On Using the Theory of Regular Functions to Prove the ε-Optimality of the Continuous Pursuit Learning Automaton

Xuan Zhang[1], Ole-Christoffer Granmo[1], B. John Oommen[2,1,⋆], and Lei Jiao[1]

[1] Dept. of ICT, University of Agder, Grimstad, Norway
[2] School of Computer Science, Carleton University, Ottawa, Canada

Abstract. There are various families of Learning Automata (LA) such as Fixed Structure, Variable Structure, Discretized etc. Informally, if the environment is stationary, their ε-optimality is defined as their ability to converge to the optimal action with an arbitrarily large probability, if the learning parameter is sufficiently small/large. Of these LA families, Estimator Algorithms (EAs) are certainly the fastest, and within this family, the set of *Pursuit* algorithms have been considered to be the pioneering schemes. The existing proofs of the ε-optimality of all the reported EAs follow the same fundamental principles. Recently, it has been reported that the previous proofs for the ε-optimality of *all* the reported EAs have a *common flaw*. In other words, people have worked with this flawed reasoning for almost three decades. The flaw lies in the condition which apparently supports the so-called "monotonicity" property of the probability of selecting the optimal action, explained in the paper. In this paper, we provide a new method to prove the ε-optimality of the Continuous Pursuit Algorithm (CPA), which was the pioneering EA. The new proof follows the same outline of the previous proofs, but instead of examining the monotonicity property of the action probabilities, it rather examines their *submartingale* property, and then, unlike the traditional approach, invokes the theory of *Regular* functions to prove the ε-optimality. We believe that the proof is both unique and pioneering, and that it can form the basis for formally demonstrating the ε-optimality of other EAs.

Keywords: Pursuit Algorithms, Continuous Pursuit Algorithm, ε-optimality.

1 Introduction

Learning automata (LA) have been studied as a typical model of reinforcement learning for decades. They have found applications in a variety of fields, including game playing [1], parameter optimization [2], solving knapsack-like problems and utilizing the solution in web polling and sampling [3], vehicle path control [4], assigning capacities in prioritized networks [5], and resource allocation [6]. They have also been used in language processing, string taxonomy [7], graph partitioning [8], and map learning [9].

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

[⋆] *Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. The Author also holds an *Adjunct Professorship* with the Dept. of ICT, University of Agder, Norway.

M. Ali et al. (Eds.): IEA/AIE 2013, LNAI 7906, pp. 262–271, 2013.

selects one action, which triggers either a reward or a penalty as a response from the Environment. Based on the response and the knowledge acquired in the past iterations, the LA adjusts its action selection strategy in order to make a "wiser" decision in the next iteration. In such a 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.

Among the families of LA, Estimator Algorithms (EAs) (a more detailed survey of the families is found in [19]) work with a noticeably different paradigm, and are certainly the fastest and most accurate. Within this family, the set of *Pursuit* Algorithms (PAs) were the pioneering schemes, whose design and analysis were initiated by Thathachar and Sastry [10]. EAs augment an action probability updating scheme with the use of estimates of the reward probabilities of the respective actions. The first Pursuit Algorithm (PA) was designed to operate by updating the action probabilities based on the L_{R-I} paradigm. By the same token, being an EA in its own right, the PA maintains running Maximum Likelihood (ML) reward probability estimates, which further determines the current "Best" action for the present iteration. The PA then pursues the current best action by linearly increasing *its* action probability. 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 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 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* 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!

The ε-optimality of the EAs have been studied and presented in [11] [12] [13] [14] [15]. The basic result stated in these papers is that by utilizing a sufficiently small value for the learning parameter, the CPA will converge to the optimal action with an arbitrarily large probability. However, these proofs have a common flaw, which involves a very fine argument. In fact, the proofs reported in these papers "deduced" the ε-optimality based on the conclusion that after a sufficiently large time instant, t_0, the probability of selecting the optimal action is monotonically increasing, which, in turn, is based on the condition that the reward probability estimates are ordered properly *forever* after t_0.

This ordering is, indeed, true by the law of large numbers if all the actions are chosen *infinitely often*. But if such an "infinite" selection does not occur, the ordering cannot be guaranteed for *all* time instants after t_0.

As a consequence of this misinterpretation, the condition supporting the monotonicity property is false, which further leads to an incorrect proof for the CPA being ε-optimal. Even though this has been the accepted argument for almost three decades (even by the third author of this present paper who was the principal author of many of the above-mentioned papers), we credit the authors of [16] for discovering this flaw. While a detailed explanation of this is found in [16], a brief explanation on this issue is also included in this paper, in Section 3.

This paper aims at correcting the above-mentioned flaw used in the earlier proofs. As opposed to these proofs, we will show that while the so-called monotonicity property is sufficient for convergence, it is not really *necessary* for proving that the CPA is ε-optimal. Rather, we will present a completely new proof methodology which is based on the convergence theory of submartingales and the theory of Regular functions [17].

2 Overview of the CPA

Since this paper concentrates on the intricate nature of the CPA, it is mandatory that the reader has a fundamental understanding of it. It is, thus, briefly surveyed here. To do this, first of all, we present below the notations used:

α_i: The i^{th} action that can be selected by the LA, and is an element from the set $\{\alpha_1, \ldots \alpha_r\}$.
p_i: The i^{th} element of the action probability vector P.
λ: The learning rate, where $0 < \lambda < 1$.
u_i: The number of times α_i has been rewarded when it has been selected.
v_i: The number of times α_i has been selected.
$\hat{d}_i$: The i^{th} element of the reward probability estimates vector $\hat{D}$, $\hat{d}_i = \frac{u_i}{v_i}$.
m: The index of the optimal action.
h: The index of the greatest element of $\hat{D}$.
R: The Environment's response, where $R = 0$ corresponds to a Reward, and $R = 1$ to a Penalty.

The CPA follows a "pursuit" paradigm of learning, which consists of three steps. Firstly, it maintains an action probability vector $P = [p_1, p_2, ..., p_r]$ to determine the issue of which action is to be selected, where the sum of the p_i's is unity, and where r is the number of actions. Secondly, it maintains running ML reward probability estimates to determine which action can be reckoned to be the "best" in the current iteration. Thus, it updates $\hat{d}_i(t)$ based on the response from the Environment as below:

$$\begin{aligned}
u_i(t) &= u_i(t-1) + (1 - R(t)); \\
v_i(t) &= v_i(t-1) + 1 \\
\hat{d}_i(t) &= \frac{u_i(t)}{v_i(t)}.
\end{aligned}$$

Thirdly, based on the response of the Environment and the knowledge of the current best action, the CPA increases the probability of selecting the current best action as per the continuous L_{R-I} rule. So, if $\hat{d}_h(t)$ is the largest element of $\hat{D}(t)$, we update $P(t)$ as:
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

$$P(t+1) = P(t)$$

We now visit the issue of the proof of the CPA's convergence.

3 Previous Proofs for CPA's ε-Optimality

The formal assertion of the ε-optimality of the CPA is stated in Theorem 1.

Theorem 1. *Given any* $\varepsilon > 0$ *and* $\delta > 0$*, there exist a* $\lambda^\star > 0$ *and a* $t_0 < \infty$ *such that for all time* $t \geq t_0$ *and for any positive learning parameter* $\lambda < \lambda^\star$*,*

$$Pr\{p_m(t) > 1-\varepsilon\} > 1-\delta.$$

The earlier reported proofs for the ε-optimality of the CPA follow the "four-step" strategy. Firstly, given a sufficiently small value for the learning parameter λ, all actions will be selected enough number of times before a finite time instant, t_0. Secondly, for all $t > t_0$, $\hat{d}_m$ will remain to be the maximum element of the reward probability estimates vector, $\hat{D}$. Thirdly, suppose $\hat{d}_m$ has been ranked as the largest element in $\hat{D}$ since t_0, the action probability sequence of $\{p_m(t)\}$, with $t > t_0$, will be monotonically increasing, whence one concludes that $p_m(t)$ converges to 1 with probability 1. Finally, given that the probability of $\hat{d}_m$ being the largest element in $\hat{D}$ is arbitrarily close to unity, and that $p_m(t) \to 1$ w.p. 1, ε-optimality is proven based on the axiom of total probability. All of these are listed below.

1. The first step of the CPA's proof of convergence is formalized by Theorem 2.

   Theorem 2. *For any given constants* $\hat{\delta} > 0$ *and* $M < \infty$*, there exist a* $\lambda^\star > 0$ *and a* $t_0 < \infty$ *such that under the CPA algorithm, for all positive* $\lambda < \lambda^\star$*,*

   Pr{All actions are selected at least M times each before $t_0\} > 1-\hat{\delta}$*, for all* $t > t_0$.
   The detailed proof for this result can be found in [14].
2. The sequence of probabilities, $\{p_m(t)_{(t>t_0)}\}$, is stated to be *monotonically* increasing. The previous proofs attempted to do this by showing that:

$$|p_m(t)| \leq 1, \text{ and}$$

$$\Delta p_m(t) = E[p_m(t+1) - p_m(t)|\bar{A}(t_0)] = d_m\lambda(1-p_m(t)) \geq 0,\ t > t_0, \quad (1)$$

   where $\bar{A}(t_0)$ is the condition that after time t_0, for any $j \in (1,2,...,r)$, $\hat{d}_j$ remains within a small enough neighborhood of d_j so that $\hat{d}_m$ remains the greatest element in $\hat{D}$. If this step of the "proof" was flawless, $p_m(t)$ can be shown to converge to 1 w.p. 1. But, as we shall see, the flaw lies here!

3. Since $p_m(t) \to 1$ w.p. 1, if it can, indeed, be proved that $Pr\{\bar{A}(t_0)\} > 1-\delta$, by the axiom of total probability, one can then see that:
$$Pr\{p_m(t) > 1-\varepsilon\} \geq Pr\{p_m(t) \to 1\} Pr\{\bar{A}(t_0)\} > 1-\delta,$$
and ε-optimality is proved.

According to the sketch of the proof above, the key is to prove $Pr\{\bar{A}(t_0)\} > 1-\delta$, i.e.,

$$Pr\{\bar{A}(t_0)\} = Pr\{\bigcap_{t>t_0} \{\hat{d}_j(t)_{\forall j} \text{ is within a } \frac{w}{2} \text{ neighborhood of } d_j \text{ at time } t\}\} > 1-\delta. \tag{2}$$

In Eq. (2), w is defined as the difference between the two *highest* reward probabilities.

In the proofs reported in the literature, Eq. (2) is considered to be true according to the law of large numbers, i.e., if each α_j has been selected enough number of times, then for $\forall j$,

$$Pr\{\hat{d}_j(t) \text{ is within a } \tfrac{w}{2} \text{ neighborhood of } d_j \text{ at time } t\} > 1-\delta', \text{ with } \delta' = 1-\sqrt[r]{1-\delta},$$

so that

$$\prod_{j=1,2,\ldots,r} Pr\{\hat{d}_j(t) \text{ is within a } \tfrac{w}{2} \text{ neighborhood of } d_j \text{ at time } t\} > 1-\delta.$$

However, there is a flaw in the above argument. In fact, if we define

$$A(t) = \{\hat{d}_j(t)_{\forall j} \text{ is within a } \tfrac{w}{2} \text{ neighborhood of } d_j \text{ at time } t\},$$

then the result that can be deduced from the law of large numbers when $t > t_0$ is that

$$Pr\{A(t)\} = \prod_{j=1,2,\ldots,r} Pr\{\hat{d}_j(t) \text{ is within a } \tfrac{w}{2} \text{ neighborhood of } d_j \text{ at time } t\} > 1-\delta.$$

But, indeed, the condition which Eq. (1) is based on is:

$$\bar{A}(t_0) = \bigcap_{t>t_0} A(t),$$

which means that for every single time instant in the future, i.e., $t > t_0$, $\hat{d}_j(t)_{(\forall j)}$ needs to be within the $\frac{w}{2}$ neighborhood of d_j. The flaw in the previous proofs reported in the literature is that they made a mistake by reckoning that $A(t)$ is equivalent to $\bar{A}(t_0)$. This renders the existing proofs for the CPA being ε-optimal, to be incorrect.

The flaw is documented in [16], which further provided a way of correcting the flaw, i.e., by proving $Pr\{\bar{A}(t_0)\} > 1-\delta$ instead of proving $Pr\{A(t)\} > 1-\delta$. However, their proof requires a sequence of *decreasing* values of the learning rate λ. We applaud the authors of [16] for discovering this flaw, and for submitting a more accurate proof for the CPA.

The proof methodology that we use here is quite distinct (and uses completely different techniques) than the proof reported in [16]. We seek an alternate proof because in their proof, the authors of [16] have required the constraint $\bar{A}(t_0)$, which is, indeed, a very strong condition. This, in turn, requires that for the CPA to achieve its ε-optimality, one must rely on an additional assumption that the parameter, λ, is gradually decreased during the learning process. We would like to remove this. Our new proof also follows a four-step sketch, but is rather based on the convergence theory of submartingales, and on the theory of Regular functions.

4 The New Proof for the CPA's ε-Optimality

4.1 The Moderation Property of CPA

The property of moderation can be described by Theorem 2, which has been proved in [14]. This implies that under the CPA, by utilizing a sufficiently small value for the learning parameter, λ, each action will be selected an arbitrarily large number of times.

4.2 The Key Condition $\bar{B}(t_0)$ for $\{p_m(t)_{t>t_0}\}$ Being a Submartingale

In our proof strategy, instead of examining the condition for $\{p_m(t)_{t>t_0}\}$ being *monotonically increasing*, we will investigate the condition for $\{p_m(t)_{t>t_0}\}$ being a *submartingale*. The latter is based on the condition, $\bar{B}(t_0)$, defined as follows:

$$\begin{aligned} &q_j(t) = Pr\{|\hat{d}_j(t) - d_j| < \frac{w}{2}\},\\ &q(t) = Pr\{|\hat{d}_j(t) - d_j| < \frac{w}{2}, \forall j \in (1,2,...,r)\} = \prod_{j=1,2,...,r} q_j(t), \qquad (3)\\ &B(t) = \{q(t) > 1-\delta\}, \delta \in (0,1),\\ &\bar{B}(t_0) = \{\bigcap_{t>t_0} \{q(t) > 1-\delta\}\}. \qquad (4)\end{aligned}$$

Our goal in this step is to prove the following result, formulated in Theorem 3.

Theorem 3. *Given a* $\delta \in (0,1)$*, there exists a time instant* $t_0 < \infty$*, such that* $Pr\{\bar{B}(t_0)\} = 1$*. In other words, for this given* δ*, there exists a* $t_0 < \infty$*, such that* $\forall t > t_0$*:* $q(t) > 1-\delta$ *w. p.* 1.

Sketch of Proof: The proof of this is quite detailed. It includes the following steps:

1. By setting $\delta' = 1 - \sqrt[r]{1-\delta}$, we observe that $\forall t > t_0$, if for $\forall j$, $q_j(t) > 1-\delta'$, then $q(t) = \prod_{j=1,2,...,r} q_j(t) > \prod_{j=1,2,...,r} (1-\delta') = 1-\delta$. Therefore, if we define $B_j(t) = \{q_j(t) > 1-\delta'\}$, and $\bar{B}_j(t_0) = \{\bigcap_{t>t_0} B_j(t)\}$, our task becomes to prove that for $\forall j$,
$$Pr\{\bar{B}_j(t_0)\} = Pr\{\bigcap_{t>t_0} B_j(t)\} = 1.$$
2. By DeMorgan's law, $Pr\{\bar{B}_j(t_0)\} = Pr\{\bigcap_{t\geq t_0} B_j(t)\} = 1 - Pr\{\bigcup_{t\geq t_0} B_j(t)^c\}$, where c denotes the complement operation. We thus need to prove $Pr\{\bigcup_{t\geq t_0} B_j(t)^c\} = 0$.
3. Let $n_j(t)$ denote the number of times α_j has been selected up to time t, then

$$\begin{aligned} Pr\{\bigcup_{t\geq t_0} B_j(t)^c\} &\leq \sum_{t\geq t_0} Pr\{B_j(t)^c\}\\ &= \sum_{t\geq t_0}\left(\sum_{n=0}^{t} Pr\{q_j(t) \leq 1-\delta' | n_j(t) = n\} \times Pr\{n_j(t) = n\}\right)\\ &= \sum_{t\geq t_0}\left(\sum_{n=0}^{t} Pr\{Pr\{|\hat{d}_j(t) - d_j| \geq \frac{w}{2}\} \geq \delta' | n_j(t){=}n\} Pr\{n_j(t){=}n\}\right).\end{aligned}$$

By applying the Hoeffding's inequality [18]: $Pr\{|\hat{d}_j(t)-d_j| \geq \frac{w}{2}\} \leq 2e^{-\frac{nw^2}{2}}$, hence,

$$\begin{aligned} Pr\{\bigcup_{t\geq t_0} B_j(t)^c\} &\leq \sum_{t\geq t_0} Pr\{B_j(t)^c\} \\ &\leq \sum_{t\geq t_0}\left(\sum_{n=0}^{t} Pr\{2e^{-\frac{nw^2}{2}} \geq \delta'\} \times Pr\{n_j(t)=n\}\right) \\ &= \sum_{t\geq t_0}\left(\sum_{n=0}^{t} Q_e \times Q_{n_j}\right), \end{aligned} \tag{5}$$

where $Q_e = Pr\{2e^{-\frac{nw^2}{2}} \geq \delta'\}$, and $Q_{n_j} = Pr\{n_j(t)=n\}$.

4. It is easy to conclude that

$$Q_e = Pr\{2e^{-\frac{nw^2}{2}} \geq \delta'\} = \begin{cases} 1, & \text{when } n \leq \frac{-2\ln\frac{\delta'}{2}}{w^2}, \\ 0, & \text{when } n > \frac{-2\ln\frac{\delta'}{2}}{w^2}. \end{cases} \tag{6}$$

Besides, the quantity Q_{n_j} is the probability of α_j being selected for n times within the given time instant t. As n could be any integer from $[0,t]$, we have $\sum_{n=0}^{t} Q_{n_j} = 1$. If we further assume that till the time instant t, α_j has been selected *at least* x_0 times, i.e. $n_j(t) \geq x_0$, then

$$Q_{n_j}\begin{cases} = 0, & \text{when } 0 \leq n < x_0, \\ \in [0,1], & \text{when } x_0 \leq n \leq t, \end{cases} \quad \text{and} \quad \sum_{n=x_0}^{t} Q_{n_j} = 1. \tag{7}$$

5. From Eq. (6) and (7), we see that if $x_0 > \lceil\frac{-2\ln\frac{\delta'}{2}}{w^2}\rceil$, then $\sum_{n=0}^{t} Q_e \times Q_{n_j} = 0$, whence

$$Pr\{\bigcup_{t\geq t_0} B_j(t)^c\} \leq \sum_{t\geq t_0} Pr\{B_j(t)^c\} \leq \sum_{t\geq t_0}\left(\sum_{n=0}^{t} Q_e \times Q_{n_j}\right) = 0.$$

Obviously, the above arguments apply to $\forall j \in (1,2,...,r)$. We thus proved that $\forall j$, $Pr\{\bar{B}_j(t_0)\} = Pr\{\bigcap_{t>t_0} B_j(t)\} = 1$, which leads to the result that $Pr\{\bar{B}(t_0)\} = 1$.

4.3 $\{p_m(t)_{t>t_0}\}$ Is a Submartingale under the CPA

We now prove the submartingale properties of $\{p_m(t)_{t>t_0}\}$ for the CPA.

Theorem 4. *Under the CPA, the quantity $\{p_m(t)_{t>t_0}\}$ is a submartingale.*

Sketch of Proof: Firstly, since $p_m(t)$ is a probability, we have $E[p_m(t)] \leq 1 < \infty$. Secondly, we proceed to explicitly calculate $E[p_m(t)]$. Using the CPA's updating rule:

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

where $p_m(t)$ and $q(t)$ are concisely written as p_m and q respectively. Then,

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

Invoking the definition of a submartingale, we know that if for all $t > t_0$, we have $Diff_{p_m(t)} > 0$, i.e., $q(t) - p_m(t) > 0$, then $\{p_m(t)_{t>t_0}\}$ is a submartingale. We now invoke the terminating condition for the CPA, in which we consider the learning process to have converged[1] if $p_j(t) > T = 1-\varepsilon, (j = 1,2,...,r)$. Therefore, if we set the quantity $(1-\delta)$ defined in Theorem 3 to be greater than the threshold T, then as per Theorem 3, there exists a time instant $t_0 < \infty$, such that for every single time instant subsequent to $t > t_0$, $q(t) > (1-\delta) > T > p_m(t)$, which, in turn, guarantees that $\{p_m(t)_{t>t_0}\}$ is a submartingale.

4.4 $Pr\{p_m(\infty) = 1\} \to 1$ under the CPA

We can now finally prove the ε-optimality of the CPA.

Theorem 5. *The CPA is* ε*-optimal in all stationary random Environments. More formally, let* $T = 1-\varepsilon$ *be a value arbitrarily close to* 1*, with* ε *being arbitrarily small. Then, given any* δ *satisfying* $(1-\delta) > T$*, there exists a positive integer* $\lambda^\star < 1$ *and a time instant* $t_0 < \infty$*, such that for all learning parameters* $\lambda < \lambda^\star$ *and for all* $t > t_0$*,* $q(t) > 1-\delta$*,* $Pr\{p_m(\infty) = 1\} \to 1$.

Sketch of Proof: According to the submartingale convergence theory [17], $p_m(\infty) =$ 0 or 1. If we denote e_j as the unit vector with the j^{th} element being 1, then our task is to prove the convergence probability

$$\Gamma_m(P) = Pr\{p_m(\infty) = 1|P(0) = P\} = Pr\{p(\infty) = e_m|P(0) = P\} \to 1. \tag{8}$$

To prove Eq. (8), we shall use the theory of Regular functions, and arguments analogous to those used in [17] for the convergence proofs of Absolutely Expedient schemes.

According to theory of Regular functions, $\Gamma_m(P)$ can be bounded from below by a subregular function of P, denoted as $\Phi(P)$, if $\Phi(P)$ meets the boundary conditions:

$$\Phi(e_m) = 1 \text{ and } \Phi(e_j) = 0, (\text{for } j \neq m). \tag{9}$$

Our task is thus to find such a subregular function of P to investigate $\Gamma_m(P)$ indirectly. If we define a function $\Phi_m(P)$ as $\Phi_m(P) = e^{-x_m p_m}$, where x_m is a positive constant, and then define an operator U as:

$$U\Phi_m(P) = E[\Phi_m(P(n+1))|P(n) = P],$$

then, under the CPA,

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

[1] In practice, T is the threshold used to determine when we say that the LA has been "absorbed" into one of the absorbing barriers. This quantity is arbitrarily close to unity, say, 0.999.

Omitting the algebraic manipulation, we get the result that if

$$0 < x_m \leq \frac{2(q(1-p_m)+p_m(1-q))}{\lambda(q-2qp_m+p_m^2)}, \tag{10}$$

then

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

which, according to the definition of (sub/super)regular functions, indicates that $\Phi_m(P)$ is superregular. Moreover, if we denote:

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

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

We now introduce another function

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

where x_m is the same as defined in $\Phi_m(P)$. According to the definition of (sub/super)regular functions in [17], the x_m, as defined in Eq. (10), renders $\Phi_m(P)$ to be superregular, also makes the $\phi_m(P)$ be subregular.

Moreover, $\phi_m(P)$ meets the boundary conditions of Eq. (9), and therefore, according to the theory of regular functions [17], we have

$$\Gamma_m(P) \geq \phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}}. \tag{11}$$

As Eq. (11) holds for every x_m bounded by Eq. (10), we take the greatest value x_{m_0}. Moreover, as $\lambda \to 0$, $x_{m_0} \to \infty$, whence $\Gamma_m(P) \to 1$. We have thus proved that $Pr\{p_m(\infty) = 1\} \to 1$, showing that the CPA is ε-optimal.

More detailed discussions about this proof and its implications are found in [19].

5 Conclusions

Estimator algorithms are acclaimed to be the fastest Learning Automata (LA), and within this family, the set of *Pursuit* algorithms have been considered to be the pioneering schemes. The ε-optimality of Pursuit algorithms are of great importance and has been studied for years. The proofs in almost all the existing papers have a common flaw which was discovered by the authors of [16], whom we applaud for this.

This paper aims at correcting the flaw by providing a new proof. Rather than examining the monotonicity property of the $\{p_m(t)_{(t>t_0)}\}$ sequence as done in the previous papers and in [16], our current proof studies the *submartingale* property of $\{p_m(t)_{(t>t_0)}\}$. Thereafter, by virtue of the submartingale property and the consequent weaker convergence condition, the new proof invokes the theory of Regular functions, and does not require the learning parameter to decrease gradually.

Further, as opposed to the proof found in [16], we believe that our proof can be easily extended to formally demonstrate the ε-optimality of other Estimator Algorithms, without the requirement of continuously changing the scheme's learning parameter.

References

1. Oommen, B.J., Granmo, O.C., Pedersen, A.: Using stochastic AI techniques to achieve unbounded resolution in finite player goore games and its applications. In: IEEE Symposium on Computational Intelligence and Games, Honolulu, HI (2007)
2. Beigy, H., Meybodi, M.R.: Adaptation of parameters of bp algorithm using learning automata. In: Sixth Brazilian Symposium on Neural Networks, JR, Brazil (2000)
3. Granmo, O.C., Oommen, B.J., Myrer, S.A., Olsen, M.G.: 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 37(1), 166–175 (2007)
4. Unsal, C., Kachroo, P., Bay, J.S.: Multiple stochastic learning automata for vehicle path control in an automated highway system. IEEE Transactions on Systems, Man, and Cybernetics, Part A 29, 120–128 (1999)
5. Oommen, B.J., Roberts, T.D.: Continuous learning automata solutions to the capacity assignment problem. IEEE Transactions on Computers 49, 608–620 (2000)
6. Granmo, O.C.: Solving stochastic nonlinear resource allocation problems using a hierarchy of twofold resource allocation automata. IEEE Transactions Computers 59(4), 545–560 (2010)
7. Oommen, B.J., Croix, T.D.S.: String taxonomy using learning automata. IEEE Transactions on Systems, Man, and Cybernetics 27, 354–365 (1997)
8. Oommen, B.J., Croix, T.D.S.: Graph partitioning using learning automata. IEEE Transactions on Computers 45, 195–208 (1996)
9. Dean, T., Angluin, D., Basye, K., Engelson, S., Aelbling, L., Maron, O.: Inferring finite automata with stochastic output functions and an application to map learning. Maching Learning 18, 81–108 (1995)
10. Thathachar, M.A.L., Sastry, P.S.: Estimator algorithms for learning automata. In: The Platinum Jubilee Conference on Systems and Signal Processing, Bangalore, India, pp. 29–32 (1986)
11. Oommen, B.J., Lanctot, J.K.: Discretized pursuit learning automata. IEEE Transactions on Systems, Man, and Cybernetics 20, 931–938 (1990)
12. Lanctot, J.K., Oommen, B.J.: On discretizing estimator-based learning algorithms. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics 2, 1417–1422 (1991)
13. Lanctot, J.K., Oommen, B.J.: Discretized estimator learning automata. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics 22(6), 1473–1483 (1992)
14. Rajaraman, K., Sastry, P.S.: Finite time analysis of the pursuit algorithm for learning automata. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 26, 590–598 (1996)
15. Oommen, B.J., Agache, M.: Continuous and discretized pursuit learning schemes: various algorithms and their comparison. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 31(3), 277–287 (2001)
16. Ryan, M., Omkar, T.: On ε-optimality of the pursuit learning algorithm. Journal of Applied Probability 49(3), 795–805 (2012)
17. Narendra, K.S., Thathachar, M.A.L.: Learning Automat: An Introduction. Prentice Hall (1989)
18. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58, 13–30 (1963)
19. Zhang, X., Granmo, O.C., Oommen, B.J., Jiao, L.: A Formal Proof of the ε-Optimality of Continuous Pursuit Algorithms Using the Theory of Regular Functions. The Unabridged Version of this Paper (Submitted for Publication. It can be made available to the Referees if needed)