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

Prokopis C. Prokopiou

Master of Engineering

Department of Electrical and Computer Engineering

McGill University

Montreal,Quebec

2014-08-15

A thesis submitted to McGill University
in partial fulfilment of the requirements of the degree of
Master of Engineering

DEDICATION

To my family, Christos, Martha, Spyros and Alexandros, and in memory of my grandmother Augusta.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest gratitude to my advisors Professors Peter E. Caines and Aditya Mahajan. I thank them for offering me the chance to be part of their research group, for their honest advises in difficult times, and for providing me with any kind of support during my so far studies at McGill. I would also like to thank my lab mates Jalal, Ali, Shuang, Tracy, Mehnaz, Jhelum and Sayani for their help, inspiration and good company. In addition I would also like to thank Nevroz who was always willing to help me with any kind of questions I had while doing research, and Rabih for the translation of the abstract in French. Furthermore, I am grateful to my friends Dimitris, Nassir, Giannis, George and Kishor, for always being there for me. Of course, I would also like to thank Irene Syrimi for supporting me in every decision I've made and pushing me to pursue my dreams. Finally, I thank my family whose lifelong love and support makes everything seem possible and means the world to me.

ABSTRACT

In this thesis, we study the stochastic multi-armed bandit problem with dependent reward processes for each bandit machine, generated by an unknown probability measure. It is assumed that for each machine there is a finite family of probability measures indexed by a (finite) number of alternative parameter values, including the true probability measure which governs the reward process of each machine.

In this framework, an index type allocation rule is proposed that employs consistent estimators and achieves a $\mathbf{o}(T^{1+\delta})$ regret, for some $\delta > 0$.

In particular, an instance of consistent estimators in finite parameter sets is investigated, as well as the properties of the proposed allocation rule under the aforementioned estimator. Conditions under which the estimates lock-on to the true parameter are also investigated.

ABRÉGÉ

Dans cette thèse, nous étudions le problème stochastique de la machine à sous à multibras avec des processus dépendants de récompense pour chaque machine. Les processus sont générés par une mesure de probabilité inconnue. Nous assumons que pour chacune des machines, il y a une famille finie de mesures de probabilités indexées par un nombre (fini) de paramètres de substitution. Cette famille doit inclure la vraie mesure de probabilité qui génère le processus de récompense de chaque machine.

Dans ce cadre, nous proposons une politique d'allocation de type index qui utilise un estimateur convergent avec un regret super-linéaire.

En particulier, une instance d'estimateurs convergents pour des ensembles de paramètres finis est étudiée. Les propriétés de la politique d'allocation découlant des estimateurs proposés sont également examinées. Les conditions sous lesquelles les estimations reflètent la vraie valeur du paramètre en temps fini sont également considérées.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 The Multi-armed bandit problem

The multi-armed bandit (MAB) problems refer to a class of sequential allocation problems in which at each instant, a unit resource is allocated to one of several competing alternative actions/projects/treatments and a reward is obtained. The goal is to maximize the total reward obtained in a sequence of allocations.

The name *bandit* derives from an imagined slot machine with $J \geq 2$ arms. In a casino, a player/gambler experiences a sequential allocation problem. That is, he is facing many slot machines at once and he has to repeatedly decide which machine to choose to insert his coin chip.

Consider a measurable space $(\Omega, \mathcal{A})$ and J independent $\mathbb{R}^1$-valued reward processes

$$\left\{y_n^j;\ j \in \{1, \ldots, J\}\right\}_{n=1}^{\infty} \tag{1.1}$$

defined in $(\Omega, \mathcal{A})$, one for each bandit machine. The probability law on $\{y_n^j; n \in \mathbb{Z}_{>0}\}$ is $\mathbb{P}_{\vartheta_j^*}^j$, where $\vartheta_j^* \in \Theta_j$ is an unknown parameter and Θ_j is a known parameter set . Furthermore, $\mathcal{A}_n \triangleq \sigma(y_1^j, \ldots, y_n^j)$ denotes the Borel σ-field generated by n observations from the j-th reward process.

Let $\mu^j(\vartheta_j)$ denote the mean under the measure $\mathbb{P}^j_{\vartheta_j}$. It is assumed that $\mu^j(\vartheta_j) < \infty$ for all $\vartheta_j \in \Theta_j$, and all $j \in \{1, \ldots, J\}$.

An allocation rule ϕ is a sequence of measurable functions $\varphi_1, \varphi_2, \ldots$, where

$$\varphi_T : \mathbb{R}^{T-1} \longrightarrow \{1, \ldots, J\}.$$

At stage T, φ_T indicates which reward process to observe on the basis of the past observations $Y_1, \ldots, Y_{T-1}$. This selection procedure is described in more detail as follows:

Let $(n_1^1, n_1^2, \ldots, n_1^J) = (0, 0, \ldots, 0)$ denote the *local* time of each machine at stage one. At stage T observe the reward process corresponding to machine j_T, specified by

$$j_T \triangleq \varphi_T(Y_1, \ldots, Y_{T-1})$$

and set the *local* time of machine $j \in \{1, \ldots, J\}$ as

$$n_T^j \triangleq \begin{cases} n_{T-1}^j + 1 & , \text{if } j = j_T \\ n_{T-1}^j & , \text{if } j \neq j_T \end{cases}$$

Then the observation Y_T at stage T takes the value of $y^{j_T}_{n_T^{j_T}}$. The local time n_T^j of machine $j \in \{1, \ldots, J\}$ counts the number of times that machine j had been used in T total plays. For that reason, T is called the *global* time of the system and satisfies $T = n_T^1 + \ldots + n_T^J$.

Let $j^* = \underset{j \in \{1, \ldots, J\}}{\operatorname{argmin}} \{\mu^j(\vartheta_j^*)\}$ denote the machine with the highest reward process mean. In the sequel, machine j^* will also be referred as "the best" machine. The

regret at stage T is a random variable defined as

$$R_T(\omega, \phi) = \sum_{t=1}^{T} \left(\mu^{j^*}(\vartheta_{j^*}^*) - Y_t \right)$$

and the expected regret at stage T is then given by

$$\mathbb{E}(R_T(\omega, \phi)) \quad = T\mu^{j^*}(\vartheta_{j^*}^*) - \sum_{t=1}^{T} \left(\mathbb{E}Y_t \right) \ . \tag{1.2}$$

Notice that $\sum_{t=1}^{T} Y_t = \sum_{j=1}^{J} \sum_{i=1}^{n_T^j} y_i^j$. Thus

$$\begin{aligned} \sum_{t=1}^{T} \left(\mathbb{E}Y_t \right) &= \mathbb{E}\left(\sum_{j=1}^{J} \sum_{i=1}^{n_T^j} y_i^j \right) \\ &= \mathbb{E}\left\{ \mathbb{E}\left(\sum_{j=1}^{J} \sum_{i=1}^{n_T^j} y_i^j \Big| n_T^j \right) \right\} \\ &= \mathbb{E}\left(\sum_{j=1}^{J} n_T^j \mathbb{E}(y_i^j) \right) \\ &= \sum_{j=1}^{J} \mu^j(\vartheta_j^*) \mathbb{E}\left(n_T^j \right). \end{aligned} \tag{1.3}$$

In addition,

$$\begin{aligned} T\mu^{j^*}(\vartheta_{j^*}^*) &= \sum_{t=1}^{T} \mathbb{E}\left(y^{j^*} \right) \\ &= \mathbb{E}\left\{ \mathbb{E}\left(\sum_{j=1}^{J} \sum_{i=1}^{n_T^j} y^{j^*} \Big| n_T^j \right) \right\} \\ &= \mathbb{E}\left(\sum_{j=1}^{J} n_T^j \mathbb{E}(y^{j^*}) \right) \\ &= \sum_{j=1}^{J} \mu^{j^*}(\vartheta_{j^*}^*) \mathbb{E}\left(n_T^j \right). \end{aligned} \tag{1.4}$$

Substituting (1.3) and (1.4) into (1.2), the expected regret at time T can be restated as

$$\mathbb{E}(R_T(\omega,\phi)) \;\; = \sum_{j=1}^{J} \left(\mu^{j^*}(\vartheta_{j^*}^*) - \mu^j(\vartheta_j^*)\right) \mathbb{E}(n_T^j). \tag{1.5}$$

Equation (1.5) reveals that the expected regret at time T is proportional to the expected local time of each machine $j \in \{1,\ldots,J\}$ up to time T, with a constant depending on the difference between the reward mean of the best machine and the reward mean of the j-th machine.

The MAB problem is to minimize the rate of growth of $\mathbb{E}(R_T(\omega;\phi))$ as $T \to \infty$, or equivalently to find functions $f_L(T)$ and $f_U(T)$ (e.g. finite,logarithmic,linear, etc.) such that there are constants $C_L, C_U > 0$, for which

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

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

Multi-armed bandit problems are paradigms of allocation problems in which the decision maker experiences the exploration versus exploitation dilemma. Precisely, the decision maker 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.

In the casino example mentioned earlier, the information about the system available for the player is limited. Thus his/her "winning" strategy necessarily involves a balance between exploring the unknown system to find profitable

machines while pulling the empirically best machine as often as possible. Such problems arise in a variety of branches of engineering and sciences.

Some other examples (among several) of the MAB problem are in advertisement (Ad) placement, in source routing and in cognitive radio communications.

In advertisement placement [15] 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 received.

In source routing [7] the MAB problem arises in terms of choosing a path between a source and a destination, among several alternatives, to transmit a packet at each transition instant. The reward in this case is associated to the transmission time or the transmission cost of the packet.

In cognitive radio communications [9], 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.

There are three fundamental formulations of the MAB problem depending on the nature of the reward process, and subsequently on the performance criterion of each formulation: the stochastic, the adversarial and the Markovian. This thesis is entirely focused on the stochastic formulation. Information about the adversarial formulation can be found in [3] and for the Markovian in [8].

1.2 Literature Review on MAB problems

The stochastic MAB problem was initially introduced by Robbins in [12]. The original motivation for studying MAB problems goes back to the 1930's in resource allocation problems that arise in clinical trials [13]. The MAB problem in clinical trials arises when different treatments are available for a disease and the decision maker must decide which treatment to use on the next patient.

Many years later Lai and Robbins in their seminal work [10], [11], considering a class of *uniformly good* allocation rules and specific families of IID reward distributions (including the exponential families), provided an asymptotic *lower bound* on the regret of $\mathbf{O}(\log T)$ with constant that depends on the Kullback-Leibler number. This result is given in the following theorem.

Theorem 1 (Asymptotic lower bound on the regret)**.** *[10, Theorem 5]*

Let ϕ by any uniformly good policy, i.e. its regret satisfies

$$\mathbb{E}(R_T(\omega,\phi)) = \boldsymbol{o}(T^{\alpha}),\ 0<\alpha<1.$$

Then,

$$\limsup_{T\to\infty} \frac{\mathbb{E}\{R_T(\omega,\phi)\}}{\log T} \geq \sum_{k<j^*} \frac{(\mu^{j^*}(\vartheta^*_{j^*}) - \mu^{j}(\vartheta^*_{j}))}{D(\mathbb{P}_{\vartheta^*_j}||\mathbb{P}_{\vartheta^*_{j^*}})},$$

where $D(\mathbb{P}_{\vartheta^*_j}||\mathbb{P}_{\vartheta^*_{j^*}}) \triangleq \int \log \frac{\mathbb{P}_{\vartheta^*_j}}{\mathbb{P}_{\vartheta^*_{j^*}}} \mathrm{d}\mathbb{P}_{\vartheta^*_j}$ is the Kullback-Leibler number between the measure $\mathbb{P}_{\vartheta^*_j}$ of the reward process of any suboptimal machine and the measure $\mathbb{P}_{\vartheta^*_{j^*}}$ of the reward process of the best machine. ■

Thereafter, they introduced the technique of *upper confidence bound* (of the mean) for the construction of index policies which are asymptotically efficient, i.e. policies whose regret achieve the lower bound described by Theorem 1.

Definition 1. *Index policy*

Let $g = \{g_{T,n} : \mathbb{R}^n \to \mathbb{R};\ T = 1, 2, \ldots;\ n = 1, \ldots, T\}$ *be a set of Borel-measurable functions, called indexes. An index policy* ϕ^g *is a policy which is specified by the collection of functions* g*, such that:*

1. *In the first* J *stages, it chooses each machine once.*
2. *Based on the rewards* $y_1^j, \ldots, y_{n_T^j}^j$ *yielded from machine* j *up to time* T *for* $T > J$*, it computes an index* g_{T,n_T^j} *and uses the machine with the highest index at* $T + 1$*.*

■

Definition 2. *Upper Confidence Bound (UCB) policy*

For each machine $j \in \{1, \ldots, J\}$*, the upper confidence bound (UCB) index, at instant* T*, is a Borel-measurable function*

$$g_{T,n} : \mathbb{R}^n \to \mathbb{R}\ (T = 1, 2, \ldots; 1 \leq n \leq T),$$

satisfying

A1. $g_{T,n}$ *is non-decreasing in* $T \geq n$*, for each fixed* $n \in \mathbb{Z}_{>0}$ *.*

A2. *Let* $y_1^j, y_2^j, \ldots, y_n^j$ *be a sequence of observations from machine* j*. Then, for any* $z < \mu^j(\vartheta_j^*)$*,*

$$\mathbb{P}_{\vartheta_j^*}\left\{g_{T,n}\left(y_1^j, \ldots, y_n^j\right) < z,\ \textit{for some } n \leq T\right\} = \boldsymbol{o}(T^{-1}) \tag{1.6}$$

Any index policy ϕ^g *whose index functions* g *satisfy A1-A2, is called upper confidence bound (UCB) policy.* ■

The notation used in the right hand side of (1.6) is the *little-o* notation where for any two functions f and g defined in $\mathbb{R}$, $f(T) = \mathbf{o}(g(T))$ means

$$\lim_{T\to\infty} \frac{f(T)}{g(T)} = 0.$$

Lai and Robbins's UCB policy. Let $h_n(y_1, \ldots, y_n) \triangleq \frac{1}{n}\sum_{i=1}^{n} y_i$ denote the sample mean of some $\mathbb{R}^1$-valued reward process $\{y_n; n \in \mathbb{Z}_{>0}\}$. For each machine $j \in \{1, \ldots, J\}$ define

$$\begin{aligned} \mu^j_{n^j_T} &\triangleq h_{n^j_T}(y^j_1, \ldots, y^j_{n^j_T}) \\ U^j_{n^j_T} &\triangleq g_{T,n^j_T}(y^j_1, \ldots, y^j_{n^j_T}) \end{aligned} \tag{1.7}$$

where $U^j_{n^j_T}$ is an upper confidence bound of the mean $\mu^j_{n^j_T}$ of the reward process of machine j. Let also $0 < \delta < J$, where J denotes the number of bandit machines in the system. Then,

For $T \leq J$: Sample each machine once.

For $T > J$: Let $j^* \triangleq \text{argmax}\{\mu^j_{n^j_T} : n^j_T > \delta T\}$. Take an observation from some machine $j \neq j^*$ only if $\mu^{j^*}_{n^{j^*}_T} \leq U^j_T$ and sample from j^* otherwise. ▲

In other words, Lai and Robbins's policy samples from machine j if the corresponding upper confidence bound $U^j_{n^j_T}$ of its mean does not fall below the estimated mean of machine j^*; otherwise it samples from machine j^*. Thus, Lai and Robbins's policy is an index policy, with index functions as described by (1.7).

During the years following Lai and Robbins's work, construction of UCB policies received considerable attention by researchers mostly because of the

simplicity of the principal idea behind them which is usually referred in the literature as *optimism in face of uncertainty*.

Optimism in face of uncertainty is described in [3] as follows: First, a set of plausible environments which are consistent with the data is constructed (i.e. index functions). Then, the most favourable environment is identified in this set. The decision which is optimal in this most favourable and plausible environment should be made.

However, Lai and Robbins's policies suffered from two drawbacks. They had been designed for specific families of probability distributions. Second, their upper confidence bounds were in general complicated and mostly relied on the entire sequence of rewards which could raise computational issues.

About a decade later, Agrawal in [1] in an attempt to address the drawbacks of Lai and Robbins's policies, constructed simple UCB policies that depend on the rewards obtained from a machine only through their sample mean, and yet achieve a $\mathbf{O}(\log(T))$ regret. These policies were applicable in a more general class of distributions including the one-parameter exponential families and members of the one-parameter shifted families.

Agrawal's UCB policy. Let $\mathbf{g} = \{g_{T,n}; T = 1, 2, \ldots; n = 1, \ldots, T\}$ be a collection of Borel-measurable functions satisfying Definition 2. Then,

For $T \leq J$**:** Sample each machine once.

For $T > J$**:** Sample the machine satisfying $j^* \triangleq \text{argmax}\{g^j_{n^j_T}; j \in \{1, \ldots, J\}\}$. ▲

In other words, Agrawal's policy compares only the upper confidence bounds indexes of each machine, and chooses to sample the machine with the highest

index. Agrawal also proved the following upper bound for the regret achieved by his policies.

Theorem 2 (Asymptotic upper bound on the regret)**.** *[1, Theorem 2.2]*

Let ϕ^g be any UCB policy. Then, the local-time of each machine $j \in \{1, \ldots, J\}$ satisfies

$$\begin{aligned} n_T^j &= 1 + \sup\{1 \leq n \leq T : \ g_{T,n}^j(y_1^j, \ldots, y_n^j) \geq \mu^{j^*}(\vartheta_{j^*}^*) - \varepsilon\} \\ &\quad + \sum_{k=1}^{T} \mathbb{1}\{g_{k,n}^{j^*}(y_1^{j^*}, \ldots, y_n^{j^*}) < \mu^{j^*}(\vartheta_{j^*}^*) - \varepsilon \text{ for some } 1 \leq n \leq T\}. \end{aligned} \tag{1.8}$$

where $\mathbb{1}_{\{\bullet\}}$ denotes the indicator function.

Moreover, its regret satisfies

$$\limsup_{T\to\infty} \frac{\mathbb{E}\{R_T(\omega, \phi^g)\}}{\log T} = \sum_{k<j^*} \frac{(\mu^{j^*}(\vartheta_{j^*}^*) - \mu^j(\vartheta_j^*)}{K(\mathbb{P}_{\vartheta_k^*}^k, \mu^{j^*}(\vartheta_{j^*}^*))}, \tag{1.9}$$

where, for all $k < j^*$

$$\frac{1}{K(\mathbb{P}_{\vartheta_k^*}^k, \mu^{j^*}(\vartheta_{j^*}^*))} \triangleq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}_{\vartheta_k^*}\left\{\sup\{1 \leq n_T^k \leq T : g^k(y_1^k, \ldots, y_{n_T^k}^k) \geq \mu^{j^*}(\vartheta_*^*) - \varepsilon\}\right\}}{\log T} \tag{1.10}$$

■

A drawback of Agrawal's policies is that although they achieve an expected regret with logarithmic rate of growth, they may not have the best constant which is the price for getting computationally easier allocation rules.

More recently, Auer, Casa-Bianchi and Fisher [2] strengthened Agawal's results by constructing UCB policies that achieve logarithmic regret *uniformly in*

time rather than only asymptotically, for arbitrary classes of reward distributions with bounded support.

The work of Auer, Casa-Bainchi and Fisher has spurred a large literature on different variations of the basic setup. We efer the reader to [3] for a detailed overview of the recent results.

A central result in their work is the construction of so-called UCB1 algorithm which is simple to implement and computationally efficient. The set of index functions employed by UCB1 is defined as

$$g = \left\{ g^j_{T,n^j_T}\left(y^j_1,\ldots,y^j_{n^j_T}\right) = \frac{1}{n^j_T}\sum_{i=1}^{n^j_T} y^j_i + \sqrt{\frac{2\log T}{n^j_T}};\ T, n^j_T \in \mathbb{Z}_{>0},\ j \in \{1,\ldots,J\} \right\}$$

The index functions of UCB1 are the sum of two terms:

The first term is the sample mean of the rewards associated with machine j. Assuming that each machine is played infinitely often, then the sample mean of machine j will eventually converge to the true mean $\boldsymbol{\mu}^j(\vartheta^*_j)$. An important remark on this is that since the reward of each machine is a real number, the mean estimation takes place in $\mathbb{R}$.

The second term of the index functions defined in [2] is related to the size of the one-sided confidence interval for the average reward (according to Chernoff-Hoeffding bounds) within which the true reward mean of each machine falls with overwhelming probability.

1.3 Contribution of the thesis

In this thesis, we study the stochastic multi-armed bandit problem with finite parameter set for each machine.

In chapter 2 we propose an index policy Φ^g which, in general, is not uniformly good. Φ^g is constructed by modifying UCB1 to include consistent estimators of the unknown parameter of each machine, and under certain assumptions on the convergence rate of the unknown parameter estimates, Φ^g is shown to be a UCB policy. This modification allows the application of the proposed policy Φ^g in a more general class of MAB problems, including those with dependent reward processes across time, which is the main contribution of this chapter.

In chapter 3, we discuss about the maximum likelihood estimator (MLE) which is a paradigm of a parameter estimator extensively used in statistics and information sciences. We present existing results in the literature about the strong consistency of the MLE for parameter estimation in finite sets. To this end, we introduce the so-called switch wrong and corrected (SWAC) condition which is a sufficient condition for the existence of finite moments up to order $2 + \alpha$, for some $\alpha > 0$, of the lock on time of the ML estimates to the true parameter, which is the main contribution of this chapter.

In chapter 4, we investigate the behaviour of the regret achieved by Φ^g in a linear and Gaussian scenario. We also present some illustrative simulation results.

CHAPTER 2
The proposed allocation rule Φ^g

As mentioned earlier, an important attribute of the allocation rule proposed in this thesis is that it employs consistent estimators of the true parameter ϑ_j^* for the measure $\mathbb{P}_{\vartheta_j^*}$ generating the reward process $\{y_1^j, \ldots, y_{n_T^j}^j\}$ of each machine $j \in \{1, \ldots, J\}$.

Before introducing the proposed policy, it is worthwhile to briefly discuss about consistency of estimates.

2.1 Consistency of estimates

Consider a reward process $y_1^j, \ldots, y_n^j$ generated by $\mathbb{P}_{\vartheta^*}^j$ where $\vartheta_j^* \in \Theta_j$ ($|\Theta_j| < \infty$), for some machine $j \in \{1, \ldots, J\}$. Any mapping $\hat{\vartheta}_n^j = \hat{\theta}_j(y_1, \ldots, y_n)$ on Ω into Θ_j which is $\mathcal{A}_n$ measurable is called an *estimator*.

Consider now the MAB problem as formulated in chapter 1, and let $\left\{\hat{\vartheta}_n^j; j \in \{1, \ldots, J\}\right\}_{n=1}^{\infty} = \hat{\vartheta}_1^j, \hat{\vartheta}_2^j, \ldots$ be a sequence of estimates of the unknown parameter, corresponding to the j-th machine. The estimator $\hat{\theta}_j$ is called strongly *consistent* if $\hat{\vartheta}_n^j \neq \vartheta_j^*$ infinitely often with $\mathbb{P}_{\vartheta_j^*}$ probability 0.

In the sequel, it is assumed that for each machine $j \in \{1, \ldots, J\}$, $\hat{\theta}_j$ is strongly consistent, and $N_j(\omega)$ is defined as its (random) *lock-on* time to the true parameter ϑ_j^*. A more precise definition for the lock-on time is given as follows.

Definition 3. *Lock-on time*

For each machine $j \in \{1, \ldots, J\}$ *there is* $\Omega_o^j \subset \Omega$ *with* $\mathbb{P}_{\vartheta_j^*}(\Omega_o^j) = 1$*, such that, for all* $\omega \in \Omega_o^j$ *the lock-on time is the least* $N_j(\omega)$ *for which for all* $n > N_j(\omega)$*,*

$$\hat{\vartheta}_n^j = \vartheta_j^*.$$

■

2.2 The allocation rule Φ^g

Motivated by the work of Auer et al.[2], we propose an index type policy, denoted by Φ^g, which at each stage T defines an index for each machine, and based on that it chooses to play the machine with the highest index in the next stage $T+1$. The set of index functions g employed by Φ^g is defined as

$$g = \left\{ g^j_{T,n^j_T}\left(y^j_1, \ldots, y^j_{n^j_T}\right) = \mu^j\left(\hat{\vartheta}^j_{n^j_T}\right) + \frac{T/C}{n^j_T};\ T, n^j_T \in \mathbb{Z}_{>0},\ C \in \mathbb{R},\ j \in \{1, \ldots, J\} \right\} \tag{2.1}$$

The index functions employed by Φ^g, instead of estimating the reward mean in the set of real numbers, they estimate the true parameter of the unknown reward measure of each process.

This type of mean value estimation takes advantage of the finiteness of the parameter space of each machine, since instead of estimating the true mean in the set of real numbers, the estimation of the unknown parameters takes place in the finite parameter sets Θ_j. That idea is illustrated in figure 2–1.

The second term, motivated by UCB1 in [2], defines a *switching rule*, preventing the player from locking on sampling a specific (possibly wrong) machine

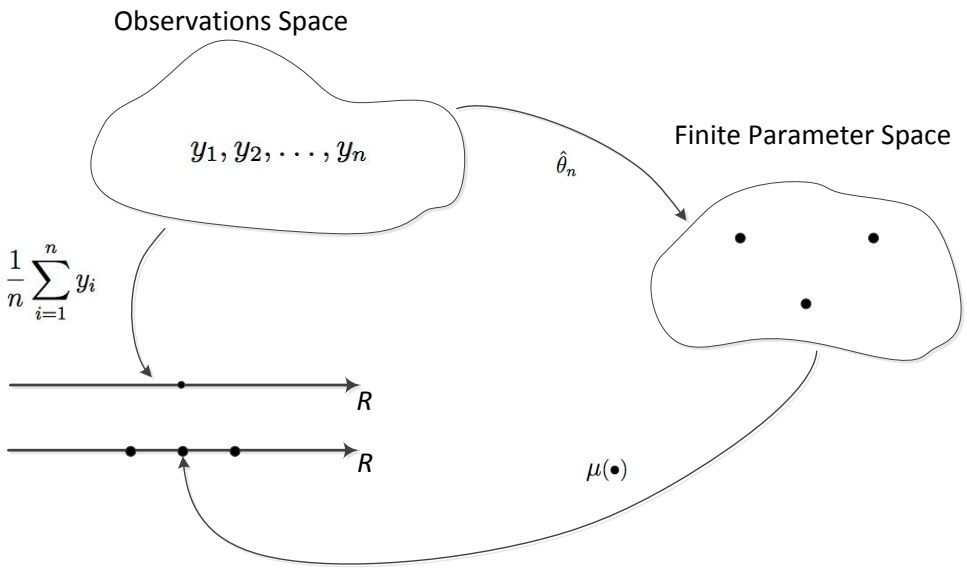

Figure 2–1: UCB1 proposed in [2] employs the sample mean of the (IID) observations which needs infinitely many samples to converge to their true mean. The policy Φ^g proposed in this thesis, is estimating the true mean of the, in general dependent, observations by performing parameter estimation in a *finite* parameter space.

indefinitely, while ignoring the others. Specifically, for the machines which had been sampled many times, this term has relatively small value compared to its value for machines that had been sampled less times.

Thus, the second term is used to increase the index value of the rarely sampled machines forcing the allocation rule to choose them at some instant in the future instead of ignoring them.

2.3 Properties of the adaptive allocation rule Φ^g

Lemma 1. *For each machine $j \in \{1, 2, \ldots, J\}$,*

$$\lim_{T\to\infty} n_T^j = \infty \quad a.s.$$

■

Proof. (By contradiction) Assume there is $\omega \in \Omega$ and a subset $\mathcal{L} \subset \mathcal{J} = \{1, \ldots, J\}$ of machines which are chosen finitely many times, while machines in $\mathcal{J} \setminus \mathcal{L}$ are chosen infinitely many times. That means

$$\forall j \in \mathcal{L},\ \exists q_j > 0 \text{ s.t. } \lim_{T\to\infty} n_T^j = q_j < \infty \tag{2.2}$$

and

$$\forall k \in \mathcal{J} \setminus \mathcal{L}, \lim_{T\to\infty} n_T^k = \infty \tag{2.3}$$

By the assumption that each machine $j \in \mathcal{L}$ had been played finitely many times, it is implied that for all $j \in \mathcal{L}$ there exists $\mathcal{T} > 1$ such that for all $T > \mathcal{T}$

$$\begin{aligned}
& \mu^j\left(\hat{\vartheta}^j_{n_T^j}\right) + \tfrac{T/C}{n_T^j} && \leq \max_{k\in\mathcal{J}\setminus\mathcal{L}} \left\{\mu^k\left(\hat{\vartheta}^k_{n_T^k}\right) + \tfrac{T/C}{n_T^k}\right\} \\
\Longrightarrow\ & \min_{\Theta_j}\left\{\mu^j\left(\hat{\vartheta}^j_{n_T^j}\right)\right\} + \tfrac{T/C}{n_T^j} && \leq \max_{k\in\mathcal{J}\setminus\mathcal{L}} \left\{\max_{\Theta_k}\left\{\mu^k\left(\hat{\vartheta}^k_{n_T^k}\right)\right\} + \tfrac{T/C}{n_T^k}\right\} \\
\Longrightarrow\ & \tfrac{\gamma-\Gamma}{T/C} + \tfrac{1}{n_T^j} && \leq \max_{k\in\mathcal{J}\setminus\mathcal{L}} \left\{\tfrac{1}{n_T^k}\right\}
\end{aligned}$$

where $\gamma = \min_{\Theta_j}\left\{\mu^j\left(\hat{\vartheta}^j_{n_T^j}\right)\right\}$ and $\Gamma = \max_{k\in\mathcal{J}\setminus\mathcal{L}}\left\{\max_{\Theta_k}\left\{\mu^k\left(\hat{\vartheta}^k_{n_T^k}\right)\right\}\right\}$.

But

$$\begin{aligned} \lim_{T\to\infty}\{\tfrac{\gamma-\Gamma}{T/C}+\tfrac{1}{n_T^j}\} &\leq \lim_{T\to\infty}\max_{k\in\mathcal{J}\setminus\mathcal{L}}\left\{\tfrac{1}{n_T^k}\right\} \\ \implies \quad \tfrac{1}{q_j} &\leq 0 \end{aligned}$$

which leads to a contradiction since q_j was assumed to be a finite positive. □

Assume the following conditions: For all $j \in \{1, \ldots, J\}$

A3. The estimator $\hat{\vartheta}^j$ of the unknown parameter ϑ_j^* is strongly consistent.

A4. For the lock-on time $N_j(\omega)$,

$$\mathbb{E}(N_j(\omega)^{2+\alpha}) < \infty, \text{ for some } \alpha > 0$$

A sufficient condition for A3 is given by Theorem 6 and a sufficient condition for A4 is the *summable wrong and corrected* (SWAC) condition, which are both discussed in Chapter 3.

Theorem 3. *Assume that A3 holds.*

1. *Then for each $j \in \{1, \ldots, J\}$, the index function g^j given in* (2.1) *is an Upper Confidence Bound (UCB).*
2. *If in addition, A4 also holds, then the regret of Φ^g satisfies $\mathbb{E}\{R_T(\omega, \Phi^g)\} = \boldsymbol{o}(T^{1+\delta})$ for some $\delta > 0$.*

■

Proof. To prove part (1) we need to show that A1 and A2 hold (see Definition 2) . Consider the index function g^j described in (2.1), for some $j \in \{1, \ldots, J\}$. For any fixed $n \leq T$, the estimate $\boldsymbol{\mu}^j(\vartheta_n^j)$ is constant, which means that (2.1) is a function

depending only on T. Thus for any fixed $n \leq T$, the index g^j is increasing in T for all $j \in \{1, \ldots, J\}$ which shows that A1 holds.

Moreover, for every $j \in \{1, 2, \ldots, J\}$, by Lemma 1 and A3, we have that for all $\omega \in \bigcap_{j=1}^{J} \Omega_o^j \subseteq \Omega$, where Ω_o^j is such that $\mathbb{P}_{\vartheta_j^*}(\Omega_o^j) = 1$, and for all $n > N_j(\omega)$

$$\hat{\vartheta}_n^j = \vartheta_j^* \tag{2.4}$$

where $N_j(\omega)$ denotes the (random) lock-on time of the j-th estimate to the true parameter $\vartheta_j^* \in \Theta_j$. In addition, define

$$\begin{aligned} B_n^j &\triangleq \{\omega : N_j(\omega) < n\}, \\ A_{T,n}^j &\triangleq \left\{\omega : g_{T,n}^j(y_1^j(\omega), \ldots, y_n^j(\omega)) < z\right\} \text{ for any } z < \mu^j(\vartheta_j^*), \text{ and} \\ A_T^j &\triangleq \left\{\omega : g_{T,n}^j(y_1^j(\omega), \ldots, y_n^j(\omega)) < z \text{ for some } n \leq T\right\} = \bigcup_{n=1}^{T} A_{T,n}^j. \end{aligned}$$

Then,

$$\mathbb{P}_{\theta_j^*}\left(A_{T,n}^j \;\middle|\; B_n^j\right) = 0 \tag{2.5}$$

In addition, let $T_j^* = \left\lfloor \frac{T}{C\left(\mu^*(\vartheta_{j^*}^*) - \min\limits_{\vartheta_j \in \Theta_j} \mu(\vartheta_j)\right)} \right\rfloor$ where $\lfloor \bullet \rfloor$ denotes the entier function. Then

$$\mathbb{P}_{\theta_j^*}\left(A_{T,n}^j\right) = 0, \quad \forall n < T_j^*. \tag{2.6}$$

.

Thereafter, consider

$$\begin{aligned} T\mathbb{P}(A_T^j) &= T\mathbb{P}\left(\bigcup_{i=1}^{T} A_{T,i}^j\right) \\ &\leq T\sum_{i=1}^{T}\mathbb{P}(A_{T,i}^j) \qquad \text{(by the union bound)} \\ &\leq T\sum_{i=T^*+1}^{T}\mathbb{P}(A_{T,i}^j) \quad \text{(by (2.6))} \end{aligned} \tag{2.7}$$

Using the law of total probability we have

$$\begin{aligned} T\sum_{i=T^*+1}^{T}\mathbb{P}(A_{T,i}^j) &\leq T\sum_{i=T^*+1}^{T}\mathbb{P}(A_{T,i}^j\Big|B_i^j)\mathbb{P}(B_i^j) + \mathbb{P}(A_{T,i}^j\Big|B_i^{j^{\complement}})\mathbb{P}(B_i^{j^{\complement}}) \\ &= T\sum_{i=T^*+1}^{T}\mathbb{P}(A_{T,i}^j\Big|B_i^{j^{\complement}})\mathbb{P}(B_i^{j^{\complement}}) \qquad \text{(by (2.5))} \end{aligned} \tag{2.8}$$

Substituting (2.8) into (2.7) we have

$$\begin{aligned} T\mathbb{P}(A_T^j) &\leq T\sum_{i=T^*+1}^{T}\mathbb{P}(A_{T,i}^j\Big|B_i^{j^{\complement}})\mathbb{P}(B_i^{j^{\complement}}) \\ &\leq T\sum_{i=T^*+1}^{T}\mathbb{P}(B_i^{j^{\complement}}) \\ &\leq T\sum_{i=T^*+1}^{T}\tfrac{\mathbb{E}(N_j(\omega)^{2+\alpha})}{n^{2+\alpha}} \qquad \text{(by Markov in.)} \\ &\leq T\mathbb{E}(N_j(\omega)^{2+\alpha})\int_{i=T^*+1}^{T}\tfrac{1}{n^{2+\alpha}}\mathrm{d}n \\ &= \tfrac{T\mathbb{E}(N_j(\omega)^{2+\alpha})}{1+\alpha}((T^*+1)^{-(1+\alpha)} - T^{-(1+\alpha)}), \quad \text{for some } \alpha > 0. \end{aligned} \tag{2.9}$$

Taking the limit as $T \to \infty$ and under A4 we have

$$\lim_{T\to\infty}\left[\frac{T\mathbb{E}(N_j(\omega)^{2+\alpha})}{1+\alpha}\left((T^*+1)^{-(1+\alpha)} - T^{-(1+\alpha)}\right)\right] = 0 \tag{2.10}$$

which shows that A2 (see chapter 1) holds, and completes the proof for part (1).

For part (2) consider the upper bound of the local time of each machine $j \in \{1, \ldots, J\}$ given in Theorem 2, which can be re-expressed as follows:

$$n_T^j \quad \leq 1 + \sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\} + \sum_{t=1}^{T} \mathbb{1}_{A_t^n} \tag{2.11}$$

Then,

$$\begin{aligned}
n_T^j \quad &\leq 1 + \sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\} + \sum_{t=1}^{T} \mathbb{1}_{A_t^n} \\
\implies \quad \mathbb{E}\{n_T^j\} \quad &\leq \mathbb{E}\{1 + \sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\} + \sum_{t=1}^{T} \mathbb{1}_{A_t^n}\} \\
\implies \quad \frac{\mathbb{E}\{n_T^j\}}{T^{1+\delta}} \quad &\leq \frac{\mathbb{E}\{1\} + \mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\}\} + \mathbb{E}\{\sum_{t=1}^{T} \mathbb{1}_{A_t^n}\}\}}{T^{1+\delta}}
\end{aligned}$$

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

$$\begin{aligned}
\limsup_{T\to\infty} \frac{\mathbb{E}\{n_T^j\}}{T^{1+\delta}} \quad &\leq \limsup_{T\to\infty} \frac{\mathbb{E}\{1\} + \mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\}\} + \mathbb{E}\{\sum_{t=1}^{T} \mathbb{1}_{A_t^n}\}\}}{T^{1+\delta}} \\
\implies \limsup_{T\to\infty} \frac{\mathbb{E}\{n_T^j\}}{T^{1+\delta}} \quad &\leq \limsup_{T\to\infty} \frac{\mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\}\}}{T^{1+\delta}} \quad \text{(by A2)} \\
\implies \limsup_{T\to\infty} \frac{\mathbb{E}\{n_T^j\}}{T^{1+\delta}} \quad &\leq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}\{\sup\{1 \leq n \leq T : g_{T,n}(y_1^j, \ldots, y_n^j \geq \mu^{j^*}(\theta_j^*) - \varepsilon)\}\}}{T^{1+\delta}}.
\end{aligned}$$

In light of (1.2), an expression involving the expected regret is then given by

$$\limsup_{T\to\infty} \frac{\mathbb{E}\{R_T(\omega, \Phi^g)\}}{T^{1+\delta}} = \sum_{k<j^*} \frac{(\mu^{j^*}(\vartheta_{j^*}^*) - \mu^k(\vartheta_k^*))}{K(\mathbb{P}_{\vartheta_k^*}^k, \mu^{j^*}(\vartheta_{j^*}^*))}, \quad \forall \delta > 0. \tag{2.12}$$

where, for all $k < j^*$

$$\frac{1}{K(\mathbb{P}_{\vartheta_k^*}^k, \mu^{j^*}(\vartheta_{j^*}^*))} \triangleq \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}_{\vartheta_k^*}\left\{\sup\{1 \leq n_T^k \leq T : g^k(y_1^k, \ldots, y_{n_T^k}^k) \geq \mu^{j^*}(\vartheta_{j^*}^*)) - \varepsilon\}\right\}}{T^{1+\delta}} \tag{2.13}$$

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

$$\begin{aligned}\frac{1}{K(\mathbb{P}^k_{\vartheta^*_k},\mu^{j^*}(\vartheta^*_{j^*}))} &= \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{\mathbb{E}_{\vartheta^*_k}\left\{\sup\{1\le n^k_T\le T:\ \mu(\hat{\vartheta}^k_{n^k_T})+\frac{T/C}{n^k_T}\ge\mu^{j^*}(\vartheta^*_{j^*}))-\varepsilon\}\right\}}{T^{1+\delta}} \\ &\le \inf_{\varepsilon>0} \limsup_{T\to\infty} \frac{T}{T^{1+\delta}} \to 0 \\ &= 0 \quad \forall\delta>0.\end{aligned}$$

But this means

$$\limsup_{T\to\infty} \frac{\mathbb{E}\{R_T(\omega,\Phi^g)\}}{T^{1+\delta}} = 0 \tag{2.14}$$

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

Theorem 3 shows that Φ^g is a UCB policy with super-linear regret which is suboptimal compared to the $\mathbf{O}(\log T)$ regret achieved by the class of uniformly good policies designed for IID reward processes in [12]. Furthermore, (2.14) implies that Φ^g is not uniformly good in the sense described in Theorem 1.

It is noted that for the case of dependent multi-armed bandits, the asymptotic lower bound of the regret function is not known. Therefore there is not any known optimality criterion for the evaluation of the asymptotic behaviour of the proposed policy under dependent reward processes.

In the following chapter we study the maximum likelihood estimator which is a particular instant of a consistent estimator under dependent observations, and is widely used in statistics and information sciences. We firstly present existing results for the consistency of the maximum likelihood estimates in finite parameter spaces. Thereafter, we investigate sufficient conditions to satisfy A4.

CHAPTER 3
Maximum Likelihood estimation

Consider a measurable space $(\Omega, \mathcal{A})$ and a finite set $\Theta = \{\vartheta^1, \ldots, \vartheta^K\}$ of parameters of cardinality $|\Theta| = K < \infty$. Let $\{\mathbb{P}_{\vartheta^k}; \vartheta^k \in \Theta\}$ be a family of probability measures defined on $\mathcal{A}$. It is assumed that for every $\vartheta^k \in \Theta$, $\{y_n; n \in \mathbb{Z}_{>0}\}$ is a $\mathbb{R}^1$ process defined on $(\Omega, \mathcal{A}, \mathbb{P}_{\vartheta^k})$

Assume that $(\Omega, \mathcal{A})$ is the Cartesian product $\prod_{i=1}^{\infty}(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the one-dimensional Borel field in $\mathbb{R}^1$ and $\mathbb{P}_{\vartheta^k}$ is the probability measure induced in $\mathcal{A}$ by a set of probability distributions $\{p_{(\vartheta^k,n)}; n \in \mathbb{Z}_{>0}\}$ on $\prod_{i=1}^{n}(\mathbb{R}, \mathcal{B})$ according to the Kolmogorov's extension theorem. Then $\mathbb{P}_{\vartheta^k,n}$ is the restriction of the probability measure $\mathbb{P}_{\vartheta^k}$ to the σ-field $\mathcal{A}_n = \mathcal{B}(y_1, \ldots, y_n); n \in \mathbb{Z}_{>0}$.

In the sequel, ϑ^* denotes the true parameter in the set Θ; in other words, the process $\{y_n; n \in \mathbb{Z}_{>0}\}$ is generated according to the set of measures $\{\mathbb{P}_{\vartheta^*,n}; n \in \mathbb{Z}_{>0}\}$. The aforementioned set of measures will be the only family of measures that governs the observed process $\{y_n; n \in \mathbb{Z}_{>0}\}$.

3.1 The likelihood function

The likelihood function f_{ϑ^k} is defined as the Radon-Nikodym derivative of $\mathbb{P}_{\vartheta^k}$ with respect to some reference measure ν as

$$f_{\vartheta^k}(y_1, \ldots, y_n) = \frac{d\mathbb{P}_{\vartheta^k,n}}{d\nu}.$$

Usually, the reference measure is the Lebesgue measure in the case of continuous random variables and the counting measure for the case of discrete random variables.

3.2 The maximum likelihood estimate

Definition 4 (Maximum-likelihood estimate (MLE))**.**

MLE is called the estimate which satisfies

$$f_{\hat{\vartheta}_n}(y_1, \ldots, y_n) \geq \max_{\vartheta^k \in \Theta} \{f_{\vartheta^k}(y_1, \ldots, y_n)\}$$

where $y_1, \ldots, y_n$ is a sequence of observations. ■

3.3 The maximum likelihood ratio

Before starting the discussion on maximum-likelihood ratios we first adopt the following assumption on the probability measures.

Assumption A5 *For each $n \in \mathbb{Z}_{>0}$ the members of the family $\{\mathbb{P}_{\vartheta,n}(\bullet); \vartheta \in \Theta\}$ are each mutually absolutely continuous with respect to the true probability measure $\mathbb{P}_{\vartheta^*,n}(\bullet)$.* ■

As a consequence, outside $\mathbb{P}_{\vartheta^*,n}$ null sets, we can have the following definition for the maximum-likelihood ratio (MLR)

Definition 5 (Maximum-likelihood ratio (MLR))**.**

MLR, denoted by x_n, is the ratio of the maximum likelihood function over the true likelihood function $f_{\vartheta^*}(y_1, \ldots, y_n)$. That is,

$$x_n(y_1, \ldots, y_n) = \frac{f_{\hat{\vartheta}_n}(y_1, \ldots, y_n)}{f_{\vartheta^*}(y_1, \ldots, y_n)} \tag{3.1}$$

■

The maximum-likelihood ratio in (3.1) would be abbreviated to $x_n(y^n)$ or x_n when no confusion is likely to occur.

An important property of the maximum likelihood ratio process, is that it forms a positive submartingale.

The submartingale property of the maximum likelihood ratio process, and subsequently its convergence property [6], along with a variation of Wald's technique suggested in [14] are crucial in the proof of the strong convergence of the MLE in the literature.

Theorem 4 ([5], pp. 327-328)**.** *Under A5, maximum likelihood ratio processes are positive submartingales.* ■

Proof. Define $h_{\vartheta^k}(y_n|y^{n-1}) \triangleq \frac{f_{\vartheta^k}(y_n|y^{n-1})}{f_{\vartheta^*}(y_n|y^{n-1})}$. By definition of the maximum likelihood function we have:

$$0 \leq h_{\hat{\vartheta}_{n-1}}(y_{n-1}|y^{n-1})x_{n-1} \leq x_n = h_{\hat{\vartheta}_n}(y_n|y^{n-1})\frac{f_{\hat{\vartheta}_n}(y^{n-1})}{f_{\vartheta^*}(y^{n-1})} \leq h_{\hat{\vartheta}_n}(y_n|y^{n-1})x_{n-1} \quad (3.2)$$

Let $\mathbb{E}_{\vartheta^*}$ denote expectation with respect to the measure $\mathbb{P}_{\vartheta^*}$. Then by the positivity of the likelihood functions it is clear that $x_n \geq 0; n \in \mathbb{Z}_{>0}$ and

$$\begin{aligned}
\mathbb{E}_{\vartheta^*}\{x_n|\mathcal{A}_{n-1}\} &\geq \mathbb{E}_{\vartheta^*}\{h_{\hat{\vartheta}_{n-1}}(y_n|y^{n-1})x_{n-1}|\mathcal{A}_{n-1}\} \\
&= x_{n-1}\mathbb{E}_{\vartheta^*}\{h_{\hat{\vartheta}_{n-1}}(y_n|y^{n-1})|\mathcal{A}_{n-1}\} \qquad \text{a.s.}\mathbb{P}_{\vartheta^*} \\
&= x_{n-1}\mathbb{E}_{\vartheta^*}\{\tfrac{f_{\hat{\vartheta}_{n-1}}(y_n|y^{n-1})}{f_{\vartheta^*}(y_n|y^{n-1})}|\mathcal{A}_{n-1}\} \qquad \text{a.s.}\mathbb{P}_{\vartheta^*} \\
&= x_{n-1}\tfrac{f_{\vartheta^*}(y^{n-1})}{f_{\hat{\vartheta}_{n-1}}(y^{n-1})}\mathbb{E}_{\vartheta^*}\{\tfrac{f_{\hat{\vartheta}_{n-1}}(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\} \qquad \text{a.s.}\mathbb{P}_{\vartheta^*}
\end{aligned} \quad (3.3)$$

Next, for

$$I(y^{n-1}) \triangleq \mathbb{E}_{\vartheta^*}\{\frac{f_{\hat{\vartheta}_{n-1}}(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\} \text{ and } A_k \triangleq \{\omega; \hat{\vartheta}_{n-1}(\omega) = \vartheta^k\},\ \forall \vartheta^k \in \Theta$$

one has

$$I(y^{n-1}) = \sum_{\Theta} \mathbb{1}_{\{A_k\}} \mathbb{E}_{\vartheta^*}\{\frac{f_k(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\}$$

But

$$\begin{aligned}
\mathbb{E}_{\vartheta^*}\{\tfrac{f_k(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\} &= \mathbb{E}_{\vartheta^*}\{\tfrac{f_k(y_n|y^{n-1})}{f_{\vartheta^*}(y_n|y^{n-1})}\tfrac{f_k(y^{n-1})}{f_{\vartheta^*}(y^{n-1})}|\mathcal{A}_{n-1}\} \\
&= \tfrac{f_k(y^{n-1})}{f_{\vartheta^*}(y^{n-1})}\mathbb{E}_{\vartheta^*}\{\tfrac{f_k(y_n|y^{n-1})}{f_{\vartheta^*}(y_n|y^{n-1})}|\mathcal{A}_{n-1}\} && \text{a.s.}\mathbb{P}_{\vartheta^*} \\
&= \tfrac{f_k(y^{n-1})}{f_{\vartheta^*}(y^{n-1})}\mathbb{E}_{\vartheta^*}\{\tfrac{f_{\vartheta^*}(y^{n-1})}{f_k(y^{n-1})}\tfrac{f_k(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\} \\
&= \frac{f_k(y^{n-1})}{\cancel{f_{\vartheta^*}(y^{n-1})}}\frac{\cancel{f_{\vartheta^*}(y^{n-1})}}{f_k(y^{n-1})}\overbrace{\mathbb{E}_{\vartheta^*}\{\tfrac{f_k(y^n)}{f_{\vartheta^*}(y^n)}|\mathcal{A}_{n-1}\}}^{1} \\
&= \tfrac{f_k(y^{n-1})}{f_{\vartheta^*}(y^{n-1})} && \text{a.s.}\mathbb{P}_{\vartheta^*},\ \forall \vartheta^k \in \Theta
\end{aligned}$$

Then

$$\begin{aligned}
I(y^{n-1}) &= \sum_{\Theta} \mathbb{1}_{\{A_k\}} \tfrac{f_k(y^{n-1})}{f_{\vartheta^*}(y^{n-1})} \quad \text{a.s.}\mathbb{P}_{\vartheta^*} \\
&= \tfrac{f_{\hat{\vartheta}_{n-1}}(y^{n-1})}{f_{\vartheta^*}(y^{n-1})}
\end{aligned}$$

and the desired result follows from (3.3). □

3.4 The strong consistency of the maximum likelihood estimate process

The sequence of ML estimates $\{\hat{\vartheta}_n; n \in \mathbb{Z}_{>0}\}$ is called strongly consistent if $\hat{\vartheta}_n \neq \vartheta^*$ infinitely often with $\mathbb{P}_{\vartheta^*}$ probability 0, i.e. $\mathbb{P}_{\vartheta^*}\left(\bigcap_{n=0}^{\infty}\bigcup_{k=n}^{\infty} \hat{\vartheta}_k \neq \vartheta^*\right) = 0$, or equivalently, $\hat{\vartheta}_n \neq \vartheta^*$ finitely often with $\mathbb{P}_{\vartheta^*}$ probability 1, i.e. $\mathbb{P}_{\vartheta^*}\left(\bigcap_{n=0}^{\infty}\bigcup_{k=n}^{\infty} \hat{\vartheta}_k = \vartheta^*\right) = 1$.

As it has already been mentioned, the family $\{\mathbb{P}_{\vartheta^*,n}; n \in \mathbb{Z}_{>0}\}$ will be the only family of measures that governs the observed process $\{y_n; n \in \mathbb{Z}_{>0}\}$. The extent

to which we require the measures $\{\mathbb{P}_{\vartheta^k,n}; n \in \mathbb{Z}_{>0}\}$ to differ from the true family $\{\mathbb{P}_{\vartheta^*,n}; n \in \mathbb{Z}_{>0}\}$ is given by the following assumption.

Assumption A6 *For any $\varepsilon > 0$, there exists $\alpha(\varepsilon) > 1$,*

$$\mathbb{P}_{\vartheta^*}\left\{0 \leq h_{\hat{\vartheta}_{n-1}}(y_n|y^{n-1}) \leq \alpha, \text{ for all } n > K\right\} < \varepsilon,$$

where $\hat{\vartheta}_n \in \Theta$ and $K = |\Theta|$. ■

Theorem 5 ([4])**.** *Under A5 and A6 the maximum likelihood estimates are strongly consistent.*

Proof. [5, pp. 328-329] □

3.5 Application to the multi-armed bandit problem and the summable wrong and corrected (SWAC) condition

The generality of the maximum likelihood estimation as described in the previous sections makes this method applicable in parameter estimation problems which arise in multi-armed bandit problems with *dependent* reward processes. In such problems it is assumed that there is one ML estimator for each machine, and that A5 and A6 are satisfied for all machines. This implies that each ML estimator is consistent, by virtue of Theorem 5.

At this point, we introduce the so-called *summable wrong and corrected condition* (SWAC). In the sequel, it will be shown that for any machine $j \in \{1, \ldots, J\}$ with reward process satisfying this condition, the estimate process $\hat{\vartheta}^j_n$ has a lock-on time $N_j(\omega)$ satisfying A4.

Assumption A7. (SWAC) *For all machines $j \in \{1, \ldots, J\}$, the estimate processes $\hat{\vartheta}_n^j$ satisfies the following condition:*

$$\mathbb{P}_{\hat{\vartheta}_j^*}(\hat{\vartheta}_{n-1} \neq \vartheta_j^*, \hat{\vartheta}_n = \vartheta_j^*) < \frac{C}{n^{3+\beta}}, \tag{3.4}$$

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

Note that SWAC does not imply strong consistency. In addition, since a necessary condition for lock-on to the true parameter $\hat{\vartheta}_j^*$ at instant n is that $\hat{\vartheta}_{n-1} \neq \vartheta^*$ and $\hat{\vartheta}_n = \vartheta^*$, SWAC is consistent with the existence of some non-zero probability under which lock-on to the true parameter may never occur.

However, having strong consistency in force, SWAC implies the $2 + \alpha$ moment, and hence the first and second moments, of the random lock-on instant $N(\omega)$ are finite for $0 < \alpha < \beta$, where β appears in the definition of SAWC condition.

Theorem 6. *Assume that for all machines $j \in \{1, \ldots, J\}$ conditions A5-A7 are satisfied. Then, the (random) lock-on time $N_j(\omega)$ of the ML estimate $\{\hat{\vartheta}_n^j\}$ satisfies*

$$\mathbb{E}N_j(\omega)^{2+\alpha} < \infty, \qquad \forall j \in \{1, \ldots, J\},\ 0 < \alpha < \beta, \tag{3.5}$$

where β appears in the definition of SWAC condition in (3.4). ■

Note that Theorem 6 shows that SWAC condition implies the $2 + \alpha$ moment property which appears in A4 in Section 2.

Proof. For every $j \in \{1, \ldots, J\}$, under A5 and A6, the sequence of the ML estimates $\{\hat{\vartheta}_n^j; n \in \mathbb{Z}_{>0}\}$ is strongly consistent, by virtue of Theorem, 5. In addition under A7, we have that for all $\omega \in \bigcap_{j=1}^{J} \Omega_o^j \subseteq \Omega$, where for all $j \in \{1, \ldots, J\}$, Ω_o^j is

such that $\mathbb{P}_{\vartheta_j^*}(\Omega_o^j) = 1$,

$$\begin{aligned}
\mathbb{E}N_j(\omega)^{2+\alpha} &= \sum_{n=1}^{\infty} n^{2+\alpha}\mathbb{P}(N_j(\omega) = n) \\
&\leq \sum_{n=1}^{\infty} n^{2+\alpha}\mathbb{P}(\hat{\vartheta}_{n-1}^j \neq \vartheta_j^*, \hat{\vartheta}_n^j \leq \vartheta_j^*) \qquad \text{(by A7)} \qquad (3.6) \\
&\leq \sum_{n=1}^{\infty} n^{2+\alpha}\tfrac{C}{n^{3+\beta}} < \infty, \quad 0 < \alpha < \beta.
\end{aligned}$$

□

CHAPTER 4
Maximum Likelihood Estimation in Gaussian ARMA systems

Consider the stationary reward process given by the following system:

$$S: \quad \begin{aligned} x_{n+1} &= \vartheta x_n + w_n \\ y_n &= x_n \end{aligned} \qquad \forall n \in \mathbb{Z}_{\geq 0} \tag{4.1}$$

where $x_n, y_n, w_n \in \mathbb{R}^1$ for $n \in \mathbb{Z}_{\geq 0}$, and where w is I.I.D. $\mathcal{N}(0, \sigma^2)$ noise process.

Evidently the system is in ARMA form. Assume that $|\vartheta| < 1$. Consider also the following assumptions:

INP1: The process ω defined on $(\Omega, \mathcal{A}), \mathbb{P}$ is a non-zero, stationary process with $\mathbb{E}(\omega_k, \omega_j^T) = \Sigma\delta_{k,j}$, for all $k, j \in \mathbb{Z}$, with $\Sigma \in \mathcal{P}$ where $\mathcal{P}$ denotes the set of all $(p \times p)$ strictly positive, symmetric matrices.

INP2: The initial conditions for (4.1) and the (full rank) orthogonal process $\{\omega_n; n \geq \mathbb{Z}_{\geq 0}\}$ are jointly Gaussian, mutually orthogonal, and have zero mean.

Under the aforementioned assumptions, INP1 and INP2, the (negative) logarithmic likelihood function 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 [4, chapter 7]:

$$-\log f(y^n; \vartheta) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\left(\frac{\sigma^{2n}}{1-\vartheta^2}\right) + \frac{1}{2}y_1^2\left(\frac{\sigma^2}{1-\vartheta^2}\right)^{-1} + \frac{1}{2}\sum_{i=2}^{n}(y_i - y_{i|i-1})^2\sigma^{-2} \tag{4.2}$$

where $y^n \triangleq y_1, \ldots, y_n$. Equation (4.2) is parametrized by both ϑ and σ.

In the sequel, we assume an ARMA process generated by (4.1), whose defining parameters are not known. However, it is known that they lie in a finite parameter set Θ. For simplicity, we assume that the parameter space Θ contains only two alternatives; $\theta^* \triangleq (\vartheta^*, \sigma^*)$ and $\theta \triangleq (\vartheta, \sigma)$ ($|\vartheta^*|, |\vartheta| < 1$), where $(\bullet)^*$ denotes the true parameter under which the observations process $\{y_n; n \in \mathbb{Z}_{\geq 0}\}$ is generated.

Thereafter, we construct two likelihood function candidates, one for each parameter in Θ to investigate the (strong) consistency of the ML estimates of the unknown parameter. To do this, we employ observations $\{y_n; n \in \mathbb{Z}_{\geq 0}\}$ from the system, as well as the decomposition property (in terms of the prediction error process) of the likelihood function candidates.

We finally present a simulation result for a simple MAB problem with (dependent) reward processes generated by (4.1), showing the behaviour of the regret under the proposed policy Φ^g.

4.1 Properties of the MLE in Gaussian ARMA system

Consider a set of ARMA processes generated by (4.1) when parametrized by $\Theta = \{\theta^*, \theta\}$.

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

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

Similarly, the prediction error process under the wrong parameter θ is:

$$\begin{aligned} e_n &= y_n - \mathbb{E}\{y_n \Big| y^{n-1}\} \\ &= y_n - \vartheta y_{n-1} + \mathbb{E}\{w_{n-1}\} \\ &= \vartheta^* y_{n-1} + w_{n-1} - \vartheta y_{n-1} \\ &= (\vartheta^* - \vartheta) y_{n-1} + w_{n-1}. \end{aligned}$$

Furthermore,

$$\begin{aligned} e_n &= (\vartheta^* - \vartheta) y_{n-1} + w_{n-1} \\ &= w_{n-1} + (\vartheta^* - \vartheta) \sum_{j=1}^{n} \vartheta^{*j-1} w_{n-1-j} \\ &= \nu_n + (\vartheta^* - \vartheta) \sum_{j=1}^{n} \vartheta^{*j-1} \nu_{n-j} \end{aligned} \tag{4.3}$$

The prediction error process of the system under the true parameter θ^* is IID and is called *innovations process*. On the other hand, the prediction error process of the system under the wrong parameter θ is in general a dependent process, and for that reason it is called *pseudo-innovations process*. The pseudo-innovations process is a regression on all past true innovations.

To claim that the estimate $\hat{\theta}_n$ is consistent one needs to verify conditions A5 and A6.

4.1.1 Verification of A5

Assuming that $\theta^* \neq \theta$, then A5 immediately holds.

4.1.2 Verification of A6

To verify A6, one needs to show that for all $\varepsilon > 0$, there is $\alpha(\varepsilon) > 1$,

$$\mathbb{P}\left\{0 \leq \frac{f(y_n|y^{n-1};\theta)}{f(y_n|y^{n-1};\theta^*)} < \alpha(\varepsilon), \forall n > |\Theta|\right\} < \varepsilon. \tag{4.4}$$

But

$$\begin{aligned} & \tfrac{f(y_n|y^{n-1};\theta)}{f(y_n|y^{n-1};\theta^*)} && < \alpha \\ \Rightarrow \quad & \tfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\tfrac{1}{2}\tfrac{e_n^2}{\sigma^2}\right) && < \alpha\tfrac{1}{\sqrt{2\pi}\sigma^*}\exp\left(-\tfrac{1}{2}\tfrac{\nu_n^2}{\sigma^{*2}}\right). \end{aligned} \tag{4.5}$$

Taking the logarithm in both sides of (4.5) one has

$$\begin{aligned} & -\log\sigma^2 - \tfrac{1}{2}\tfrac{e_n^2}{\sigma^2} && < \log\alpha - \log\sigma^{*2} - \tfrac{1}{2}\tfrac{\nu_n^2}{\sigma^{*2}} \\ \Rightarrow \quad & \log\tfrac{\sigma^{*2}}{\sigma^2} - \log\alpha && < \tfrac{1}{2}\left(\tfrac{e_n^2}{\sigma^2} - \tfrac{\nu_n^2}{\sigma^{*2}}\right) \\ \Rightarrow \quad & \log\tfrac{\sigma^{*2}}{\sigma^2} - \log\alpha && < \tfrac{1}{2}\left(\frac{(\nu_n + (\vartheta^* - \vartheta)\sum_{j=1}^{n}\vartheta^{*j-1}\nu_{n-j})^2}{\sigma^2} - \tfrac{\nu_n^2}{\sigma^{*2}}\right) \quad \text{(by (4.3))}. \end{aligned}$$

Returning to (4.4) we have

$$\mathbb{P}\left\{\log\frac{\sigma^{*2}}{\sigma^2} - \log\alpha < \frac{1}{2}\left(\frac{(\nu_n + (\vartheta^* - \vartheta)\sum_{j=1}^{n}\vartheta^{*j-1}\nu_{n-j})^2}{\sigma^2} - \frac{\nu_n^2}{\sigma^{*2}}\right), \forall n > |\Theta|\right\} < \varepsilon \tag{4.6}$$

The event inside the probability measure in (4.6) is in general hard to be evaluated since it involves summation of squares of the 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 impose the following conjecture.

Conjecture 1. *For the set of likelihood functions specified by the parameter set Θ, condition A6 is satisfied.* ■

With conditions A5 and A6 being verified (with the support of Conjecture 1), we have that the maximum likelihood estimate $\hat{\theta}_n$ of the true parameter θ^*

is strongly consistent by virtue of Theorem 5. This in turn implies that A3 (see chapter 2) also holds.

Next, we investigate whether A4 holds (i.e. $\mathbb{E}\left(N^{2+\alpha}(\omega)\right) < \infty$, for some $\alpha > 0$). Since we know that the ML estimate is consistent, it is sufficient to verify A7 (SWAC), by virtue of Theorem 6.

4.1.3 Verification of A7

Firstly, consider the event $\{\hat{\theta}_n \neq \theta^*\}$ for which we have

$$\begin{aligned} \{\hat{\theta}_n \neq \theta^*\} = & \{f(y^n;\theta) > f(y^n;\theta^*)\} \\ = & \{-\log f(y^n;\theta) < -\log f(y^n;\theta^*)\} \end{aligned} \tag{4.7}$$

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

$$\begin{aligned} \{\hat{\theta}_n \neq \theta^*\} = & \{\log(\tfrac{\sigma^{2n}}{1-\vartheta^2}) + y_1^2(\tfrac{\sigma^2}{1-\vartheta^2})^{-1} + \sum_{i=2}^{n}\tfrac{e_i^2}{\sigma^2} < \log(\tfrac{\sigma^{*2n}}{1-\vartheta^{*2}}) + y_1^2(\tfrac{\sigma^{*2}}{1-\vartheta^{*2}})^{-1} + \sum_{i=2}^{n}\tfrac{\nu_i^2}{\sigma^{*2}}\} \\ = & \{n\log(\tfrac{\sigma^2}{\sigma^{*2}}) + \log(\tfrac{1-\vartheta^{*2}}{1-\vartheta^2}) + y_1^2(\tfrac{\sigma^2}{1-\vartheta^2})^{-1} - y_1^2(\tfrac{\sigma^{*2}}{1-\vartheta^{*2}})^{-1} + \sum_{i=2}^{n}\tfrac{e_i^2}{\sigma^2} < \sum_{i=2}^{n}\tfrac{\nu_i^2}{\sigma^{*2}}\} \end{aligned} \tag{4.8}$$

Thereafter, consider the event $\{\hat{\theta}_{n+1} = \theta^*\}$ for which, similarly to (4.8) we have

$$\begin{aligned} \left\{\hat{\theta}_{n+1} = \theta^*\right\} = \Bigg\{ & (n+1)\log(\frac{\sigma^2}{\sigma^{*2}}) + \log(\frac{1-\vartheta^{*2}}{1-\vartheta^2}) + y_1^2(\frac{\sigma^2}{1-\vartheta^2})^{-1} \\ & - y_1^2(\frac{\sigma^{*2}}{1-\vartheta^{*2}})^{-1} + \sum_{i=2}^{n}\frac{e_i^2}{\sigma^2} + \frac{e_{n+1}^2}{\sigma^2} > \sum_{i=2}^{n}\frac{\nu_i^2}{\sigma^{*2}} + \frac{\nu_{n+1}^2}{\sigma^{*2}} \Bigg\} \end{aligned} \tag{4.9}$$

To construct the joint event $\{\hat{\theta}_n \neq \theta^*, \hat{\theta}_{n+1} = \theta^*\}$, we substitute (4.8) into (4.9) to get

$$\begin{aligned}\{\hat{\theta}_n \neq \theta^*, \hat{\theta}_{n+1} = \theta^*\} = & \ \{\tfrac{e_{n+1}^2}{\sigma^2} - \tfrac{\nu_{n+1}^2}{\sigma^{*2}} > \log\tfrac{\sigma^{*2}}{\sigma^2} \cap \{\hat{\theta}_n \neq \theta^*\}\} \\ = & \ \{\tfrac{e_{n+1}^2}{\sigma^2} - \tfrac{\nu_{n+1}^2}{\sigma^{*2}} > \log\tfrac{\sigma^{*2}}{\sigma^2} \cap f(y^n;\theta) > f(y^n;\theta^*)\} \\ = & \ \Big\{\{\tfrac{e_{n+1}^2}{\sigma^2} - \tfrac{\nu_{n+1}^2}{\sigma^{*2}} > \log\tfrac{\sigma^{*2}}{\sigma^2} \ \bigcap \\ & \ \{n\log(\tfrac{\sigma^{*2}}{\sigma^2}) + \log(\tfrac{1-\vartheta^2}{1-\vartheta^{*2}}) - y_1^2(\tfrac{\sigma^2}{1-\vartheta^2})^{-1} + y_1^2(\tfrac{\sigma^{*2}}{1-\vartheta^{*2}})^{-1} + \sum_{i=2}^{n}\left(\tfrac{\nu_i^2}{\sigma^{*2}} - \tfrac{e_i^2}{\sigma^2}\right) > 0\}\Big\}\end{aligned} \tag{4.10}$$

where the event in the right hand side of the intersection is (4.8) properly rearranged.

As for the event inside the probability measure in (4.6), the event described by (4.10) involves a linear combination of χ^2 random variables whose probability density function is not known. This leads us to impose the following conjecture.

Conjecture 2. *Let*

$$M_n(\omega) \triangleq \left\{n\log(\frac{\sigma^{*2}}{\sigma^2}) + \log(\frac{1-\vartheta^2}{1-\vartheta^{*2}}) - y_1^2(\frac{\sigma^2}{1-\vartheta^2})^{-1} + y_1^2(\frac{\sigma^{*2}}{1-\vartheta^{*2}})^{-1} + \sum_{i=2}^{n}\left(\frac{\nu_i^2}{\sigma^{*2}} - \frac{e_i^2}{\sigma^2}\right) > 0\right\} \tag{4.11}$$

Then,

$$\mathbb{P}\left\{\frac{e_{n+1}^2}{\sigma^2} - \frac{\nu_{n+1}^2}{\sigma^{*2}} > \log\frac{\sigma^{*2}}{\sigma^2} \cap M_n(\omega)\right\} < \frac{C}{n^{3+\beta}}, \tag{4.12}$$

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

The conjecture is supported by the fact that the event $M_n(\omega)$ is expected to be decreasing as n increases. This is because the summation of the pseudo-innovations appears with a negative sign on the left hand side of (4.11). Thus the

probability of any intersection with that event will also be decreasing. In addition, since χ^2 distribution is exponential, it is plausible to assume that the rate of decrease of the probability in (4.12) is faster than $\frac{C}{n^{3+\beta}}$, for some $C \in \mathbb{Z}_{>0}$ and $\beta \in \mathbb{R}_{>0}$.

4.2 Simulations

Consider a 2-bandit system whose observations processes are generated by ARMA systems described by (4.1) with parameter spaces $\Theta_j = \{\theta_j^i\}; \theta_j^i = (\vartheta_j^i, \sigma_j^i)$, where $j \in \{1, 2\}$ denotes the machine index, and $i \in \{a, b\}$ denotes the parameter of machine j. With no loss of generality, assume that $\theta_1^* = \theta_1^a$ and $\theta_2^* = \theta_2^b$, where θ_j^* denotes the true parameter of machine j.

Consider also the following scenario: at each step T the player chooses to observe a sample from machine $j \in \{1, 2\}$ and receives a cost equal to the minimum one step prediction error of the next observation $y^j_{n_T^j}$ given the past observations $y_1^j, \ldots, y^j_{n_T^j - 1}$, where n_T^j denotes the local time of machine j.

For linear and Gaussian systems, the minimum one step prediction error process is equal to the (true) innovations process. This allows us to define the regret at time T as follows.

$$\begin{aligned} R_T(\omega, \phi) &= \sum_{i=1}^{T} -(\min_{j \in \{1,2\}} \mathbb{E}(\nu^{j\,2}) - \nu^{j_i}_{n_T^{j_i}}{}^2) \\ &= \sum_{i=1}^{T} -(\min_{j \in \{1,2\}} \sigma_j^{*2} - \nu^{j_i}_{n_T^{j_i}}{}^2) \end{aligned} \tag{4.13}$$

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

The goal of the player is to minimize the *rate of growth* of the expected regret $\mathbb{E}\{R_T(\omega,\phi\}$ as $T \to \infty$. This corresponds to fairly realistic cases where while one wants to "learn" the unknown system in terms of identifying the true parameter of each machine, he/she wants to hit a target (physical or financial etc) with the greatest accuracy based on his/her so far knowledge about the system.

In light of (1.5), the expected total regret at time T is given by

$$\mathbb{E}\{R_T(\omega,\phi)\} = \sum_{j=1}^{2} -(\min_{j\in\{1,2\}} \sigma_j^{*2} - \sigma_j^{*2})\mathbb{E}(n_T^j) \tag{4.14}$$

where $\sigma_j^{*2} = \mathbb{E}\{{\nu^j}^2\}$. Equation (4.14) reveals that the expected regret is proportional to the difference between the innovations process variance of each machine j and the innovations process variance of the best machine j^*.

The index functions in this case can be defined as

$$g^j_{T,n^j_T} = \frac{2}{\hat{\sigma}_j} + \frac{T}{Cn_T^j},\ j \in \{1,2\} \tag{4.15}$$

where $\hat{\sigma}_j$ is the ML estimate of the innovations process variance of machine j. This implies that the first term in the summation of (4.17) will be bigger for the machine j^* with the smallest one step prediction error variance compared to the same term of other machines. This in turn implies that machine j^* will be chosen more often than other machines, as long as the estimate $\hat{\sigma}_{j^*}$ is close to the true value $\sigma^*_{j^*}$.

For the computation of $\hat{\sigma}_T^j$ at stage T, we proceed as follows. Firstly, we compute the maximized likelihood ratio (MLR) given by

$$x^j(T) = \frac{f_\theta(y_1^j, \ldots, y_n^j)}{f_{\theta^*}(y_1^j, \ldots, y_n^j)}. \tag{4.16}$$

Then, the ML estimate $\hat{\sigma}_T^j$ is given by

$$\hat{\sigma}_T^j = \begin{cases} \sigma_j, & \text{if } x_T \geq 1 \\ \sigma_j^*, & \text{if } x_T < 1 \end{cases} \tag{4.17}$$

Remark 1. *We note that the MAB model considered for the simulations does not fit with the model described in chapter 1. This is because in the MAB model defined in chapter 1, the reward yielded from machine $j \in \{1, \ldots, J\}$ at instant T depends only on the observation $y^j_{n_T^j}$ made at the same instant, while in the MAB model considered for the simulations, the reward yielded from machine j ($\nu^j_{n_T^j}$) depends on the past observations $y_1^j, \ldots, y^j_{n_T^j - 1}$ as well.*

However, we can make the later model fit with the former model by using a simple transformation of the observations. Specifically, we can assume that whenever the player plays machine j, he/she observes a vector $\left(y^j_{n_T^j}, \nu^j_{n_T^j}\right)^T$. By using this transformation, the scenario described earlier remains valid in terms of estimation. This is because $\nu^j_{n_T^j}$ is a function of the past and present observation $\left(y_1^j, \ldots, y^j_{n_T^j}\right)$, and thus employing $\nu^j_{n_T^j}$ does not improve the parameter estimation of machine j. ▶

In the sequel, we initially investigate the validity of A6 (and thus of Conjecture 1) under the adopted scenario described above. A simulation of 10000

(independent) experiments had been carried out in which $\theta^* = (0.145, 8)$, and $\theta = (0.09, 10)$.

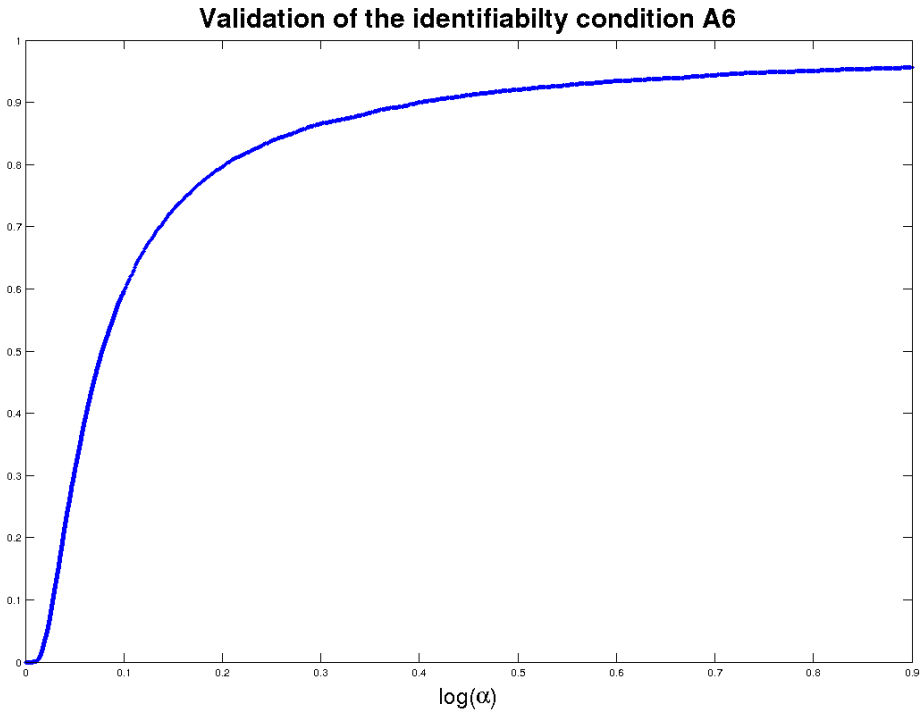

Figure 4–1: The values of $\mathbb{P}\left\{\log\frac{\sigma^{*2}}{\sigma^2} - \log\alpha < \frac{1}{2}\left(\frac{(\nu_n + (\vartheta^* - \vartheta)\sum_{j=1}^{n}\vartheta^{*j-1}\nu_{n-j})^2}{\sigma^2} - \frac{\nu_n^2}{\sigma^{*2}}\right), \forall n > |\Theta|\right\}$ (vertical axis) over a range of values for $\alpha > 1$ (horizontal axis).

In figure 4–1 the vertical axis represents the values of $\mathbb{P}\left\{\log\frac{\sigma^{*2}}{\sigma^2} - \log\alpha < \frac{1}{2}\left(\frac{(\nu_n + (\vartheta^* - \vartheta)\sum_{j=1}^{n}\vartheta^{*j-1}\nu_{n-j})^2}{\sigma^2} - \frac{\nu_n^2}{\sigma^{*2}}\right), \forall n > |\Theta|\right\}$, and the horizontal axis a range of possible values for $\alpha > 1$. Evidently, for any value $\varepsilon > 0$ we can always find a some $\alpha > 1$ such that the identifiability condition holds.

Thereafter, we investigate the lock-on time of $\hat{\theta}^j$ to θ_j^*.

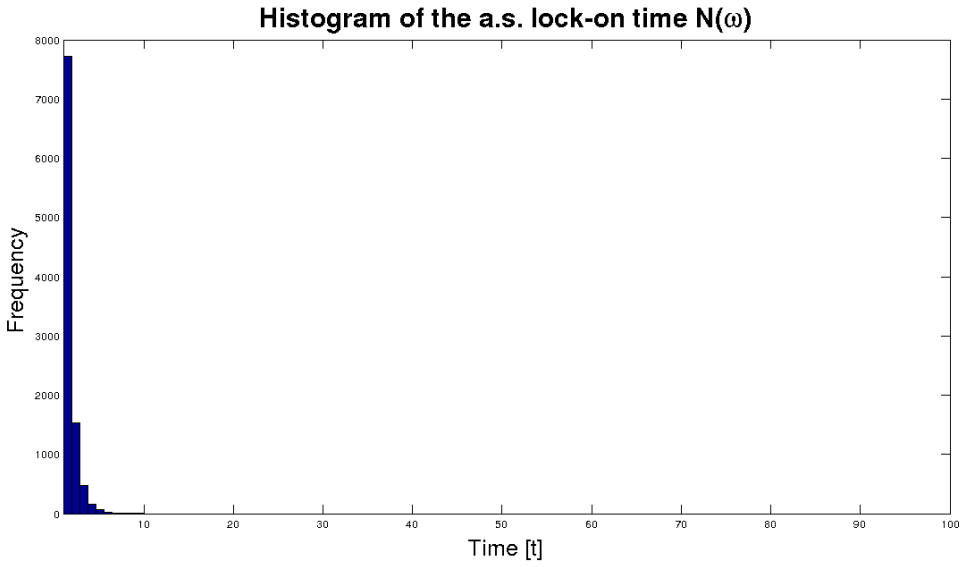

Figure 4–2: Histogram of the a.s. lock-on time $N(\omega)$ of the parameter estimate to the true parameter, for a sample of 10000 independent random experiments. The figure suggests that the probability that of lock-on decreases exponentially in time.

In figure 4–2 the horizontal axis represents the range of the lock-on time, and the vertical distribution of the lock-on time among the 10000 independent random experiments. We observe that in most of the experiments, the lock-on time occurs quite early. This simulation result supports Conjecture 2 since it suggests that the distribution of the lock-on time drops of exponentially in time.

Finally, we investigate the behaviour of the expected regret achieved by the proposed policy Φ^g introduced in section 2.2, considering the MAB problem described by (4.12).The constant C employed by the index functions g^j was arbitrarily set to 1000 for all arms $j \in \{1, 2\}$. A simulation of 100 (independent) random experiments had been carried out. The horizon was set to $T = 10000$ for all random experiments.

The parameter values considered for the simulation are as follows:

$$\text{Machine 1}: \quad \theta_1^1 = (0.145, 8), \quad \theta_1^2 = (0.09, 10)$$
$$\text{Machine 2}: \quad \theta_2^1 = (0.2, 5), \quad \theta_2^2 = (0.19, 15)$$

According to these values, the best bandit machine in the system, is machine 1, since $\sigma_1^* = 8 < \sigma_2^* = 15$. From equation (1.5), it is implied that the regret is proportional to $\mathbb{E}\{n_T^2\}$, with constant equal to $(\sigma_2^* - \sigma_1^*) = 3$.

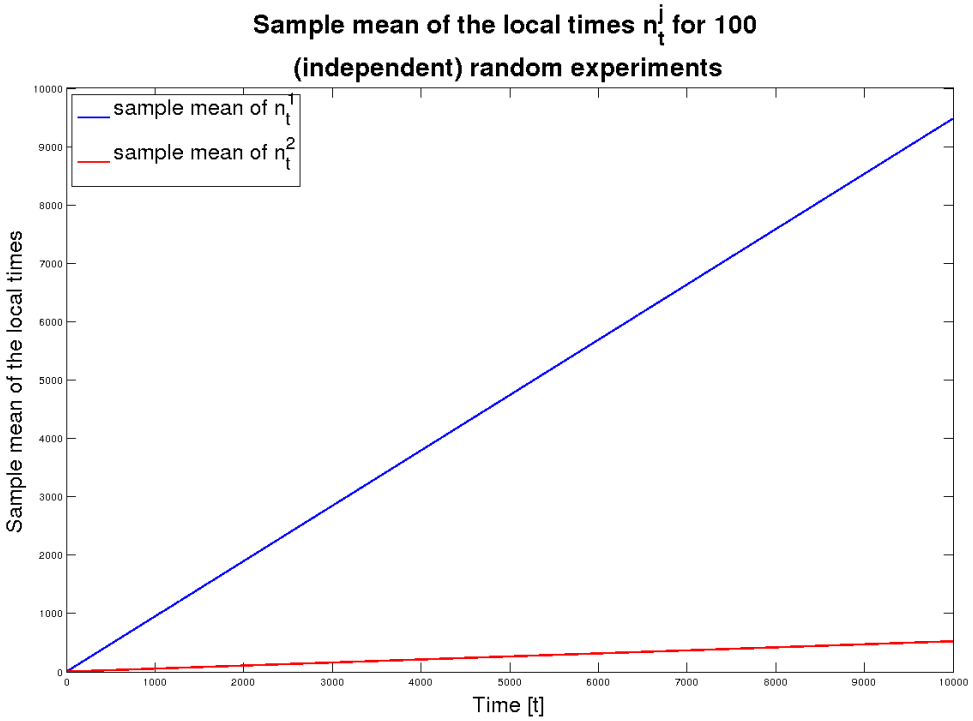

Figure 4–3: Simulation of 100 random experiments, where $T = 10000$ (vertical axis length). n_T^1: the local time of machine 1, n_T^2: the local time of machine 2.

The sample mean of the local time of each machine are shown in figure 4–3. The blue plot represents the sample mean of the local time of machine 1, $\bar{n}_t^1$, and the red plot represent the sample mean of the local time of machine 2, $\bar{n}_t^2$, where

$$\bar{n}_t^j = \frac{1}{100} \sum_{i=1}^{100} n_t^j$$

In figure 4–3 we observe that the local time of machine 2, i.e. $\bar{n}_t^2$, is increasing uniformly linearly in time. This is a property of the proposed allocation rule Φ^g, inherited by UCB1. It suggests that $\mathbb{E}\{n_T^2\} \in \mathbf{O}(T)$ as $T \to \infty$, or equivalently that $\mathbb{E}\{R_T(\omega, \Phi^g)\} \in \mathbf{O}(T)$ as $T \to \infty$. This result was expected, since the index function g^2, for fixed n_T^2, is increasing linearly in T, forcing the allocation rule Φ^g to choose it in a linear rate.

Furthermore, $\mathbb{E}(R_T(\omega, \Phi^g)) \in \mathbf{O}(T)$ implies that $\mathbb{E}(R_T(\omega, \Phi^g)) \in \mathbf{o}(T^{1+\beta})$, for some $\beta > 0$. That means

$$\lim T \to \infty \frac{\mathbb{E}(R_T(\omega, \Phi^g))}{T^{1+\beta}} = 0, \quad \beta > 0,$$

which verifies the result of Theorem 3(b).

CHAPTER 5
Conclusion

5.1 Summary

In this thesis, we consider the stochastic multi-armed bandit (MAB) problem in finite parameter sets. That is, for each machine there is an unknown parameter which lies in a known and finite, albeit arbitrarily large, parameter set.

We propose an index-type policy Φ^g, whose index functions are modified versions of the indexes employed by UCB1 in [2]. In particular, the index function of each machine, employs a strongly consistent estimator of the unknown parameter corresponding to that machine.

To this end, we consider the maximum likelihood estimator which had been shown in the literature to be strongly consistent for parameter estimation in finite sets. In addition, we introduce the summable wrong and corrected (SWAC) condition which implies the $2 + \alpha$ moment, and hence the first and second moments, of the random lock-on instant $N(\omega)$ are finite for $0 < \alpha < \beta$, where β appears in the definition of SAWC condition.

Under this framework, we show that Φ^g is UCB-index type, and achieves a super-linear regret. While the regret of Φ^g does not achieve the optimal lower bound of the regret of uniformly good policies which was given in [12], it is attractive because of its generality, since it can be applied in MAB problems with dependent reward processes across time.

5.2 Open Problems and Future Research

As it had already been mentioned, the regret achieved by the proposed policy Φ^g, which is not necessarily considered as uniformly good, is suboptimal compared to the regret achieved by the class of uniformly good policies described in [12].

A considerable improvement of the proposed policy Φ^g in future work, will be to re-design the switching rule employed by the index functions g taking into consideration the finite, although not uniform, lock-on time of the consistent parameter estimate to the true parameter value for each arm.

This modification would provide us with an improved version of Φ^g which locks-on choosing the best machine in finite time and hence achieves finite expected regret.

References

[1] Rajeev Agrawal. Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078, 1995.

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[3] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

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

[5] Peter E Caines. *Linear stochastic systems*. John Wiley & Sons, Inc., 1987.

[6] Joseph L Doob. *Stochastic processes*, volume 101. New York Wiley, 1953.

[7] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478, 2012.

[8] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. Wiley Online Library, 1989.

[9] Lifeng Lai, Hesham El Gamal, Hai Jiang, and H Vincent Poor. Cognitive medium access: Exploration, exploitation, and competition. *Mobile Computing, IEEE Transactions on*, 10(2):239–253, 2011.

[10] Tze Leung Lai and Herbert Robbins. Asymptotically optimal allocation of treatments in sequential experiments, 1984.

[11] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[12] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

[13] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[14] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, pages 595–601, 1949.

[15] John White. *Bandit algorithms for website optimization.* " O'Reilly Media, Inc.", 2012.

Abbreviations

IID Independent and identically distributed

MAB Malti-armed Bandit

MLE Maximum-likelihood estimate

MLR Maximum-likelihood ratio

SWAC Summable Wrong and Corrected Condition