Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

A Marginal Productivity Index Policy for the Finite-Horizon Multiarmed Bandit Problem

José Niño-Mora
Department of Statistics
Universidad Carlos III de Madrid
Av. Universidad 30
28911 Leganés (Madrid), Spain
Email: jnimora@alum.mit.edu
Web: alum.mit.edu/www/jnimora

Abstract— **This paper addresses the discounted discrete-state finite-horizon multiarmed bandit problem. The goal is to design a well-grounded and tractable scheduling policy of index type. The approach is based on reformulating the model as a restless bandit problem, and then deploying the marginal productivity index (MPI) theory developed by the author in recent work, which extends the approach of Klimov (1974), Gittins (1979) and Whittle (1988). It is shown that the model satisfies the author's sufficient conditions for existence of the MPI. An efficient recursive procedure is proposed for computing the finite-horizon MPI. This further furnishes a new method for approximating the infinite-horizon Gittins index.**

I. Introduction

The *finite-horizon multiarmed bandit problem* concerns the optimal allocation of effort to a finite collection of stochastic projects, one of which must be engaged at each period of discrete time. Projects are modelled as binary-action (work/rest) finite-state Markov decision chains (MDCs), which can only change state when active. The goal is to find a scheduling policy which maximizes the expected total discounted value of rewards earned over a finite time horizon.

Given the intractability of such problem, we restrict attention to the class of *index policies*, which is particularly well-suited for practical implementation. Such policies are based on constructing an index for each type of activity, e.g. working on a project, which is a function of its current state. The index policy dynamically prescribes to allocate the resource of concern to the activity having a largest current index value. See e.g. [1] and the references therein. This raises the issue of how to design and construct appropriate index functions.

A powerful approach to index design, introduced in [3] and [2], was extended in [4] and recently developed in [5], [6], [8]. Such approach defines what we have termed a *marginal productivity index (MPI)*, which has a sound economic interpretation, as it measures the marginal productivity of work at every state. The MPI exists in models that satisfy the economic law of diminishing marginal returns to effort. Hence, MPI-based resource allocation policies seek to dynamically allocate a resource to its currently more productive use, using the MPI as a proxy measure of the true marginal productivity of effort.

In this paper we deploy such approach to design and construct new dynamic MPI policies for scheduling a multiarmed bandit over a finite time horizon. It is clear that such model is more realistic for a variety of applications than its infinite-horizon counterpart.

Research on scheduling problems over a finite horizon is scarce, in contrast with the large literature available on corresponding infinite-horizon models. In the finite-horizon case, the analysis is complicated by transient effects, which typically has prevented a characterization of the structure of optimal policies.

The rest of the paper is structured as follows. Section II describes the model of concern. Section III reformulates the model as an infinite horizon restless bandit problem. Section IV outlines the MPI indexability theory, as it applies to the model. Section V outlines the relevant PCL-indexability conditions and the corresponding MPI-computing algorithm. Section VI discusses a more efficient method to compute the finite-horizon MPI. Section VII discusses application of the above results to compute approximately the Gittins index for infinite-horizon projects. Finally, Section VIII ends the paper with some concluding remarks.

The full version of this paper contains proofs of all results and the results of an extensive computational study. See [9].

II. The finite-horizon multiarmed bandit problem

Consider a a finite collection of stochastic projects labelled by $k \in \mathbb{K} \triangleq \{1, \ldots, K\}$. Project k is modelled as a discrete-time, finite-state bandit process, i.e. a binary-action (active/passive) Markov decision chain which does not change state when passive. We denote by $X_k(t)$, $a_k(t)$ and $\mathscr{N}_k$ the project's state and the action taken at time t, and its state space, respectively. The project is active/worked on at time t when $a_k(t) = 1$, and is passive/rested when $a_k(t) = 0$.

The system controller must choose at each of a finite number of time periods $t \in \mathscr{T} \triangleq \{0, 1, \ldots, T\}$ a project to be worked on, while other projects are rested. If project k occupying state i_k is worked on, an immediate active reward $r_k^1(i_k)$ is earned, discounted at the geometric rate $0 < \beta < 1$, and its state evolves according to an active Markovian transition rule, being j_k at the next time period

0-7803-9568-9/05/$20.00 ©2005 IEEE

with probability $p_k(i_k, j_k)$. If the project is rested, a passive reward $r_k^0(i_k)$ is earned, and the state does not change.

Actions are dynamically prescribed through adoption of a *scheduling policy* π, chosen from the class Π of *nonanticipative policies*.

In such setting, it is of interest to consider the problem of finding a discount-optimal scheduling policy, i.e. one maximizing the expected total discounted value of rewards earned over the stated horizon:

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{T} \sum_{k \in \mathbb{K}} r_k^{a_k(t)}(X_k(t))\beta^t\right]. \tag{1}$$

Since problem (1) belongs in the class of finite-state and -action finite-horizon Markov decision processes, an optimal deterministic policy could be obtained in principle by formulating the corresponding Bellman equations, and solving them by backward recursion. Such solution approach, however, becomes impractical in large-scale instances due to the combinatorial explosion of the overall state space.

We will thus address the more practical goal of designing and constructing a well-grounded and tractable scheduling policy. We will focus attention on the class of *(priority) index policies*, which have played a central role in the solution of a wide varity of *stochastic scheduling* problems, including the infinite-horizon version of problem (1), solved by Gittins in [2]. An index policy for problem (1) is characterized by a separate *index* $\nu_k\colon \mathscr{T} \times \mathscr{N}_k \to \mathbb{R}$ attached to each project k. When $t \in \mathscr{T}$ periods remain, the index policy prescribes to work on a project k with largest value of the index $\nu_k(t, X_k(T-t))$. We are thus faced with the issue of how to design and construct well-grounded index functions for problem (1).

III. Reformulation as a restless bandit problem

Our proposed approach to the design and construction of an index policy for problem (1) is based on reformulating it as an infinite-horizon *restless bandit problem (RBP)*, and then deploying in the latter the methods introduced in [4] and developed in [5], [6], [8]. In the RBP, the projects to be scheduled are restless, meaning that they can change state when rested. Such is the case with the model of concern if one considers for each project k the *augmented state*

$$\hat{X}(t) \triangleq \begin{cases} (T-t, X_k(t)) & \text{if } 0 \le t \le T \\ * & \text{if } t \ge T+1, \end{cases}$$

which evolves over the *augmented state space*

$$\hat{\mathscr{N}_k} \triangleq \hat{\mathscr{N}_k}^{\{0,1\}} \cup \{*\},$$

where

$$\hat{\mathscr{N}_k}^{\{0,1\}} \triangleq \mathscr{T} \times \mathscr{N}_k.$$

is the project's *controllable state space*.

States $(t, i_k) \in \hat{\mathscr{N}_k}^{\{0,1\}}$ are *controllable*, in that the active and passive actions are available and differ. We further introduce the *uncontrollable state* $*$, which is absorbing, corresponding to a terminated system after the given time horizon has elapsed, where both actions are identical.

The active and passive transition probabilities are given by

$$\hat{p}_k^1((t, i_k), (t-1, j_k)) = p_k(i_k, j_k)$$
$$\hat{p}_k^0((t, i_k), (t-1, i_k)) = 1,$$

for $1 \le s \le T$, and

$$\hat{p}_k^1((0, i_k), *) = \hat{p}_k^0((0, i_k), *) = 1$$
$$\hat{p}_k^1(*, *) = \hat{p}_k^0(*, *) = 1.$$

All other probabilities are zero.

The one-period rewards of the restless project are

$$\hat{r}_k^{a_k}(i_k, s) = r_k^{a_k}(i_k)$$
$$\hat{r}_k^{a_k}(*) = 0.$$

We can thus reformulate finite-horizon multiarmed bandit problem (1) as the infinite-horizon RBP

$$\max_{\pi \in \Pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \sum_{k \in \mathbb{K}} \hat{r}_k^{a_k(t)}(\hat{X}_k(t))\beta^t\right]. \tag{2}$$

IV. Indexability and the MPI

We outline in this section the definition, interpretation and construction of the discounted MPI for restless projects, as it applies specifically to a reformulated project k of the model of concern. We hence consider the subsystem obtained by considering project *in isolation*. We drop henceforth the project label k. Thus, e.g. now Π denotes the space of admissible effort allocation policies for operating the project, prescribing the action $a(t) \in \{0, 1\}$ to be taken at each time period t.

To evaluate the value of costs incurred under a policy $\pi \in \Pi$, starting at $\hat{\imath} \in \hat{\mathscr{N}}$, we use the *discounted reward measure*

$$f^{\pi}(\hat{\imath}) \triangleq \mathbb{E}_{\hat{\imath}}^{\pi}\left[\sum_{t=0}^{\infty} \hat{r}^{a(t)}(\hat{X}(t))\beta^t\right],$$

where $\mathbb{E}_{\hat{\imath}}^{\pi}[\cdot]$ denotes the corresponding conditional expectation.

We further evaluate the amount of work expended, by the *discounted work measure*

$$g^{\pi}(\hat{\imath}) \triangleq \mathbb{E}_{\hat{\imath}}^{\pi}\left[\sum_{t=0}^{\infty} a(t)\beta^t\right],$$

To avoid technical issues arising from the choice of initial state, we consider this to be drawn from a distribution having a positive probability mass function over $\hat{\mathscr{N}}$. We denote the resulting cost and work measures by f^{π} and g^{π}.

Suppose now that effort must be paid for at a *wage rate* of ν € per unit work performed. We will address the project's *ν-wage subproblem*

$$\max_{\pi \in \Pi} f^{\pi} - \nu g^{\pi}, \tag{3}$$

which is to find an admissible policy maximizing the value of its net rewards (i.e. subtracting working costs).

Since problem (3) is a finite MDP, we know that there exists an optimal deterministic policy which does not depend on the initial state's distribution. We will represent such policies by the set of augmented states $\hat{S}$ where they take the active action, or *active-state sets*, and thus refer, e.g. to the *$\hat{S}$-active policy*.

We will find it convenient to represent such policies in the form

$$\hat{S} = S^0 \oplus \cdots \oplus S^T,$$

meaning that the project is engaged in state i when t periods remain if $i \in S^t$, for $t = 0, \ldots, T$. Intuitive considerations lead us to define the relevant family of *feasible active-state sets* by

$$\hat{\mathscr{F}} \triangleq \left\{ \hat{S} = S^0 \oplus \cdots \oplus S^T : S^0 \subseteq S^1 \subseteq \cdots \subseteq S^T \subseteq \mathscr{N} \right\}.$$

We will henceforth refer to such policies as *$\hat{\mathscr{F}}$-policies*, writing e.g. $f^{\hat{S}}$, $g^{\hat{S}}$, for $\hat{S} \in \hat{\mathscr{F}}$.

We will further say that an ordering $\hat{\boldsymbol{\imath}} = (\hat{\imath}^1, \ldots, \hat{\imath}^n)$ of the n controllable states in $\mathscr{N}^{\{0,1\}}$ is an *$\hat{\mathscr{F}}$-string* if

$$\hat{S}_k(\hat{\boldsymbol{\imath}}) \triangleq \{\hat{\imath}^1, \ldots, \hat{\imath}^{k-1}\} \in \hat{\mathscr{F}}, \quad k = 2, \ldots, n+1.$$

We will further write $\hat{S}_1(\hat{\boldsymbol{\imath}}) = \hat{S}_1 \triangleq \emptyset$.

We next define a key property of the project based on the structure of optimal policies for problem (3) as the *prevailing wage* ν varies over $\mathbb{R}$

Definition 4.1: We say that the restless project is *$\hat{\mathscr{F}}$-indexable* if there exists an index $\nu^*\colon \hat{\mathscr{N}}^{\{0,1\}} \to \mathbb{R}$ which is nonincreasing along an $\hat{\mathscr{F}}$-string $\hat{\boldsymbol{\imath}}$, such that, for every controllable state $\hat{\imath} \in \hat{\mathscr{N}}^{\{0,1\}}$, the $\hat{S}_k(\hat{\boldsymbol{\imath}})$-active policy is optimal for ν-wage problem (3) iff $\nu \in [\nu^*(\hat{\imath}^k), \nu^*(\hat{\imath}^{k-1})]$. We term $\nu^*(\cdot)$ the project's *MPI*.

Such definition suggests, drawing on the theory of optimal resource allocation in economics, that $\nu^*(\hat{\imath})$ must measure the *marginal productivity of work at state* $\hat{\imath}$. Such is indeed the case, as established in [8].

V. PCL-INDEXABILITY CONDITIONS AND MPI CALCULATION

We draw below on the framework for establishing indexability, based on satisfaction of *partial conservation laws (PCLs)*, which we have introduced and developed in [5], [6], [8]. Our main result is Theorem 5.2, which identifies a tractable class of indexable projects, termed *PCL($\hat{\mathscr{F}}$)-indexable*. For our present purposes, it will suffice to formulate the relevant *PCL-indexability* conditions that need to be checked to ensure indexability and calculate the MPI.

Let $\hat{S} \in \hat{\mathscr{F}}$ be a feasible active-state set, and let $\hat{\imath} \in \hat{\mathscr{N}}^{\{0,1\}}$ be a controllable state. We now define the *discounted $(\hat{\imath}, \hat{S})$-marginal workload* by

$$w^{\hat{S}}(\hat{\imath}) \triangleq 1 + \beta \sum_{\hat{\jmath} \in \hat{\mathscr{N}}} \left(\hat{p}^1(\hat{\imath}, \hat{\jmath}) - \hat{p}^0(\hat{\imath}, \hat{\jmath}) \right) g^{\hat{S}}(\hat{\jmath}). \tag{4}$$

Notice that $w^{\hat{S}}(\hat{\imath})$ measures the marginal increment in work expended which results from having the project active instead of passive in the initial period, provided the $\hat{S}$-active policy is adopted thereafter.

```
ALGORITHM AG($\hat{\mathscr{F}}$):
Input: $\hat{\mathbf{r}}^a$
Output: $(\hat{\boldsymbol{\imath}}, \boldsymbol{\nu}^*)$
set  $\hat{S}_1 = \emptyset$
for $k := 1$ to $n$ do
    pick $\hat{\imath}^k \in \arg\max \left\{ \nu^{\hat{S}_k}(\hat{\imath}) \colon \hat{\imath} \in \hat{S}_k^c, \hat{S}_k \cup \{\hat{\imath}\} \in \hat{\mathscr{F}} \right\}$
    set $\nu^*(\hat{\imath}^k) := \nu^{\hat{S}_k}(\hat{\imath}^k)$
    set $\hat{S}_{k+1} := \hat{S}_k \cup \{\hat{\imath}^k\}$
end
```

Fig. 1. Adaptive-greedy algorithm AG($\hat{\mathscr{F}}$).

We further define the *discounted $(\hat{\imath}, \hat{S})$-marginal reward* by

$$r^{\hat{S}}(\hat{\imath}) \triangleq r^1(\hat{\imath}) - r^0(\hat{\imath}) + \beta \sum_{\hat{\jmath} \in \hat{\mathscr{N}}} \left(\hat{p}^1(\hat{\imath}, \hat{\jmath}) - \hat{p}^0(\hat{\imath}, \hat{\jmath}) \right) f^{\hat{S}}(\hat{\jmath}). \tag{5}$$

Thus, $r^{\hat{S}}(\hat{\imath})$ is a measure of the marginal increment in rewards earned which results from having the project active instead of passive in the initial period, provided the $\hat{S}$-active policy is adopted thereafter.

Define now the *discounted $(\hat{\imath}, \hat{S})$-marginal productivity rate*, by

$$\nu^{\hat{S}}(\hat{\imath}) \triangleq \frac{r^{\hat{S}}(\hat{\imath})}{w^{\hat{S}}(\hat{\imath})} \tag{6}$$

provided the denominator does not vanish.

Consider now the *adaptive-greedy algorithm* AG($\hat{\mathscr{F}}$) described in Figure 1, which we introduced and analyzed in [5], [6], drawing on Klimov's [3] algorithm. Notice that we use the notation $\hat{S}^c \triangleq \mathscr{N}^{\{0,1\}} \setminus \hat{S}$. The algorithm is fed with the reformulated restless project's one-period reward vectors $\hat{\mathbf{r}}^a$. It produces as output: (i) an $\hat{\mathscr{F}}$-string $\hat{\boldsymbol{\imath}}$; and (ii) index values $\nu^*(\hat{\imath}^k)$, for $k = 1, \ldots, n$.

We are now ready to state the key result we will use to establish indexability, which we have introduced and proven in [5], [6], [8] in increasingly general settings. We need the following definition.

Definition 5.1: We say that the restless project is *PCL($\hat{\mathscr{F}}$)-indexable* if the following conditions hold:

(i) $w^{\hat{S}}(\hat{\imath}) > 0$, for $\hat{\imath} \in \hat{\mathscr{N}}^{\{0,1\}}$ and $\hat{S} \in \hat{\mathscr{F}}$.

(ii) The sequence of index values produced by the algorithm is nonincreasing:

$$\nu^*(\hat{\imath}^1) \geq \nu^*(\hat{\imath}^2) \geq \cdots \geq \nu^*(\hat{\imath}^n).$$

Theorem 5.2: If the restless project is PCL($\hat{\mathscr{F}}$)-indexable, then it is $\hat{\mathscr{F}}$-indexable, with MPI $\nu^*(\hat{\imath})$.

Thus, Theorem 5.2 gives tractable sufficient conditions for indexability, as it identifies a tractable class of indexable projects.

A key result of this paper, proven in the full version [9], is the following.

Theorem 5.3: The reformulated restless project is PCL($\hat{\mathscr{F}}$)-indexable, having MPI $\nu^*(\hat{\imath})$.

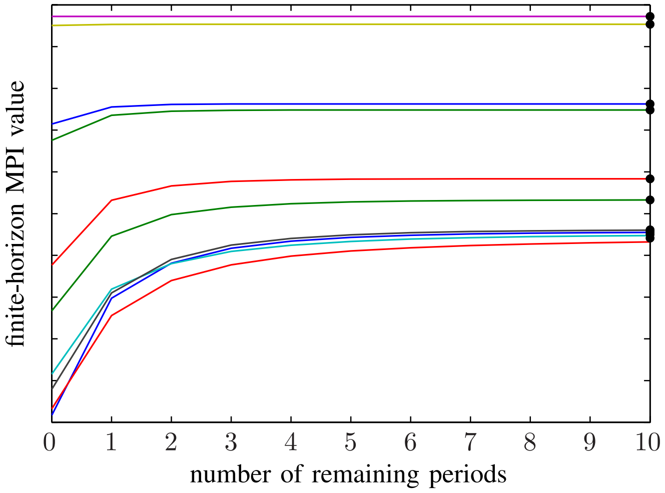

Fig. 2. Experiment 1.

VI. Efficient computation of the finite-horizon MPI

Direct application of the adaptive-greedy algorithm $\mathrm{AG}(\hat{\mathscr{F}})$ to compute the finite-horizon MPI becomes prohibitively expensive as the horizon T grows. This raises the need for a more efficient method for computing the finite-horizon MPI $\nu^*(\hat{\imath})$. A further contribution of this work is to furnish such a method. Here we only outline the key ideas. For full details, see [9].

In short, the improved method identifies and exploits certain recursions relating marginal work, cost and productivity measures over consecutive time periods. Instead of running the above adaptive-greedy $\mathrm{AG}(\hat{\mathscr{F}})$, our improved method runs $T+1$ smaller versions of it, the first one corresponding to a time horizon of $T = 0$ (where the resulting MPI is just the greedy one), the second one corresponding to $T = 1$, and so on until the last one corresponding to the time horizon T of interest. Each algorithm run in these sequence feeds the next one key quantities (marginal work, cost and productivity measures) it will use in its computations, thus reducing its computational burden. The computational savings achieved by such decomposition are dramatic.

VII. A new method for approximating the Gittins index

A byproduct of the above result is a new method for computing approximately the Gittins index $\nu^*(i)$ for an infinite-horizon bandit, by computing the finite-horizon MPI $\nu^*(T, i)$ for T large enough. We have performed several experiments which indicate that it may be enough to use relative small values of T, as convergence to the Gittins index appears to be relatively fast. Thus Figure 2 plots, for a randomly generated instance, the finite-horizon MPI values versus the number of remaining periods. We see that after a few periods the former become very close to the corresponding Gittins indices.

In Figure 3, we have plotted, for a 20-state project, the maximum relative error for using the finite-horizon MPI as an approximation of the Gittins index. Again, we observe that the magnitude of such error drops to acceptable levels for relative small time horizons.

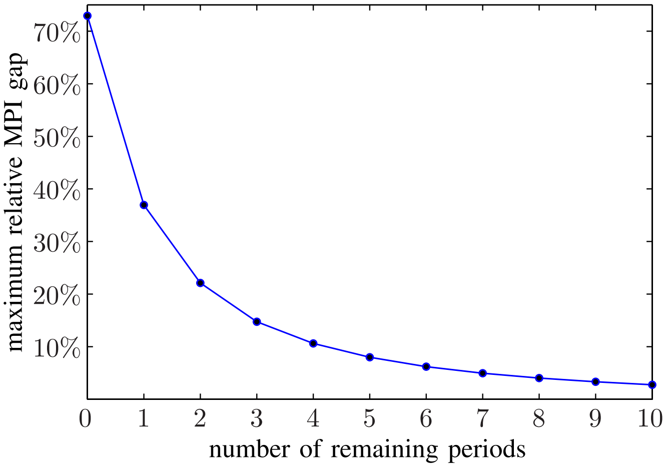

Fig. 3. Experiment 2.

VIII. Conclusion

We have deployed in the finite-horizon multiarmed bandit problem an approach for designing and constructing well-grounded and tractable dynamic index policies based on the theory of MPIs and PCL-indexability conditions, developed by the author in recent work. Such approach does indeed apply to the model of concern, and it produces a relatively efficient method for computing finite-horizon MPI, which can be used both as priority indices for the finite-horizon problem, and as approximations for the infinite-horizon Gittins index. Preliminary computational experience suggests that the suboptimality gap of the finite-horizon MPI policy on the finite-horizon model remains within reasonably small limits. The results of an extensive computational study on the performance of such policy will be reported in the full paper's version [9].

Acknowledgment

The author's work has been supported in part by the Spanish Ministry of Education and Science under a *Ramón y Cajal Investigator Award* and project MTM2004-02334 of the National Research Program in Mathematics, and by the European Union Network of Excellence Euro-NGI.

References

[1] J. Niño-Mora, "Stochastic scheduling," in *Encyclopedia of Optimization*, C.A. Floudas and P.M. Pardalos, Eds., vol. 5, pp. 367–372. Kluwer, Dordrecht, 2001.

[2] J.C. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 148–177, 1979, With discussion.

[3] G.P. Klimov, "Time-sharing queueing systems. I," *Theory Probab. Appl.*, vol. 19, no. 3, pp. 532–551, 1974.

[4] P. Whittle, "Restless bandits: Activity allocation in a changing world," in *A Celebration of Applied Probability*, J. Gani, Ed., *J. Appl. Probab. Special Vol. 25A*, pp. 287–298. Applied Probability Trust, 1988.

[5] J. Niño-Mora, "Restless bandits, partial conservation laws and indexability," *Adv. in Appl. Probab.*, vol. 33, no. 1, pp. 76–98, 2001.

[6] J. Niño-Mora, "Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach," *Math. Program.*, vol. 93, no. 3, Ser. A, pp. 361–413, 2002.

[7] J. Niño-Mora, "Marginal productivity index policies for scheduling multiclass delay-/loss-sensitive traffic," in *Proc. 1st Euro-NGI Conference on Next Generation Internet Networks — Traffic Engineering (NGI 05)*, Rome, Apr. 2005, pp. 61–66.

[8] J. Niño-Mora, "Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock $M/G/1$ queues," *Math. Oper. Res.*, forthcoming.

[9] J. Niño-Mora, "A marginal productivity index policy for the finite-horizon multiarmed bandit problem," unpublished.