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

McGill University

May 6, 2015

The Multi-Armed Bandit (MAB) Problem

- At each step a Decision Maker (DM) faces the following sequential allocation problem:
  - must allocate a unit resource between several competing actions/projects.
  - obtains a random reward with unkown probability distribution.
- The DM must design a policy to maximize the cumulative expected reward asymptotically in time.

Stylized model to understand exploration-exploitation trade-off

- Imagined slot machine with multiple arms.
- The gambler must choose one arm to pull at each time instant.
- He/she wins a random reward following some unknown probability distribution.
- His/her objective is to choose a policy to maximize the cumulative expected reward over the long term.

Real examples

- In Internet routing:
  - Sequential transmission of packets between a source and a destination.
  - The DM must choose one route among several alternatives.
  - Reward = transmission time or transmission cost of the packet.
- In cognitive radio communications:
  - The DM must choose which channel to use in different time slots among several alternatives.
  - Reward = Number of bits sent at each slot
- In advertisement placement:
  - The DM must choose which advertisement to show to the next visitor of a web-page among a finite set of alternatives.
  - Reward = Number of click-outs.

Literature Overview

i.i.d. rewards

- Lai and Robbins (1985) constructed a policy that achieves the asymptotically optimal regret of $O(\log T)$.
- Agrawal (1995) constructed index type policies that depend on the sample mean of the reward process, and they achieve asymptotically optimal regret of $O(\log T)$.
- Auer et. al. (2002), constructed an index type policy, called UCB1, which whose regret is $O(\log T)$ uniformly in time.

Markov rewards

- Tekin et. al. (2010) proposed an index-based policy that achieves an asymptotically optimal regret of $O(\log T)$.

The Reward Process and the Regret

Reward processes for each machine	$\{Y_n^k\}_{n=1}^{\infty}$; $k = 1, \dots, K$, defined on a common measurable space $(\Omega, \mathcal{A})$.
Set of probability measures	$\{\mathbb{P}_\theta^k; \theta \in \Theta_k\}$, where Θ_k is a known finite set, for which:

- f_θ^k denotes probability density,
- μ_θ^k denotes mean.

Best machine

$$k^* \triangleq \underset{k \in \{1,\dots,K\}}{\operatorname{argmax}} \{\mu_{\theta_k^*}^k\}.$$

- true parameter for machine k is denoted θ_k^*.

Allocation policy and Expected Regret

Allocation policy

A mapping $\phi_t : \mathbb{R}^{t-1} \to \{1, \dots, K\}$ that indicates the arm to be selected at the instant t

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

where $Z_1, \dots, Z_{t-1}$ denote the rewards gained up until $t-1$.

Expected Regret

$$R_T(\phi) = \sum_{k=1}^{K} \left(\mu^{k^*}_{\theta^*_{k^*}} - \mu^{k}_{\theta^*_k} \right) \mathbb{E}(n_T^k),$$

where

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

The Multi-Armed Bandit Problem

Definition

The MAB problem is to define a policy

$$\phi = \{\phi_t;\ t \in \mathbb{Z}_{>0}\}$$

in order to minimize the rate of growth of

$$R_T(\phi) \text{ as } T \to \infty.$$

Index policies and Upper Confidence Bounds

Index policy ϕ^g

A policy that depends on a set g of indices for each arm and chooses the arm with the highest index at each time.

Upper Confidence Bounds (UCB) [Agrawal (1985)]

A set g of indices is a UCB, if it satisfies the following conditions:

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 $z < \mu_t^k$,

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

The Proposed Allocation (UCB) policy

Consider a set of index functions g with

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

where $t \in \mathbb{Z}_{>0}$, $n \triangleq n_t^k \in \{1, \ldots, t\}$, $C \in \mathbb{R}$ and $k \in \{1, \ldots, K\}$, and $\hat{\mu}_n^k$ is the maximum likelihood estimate of the mean of Y^k.

Then,

- if $t \leq K$: ϕ^g samples from each process Y^k once
- if $t > K$: ϕ^g samples from Y^{u_t}, where

$$u_t = \operatorname{argmax}\{g_{t,n_t^k}^{k};\ k \in \{1, \ldots, K\}\}$$

The main results

Theorem

Under suitable technical assumptions, the regret of the proposed policy satisfies

$$R_T(\phi^g) = \mathbf{o}(T^{1+\delta})$$

for some $\delta > 0$.

- The proposed index policy works when the rewards processes are ARMA processes with unknown means and variance.

Preliminaries on MLE

Definition

A sequence of estimates $\{\hat{\theta}_n\}_{n=1}^{\infty}$ is called a maximum likelihood estimate if

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

Definition

$\{\hat{\theta}_n\}_{n=1}^{\infty}$ is called a (strongly) consistent estimator if $\hat{\theta}_n \neq \theta^*$ finitely often, $\mathbb{P}_{\theta^*}$ *a.s.*

Assumption 1

Let $\mathbb{P}_{\theta,n}$ denote the restriction of $\mathbb{P}_\theta$ to the σ-field $\mathcal{A}_n$, $n \geq 0$. Then, for all $\theta \in \Theta$ and $n \geq 0$, $\mathbb{P}_{\theta,n}$ 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

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

where $y^n \triangleq (y_1, \ldots, y_n)$.
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 (PEC, 1975)

Under Assumptions 1 and 2, the sequence of the maximum likelihood estimates is (strongly) consistent.

Assumptions on the model

Assumption 3

For every arm k, there is a consistent estimator $\hat{\vartheta}^k = \{\hat{\vartheta}^k_1, \hat{\vartheta}^k_2, \ldots\}$.

Assumption 4 (The summable Wrong and Corrected Condition (SWAC))

For all machines $k \in \{1, \ldots, K\}$, the sequence of estimates $\hat{\theta}^k_1, \ldots, \hat{\theta}^k_n, \ldots$ satisfies the following condition:

$$\mathbb{P}^k_{\theta^*_k}(\hat{\theta}^k_{n-1} \neq \theta^*_k, \hat{\theta}^k_m = \theta^*_k, \ \forall m \geq n) < \frac{C}{n^{3+\beta}},$$

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

The Lock-on time

Definition

For a consistent sequence of estimates $\hat{\theta}_1^k, \ldots, \hat{\theta}_n^k, \ldots$, the *lock-on time* refers to the least N such that for all $n \geq N$, $\hat{\theta}_n = \theta^*$, $\mathbb{P}_{\theta^*}$ a.s.

Lemma 1

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

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

where β appears in Assumption 4.

Performance of ϕ^g

Theorem 2

If Assumptions 3 and 4 hold, then for each $k \in \{1, \ldots, K\}$, the proposed index function

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

is an Upper Confidence Bound (UCB)

Theorem 3

If Assumptions 3 and 4 hold, then the regret of the proposed policy ϕ^g satisfies

$$R_T(\phi^g) = \mathbf{o}(T^{1+\delta}),$$

for some $\delta > 0$.

A MAB Problem for ARMA Processes

Consider a bandit system with reward process generated by the following ARMA process

$$S: \quad \begin{aligned} x_{n+1}^k &= \lambda_k x_n^k + w_n^k \\ y_n^k &= x_n^k \end{aligned} \qquad \forall n \in \mathbb{Z}_{\geq 0}, k \in \{1,2\}$$

where $x_n^k, y_n^k, w_n^k \in \mathbb{R}$, $n \in \mathbb{Z}_{\geq 0}$, and w^k is i.i.d. $\sim \mathcal{N}(0, {\sigma_k}^2) \perp\!\!\!\perp x_0^k$.

Assumptions:

- The parameter space of the system contains two alternatives:

$$\Theta_k = \{\theta_k^*, \theta_k\}; \ \theta_k \triangleq (\lambda_k, \sigma_k), \ k \in \{1,2\}.$$

- For each system $|\lambda| < 1$ and each process y_n^k is stationary.

A MAB Problem for ARMA Processes

Problem Description

At each step t,

- the player chooses to observe a sample from machine $k \in \{1, 2\}$
- pays a cost υ_t^k equal to the squared minimum one step prediction error of the next observation $y_{n_t^k}^k$ given the past observations $y_1^k, \ldots, y_{n_t^k - 1}^k$.

The Expected Regret

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

where u_i denotes the arm that is needed to be chosen at time i, specified by the proposed index policy ϕ^g.

Preliminary results for ARMA Processes

The negative logarithmic likelihood function of the reward process can be described as follows:

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

where

- $y_{i|i-1} \triangleq \mathbb{E}(y_i|y^{i-i}) = \lambda y_{i-1}$, and
- $y_i - y_{i|i-1}$ is the prediction error process.

Prediction error process the true parameter under θ^*

$$\nu_n = y_n - y_{i|i-1} = w_{n-1}, \quad w_{n-1} \sim \mathcal{N}(0, {\sigma^*}^2).$$

The prediction error process under the incorrect parameter θ

$$e_n = y_n - y_{i|i-1} = \nu_n + (\lambda^* - \lambda) \sum_{j=1}^{n} {\lambda^*}^{j-1} \nu_{n-j},$$

Remarks:

- ν_n is called the innovations process of y_n, and it is i.i.d.
- e_n is called the pseudo-innovations process of y_n, and it is a dependent process.

Verification of Assumptions 1,2, and 4

Concerning Assumption 1

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

Concerning Assumption 2

We make the conjecture that for the set of likelihood functions specified by the parameter set Θ, Assumption 2 is satisfied.

Verification of Assumptions 1,2, and 4

Assumption 4

- Consider each machine separately.
- Define

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

- Let $V_n = \sum_{i=2}^{n} \frac{\nu_i^2}{{\sigma^*}^2}$.
- Define

$$E_n \triangleq \{\hat{\theta}_n \neq \theta^*, \hat{\theta}_m = \theta^*,\ \forall m \geq n\}$$
$$= \left\{\sum_{i=2}^{n}\frac{\nu_i^2}{{\sigma^*}^2} > A_n\right\} \cap \{A_{n+1} \geq V_{n+1}\} \cap \{A_{n+2} \geq V_{n+2}\} \cap \dots,$$

Verification of Assumptions 1,2, and 4

Assumption 4

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

$$\mathbb{P}\{E_n\} < \frac{a}{n^{3+\beta}}.$$

and hence Assumption 4 is satisfied.

The index functions

Definition

$$g^k_{T,n^k_T} = \frac{2}{\hat{\sigma}^2_k} + \frac{T}{Cn^k_T},\ \ k \in \{1, 2\}$$

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

Computation of $\hat{\sigma}^k_T$ at stage T

$$\hat{\sigma}^k_T = \underset{\psi^k \in \Theta_k}{\operatorname{argmax}} \frac{f_{\psi^k}(y^k_1, \ldots, y^k_T)}{f_{\theta^k_0}(y^k_1, \ldots, y^k_T)}.$$

where θ^k_0 is arbitrary.

The Asymptotic Behaviour of the Expected Regret

Theorem 4

For the ARMA problem under consideration, subject to Assumptions 2 and 4, the index policy ϕ^g specified by

$$u_t = \begin{cases} \text{sample from each process once} & \text{if } t \leq K, \\ \text{argmax}\{g^k_{t,n^k_t};\ k \in \{1,\ldots,K\}\} & \text{if } t > K, \end{cases}$$

is a UCB, and hence

$$R_T(\phi^g) = -\sum_{i=1}^{T} \Big(\min_{k \in \{1,2\}} \mathbb{E}{\upsilon^k_{n^k_i}}^2 - \mathbb{E}{\upsilon^{u_i}_{n^{u_i}_i}}^2 \Big) = \mathbf{o}(T^{1+\delta})$$

is obtained, for some $\delta > 0$.

Simulation of 10000 realizations for System 1 for 3 values of C

System 1 (S1)

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

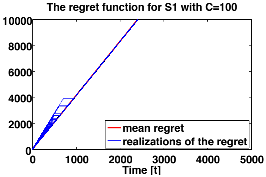

Figure : $C = 100$

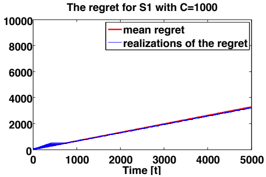

Figure : $C = 1000$

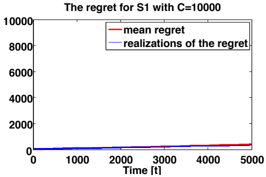

Figure : $C = 10000$

The regret resulted from each realization is plotted in blue, and the regret over all realizations in red.

Simulation of 10000 realizations for System 2 for 3 values of C

System 2 (S2)

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

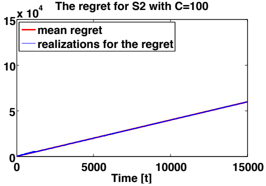

Figure : $C = 100$

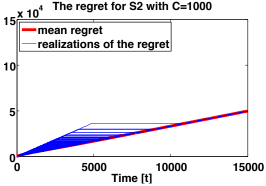

Figure : $C = 1000$

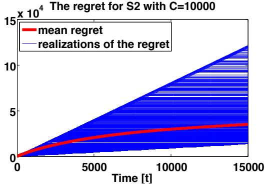

Figure : $C = 10000$

The regret resulted from each realization is plotted in blue, and the regret over all realizations in red.

Simulation of 10000 realizations for System 3 for 3 values of C

System 3 (S3)

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

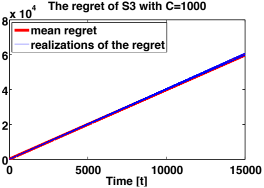

Figure : $C = 1000$

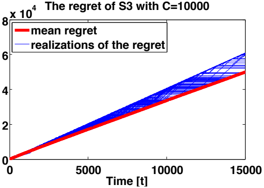

Figure : $C = 10000$

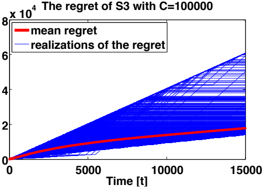

Figure : $C = 100000$

The regret resulted from each realization is plotted in blue, and the regret over all realizations in red.

Conclusion

- We consider the MAB problem with time-dependent rewards that depend on single parameters which lie in a known, finite parameter space.
- We propose the allocation rule ϕ^g that depends on consistent estimators of the unknown parameters.
- Under some assumptions, we have shown that ϕ^g is a UCB and $R_T(\phi^g) \in \mathbf{O}(T^{1+\delta})$ for some $\delta > 0$.
- This result is suboptimal compared to other results in the literature, but there an i.i.d. rewards condition is imposed.
- ϕ^g is more flexible because it can be applied to a more general class of MAB problems, including those with stochastically dependent and time dependent reward processes.