A Formal Proof of the ε-Optimality of *Discretized* Pursuit Algorithms*

Xuan Zhang†, B. John Oommen‡, Ole-Christoffer Granmo§, Lei Jiao¶

Abstract

Learning Automata (LA) can be reckoned to be the founding algorithms on which the field of Reinforcement Learning has been built. Among the families of LA, Estimator Algorithms (EAs) are certainly the fastest, and of these, the family of *discretized* algorithms are proven to converge even faster than their *continuous* counterparts. However, it has recently been reported that the previous proofs for ε-optimality for *all* the reported algorithms *for the past three decades* have been flawed[1]. We applaud the researchers who discovered this flaw, and who further proceeded to rectify the proof for the Continuous Pursuit Algorithm (CPA). The latter proof examines the monotonicity property of the probability of selecting the optimal action, and requires the learning parameter to be continuously changing. In this paper, we provide a new method to prove the ε-optimality of the Discretized Pursuit Algorithm (DPA) which does not require this constraint, by virtue of the fact that the DPA has, in and of itself, absorbing barriers to which the LA can jump in a discretized manner. Unlike the proof given [3] for an absorbing version of the CPA, which utilizes the single-action Hoeffding's inequality, the current proof invokes, what we shall refer to, as the "multi-action" version of the Hoeffding's inequality. We believe that our proof is both unique and pioneering. It can also form the basis for formally showing the ε-optimality of the other EAs that possess absorbing states.

Keywords : *Machine Learning, Learning Automata, Pursuit algorithms, DPA, Convergence, ε-optimality.*

*This work was partially supported by NSERC, the Natural Sciences and Engineering Research Council of Canada. A preliminary version of some of the results of this paper was presented at IEAAIE-2014, the 27th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Kaohsiung, Taiwan, in June 2014 [1]. This paper won the *Best Paper Award* at the conference.

†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.

¶This author can be contacted at: Department of ICT, University of Agder, Grimstad, Norway. E-mail address: lei.jiao@uia.no.

[1]This flaw also renders the finite time analysis of these algorithms [2] to be incorrect. This is because the latter analysis relied on the same condition used in the flawed proofs, i.e., they considered the *monotonicity* property of the probability of selecting the optimal action.

1 Introduction

Learning automata (LA) have been studied as a typical model of reinforcement learning for decades. 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, which triggers either a *stochastic* 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.

Initial LA were designed to be Fixed Structure Stochastic Automata (FSSA), whose state update and decision functions are time invariant. Later, Variable Structure Stochastic Automata (VSSA) were developed, which are characterized by functions that update the probability of selecting the various actions. Representatives of VSSA include the Linear Reward-Penalty (L_{R-P}) scheme, the Linear Reward-Inaction (L_{R-I}) scheme, the Linear Inaction-Penalty (L_{I-P}) scheme and the Linear Reward-εPenalty ($L_{R-\varepsilon P}$) scheme [4]. As one observes, the L_{R-I} and $L_{R-\varepsilon P}$ schemes assign more importance to reward responses than to penalties; they are also ε-optimal in all stationary environments.

According to their Markovian representation, automata fall into two categories: Ergodic automata and automata possessing absorbing barriers. The latter automata get locked into a barrier state – sometimes after even a *finite* number of iterations. Many families of automata that posses absorbing barriers have been reported [4, 5]. Ergodic automata have also been investigated in [4, 5]. These ergodic automata converge in distribution and thus, the asymptotic distribution of the action probability vector has a value that is independent of the corresponding initial vector. While ergodic LA are suitable for non-stationary environments, absorbing automata are preferred in stationary environments. In fact, ergodic automata are known to better adapt to non-stationary environments where the reward probabilities are time dependent.

Among the families of LA, Estimator Algorithms (EAs) work with a noticeably different paradigm. During each learning cycle, these algorithms incorporate an estimation phase, in which they estimate the reward probability of each action. This renders the learning process to be more goal-directed, leading to a much faster convergence and to more accurate learning results [6] [7] [8] [9] when compared to non-estimator algorithms. Within the family of EAs, the set of *Pursuit* Algorithms (PAs) were the pioneering schemes, whose design and analysis were initiated by Thathachar and Sastry [6]. EAs augment an action probability updating scheme with the use of estimates (typically, Maximum Likelihood (ML)) of the actions' reward probabilities. In each iteration, the PA determines the current "Best" action based on the estimates of the reward probabilities, and then pursues *the "Best" action* by linearly increasing *its* action probability. Families of Pursuit and Estimator-based LA have been shown to be faster than VSSA, and the Continuous Pursuit Algorithm (CPA) was the pioneering member of these EAs.

Moving now to a disjoint vein, with respect to the values that the action probabilities can take, the

families of LA typically fall into one of the two categories, namely, Continuous or Discretized. Continuous LA permit the action probabilities to take any value in the interval $[0, 1]$. In practice, the relatively slow rate of convergence of these algorithms constituted a limiting factor in their applicability. In order to increase their speed of convergence, the concept of discretizing the probability space was introduced in [10–12]. This concept is implemented by restricting the probability of choosing an action to be one of a finite number of values in the interval $[0, 1]$. If the values allowed are equally spaced in this interval, the discretization is said to be *linear*, otherwise, the discretization is called *non-linear*. Following the discretization concept, many of the continuous VSSA have been discretized; indeed, discretized versions of almost all continuous automata have been reported [10, 13, 14].

Historically, Oommen and Lanctot [10] 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 within the family of PAs, the continuous version is referred to as the CPA[2]. We briefly mention that discretized versions of all the reported EA schemes have been devised [9, 13, 14].

LA have found applications in a variety of fields, including game playing [15], parameter optimization [16], solving knapsack-like problems and utilizing the solution in web polling and sampling [17]. LA have also been used in vehicle path control [18], assigning capacities in prioritized networks [19], resource allocation [20], string taxonomy [21], graph partitioning [22], and map learning [23]. To exemplify the importance of the DPA, it is worth mentioning an application where one of its variants has been applied. Consider the research field of Cognitive Radio Networks (CRNs), in which the so-called Secondary Users (SUs) attempt to use the channels that are allocated but which are not, at any given time instant, being occupied by the Primary Users (PUs). By virtue of the stochastic property of the PUs, it is meaningful that the SUs attempt to determine the "best" channel, i.e., the one which is being least used by the PUs, and to select this "best" channel for possible package transmission. This is precisely the scheme advocated in [24], where the authors have used a modified version of the DPA, i.e., the Discretized Generalized Pursuit Algorithm (DGPA), to assist the SU to adapt itself to the "best" channel, and to thus achieve smart stochastic channel selection.

1.1 Problem Statement

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, and the details of these proof methodologies for these various families are described extensively in the literature.

[2]PAs have also been extended by allowing them to be of the Reward-Penalty paradigms [13]. We do not consider these here.

Proof Complexity for EAs: Understandably, the most difficult proofs involve the family of EAs. This is because the convergence involves two intertwined phenomena, i.e., the convergence of the reward estimates *and* the convergence of the action probabilities themselves. Ironically, the *combination* of these 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., the sub-optimal actions are not sampled "enough number of times"), the convergence accuracy can be diminished. Hence the dilemma!

Prior Proofs: The ϵ-optimality of the families of PAs have been presented in [2], [13], [10], [11], and [12]. The basic result stated in these papers is that by utilizing a sufficiently small value for the learning parameter (or resolution), both the CPA and the DPA will converge to the optimal action with an arbitrarily large probability.

Flaws in the Existing Proofs: The premise for this paper is that the proofs reported *for almost three decades* for all these schemes 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, which, consequently, renders the time instant, t_0, to also be infinite. If such an "infinite" selection does not occur, the ordering cannot be guaranteed for *all* time instants after t_0. In other words, the authors of these papers misinterpreted the concept ordering "forever" with the ordering "most of the time" after t_0. As a consequence of this misinterpretation, the condition supporting the monotonicity property is false, which leads to an incorrect "proof" for the CPA and DPA being ϵ-optimal[3].

Discovery of the Flaw: Even though this has been the accepted argument for almost *three decades*, (even by the second author of this present paper who was the principal author of many of the earlier papers), we credit the authors of [25] for discovering this flaw. While a detailed exegesis of this is found in [25], in the interest of completeness, a brief explanation on this issue is also included in this paper, in Section 3.

Rationale for this Paper: This paper aims at correcting the above-mentioned flaw found in the earlier proofs by providing a new proof for the convergence of the DPA. As opposed to previous proofs, we will show that because the DPA possesses absorbing barriers, the so-called monotonicity property, though it is sufficient for convergence, it is not really *necessary* for proving that the DPA 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 [4]. Our proof is distinct in principle and argument from the proof reported in [25], which while it is valid for the CPA, also requires that the learning parameter is continuously decreasing.

[3]In addition, like the proofs for the asymptotic convergence, the finite time analysis of both the CPA and the DPA were also done in [2]. Unfortunately, these analyses are also flawed inasmuch as they are also based on the reasoning of the above-mentioned "monotonicity" assumption of the probability of selecting the optimal action.

Proofs for Absorbing Continuous Pursuit Algorithms (ACPAs): An accurate and formal proof for the ACPA's convergence was given in [26] and [3]. This proof is only valid for the ACPA, and it does not require the learning parameter to be continuously decreasing. It also uses what we shall call the "single-action" version of the Hoeffding's inequality [27], invoking which one can bound how much the estimate of any *single* reward probability differs from its true value.

Salient feature of this Present Proof as opposed to the ACPA's: This present paper specifically invokes the "multi-action" version of the Hoeffding's inequality using which one can bound how far the estimate of any *single* reward probability differs from the estimate of the reward probability of any other action. We believe that the latter can be utilized to formally demonstrate the ε-optimality of other absorbing EAs, including, of course, the ACPA.

2 Overview of the DPA

As mentioned earlier, in this paper, we consider only the DPA. The problem of correctly analyzing the discretized versions of the other reported EAs remains open.

We first present the notations used for the DPA:

r: The number of actions.

d_i: The i^{th} element of the reward probability vector D.

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.

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 largest element of $\hat{D}$.

R: The response from the Environment, where $R = 0$ corresponds to a Reward, and $R = 1$ to a Penalty.

Δ: The discretized step size, where $\Delta = \frac{1}{rN}$, with N being a positive integer.

The DPA 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 $\sum_{j=1\ldots r} p_j = 1$. 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 Environment's response as:

$$u_i(t) = u_i(t-1) + (1 - R(t))$$
$$v_i(t) = v_i(t-1) + 1$$
$$\hat{d}_i(t) = \frac{u_i(t)}{v_i(t)}.$$

Thirdly, based on the response of the Environment and the knowledge of the current best action, the DPA

increases the probability of selecting *this* action as per the Discretized L_{R-I} rules. So if $\hat{d}_h(t)$ is the largest element in $\hat{D}(t)$, we update $p(t)$ as:

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

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

EndIf

We now visit the proofs of the DPA's convergence.

3 Previous "Proofs" for DPA's ε-optimality

The formal assertion of the ε-optimality of the DPA [13] is stated in Theorem 1, where, 't' is measured in terms of the number of iterations.

Theorem 1 *Given any* $\varepsilon > 0$ *and* $\delta > 0$, *there exist an* $N_0 > 0$ *and a* $t_0 < \infty$ *such that for all time* $t \geq t_0$ *and for any positive learning parameter* $N > N_0$,

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

The earlier reported proofs for the ε-optimality of the CPA and the DPA follow the same strategy, which consists of four steps[4]. Firstly, given a sufficiently small (large) value for the learning parameter λ (N), 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. In that case, 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 $p_m(t) \to 1$ with probability (w.p.) 1, ε-optimality is proven based on the axiom of total probability.

The formal assertions of these steps are catalogued below.

1. The first step of the proof can be described mathematically by Theorem 2.

[4]We state the conditions and parameters for the CPA, while the analogous counterpart conditions and parameters for the DPA are stated in parenthesis to avoid repetition. The reader should observe that we, really, did not have to mention the conditions and parameters for the CPA. But we have opted to show it because the proof given in [25], which demonstrated the flaw, is based on the CPA, and we believe that this will improve the readability of the present paper. But this can be omitted if requested by the Referees.

Theorem 2 *For any given constants $\delta > 0$ and $M < \infty$, there exist an $N_0 > 0$ and $t_0 < \infty$ such that under the DPA algorithm, for all positive $N > N_0$,*

$$Pr\{\textit{All actions are selected at least M times each before time } t_0\} > 1-\delta,\ \forall t > t_0.$$

The detailed proof for this result can be found in [2], [13] and [6].

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:

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

$$E[p_m(t+1) - p_m(t)|\bar{K}(t_0)] = \begin{cases} d_m\lambda(1-p_m(t)) \geq 0,\ t > t_0, \text{ for the CPA} \\ p_m(t) + d_m c_t \Delta \geq 0,\ t > t_0, \text{ for the DPA,} \end{cases} \tag{2}$$

where $c_t = 1, 2, ..., r-1$, and $\bar{K}(t_0)$ is the condition that $\hat{d}_m$ remains the greatest element in $\hat{D}$ after time t_0.

It is worth mentioning that the combination of $|p_m(t)| \leq 1$ and $E[p_m(t+1) - p_m(t)|\bar{K}(t_0)] \geq 0$ in Eq. (1) and Eq. (2) makes the sequence $\{p_m(t)_{t>t_0}\}$ a submartingale, which is a weaker convergence than $\{p_m(t)_{t>t_0}\}$ being monotonically increasing. However, the result of the expectation, i.e., the right hand side of Eq. (2), was obtained by invoking the condition $\bar{K}(t_0)$, which indeed ensures the *monotonicity* property of the sequence of $\{p_m(t)_{t>t_0}\}$. Now that $\bar{K}(t_0)$ is invoked, it is, in fact, unnecessary to examine the submartingale property, as the monotonicity property can be simply proven by

$$[p_m(t+1) - p_m(t)]|\bar{K}(t_0) = \begin{cases} d_m\lambda(1-p_m(t)) \geq 0,\ t > t_0, \text{ for the CPA} \\ p_m(t) + d_m c_t \Delta \geq 0,\ t > t_0, \text{ for the DPA.} \end{cases}$$

Since $\{p_m(t)_{(t>t_0)}\}$ can be proven to be monotonically increasing based on the condition, $\bar{K}(t_0)$, $p_m(t)$ converges to 1 w.p. 1.

3. As $p_m(t) \to 1$ w.p. 1, if it can, indeed, be proven that $Pr\{\bar{K}(t_0)\} > 1-\delta$, by the axiom of total probability, one can then see that:

$$Pr\{p_m(t) > 1-\varepsilon\} \geq Pr\{p_m(t) \to 1\} Pr\{\bar{K}(t_0)\} > 1 \cdot (1-\delta) = 1-\delta,$$

and ε-optimality is proven.

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

$$Pr\{\hat{d}_m(t) > \hat{d}_j(t)_{j\neq m, \forall t: t>t_0}\} > 1-\delta. \tag{3}$$

In the proofs reported in the literature, Eq. (3) is considered to be true if the following result, formalized in Theorem 3, is true under the CPA and DPA algorithms.

Theorem 3 *Let $n_i(t)$ be the number of times α_i has been selected up to time t, and w be the difference between the two* highest *reward probabilities. Suppose that for a given $\delta > 0$, and for all $i \in (1...r)$, there exists an $M_i < \infty$, such that if α_i is selected at least M_i times,*

$$Pr\{|\hat{d}_i(t) - d_i| < \tfrac{w}{2}\} > 1 - \delta.$$

Then, as per Theorem 1, if we let $M = max_{1 \le i \le r}\{M_i\}$, and for all $t > t_0$, if $min_{1 \le i \le r}\{n_i(t)\} > M$,

$$Pr\{|\hat{d}_i(t) - d_i| < \tfrac{w}{2}\} > 1 - \delta.$$

The rationale of Theorem 3 is that since $n_i(t)$ is the number of times α_i has been selected up to time t, and w is the difference between the two *highest* reward probabilities, it implies that if all actions are selected at least M times, each of the $\hat{d}_i$ will be in a $\frac{w}{2}$ neighborhood of d_i with an arbitrarily large probability. In other words, the probability of $\hat{d}_m(t)$ being greater than $\hat{d}_j(t)_{j \neq m}$ will be arbitrarily close to unity. This result can be easily "proven" by the weak law of large numbers, and the "proof" can be found in [13].

However, there is a flaw in the above argument. In fact, Theorem 3 does not guarantee $Pr\{\bar{K}(t_0)\} > 1 - \delta$. To be specific, let us define

$$K(t) = \{\hat{d}_m(t) \text{ is the largest element in } \hat{D}(t)\}.$$

Then the result that can be deduced from Theorem 3 is that $Pr\{K(t)\} > 1 - \delta$ when $t > t_0$. But, indeed, the condition which Eq. (2) is based on is:

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

which means that for every single time instant in the future, i.e., $t > t_0$, $\hat{d}_m(t)$ needs to be the largest element in $\hat{D}(t)$. The flaw in the previous proofs reported in the literature is that the authors made a mistake by reckoning that $[Pr\{K(t)\} > 1 - \delta]_{(t > t_0)}$ is equivalent to $\bar{K}(t_0)$. This renders the existing proofs for the CPA and the DPA being ε-optimal, to be incorrect.

The flaw is documented in [25], which focused on the CPA, and further provided a way of correcting the flaw, i.e., by proving $Pr\{\bar{K}(t_0)\} > 1 - \delta$ instead of proving $Pr\{K(t)\}_{(t > t_0)} > 1 - \delta$. Although their proof requires a sequence of *decreasing* values for the learning parameter λ, to the best of our knowledge, it currently stands as the only correct way to prove the ε-optimality of the CPA. We applaud the authors of [25] for discovering this flaw, and for submitting an accurate proof for the CPA for the scenario when the λ's are changing with time.

However, the proof methodology that we have used here for the DPA is quite distinct (and uses completely different techniques) than the proof reported in [25]. The reasons why we have sought an alternate proof are the following:

1. The monotonicity property which all the previous flawed proofs and the proof in [25] were based on, is, indeed, a very strong condition. The condition requires that $\hat{d}_m(t)$ is ranked as the largest element in $\hat{D}(t)$ *at every single point of time* for all $t > t_0$, which, in turn, requires that for the CPA to achieve its ε-optimality, one must rely on an additional external assumption that the learning parameter, λ, is gradually decreasing during the learning process. Though there is, currently, no way to circumvent this external constraint so as to prove the CPA's ε-optimality, the essential difference between the DPA and the CPA, i.e., that the former possesses states that are explicitly absorbing to which the LA can *jump* to, makes it possible for us to prove the DPA's ε-optimality without requiring the constraint of decreasing the learning parameter over time.

2. In our earlier proof for the convergence of the ACPA [3], we had relied on the use of Hoeffding's inequality [27]. The application of the inequality in such a setting was only able to bound how much the estimate of any *single* reward probability differs from its true value. We refer to this as the "single-action" version of the Hoeffding's inequality. The present application of the inequality is able to bound how far the estimate of any *single* reward probability (for example, of the "best" action) differs from the estimate of the reward probability of any other action (i.e., of the "second best" action). We refer to this as the "multi-action" version of the Hoeffding's inequality. This version is far more powerful, but by the same token, it is also more difficult to both apply and invoke. This version of the inequality is fundamental to our present proof, and we believe that the consequent properties can be utilized to formally and more elegantly demonstrate the ε-optimality of other EAs, including the ACPA.

In the next two sections, we shall correct the above-mentioned flaw that exists in the previous proofs of EAs. We do this by providing a new proof strategy for the ε-optimality of the DPA, and which does not require that the learning parameter, N, is gradually increased. The 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 DPA's ε-optimality: A New Proof

Our proof for the DPA's ε-optimality consists of four steps, which we first explain informally here. Firstly, we prove that by properly setting the learning parameter, N, each action will be selected a large number of times within a finite time instant. Secondly, based on the fact that each action has been selected (or probed) a large number of times, according to the law of large numbers, the probability that the estimate of each reward probability being within a narrow neighborhood of its true value, can be made arbitrarily large. In other words, the reward estimation for each action is characterized by an arbitrarily high accuracy. Thirdly, we prove that the sequence of $\{p_m(t)\}$, the probability of choosing the best action, with t being greater than a certain time instant, is a submartingale, given the condition that the accuracy of the estimation of the reward probability is arbitrarily high. Finally, by invoking the submartingale convergence theory and the theory of

Regular functions, $\{p_m(t)\}$ can be proven to converge to unity in probability, implying the ε-optimality of the DPA.

Each of the steps of the proof will be detailed in the following subsections.

4.1 The Moderation Property of the DPA

The property of moderation can be described precisely by Theorem 2, which has been proven in [10]. This implies that under the DPA, by utilizing a sufficiently large value for the learning parameter, N, each action will be selected an arbitrarily large number of times.

4.2 The Key Condition $\bar{G}(t_0)$ 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*. By doing this, the previously mentioned strong condition required by the authors of [25] represented by $\bar{K}(t_0)$, i.e., of ranking $\hat{d}_m(t)$ as the largest element in $\hat{D}(t)$ at *every single time instant* after time t_0, will not be necessary any longer. Instead, we base our arguments on the weaker *submartingale* phenomenon, $\bar{G}(t_0)$, defined as follows[5]:

$$
\begin{aligned}
&q_j(t) = Pr\{\hat{d}_m(t) > \hat{d}_j(t), j \neq m\},\\
&q(t) = Pr\{\hat{d}_m(t) > \hat{d}_j(t), \forall j \neq m\} = \prod_{j \neq m} q_j(t), \qquad (4)\\
&G(t) = \{q(t) > 1-\delta\}, \delta \in (0,1),\\
&\bar{G}(t_0) = \{\bigcap_{t>t_0} \{q(t) > 1-\delta\}\}, \delta \in (0,1). \qquad (5)
\end{aligned}
$$

Note that $\bar{K}(t_0)$ is stronger than $\bar{G}(t_0)$ in the sense that $\bar{K}(t_0) = \{\bigcap_{t>t_0} \{\hat{d}_m(t) > \hat{d}_j(t), \forall j \neq m\}\}$, is stronger than $\{\bigcap_{t>t_0} \{q(t) = 1\}\}$, which is, in turn, stronger than $\bar{G}(t_0)$.

Our goal in this step is to prove the following result, formulated in Theorem 4.

Theorem 4 *Given a $\delta \in (0,1)$, there exists a time instant $t_0 < \infty$, such that the condition $\bar{G}(t_0)$ holds. In other words, for this given δ, there exists a $t_0 < \infty$, such that $\forall t > t_0$:*

$$q(t) > 1-\delta. \qquad (6)$$

Proof: First of all, to make the proof easier and to help clarify arguments, we initially consider a two-action Environment, whence the enhanced arguments for the r-action Environment can be generalized. With-

[5] In the interest of simplicity, at this juncture we have assumed in Eq. (4) that the $\hat{d}_j$'s are independent of each other. We believe that this assumption can be easily relaxed by considering only the individual d_j's and not all of them together.

out loss of generality, let α_1 be the optimal action and α_2 the inferior one. We are to prove that:

$$Pr\{\hat{d}_1(t) - \hat{d}_2(t) > 0\} > 1 - \delta. \tag{7}$$

If we further define

$$\begin{aligned} H &= d_1 - d_2, \text{ and} \\ \hat{H}(t) &= \hat{d}_1(t) - \hat{d}_2(t), \end{aligned} \tag{8}$$

then

$$\begin{aligned} Pr\{\hat{d}_1(t) - \hat{d}_2(t) > 0\} &\Leftrightarrow 1 - Pr\{\hat{H}(t) - H \leq -H\}, \text{ and} \\ Pr\{\hat{d}_1(t) - \hat{d}_2(t) > 0\} > 1 - \delta &\Leftrightarrow Pr\{\hat{H}(t) - H \leq -H\} \leq \delta. \end{aligned} \tag{9}$$

Hence, we can equivalently prove

$$Pr\{\hat{H}(t) - H \leq -H\} \leq \delta. \tag{10}$$

If we denote $n_1(t)$ as the number of times α_1 has been selected up to time t, by invoking the "two-action" version of Hoeffding's inequality [27], we have:

$$Pr\{\hat{H}(t) - H \leq -H | n_1(t) = n\} \leq e^{-\frac{2H^2}{n^{-1} + (t-n)^{-1}}}. \tag{11}$$

We thus have to find an appropriate value for n such that

$$e^{-\frac{2H^2}{n^{-1} + (t-n)^{-1}}} \leq \delta, \tag{12}$$

which guarantees that $Pr\{\hat{H}(t) - H \leq -H\} \leq \delta$.

1. It is easy to see that

$$\begin{aligned} & e^{-\frac{2H^2}{n^{-1} + (t-n)^{-1}}} \leq \delta \\ \Longleftrightarrow & -\frac{2H^2}{n^{-1} + (t-n)^{-1}} \leq \ln \delta \\ \Longleftrightarrow & 2H^2 n^2 - 2H^2 tn - t \ln \delta \leq 0. \end{aligned} \tag{13}$$

   Consider now the equation

$$2H^2 n^2 - 2H^2 tn - t \ln \delta = 0, \tag{14}$$

where n is the variable to be solved for.

One can easily observe that Eq. (14) is a quadratic equation of n. By applying the formula for obtaining the roots of this quadratic equation, we can derive the expressions for its two *real* roots of Eq. (14) as:

$$\begin{aligned} n_{r1} &= \frac{t}{2} - \frac{\sqrt{H^2t^2 + 2t\ln\delta}}{2H}, \text{ and} \\ n_{r2} &= \frac{t}{2} + \frac{\sqrt{H^2t^2 + 2t\ln\delta}}{2H}. \end{aligned} \tag{15}$$

From Eq. (15), we see that if

$$H^2t^2 + 2t\ln\delta < 0, \tag{16}$$

i.e., if

$$t < \frac{-2\ln\delta}{H^2}, \tag{17}$$

then Eq. (14) has no real roots. Besides, as the quadratic coefficient $2H^2 > 0$, we have that for all n,

$$2H^2n^2 - 2H^2tn - t\ln\delta > 0. \tag{18}$$

Consequently, we conclude that

$$e^{-\frac{2H^2}{n^{-1}+(t-n)^{-1}}} > \delta \text{ when } t < \frac{-2\ln\delta}{H^2}. \tag{19}$$

Conversely, if $t \geq \frac{-2\ln\delta}{H^2}$,

$$e^{-\frac{2H^2}{n^{-1}+(t-n)^{-1}}} = \begin{cases} \leq \delta, & \text{when } n_{r1} \leq n \leq n_{r2}, \\ > \delta, & \text{otherwise.} \end{cases} \tag{20}$$

2. We now investigate $n_{r1} \leq n \leq n_{r2}$, under the condition of $t \geq \frac{-2\ln\delta}{H^2}$. Briefly speaking, to make sure that Eq. (12) holds, we need to find an appropriate value for n such that $\forall t : t \geq \frac{-2\ln\delta}{H^2}$, n is greater than n_{r1} and at the same time, n is less than or equal to n_{r2}.

   Firstly, we investigate $n_{r1} \leq n$. Consider the function

$$f_1(t) = n_{r1} = \frac{t}{2} - \frac{\sqrt{H^2t^2 + 2t\ln\delta}}{2H}. \tag{21}$$

If we define $f_1'(t) = \frac{\partial f_1(t)}{\partial t}$, then

$$\begin{aligned} f_1'(t) &= \frac{1}{2} - \frac{H^2 t + \ln\delta}{2H\sqrt{H^2t^2 + 2t\ln\delta}} \\ &= \frac{1}{2}\left(1 - \frac{H^2 t + \ln\delta}{\sqrt{H^4t^2 + 2tH^2\ln\delta}}\right) \\ &= \frac{1}{2}\left(1 - \frac{H^2 t + \ln\delta}{\sqrt{(H^2t + \ln\delta)^2 - (\ln\delta)^2}}\right). \end{aligned} \tag{22}$$

Since

$$\frac{H^2 t + \ln\delta}{\sqrt{(H^2t + \ln\delta)^2 - (\ln\delta)^2}} > 1, \tag{23}$$

we can see that $f_1'(t) < 0$, i.e., $f_1(t)$ decreases monotonically as t grows. The maximum of $f_1(t)$ thus occurs when $t = \frac{-2\ln\delta}{H^2}$. This leads us to the conclusion that $\forall t \geq \frac{-2\ln\delta}{H^2}$,

$$n_{r1} = f_1(t) \leq f_1(\frac{-2\ln\delta}{H^2}) = \frac{-\ln\delta}{H^2}. \tag{24}$$

Consequently, if we choose

$$n \geq \frac{-\ln\delta}{H^2}, \tag{25}$$

i.e., α_1 is selected more than $\frac{-\ln\delta}{H^2}$ times, then $\forall t > \frac{-2\ln\delta}{H^2}$, $n \geq n_{r1}$.

Secondly, we investigate $n \leq n_{r2}$. To accomplish this, we consider the analogous function

$$f_2(t) = n_{r2} = \frac{t}{2} + \frac{\sqrt{H^2t^2 + 2t\ln\delta}}{2H}. \tag{26}$$

Arguing in a manner analogous to the above, we observe that

$$f_2'(t) = \frac{\partial f_2(t)}{\partial t} = \frac{1}{2}\left(1 + \frac{H^2 t + \ln\delta}{\sqrt{(H^2t + \ln\delta)^2 - (\ln\delta)^2}}\right) > 0. \tag{27}$$

Hence $f_2(t)$ increases monotonically as t grows and the minimum of $f_2(t)$ occurs when $t = \frac{-2\ln\delta}{H^2}$. In other words, $\forall t \geq \frac{-2\ln\delta}{H^2}$,

$$n_{r2} = f_2(t) \geq f_2(\frac{-2\ln\delta}{H^2}) = \frac{-\ln\delta}{H^2}. \tag{28}$$

If we further define:

$$f_3(t) = t, \tag{29}$$

we have

$$f_3'(t) = \frac{\partial f_3(t)}{\partial t} = 1, \text{ and}$$
$$f_2'(t) - f_3'(t) = \frac{1}{2}\left(\frac{H^2 t + \ln\delta}{\sqrt{(H^2 t + \ln\delta)^2 - (\ln\delta)^2}} - 1\right) > 0, \tag{30}$$

implying that $\frac{\partial f_2(t)}{dt} > \frac{\partial f_3(t)}{dt}$. Consequently, as t increases, the increment of n_{r2} is always greater than that of t. Moreover, if we increase t by unity, n will be either increased by unity or remain the same, as α_1 will be either selected or not selected. This further implies that the speed of the increment of t is greater than that of n. This reasoning leads to the conclusion that:

(a) when $t \geq \frac{-2\ln\delta}{H^2}$, $n_{r2} \geq \frac{-\ln\delta}{H^2}$,

(b) and when t increases, n_{r2} increases faster than n.

Therefore, we can see that $\forall t \geq \frac{-2\ln\delta}{H^2}$, we need n to only satisfy Eq. (25), i.e., $n \geq \frac{-\ln\delta}{H^2}$, to ensure that $n_{r1} \leq n \leq n_{r2}$.

Note that the above analysis is also applicable to α_2, which is, indeed, symmetric to α_1 in this two action environment considered. In other words, if we substitute α_1 with α_2, and further define $n_1(t)$ as the number of times α_2 has been selected within time t, the corresponding results with regard to n can be directly applied to α_2. The consequence of the above arguments is the following: Let us suppose that we define the time instant t_0 such that within the time defined by t_0, α_1 and α_2 have each been selected more than $\left\lceil \frac{-\ln\delta}{H^2} \right\rceil$ times. In that case:

$$e^{-\frac{2H^2}{n^{-1}+(t-n)^{-1}}} \leq \delta,$$

whence we can conclude that for the given $\delta \in (0,1)$, $\forall t > t_0$,

$$q(t) = Pr\{\hat{d}_1(t) - \hat{d}_2(t) > 0\} > 1 - \delta.$$

The result follows because the above arguments can be easily seen to be true for any r-action Environment inasmuch as it is true for every pair of actions[6]. Theorem 4 is thus proven.

[6]If one is interested in pursuing the general r-action scenario in greater detail without invoking the 2-action results, the arguments involved are almost identical, except that the algebra is a little more cumbersome. Without going into the detailed algebraic manipulations, we can submit the arguments as follows. For the specific Environment, we define:

$$H_j = d_m - d_j, j \neq m,$$
$$\hat{H}_j(t) = \hat{d}_m(t) - \hat{d}_j(t), j \neq m.$$

Then, given any $\delta \in (0,1)$, if we denote $\delta^\star = 1 - \sqrt[r-1]{1-\delta}$, we can show that there exists a time instant t_0, such that within the time defined by t_0, α_m has been selected more than $\left\lceil \frac{-\ln\delta^\star}{(min\{H_j\})^2} \right\rceil$ times, and $\alpha_{j,(j\neq m)}$ has been selected more than $\left\lceil \frac{-\ln\delta^\star}{H_j^2} \right\rceil$ times.

Table 1: The various possibilities for updating p_m for the next iteration under the DPA.

	Responses	The greatest element in $\hat{D}$	Updating p_m
$p_m(t+1)$	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)	$p_m(t)$

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

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

Theorem 5 *Under the DPA, the quantity $\{p_m(t)_{t>t_0}\}$ is a submartingale.*

Proof: If we denote a sequence of random variables as $X_1, X_2, ..., X_t, ...$, then the sequence is a submartingale, if it satisfies the property that for any time instant t,

$$E[X_t] < \infty, \text{ and}$$
$$E[X_{t+1}|X_1, X_2, ..., X_t] \geq X_t.$$

Firstly, as $p_m(t)$ is a probability, we have

$$E[p_m(t)] \leq 1 < \infty. \tag{31}$$

Secondly, we explicitly calculate $E[p_m(t)]$. Using the DPA's updating rule, we can describe the update of $p_m(t)$ as per Table 1. Thus, we have:

$$\begin{aligned}
&E[p_m(t+1)|P(t)] \\
=&\sum_{j=1...r} p_j\left(d_j\left(q(p_m+c_t\Delta)+(1-q)(p_m-\Delta)\right)+(1-d_j)p_m\right) \\
=&\sum_{j=1...r}(p_j d_j q c_t\Delta) - \sum_{j=1...r} p_j d_j\Delta + \sum_{j=1...r} p_j d_j q\Delta + \sum_{j=1...r} p_j p_m \\
=&p_m + \sum_{j=1...r} p_j d_j\left(q(c_t\Delta+\Delta)-\Delta\right).
\end{aligned} \tag{32}$$

In the above, $p_m(t)$ and $q(t)$ are respectively written as p_m and q in the interest of conciseness. The difference between $E[p_m(t+1)]$ and $p_m(t)$ can be expressed as:

$$\begin{aligned}
Diff(t) &= E[p_m(t+1)|P(t)] - p_m(t) \\
&= \sum_{j=1...r} p_j(t) d_j\left(q(t)(c_t\Delta+\Delta)-\Delta\right).
\end{aligned} \tag{33}$$

Consequently, for $\forall t > t_0$, $q_j(t) > 1-\delta^\star$ and $q(t) \geq \prod_{j=1...r, j\neq m} q_j(t) > 1-\delta$.

Given that $p_j(t) > 0$ and $d_j > 0$, if we denote $Z_t = \frac{\Delta}{c_t\Delta+\Delta} = \frac{1}{c_t+1}$, we see that if $\forall t > t_0, \quad q(t) > Z_t$, then, $Diff(t) > 0$, and the sequence $\{p_m(t)_{t>t_0}\}$ is a submartingale.

As per the action probability updating rules of the DPA, $c_t = 1, 2, ..., r-1$, implying that $Z_t \in [\frac{1}{r}, \frac{1}{2}]$. Let

$$1 - \delta = max\{Z_t\} = \frac{1}{2}, \tag{34}$$

then, according to Theorem 4, there exists a time instant t_0 such that $\forall t > t_0$,

$$q(t) > \tfrac{1}{2} \geq Z_t.$$

Consequently, $\{p_m(t)_{t>t_0}\}$ is a submartingale, and the theorem is proven.

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

We now prove the ε-optimality of the DPA.

Theorem 6 *The DPA is ε-optimal in all stationary random Environments. More formally, given any $1 - \delta \geq \frac{1}{2}$, there exists a positive integer $N_0 < \infty$ and a time instant $t_0 < \infty$, such that for all resolution parameters $N > N_0$ and for all $t > t_0$, the quantities $q(t) > 1 - \delta$, and $Pr\{p_m(\infty) = 1\} \to 1$.*

Proof: According to the submartingale convergence theory [4],

$$p_m(\infty) = 0 \text{ or } 1. \tag{35}$$

If we denote e_j as the unit vector with the j^{th} element being unity, then $p_m(\infty) = 1$ is equivalent to the assertion that $P(\infty) = e_m$. If we define the convergence probability

$$\Gamma_m(P) = Pr\{P(\infty) = e_m | P(0) = P\}, \tag{36}$$

our task is to now prove:

$$\Gamma_m(P) \to 1. \tag{37}$$

To prove Eq. (37), we shall use the theory of Regular functions, and the arguments used follow the lines of the arguments found in [4] for the convergence proofs of Absolutely Expedient schemes.

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

$$U\Phi(P) = E[\Phi(P(n+1))|P(n) = P]. \tag{38}$$

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

$$U^n\Phi(P) = E[\Phi(P(n))|P(0) = P]. \tag{39}$$

We refer to the function $\Phi(P)$ as being:

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

$$\Phi(P) \geq U\Phi(P) \geq U^2\Phi(P) \geq ... \geq U^\infty\Phi(P). \tag{40}$$

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

$$\Phi(P) \leq U\Phi(P) \leq U^2\Phi(P) \leq ... \leq U^\infty\Phi(P). \tag{41}$$

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

$$\Phi(P) = U\Phi(P) = U^2\Phi(P) = ... = U^\infty\Phi(P). \tag{42}$$

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

$$\Phi(e_m) = 1 \text{ and } \Phi(e_j) = 0, (\text{for } j \neq m), \tag{43}$$

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

$$\begin{aligned} U^\infty\Phi(P) &= E[\Phi(P(\infty))|P(0) = P] \\ &= \sum_{j=1}^{r} \Phi(e_m) Pr\{P(\infty) = e_j|P(0) = P\} \\ &= Pr\{P(\infty) = e_m|P(0) = P\} \\ &= \Gamma_m(P). \end{aligned} \tag{44}$$

Comparing Eq. (44) with Eq. (42), 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. (40) and Eq. (41) 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. (43), which will then guarantee to bound $\Gamma_m(P)$ from below.

To find such a Subregular function of P, we will firstly find a corresponding Superregular of P. Consider a specific instantiation of Φ to be the function Φ_m, defined below as:

$$\Phi_m(P) = e^{-x_m p_m}, \tag{45}$$

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

$$\begin{aligned} U(\Phi_m(P)) - \Phi_m(P) &= E[\Phi_m(P(n+1))|P(n)=P] - \Phi_m(P) \\ &= E[e^{-x_m p_m(n+1)}|P(n)=P] - e^{-x_m p_m} \\ &= \sum_{j=1\ldots r} e^{-x_m(p_m+c_t\Delta)} p_j d_j q + \sum_{j=1\ldots r} e^{-x_m(p_m-\Delta)} p_j d_j (1-q) \\ &+ \sum_{j=1\ldots r} e^{-x_m p_m} p_j (1-d_j) - e^{-x_m p_m} \\ &= \sum_{j=1\ldots r} p_j d_j e^{-x_m p_m} \left(q(e^{-x_m c_t \Delta} - e^{x_m \Delta}) + (e^{x_m \Delta} - 1) \right). \end{aligned} \tag{46}$$

Our task is to determine a proper value for x_m such that $\Phi_m(P)$ is Superregular, i.e.,

$$U(\Phi_m(P)) - \Phi_m(P) \leq 0. \tag{47}$$

This is equivalent to solving the following inequality:

$$q(e^{-x_m c_t \Delta} - e^{x_m \Delta}) + (e^{x_m \Delta} - 1) \leq 0. \tag{48}$$

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

$$b^x \doteq 1 + (\ln b)x + \frac{(\ln b)^2}{2} x^2. \tag{49}$$

If we set $b = e^{-x_m}$, when $\Delta \to 0$, Eq. (48) can be re-written as

$$q\left((\ln b)(c_t+1)\Delta + \frac{(\ln b)^2}{2}(c_t^2-1)^2\Delta^2 \right) - (\ln b)\Delta + \frac{\ln b^2}{2}\Delta^2 \leq 0. \tag{50}$$

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

$$x_m \left(x_m - \frac{2(q(c_t+1)-1)}{\Delta(q(c_t^2-1)+1)} \right) \leq 0. \tag{51}$$

As x_m is defined as a positive constant, we have

$$0 < x_m \leq \frac{2(q(c_t+1)-1)}{\Delta(q(c_t^2-1)+1)}. \tag{52}$$

If we denote

$$x_{m_0} = \frac{2(q(c_t+1)-1)}{\Delta(q(c_t^2-1)+1)}, \tag{53}$$

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 $\Delta \to 0$, $x_{m_0} \to \infty$.

We now introduce another function

$$\phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}}, \tag{54}$$

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 Subregular (Superregular) [4]. Therefore, the x_m, as defined in Eq. (52), which renders $\Phi_m(P)$ to be Superregular, causes the function $\phi_m(P)$ to be Subregular.

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

$$\phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}} = \begin{cases} 1, \text{ when } P = e_m, \\ 0, \text{ when } P = e_j. \end{cases} \tag{55}$$

Therefore, according to Eq. (41),

$$\Gamma_m(P) \geq \phi_m(P) = \frac{1-e^{-x_m p_m}}{1-e^{-x_m}}. \tag{56}$$

As Eq. (56) holds for every x_m bounded by Eq. (52), we can choose the largest value x_{m_0}, and when $x_{m_0} \to \infty$, $\Gamma_m(P) \to 1$. We have thus proved that $Pr\{p_m(\infty) = 1\} \to 1$ under the DPA, implying its ε-optimality.

Remark: Having completed the proof of the DPA's ε-optimality, we are able to give firm figures for t_0 and the number of times each action needs to be selected. To actually determine the value of t_0, we summarize the result of the above arguments as follows: Let $\delta^\star = 1 - \sqrt[r-1]{1-\delta}$, where δ is the quantity specified in the statement of Theorem 6. Then there exists a time instant, t_0, such that

$$t_0 > \left\lceil \frac{-\ln \delta^\star}{(min\{H_j\})^2} \right\rceil + \sum_{j \neq m} \left\lceil \frac{-\ln \delta^\star}{H_j^2} \right\rceil, \tag{57}$$

and up to the time instant specified by t_0, we can guarantee that α_m has been selected more than $\left\lceil \frac{-\ln \delta^\star}{(min\{H_j\})^2} \right\rceil$ times, and $\alpha_{j,(j\neq m)}$ has been selected more than $\left\lceil \frac{-\ln \delta^\star}{H_j^2} \right\rceil$ times, guaranteeing all the conditions imposed by

the corresponding theorems.

5 The Difference between the Proofs Requiring Monotonicity and the Submartingale Properties

To highlight the difference between the proof presented in this paper and the proof in [25], we define

$$\bar{C}(t_0) = \{\bigcap_{t>t_0} \{q(t) = 1\}\}. \tag{58}$$

If we compare Eq. (58) with the definition of $\bar{G}(t_0)$ in Eq. (5), we see that if we enforce the value $\delta = 0$, $\bar{G}(t_0)$ becomes equivalent to $\bar{C}(t_0)$, and the *submartingale* property of $\{p_m(t)_{(t>t_0)}\}$ becomes precisely the *monotonicity* property. Accordingly, we see that the condition $\delta^\star = 0$ yields the lower bound of t_0 to become:

$$t_0 > \left\lceil \frac{-\ln \delta^\star}{(min\{H_j\})^2} \right\rceil + \sum_{j \neq m} \left\lceil \frac{-\ln \delta^\star}{H_j^2} \right\rceil = \infty. \tag{59}$$

Consequently, there will be no such time instant $t_0 < \infty$, after which $q(t)_{(t>t_0)} = 1$. This is precisely the reason why in the case of the CPA in [25], one requires an additional assumption that the learning parameter λ has to be gradually decreasing, as explained in greater detail in [25].

It should therefore be very clear to the reader that the analysis and the new proof presented here are significantly different than the corresponding analysis and proof in [25]. They are based on the weaker condition $\bar{G}(t_0)$ instead of $\bar{K}(t_0)$, because of which the ε-optimality does not require that the scheme's learning parameter gradually decreases.

6 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 (PAs) are of great importance and has been studied for decades. The convergence proofs for the PAs in all the reported papers have a common flaw which was discovered by the authors of [25], whom we applaud. This paper corrects the flaw and provides a new proof for the Discretized Pursuit Algorithm (DPA).

Rather than examining the monotonicity property of the $\{p_m(t)_{(t>t_0)}\}$ sequence as done in the previous papers and in [25], our current proof studies the *submartingale* property of $\{p_m(t)_{(t>t_0)}\}$. Thereafter, by virtue of the submartingale property and the weaker condition, the new proof invokes the theory of Regular functions, and does not require the resolution/parameter to decrease/increase gradually.

Our analysis constitutes the only result for the DPA. We submit that it is both novel and pioneering. Further, as opposed to the proof found in [25], we do not require the parameter to change continuously.

Also, since we have invoked the "multi-action" version of Hoeffding's inequality, we believe that our proof can be extended to formally demonstrate the ε-optimality of other EAs which possess absorbing states.

The formal "corrected" proof for the finite time analysis of the DPA [2] remains open. It is currently being investigated.

References

[1] 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 2014*. Kaohsiung, Taiwan: Springer, Jun. 2014.

[2] 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.

[3] 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," *to appear in Applied Intelligence*, 2014.

[4] K. S. Narendra and M. A. L. Thathachar, *Learning Automata: An Introduction*. Prentice Hall, 1989.

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

[6] 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.

[7] M. Agache and B. J. Oommen, "Generalized pursuit learning schemes: new families of continuous and discretized learning automata," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 32, no. 6, pp. 738–749, 2002.

[8] 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.

[9] ——, "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.

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

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

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

[13] 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.

[14] X. Zhang, O.-C. Granmo, and B. J. Oommen, "Discretized Bayesian pursuit - a new scheme for reinforcement learning," in *Proceedings of IEA-AIE 2012*, Dalian, China, Jun. 2012, pp. 784–793.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] O.-C. Granmo, "Solving stochastic nonlinear resource allocation problems using a hierarchy of twofold resource allocation automata," *IEEE Transactions on Computers*, vol. 59, no. 4, pp. 545–560, 2010.

[21] 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.

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

[23] 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.

[24] Y. Song, Y. Fang, and Y. Zhang, “Stochastic channel selection in cognitive radio networks,” in *Proceedings of IEEE Global Telecommunications Conference*, Washington DC, USA, Nov. 2007, pp. 4878–4882.

[25] 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.

[26] 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.

[27] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.