABPA and the DBPA so that the readers can possess a fundamental understanding of them both.

Both the ABPA and the DBPA are based on the "pursuit" paradigm of learning. Firstly, they maintain an action probability vector $P = [p_1, p_2, ..., p_r]$, where $\sum_{j=1...r} p_j = 1$, and where r is the number of actions. In each iteration, the question of which action is to be selected is determined by randomly sampling the action probability vector. Secondly, they maintain running *Bayesian* (as opposed to ML) estimates for the reward probabilities. The action associated with the largest reward probability estimate is considered as the "best" action in each current iteration. Thirdly, given the response of the Environment and the knowledge of the current best action, the ABPA increases the probability of selecting the current best action as per the *continuous* L_{R-I} rules, while the DBPA increases the action probability of the current best action as per the *discretized* L_{R-I} rules.

In all brevity, we mention that far more details about the algorithms and their simulated performances have already been reported in the literature. To avoid repetition, they are not included here. Rather, we refer the reader to [1] for additional details about the BPA, and to [2, 3] for additional details about the DBPA. The only difference between the ABPA and the BPA is that if any one of the action probabilities, $p_j(t+1)$, exceeds a pre-defined Threshold, T, which is a user-defined quantity set to be very close to *unity*, $p_j(t+1)$ will jump directly to *unity* and the learning process is terminated. At this juncture, we say that the LA has been "absorbed" into one of the absorbing barriers, where the r unit vectors are the absorbing states.

We first present the notation used in the ABPA and the DBPA.

Table 1: Notations used in the ABPA and the DBPA

Parameters	**Descriptions**
α	The action selected by LA.
p_j	The j^{th} element of the action selection probability vector, P.
a_j, b_j	The two positive parameters of the *Beta* distribution for Action j.
$\hat{d}_j$	The j^{th} element of the Bayesian estimates vector $\hat{D}$, given by the 95% upper bound of the cumulative distribution function of the corresponding *Beta* distribution.
h	The index of the maximal component of the reward probability estimates vector $\hat{D}$.
m	The index of the optimal action.
R	The response from the environment, where $R = 0$ (reward) or $R = 1$ (penalty).
Δ	The minimum stepsize; $\Delta = \frac{1}{rN}$, with N being a positive integer; N is also considered as the learning parameter for the DBPA.
λ	The learning parameter for the ABPA; $0 < \lambda < 1$.

We now formally describe the ABPA and DBPA algorithms. The reader must observe that both of these algorithms have identical steps in the estimation and learning phases, but differ only in the manner by which

the action probabilities are updated - in a continuous manner for the ABPA and in a discretized manner in the DBPA. Consequently, to avoid repetition, both of them are presented in a single schema below. After describing the algorithms, we consider the primary contribution of the paper, namely the definition of an LA being ε-optimal, and the proof of the ABPA and DBPA being ε-optimal.

Algorithms: ABPA & DBPA
Notations: Refer to Table 1.
Initialization:
1. $p_j(t) = 1/r$, where r is the number of actions.
2. Set $a_j = b_j = 1$.
Method:
For t:=1 to ForEver **Do**

1. Pick $\alpha(t)$ randomly as per $P(t)$. Suppose $\alpha(t) = \alpha_j$, for which the Environment's response is R(t).
2. Based on the Bayesian nature of the conjugate distributions, update $a_j(t)$ and $b_j(t)$ as below:

$$\begin{array}{ll} \textbf{If } R(t) = 0, \textbf{ Then} & a_j(t) = a_j(t-1) + 1; b_j(t) = b_j(t-1); \\ \textbf{Else} & a_j(t) = a_j(t-1); b_j(t) = b_j(t-1) + 1; \end{array}$$

3. Define $f_j(t) = f_j(v; a_j(t), b_j(t)) = \frac{v^{(a_j(t)-1)}(1-v)^{(b_j(t)-1)}}{\int_0^1 u^{(a_j(t)-1)}(1-u)^{(b_j(t)-1)}du}$.
   $f_j(t)$ is the probability distribution function of the Beta distribution of the j^{th} action at time t.
4. Identify the upper 95% reward probability bound of $\hat{d}_j(t)$ for each action j as:

$$\int_0^{\hat{d}_j(t)} f_j(v; a_j(t), b_j(t))dv = \frac{\int_0^{\hat{d}_j(t)} v^{(a_j(t)-1)}(1-v)^{(b_j(t)-1)}dv}{\int_0^1 u^{(a_j(t)-1)}(1-u)^{(b_j(t)-1)}du} = 0.95.$$

5. If $\hat{d}_h(t)$ is the largest element of all the estimates at time t, then update $P(t+1)$ as follows:

   - Continuous linear rules for the ABPA,

   If $R(t) = 0$, **Then**
   $p_j(t+1) = (1-\lambda)p_j(t), j \neq h,$
   $p_h(t+1) = 1 - \sum_{j \neq h} p_j(t+1).$
   Else
   $P(t+1) = P(t).$
   /*If any $p_j(t+1) \geq T$, make $p_j(t+1)$ jump to 1 and break the loop*/
   If $p_j(t+1) \geq T, \forall j \in (1, 2, ..., r)$, **Then**
   $p_j(t+1) = 1,$
   Break
   EndIf

   - Discrete linear rules for the DBPA,

   If $R(t) = 0$, **Then**
   $p_j(t+1) = max\{p_j(t) - \Delta, 0\}, j \neq h,$
   $p_h(t+1) = 1 - \sum_{j \neq h} p_j(t+1).$
   Else
   $P(t+1) = P(t).$

EndFor
End Algorithm: ABPA & DBPA

3 The ε-optimality of the ABPA and the DBPA

The ε-optimality[5] of the ABPA (or the DBPA) are defined by the following statement, where 't' is measured in terms of the number of iterations.

Definition of the ABPA (or the DBPA) being ε-optimal: *Given any* $\varepsilon > 0$ *and* $\delta > 0$*, there exist a* $\lambda_0 > 0$ *(or a* $N_0 > 0$*) and a* $t_0 < \infty$ *such that for all time* $t \geq t_0$ *and for any positive learning parameter* $\lambda < \lambda_0$ *(or* $N > N_0$*),* $Pr\{p_m(t) > 1-\varepsilon\} > 1-\delta$.

Informally speaking, this implies that given a sufficiently small (large) value for the learning parameter, the LA will converge to the optimal action with an arbitrarily high probability.

We now prove that the above statement is true. From the perspective of Martingale convergence theory and the theory of Regular functions, the proof follows the lines of the arguments for the convergence proofs of the ACPA [29], which consists of four steps.

1. Firstly, given a sufficiently small (large) value for the learning parameter λ (or N), all the LA's actions will be selected an sufficiently large number of times before a finite time instant, t_0.

2. Secondly, for all $t > t_0$, with an arbitrarily high probability, $\hat{d_j}$, estimated from a Bayesian perspective, will remain in a small enough neighborhood of d_j, implying that $\hat{d_m}$ will be the maximal element in $\hat{D}$ with an arbitrarily high probability.

3. Thirdly, suppose that for $t > t_0$, the probability that $\hat{d_m}$ is ranked as the largest element in $\hat{D}$ is large enough. Then the action probability sequence of $\{p_m(t)\}$, with $t > t_0$, will be a submartingale.

4. Finally, if $\{p_m(t)_{t>t_0}\}$ is a submartingale, by the submartingale convergence theorem and the theory of Regular functions, the probability of the LA converging to the optimal action converges to 1, i.e., $Pr\{p_m(\infty) \to 1\} \to 1$.

Obviously, each of the above steps of the proof relies on its previous step, and so the relationship between them can be depicted as $Step\ 1 \Rightarrow Step\ 2 \Rightarrow Step\ 3 \Rightarrow Step\ 4$. In the following subsections, we will formalize the proofs of each of the four steps one by one.

3.1 *Step* 1: All actions will be selected enough number of times before t_0

The step asserts that by utilizing a sufficiently small value for the learning parameter, λ, for the ABPA, (or by using a sufficiently large value for the learning parameter, N, for the DBPA), each action will have been selected a sufficiently large number of times by a finite time instant t_0.

The reader will observe that this claim intrinsically depends on the respective probability updating rules of the ABPA and the DPBA, which are exactly the same updating rules invoked by the ACPA and DPA

[5] In the interest of compactness and to avoid repetition, the definition and explanations/statements are given for the ABPA in the main text, and described in a parenthesized manner separately for the DBPA.

respectively, that utilize the ML estimates. In other words, the proofs of these results are already found in the literature, namely in [26] and in [23] respectively. The proofs are thus not repeated here.

3.2 *Step* 2 : $\hat{d}_m$ will be ranked the largest element in $\hat{D}$ with an arbitrarily high probability.

We define $A_1(t)$ as the event that at the time instant t, $\hat{d}_m(t)$ is the largest element in $D(t)$, i.e.,

$$A_1(t) = \{\hat{d}_m(t) > \hat{d}_j(t), \forall j \neq m\}.$$

In that case:

$$q(t) = Pr[A_1(t)] = Pr\{\hat{d}_m(t) > \hat{d}_j(t), \forall j \neq m\},$$

is the probability that $\hat{d}_m(t)$ is the largest element in $D(t)$. The goal of this step is to prove that for any small value $\delta \in (0,1)$, if each action has been selected a sufficiently large number of times by the time instant t_0, then

$$q(t)_{t>t_0} > 1 - \delta. \tag{1}$$

In other words, we prove that if $t > t_0$, the probability of the optimal action being estimated as the best action is arbitrarily close to unity.

We now define another event $A_2(t)$ as

$$A_2(t) = \{|\hat{d}_j(t) - d_j| < \frac{w}{2}, \forall j = 1, ..., r\},$$

where w is the absolute difference between the two largest reward probabilities. Then $A_2(t)$ indicates that at the time instant t, for each action, the reward probability estimate is within the $\frac{w}{2}$ neighborhood of its real value. We can thus see that

$$A_2(t) \Rightarrow A_1(t). \tag{2}$$

If we further define

$$q'(t) = Pr[A_2(t)] = Pr\{|\hat{d}_j(t) - d_j| < \frac{w}{2}, \forall j = 1, ..., r\}, \tag{3}$$

then on the basis of Eq. (2), we have

$$q'(t) < q(t).$$

Therefore, if we can prove that

$$q'(t)_{t>t_0} = Pr\{|\hat{d}_j(t)_{t>t_0} - d_j| < \frac{w}{2}, \forall j = 1,...,r\} > 1-\delta, \tag{4}$$

then Eq. (1) holds. In other words, we are to prove that whenever $t > t_0$, the probability of every reward probability estimate, $\hat{d}_j(t), \forall j = 1,...,r$, being within a $\frac{w}{2}$ neighborhood of its real value, d_j, is greater than $1-\delta$.

For any action α_j, Figure 1 illustrates the relationships between the Beta distribution of the Bayesian estimate of the reward probability, $f_j(t)$, the reward probability estimate, $\hat{d}_j(t)$, the mean, $\bar{d}_j(t)$, and the real reward probability, d_j. The basic idea of proving Eq. (4) consists of two steps. Firstly, we prove that for all j, the mean, $\bar{d}_j(t)$, will be arbitrarily close to the real reward probability, d_j, if each action has been selected an arbitrarily large number of times. Secondly, given that each action has been selected an arbitrarily large number of times, we prove that the quantity $\hat{d}_j(t)$, i.e., the 95% percentile of the posterior Beta distribution at time t, will be arbitrarily close to the mean, $\bar{d}_j(t)$.

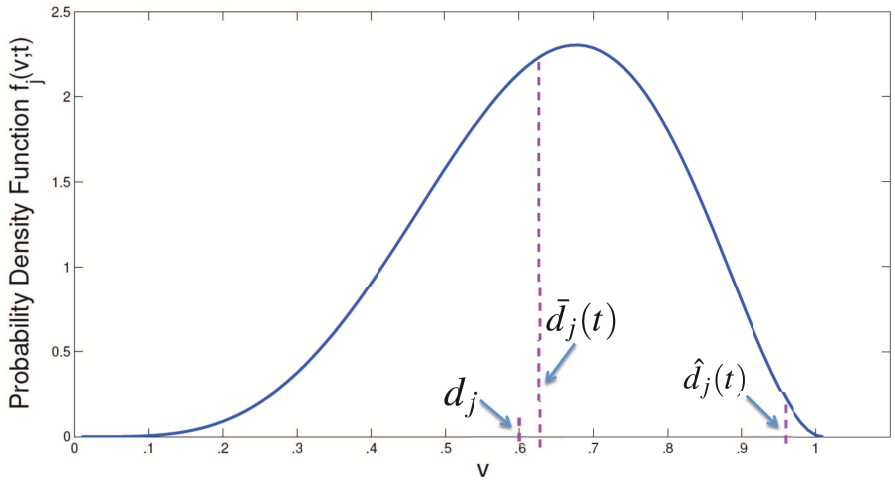

Figure 1: The relationships between the reward probability d_j, the mean $\bar{d}_j$, the 95% percentile $\hat{d}_j$.

1. We prove that $\bar{d}_j(t) \to d_j$ for all $j = 1,...,r$.

   To achieve this, we firstly note that based on the result of Section 3.1, there exists a time instant, denoted in this section as t_1, such that when $t > t_1$, each action will be selected an arbitrarily large number of times. Secondly, as the Beta distribution, $f_j(t)$, is the posterior probability distribution of d_j, and $\bar{d}_j(t)$ is the mean of the Beta distribution, this mean (of the posterior Beta distribution) converges, in the limit to the ML estimate because of the weak law of large numbers, when the number of samples goes to infinity. One should observe that for any finite time instant, the initial values of $a_j(0)$ and $b_j(0)$ will lead to a term that contributes to the mean, $\bar{d}_j(t)$, but this contribution goes to zero as $t \to \infty$. Therefore, proving $\bar{d}_j(t) \to d_j$ is equivalent to proving the convergence of the corresponding

ML estimates which has been presented in [29][6]. Consequently, when $t > t_1$ and $t \to \infty$, $\bar{d_j}(t) \to d_j$. In other words, there exists a time instant $t_1 < \infty$, after which each action will have been selected an arbitrarily large number of times, and

$$Pr\{|\bar{d_j}(t) - d_j| < \frac{w}{4}\} > \sqrt[r]{1-\delta}. \tag{5}$$

If we now consider *all* the actions, the above implies that[7]:

$$Pr\{|\bar{d_j}(t) - d_j| < \frac{w}{4}, \forall j = 1,...,r\} > 1-\delta. \tag{6}$$

2. We prove that $\hat{d_j}(t) \to \bar{d_j}(t)$ for all $j = 1,...,r$.

   To do this, we define $y \in [0,1]$ as a random variable subject to the Beta distribution, $f_j(t)$. Let $\sigma(t)$ be the standard deviation of *this* distribution at time t. Then according to Tchebychev's inequality, for any real number k, we have

$$Pr\{y - \bar{d_j}(t) \geq k\sigma(t)\} \leq \frac{1}{k^2+1}. \tag{7}$$

   As $\hat{d_j}(t)$ is the 95% percentile of the Beta distribution, we have

$$Pr\{y \geq \hat{d_j}(t)\} = 0.05.$$

   If we define a quantity Y as

$$Pr\{y \geq Y\} \leq 0.05, \tag{8}$$

   as per the definition of the *percentile*,

$$Y \geq \hat{d_j}(t),$$

   i.e., Y is the 95% and above percentile.

   We now fix k as

$$\frac{1}{k^2+1} = 0.05. \tag{9}$$

[6] We emphasize that proving the convergence in the case of utilizing the corresponding ML estimates is not merely a consequence of the weak law of large numbers. Indeed, one has to also take into consideration the specific details of the LA updating rules using which the actions are chosen for the estimation purposes. The arguments to do this are quite intricate, and they have been presented in fine detail in [29]. This proof is not repeated here, but can be included if requested by the Referees.

[7] In the interest of simplicity, at this juncture we have assumed that $\bar{d_j}$ are independent of each other. We believe that this assumption can be easily relaxed by considering only the individual d_j's as in Eq. (5), and not all of them together, as in Eq. (6).

In that case, by comparing Eq. (7) and Eq. (8), we have

$$Y = \bar{d_j}(t) + k\sigma(t),$$

indicating that the distance between Y and $\bar{d_j}(t)$ is $k\sigma(t)$. As k is fixed by Eq. (9), and $\sigma(t) \to 0$ as $t \to \infty$, there exists a time instant $t_2 < \infty$, such that when $t > t_2$, $k\sigma(t) < \frac{w}{4}$. Since the 95% percentile value and the percentile values larger than 95% are within a $\frac{w}{4}$ neighborhood of the mean, the 95% percentile $\hat{d_j}(t)_{t>t_2}$ will certainly be within this neighborhood also.

Based on the proofs of the above two steps, if we let $t_0 = max\{t_1, t_2\}$, we can assert that when $t > t_0$:

$$Pr\{|\bar{d_j}(t) - d_j| < \frac{w}{4}, \forall j = 1, ..., r\} > 1 - \delta, \text{ and}$$
$$|\hat{d_j}(t) - \bar{d_j}(t)| < \frac{w}{4}, \forall j = 1, ..., r.$$

Therefore,

$$q'(t) = Pr\{|\hat{d_j}(t) - d_j| < \frac{w}{2}, \forall j = 1, ..., r\} > 1 - \delta,$$

whence

$$q(t) > q'(t) > 1 - \delta,$$

and the result is proven.

3.3 *Step* 3 : $\{p_m(t)_{t>t_0}\}$ is a Submartingale

Before we proceed with the proof of this assertion, we clarify the basic definition of submartingales. Given a sequence of random variables $X_1, X_2, ..., X_t, ...$, if the sequence satisfies the condition that for any time instant t,

$$E[|X_t|] < \infty, \text{ and}$$
$$E[X_{t+1}|X_t, X_{t-1}, ..., X_1] \geq X_t,$$

then the sequence is a submartingale. We now prove that subject to a specific condition, the sequence of $\{p_m(t)_{t>t_0}\}$, i.e., the sequence of $\{p_m(t)\}$ after t_0, indeed is a submartingale.

Firstly, as $p_m(t)$ is a probability, $0 < E[p_m(t)] \leq 1 < \infty$.

Secondly, the description of the algorithms tell us that at the iteration t, an action, say α_j, is selected by the LA. According to the updating rules of the ABPA and the DBPA, if the environment gives a penalty as a

Table 2: The various possibilities for updating p_m for the next iteration under the ABPA (whenever any $p_j(t) < T$) and the DBPA.

	Algorithms	Responses	The greatest element in $\hat{D}$	Updating p_m
$p_m(t+1)$	ABPA	Reward, (w.p. d_j)	$\hat{d}_m$, (w.p. $q(t)$)	$(1-\lambda)p_m(t)+\lambda$
			$\hat{d}_j, j \neq m$, (w.p. $1-q(t)$)	$(1-\lambda)p_m(t)$
		Penalty, (w.p. $1-d_j$)	$\hat{d}_j, j = 1...r$, (1)	$p_m(t)$
	DBPA	Reward, (w.p. d_j)	$\hat{d}_m$, (w.p. $q(t)$)	$p_m(t)+c_t\Delta$
			$\hat{d}_j, j \neq m$, (w.p. $1-q(t)$)	$p_m(t)-\Delta$
		Penalty, (w.p. $1-d_j$)	$\hat{d}_j, j = 1...r$, (1)	$p_m(t)$

response (with probability $1-d_j$), p_m remains the same. But if the LA receives a reward (with probability d_j), one of the following two possibilities follows:

1. If $\hat{d}_m(t)$ is the largest element in $D(t)$, i.e., action m is estimated as the best action at iteration t, which happens with probability $q(t)$, then $p_m(t+1)$ will be increased according to the linear rules with the learning parameter being λ under the ABPA, and with Δ under the DBPA.

2. If $\hat{d}_m(t)$ is not the largest element in $D(t)$, which happens with probability $1-q(t)$, then $p_m(t+1)$ will be decreased according to the corresponding linear rules.

Let us now proceed with computing the expectation of $p_m(t+1)$. To do this, we catalogue the details of the respective updating possibilities in Table 2, based on which we can calculate the expectation of $p_m(t+1)$ explicitly for each scenario, as the following:

- **For the ABPA:**

$$\begin{aligned}
E[p_m(t+1)|P(t)] &= \sum_{j=1...r} p_j\left(d_j\left(q[(1-\lambda)p_m+\lambda]+(1-q)[(1-\lambda)p_m]\right)+(1-d_j)p_m\right)\\
&= p_m d_m q\lambda - d_m\lambda p_m^2 + p_m + \lambda(q-p_m)\sum_{j\neq m} p_j d_j\\
&= p_m + \lambda(q-p_m)\sum_{j=1...r} p_j d_j.
\end{aligned}$$

- **For the DBPA:**

$$\begin{aligned}
E[p_m(t+1)|P(t)] &= \sum_{j=1...r} p_j\left(d_j\left(q(p_m+c_t\Delta)+(1-q)(p_m-\Delta)\right)+(1-d_j)p_m\right)\\
&= \sum_{j=1...r}(p_j d_j q c_t\Delta) - \sum_{j=1...r} p_j d_j\Delta + \sum_{j=1...r} p_j d_j q\Delta + \sum_{j=1...r} p_j p_m\\
&= p_m + \sum_{j=1...r} p_j d_j\left(q(c_t\Delta+\Delta)-\Delta\right).
\end{aligned}$$

In the interest of conciseness, in the above two equations we have omitted the reference to the time index, 't', and hence $p_j(t)$, $p_m(t)$ and $q(t)$ were written as p_j, p_m and q, respectively. Given these explicit expressions

of the expectation of $p_m(t+1)$, we see that for the ABPA,

$$Diff_1(t) = E[p_m(t+1)|P(t)] - p_m(t) = \lambda(q(t) - p_m(t)) \sum_{j=1...r} p_j(t)d_j. \tag{10}$$

The corresponding expression for the DBPA is:

$$Diff_2(t) = E[p_m(t+1)|P(t)] - p_m(t) = \sum_{j=1...r} p_j(t)d_j\,(q(t)(c_t\Delta+\Delta) - \Delta). \tag{11}$$

Invoking the definition of a random variable being a submartingale, we see that in order for the sequence $\{p_m(t)_{t>t_0}\}$ to be a submartingale, we need $Diff_1(t)$ and $Diff_2(t)$ to be greater than or equal to 0 after t_0. We shall examine both these individually.

With regard to Eq. (10), we invoke the terminating condition for the continuous version of Pursuit algorithms, in which we consider the learning process to have converged if $p_j(t) > T = 1-\varepsilon, (j = 1,2,...,r)$. Therefore, if we set the quantity $(1-\delta)$ defined in Section 3.2 to be greater than the threshold T, as per the result in Section 3.2, there exists a time instant $t_0 < \infty$, such that when $t > t_0$, $q(t) > 1-\delta > T > p_m(t)$, which, in turn, guarantees that $Diff_1(t)_{t>t_0} > 0$, proving that $\{p_m(t)_{t>t_0}\}$ is a submartingale under the ABPA.

Now with regard to Eq. (11), we see that we need $q(t) > \frac{\Delta}{c_t\Delta+\Delta}$ for $\{p_m(t)_{t>t_0}\}$ to be a submartingale. As per the action probability updating rules of the DBPA, $c_t = 1,2,...,r-1$, implying that $\frac{\Delta}{c_t\Delta+\Delta} = \frac{1}{r}, \frac{1}{r-1}, ..., \frac{1}{2}$. Therefore, if we set $1-\delta > \frac{1}{2}$, then as per the result in Section 3.2, there exists a time instant $t_0 < \infty$, such that when $t > t_0$, $q(t) > 1-\delta > \frac{1}{2} \geq \frac{\Delta}{c_t\Delta+\Delta}$. Consequently, $\{p_m(t)_{t>t_0}\}$ is a submartingale under the DBPA.

The Claim of Step 3 is thus proven for both the ABPA and the DBPA.

3.4 *Step* 4 : $Pr\{p_m(\infty) = 1\} \to 1$ under the ABPA and the DBPA

Since $\{p_m(t)_{t>t_0}\}$ is a submartingale for both the ABPA and the DBPA, according to the submartingale convergence theory [5],

$$p_m(\infty) = 0 \text{ or } 1.$$

Denoting e_j as the unit vector with the j^{th} element being 1, then

$$p_m(\infty) = 1 \Leftrightarrow p(\infty) = e_m.$$

If we define the convergence probability

$$\Gamma_m(P) = Pr\{P(\infty) = e_m|P(0) = P\},$$

our task is to prove:

$$\Gamma_m(P) \to 1. \tag{12}$$

To prove Eq. (12), we shall use the theory of Regular functions, and the arguments used follow the lines of the arguments found in [5] for the convergence proofs of Absolutely Expedient schemes.

Let $\Phi(P)$ as a function of P. We define an operator U as

$$U\Phi(P) = E[\Phi(P(n+1))|P(n) = P].$$

If we repeatedly apply U, we get the result of the n-step invocation of U as:

$$U^n\Phi(P) = E[\Phi(P(n))|P(0) = P].$$

The function $\Phi(P)$ is referred to as being:

- **Superregular**: If $U\Phi(P) \leq \Phi(P)$. Then applying U repeatedly yields:

$$\Phi(P) \geq U\Phi(P) \geq U^2\Phi(P) \geq ... \geq U^\infty\Phi(P). \tag{13}$$

- **Subregular**: If $U\Phi(P) \geq \Phi(P)$. In this case, if we apply U repeatedly, we have

$$\Phi(P) \leq U\Phi(P) \leq U^2\Phi(P) \leq ... \leq U^\infty\Phi(P). \tag{14}$$

- **Regular**: If $U\Phi(P) = \Phi(P)$. In such a case, it follows that:

$$\Phi(P) = U\Phi(P) = U^2\Phi(P) = ... = U^\infty\Phi(P). \tag{15}$$

Moreover, if $\Phi(P)$ satisfies the boundary conditions

$$\Phi(e_m) = 1 \text{ and } \Phi(e_j) = 0, (\text{for } j \neq m), \tag{16}$$

as per the definition of Regular functions and the submartingale convergence theory, we have

$$\begin{aligned} U^\infty\Phi(P) &= E[\Phi(P(\infty))|P(0) = P] \\ &= \sum_{j=1}^{r} \Phi(e_m) Pr\{P(\infty) = e_j|P(0) = P\} \\ &= Pr\{P(\infty) = e_m|P(0) = P\} \\ &= \Gamma_m(P). \end{aligned} \tag{17}$$

Comparing Eq. (17) with Eq. (15), we see that $\Gamma_m(P)$ is exactly the function $\Phi(P)$ upon which if U is applied an infinite number of times, the sequence of operations will lead to a function that equals the function $\Phi(P)$ itself, because it would then be a *Regular* function. This observation readily leads us to the conclusion that $\Gamma_m(P)$ can be indirectly obtained by investigating a Regular function of P. However, as in the case

of Absolutely Expedient LA, a Regular function is not easily found, although its *existence* is guaranteed. Fortunately, Eq. (13) and Eq. (14) tell us that $\Gamma_m(P)$, i.e., the Regular function of P, can be bounded from above (below) by the superregular (subregular) function of P. Furthermore, as we are most interested in the lower bound of $\Gamma_m(P)$, our goal is to find a proper *subregular* function of P, which also satisfies the boundary conditions given by Eq. (16), which then will guarantee to bound $\Gamma_m(P)$ from below.

Consider a specific instantiation of Φ to be the function Φ_m, defined below as:

$$\Phi_m(P) = e^{-x_m p_m},$$

where x_m is a positive constant. Then, under the ABPA,

$$\begin{aligned}
U(\Phi_m(P)) - \Phi_m(P) &= E[\Phi_m(P(t+1))|P(n) = P] - \Phi_m(P) \\
&= E[e^{-x_m p_m(t+1)}|P(t) = P] - e^{-x_m p_m} \\
&= \sum_{j=1\ldots r} e^{-x_m(p_m(1-\lambda)+\lambda)} p_j d_j q + \sum_{j=1\ldots r} e^{-x_m(p_m(1-\lambda))} p_j d_j (1-q) \\
&\quad + \sum_{j=1\ldots r} e^{-x_m p_m} p_j (1-d_j) - e^{-x_m p_m} \\
&= \sum_{j=1\ldots r} p_j d_j e^{-x_m p_m} \left(q e^{-x_m(1-p_m)\lambda} + (1-q) e^{x_m p_m \lambda} - 1 \right).
\end{aligned}$$

We are, first of all, to find a proper value for x_m such that $\Phi_m(P)$ is superregular, i.e.,

$$U(\Phi_m(P)) - \Phi_m(P) \leq 0.$$

We will see that by determining a suitable superregular function, the corresponding subregular function which satisfies the boundary conditions, can be easily determined.

Determining such an x_m is equivalent to solving the following inequality:

$$q e^{-x_m(1-p_m)\lambda} + (1-q) e^{x_m p_m \lambda} - 1 \leq 0. \tag{18}$$

We know that when $b > 0$ and $x \to 0$,

$$b^x \doteq 1 + (\ln b)x + \tfrac{(\ln b)^2}{2} x^2.$$

If we set $b = e^{-x_m}$, when $\lambda \to 0$, Eq. (18) can be re-written as

$$q\left(1 + (\ln b)(1-p_m)\lambda + \frac{(\ln b)^2}{2}(1-p_m)^2\lambda^2\right) + (1-q)\left(1 + (\ln b)p_m\lambda + \frac{\ln b^2}{2}p_m^2\lambda^2\right) - 1 \leq 0.$$

Substitute b with e^{-x_m}, we see that

$$x_m\left(x_m - \tfrac{2(q(1-p_m)+p_m(1-q))}{\lambda(q-2qp_m+p_m^2)}\right) \leq 0.$$

As x_m is defined as a positive constant, we have

$$0 < x_m \leq \frac{2(q(1-p_m)+p_m(1-q))}{\lambda(q-2qp_m+p_m^2)}. \tag{19}$$

Denoting

$$x_{m_0} = \frac{2(q(1-p_m)+p_m(1-q))}{\lambda(q-2qp_m+p_m^2)},$$

we have $x_{m_0} > 0$, implying that when $\lambda \to 0$, $x_{m_0} \to \infty$.

We now introduce another function

$$\phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}},$$

where x_m is the same as defined in $\Phi_m(P)$. Moreover, we observe the property that if $\Phi_m(P) = e^{-x_m p_m}$ is a superregular (subregular), then $\phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}}$ is a subregular (superregular) [5]. Therefore, the x_m, as defined in Eq. (19), which renders $\Phi_m(P)$ to be superregular, forces the $\phi_m(P)$ to be subregular.

Obviously, $\phi_m(P)$ meets the boundary conditions, i.e.,

$$\phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}} = \begin{cases} 1, \text{ when } P = e_m, \\ 0, \text{ when } P = e_j. \end{cases}$$

Therefore, according to Eq. (14),

$$\Gamma_m(P) \geq \phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}}. \tag{20}$$

As Eq. (20) holds for every x_m bounded by Eq. (19), we take the greatest value x_{m_0}. Moreover, as $\lambda \to 0$, $x_{m_0} \to \infty$, whence $\Gamma_m(P) \to 1$ under the ABPA, proving the ε-optimality of the ABPA.

We now consider the DBPA. For the same function $\Phi_m(P) = e^{-x_m p_m}$, under the DBPA,

$$\begin{aligned} U(\Phi_m(P)) - \Phi_m(P) &= E[\Phi_m(P(t+1))|P(t) = P] - \Phi_m(P) \\ &= E[e^{-x_m p_m(t+1)}|P(t) = P] - e^{-x_m p_m} \\ &= \sum_{j=1\ldots r} e^{-x_m(p_m+c_t\Delta)} p_j d_j q + \sum_{j=1\ldots r} e^{-x_m(p_m-\Delta)} p_j d_j (1-q) \\ &+ \sum_{j=1\ldots r} e^{-x_m p_m} p_j (1-d_j) - e^{-x_m p_m} \\ &= \sum_{j=1\ldots r} p_j d_j e^{-x_m p_m} \left(q(e^{-x_m c_t \Delta} - e^{x_m \Delta}) + (e^{x_m \Delta} - 1) \right). \end{aligned} \tag{21}$$

If we now follow the same algebraic steps as in the ABPA[8], we determine the x_m that renders $U(\Phi_m(P)) - \Phi_m(P) \leq 0$ as:

$$0 < x_m \leq \frac{2(q(c_t+1)-1)}{\Delta(q(c_t^2-1)+1)}. \tag{22}$$

Denoting

$$x_{m_0} = \frac{2(q(c_t+1)-1)}{\Delta(q(c_t^2-1)+1)},$$

we see that $x_{m_0} > 0$ because $c_t = 1, 2, ..., r-1$ and $q(t)_{(t>t_0)} > \frac{1}{2}$. Thus, when $\Delta \to 0$, $x_{m_0} \to \infty$. Substituting x_{m_0} into Eq. (20), we see that as $\Delta \to 0$, $x_{m_0} \to \infty$, whence $\Gamma_m(P) \to 1$ under the DBPA, thus proving its ε-optimality.

We have hereby proved that both the ABPA and DBPA are ε-optimal in all stationary environments.

4 Conclusions

Estimator Algorithms (EAs) which use Maximum Likelihood (ML) estimates have been acclaimed to be the fastest Learning Automata (LA). Discretized versions of all these schemes have also been proposed. More recently, we have further enhanced EAs by replacing the ML estimates with their corresponding Bayesian counterparts that incorporate the properties of the conjugate priors [1–3]. Further, since the Bayesian estimates take into account the information in the higher order moments of the underlying distributions of the estimates, they provide a more accurate estimation strategy than the ML estimates – which only consider the information contained in the first order moment, i.e., the mean. The consequent algorithms that we have proposed are the Bayesian Pursuit Algorithm (BPA) [1], and the Discretized Bayesian Pursuit Algorithm (DBPA) [2, 3]. Although these algorithms have been designed and efficiently implemented, and are, arguably, the fastest and most accurate LA reported in the literature, the proofs of their ε-optimality have been unsolved.

In this paper, we have formally proven, with a single unified proof, the ε-optimality of both the Absorbing Continuous Bayesian Pursuit Algorithm and the Discretized Bayesian Pursuit Algorithm, where the ABPA is the BPA with artificially rendered absorbing states. From the perspective of the Martingale convergence theory and the theory of Regular functions, the proof which we have submitted is akin to the proof for the convergence of the family of PAs [28–31]. However, unlike the ML-based pursuit schemes, the Bayesian schemes have to not only consider the estimates themselves but also the *distributional forms* of their conjugate posteriors and their higher order moments – all of which render the proofs to be particularly challenging. The interesting feature of this current proof for the family of Bayesian-based EAs is that it takes into consideration both the means and the standard deviations of the estimates.

[8] In order to not burden the reader with cumbersome algebraic manipulations, we omit the straightforward steps.

We believe that the present proof for the convergence accuracy of the family of BPAs will add more insight into the mechanism by which EAs can be both improved and analyzed.

References

[1] X. Zhang, O.-C. Granmo, and B. J. Oommen, "The Bayesian pursuit algorithm: A new family of estimator learning automata," in *Proceedings of IEA-AIE 2011*. New York, USA: Springer, Jun. 2011, pp. 608–620.

[2] ——, "On incorporating the paradigms of discretization and Bayesian estimation to create a new family of pursuit learning automata," *Applied Intelligence*, vol. 39, pp. 782–792, 2013.

[3] ——, "Discretized Bayesian pursuit - a new scheme for reinforcement learning," in *Proceedings of IEA-AIE 2012*, Dalian, China, Jun. 2012, pp. 784–793.

[4] K. S. Narendra and M. A. L. Thathachar, "Learning automata - a survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 4, pp. 323–334, 1974.

[5] ——, *Learning Automata: An Introduction*. Prentice Hall, 1989.

[6] B. J. Oommen and M. Agache, "Continuous and discretized pursuit learning schemes: various algorithms and their comparison," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 3, pp. 277–287, 2001.

[7] B. J. Oommen, O.-C. Granmo, and A. Pedersen, "Using stochastic AI techniques to achieve unbounded resolution in finite player Goore Games and its applications," in *Proceedings of IEEE Symposium on Computational Intelligence and Games*, Honolulu, HI, Apr. 2007, pp. 161–167.

[8] H. Beigy and M. R. Meybodi, "Adaptation of parameters of BP algorithm using learning automata," in *Proceedings of Sixth Brazilian Symposium on Neural Networks*, JR, Brazil, Nov. 2000, pp. 24–31.

[9] X. Zhang, L. Jiao, O.-C. Granmo, and B. J. Oommen, "Channel selection in cognitive radio networks: A switchable bayesian learning automata approach," in *Proceedings of PIMRC*, London, United Kingdom, Sept. 2013, pp. 2362–2367.

[10] L. Jiao, X. Zhang, O.-C. Granmo, and B. J. Oommen, "A bayesian learning automata-based distributed channel selection scheme for cognitive radio networks," in *Proceedings of IEA-AIE*, Kaohsiung, Taiwan, Jun. 2014, pp. 48–57.

[11] O.-C. Granmo, B. J. Oommen, S.-A. Myrer, and M. G. Olsen, "Learning automata-based solutions to the nonlinear fractional knapsack problem with applications to optimal resource allocation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, no. 1, pp. 166–175, 2007.

[12] ——, “Determining Optimal Polling Frequency Using a Learning Automata-based Solution to the Fractional Knapsack Problem,” in *Proceedings of the 2006 IEEE International Conferences on Cybernetics and Intelligent Systems (CIS) and Robotics, Automation and Mechatronics (RAM)*, Bangkok, Thailand, Jun. 2006, pp. 1–7.

[13] O.-C. Granmo and B. J. Oommen, “Learning automata-based solutions to the optimal web polling problem modeled as a nonlinear fractional knapsack problem,” *Engineering Applications of Artificial Intelligence*, vol. 24, no. 7, pp. 1238–1251, 2011.

[14] ——, “On Allocating Limited Sampling Resources Using a Learning Automata-based Solution to the Fractional Knapsack Problem,” in *Proceedings of the 2006 International Intelligent Information Processing and Web Mining Conference, Advances in Soft Computing*, vol. 35, Ustron, Poland, Jun. 2006, pp. 263–272.

[15] ——, “Optimal sampling for estimation with constrained resources using a learning automaton-based solution for the nonlinear fractional knapsack problem,” *Applied Intelligence*, vol. 33, no. 1, pp. 3–20, 2010.

[16] A. Yazidi, O.-C. Granmo, and B. J. Oommen, “Service selection in stochastic environments: A learning-automaton based solution,” *Applied Intelligence*, vol. 36, pp. 617–637, 2012.

[17] C. Unsal, P. Kachroo, and J. S. Bay, “Multiple stochastic learning automata for vehicle path control in an automated highway system,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 29, pp. 120–128, 1999.

[18] B. J. Oommen and T. D. Roberts, “Continuous learning automata solutions to the capacity assignment problem,” *IEEE Transactions on Computers*, vol. 49, pp. 608–620, Jun. 2000.

[19] B. J. Oommen and T. D. S. Croix, “String taxonomy using learning automata,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 27, pp. 354–365, Apr. 1997.

[20] ——, “Graph partitioning using learning automata,” *IEEE Transactions on computers*, vol. 45, pp. 195–208, 1996.

[21] T. Dean, D. Angluin, K. Basye, S. Engelson, L. Aelbling, and O. Maron, “Inferring finite automata with stochastic output functions and an application to map learning,” *Maching Learning*, vol. 18, pp. 81–108, 1995.

[22] M. A. L. Thathachar and P. S. Sastry, “Estimator algorithms for learning automata,” in *Proceedings of the Platinum Jubilee Conference on Systems and Signal Processing*, Bangalore, India, Dec. 1986, pp. 29–32.

[23] B. J. Oommen and J. K. Lanctot, "Discretized pursuit learning automata," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, pp. 931–938, 1990.

[24] J. K. Lanctot and B. J. Oommen, "Discretized estimator learning automata," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 22, no. 6, pp. 1473–1483, 1992.

[25] ——, "On discretizing estimator-based learning algorithms," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 2, pp. 1417–1422, 1991.

[26] K. Rajaraman and P. S. Sastry, "Finite time analysis of the pursuit algorithm for learning automata," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 26, pp. 590–598, 1996.

[27] M. Ryan and T. Omkar, "On ε-optimality of the pursuit learning algorithm," *Journal of Applied Probability*, vol. 49, no. 3, pp. 795–805, 2012.

[28] X. Zhang, O.-C. Granmo, B. J. Oommen, and L. Jiao, "On using the theory of regular functions to prove the ε-optimality of the continuous pursuit learning automaton," in *Proceedings of IEA-AIE 2013*. Amsterdan, Holland: Springer, Jun. 2013, pp. 262–271.

[29] ——, "A formal proof of the ε-optimality of absorbing continuous pursuit algorithms using the theory of regular functions," *Applied Intelligence*, vol. 41, pp. 974–985, 2014.

[30] X. Zhang, B. J. Oommen, O.-C. Granmo, and L. Jiao, "Using the theory of regular functions to formally prove the ε-optimality of discretized pursuit learning algorithms," in *Proceedings of IEA-AIE*. Kaohsiung, Taiwan: Springer, Jun. 2014, pp. 379–388.

[31] ——, "A formal proof of the ε-optimality of *discretized* pursuit algorithms," *submitted to Applied Intelligence*, 2014.

[32] B. J. Oommen, "Absorbing and ergodic discretized two-action learning automata," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, pp. 282–296, 1986.

[33] J. K. Lanctot and B. J. Oommen, "Discretized estimator learning automata," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 6, pp. 1473–1483, 1992.