An Index Policy for Multiarmed Multimode Restless Bandits

(Invited paper)

José Niño-Mora
Department of Statistics
Universidad Carlos III de Madrid
Avda. Universidad 30
28911 Leganés (Madrid), Spain
jnimora@alum.mit.edu

ABSTRACT

This paper introduces and addresses the multiarmed multimode restless bandit problem, concerning the optimal dynamic allocation of a shared resource to a collection of projects which can be operated in multiple modes, subject to a peak resource consumption constraint, thus extending the conventional multiarmed restless bandit problem where projects are restricted to a binary-mode (active or passive) operation. After discussing a motivating application, concerning the optimal dynamic power allocation to multiple users sharing a wireless downlink communication channel subject to a peak energy constraint, a general approach is developed to design and compute a tractable heuristic policy based on marginal productivity indices (MPIs) defined separately for each project. Sufficient conditions are given which ensure both the existence of such an index and the validity of an adaptive-greedy algorithm for its computation. Such conditions extend to multimode projects those introduced by the author in earlier work for binary-mode projects.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Markov processes; F.2.1 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems

General Terms

Algorithms, Theory

Keywords

Markov decision processes, dynamic resource allocation, multimode projects, restless bandits, index policies

1. INTRODUCTION

ValueTools 2008 October 21–23, 2008, Athens, GREECE

1.1 Motivation and Background

A decision-maker aims to extract the maximum gain out of a portfolio of M dynamic and stochastic projects labeled by $m = 1, \ldots, M$. Project m is modeled as a discrete-time *Markov decision process* (MDP) moving on the finite state space $\mathbb{X}_m$, which at each time is to be operated in one of a finite set of *modes* $a_m \in \mathbb{A}_m \triangleq \{0, 1, \ldots, A_m\}$ representing intensity or effort levels. Thus, $a_m = 0$ corresponds to idling the project, and $a_m = A_m$ to operating it at full gear. If engaged in mode a_m when it occupies state i_m, the project yields a one-period reward $R_m^{a_m}(i_m)$ and changes state at the next time period to j_m with probability $p_m^{a_m}(i_m, j_m)$. It further expends an amount $0 \leq Q_m^{a_m}(i_m) \leq q$ of a certain shared resource, where $q > 0$ is the resource amount available per period for all projects. We assume that, for each state $i_m \in \mathbb{X}_m$,

$$0 \leq Q_m^0(i_m) \leq Q_m^1(i_m) \leq \cdots \leq Q_m^{A_m}(i_m) \leq q, \qquad (1)$$

i.e., the higher the intensity level the larger the resource expenditure. Rewards and resource expenditures are discounted over time with factor $0 < \beta < 1$.

It will be convenient to partition the state space $\mathbb{X}_m$ for each project m into the (possibly empty) set $\mathbb{X}_m^{\{0\}}$ of *uncontrollable* states, where all actions are effectively identical to idling the project (hence the notation), i.e., for $i_m \in \mathbb{X}_m^{\{0\}}$,

$$\begin{aligned} Q_m^0(i_m) &= Q_m^1(i_m) = \cdots = Q_m^{A_m}(i_m) \\ R_m^0(i_m) &= R_m^1(i_m) = \cdots = R_m^{A_m}(i_m), \end{aligned} \qquad (2)$$

and the remaining set $\overline{\mathbb{X}}_m$ of *controllable* states. We assume that the latter is nonempty, and that, for such states, the resource consumption measure is increasing in the mode:

$$Q_m^0(i_m) < Q_m^1(i_m) < \cdots < Q_m^{A_m}(i_m). \qquad (3)$$

Decisions as to the mode in which to operate each project at each time are based on adoption of a *scheduling policy* $\boldsymbol{\pi}$. In addition to being drawn from the class $\mathbf{\Pi}$ of *nonanticipative randomized policies* that use only past or present information on states and actions, in order to be *admissible* a policy must further not expend more than the available q units of resource per period. The *multiarmed multimode restless bandit problem* (MAMRBP) introduced herein is to find an admissible policy that maximizes the expected total discounted reward earned over an infinite horizon. Denoting by $X_m(t)$ and $a_m(t)$ the prevailing state and action (mode)

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410

on project m at time t, we can formulate such a problem as

$$\begin{aligned} v^*(\mathbf{i}) = \max \mathbb{E}_{\mathbf{i}}^{\boldsymbol{\pi}} & \left[\sum_{t=0}^{\infty} \sum_{m=1}^{M} R_m^{a_m(t)}\big(X_m(t)\big)\beta^t \right] \\ & \text{subject to} \\ & \sum_{m=1}^{M} Q_m^{a_m(t)}\big(X_m(t)\big) \le q, \quad t = 0, 1, 2, \ldots \\ & \boldsymbol{\pi} \in \boldsymbol{\Pi}, \end{aligned} \tag{4}$$

where $\mathbb{E}_{\mathbf{i}}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under policy $\boldsymbol{\pi}$ conditioned on the initial joint state being equal to $\mathbf{i} = (i_m)$.

We will further consider the MAMRBP under the (long-run) average criterion:

$$\begin{aligned} \bar{v}^*(\mathbf{i}) = \max \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}_{\mathbf{i}}^{\boldsymbol{\pi}} & \left[\sum_{t=0}^{T} \sum_{m=1}^{M} R_m^{a_m(t)}\big(X_m(t)\big)\beta^t \right] \\ & \text{subject to} \\ & \sum_{m=1}^{M} Q_m^{a_m(t)}\big(X_m(t)\big) \le q, \quad t = 0, 1, 2, \ldots \\ & \boldsymbol{\pi} \in \boldsymbol{\Pi}. \end{aligned} \tag{5}$$

Problems (4)–(5) furnish a new, powerful modeling framework for optimal dynamic resource allocation in multiproject settings that substantially expands the scope of previous frameworks considered in the literature. Consider the following special cases:

(a) The classic *multiarmed bandit problem* (MABP). The projects of concern are *classic*, in the sense that they only allow two actions ($a_m = 0$: passive; $a_m = 1$: active) and their states do not change while they are passive. The resource to be allocated at each time period is the attention of a single operator, which is modeled by setting $q = 1$, $Q_m^0(i_m) \equiv 0$ and $Q_m^1(i_m) \equiv 1$. Note that while the conventional formulation of the MABP requires that exactly one project be engaged at each time, the case where at most one project is to be engaged is easily reduced to such a binding case. This model was solved in [4] by means of an *index policy* based on attaching a certain (Gittins) index $\nu_m^*(i_m)$ to each project m as a function of its state i_m. For the conventional case where the resouce constraint is binding, it is optimal to engage at each time a project of largest index; for the nonbinding case, it is optimal to engage at each time a project of largest positive index, if any, and otherwise to idle all projects.

(b) The *MABP with multiple plays* (cf. [9]). This model extends the classic MABP by allowing up to $q \ge 2$ projects to be engaged at each time. The parameters $Q_m^{a_m}(i_m)$ are as in (a). Although in this case the Gittins index policy is generally suboptimal, [9] shows that it is optimal under certain conditions.

(c) The conventional *multiarmed restless bandit problem* (MARBP). The projects of concern are *restless*, in the sense that they can change state while passive, although they only allow two modes of operation (active and passive). Again, the parameters $Q_m^{a_m}(i_m)$ are as in the MABP. The MARBP was introduced by Whittle in [13]. Although the problem is generally intractable, he proposed a heuristic index policy for the long-run average criterion case with a binding resource constraint, based on Lagrangian relaxation ideas, which in the classic MABP recovers the optimal policy.

(d) The problem of dynamic power allocation to multiple users sharing a wireless downlink communication channel subject to a peak energy constraint, in the setting of the Markovian model addressed in [2]. This problem represents a concrete and relevant motivating application for the new modeling framework in (4)–(5). In such a setting, the projects represent users sharing a downlink fading wireless channel, the state of a project comprises both the number of packets awaiting transmission to the corresponding user, with packets being held in a separate finite buffer for each user, and the user's binary channel state (good or bad). The modes for a project/user correspond to a set of possible transmission rates (which in [2] is taken as an interval, which would have to be discretized to fit the model in the present framework). The reward for a project/user is its average throughput. As for the shared limited resource, it represents transmission power, so that the quantities $Q_m^{a_m}(i_m)$ are amounts of power expended. Actually, the problem addressed and solved (via Lagrangian methods) in [2] under the average criterion is not (5), but the simpler problem obtained by replacing the *peak power consumption constraint* in (5) by the *average power consumption constraint*

$$\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\mathbf{i}}^{\boldsymbol{\pi}} \left[\sum_{t=0}^{T} \sum_{m=1}^{M} Q_m^{a_m(t)}\big(X_m(t)\big) \right] \le q.$$

The authors state in [2, p. 41] why they choose to focus on the latter problem instead of on (5):

> "Undoubtedly, a constraint on the total power consumption rate, that is, a peak power constraint, as opposed to a constraint on average power consumption would be more appropriate as the users may be allocated very high power levels at short time spans under an average power constraint. However, such a peak power constraint renders our formulation intractable."

Since problems (4)–(5) are generally intractable, we will address the goals of designing and computing a well-grounded and tractable heuristic scheduling policy that performs well. The availability and success of index policies for the special cases (a)–(c) of binary-mode projects (cf. [8] and the accompanying discussions), motivates the interest of extending such results to design tractable index-based heuristic policies for multimode projects, which would be applicable to more complex models such as that in case (d) above.

Yet, carrying out such an extension raises fundamental research challenges, as the concept of "index policy" for problems (4)–(5) in the general multimode project case has not been previously defined in the literature. Thus, two issues need to be addressed: (1) How to define appropriate indices for multimode projects that extend those known for binary-mode projects?; and (2) How to use indices defined for individual projects to construct scheduling policies for the multiproject problems?

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410

Concerning the first issue, the only prior work is a proposal outlined by Weber in [12], in the setting of a particular queueing admission control model, for defining indices $\nu_m^{*,a_m}(i_m)$ attached to a multimode project m depending on both its mode a_m and state i_m, which extend those introduced by Whittle [13] for binary-mode projects. As for the second issue, to the author's knowledge it has not been previously addressed in the literature. A third related issue of interest is to obtain tractable upper *bounds* on the optimal performance objective for problems (4)–(5), which can be used to assess in large-scale instances the degree of suboptimality of proposed heuristic policies.

Motivated by such issues, this paper presents the following contributions, which extend to the MAMRBP the approach introduced in [13] for the MARBP with binary-mode projects: it develops a general approach to the design and computation of a tractable heuristic policy based on marginal productivity indices (MPIs) defined separately for each project. Sufficient conditions are given which ensure both the existence of such an index and the validity of an adaptive-greedy algorithm for its computation. Such conditions extend to multimode projects those introduced by the author in earlier work for binary-mode projects.

We remark that the results for multimode projects offer a new approach to a variety of problems addressed in the literature on optimal control of queues by conventional methods. See, e.g., [3], [10], [5], and [11]. We further remark that the approach to indexation of a multimode restless project developed herein extends the ideas outlined by Weber in [12], and builds on the author's earlier results on indexation for binary-mode projects starting in [6], which are reviewed in [8].

The remainder of the paper is organized as follows. Section 2 develops a Lagrangian relaxation approach to the MAMRBP, focusing on the discounted problem (4). Section 3 develops the indexation theory for a multimode restless project. Section 4 ends the paper with some concluding remarks.

Proofs of the results announced in this paper will be given in the full version, currently under preparation, along with the results of a computational study.

2. BOUNDS AND INDEX POLICIES

2.1 Centralized Problem Relaxation

We focus the ensuing presentation on the discounted MAMRBP (4). This is with no loss of generality, since the following discussion can be readily adapted to the time-average criterion. To obtain a simpler *relaxed problem*, imagine that each project m is autonomously operated by a dedicated operator, who implements a policy π_m drawn from the class Π_m of nonanticipative randomized policies for operating the project as if it were in isolation. Overall control is *centralized* by a central controller who is in charge of choosing and assigning to each project operator his corresponding policy, subject to the coordinating constraint that the expected total discounted amount of resource expended does not exceed $q/(1-\beta)$:

$$\sum_{m=1}^{M} \mathbb{E}_{i_m}^{\pi_m}\left[\sum_{t=0}^{\infty} Q_m^{a_m(t)}\big(X_m(t)\big)\beta^t\right] \le \frac{q}{1-\beta}. \tag{6}$$

Such a discounted average resource consumption constraint *relaxes* the peak resource consumption constraint in (4).

The relaxed problem so obtained is

$$\begin{aligned} v^{\mathrm{R}}(\mathbf{i}) = \max & \sum_{m=1}^{M} \mathbb{E}_{i_m}^{\pi_m}\left[\sum_{t=0}^{\infty} R_m^{a_m(t)}\big(X_m(t)\big)\beta^t\right], \\ & \text{subject to} \\ & (6) \\ & \pi_m \in \Pi_m, \quad m = 1, \ldots, M. \end{aligned} \tag{7}$$

Note that the optimal value of (7), $v^{\mathrm{R}}(\mathbf{i})$, furnishes an upper bound on the optimal value of (4), i.e., $v^*(\mathbf{i}) \le v^{\mathrm{R}}(\mathbf{i})$.

2.2 Decentralized Lagrangian Relaxation

To address (7) we deploy a Lagrangian approach, by attaching a Lagrange multiplier ν to constraint (6). This allows us to dualize such a constraint, bringing it into the objective, which yields, after eliminating the constant $\nu q/(1-\beta)$ from the objective, the problem

$$\begin{aligned} & \max \sum_{m=1}^{M} \mathbb{E}_{i_m}^{\pi_m}\left[\sum_{t=0}^{\infty}\Big\{R_m^{a_m(t)}\big(X_m(t)\big) - \nu Q_m^{a_m(t)}\big(X_m(t)\big)\Big\}\beta^t\right], \\ & \text{subject to} \\ & \pi_m \in \Pi_m, \quad m = 1, \ldots, M. \end{aligned} \tag{8}$$

Denote by $v^{\mathrm{L}}(\mathbf{i};\nu)$ the optimal value of (8) *plus such a constant*. Note that, for every nonnegative value $\nu \ge 0$ of the multiplier, (8) is a *Lagrangian relaxation* of (7), in the sense that $v^{\mathrm{R}}(\mathbf{i}) \le v^{\mathrm{L}}(\mathbf{i};\nu)$.

In the setting of the economic interpretation outlined above, (8) is a *decoupled* or *decentralized problem* where the central controller's role is reduced to informing the project operators of the *resource unit price* ν. Based on such information, operators are then left free to seek policies that maximize their own projects' objective values (net of resource expenditures), by solving the *individual project subproblems*

$$\max_{\pi_m \in \Pi_m} \mathbb{E}_{i_m}^{\pi_m}\left[\sum_{t=0}^{\infty}\Big\{R_m^{a_m(t)}\big(X_m(t)\big) - \nu Q_m^{a_m(t)}\big(X_m(t)\big)\Big\}\beta^t\right], \tag{9}$$

for $m = 1, \ldots, M$.

Denoting by $v_m^*(i_m;\nu)$ the optimal value of subproblem (9), the relaxed Lagrangian value $v^{\mathrm{L}}(\mathbf{i};\nu)$ is represented as

$$v^{\mathrm{L}}(\mathbf{i};\nu) = \frac{q}{1-\beta}\nu + \sum_{m=1}^{M} v_m^*(i_m;\nu). \tag{10}$$

Consider now the issue of whether relaxed resource allocation problem (7) can be effectively decentralized, i.e., reduced to (8), by pitching the resource unit price ν to an appropriate critical level ν^* for which $v^{\mathrm{R}}(\mathbf{i}) = v^{\mathrm{L}}(\mathbf{i};\nu^*)$. Note that, if such a value ν^* exists, it can be obtained by solving the *Lagrangian problem*

$$\min_{\nu \ge 0} v^{\mathrm{L}}(\mathbf{i};\nu), \tag{11}$$

which is a *convex optimization problem* —since each function $v_m^*(i_m;\nu)$ is convex in ν, being the maximum of linear functions in ν. Now, in the case of finite state and action projects of concern herein, we can indeed ensure existence of such a critical value ν^*, as it follows from the *strong duality theorem* of linear programming (LP), noting that all

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410

of the above problems, being finite-state and -action constrained MDPs, are readily formulated as LP problems (cf. [1]). Note further that, generally, the critical multiplier ν^* will depend on the initial joint project state, i.e., $\nu^* = \nu^*(\mathbf{i})$.

2.3 Indexable Projects and the Marginal Productivity Index

In order to deploy the above approach, one must be able to solve efficiently the individual project subproblems (9) as the resource price parameter ν varies. We will refer henceforth to (9) as the *ν-price subproblem* for project m.

It appears reasonable to expect that, in models arising from relevant applications, as the resource price ν gets higher the optimal project operating mode should get smaller (as a higher mode incurs a larger consumption; cf. (1)). Such an intuitive idea is formalized in the following definition.

Extending the approach outlined in [12] in an admission control model to the present general setting, let us say that project m (or its ν-price subproblem (9)) is *indexable* if, as the resource unit price ν increases from $-\infty$ to ∞, the lowest optimal action (mode) in each controllable state $j_m \in \overline{\mathbb{X}}_m$ decreases monotonically from A_m to 0. In such a case, optimal policies for (9) are determinined by an *index* $\nu^*_m(a_m, j_m)$ that is monotone nonincreasing in a_m, i.e.,

$$\nu^*_m(A_m, j_m) \leq \cdots \leq \nu^*_m(1, j_m), \tag{12}$$

as follows: under any initial state i_m, action $a_m = A_m$ is optimal in state j_m iff $\nu \leq \nu^*_m(A_m, j_m)$; action $a_m \in \{1, \ldots, A_m - 1\}$ is optimal iff $\nu^*_m(a_m+1, j_m) \leq \nu \leq \nu^*_m(a_m, j_m)$; and action $a_m = 0$ is optimal iff $\nu \geq \nu^*_m(1, j_m)$.

Thus, for an indexable project, index value $\nu^*_m(a_m, j_m)$ is the unique critical value of the unit resource price ν under which one should be indifferent between engaging the project in mode a_m or in mode $a_m - 1$ when it occupies state j_m, as both actions are then optimal. Extending to the present setting the terminology introduced in [7], based on the economic interpretation of the index, we will refer to $\nu^*_m(a_m, j_m)$ as the project's *marginal productivity index* (MPI).

2.4 An Index-based Scheduling Policy

Supppose now that each project in (4) is indexable. We next address the issue of how to use the indices $\nu^*_m(a_m, j_m)$ which are defined separately for each project to obtain a scheduling policy for the multiproject problem (4). In the case of binary-mode projects on which prior work has focused, the appropriate scheduling policy is obtained by engaging (in the active mode: 1) at each time those projects, up to a maximum of q, which occupy a controllable state j_m, and have higher positive index values $\nu^*_m(1, j_m)$, if any, and idling (passive mode: 0) other projects. Yet, extending such an index policy to the case of multimode projects is not trivial, since the policy must prescribe the mode in which each project is to be run.

The proposal we introduce next is based on the economic interpretation of index $\nu^*_m(a_m, j_m)$ as the marginal productivity on project m that corresponds to taking action a_m when it occupies state j_m. Thus, suppose that at time t project m lies in state $X_j(t) = j_m$, for $m = 1, \ldots, M$. We then propose to obtain the mode a^*_m in which to run each project m in time period t by solving the problem

$$\begin{aligned} &\max \sum_{m=1}^{M} \nu^*_m(a_m, j_m) \\ &\text{subject to} \\ &\sum_{m=1}^{M} Q^{a_m}_m(j_m) \leq q \\ &a_m \in \mathbb{A}_m, \quad m = 1, \ldots, M, \\ &a_m = 0, \quad j_m \in \mathbb{X}^{\{0\}}_m, \end{aligned} \tag{13}$$

where we take $\nu^*_m(0, j_m) \equiv 0$. Note that (13) is a *nonlinear integer knapsack problem*, which in the case of binary-mode projects does yield the appropriate index policy. The motivation and intuition behing problem (13) is that it seeks to obtain a feasible resource allocation that maximizes the total marginal productivity at each joint state $\mathbf{j} = (j_m)$.

3. MULTIMODE PROJECT INDEXATION: THEORY AND COMPUTATION

This section extends to the setting of multimode restless projects the approach introduced and developed by the author in previous work for the special binary-mode case (cf. [8]) for establishing that a given project model is indexable, and computing its MPI.

We will thus consider a single multimode restless project as above, for which we will drop henceforth the label m from the notation. We evaluate the value of rewards earned under policy π starting in state i by the discounted *reward measure*

$$f^\pi(i) \triangleq \mathbb{E}^\pi_i \left[\sum_{t=0}^{\infty} R^{a(t)}\big(X(t)\big) \right], \tag{14}$$

and similarly evaluate the corresponding amount of resource expended by the discounted *resource consumption measure*

$$g^\pi(i) \triangleq \mathbb{E}^\pi_i \left[\sum_{t=0}^{\infty} Q^{a(t)}\big(X(t)\big) \right]. \tag{15}$$

For a given resource unit price $\nu \in \mathbb{R}$, consider the project's *ν-price problem*,

$$\max_{\pi \in \Pi} f^\pi(i) - \nu g^\pi(i), \tag{16}$$

which is to find an operating policy that maximizes the value of rewards earned minus resource consumption expenses.

We say that the project's *ν-price problem* (16) is *indexable* if, as the resource unit price ν increases from $-\infty$ to ∞, the lowest optimal action (mode) in each controllable state $j \in \overline{\mathbb{X}}$ decreases monotonically from A to 0. In such a case, the optimal policies for (16) are determinined by an *index* $\nu^*(a, j)$ that is monotone nonincreasing in the mode a, i.e.,

$$\nu^*(A, j) \leq \cdots \leq \nu^*(1, j), \tag{17}$$

as follows: under any initial state i, action $a = A$ is optimal in state j iff $\nu \leq \nu^*(A, j)$; action $a \in \{1, \ldots, A-1\}$ is optimal iff $\nu^*(a+1, j) \leq \nu \leq \nu^*(a, j)$; and action $a = 0$ is optimal iff $\nu \geq \nu^*(1, j)$.

3.1 Exploiting Special Structure: Indexability relative to a Family of Policies

While one can readily test numerically whether a given multimode restless bandit instance is or not indexable, a

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410

researcher investigating a particular model will instead be concerned with establishing analytically its indexability under an appropriate range of model parameters. The key to achieving such a goal is to exploit special structure by *guessing* an appropriate family of stationary deterministic policies among which an optimal policy for (16) exists for every resource unit price $\nu \in \mathbb{R}$.

We will represent a stationary deterministic policy by its corresponding *partition* $\mathbf{S} = (S_0, S_1, \ldots, S_A) \in \bar{\mathscr{P}}(\mathbb{X})$, where S_a is the set of states where the policy prescribes to use mode a, and $\bar{\mathscr{P}}(\mathbb{X})$ denotes the class of partitions which idle the project in uncontrollable states, i.e., $\mathbb{X}^{\{0\}} \subseteq S_0$. A family of such policies is thus given as a family $\mathscr{F}$ of partitions $\mathbf{S} \in \bar{\mathscr{P}}(\mathbb{X})$, and hence we will refer to the family of $\mathscr{F}$*-policies*. Relative to such a family, we will say that the project is $\mathscr{F}$*-indexable* if (i) it is indexable, and (ii) $\mathscr{F}$-policies are optimal for ν-price problem (16), i.e., for every resource price $\nu \in \mathbb{R}$ there is an optimal partition $\mathbf{S}^* \in \mathscr{F}$ for (16).

We next introduce a *partial order* relation on such partitions as follows: $\mathbf{S} = (S_0, S_1, \ldots, S_A) \preceq \mathbf{S}' = (S'_0, S'_1, \ldots, S'_A)$ if, for every controllable state $i \in \overline{\mathbb{X}}$ it holds that, if $i \in S_a$ then $i \in S'_b$ for some $b \geq a$. In other words, $\mathbf{S} \preceq \mathbf{S}'$ if $S_a \subseteq \cup_{b=a}^{A} S'_b$.

For a pair of such partitions $\mathbf{S} = (S_0, S_1, \ldots, S_A)$ and $\mathbf{S}' = (S'_0, S'_1, \ldots, S'_A)$, the *join* $\mathbf{S}'' = \mathbf{S} \vee \mathbf{S}'$ relative to such a partial order relation is given by

$$S''_a = \Big(S_a \cap \big(\cup_{b=0}^{a} S'_b \big) \Big) \cup \Big(S'_a \cap \big(\cup_{b=0}^{a} S_b \big) \Big),$$

and the *meet* $\mathbf{S}'' = \mathbf{S} \wedge \mathbf{S}'$ is given by

$$S''_a = \Big(S_a \cap \big(\cup_{b=a}^{A} S'_b \big) \Big) \cup \Big(S'_a \cap \big(\cup_{b=a}^{A} S_b \big) \Big).$$

We further introduce an operator $T_i^{a,b}$ on such partitions as follows. Given a partition $\mathbf{S} = (S_0, S_1, \ldots, S_A)$ with $i \in S_a$, and a mode $b \neq a$, we say that $\mathbf{S}' = (S'_0, S'_1, \ldots, S'_A) = T_i^{a,b}\mathbf{S}$ if such partitions differ only in the states assigned to modes a and b as follows: $S'_a = S_a \setminus \{i\}$, $S'_b = S_b \cup \{i\}$. Namely, if partition $\mathbf{S}'$ is obtained from $\mathbf{S}$ by assigning to state i action b instead of action a.

We say that two partitions $\mathbf{S}$ and $\mathbf{S}'$ are *adjacent* if they are of the form $\mathbf{S}' = T_i^{a,a-1}\mathbf{S}$ or $\mathbf{S}' = T_i^{a,a+1}\mathbf{S}$ for some i, a.

We further denote by $\mathbf{S}_{\min} \triangleq (\mathbb{X}, \emptyset, \ldots, \emptyset)$ and by $\mathbf{S}_{\max} \triangleq (\mathbb{X}^{\{0\}}, \emptyset, \ldots, \emptyset, \overline{\mathbb{X}})$ the minimum and the maximum partitions relative to such a partial ordering, respectively.

We impose the following connectivity requirements on $\mathscr{F}$.

ASSUMPTION 3.1. *$\mathscr{F}$ satisfies the following conditions:*

(i) $\mathbf{S}_{\min}, \mathbf{S}_{\max} \in \mathscr{F}$*;*

(ii) *for every pair $\mathbf{S}, \mathbf{S}' \in \mathscr{F}$ with $\mathbf{S} \prec \mathbf{S}'$ there exist (i, a) and (i', a') such that $T_i^{a,a+1}\mathbf{S} \preceq \mathbf{S}'$, $\mathbf{S} \preceq T_{i'}^{a',a'-1}\mathbf{S}'$, and $T_i^{a,a+1}\mathbf{S}, T_{i'}^{a',a'-1}\mathbf{S}' \in \mathscr{F}$;*

(iii) *for any $S, S' \in \mathscr{F}$, $S \vee S', S \wedge S' \in \mathscr{F}$.*

Note that condition (iii) in Assumption 3.1 means that $\mathscr{F}$ is a *lattice* relative to the given partial order. As for condition (ii), it ensures that any two nested partitions $\mathbf{S}, \mathbf{S}' \in \mathscr{F}$ with $\mathbf{S} \prec \mathbf{S}'$ can be connected by an increasing *chain* $\mathbf{S} = \mathbf{S}_0 \prec \cdots \prec \mathbf{S}_k = \mathbf{S}'$ of adjacent partitions in $\mathscr{F}$. Further, condition (i) ensures that one can connect in such a fashion $(\mathbb{X}, \emptyset, \ldots, \emptyset)$ to, $(\mathbb{X}^{\{0\}}, \emptyset, \ldots, \emptyset, \overline{\mathbb{X}})$. We will call a set family $\mathscr{F}$ satisfying Assumption 3.1(ii, iii) a *monotonically connected lattice*.

3.2 Indexability Conditions and Index Algorithm.

Suppose that, for a given multimode restless bandit model, a suitable family of partitions $\mathscr{F}$ as above has been identified relative to which one seeks to establish analytically $\mathscr{F}$-indexability. We next introduce sufficient conditions and an index algorithm which extend to the present multimode setting those introduced by the author in earlier work for binary-mode projects.

To formulate the conditions and the index algorithm we need to define certain *marginal measures*, as follows. For an action $a \in \mathbb{A}$ and a partition $\mathbf{S} \in \bar{\mathscr{P}}(\mathbb{X})$, denote by $\langle a, \mathbf{S} \rangle$ the policy that takes action a in the initial stage, and adopts the $\mathbf{S}$-*policy* thereafter. For a state i, an action a, and a partition $\mathbf{S}$, define the *marginal resource measure* $w^{\mathbf{S}}(a, i)$ by

$$w^{\mathbf{S}}(a, i) \triangleq g^{\langle a, \mathbf{S} \rangle}(i) - g^{\mathbf{S}}(i), \tag{18}$$

i.e., as the marginal increase in the amount of resource expended which results from using initially mode a in state i, provided that the $\mathbf{S}$-policy is adopted thereafter. Note that such a marginal resource measure vanishes at uncontrollable states:

$$w^{\mathbf{S}}(a, i) \equiv 0, \quad i \in \mathbb{X}^{\{0\}}. \tag{19}$$

Further, define the *marginal reward measure* $r^{\mathbf{S}}(a, i)$ by

$$r^{\mathbf{S}}(a, i) \triangleq f^{\langle a, \mathbf{S} \rangle}(i) - f^{\mathbf{S}}(i), \tag{20}$$

i.e., as the corresponding marginal increase in the value of rewards earned. Finally, for $w^{\mathbf{S}}(a, i) \neq 0$ define the *marginal productivity measure* $\nu^{\mathbf{S}}(a, i)$ by

$$\nu^{\mathbf{S}}(a, i) \triangleq \frac{r^{\mathbf{S}}(a, i)}{w^{\mathbf{S}}(a, i)}. \tag{21}$$

We will further refer to the *adaptive-greedy index algorithm* given in Table 1 —in its *top-down* version, where index values are meant to be computed from highest to lowest; one could similarly consider the symmetric *bottom-up* version. Such an algorithm has a very simple structure, as it constructs in n steps, where $n \triangleq A|\overline{\mathbb{X}}|$ is the number of nonzero modes times the number of controllable states, an increasing chain of adjacent partitions $\mathbf{S}_0 = \mathbf{S}_{\min} \prec \mathbf{S}_1 \prec \cdots \prec \mathbf{S}_n = \mathbf{S}_{\max}$ in $\mathscr{F}$, proceeding at each step in a greedy fashion. Thus, once partition $\mathbf{S}_{k-1} \in \mathscr{F}$ has been constructed, the next adjacent partition $\mathbf{S}_k$ is obtained by gearing up one notch from $a - 1$ to a the mode of some controllable state i, in such a way that the choice of such a mode and state maximizes the marginal productivity rate $\nu^{\mathbf{S}_{k-1}}(a, i)$, while restricting attention to action-state pairs for which the partition S_k so obtained remains in $\mathscr{F}$. Ties are broken arbitrarily.

The main result of this section, giving the new indexability conditions and ensuring the validity of the adaptive-greedy index algorithm to compute the MPI, is stated next.

THEOREM 3.2. *The following holds:*

(a) *Suppose that:*

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410

Table 1: Adaptive-greedy Index Algorithm $\mathrm{AG}_{\mathscr{F}}$.

ALGORITHM $\mathrm{AG}_{\mathscr{F}}$:
Output:
$\{(a^k, i^k), \nu^*(a^k, i^k)\}_{k=1}^n$

$\mathbf{S}^0 := \mathbf{S}^{\min}$
for $k := 1$ **to** n **do**
  pick $(a^k, i^k) \in \underset{i \in S_{a-1}^{k-1}, T_i^{a-1,a}\mathbf{S}^{k-1} \in \mathscr{F}}{\arg\max} \nu^{\mathbf{S}^{k-1}}(a, i)$
  $\nu^*(a^k, i^k) := \nu^{\mathbf{S}^{k-1}}(a^k, i^k)$; $\mathbf{S}^k := T_{i^k}^{a^k-1,a^k}\mathbf{S}^{k-1}$
end { for }

(i) *for every partition* $\mathbf{S} \in \mathscr{F}$,

$$\begin{aligned} w^{\mathbf{S}}(a,i) > 0, \quad i \in S_{a-1}, T_i^{a-1,a}\mathbf{S} \in \mathscr{F}, \\ w^{\mathbf{S}}(a,i) < 0, \quad i \in S_{a+1}, T_i^{a+1,a}\mathbf{S} \in \mathscr{F}; \end{aligned} \tag{22}$$

or, equivalently, for every nested partition pair $\mathbf{S} \prec \mathbf{S}'$ *with* $\mathbf{S}, \mathbf{S}' \in \mathscr{F}$,

$$\left(g_i^{\mathbf{S}}\right)_{i\in\mathbb{X}} \lneq \left(g_i^{\mathbf{S}'}\right)_{i\in\mathbb{X}}. \tag{23}$$

(ii) *for every resource price* $\nu \in \mathbb{R}$ *there exists an optimal* $\mathscr{F}$*-policy for* ν*-price problem* (16).

Then, the project is $\mathscr{F}$*-indexable and algorithm* $\mathrm{AG}_{\mathscr{F}}$ *gives its MPI* $\nu^*(a_k, i_k)$ *in nonincreasing order.*

(b) *If the project is indexable then it satisfies conditions* (i) *and* (ii) *in part* (a) *for some nested family* $\mathscr{F}$ *of adjacent partitions.*

Note that the reformulation of condition (ii) in (23) clarifies its intuitive meaning: it means that resource consumption measure $g_i^{\mathbf{S}}$ is monotone nondecreasing in the partition $\mathbf{S}$ within the domain $\mathscr{F}$, and that two nested partitions $\mathbf{S} \prec \mathbf{S}'$ in $\mathscr{F}$ give different resource consumption vectors $\left(g_i^{\mathbf{S}}\right)_{i\in\mathbb{X}}$ and $\left(g_i^{\mathbf{S}'}\right)_{i\in\mathbb{X}}$.

4. CONCLUSIONS

This paper has introduced a significant extension to the conventional multiarmed restless bandit problem, by allowing projects to have multiple operating modes instead of only two. A relevant application, concerning the optimal dynamic power allocation to multiple users sharing a wireless downlink communication channel subject to a peak energy constraint, has been presented. The paper has introduced a tractable heuristic policy based on indices defined individually for each project, and has further advanced the key results on the underlying theory and computation of such indices. The author is currently engaged in work aimed at testing the effectiveness of such results, believing them to be widely applicable to a variety of relevant resource allocation problems fitting into the new multiarmed multimode restless bandit problem framework.

Acknowledgments

The author's research has been supported in part by the Spanish Ministry of Education and Science under project MTM2007-63140 and an I3 faculty endowment grant, by the European Union's Network of Excellence Euro-FGI, and by the Autonomous Community of Madrid under grant CCG07-UC3M/ESP-3389.

5. REFERENCES

[1] E. Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Boca Raton, FL, 1999.

[2] B. Ata and K. E. Zachariadis. Dynamic power control in a fading downlink channel subject to an energy constraint. *Queueing Syst.*, 55:41–69, 2007.

[3] T. B. Crabill. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Sci.*, 18:560–566, 1972.

[4] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi, and I. Vincze, editors, *Progress in Statistics (European Meeting of Statisticians, Budapest, 1972)*, pp. 241–266. North-Holland, Amsterdam, The Netherlands, 1974.

[5] M. E. Mayorga, H.-S. Ahn, and J. G. Shanthikumar. Optimal control of a make-to-stock system with adjustable service rate. *Probab. Engrg. Inform. Sci.*, 20:609–634, 2006.

[6] J. Niño-Mora. Restless bandits, partial conservation laws and indexability. *Adv. Appl. Probab.*, 33:76–98, 2001.

[7] 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.*, 31:50–84, 2006.

[8] J. Niño-Mora. Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15:161–198, 2007. Followed by six discussions by I. J. B. F. Adan and O. J. Boxma, E. Altman, O. Hernández-Lerma, R. Weber, P. Whittle, and D. D. Yao.

[9] D. G. Pandelis and D. Teneketzis. On the optimality of the Gittins index rule for multi-armed bandits with multiple plays. *Math. Methods Oper. Res.*, 50:449–461, 1999.

[10] H. Sabeti. Optimal selection of service rates in queueing with different cost. *J. Operations Res. Soc. Japan*, 16:15–35, 1973.

[11] R. Serfozo. Optimal control of random walks, birth and death processes, and queues. *Adv. in Appl. Probab.*, 13:61–83, 1981.

[12] R. Weber. Comments on: "Dynamic priority allocation via restless bandit marginal productivity indices" [TOP 15:161–198, 2007] by J. Niño-Mora. *TOP*, 15:211–216, 2007.

[13] P. Whittle. Restless bandits: Activity allocation in a changing world. In J. Gani, editor, *A Celebration of Applied Probability*, spec. vol. 25A of *J. Appl. Probab.*, pp. 287–298. Applied Probability Trust, Sheffield, UK, 1988.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4410
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4410