An Estimation Based Allocation Rule with Super-linear Regret and Finite Lock-on Time for Time-dependent Multi-armed Bandit Processes

Prokopis C. Prokopiou, Peter E. Caines and Aditya Mahajan

***Abstract*— The multi-armed bandit (MAB) problem has been an active area of research since the early 1930s. The majority of the literature restricts attention to i.i.d. or Markov reward processes. In this paper, the finite-parameter MAB problem with time-dependent reward processes is investigated. An upper confidence bound (UCB) based index policy, where the index is computed based on the maximum-likelihood estimate of the unknown parameter, is proposed. This policy locks on to the optimal arm in finite expected time but has a super-linear regret. As an example, the proposed index policy is used for minimizing prediction error when each arm is a auto-regressive moving average (ARMA) process.**

I. INTRODUCTION

A. Motivation

The multi-armed bandit (MAB) problem refers to a sequential allocation problem in which a unit resource is allocated to one of several competitive alternative actions/projects and a random reward (dependent on the chosen alternative) is obtained. The decision maker is not aware of the probability distribution of the reward process of each alternative and must use some allocation rule (or policy) to maximize the cumulative expected reward asymptotically in time.

The name multi-armed bandit, derives from an imagined slot machine with multiple arms. When an arm is pulled, the player wins a random reward following some unknown probability distribution. The objective of the player is to choose a policy to maximize the cumulative expected reward over the long term.

Multi-armed bandit problems are paradigms of allocation problems in which the decision maker experiences the *exploration versus exploitation dilemma*: the player must balance the exploitation of actions that did well in the past and the exploration of actions that might give higher rewards in the future. Some motivating examples of MAB problems are advertisement placement, internet routing, and cognitive radio communications.

In advertisement placement [9], the MAB problem arises in terms of deciding which advertisement to show to the next visitor of some web-page, among a finite set of advertisements. The total reward in this case is associated to the number of click-outs that the advertisement receives.

P.C. Prokopiou is with the Department of Neurology and Neurosurgery, Integrated Program in Neuroscience, McGill University, Montreal, QC, Canada prokopis.prokopiou@mail.mcgill.ca

P.E. Caines and A. Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada peter.caines,aditya.mahajan@mcgill.ca

In Internet routing [10], the MAB problem arises in terms of choosing a route between a source and a destination, among several alternatives, to transmit a packet at each transmission instant. The reward in this case is associated to the transmission time or the transmission cost of the packet.

In cognitive radio communications [11], the MAB problem arises in terms of choosing which channel a cognitive user should attempt to use in different time slots. The reward in this case is associated to the number of bits that the cognitive user is able to send at each time slot.

B. Literature Overview

Multi-armed bandit problems have been investigated since the 1930s [14]. See [16] for a historical survey. The break through in the solution to the MAB problem was achieved by Lai and Robbins [5] who constructed an allocation policy that asymptotically achieves the optimal regret of $O(\log T)$, where T is the total number of plays. Their results assumed K independent arms, each generating an i.i.d. sequence of rewards. These assumptions are also present in the vast majority of subsequent literature. We refer the reader to [15] for a survey.

One important line of work is to identify computationally efficient strategies that are also optimal in some sense. Agrawal [4] proposed a class of index type policies (i.e, policies that choose the arm with the highest index at each time), where the index depends on the sample mean of the reward process, and showed that these policies are asymptotically optimal, i.e. they achieve a regret of $O(\log T)$. Auer et. al. [3], proposed a similar index type policy, which they called UCB1, and showed that the regret is $O(\log T)$ uniformly in time (rather than asymptotically).

There has been less work on the MAB problems with independent arms and time-dependent reward process, most of which is restricted to Markov processes. The earliest work with such a setting is [7], where the reward of each arm is generated by Markov chain with a parametrized transition probability matrix. The authors proposed an index-based policy that achieves an asymptotically optimal regret. A more recent work in this setting is [12] in which a policy based on UCB1 is proposed; this policy achieves a logarithmic regret bound, uniformly in time.

To the best of our knowledge, there has been no study of the MAB problem with independent arms and general time-dependent reward processes such as an ARMA process, etc. The purpose of this paper is to investigate this set up.

C. Contributions

In this paper we consider the MAB problem with finite parameter spaces. Our main results are the following:

- We propose an allocation rule that depends on maximum-likehood estimate of the unknown parameter for each arm; as such, it applies to general time-dependent reward processes.
- Under some assumptions on the reward process, we show that the proposed policy is an upper confidence bound as defined in [4].
- The proposed allocation rule almost surely locks on to the optimal arm in finite expected time but has a super-linear regret. The finite lock-on time is an extremely desirable property of learning algorithms. The standard allocation rules for MAB, such as UCB and UCB1, do not have this property.

II. Model and Problem Formulation

Consider K mutually independent real-valued processes $\{Y_n^k\}_{n=1}^{\infty}$, $k=1,\dots,K$, defined on a common measurable space $(\Omega,\mathscr{A})$. The probability law on the k-th reward process $\{Y_n^k\}_{n=1}^{\infty}$ belongs to a finite set of probability measures $\{\mathbb{P}_\theta^k;\theta\in\Theta_k\}$, where Θ_k is a known finite set. Let $\mathbb{P}_{\theta_k^*}^k$ denote the true probability law and θ_k^* denote the unique true parameter for the k-th reward process. Let f_θ^k and μ_θ^k denote the probability density and mean, respectively, of the law $\mathbb{P}_\theta^k$, $\theta\in\Theta_k$. We assume that for all k and all $\theta\in\Theta_k$, f_θ^k exists and $|\mu_\theta^k|<\infty$. Let $k^* \triangleq \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\{\mu_{\theta_k^*}^k\}$ denote the arm with the highest mean reward.

A decision maker sequentially samples one of the K reward processes. Let Z_t denote the t-th sample. At time t, the decision maker samples the arm

$$u_t=\phi_t(Z_1,\dots,Z_{t-1}),$$

where $\phi_t:\mathbb{R}^{t-1}\to\{1,\dots,K\}$ is called the allocation policy.

Let n_t^k denote the number of times arm k has been sampled up to time t. Note that n_t^k satisfies the following recursion:

$$n_t^k=\begin{cases} n_{t-1}^k+1 & \text{if } u_t=k,\\ n_{t-1}^k & \text{if } u_t\neq k.\end{cases}$$

The t-th sample is then given by $Z_t=Y_{n_t^{u_t}}^{u_t}$.

The quality of an allocation policy is measured by its expected regret, which is defined as follows

$$\begin{aligned}R_T(\phi)&=T\mu_{\theta_{k^*}^*}^{k^*}-\sum_{t=1}^{T}\mathbb{E}(Z_t)\\&=\sum_{j=1}^{K}\left(\mu_{\theta_{k^*}^*}^{k^*}-\mu_{\theta_k^*}^{k}\right)\mathbb{E}(n_T^k).\end{aligned}\tag{1}$$

The MAB problem is to minimize the rate of growth of $R_T(\phi)$ as $T\to\infty$, and furthermore to find functions $f_L(T)$ and $f_U(T)$ such that there exist constants $C_L,C_U>0$, for which

$$C_L f_L(T)\le\mathbb{E}\{R_T(\phi)\}\le C_U f_U(T).$$

If $f_L=f_U=f$, then we say that the regret is of order f.

III. Main Results

In this paper, we use maximum likelihood estimates to determine an allocation policy and evaluate its asymptotic performance. The maximum likelihood estimate (MLE) has the property of strong convergence in finite parameter spaces, even when the observations are dependent over the time. We start by defining ML estimators and their properties.

A. Preliminaries on Maximum Likelihood Estimation

Consider a process $\{Y_n\}_{n=1}^{\infty}$ over a measurable space $(\Omega,\mathscr{A})$ and let $\mathscr{A}_n=\sigma(Y_1,\dots,Y_n)$ be the Borel σ-field generated by n observations. The probability law $\mathbb{P}_{\theta^*}$ of the process $\{Y_n\}_{n=1}^{\infty}$ belongs to a set $\{\mathbb{P}_\theta,\theta\in\Theta\}$, where Θ is a known finite set and $\theta^*\in\Theta$ denotes the true parameter. We assume that under each parameter $\theta\in\Theta$ there exists a density f_θ corresponding to $\mathbb{P}_\theta$.

An estimator $\hat{\vartheta}=\{\hat{\vartheta}_1,\hat{\vartheta}_2,\dots\}$ is sequence of mappings $\hat{\vartheta}_n:\Omega\to\Theta$ that are $\mathscr{A}_n$ measurable. Given any sequence $\{y_n\}_{n=1}^{\infty}$ of observations, let $\{\hat{\theta}_n\}_{n=1}^{\infty}$ be the sequence of estimates corresponding to the estimator $\hat{\vartheta}$. Then, $\hat{\vartheta}$ is called the *maximum likelihood estimator* if

$$f_{\hat{\theta}_n}(y_1,\dots,y_n)\ge\max_{\theta\in\Theta}\{f_\theta(y_1,\dots,y_n)\},\ \mathbb{P}_{\theta^*}a.s.$$

Moreover, $\hat{\vartheta}$ is called a (strongly) consistent estimator if $\hat{\theta}_n\neq\theta^*$ finitely often, $\mathbb{P}_{\theta^*}$ almost surely. For a consistent estimator $\hat{\vartheta}$, the *lock-on time* refers to least N such that for all $n>N$, $\hat{\theta}_n=\theta^*$, $\mathbb{P}_{\theta^*}$ almost surely.

Assume that the probability law $\mathbb{P}_\theta$, $\theta\in\Theta$, satisfies the following assumptions:

Assumption 1: Let $\mathbb{P}_{\theta,n}$ denote the restriction of $\mathbb{P}_\theta$ to the σ-field $\mathscr{A}_n$, $n\ge0$. Then, for all $\theta\in\Theta$, $\mathbb{P}_{\theta,n}$ is absolutely continuous with respect to $\mathbb{P}_{\theta^*,n}$.

Assumption 2: For every $\theta\in\Theta$, let $f_{\theta,n}$ be the density function associated with $\mathbb{P}_{\theta,n}$. Define

$$f_{\theta,n}(y_n|y^{n-1})=\frac{f_{\theta,n}(y^n)}{f_{\theta,n-1}(y^{n-1})},\qquad f_{\theta,0}(y_0|y^{-1})=f_\theta(y_0),$$

$$\text{and}\qquad h_{\theta,n}(y_n|y^{n-1})=\frac{f_{\theta,n}(y_n|y^{n-1})}{f_{\theta^*,n}(y_n|y^{n-1})},$$

where $y^n\triangleq y_1,\dots,y_n$.

Then for every $\varepsilon>0$, there exists $\alpha(\varepsilon)>1$, such that

$$P_{\theta^*}\left\{0\le h_{\hat{\theta}_{n-1}}(y_n|y^{n-1})\le\alpha,\ \text{for all } n>|\Theta|\right\}<\varepsilon,$$

where $\hat{\theta}_n\in\Theta$.

Theorem 1 ([2], pp. 327–328): Suppose that Assumptions 1 and 2 are satisfied, then maximum likelihood estimates are consistent.

B. The Proposed Allocation Rule

Coming back to the MAB problem, assume that the following assumption is satisfied:

Assumption 3: For every arm k, there is a consistent estimator $\hat{\vartheta}^k=\{\hat{\vartheta}_1^k,\hat{\vartheta}_2^k,\dots\}$.

We observe that if $\{\mathbb{P}_\theta^k\}$, $\theta\in\Theta^k$, satisfies Assumptions 1 and 2, then the maximum likelihood estimator $\hat{\vartheta}^k$ is consistent and hence Assumption 3 is satisfied.

Given any sequence $\{y_n^k\}_{n=1}^{\infty}$ of observations from arm k, let $\{\hat{\theta}_n^k\}_{n=1}^{\infty}$ be the sequence of estimates corresponding to the estimator $\hat{\vartheta}^k$.

Motivated by UCB1 in [3], we propose the following allocation rule Φ^g. For ease of notation, let $\hat{\mu}_t^k = \mu_{\hat{\theta}_{n_t^k}^k}^k$. Consider a set of index functions $g = \left\{g_{t,n}^k\left(y_1^k,\ldots,y_n^k\right)\right\}$ with

$$g_{t,n}^k\left(y_1^k,\ldots,y_n^k\right) \triangleq \hat{\mu}_t^k + \frac{t/C}{n}, \tag{2}$$

where $t \in \mathbb{Z}_{>0},\ n \in \{1,\ldots,t\},\ C \in \mathbb{R}$ and $k \in \{1,\ldots,K\}$.

If $t < K$ then ϕ^g samples from each process y_n^k once, otherwise it samples from the process indexed by $u_t = \max\{g_{t,n_t^k}^k;\ k \in \{1,\ldots,K\}\}$.

The main difference between the index in Φ^g and the index in UCB1 [2] is that the index there depends upon the sample mean, while the index in Φ^g depends upon the mean of the maximum likelihood estimate or, in general, any consistent estimator. This modification makes Φ^g applicable to more general MAB problems in which the reward process for each machine is time dependent.

Lemma 1: For each machine $k \in \{1,2,\ldots,K\}$,

$$\lim_{t\to\infty} n_t^k = \infty \ \text{ a.s. } \mathbb{P}_{\theta_k^*}^k.$$

The proof in presented in Appendix I.

Theorem 2: If Assumption 3 holds, then for each $k \in \{1,\ldots,K\}$, the index function g^k given in (2) is an Upper Confidence Bound (UCB), i.e, it satisfies the following conditions ([4]):

1) $g_{t,n}$ is non-decreasing in $t \geq n$, for each fixed $n \in \mathbb{Z}_{>0}$.
2) Let $y_1^k, y_2^k, \ldots, y_n^k$ be a sequence of observations from machine k. Then, for any $z < \hat{\mu}_t^k$,

$$\mathbb{P}_{\theta_k^*}\left\{g_{t,n}\left(y_1^k,\ldots,y_n^k\right) < z, \text{ for some } n \leq t\right\} = \mathbf{o}(t^{-1})$$

The proof in presented in Appendix II.

C. Performance of Φ^g

Assumption 4: (Summable Wrong and Corrected Condition (SWAC)) For all machines $k \in \{1,\ldots,K\}$, the sequence of estimates $\hat{\theta}_1^k,\ldots,\hat{\theta}_n^k,\ldots$ satisfies the following condition:

$$\mathbb{P}_{\theta_k^*}^k(\hat{\theta}_{n-1}^k \neq \theta_k^*, \hat{\theta}_m^k = \theta_k^*,\ \forall m \geq n) < \frac{C}{n^{3+\beta}}, \tag{3}$$

for some $C \in \mathbb{R}_{>0},\ \beta \in \mathbb{R}_{>0}$, and for all $n \in \mathbb{Z}_{>0}$.

Under Assumption 4 it becomes geometrically more difficult to change a false decision to a true one over time. We note that Assumption 4 does not imply consistency.

However, when consistency holds, Assumption 4 implies the $2+\alpha$ moment, and hence the first and second moments, of the random lock-on instant N are finite for $0 < \alpha < \beta$, where β appears in Assumption 4.

Lemma 2: Let N_k be the lock-on time for estimator $\hat{\theta}^k$. Then, under Assumption 4,

$$\mathbb{E}\{N_k^{2+\alpha}\} < \infty, \qquad \forall k \in \{1,\ldots,K\},\ 0 < \alpha < \beta, \tag{4}$$

where β appears in Assumption 4.

Proof: Under Assumption 4 the lock on time N_k satisfies

$$\begin{aligned}\mathbb{E}N_k^{(2+\alpha)} &= \sum_{n=1}^{\infty} n^{(2+\alpha)}\mathbb{P}(N_k = n)\\ &= \sum_{n=1}^{\infty} n^{(2+\alpha)}\mathbb{P}(\hat{\theta}_{n-1}^k \neq \theta_k^*, \hat{\theta}_m^k = \theta_k^*,\ \forall m \geq n)\\ &< \sum_{n=1}^{\infty} n^{(2+\alpha)}\frac{C}{n^{3+\beta}} < \infty, \quad 0 < \alpha < \beta.\end{aligned}$$

■

Theorem 3: If Assumptions 3 and 4 hold, then the regret of ϕ^g satisfies $R_T(\phi^g) = \mathbf{o}(T^{1+\delta})$ for some $\delta > 0$.
The proof in presented in Appendix III.

IV. MAB PROBLEMS FOR LINEAR SYSTEMS

A. An MAB Problem for ARMA Processes

Consider a bandit system with reward process generated by the following ARMA (or equivalently linear state space) system:

$$S: \qquad \begin{aligned} x_{n+1}^k &= \lambda_k x_n^k + w_n^k \\ y_n^k &= x_n^k \end{aligned} \qquad \forall n \in \mathbb{Z}_{\geq 0} \tag{5}$$

where $x_n^k, y_n^k, w_n^k \in \mathbb{R}$ for $n \in \mathbb{Z}_{\geq 0}$, and w^k is an i.i.d. $\mathcal{N}(0,{\sigma_k}^2)$ process independent of x_0^k. For simplicity assume that there are only two machines, i.e. $k \in \{1,2\}$, and that the parameter space of the system contains two alternatives: $\Theta_k = \{\theta_k^*, \theta_k\};\ \theta_k \triangleq (\lambda_k, \sigma_k)$. In addition, assume that (5) is in steady state; in steady state, the system must be asymptotically stable, that is, $|\lambda| < 1$.

At each step t the player chooses to observe a sample from machine $k \in \{1,2\}$ and pays a cost equal to the squared minimum one step prediction error of the next observation $y_{n_t^k}^k$ given the past observations $y_1^k,\ldots,y_{n_t^k-1}^k$, where n_t^k denotes the local time of machine k. Denote this cost at time t as υ_t^k.

For this example, we define the expected regret as follows:

$$R_T(\phi) = -\sum_{i=1}^{T}\left(\min_{k\in\{1,2\}} \mathbb{E}{\upsilon_{n_i^k}^k}^2 - \mathbb{E}{\upsilon_{n_i^{u_i}}^{u_i}}^2\right), \tag{6}$$

and we recall that the player aims to minimize the rate of growth of the expected regret $R_T(\phi)$ as $T \to \infty$.

The problem under consideration corresponds to fairly realistic cases where one wants to learn the dynamics of both machines (by identifying the true unknown parameter of each machine) in order to hit a target (e.g. a physical or financial target) based on the knowledge so far gained.

Remark 1: We note that the MAB model as described above does not fit with the model described in Section I. This is because in the MAB model described in Section I, the reward yielded from machine $k \in \{1,\ldots,K\}$ at instant t depends only on the observation $y_{n_t^k}^k$ made at the same instant, while in the MAB model described above, the reward yielded from machine k ($\upsilon_{n_t^k}^k$) depends on the past observations $y_1^k,\ldots,y_{n_t^k-1}^k$ as well.

However, we can make the latter model fit with the former model by using a simple transformation of the observations.

Specifically, we can assume that whenever an agent plays machine k, it observes a *vector* $(y^k_{n^k_t}, \upsilon^k_{n^k_t})$. By using this transformation, the scenario described earlier remains valid in terms of estimation. This is because $\upsilon^k_{n^k_t}$ is a function of the past and present observation $(y^k_1,\ldots,y^k_{n^k_t})$, and thus employing $\upsilon^k_{n^k_t}$ does not alter the selection of the ML estimator for machine k.

In the sequel, we examine the behaviour of Φ^g in such a system. To do so, we first need to verify Assumptions 1, 2, and 4 for each machine separately. These can be translated in terms of properties of the innovations processes. So we first start with a recap of some standard results for ARMA processes.

B. Preliminary Results on ARMA Processes

Consider the reward process of each machine described by (5). The state process x_n, and hence the output process y_n, shall be assumed to be zero mean stationary under both parametrizations and consequently it is assumed that, $|\lambda^*| < 1$ and $|\lambda| < 1$.

The negative logarithmic likelihood function of the reward process can be decomposed in terms of the prediction error process $y_i - \mathbb{E}(y_i|y^{i-i}) = y_i - y_{i|i-1}$ as follows:

$$-\log f(y^n;\lambda) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\left(\frac{\sigma^{2n}}{1-\lambda^2}\right) + \frac{1}{2}y_1^2\left(\frac{\sigma^2}{1-\lambda^2}\right)^{-1} + \frac{1}{2}\sum_{i=2}^{n}(y_i - y_{i|i-1})^2\sigma^{-2} \quad (7)$$

where $y_{i|i-1} \triangleq \mathbb{E}(y_i|y^{i-i}) = \lambda y_{i-1}$ is the optimal non-linear least squares estimate of y_i given the past observations $y_1,\ldots,y_{i-1}$, which is generated by the Kalman filter.

The prediction error process under the true parameter θ^* is:

$$\begin{aligned} \nu_n &= y_n - \mathbb{E}\{y_n \big| y^{n-1}\} = (\lambda^* y_{n-1} + w_{n-1}) - \lambda^* y_{n-1} \\ &= w_{n-1}, \quad w_{n-1} \sim \mathcal{N}(0,\sigma^{*2}). \end{aligned}$$

The prediction error process under the incorrect parameter θ is:

$$\begin{aligned} e_n &= y_n - \mathbb{E}\{y_n \big| y^{n-1}\} = (\lambda^* - \lambda) y_{n-1} + w_{n-1} \\ &= w_{n-1} + (\lambda^* - \lambda)\sum_{j=1}^{n} \lambda^{*j-1} w_{n-1-j} \\ &= \nu_n + (\lambda^* - \lambda)\sum_{j=1}^{n} \lambda^{*j-1}\nu_{n-j}, \end{aligned}$$

where the fact that under both hypotheses $y_{1|0} = 0$ is used to obtain the third line.

The prediction error process of the system under the true parameter θ^* is called the innovations process of y_n; in general, it is an independent process (see [2]) and in the case under consideration it is i.i.d.. On the other hand, the prediction error process of the system under the incorrect parameter θ is in general a dependent process is called the pseudo-innovations process. Note that $\mathbb{E}(\nu^2) = \sigma^{*2} < \mathbb{E}(e^2)$.

We now consider in turn the validity in the current case of each of the general assumptions introduced earlier.

1) Assumption 1: Assuming that $\theta^* \neq \theta$ for each linear system, Assumption 1 follows in each case.

2) Assumption 2: We consider each machine separately. To verify Assumption 2, one needs to show that for all $\varepsilon > 0$, there exists $\alpha(\varepsilon) > 1$ such that,

$$\mathbb{P}_{\theta^*}\left\{0 \leq \frac{f(y_n|y^{n-1};\theta)}{f(y_n|y^{n-1};\theta^*)} < \alpha(\varepsilon), \forall n > |\Theta|\right\} < \varepsilon. \quad (8)$$

But

$$\begin{aligned} &\frac{f(y_n|y^{n-1};\theta)}{f(y_n|y^{n-1};\theta^*)} < \alpha \\ \implies &\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{e_n^2}{\sigma^2}\right) < \alpha\frac{1}{\sqrt{2\pi}\sigma^*}\exp\left(-\frac{1}{2}\frac{\nu_n^2}{\sigma^{*2}}\right). \quad (9) \end{aligned}$$

Taking the logarithm in both sides of (9), one has

$$\log\frac{\sigma^{*2}}{\sigma^2} - \log\alpha < \frac{1}{2}\left[\frac{(\nu_n + (\lambda^* - \lambda)\sum_{j=1}^{n}\lambda^{*j-1}\nu_{n-j})^2}{\sigma^2} - \frac{\nu_n^2}{\sigma^{*2}}\right].$$

So substituting in (8) yields the following expression for Assumption 2:

$$\mathbb{P}_{\theta^*}\left\{\log\frac{\sigma^{*2}}{\sigma^2} - \log\alpha < \frac{1}{2}\left[\frac{(\nu_n + (\lambda^* - \lambda)\sum_{j=1}^{n}\lambda^{*j-1}\nu_{n-j})^2}{\sigma^2} - \frac{\nu_n^2}{\sigma^{*2}}\right], \forall n > |\Theta|\right\} < \varepsilon. \quad (10)$$

The event inside the probability measure in (10) is in general hard to evaluate since it involves the sum of squared past innovations which are Gaussian random variables. Thus, this summation follows some generalized form of χ^2-distribution whose close form expression is not known. This leads us to adopt the following conjecture which is seen to be very plausible when the Strong Law of Large Numbers and the Law of the Iterated Logarthim are applied to the sums appearing in the expansion of the quadratic expression in (10):

Conjecture 1: For the set of likelihood functions specified by the parameter set Θ, Assumption 2 is satisfied.

3) Assumption 4: We consider each machine separately. Consider the event $\{\hat{\theta}_n \neq \theta^*\}$ for which we have

$$\begin{aligned} \{\hat{\theta}_n \neq \theta^*\} &= \{f(y^n;\theta) > f(y^n;\theta^*)\} \\ &= \{-\log f(y^n;\theta) < -\log f(y^n;\theta^*)\} \quad (11) \end{aligned}$$

Using the decomposition property of the likelihood function [2] in terms of the innovations process we have

$$\{\hat{\theta}_n \neq \theta^*\} = \left\{A_n < \sum_{i=2}^{n}\frac{\nu_i^2}{\sigma^{*2}}\right\} \quad (12)$$

where

$$A_n \triangleq n\log\left(\frac{\sigma^2}{\sigma^{*2}}\right) + \log\left(\frac{1-\lambda^{*2}}{1-\lambda^2}\right) + y_1^2\left(\frac{\sigma^2}{1-\lambda^2}\right)^{-1} - y_1^2\left(\frac{\sigma^{*2}}{1-\lambda^{*2}}\right)^{-1} + \sum_{i=2}^{n}\frac{e_i^2}{\sigma^2}. \quad (13)$$

Consider also the event $\{\hat{\theta}_{n+1} = \theta^*\}$ for which, by similar reasoning to that yielding (12), we have

$$\{\hat{\theta}_{n+1} = \theta^*\} = \Big\{A_{n+1} \geq \sum_{i=2}^{n+1} \frac{\nu_i^2}{\sigma^{*2}}\Big\}, \tag{14}$$

which we shall denote by $\{A_{n+1} \geq V_{n+1}\}$.

To analyse the joint event $E_n \triangleq \{\hat{\theta}_n \neq \theta^*, \hat{\theta}_m = \theta^*,\ \forall m \geq n\}$, we substitute (12) into (14) to see that it is equivalent to the event F_n given in the following expression:

$$\begin{aligned} F_n &= \{\hat{\theta}_n \neq \theta^*\} \cap \{A_{n+1} \geq V_{n+1}\} \cap \{A_{n+2} \geq V_{n+2}\} \cap \dots \\ &= \Big\{\sum_{i=2}^{n} \frac{\nu_i^2}{\sigma^{*2}} > A_n\Big\} \cap \{A_{n+1} \geq V_{n+1}\} \cap \{A_{n+2} \geq V_{n+2}\} \cap \dots \end{aligned} \tag{15}$$

As for the event inside the probability measure in (10), the event described by (15) involves a linear combination of χ^2 random variables whose probability density function is not known. This leads us to adopt the following conjecture which, again, is very plausible when considered in terms of the Strong Law of Large Numbers and the Law of the Iterated Logarthim.

Conjecture 2: There exists a a, $\beta \in \mathbb{R}_{>0}$ such that for all $n \in \mathbb{Z}_{>0}$,

$$\mathbb{P}\{F_n\} < \frac{a}{n^{3+\beta}}, \tag{16}$$

and hence Assumption 4 is satisfied.

C. The regret function

Based on the preliminary results on the ARMA processes introduced above, the expected regret function of the MAB system S given by (6) can be restated as follows:

$$\begin{aligned} R_T(\phi) &= -\sum_{i=1}^{T} \big(\min_{j\in\{1,2\}} \mathbb{E}\nu^{j\,2} - \mathbb{E}\nu_{n_T^{u_i}}^{u_i\,2}\big) \\ &= -\sum_{k=1}^{2} \big(\min_{j\in\{1,2\}} \sigma_j^{*2} - \sigma_k^{*2}\big)\mathbb{E}(n_T^k) \end{aligned} \tag{17}$$

where $\nu_{n_T^{u_i}}^{u_i\,2}$ is the squared innovations process of machine $u_i \in \{1,2\}$ played at instant i and σ_j^{*2} denotes the innovations process variance of machine $k \in \{1,2\}$.

D. Index policies

The index functions in (2) can be expressed as

$$g_{T,n_T^k}^k = \frac{2}{\hat{\sigma}_k^2} + \frac{T}{Cn_T^k},\ k \in \{1,2\} \tag{18}$$

where $\hat{\sigma}_k^2$ is the ML estimate of the innovations process variance of machine k.

For the computation of $\hat{\sigma}_T^k$ at stage T, we compute the maximized likelihood ratio (MLR) given by

$$l^k(T) = \max_{\psi^k \in \Theta_k} \frac{f_{\psi^k}(y_1^k, \dots, y_T^k)}{f_{\theta_0^k}(y_1^k, \dots, y_T^k)}. \tag{19}$$

where θ_0^k in (19) is arbitrary and the ML estimate is given by the value of the parameter ψ^k giving the greatest value to the ratio in (19).

TABLE I: The parameter values of the 3 systems considered for simulations. The simulations results are shown in Fig. 1,2,3.

System 1 (S1)	
$\Theta_1 = \{\theta_1^1 = (0.145, 8), \theta_1^2 = (0.09, 10)\}$	$\theta_1^* = \theta_1^1$
$\Theta_2 = \{\theta_2^1 = (0.2, 5), \theta_2^2 = (0.19, 15)\}$	$\theta_2^* = \theta_2^2$

System 2 (S2)	
$\Theta_1 = \{\theta_1^1 = (0.145, 8), \theta_1^2 = (0.09, 10)\}$	$\theta_1^* = \theta_1^1$
$\Theta_2 = \{\theta_2^1 = (0.2, 5), \theta_2^2 = (0.19, 8.1)\}$	$\theta_2^* = \theta_2^2$

System 3 (S3)	
$\Theta_1 = \{\theta_1^1 = (0.145, 8.09), \theta_1^2 = (0.09, 8.1)\}$	$\theta_1^* = \theta_1^1$
$\Theta_2 = \{\theta_2^1 = (0.2, 8.11), \theta_2^2 = (0.19, 8.1)\}$	$\theta_2^* = \theta_2^2$

E. Simulation Results

Consider the MAB problem with ARMA reward processes and the parameter values given in Table I, where Θ_k corresponds to the parameter space of the k-th machine and $k \in \{1,2\}$:

For each of these systems, Figures 1, 2, and 3 show the sample regret for different realizations as well as the sample mean for Systems (S1)–(S3) and for different values of the parameter C.

When C is small, the switching between the arms takes place with a higher frequency. Consequently, more time is spent in exploration, and the regret increases at a larger rate. When C is large, the switching between arms takes place at a lower frequency. Consequently, more time is spent in exploitation, and the regret increases at a slower rate. Note, however, that the lock-on to the true parameter is also slower when C is larger. This is clearly illustrated in Systems (S2) and (S3), where the parameters were chosen close to one another to make estimation harder. Figures 2 and 3 show that for $C = 10000$ there are some sample paths for which the lock-on to the true parameter has not taken place until the end of the simulation.

It should be noted that although the algorithm locks-on to the true arm in finite expected time, the regret keeps on increasing because of the T/Cn_T term in our index rule. In UCB1-type algorithms, this index if of the form $\hat{\mu} + \log T/Cn_T$, and hence the regret increases as $\mathbf{O}(\log T)$.

V. CONCLUSIONS

In this paper, we consider the MAB problem with time-dependent rewards that depend on single parameters which lie in a known, finite parameter space. We propose an allocation rule, Φ^g, which employs a set of functions that depend on consistent estimators of the unknown parameters. In particular, we consider the Maximum Likelihood Estimators on finite parameter sets which under Assumptions 1-2 are known to be consistent [1]. Further given Assumptions 3-4, we show that Φ^g is of index type and $R_T(\Phi^g) \in \mathbf{o}(T^\delta)$ for some $\delta > 1$. Although this result is suboptimal compared to other results in the literature for MAB problems with i.i.d.

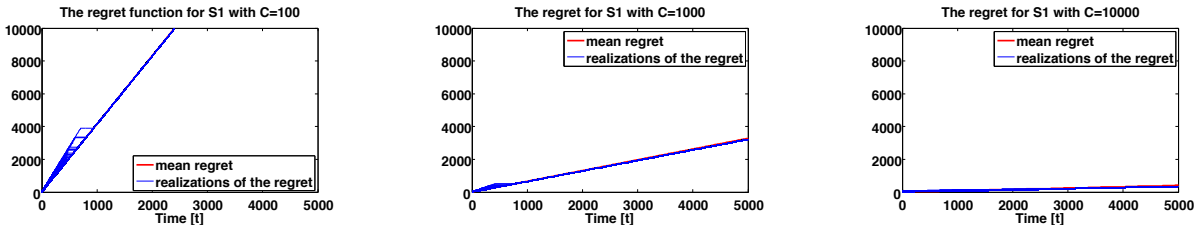

Fig. 1: Simulation of 10000 realizations for System 1 for 3 values of C. Left panel: $C = 100$. Middle panel: $C = 1000$. Right panel: $C = 10000$. The regret resulted from each realizition is plotted in blue, and the regret over all realizations in red. The parametres specifying System 1 are given in Table I.

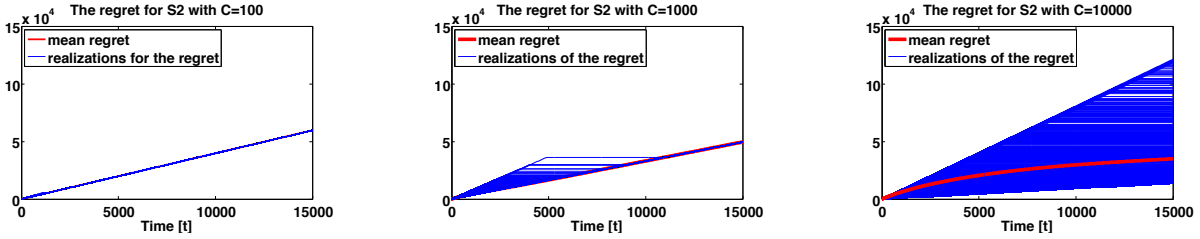

Fig. 2: Simulation of 10000 realizations for System 2 for 3 values of C. Left panel: $C = 100$. Middle panel: $C = 1000$. Right panel: $C = 10000$. The regret resulted from each realizition is plotted in blue, and the regret over all realizations in red. The parametres specifying System 2 are given in Table I.

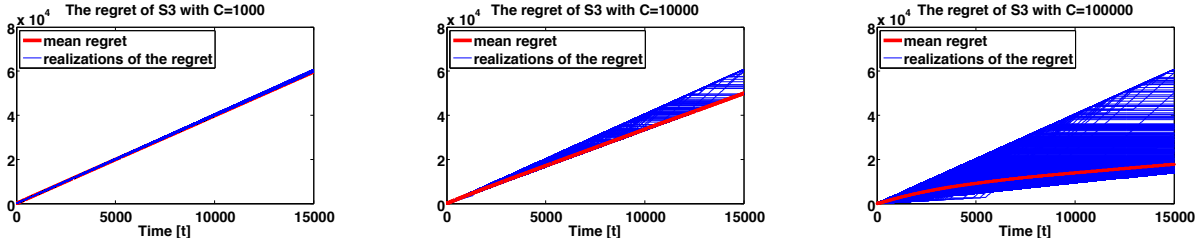

Fig. 3: Simulation of 10000 realizations for System 3 for 3 values of C. Left panel: $C = 1000$. Middle panel: $C = 10000$. Right panel: $C = 100000$. The cost function resulted from each realizition is plotted in blue, and the regret over all realizations in red. The parametres specifying System 3 are given in Table I.

rewards, the proposed policy Φ^g is more flexible because it can be applied to a more general class of MAB problems, including those with time-dependent rewards.

VI. ACKNOWLEDGMENTS

The support of NSERC research grants to the second two authors is gratefully acknowledged.

REFERENCES

[1] P.E. Caines, *A note on the consistency of maximum likelihood estimates for finite families of stochastic processes*, The Annals of Statistics, JSTOR, 1975, pp. 539–546.

[2] P.E. Caines, *Linear stochastic systems*, John Wiley & Sons, Inc., 1987, pp 265-271.

[3] P. Auer and N. Cesa-Bianchi and P. Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine learning, Springer, Englewood Cliffs, NJ; 2002, (47)2-3, pp 235–256.

[4] R. Agrawal, *Sample mean based index policies with O(log n) regret for the multi-armed bandit problem*, Advances in Applied Probability, JSTOR, 1995, pp. 1054–1078.

[5] T.L. Laiand H. Robbins, *Asymptotically efficient adaptive allocation rules* , Advances in applied mathematics, 1985, (22), pp. 4–22

[6] V. Anantharam and P. Varaiya and J. Walrand,*Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: IID rewards*, Automatic Control, IEEE Transactions, (32)11, IEEE, 1987, pp. 968–976

[7] V. Anantharam and P. Varaiya and J. Walrand,*Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: IID rewards*, Automatic Control, IEEE Transactions, (32)11, IEEE, 1987, pp. 977–982

[8] D. Boley and R. Maier, "A Parallel QR Algorithm for the Non-Symmetric Eigenvalue Algorithm", *in Third SIAM Conference on Applied Linear Algebra*, Madison, WI, 1988, pp. A20.

[9] J. White, *Bandit algorithms for website optimization*, O'Reilly Media, Inc., 2012

[10] Y. Gai and B. Krishnamachari and R. Jain, *Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations*, IEEE/ACM Transactions on Networking (TON), IEEE Press, (20)5, pp 1466–1478

[11] L. Lai and H. El Gamal and H. Jiang and V. H Poor, *Cognitive medium access: Exploration, exploitation, and competition*, Mobile Computing, IEEE Transactions, IEEE, 2011, (10)2, pp 239–253

[12] C. Tekin and M. Liu, *Online algorithms for the multi-armed bandit problem with markovian rewards*,Conference on Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton , 2010, IEEE, pp. 1675–1682

[13] P.C. Prokopiou, *An estimation based allocation rule with super-*

linear regret and finite lock-on time for dependent multi-armed bandit processes , MEng Theisis, McGill University, May, 2014

[14] W.R. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*,Biometrika,1933, JSTOR, pp. 285–294

[15] S. Bubeck and N. Cesa-Bianchi, *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, Foundations and Trends in Machine Learning, 2012, (5)1 IEEE, pp. 1–122

[16] H. Robbins, *Some Aspects of the Sequential Design of Experiments*, Bulletin of the American Mathematical Society, 1952, (58), pp. 527–535

APPENDIX I

Proof: [Lemma 1] We prove th result by contradiction. Assume that there is $\omega \in \Omega$ and a subset $\mathscr{S} \subset \mathscr{K} = \{1,\dots,K\}$ of machines which are chosen finitely many times, while machines in $\mathscr{K} \setminus \mathscr{S}$ are chosen infinitely many times. That means

$$\forall k \in \mathscr{S},\ \exists q_k > 0 \text{ s.t. } \lim_{t\to\infty} n_t^k = q_k < \infty \tag{20}$$

and

$$\forall j \in \mathscr{K} \setminus \mathscr{S}, \lim_{t\to\infty} n_t^j = \infty. \tag{21}$$

By the assumption that each machine $k \in \mathscr{S}$ had been played finitely many times, it is implied that for all $k \in \mathscr{S}$ there exists $t' > 1$ such that for all $t > t'$

$$\begin{aligned}
&\mu^k_{\hat{\theta}^k_{n_t^k}} + \frac{t/C}{n_t^k} \le \max_{j\in\mathscr{K}\setminus\mathscr{S}} \left\{ \mu^j_{\hat{\theta}^j_{n_t^j}} + \frac{t/C}{n_T^j} \right\} \\
\implies &\min_{\theta_k\in\Theta_k}\left\{\mu^k_{\theta_k}\right\} + \frac{t/C}{n_t^k} \le \max_{j\in\mathscr{K}\setminus\mathscr{S}} \left\{ \max_{\theta_j\in\Theta_j}\left\{\mu^j_{\theta_j}\right\} + \frac{t/C}{n_t^j} \right\} \\
\implies &\frac{\gamma_k - \Gamma}{t/C} + \frac{1}{n_t^k} \le \max_{j\in\mathscr{K}\setminus\mathscr{S}} \left\{ \frac{1}{n_t^j} \right\}
\end{aligned}$$

where $\gamma_k = \min_{\theta_k\in\Theta_k}\left\{\mu^k_{\theta_k}\right\}$ and $\Gamma = \max_{j\in\mathscr{K}\setminus\mathscr{S}}\left\{\max_{\theta_j\in\Theta_j}\left\{\mu^j_{\theta_j}\right\}\right\}$.

But

$$\begin{aligned}
\lim_{t\to\infty}\{\frac{\gamma_k - \Gamma}{t/C} + \frac{1}{n_t^k}\} &\le \lim_{t\to\infty} \max_{j\in\mathscr{K}\setminus\mathscr{S}} \left\{ \frac{1}{n_t^j} \right\} \\
\implies \quad \frac{1}{q_k} &\le 0
\end{aligned}$$

which leads to a contradiction since q_k was assumed to be a finite positive. ∎

APPENDIX II

Proof: [Theorem 2] For any $k \in \{1,\dots,K\}$, and any fixed $n < t$, the estimate $\hat{\mu}_t^k$ is constant. That means g^k is a linear function in t, which proves condition 1.

Moreover, by Lemma 1 and under Assumption 3 we have that for every $k \in \{1,\dots,K\}$, for all $\omega \in \Omega_o^k \subseteq \Omega$ where Ω_o^k is such that $\mathbb{P}^{\Omega_o^k}_{\theta_k^*} = 1$, and all $n > N_k$

$$\hat{\theta}_n^k = \theta_k^* \tag{22}$$

In addition, define

$$\begin{aligned}
&B_n^k \triangleq \{\omega : N_k < n\}, \\
&A_{t,n}^k \triangleq \left\{\omega : g_{t,n}^k(y_1^k,\dots,y_n^k) < z\right\} \text{ for any } z < \mu^k_{\theta_k^*} \text{ and} \\
&A_t^k \triangleq \left\{\omega : g_{t,n}^j(y_1^k,\dots,y_n^k) < z \text{ for some } n \le t\right\} = \bigcup_{n=1}^{t} A_{t,n}^k.
\end{aligned}$$

Then,

$$\mathbb{P}_{\theta_k^*}\left(A_{t,n}^k \;\middle|\; B_n^k\right) = 0 \tag{23}$$

In addition, let $t_k^* = \left\lfloor t/C\left(\mu^k_{\theta_k^*} - \min_{\theta_k\in\Theta_k}\mu^k_{\theta_k}\right)\right\rfloor$ where $\lfloor\bullet\rfloor$ denotes the floor function. Then,

$$\mathbb{P}_{\theta_k^*}\left(A_{t,n}^k\right) = 0, \quad \forall n < t_k^*. \tag{24}$$

Thereafter, consider

$$\begin{aligned}
t\mathbb{P}(A_t^k) &= T\mathbb{P}\left(\bigcup_{i=1}^{t} A_{t,i}^k\right) \\
&\le t\sum_{i=1}^{t}\mathbb{P}(A_{t,i}^k) \quad \text{(by the union bound)} \\
&\le t\sum_{i=t^*+1}^{t}\mathbb{P}(A_{t,i}^k) \quad \text{(by eqn. (24))}
\end{aligned} \tag{25}$$

Using the law of total probability we have

$$\begin{aligned}
t\sum_{i=t_k^*+1}^{t}\mathbb{P}(A_{t,i}^k) &\le t\sum_{i=t^*+1}^{t}\mathbb{P}(A_{t,i}^k\big|B_i^k)\mathbb{P}(B_i^k) + \mathbb{P}(A_{t,i}^k\big|B_i^{k\complement})\mathbb{P}(B_i^{k\complement}) \\
&= t\sum_{i=t^*+1}^{t}\mathbb{P}(A_{t,i}^k\big|B_i^{k\complement})\mathbb{P}(B_i^{k\complement}) \quad \text{(by eqn. (23))}
\end{aligned} \tag{26}$$

Substituting (26) into (25) we have

$$\begin{aligned}
t\mathbb{P}(A_t^k) &\le t\sum_{i=t^*+1}^{t}\mathbb{P}(A_{t,i}^k\big|B_i^{k\complement})\mathbb{P}(B_i^{k\complement}) \\
&\le t\sum_{i=t^*+1}^{t}\mathbb{P}(B_i^{k\complement}) \\
&\le t\sum_{i=t^*+1}^{t}\frac{\mathbb{E}(N_k(\omega)^{2+\alpha})}{n^{2+\alpha}} \quad \text{(by Markov ineq.)} \\
&\le t\mathbb{E}(N_k(\omega)^{2+\alpha})\int_{i=t^*+1}^{t}\frac{1}{n^{2+\alpha}}\mathrm{d}n \\
&= \frac{t\mathbb{E}(N_k(\omega)^{2+\alpha})}{1+\alpha}((t^*+1)^{-(1+\alpha)} - t^{-(1+\alpha)}),
\end{aligned}$$

for some $\alpha > 0$. (27)

Taking the limit as $t \to \infty$ and under Assumption 3 we have

$$\lim_{t\to\infty}\left[\frac{t\mathbb{E}(N_k^{2+\alpha})}{1+\alpha}\left((t_k^*+1)^{-(1+\alpha)} - t^{-(1+\alpha)}\right)\right] = 0 \tag{28}$$

which shows that part 2 holds, and completes the proof of Theorem 2. ■

APPENDIX III

Proof: [Theorem 3] Consider the following upper bound of the local time of each machine $k \in \{1,\dots,K\}$ [4]:

$$n_T^k \leq 1 + \sup\{1 \leq n \leq T : g_{T,n}^k(y_1^k,\dots,y_n^k \geq \mu_{\theta_k^*}^{k^*} - \varepsilon)\} + \sum_{i=1}^{t} \mathbb{1}_{A_i^n} \tag{29}$$

Then,

$$\frac{\mathbb{E}\{n_T^k\}}{T^{1+\delta}} \leq \frac{\mathbb{E}\{1\} + \mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}^k(y_1^k,\dots,y_n^k \geq \mu_{\theta_k^*}^{k^*} - \varepsilon)\}\} + \mathbb{E}\{\sum_{i=1}^{T} \mathbb{1}_{A_i^n}\}\}}{T^{1+\delta}}$$

Taking the lim sup as $T \to \infty$ and the infimum over $\varepsilon > 0$ in both sides, and by condition 2 in Theorem 2, we end up with

$$\limsup_{T\to\infty} \frac{\mathbb{E}\{n_T^k\}}{T^{1+\delta}} \leq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}^k(y_1^k,\dots,y_n^k \geq \mu_{\theta_k^*}^{k^*} - \varepsilon)\}\}}{T^{1+\delta}}.$$

Then, an expression involving the expected regret is given by

$$\limsup_{T\to\infty} \frac{R_T(\Phi^g)}{T^{1+\delta}} = \sum_{j<k^*} \frac{(\mu_{\theta_{k^*}^*}^{k^*} - \mu_{\theta_j^*}^{j})}{L(\mathbb{P}_{\theta_j^*}^{j}, \mu_{\theta_{k^*}^*}^{k^*})}, \quad \forall \delta > 0. \tag{30}$$

where, for all $j < k^*$

$$\frac{1}{L(\mathbb{P}_{\theta_j^*}^{j}, \mu_{\vartheta_{k^*}^*}^{k^*})} \triangleq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}_{\theta_k^*}\left\{\sup\{1 \leq n_T^k \leq T : g^k(y_1^k,\dots,y_{n_T^k}^k) \geq \mu_{\theta_{k^*}^*}^{k^*}) - \varepsilon\}\right\}}{T^{1+\delta}} \tag{31}$$

Evaluating this result under the proposed index policy Φ^g we have that

$$\frac{1}{L(\mathbb{P}_{\theta_j^*}^{j}, \mu_{\vartheta_{k^*}^*}^{k^*})} \triangleq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}_{\theta_k^*}\left\{\sup\{1 \leq n_T^k \leq T : g^j(y_1^k,\dots,y_{n_T^k}^k) \geq \mu_{\theta_{k^*}^*}^{k^*}) - \varepsilon\}\right\}}{T^{1+\delta}}$$

$$\leq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{T}{T^{1+\delta}} = 0 \quad \forall \delta > 0.$$

But this means

$$\limsup_{T\to\infty} \frac{R_T(\Phi^g)}{T^{1+\delta}} = 0 \tag{32}$$

which is equivalent to $R_T(\omega, \Phi^g) \in \mathbf{o}(T^{1+\delta})$. ■