

RESTLESS BANDITS, PARTIAL CONSERVATION LAWS AND INDEXABILITY

JOSÉ NIÑO-MORA,* *Universitat Pompeu Fabra, Barcelona*

Abstract

We show that if performance measures in a general stochastic scheduling problem satisfy partial conservation laws (PCL), which extend the generalized conservation laws (GCL) introduced by Bertsimas and Niño-Mora (1996), then the problem is solved optimally by a priority-index policy under a range of *admissible* linear performance objectives, with both this range and the optimal indices being determined by a one-pass adaptive-greedy algorithm that extends Klimov's: we call such scheduling problems *PCL-indexable*. We further apply the PCL framework to investigate the indexability property of restless bandits (two-action finite-state Markov decision chains) introduced by Whittle, obtaining the following results: (i) we present conditions on model parameters under which a single restless bandit is PCL-indexable, and hence indexable; membership of the class of PCL-indexable bandits is tested through a single run of the adaptive-greedy algorithm, which further computes the Whittle indices when the test is positive; this provides a tractable sufficient condition for indexability; (ii) we further introduce the subclass of *GCL-indexable* bandits (including classical bandits), which are indexable under arbitrary linear rewards. Our analysis is based on the achievable region approach to stochastic optimization, as the results follow from deriving and exploiting a new linear programming reformulation for single restless bandits.

Keywords: Stochastic scheduling; index policies; Markov decision chains; bandit problems; achievable region

AMS 2000 Subject Classification: Primary 90B35
Secondary 90C40

1. Introduction

The exact solution of stochastic scheduling problems, which involves designing a dynamic resource allocation policy in order to optimize a performance objective, appears to be, in most relevant models, an unreachable goal. Yet the identification of problem classes whose special structure yields a tractable solution remains of prime research interest: not only are such well-solved problems often of intrinsic interest, but their optimal solutions may provide building blocks for constructing well-grounded heuristics for more complex models. The latter situation is epitomized by Whittle's [17] pioneering approach to what is arguably the most promising extension of the classical multiarmed bandit problem: the *restless bandit problem*.

Both bandit models, classical and restless, are concerned with designing an optimal policy for sequential resource allocation to a collection of stochastic projects (the terms bandit and project will be used interchangeably in this paper), each being modelled as a finite Markov decision chain (MDC) having two actions available in each state: an *active* action (engaging

Received 20 December 1999; revision received 25 October 2000.

* Postal address: Department of Economics and Business, Universitat Pompeu Fabra, 08005 Barcelona, Spain.
Email address: jose.nino-mora@econ.upf.es

the project) and a *passive* action (letting it rest). Such models are paradigms of a fundamental conflict between taking actions that yield high current rewards, or instead taking actions that sacrifice current gains with the prospect of reaping better future returns. In the classical model, one project is to be engaged at each time, and passive projects do not change state. In the restless model, a fixed number of projects (possibly more than one) must be engaged at each time, and passive projects can change state. The performance objective of concern involves maximizing either the infinite horizon expected discounted reward, or the time-average reward rate.

While it is well known that the classical model is solved optimally by Gittins's [8] *priority-index policy* (an index is computed for each project state; then a project with larger current index is engaged at each time), its restless extension has been shown [13] to be PSPACE-hard, which most likely rules out the possibility of describing an optimal policy.

Yet the rich modelling power of restless bandits makes the development and analysis of a sound heuristic policy a problem of significant research interest: Whittle [17] has proposed applications in the areas of clinical trials, aircraft surveillance and worker scheduling. Other application areas reported in the literature include control of a make-to-stock queue [14] and behaviour coordination in robotics [7].

In his seminal paper on the subject, Whittle [17] presented a simple heuristic policy, together with an upper bound on the optimum problem value, both of which can be efficiently computed. Whittle's policy, like Gittins's, is a priority-index rule: an index is computed for each project state; one then engages at each time the required number of projects, say K out of the M available, with larger indices. The heuristic is grounded on the tractable optimal solution to a relaxed problem, whose optimum value gives the aforementioned bound: the sample-path constraint that K projects be active *at each time* is replaced by the relaxed constraint that K projects be active *on average*.

Whittle showed that such problem relaxation is solved optimally by a policy characterized by a set of indices attached to project states, provided each project in isolation satisfies a certain indexability property. For a given restless project, such a property refers to a parametric family of single-project subproblems, where all passive rewards (obtained when the passive action is chosen) are subsidized by a constant amount γ : a restless project is said to be *indexable* if, as the passive subsidy parameter γ grows from minus to plus infinity, the set of states where it is optimal to take the passive action increases monotonically from the empty set to the full state space. Under indexability, Whittle's priority index for a project's state is defined as the unique break-even value of γ for which both the active and the passive actions are optimal in that state. The appealing properties of the resulting index policy include the following: (i) it extends the Gittins optimal policy for classical bandits; (ii) the indices, like Gittins's, can be computed separately for each project; and (iii) it has asymptotic optimality, under regularity conditions, when K and M tend to infinity in a constant ratio, as established by Weber and Weiss [15], [16]. The fact that the Whittle index is only defined for such indexable projects motivated the development of an alternative index policy, which is always well defined [3], together with stronger performance bounds based on linear programming (LP).

Given a restless project, the tasks of testing whether it is indexable, and if so of computing its Whittle indices, can be efficiently accomplished through direct application of the definition of indexability, combined with the standard LP formulations for MDC introduced in [6], [11]: they involve n simplex pivot steps, carried out on an LP problem having $2n$ variables and n constraints, n being the number of project states. Such numerical procedure, however, does not provide any qualitative insight into which restless projects are indexable.

The above discussion motivates our main research goal: to identify and analyse relevant classes of indexable restless bandits, defined by testable conditions on model parameters. Whittle [17] stated that

...one would very much like to have simple sufficient conditions for indexability; at the moment, none are known.

To the best of our knowledge, no such indexability conditions have been obtained prior to those we present in this paper.

Our approach to the study of the indexability of restless bandits is based on the achievable region method (cf. [5]): we shall show that the indexability property can be explained in part as a consequence of a more general, underlying structural property on the system's polyhedral achievable performance region (i.e., the region spanned by the relevant performance measures under all admissible policies).

Specifically, our contribution in this paper is twofold:

1. We develop a general polyhedral framework for establishing the optimality of priority-index policies with special structure in stochastic scheduling problems under *some* linear performance objectives (*partial indexability*). This extends the framework of Bertsimas and Niño-Mora [1] for proving the optimality of arbitrary priority-index policies under *any* linear objective (*complete indexability*), which was based on the satisfaction by performance measures of *generalized conservation laws* (GCL). The new framework is based on the satisfaction of a relaxed version of the GCL, which we call *partial conservation laws* (PCL). We show that under PCL the problem is solved optimally by a priority-index policy with the required structure under a range of admissible linear performance objectives, with both the optimal indices and the admissible linear objectives being characterized by a one-pass adaptive-greedy algorithm that extends Klimov's [10]. We call such scheduling problems *PCL-indexable*. The former framework [1] was based on obtaining a complete polyhedral characterization of the system's achievable performance region, resulting from the satisfaction of GCL. The extension we present here is based instead on exploiting a partial polyhedral characterization of the achievable performance region, arising from the satisfaction of PCL.
2. We apply the PCL framework to investigate the indexability property of restless bandits, obtaining the following results: (i) we identify a class of restless bandits (PCL-indexable) that are partly indexable; membership in the class is tested through a single run of the adaptive-greedy algorithm, which further computes the Whittle indices when the test is positive; (ii) we further identify the subclass of *GCL-indexable* bandits, of which classical bandits are the main example, which are completely indexable. Our analysis is based on deriving and exploiting a new LP reformulation of the standard LP formulation for single restless bandits.

The use of LP formulations to analyse restless bandits can be regarded as an extension of previous work where LP formulations have played a major role in the study of classical multiarmed bandits (cf. [4], [9], [1]).

The rest of the paper is structured as follows: in Section 2, we describe the restless bandit problem, and review Whittle's relaxation and index heuristic. In Section 3, we present the general PCL framework for partial indexability. The PCL framework is applied to the analysis of the indexability property in discounted restless bandits in Section 4. The corresponding analysis under the time-average criterion is developed in Section 5. Section 6 presents some examples where the previous results are applied. Finally, Section 7 ends the paper with some concluding remarks.

2. Whittle's relaxation and index heuristic

In this section, we formulate the restless bandit problem of concern, and review the relaxation and the heuristic index policy proposed by Whittle [17]. Consider the problem faced by a decision maker seeking to maximize the average reward earned from a collection of M stochastic projects, of which $1 \leq K < M$ must be engaged at each discrete time epoch $t \geq 0$. Project m is modelled as an MDC evolving through a finite state space N_m , and having two actions $a \in \{0, 1\}$ available in each state $i \in N_m$, for $1 \leq m \leq M$. We shall assume, for convenience of notation, that the state spaces for the M projects are disjoint, and denote their *aggregate state space* by $N = \bigcup_{m=1}^M N_m$. The *active* ($a = 1$) and *passive* ($a = 0$) actions correspond to engaging a project or letting it rest, respectively. Taking action $a \in \{0, 1\}$ on a project in state i has two effects: first, it yields a current reward R_i^a ; then, it causes the state to evolve in a Markovian fashion, moving at the next time epoch into state j with probability p_{ij}^a . We write $\mathbf{R}^a = (R_i^a)_{i \in N}$ and $\mathbf{P}^a = (p_{ij}^a)_{i, j \in N}$, for $a \in \{0, 1\}$. Both the time-discounted and the time-average criteria will be considered.

Under the discounted criterion, rewards are time-discounted by a factor $0 < \beta < 1$, and the problem consists in finding a *scheduling policy*, u^{opt} , belonging in the space \mathcal{U} of stationary policies (which base decisions on current project states), that maximizes the expected net present value of rewards earned over an infinite horizon:

$$Z^{\text{opt}}(\beta) = \max_{u \in \mathcal{U}} \mathbb{E}_u \left[\sum_{t=0}^{\infty} (R_{i_1(t)}^{a_1(t)} + \dots + R_{i_M(t)}^{a_M(t)}) \beta^t \right].$$

Here, $Z^{\text{opt}}(\beta)$ denotes the optimum problem value, $i_m(t)$ and $a_m(t)$ denote the state and the action corresponding to project m at time t , respectively, and $\mathbb{E}_u[\cdot]$ represents the expectation operator under policy u . Such expectation is conditional on initial project states, as given by a known vector $\boldsymbol{\alpha} = (\alpha_i)_{i \in N}$, where

$$\alpha_i = \begin{cases} 1 & \text{if a project is initially in state } i, \\ 0 & \text{otherwise.} \end{cases}$$

Under the time-average criterion, we are concerned with finding a stationary scheduling policy u^{opt} that maximizes the long-run time-average reward rate (which is well defined under suitable regularity conditions):

$$Z^{\text{opt}}(1) = \max_{u \in \mathcal{U}} \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_u \left[\sum_{t=0}^T (R_{i_1(t)}^{a_1(t)} + \dots + R_{i_M(t)}^{a_M(t)}) \right].$$

Note that here we have denoted the optimal problem value by $Z^{\text{opt}}(1)$, thus identifying, for notational convenience, the time-average criterion with the value $\beta = 1$ in the notation $Z^{\text{opt}}(\beta)$ for the discounted case. Using that convention will allow us to discuss below the time-discounted and the time-average cases in parallel in a common framework.

Whittle proposed the following relaxed version of the problem: the sample-path requirement that K projects be active *at each time* is relaxed by requiring instead that K projects be active *on average*. The optimum value of the relaxed problem is hence an upper bound for that of the original problem. Furthermore, such bound can be efficiently computed by solving a polynomial-size linear program (cf. [3]). To see this, let us associate a *performance measure*,

$x_i^a(\beta, u)$, with each stationary scheduling policy $u \in \mathcal{U}$, project state $i \in N$, and action $a \in \{0, 1\}$, defined in the discounted case by

$$x_i^a(\beta, u) = \mathbb{E}_u \left[\sum_{t=0}^{\infty} I_i^a(t) \beta^t \right], \quad (2.1)$$

and in the time-average case by

$$x_i^a(1, u) = \lim_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}_u \left[\sum_{t=0}^T I_i^a(t) \right],$$

where

$$I_i^a(t) = \begin{cases} 1 & \text{if action } a \text{ is taken at a project in state } i \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the performance measure $x_i^a(\beta, u)$ is a standard state-action frequency measure in MDC theory, representing the total expected discounted number of times, or, when $\beta = 1$, the expected long-run fraction of time, that action a is taken at a project in state i under policy u . It follows directly from the standard LP formulations for discounted and time-average MDC (introduced by d'Épéroux [6] and by Manne [11]), respectively) that Whittle's relaxation can be formulated as the LP problem

$$Z^W(\beta) = \max \sum_{j \in N} R_j^0 x_j^0 + \sum_{j \in N} R_j^1 x_j^1 \quad (2.2)$$

subject to

$$(x_j^0, x_j^1)_{j \in N_m} \in \mathcal{P}_m(\beta), \quad 1 \leq m \leq M,$$

$$\sum_{j \in N} x_j^1 = Ks(\beta), \quad (2.3)$$

where

$$s(\beta) = \begin{cases} \frac{1}{1-\beta} & \text{if } 0 < \beta < 1, \\ 1 & \text{if } \beta = 1, \end{cases}$$

$$\mathcal{P}_m(\beta) = \left\{ (x_j^0, x_j^1)_{j \in N_m} \geq \mathbf{0} : \sum_{a \in \{0,1\}} \left[x_j^a - \beta \sum_{i \in N_m} p_{ij}^a x_i^a \right] = \alpha_j, \quad j \in N_m \right\},$$

for $0 < \beta < 1$, is the bounded polyhedron (i.e., the polytope) defined by the standard LP constraints for the discounted MDC modelling project m , and

$$\mathcal{P}_m(1) = \left\{ (x_j^0, x_j^1)_{j \in N_m} \geq \mathbf{0} : \sum_{a \in \{0,1\}} \left[x_j^a - \sum_{i \in N_m} p_{ij}^a x_i^a \right] = 0, \quad j \in N_m, \quad \sum_{a \in \{0,1\}, j \in N_m} x_j^a = 1 \right\}$$

is the corresponding polytope for the time-average case. In (2.2), the variables x_j^a correspond to performance measures $x_j^a(\beta, u)$, and the constraint (2.3) formulates the relaxed requirement that K projects be active on average. Note that this LP problem has polynomial size on the problem's defining data, and is therefore known to be solvable in polynomial time by LP interior point algorithms.

Whittle applied instead a Lagrangian approach to elucidate the structure of the relaxed problem's optimal solution. By dualizing the linking constraint (2.3) in (2.2), denoting by γ the corresponding Lagrange multiplier, and using the implied constraint

$$\sum_{j \in N} x_j^0 + \sum_{j \in N} x_j^1 = Ms(\beta),$$

we obtain the *Lagrangian relaxation*

$$\mathcal{L}(\beta, \gamma) = \max \sum_{j \in N} (R_j^0 + \gamma)x_j^0 + \sum_{j \in N} R_j^1 x_j^1 - (M - K)s(\beta)\gamma \quad (2.4)$$

subject to

$$(x_j^0, x_j^1)_{j \in N_m} \in \mathcal{P}_m(\beta), \quad 1 \leq m \leq M.$$

We thus see that, as observed by Whittle, the multiplier γ plays the economic role of a constant subsidy for passivity. Problem (2.4) is naturally decoupled into the M single-project subproblems

$$\mathcal{L}_m(\beta, \gamma) = \max \sum_{j \in N_m} (R_j^0 + \gamma)x_j^0 + \sum_{j \in N_m} R_j^1 x_j^1 \quad (2.5)$$

subject to

$$(x_j^0, x_j^1)_{j \in N_m} \in \mathcal{P}_m(\beta),$$

for $m = 1, \dots, M$, so that

$$\mathcal{L}(\beta, \gamma) = \sum_{m=1}^M \mathcal{L}_m(\beta, \gamma) - (M - K)s(\beta)\gamma.$$

Note that the subproblem (2.5) is the LP formulation for the MDC corresponding to a modified version of project m , where passive rewards (earned under the passive action) are subsidized by the amount γ . Note further that, for each value of the multiplier/passive subsidy γ , $\mathcal{L}(\beta, \gamma) \geq Z^W(\beta)$. Now, strong LP duality yields that there exists a multiplier γ^* (which, in the discounted case, depends on the initial state vector α) such that $\mathcal{L}(\beta, \gamma^*) = Z^W(\beta)$. In the regular case where $\gamma^* \neq 0$, LP complementary slackness ensures that *any* optimal solution to Lagrangian relaxation (2.4) (with $\gamma = \gamma^*$) must satisfy the linking constraint (2.3), and will therefore also be optimal for the LP formulation of Whittle's relaxed problem in (2.2): in that case ($\gamma^* \neq 0$), subsidizing passive actions by the amount γ^* causes the decision maker to engage K projects on average when acting optimally.

Whittle identified a key property that makes the solution to the family of single-project subproblems in (2.5) depend on the parameter γ in a particularly simple fashion.

Definition 1. (Whittle [17].) A project is said to be *indexable* for a given discount factor $0 < \beta \leq 1$ if the set of states where the passive action is optimal in the single-project subproblem (2.5) increases monotonically from the empty set to the full set of states as the passive subsidy γ increases from $-\infty$ to $+\infty$.

It follows from Definition 1 that, for an indexable project, there exist break-even indices γ_i for each state i , such that an optimal policy for the corresponding subproblem (2.5) can be given

as follows: take the passive action in states i with $\gamma_i < \gamma$, and the active action otherwise. Note that for $\gamma = \gamma_i$ both the active and the passive actions are optimal in state i . If each project is indexable, and $\gamma^* \neq 0$, it follows from the above that an optimal policy for Whittle's relaxation is obtained by applying independently to each project the single-project policy just described, letting $\gamma = \gamma^*$.

Whittle proposed to use the indices γ_i as priority indices to define a heuristic policy for the original problem, as follows: at each time K engage projects with larger indices. Note that it follows from their definition that Whittle's indices reduce to Gittins's when applied to classical multiarmed bandits, and therefore the heuristic is optimal in that special case (where $K = 1$).

3. Partial conservation laws

In this section we develop a general framework for investigating the partial indexability property in stochastic scheduling problems, outlined in Section 1. This framework extends that introduced in [1] for studying the complete indexability property in stochastic scheduling.

Consider a general dynamic and stochastic service system catering to a finite set $N = \{1, \dots, n\}$ of job classes. Service resources (e.g., servers) are to be allocated over time to jobs vying for their attention, on the basis of a *scheduling policy* u , which belongs in a space \mathcal{U} of *admissible policies*. The performance of a policy $u \in \mathcal{U}$ over a job class $i \in N$ is evaluated by a performance measure $x_i(u) \geq 0$, which we assume to be an expectation. We denote by $\mathbf{x}(u) = (x_i(u))_{i \in N}$ the corresponding performance vector. We further assume the system admits a consistent notion of service priority among classes: to each ordered string $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ spanning all the classes is associated a corresponding *$\boldsymbol{\pi}$ -priority policy*, which assigns higher priority to class π_i over class π_j if $i < j$, so that class π_1 has top priority. We refer to all such policies as *priority policies*. We further say that a policy gives priority to classes in a subset $S \subseteq N$ (*S-jobs*) if it gives priority to any job class $i \in S$ over any job class $j \in S^c = N \setminus S$.

Consider now, for a given *reward vector* $\mathbf{R} = (R_i)_{i \in N} \in \mathbb{R}^n$, the *optimal scheduling problem*

$$Z^{\text{opt}}(\mathbf{R}) = \max \left\{ \sum_{i \in N} R_i x_i(u) : u \in \mathcal{U} \right\}, \quad (3.1)$$

which involves finding an admissible scheduling policy that maximizes the linear performance objective in (3.1), and computing the corresponding optimum value $Z^{\text{opt}}(\mathbf{R})$.

A wide variety of scheduling problems fitting (3.1), such as classical multiarmed bandits, possess the following structural property, which we call *complete indexability*: to each reward vector $\mathbf{R} \in \mathbb{R}^n$ there corresponds an index vector $\boldsymbol{\gamma}(\mathbf{R}) = (\gamma_i(\mathbf{R}))_{i \in N}$ such that the corresponding priority-index policy (which gives higher priority to classes with larger indices) is optimal for problem (3.1). A general framework providing a sufficient condition for problem (3.1) to be completely indexable was presented in [1]: satisfaction by a performance vector $\mathbf{x}(u)$ of GCL implies complete indexability; furthermore, under GCL the optimal indices can be efficiently computed through an n -step adaptive-greedy algorithm first introduced by Klimov [10] (in his landmark derivation of the optimal priority-index policy for a multiclass $M/G/1$ queue with feedback, under the time-average criterion).

In the context of certain models, however, complete indexability appears as too strong a requirement: the relevant concern may be instead to establish the optimality of a restricted family of priority policies under a limited range of linear performance objectives. Namely, the optimal index vector $\boldsymbol{\gamma}(\mathbf{R})$ is defined only over a domain $\mathcal{D} \subseteq \mathbb{R}^n$ of admissible rewards \mathbf{R} , and

it yields priorities having a certain special structure. We call this property *partial indexability*. Such is the case, e.g., in models involving the control of arrivals into a single queue, where researchers typically aim to establish the optimality of threshold policies (shut-off arrivals when the queue length is above a threshold value) under strong structural assumptions on reward/cost coefficients. More generally, as we shall demonstrate in Sections 4 and 5, the problem of determining whether a restless bandit is indexable may be formulated in terms of checking the partial indexability of a problem of the form (3.1).

We present next a general framework for establishing partial indexability, based on the satisfaction by performance measures of a relaxed version of the GCL, that is, PCL. We shall represent a given family of priority policies with special structure as a *set system* (N, \mathcal{F}) defined over the ground set N of job classes, so that $\mathcal{F} \subseteq 2^N$ is the family of job class subsets that may be assigned *higher* priority under the policy family of concern. It will be convenient to consider also the complementary family $\mathcal{F}^c = \{S^c : S \in \mathcal{F}\}$ of job class subsets that may receive *lower* priority. We impose on the set system (N, \mathcal{F}) the requirements stated next, which are motivated by algorithmic considerations.

Assumption 1. *The set system (N, \mathcal{F}) satisfies the following conditions:*

- (i) (Accessibility.) *If $\emptyset \neq S \in \mathcal{F}$, then there exists $i \in S$ such that $S \setminus \{i\} \in \mathcal{F}$.*
- (ii) (Augmentability.) *If $N \neq S \in \mathcal{F}$, then there exists $i \in S^c$ such that $S \cup \{i\} \in \mathcal{F}$.*

Note that if (N, \mathcal{F}) satisfies Assumption 1, then so does its complementary set system (N, \mathcal{F}^c) ; it further follows that \emptyset and N are members of both \mathcal{F} and \mathcal{F}^c .

Definition 2. (*Partial conservation laws.*) The performance vector $\mathbf{x}(u)$ satisfies *partial conservation laws* with respect to the set system (N, \mathcal{F}) if there exist coefficients $A_i^S > 0$ for $i \in S$ and $S \in \mathcal{F}^c$ such that, letting

$$b(S) = \inf \left\{ \sum_{i \in S} A_i^S x_i(u) : u \in \mathcal{U} \right\}, \quad S \in \mathcal{F}^c,$$

the following identities hold: for $S \in \mathcal{F}^c \setminus \{N\}$,

$$\sum_{i \in S} A_i^S x_i(u) = b(S),$$

under any priority policy u giving priority to S^c -jobs; and, for $S = N$,

$$\sum_{i \in N} A_i^N x_i(u) = b(N),$$

under any policy $u \in \mathcal{U}$.

Note that the GCL in [1] correspond to the special case of the PCL above where $\mathcal{F} = 2^N$. In words, a performance vector $\mathbf{x}(u)$ satisfies PCL with respect to a set system (N, \mathcal{F}) if, for each low priority class subset $S \in \mathcal{F}^c$, there exist weights $A_i^S > 0$, for $i \in S$, such that the corresponding weighted performance objective $\sum_{i \in S} A_i^S x_i(u)$ is minimized by any priority policy that gives priority to S^c -jobs, and is invariant under all admissible policies when $S = N$. The PCL thus state that the family of priority policies that give higher priority to S -jobs, for $S \in \mathcal{F}$, minimizes a certain finite set of linear performance objectives. As we shall demonstrate in what follows, the PCL further imply the optimality of such priority policies for a larger family of linear objectives, where the optimal priorities are determined by efficiently computed class-ranking indices.

For a scheduling problem that satisfies the PCL in Definition 2, consider the LP problem

$$\begin{aligned}
 Z^{\text{LP}}(\mathbf{R}) &= \max \sum_{i \in N} R_i x_i & (3.2) \\
 &\text{subject to} \\
 y^S &: \sum_{i \in S} A_i^S x_i \geq b(S), & S \in \mathcal{F}^c \setminus \{N\}, \\
 y^N &: \sum_{i \in N} A_i^N x_i = b(N), \\
 x_i &\geq 0, & i \in N,
 \end{aligned}$$

where we have associated with each constraint a corresponding dual variable y^S . For any reward vector $\mathbf{R} \in \mathbb{R}^n$, the PCL imply that (3.2) is an *LP relaxation* of the scheduling problem (3.1), so that $Z^{\text{LP}}(\mathbf{R}) \geq Z^{\text{opt}}(\mathbf{R})$. We shall identify next a range of reward vectors \mathbf{R} for which (3.2) is an *exact* LP formulation of (3.1), in that $Z^{\text{LP}}(\mathbf{R}) = Z^{\text{opt}}(\mathbf{R})$. Let us thus define the region $\mathcal{D}(\mathcal{F}) \subseteq \mathbb{R}^n$ of *admissible rewards* as the domain of the adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$ described in the Appendix. Namely, $\mathcal{D}(\mathcal{F})$ is the set of rewards \mathbf{R} for which the algorithm returns an output having $\text{ADMISSIBLE} = \text{YES}$ when fed with input \mathbf{R} . Note that $\mathcal{D}(2^N) = \mathbb{R}^n$.

The next definition draws on the notion of PCL to define a class of scheduling problems that, as will be shown in Theorem 1 below, are partly indexable.

Definition 3. (*PCL/GCL-indexability.*) The scheduling problem (3.1) is said to be *PCL-indexable with respect to a set system* (N, \mathcal{F}) if

- (i) it satisfies PCL with respect to (N, \mathcal{F}) ; and
- (ii) $\mathbf{R} \in \mathcal{D}(\mathcal{F})$.

If (i) holds with $\mathcal{F} = 2^N$, we say the problem is *GCL-indexable*.

The next result shows that PCL-indexable problems are indeed indexable. Let the output of the adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$ corresponding to input \mathbf{R} be given by the triple $(\text{ADMISSIBLE}, \boldsymbol{\gamma}, \boldsymbol{\pi})$. Note that if $\text{ADMISSIBLE} = \text{YES}$ then the algorithm returns an index vector $\boldsymbol{\gamma} = (\gamma_i)_{i \in N}$ and an ordered string $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ spanning all the classes with

$$\gamma_{\pi_1} \geq \dots \geq \gamma_{\pi_n}.$$

Theorem 1. (*Indexability under PCL.*) *Assume that (3.1) is PCL-indexable with respect to (N, \mathcal{F}) . Then, the problem is solved optimally by any priority policy that gives higher priority to class i over class j if $\gamma_i \geq \gamma_j$, for $i, j \in N$.*

Proof. The dual of the LP problem (3.2) is

$$\begin{aligned}
 Z^{\text{LP}}(\mathbf{R}) &= \min \sum_{S \in \mathcal{F}^c} b(S) y^S \\
 &\text{subject to} \\
 \sum_{i \in S \in \mathcal{F}^c} A_i^S y^S &\geq R_i, & i \in N, \\
 y^S &\leq 0, & S \in \mathcal{F}^c \setminus \{N\}.
 \end{aligned}$$

Now, it is easily checked that the vector \bar{y} computed in the course of the adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$ is a feasible solution to the dual of the LP problem, which further satisfies

$$R_{\pi_i} = \sum_{k=1}^i A_{\pi_i}^{S_k} \bar{y}^{S_k}, \quad i \in N,$$

where $S_k = \{\pi_k, \dots, \pi_n\}$ (note that $S_1 = N$) and

$$\bar{y}^{S_1} = \gamma_{\pi_1}, \quad \bar{y}^{S_k} = \gamma_{\pi_k} - \gamma_{\pi_{k-1}}, \quad 2 \leq k \leq n.$$

It thus follows that the performance objective can be expressed, under any policy $u \in \mathcal{U}$, as

$$\sum_{i \in N} R_i x_i(u) = \gamma_{\pi_1} \sum_{i \in S_1} A_i^{S_1} x_i(u) + \sum_{k=2}^n (\gamma_{\pi_k} - \gamma_{\pi_{k-1}}) \sum_{i \in S_k} A_i^{S_k} x_i(u).$$

The fact that the indicated priority-index policy is optimal follows directly from this last identity, the PCL in Definition 2, and the fact that $\gamma_{\pi_1} \geq \dots \geq \gamma_{\pi_n}$. This completes the proof.

Note that in the special case where $\mathcal{F} = 2^N$, the algorithm $\text{AG}(\cdot \mid \mathcal{F})$ reduces to the well-known adaptive-greedy algorithm due to Klimov [10], and Theorem 1 then yields the optimality of priority-index policies under arbitrary rewards, as established in the GCL framework (cf. [1]).

3.1. Project scheduling and index decomposition

It was shown in [1] that in the GCL framework ($\mathcal{F} = 2^N$ in the PCL) a stronger index decomposition property holds under certain conditions. That property explains that in certain scheduling models with multiple projects, such as classical multiarmed bandits, the priority indices for a project depend only on its defining parameters, and not on those of other projects. See [8], [9]. An analogous index decomposition property (which will be applied in Section 4) holds in the PCL framework, as outlined next. Consider the special case of the general model above which corresponds to the problem of scheduling a collection of m projects, where project k has state space N_k , for $1 \leq k \leq m$. Assume that the project state spaces are disjoint, and consider the aggregate state space $N = \bigcup_{k=1}^m N_k$. Note that in this setting, project states play the role of job classes in the general framework above. Suppose that it can be established that a performance vector $\mathbf{x}(u) = (x_i(u))_{i \in N}$ for the overall model satisfies PCL with respect to set system (N, \mathcal{F}) , with parameters A_i^S . Let

$$\mathcal{F}_k = \{S \in \mathcal{F} : S \subseteq N_k\} \quad \text{and} \quad \mathcal{F}_k^c = \{S \in 2^{N_k} : N_k \setminus S \in \mathcal{F}_k\}, \quad 1 \leq k \leq m,$$

and suppose that the conditions given next hold.

Assumption 2. For $1 \leq k \leq m$,

- (i) $\mathcal{F}_k = \{S \cap N_k : S \in \mathcal{F}\}$;
- (ii) the set system (N_k, \mathcal{F}_k) satisfies Assumption 1;
- (iii) $A_i^{S \cap N_k} = A_i^S$, $i \in S \cap N_k$, $S \in \mathcal{F}^c$.

A simple extension to the argument given to prove Theorem 6 in [1] yields the following result. Let $(\text{ADMISSIBLE}_k, \boldsymbol{\gamma}^k, \boldsymbol{\pi}^k)$ be the output of the adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F}_k)$ on input $\mathbf{R}^k = (R_i)_{i \in N_k}$, for $1 \leq k \leq m$, and let $(\text{ADMISSIBLE}, \boldsymbol{\gamma}, \boldsymbol{\pi})$ be the output of adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$ on input $\mathbf{R} = (R_i)_{i \in N}$.

Theorem 2. (Index decomposition.) *Under Assumption 2:*

- (i) $\text{ADMISSIBLE} = \text{YES}$ if and only if $\text{ADMISSIBLE}_k = \text{YES}$ for $1 \leq k \leq m$. If $\text{ADMISSIBLE} = \text{YES}$, then:
- (ii) $\gamma_i = \gamma_i^k$, for $1 \leq k \leq m$; and
- (iii) it is optimal to give higher priority to projects whose current states have larger indices.

4. PCL for restless bandits: discounted criterion

In this section we apply the PCL framework developed in Section 3 to identify sufficient conditions for the restless bandit's indexability property introduced by Whittle. We focus here on the discounted criterion, which will be the basis for our treatment of the time-average criterion in the next section.

We thus consider a single restless bandit, as described in Section 2: it is modelled as a discrete-time MDC, having state space $N = \{1, \dots, n\}$, transition probability matrices $\mathbf{P}^a = (p_{ij}^a)_{i, j \in N}$, and reward vectors $\mathbf{R}^a = (R_i^a)_{i \in N}$, corresponding to the active ($a = 1$) and passive ($a = 0$) actions, respectively. Rewards are discounted in time by the factor $0 < \beta < 1$. The initial state information is given by the 0/1 indicator vector $\boldsymbol{\alpha} = (\alpha_i)_{i \in N}$. Our concern is to identify sufficient conditions on model parameters under which the bandit is indexable. We thus need to investigate the parametric family of MDC subproblems whose LP formulation was given in (2.5). We write them here more explicitly as

$$Z^{\text{opt}}(\gamma; \mathbf{R}^0, \mathbf{R}^1) = \max_{u \in \mathcal{U}} \mathbb{E}_u \left[\sum_{t=0}^{\infty} \sum_{i \in N} ((R_i^0 + \gamma)I_i^0(t) + R_i^1 I_i^1(t)) \beta^t \right], \quad (4.1)$$

for each value of the passive subsidy parameter $\gamma \in \mathbb{R}$. In (4.1), $I_i^a(t)$ represents the 0/1 indicator corresponding to taking action $a \in \{0, 1\}$ in state i at time t , and \mathcal{U} denotes the space of stationary policies. The standard LP formulation of problem (4.1), which was given in (2.5), can be expressed using vector notation as

$$\begin{aligned} Z^{\text{opt}}(\gamma; \mathbf{R}^0, \mathbf{R}^1) &= \max \mathbf{x}^0(\mathbf{R}^0 + \gamma \mathbf{1}) + \mathbf{x}^1 \mathbf{R}^1 & (4.2) \\ &\text{subject to} \\ &\mathbf{x}^0(\mathbf{I} - \beta \mathbf{P}^0) + \mathbf{x}^1(\mathbf{I} - \beta \mathbf{P}^1) = \boldsymbol{\alpha} \\ &\mathbf{x}^0, \mathbf{x}^1 \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{x}^a = (x_j^a)_{j \in N}$, for $a \in \{0, 1\}$, and $\boldsymbol{\alpha}$ are taken to be row vectors, $\mathbf{1}$ denotes an n -vector of ones, and \mathbf{I} is the identity matrix. Note further that the constraints of the LP problem (4.2) imply that

$$\mathbf{x}^0 = \boldsymbol{\alpha}(\mathbf{I} - \beta \mathbf{P}^0)^{-1} - \mathbf{x}^1(\mathbf{I} - \beta \mathbf{P}^1)(\mathbf{I} - \beta \mathbf{P}^0)^{-1}.$$

An immediate consequence of this observation and the LP formulation (4.2) is the identity

$$Z^{\text{opt}}(\gamma; \mathbf{R}^0, \mathbf{R}^1) = \boldsymbol{\alpha}(\mathbf{I} - \beta \mathbf{P}^0)^{-1} \mathbf{R}^0 + Z^{\text{opt}}(\gamma; \mathbf{0}, \mathbf{R}^1 - (\mathbf{I} - \beta \mathbf{P}^1)(\mathbf{I} - \beta \mathbf{P}^0)^{-1} \mathbf{R}^0).$$

In light of this identity, we shall focus our subsequent analysis on the normalized case where $\mathbf{R}^0 = \mathbf{0}$, without loss of generality.

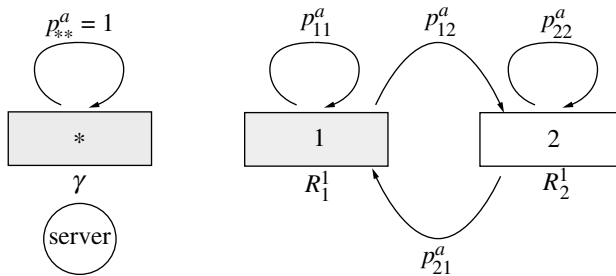


FIGURE 1: A single restless bandit seen as a multiclass service system.

In order to apply the PCL framework to analyse (4.1), we shall reformulate it as an equivalent scheduling problem over a service system with multiple job classes. The latter problem involves the scheduling of two projects on a single server: to the original project we add an auxiliary single-state calibrating project, which returns a reward of γ when engaged, and no reward otherwise. Engaging this auxiliary project corresponds to resting the original project.

We thus identify a class i job with a project in state $i \in N_* = \{*\} \cup N$ (where the label of the calibrating project's single state is denoted by $*$). To each policy $u \in \mathcal{U}$ we associate performance measures $x_i^a(u)$, for $i \in N_*$ and $a \in \{0, 1\}$, as defined in (2.1). Note that $x_*^1(u)$ represents the total expected discounted time that the original project is passive (equivalently, that the calibrating single-state project is active).

Figure 1 represents a simple example, where the original project has two states ($N = \{1, 2\}$), and is currently in state 1: that is, there is currently one job of class $*$ and another of class 1.

It will be convenient in what follows to use the following additional notation: given a vector $\mathbf{x} = (x_i)_{i \in N}$, a matrix $\mathbf{P} = (p_{ij})_{i,j \in N}$, and subsets $S, T \subseteq N$, we shall write $\mathbf{x}_S = (x_i)_{i \in S}$ and $\mathbf{P}_{ST} = (p_{ij})_{i \in S, j \in T}$. Recall that $S^c = N \setminus S$, for $S \subseteq N$.

We next define certain project parameters, derived from model primitives, which will be required in our analysis. We start by considering, for each class/state subset $S \subseteq N$, a corresponding S -active policy: this takes the active action on the original project when its state lies in S , and the passive action otherwise. Let us further define T_i^S , for $i \in N$, as the total expected discounted time the project state lies in S under the S -active policy, provided the initial state is i . For a given $S \subseteq N$, the times T_i^S are determined as the unique solution to the system of linear equations

$$\begin{aligned} T_i^S &= 1 + \beta \sum_{j \in N} p_{ij}^1 T_j^S, & i \in S, \\ T_i^S &= \beta \sum_{j \in N} p_{ij}^0 T_j^S, & i \in S^c. \end{aligned}$$

It will be convenient for our analysis to rewrite this system, using the matrix notation introduced above, as

$$\begin{aligned} \mathbf{T}_S^S &= \mathbf{1}_S + \beta \mathbf{P}_{SS}^1 \mathbf{T}_S^S + \beta \mathbf{P}_{SS^c}^1 \mathbf{T}_{S^c}^S, \\ \mathbf{T}_{S^c}^S &= \beta \mathbf{P}_{S^c S}^0 \mathbf{T}_S^S + \beta \mathbf{P}_{S^c S^c}^0 \mathbf{T}_{S^c}^S, \end{aligned} \tag{4.3}$$

where $\mathbf{1}_S$ denotes a vector of ones indexed by classes/states in S , $\mathbf{T}_S^S = (T_i^S)_{i \in S}$, and $\mathbf{T}_{S^c}^S = (T_i^S)_{i \in S^c}$.

We next use the times T_i^S as building blocks to define quantities A_i^S , for $i \in N$ and $S \subseteq N$, by

$$A_i^{S^c} = 1 + \beta \sum_{j \in N} (p_{ij}^1 - p_{ij}^0) T_j^S. \quad (4.4)$$

Note that

$$A_i^\emptyset = A_i^N = 1, \quad i \in N. \quad (4.5)$$

Furthermore, it is straightforward from (4.3) and (4.4) that

$$\begin{aligned} A_S^S &= \mathbf{1}_S + \beta \mathbf{P}_{SN}^1 \mathbf{T}_N^{S^c} - \mathbf{T}_S^{S^c}, \\ A_{S^c}^{S^c} &= \mathbf{T}_{S^c}^{S^c} - \beta \mathbf{P}_{S^c N}^0 \mathbf{T}_N^{S^c}, \end{aligned} \quad (4.6)$$

where $A_S^S = (A_i^S)_{i \in S}$ and $A_{S^c}^{S^c} = (A_i^{S^c})_{i \in S^c}$. We further define $A_i^{\{*\} \cup S}$, for $i \in \{*\} \cup S$ and $S \subseteq N$, by

$$A_i^{\{*\} \cup S} = \begin{cases} A_i^S & \text{if } i \in S, \\ 1 & \text{if } i = *. \end{cases}$$

We complete the definitions by letting $b(S_*)$, for $S_* \subseteq N_*$, be given by

$$b(S_*) = \begin{cases} \frac{1}{1 - \beta} - \sum_{i \in N} \alpha_i T_i^{S^c} & \text{if } * \in S_* \text{ and } S = S_* \cap N \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Our next result will play a central role in our analysis of the indexability property of restless bandits via PCL. It formulates a family of *decomposition laws*, where a linear combination of active performance measures (the $x_i^1(u)$) is shown to decompose, under any admissible policy, as the sum of a policy-invariant term plus a linear combination of passive performance measures (the $x_i^0(u)$). We note that such identities are analogous to the work decomposition laws satisfied by certain single-server multiclass queueing systems (cf. [2, Theorem 5]).

Theorem 3. (Decomposition laws.) *Under any admissible policy u and for any $S \subseteq N$,*

$$x_*^1(u) + \sum_{i \in S} A_i^S x_i^1(u) = b(\{*\} \cup S) + \sum_{i \in S^c} A_i^S x_i^0(u). \quad (4.7)$$

Proof. To simplify notation, we shall write in what follows $\mathbf{x}^a(u) = \mathbf{x}^a$, for $a = 0, 1$, and consider the \mathbf{x}^a to be a row vector. We first note that the standard LP formulation for discounted MDC, applied to the restless project being studied, yields that the performance vectors \mathbf{x}^a , for $a = 0, 1$, satisfy the matrix equation

$$\mathbf{x}^0(\mathbf{I} - \beta \mathbf{P}^0) + \mathbf{x}^1(\mathbf{I} - \beta \mathbf{P}^1) = \boldsymbol{\alpha}.$$

We now rewrite this system in terms of a given subset $S \subset N$, as

$$\begin{aligned} \begin{bmatrix} \mathbf{x}_S^0 & \mathbf{x}_{S^c}^0 \end{bmatrix} \begin{bmatrix} \mathbf{I}_S - \beta \mathbf{P}_{SS}^0 & -\beta \mathbf{P}_{SS^c}^0 \\ -\beta \mathbf{P}_{S^c S}^0 & \mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^0 \end{bmatrix} + \begin{bmatrix} \mathbf{x}_S^1 & \mathbf{x}_{S^c}^1 \end{bmatrix} \begin{bmatrix} \mathbf{I}_S - \beta \mathbf{P}_{SS}^1 & -\beta \mathbf{P}_{SS^c}^1 \\ -\beta \mathbf{P}_{S^c S}^1 & \mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^1 \end{bmatrix} \\ = \begin{bmatrix} \boldsymbol{\alpha}_S & \boldsymbol{\alpha}_{S^c} \end{bmatrix}, \end{aligned}$$

or, equivalently,

$$\begin{aligned} \mathbf{x}_S^0(\mathbf{I}_S - \beta \mathbf{P}_{SS}^0) &= \boldsymbol{\alpha}_S + \beta \mathbf{x}_{S^c}^0 \mathbf{P}_{S^c S}^0 + \beta \mathbf{x}_{S^c}^1 \mathbf{P}_{S^c S}^1 - \mathbf{x}_S^1(\mathbf{I}_S - \beta \mathbf{P}_{SS}^1), \\ \mathbf{x}_{S^c}^1(\mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^1) &= \boldsymbol{\alpha}_{S^c} + \beta \mathbf{x}_S^0 \mathbf{P}_{SS^c}^0 + \beta \mathbf{x}_S^1 \mathbf{P}_{SS^c}^1 - \mathbf{x}_{S^c}^0(\mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^0). \end{aligned}$$

Solving for \mathbf{x}_S^0 in the first of the last two equations, substituting for it in the second, and letting

$$\mathbf{B}(S^c) = \mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^1 - \beta^2 \mathbf{P}_{S^c S}^1 (\mathbf{I}_S - \beta \mathbf{P}_{SS}^0)^{-1} \mathbf{P}_{SS^c}^0,$$

yields

$$\begin{aligned} \mathbf{x}_{S^c}^1 \mathbf{B}(S^c) &= \boldsymbol{\alpha}_{S^c} + \beta \boldsymbol{\alpha}_S (\mathbf{I}_S - \beta \mathbf{P}_{SS}^0)^{-1} \mathbf{P}_{SS^c}^0 \\ &\quad + \beta \mathbf{x}_S^1 [\mathbf{P}_{SS^c}^1 - (\mathbf{I}_S - \beta \mathbf{P}_{SS}^1)(\mathbf{I} - \beta \mathbf{P}_{SS}^0)^{-1} \mathbf{P}_{SS^c}^0] \\ &\quad - \mathbf{x}_{S^c}^0 [\mathbf{I}_{S^c} - \beta \mathbf{P}_{S^c S^c}^0 - \beta^2 \mathbf{P}_{S^c S}^0 (\mathbf{I}_S - \beta \mathbf{P}_{SS}^0)^{-1} \mathbf{P}_{SS^c}^0]. \end{aligned}$$

Now, postmultiplying both sides of the above equation by $\mathbf{T}_{S^c}^{S^c}$, and simplifying the resulting expression using (4.3)–(4.6), and the identity (which follows from the definition of \mathbf{A}_S^S)

$$\mathbf{A}_S^S = \mathbf{1}_S + \beta [\mathbf{P}_{SS^c}^1 - (\mathbf{I}_S - \beta \mathbf{P}_{SS}^1)(\mathbf{I} - \beta \mathbf{P}_{SS}^0)^{-1} \mathbf{P}_{SS^c}^0] \mathbf{T}_{S^c}^{S^c},$$

we obtain

$$\mathbf{x}_{S^c}^1 \mathbf{1}_{S^c} = \boldsymbol{\alpha} \mathbf{T}^{S^c} + \mathbf{x}_S^1 [\mathbf{A}_S^S - \mathbf{1}_S] - \mathbf{x}_{S^c}^0 \mathbf{A}_{S^c}^S.$$

Note further that the requirement that at each time one of the two projects be active implies that

$$\mathbf{x}_*^1 + \mathbf{x}_S^1 \mathbf{1}_S + \mathbf{x}_{S^c}^1 \mathbf{1}_{S^c} = \mathbf{x}_*^1 + \sum_{i \in N} x_i^1 = \frac{1}{1 - \beta}.$$

Combining the last two equations yields

$$\mathbf{x}_*^1 + \mathbf{x}_S^1 \mathbf{A}_S^S = \frac{1}{1 - \beta} - \boldsymbol{\alpha} \mathbf{T}^{S^c} + \mathbf{x}_{S^c}^0 \mathbf{A}_{S^c}^S,$$

which is precisely (4.7). This completes the proof.

Corollary 1. *The following identities hold, for any $S \subseteq N$:*

- (i) *for any admissible policy u that gives priority to classes in S^c over class $*$ (i.e., that takes the active action over S^c),*

$$x_*^1(u) + \sum_{i \in S} A_i^S x_i^1(u) = b(\{*\} \cup S);$$

- (ii) *for any admissible policy u that gives priority to class $*$ over classes in S (i.e., that takes the passive action over S),*

$$\sum_{i \in S} A_i^S x_i^1(u) = b(S) = 0.$$

Proof. (i) Note that, under any such policy, $x_i^0(u) = 0$, for $i \in S^c$. This fact, together with Lemma 3, proves this part.

(ii) The result follows by observing that, under any such policy, $x_i^1(u) = 0$, for $i \in S$. This completes the proof.

Now let $\mathcal{F} \subseteq 2^N$ be a family of project state subsets that satisfies the following requirements.

Assumption 3. *The following conditions hold:*

- (i) *the set system (N, \mathcal{F}) satisfies Assumption 1;*
- (ii) *for each $S \in \mathcal{F}^c$, $A_i^S > 0$, for $i \in S$ and $A_i^S \geq 0$, for $i \in S^c$.*

Let us define

$$\mathcal{F}_* = \mathcal{F} \cup \{\{*\} \cup S, S \in \mathcal{F}\}.$$

Note that it follows from the above that the set system (N_*, \mathcal{F}_*) satisfies Assumption 1, and it is therefore appropriate to define the corresponding PCL. We have the following result.

Theorem 4. (PCL: discounted restless bandits.) *The performance measures $x_i^1(u)$, for $i \in N_*$, satisfy the following PCL with respect to the set system (N_*, \mathcal{F}_*) : for each state subset $S \in \mathcal{F}^c$ and policy $u \in \mathcal{U}$, the inequalities*

$$x_*^1(u) + \sum_{i \in S} A_i^S x_i^1(u) \geq b(\{*\} \cup S), \quad \sum_{i \in S} A_i^S x_i^1(u) \geq b(S) \quad (4.8)$$

hold, together with the identities

$$x_*^1(u) + \sum_{i \in S} A_i^S x_i^1(u) = b(\{*\} \cup S), \quad \text{if } u \text{ gives priority to } S^c \text{ over } *, \quad (4.9)$$

$$\sum_{i \in S} A_i^S x_i^1(u) = b(S), \quad \text{if } u \text{ gives priority to } * \text{ over } S, \quad (4.10)$$

and

$$x_*^1(u) + \sum_{i \in N} x_i^1(u) = b(N_*). \quad (4.11)$$

Proof. First, it is easy to see that, due to the special structure of the two-project model at hand, the PCL in Definition 2 can be formulated equivalently as (4.8)–(4.11). Note next that the requirement that all the A_i^S coefficients arising in the PCL be positive is guaranteed to hold by Assumption 3. Furthermore, Assumption 3(ii), combined with the decomposition laws in Theorem 3, yields directly the inequalities (4.8). The required identities (4.9)–(4.11) were established in Corollary 1. This completes the proof.

We present next an adaptation of Definition 3 to the specific restless bandit model under consideration. Its correspondence with Definition 3 is apparent from Theorem 4. As in Section 3, we let $\mathcal{D}^1(\mathcal{F}) \subseteq \mathbb{R}^n$ denote the set of *admissible active rewards* \mathbf{R}^1 such that, when the adaptive-greedy algorithm $\text{AG}(\cdot | \mathcal{F})$ in the Appendix is fed with input \mathbf{R}^1 , it returns an output having $\text{ADMISSIBLE} = \text{YES}$. Namely, $\mathcal{D}^1(\mathcal{F})$ is the *domain* of the algorithm $\text{AG}(\cdot | \mathcal{F})$. Recall that $\mathcal{F}^c = \{S \subseteq N : S^c \in \mathcal{F}\}$.

Definition 4. (PCL/GCL-indexable restless bandits.) A restless bandit (normalized so that $\mathbf{R}^0 = \mathbf{0}$) is said to be *PCL-indexable with respect to the set system (N, \mathcal{F})* if the following conditions hold:

- (i) $A_i^S > 0$, $i \in N$, $S \in \mathcal{F}^c$;
- (ii) $\mathbf{R}^1 \in \mathcal{D}^1(\mathcal{F})$.

If $\mathcal{F} = 2^N$, we say the restless bandit is *GCL-indexable*.

As we show next, PCL-indexable restless bandits are indeed indexable, and their Whittle indices are precisely those returned by the adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$.

Corollary 2. (Indexability conditions.) *The following sufficient indexability conditions hold:*

- (i) *if a restless bandit is PCL-indexable with respect to the set system (N, \mathcal{F}) , then it is indexable under any reward vector $\mathbf{R}^1 \in \mathcal{D}^1(\mathcal{F})$. Its Whittle indices are given by the index vector $\boldsymbol{\gamma} = (\gamma_i)_{i \in N}$ computed by $\text{AG}(\cdot \mid \mathcal{F})$ when fed with the input \mathbf{R}^1 ;*
- (ii) *a GCL-indexable bandit is indexable under any reward vector.*

Proof. The fact that, for $\mathbf{R}^1 \in \mathcal{D}^1(\mathcal{F})$, the problem discussed above is solved optimally by a priority-index policy, where the optimal indices are computed by $\text{AG}(\cdot \mid \mathcal{F})$, is a straightforward consequence of combining Theorem 1 and Theorem 4. The result now follows by observing that the required conditions for index decomposition (Assumption 2) hold, as applied to the natural decomposition of class set N_* into $\{*\}$ and N , and therefore Theorem 2 applies. This yields the result that the priority index for auxiliary class/state $*$ is simply γ , while the priority indices for the classes/states in N are computed as described above.

Note that Corollary 2 provides an efficient algorithmic test for the indexability of a restless bandit: given that we have established that Assumption 3 holds, so that the bandit satisfies the PCL in Theorem 4, the test is based on checking whether the active reward vector of concern is admissible (whether $\mathbf{R}^1 \in \mathcal{D}^1(\mathcal{F})$), which involves a single run of $\text{AG}(\cdot \mid \mathcal{F})$. Note that this test represents a sufficient, yet not necessary, indexability condition.

We presented in (4.2) the standard LP formulation of (4.1), well known from MDC theory. The PCL framework provides a new, equivalent LP reformulation for PCL-indexable bandits, as shown next.

Corollary 3. *Suppose that a bandit (normalized so that $\mathbf{R}^0 = \mathbf{0}$) is PCL-indexable with respect to the set system (N, \mathcal{F}) as above. Then, the problem (4.1) can be reformulated as the LP problem*

$$\begin{aligned}
 Z^{\text{opt}}(\boldsymbol{\gamma}; \mathbf{0}, \mathbf{R}^1) &= \max \gamma x_*^1 + \sum_{i \in N} R_i^1 x_i^1 \\
 &\text{subject to} \\
 x_*^1(u) + \sum_{i \in S} A_i^S x_i^1(u) &\geq b(\{*\} \cup S), \quad S \in \mathcal{F}^c \setminus \{N\}, \\
 \sum_{i \in S} A_i^S x_i^1(u) &\geq b(S), \quad S \in \mathcal{F}^c \setminus \{N\}, \\
 x_*^1(u) + \sum_{i \in N} A_i^N x_i^1(u) &= b(N_*), \\
 x_i^1 &\geq 0, \quad i \in N_*.
 \end{aligned}$$

In our next result, we verify that projects corresponding to classical bandits, where $\mathbf{P}^0 = \mathbf{I}$, are GCL-indexable.

Corollary 4. *Classical bandits are GCL-indexable.*

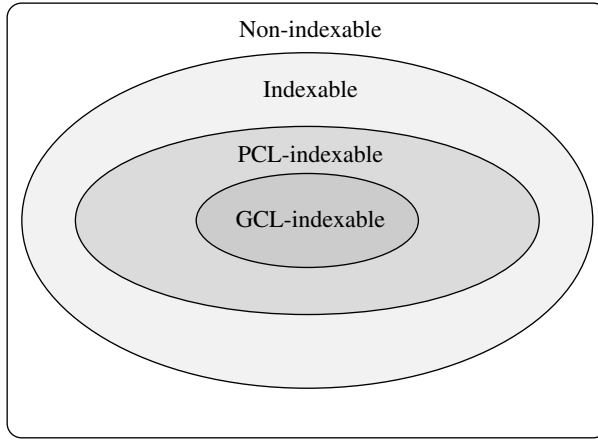


FIGURE 2: Classification of restless bandits.

Proof. In the classical case we have the identities, for $S \subseteq N$,

$$T_i^S = 1 + \beta \sum_{j \in S} p_{ij}^1 T_j^S, \quad \text{for } i \in S,$$

$$T_i^S = 0, \quad \text{for } i \in S^c.$$

The inequalities in Definition 4(i) follow immediately. This completes the proof.

The next result is concerned with the role of the discount factor on indexability. It is a direct consequence of the definition of GCL-indexability combined with Corollary 2.

Corollary 5. (GCL-indexability for small discounts.) *Any restless bandit is GCL-indexable when the discount factor β is small enough.*

Figure 2 illustrates the classification of restless bandits that results from the introduction of the classes of PCL- and GCL-indexable bandits.

Interpretation of GCL-indexability

We next consider the intuitive interpretation of GCL-indexability. Note that, for any $S \subset N$, we can write

$$A_i^{S^c} = \begin{cases} T_i^S - \beta \sum_{j \in N} p_{ij}^0 T_j^S, & i \in S, \\ 1 + \beta \sum_{j \in N} p_{ij}^1 T_j^S - T_i^S, & i \in S^c. \end{cases}$$

Let us focus on a given $S \subset N$. Recall the interpretation of T_i^S as the total expected discounted time that the project is active (*active time*, for short) under the S -active policy (which takes the active action on states in S , and the passive action otherwise), when starting in state $i \in N$. The condition $A_i^{S^c} > 0$, for a given $i \in S^c$, means that the effect of modifying the S -active policy by changing only the initial action (in state i) from passive to active is to increase the project expected active time. Similarly, the condition $A_i^{S^c} \geq 0$, for a given $i \in S$, means that the effect of modifying the S -active policy by changing only the initial action (in state i) from active to

passive is to decrease or let equal the project expected active time. We may thus regard the coefficients $A_i^{S^c}$ as active time differentials corresponding to changing the initial action in state i under the S -active policy.

5. PCL for restless bandits: time-average criterion

In this section we outline how the results in Section 4 for the time-discounted criterion can be extended to analyse the indexability of restless bandits under the time-average criterion.

We shall thus consider a general restless bandit, as described in Section 4, on which we shall impose the following additional requirement.

Assumption 4. (Ergodicity.) *For any subset $S \subseteq N$ of states, the Markov chain over N having transition probability matrix*

$$\mathbf{P}(S) = \begin{bmatrix} \mathbf{P}_{SN}^1 \\ \mathbf{P}_{S^c N}^0 \end{bmatrix}$$

is ergodic.

Note that $\mathbf{P}(S)$ is the transition probability matrix for the S -active policy considered in our study of the discounted case. Furthermore, as in the previous section, we need only consider the case where the passive reward vector is $\mathbf{R}^0 = \mathbf{0}$, since the general case can be reduced to it.

Our approach to the time-average case is based on analysing the asymptotic behaviour, as the discount factor β converges to one, of the expressions obtained in our analyses of the discounted case. Hence, to avoid confusion, in what follows we shall make explicit the dependence on the discount factor β in the quantities defined in Section 4, writing, e.g., $T_i^S(\beta)$, $A_i^S(\beta)$, and $x_i^u(u, \beta)$.

To relate the discounted and the time-average cases we shall apply the result that, under Assumption 4, for any state $i \in N$ and state subset $S \subseteq N$, we can express $T_i^S(\beta)$ as

$$T_i^S(\beta) = \frac{T^S}{1 - \beta} + t_i^S + O(1 - \beta) \quad \text{as } \beta \nearrow 1. \quad (5.1)$$

Here, T^S represents the long-run time-average fraction of time the bandit is active under the S -active policy, whereas t_i^S represents the corresponding total expected active time differential due to starting in state i . Substituting for $T_i^S(\beta)$ in (4.3) using (5.1), and letting $\beta \nearrow 1$, we obtain the system of linear equations

$$\begin{aligned} T^S + t_i^S &= 1 + \sum_{j \in N} p_{ij}^1 t_j^S, \quad \text{for } i \in S, \\ T^S + t_i^S &= \sum_{j \in N} p_{ij}^0 t_j^S, \quad \text{for } i \in S^c. \end{aligned} \quad (5.2)$$

Note that (5.2) determines t_i^S , for $i \in N$, in terms of T^S . Furthermore, T^S can be calculated as the sum of the equilibrium probabilities for the states in S corresponding to the ergodic Markov chain having transition probability matrix $\mathbf{P}(S)$.

We next apply (5.1) and (5.2) to study the asymptotics of the coefficients $A_i^S(\beta)$ appearing in the PCL obtained in the discounted case. Substituting for $T_i^S(\beta)$ in (4.4) using (5.1), and letting $\beta \nearrow 1$, yields the result that each $A_i^S(\beta)$ converges to a limit A_i^S , given by

$$A_i^{S^c} = 1 + \sum_{j \in N} (p_{ij}^1 - p_{ij}^0) t_j^S, \quad i \in N, S \subseteq N.$$

The performance measures of interest are now the time-average state-action frequencies: we denote by $x_i^a(u)$ the time-average fraction of time that the project is in state i and action a is taken under policy u . We further write $\mathbf{x}^a(u) = (x_i^a(u))_{i \in N}$. It is well known that in the ergodic case under discussion, under any stationary policy u ,

$$\mathbf{x}^a(u) = \lim_{\beta \nearrow 1} (1 - \beta) \mathbf{x}^a(u, \beta).$$

In order to investigate the indexability property under the time-average criterion, we consider the equivalent two-project restless bandit model discussed in Section 4, where an auxiliary calibrating project having a single state $*$ is introduced, yielding a reward of γ when active (or, equivalently, when the original project is passive).

We further define, as in the discounted case, quantities $A_i^{\{*\} \cup S}$, for $i \in \{*\} \cup S$ and $S \subseteq N$, by

$$A_i^{\{*\} \cup S} = \begin{cases} A_i^S & \text{if } i \in S, \\ 1 & \text{if } i = *. \end{cases}$$

We complete the definitions by letting $b(S_*)$, for $S_* \subseteq \{*\} \cup N$, be given by

$$b(S_*) = \begin{cases} 1 - T^{S^c} & \text{if } * \in S_* \text{ and } S = S_* \cap N \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

With these definitions, all of the results of the previous section carry over, in a verbatim fashion, to the time-average case under discussion by taking limits as $\beta \nearrow 1$.

6. Examples

In this section we analyse several special cases of restless bandits using the results developed above.

6.1. Two-state restless bandits are GCL-indexable

We first consider the case of restless bandits having two states. The corresponding two-project restless bandit problem discussed in Section 4 is precisely that represented in Figure 1. In the discounted case, the relevant A_i^S coefficients are readily calculated to be

$$\begin{aligned} A_1^{\{1\}} &= \frac{1 + \beta - \beta p_{11}^1 - \beta p_{22}^1}{1 + \beta - \beta p_{11}^0 - \beta p_{22}^1}, \\ A_2^{\{1\}} &= \frac{1 + \beta - \beta p_{11}^0 - \beta p_{22}^0}{1 + \beta - \beta p_{11}^0 - \beta p_{22}^1}, \\ A_1^{\{2\}} &= \frac{1 + \beta - \beta p_{11}^0 - \beta p_{22}^0}{1 + \beta - \beta p_{11}^1 - \beta p_{22}^0}, \\ A_2^{\{2\}} &= \frac{1 + \beta - \beta p_{11}^1 - \beta p_{22}^1}{1 + \beta - \beta p_{22}^0 - \beta p_{11}^1}, \end{aligned}$$

whereas $A_i^\emptyset = A_i^N = 1$, for $i = 1, 2$ (cf. (4.5)). It follows that, for any discount factor $0 < \beta < 1$ and set $S \in \mathcal{F} = 2^N$, $A_i^S > 0$, for $i = 1, 2$. Therefore, we find that discounted two-state restless bandits are GCL-indexable, and hence indexable under any reward vector. By taking

the limit as $\beta \nearrow 1$, the corresponding result is obtained for the time-average case, provided the ergodicity requirement in Assumption 4 holds.

Furthermore, if $R_1^1 \geq R_2^1$ (we assume, as before, that passive rewards are 0), the Whittle indices produced by the adaptive-greedy algorithm $\text{AG}(2^{\{1,2\}})$ when fed with input $\mathbf{R}^1 = (R_i^1)_{i \in N}$ are $\gamma_1 = R_1^1$ and $\gamma_2 = R_1^1 + (R_2^1 - R_1^1)/A_2^{\{2\}}$.

6.2. An indexable three-state restless bandit that is not PCL-indexable

Consider now a discounted restless bandit having state space $N = \{1, 2, 3\}$ and transition probability matrices given by

$$\mathbf{P}^1 = \begin{bmatrix} \varepsilon & 0 & 1 - \varepsilon \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \quad \text{and} \quad \mathbf{P}^0 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \varepsilon & 0 & 1 - \varepsilon \end{bmatrix},$$

for some $0 < \varepsilon < 1$. We have

$$\begin{aligned} A_1^{\{3\}} &= \frac{3 - (4 - 3\varepsilon)\beta - (1 - 2\varepsilon)\beta^2}{3 - \beta}, \\ A_2^{\{3\}} &= 1, \\ A_3^{\{3\}} &= \frac{3 + (1 - 3\varepsilon)\beta}{3 - \beta}. \end{aligned}$$

Therefore, $A_2^{\{3\}}, A_3^{\{3\}} > 0$ for any $\beta, \varepsilon \in (0, 1)$. Note that, however, $A_1^{\{3\}}$ can be made negative for any $0 < \varepsilon < \frac{2}{5}$ by choosing the discount factor β close enough to unity. We shall consider the case $\varepsilon = \frac{1}{9}$ and $\beta = \frac{3}{4}$, for which $A_1^{\{3\}} = -\frac{1}{12}$, $A_2^{\{3\}} = 1$ and $A_3^{\{3\}} = \frac{14}{9}$. Since $A_1^{\{3\}} < 0$, this restless bandit (under any reward objective) is *not* GCL-indexable (see Definition 4). Consider now the decomposition identity (4.7) in Theorem 3 corresponding to $S = \{3\}$: under any stationary policy u ,

$$x_*^1(u) + A_3^{\{3\}} x_3^1(u) = b(\{*, 3\}) + A_1^{\{3\}} x_1^0(u) + A_2^{\{3\}} x_2^0(u).$$

Note, in particular, that the PCL (4.9) in Theorem 1 (with $S = \{3\}$) does not hold, since the problem of minimizing the objective

$$x_*^1(u) + A_3^{\{3\}} x_3^1(u)$$

under stationary policies u is solved by the $\{2, 3\}$ -active policy, and not by the $\{1, 2\}$ -active policy. Consider the case where the reward coefficients are $R_1^1 = R_2^1 = 0$, $R_3^1 = -A_3^{\{3\}}$, and passive rewards are zero. Then, it can be checked that the optimal active sets for the problem of maximizing the objective $\gamma x_*^1(u) - A_3^{\{3\}} x_3^1(u)$, or, equivalently, of maximizing $-\gamma x_1^1(u) - \gamma x_2^1(u) + (-\gamma - A_3^{\{3\}}) x_3^1(u)$, as the passive subsidy γ decreases from $+\infty$ to $-\infty$ are successively given by \emptyset (i.e., the \emptyset -active policy is optimal for γ large enough), $\{2\}$, $\{2, 3\}$, and $\{1, 2, 3\}$. Therefore, under such rewards the bandit is indexable, though not GCL-indexable. Is it at least PCL-indexable? The answer is no: if it were PCL-indexable, it would have to be PCL-indexable with respect to the family of optimal active sets, which was found to be

$$\mathcal{F} = \{\emptyset, \{2\}, \{2, 3\}, \{1, 2, 3\}\}.$$

We have, however, that $A_1^{\{1,3\}} = -\frac{1}{11} < 0$ (note that $\{1, 3\} \in \mathcal{F}^c$; see Definition 4). Therefore, this restless bandit is indexable, yet it is not PCL-indexable.

6.3. Threshold optimality in admission control through PCL

We illustrate next through a small example how PCL provide an appropriate framework for investigating the optimality of threshold policies in admission control problems on queueing systems. This issue is studied in general in [12]. Consider a discrete-time single-server queue which can hold at most 2 customers. The set of system states, i.e., of the possible number of customers in the system (waiting or in service), is thus $N = \{0, 1, 2\}$. At each time epoch the system manager can choose to accept potential arrivals (active action), if there are fewer than 2 customers in the system, or to reject them (passive action). Corresponding state-dependent rewards are earned under the active action, which are discounted in time with factor $0 < \beta < 1$. The goal is to design an admission control policy that maximizes the total expected discounted reward earned over an infinite horizon. We model this problem as a three-state restless bandit, where the active and passive transition probability matrices are given by

$$P^1 = \begin{bmatrix} \mu & \lambda & 0 \\ \mu & 0 & \lambda \\ 0 & \mu & \lambda \end{bmatrix}, \quad P^0 = \begin{bmatrix} 1 & 0 & 0 \\ \mu & \lambda & 0 \\ 0 & \mu & \lambda \end{bmatrix},$$

where $\lambda, \mu > 0$, with $\lambda + \mu = 1$. The state-dependent active rewards are R_0^1, R_1^1 , and R_2^1 . Passive rewards are assumed to be zero.

The optimality of threshold policies (reject arrivals when the number in the system exceeds a given threshold value) will be deduced, under appropriate conditions on reward coefficients, from the satisfaction by this restless bandit of PCL with respect to the family of active sets

$$\mathcal{F} = \{\emptyset, \{0\}, \{0, 1\}, \{0, 1, 2\}\}$$

(see Definition 4). Let us calculate A_i^S , for $S \in \mathcal{F}^c$ and $i \in S$. For $S = \{2\}$, we have

$$\begin{aligned} A_0^{\{2\}} &= \frac{1 - \beta^2 \lambda}{1 - \beta^2 \lambda \mu} > 0, \\ A_1^{\{2\}} &= \frac{1 - \beta \lambda (1 + \beta \mu)}{1 - \beta^2 \lambda \mu} > 0, \\ A_2^{\{2\}} &= 1 > 0. \end{aligned}$$

For $S = \{1, 2\}$,

$$\begin{aligned} A_0^{\{1,2\}} &= 1 - \beta \lambda > 0, \\ A_1^{\{1,2\}} &= 1 - \frac{\beta^2 \lambda \mu}{1 - \beta \lambda} > 0, \\ A_2^{\{1,2\}} &= 1 > 0. \end{aligned}$$

For $S = \{0, 1, 2\}$, we have

$$A_i^{\{0,1,2\}} = 1, \quad i \in \{0, 1, 2\}.$$

Therefore, this restless bandit satisfies PCL with respect to \mathcal{F} . Let us now characterize the corresponding set of admissible rewards $\mathcal{D}(\mathcal{F})$, characterized as the domain of $AG(\cdot \mid \mathcal{F})$. The conditions defining $\mathcal{D}(\mathcal{F})$ are easily seen to be

$$R_0^1 \geq \max(R_1^1, R_2^1) \quad \text{and} \quad \frac{R_1^1 - R_0^1}{A_1^{\{1,2\}}} \geq \frac{R_2^1 - R_0^1}{A_2^{\{1,2\}}}.$$

Therefore, under these conditions on reward coefficients, the bandit is PCL-indexable with respect to \mathcal{F} . The corresponding Whittle indices produced by $\text{AG}(\cdot \mid \mathcal{F})$ are

$$\gamma_0 = R_0^1, \quad \gamma_1 = R_0^1 + \frac{R_1^1 - R_0^1}{A_1^{\{1,2\}}}, \quad \gamma_2 = \gamma_1 + R_2^1 - \gamma_0 - A_2^{\{1,2\}}(\gamma_1 - \gamma_0).$$

7. Concluding remarks

We introduced the class of PCL-indexable restless bandits, which are guaranteed to be indexable. Membership of a given restless bandit in this class can be efficiently tested through a single run of an adaptive-greedy algorithm based on Klimov's, which further computes the Whittle indices when the test is positive. The notion of PCL-indexability explains the indexability property in several important models, including classical bandits (which are GCL-indexable) and input control in queueing systems (see [12]). Given the rich modelling power of restless bandits, and the interest of having simple conditions for their indexability, we believe the notion of PCL-indexability introduced in this paper has the potential of providing a unifying framework to analyse a variety of specific restless bandit models arising in applications.

Our analysis further reveals the power of the achievable region method (cf. [5]) to analyse stochastic optimization problems, as our analyses are based on deriving and exploiting new linear programming formulations of the problems investigated.

Acknowledgements

The author thanks Professor K. Glazebrook and Professor G. Weiss for interesting discussions that helped improve the presentation of the paper. A preliminary version of the paper was presented at the 10th INFORMS Applied Probability Conference, Ulm, Germany, 26–28 July 1999. This work was supported in part by Spanish National R & D Program Grant CICYT TAP98-0229, and by NATO Collaborative Linkage Grant PST.CLG.976568.

Appendix. Adaptive-greedy algorithm $\text{AG}(\cdot \mid \mathcal{F})$

Input: $R = (R_i)_{i \in N}$

Output: ($\text{ADMISSIBLE} \in \{\text{YES}, \text{NO}\}$, $\boldsymbol{\gamma} = (\gamma_i)_{i \in N}$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$)

Initialization: let $\text{ADMISSIBLE} := \text{YES}$

let $\bar{y}^N := \max\{R_i/A_i^N : i \in N, N \setminus \{i\} \in \mathcal{F}^c\}$

choose $\pi_1 \in N$ attaining the above maximum

let $\gamma_{\pi_1} := \bar{y}^N$

let $S_1 := N$; let $k := 1$

Loop: while $\text{ADMISSIBLE} = \text{YES}$ and $k \leq n - 1$ do

begin

let $S_{k+1} := S_k \setminus \{\pi_k\}$

let $\bar{y}^{S_{k+1}} := \max\{[R_i - \sum_{j=1}^k A_i^{S_j} \bar{y}^{S_j}]/A_i^{S_{k+1}} : i \in S_{k+1}, S_{k+1} \setminus \{i\} \in \mathcal{F}^c\}$

choose $\pi_{k+1} \in S_{k+1}$ attaining the above maximum

let $\gamma_{\pi_{k+1}} := \gamma_{\pi_k} + \bar{y}^{S_{k+1}}$

if $\gamma_{\pi_{k+1}} > \gamma_{\pi_k}$ then let $\text{ADMISSIBLE} := \text{NO}$

let $k := k + 1$

end {while}

References

- [1] BERTSIMAS, D. AND NIÑO-MORA, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a unified approach to indexable systems. *Math. Operat. Res.* **21**, 257–306.
- [2] BERTSIMAS, D. AND NIÑO-MORA, J. (1999). Optimization of multiclass queueing networks with changeover times via the achievable region method: Part I, the single-station case. *Math. Operat. Res.* **24**, 306–330.
- [3] BERTSIMAS, D. AND NIÑO-MORA, J. (2000). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operat. Res.* **48**, 80–90.
- [4] CHEN, Y. R. AND KATEHAKIS, M. N. (1986). Linear programming for finite state multi-armed bandit problems. *Math. Operat. Res.* **11**, 180–183.
- [5] DACRE, M., GLAZEBROOK, K. D. AND NIÑO-MORA, J. (1999). The achievable region approach to the optimal control of stochastic systems (with discussion). *J. R. Statist. Soc. B* **61**, 747–791.
- [6] D'EPÉNOUX, F. (1960). Sur un problème de production et de stockage dans l'aléatoire. *RAIRO Rech. Opérat.* **14**, 3–16. English translation: A probabilistic production and inventory problem. *Management Sci.* **10** (1963), 98–108.
- [7] FAIHE, Y. AND MÜLLER, J.-P. (1998). Behaviors coordination using restless bandits allocation indexes. In *From Animals to Animats 5 (Proc. 5th Int. Conf. Simulation of Adaptive Behavior)*, eds R. Pfeifer, B. Blumberg, J. A. Meyer and S. W. Wilson. MIT Press, Cambridge, MA.
- [8] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. R. Statist. Soc. B* **41**, 148–177.
- [9] KATEHAKIS, M. N. AND VEINOTT, A. F., JR (1987). The multi-armed bandit problem: decomposition and computation. *Math. Operat. Res.* **12**, 262–268.
- [10] KLIMOV, G. P. (1974). Time sharing service systems I. *Theory Prob. Appl.* **19**, 532–551.
- [11] MANNE, A. S. (1960). Linear programming and sequential decisions. *Management Sci.* **6**, 259–267.
- [12] NIÑO-MORA, J. (2000). Admission control to birth–death queueing systems, restless bandit indices, and partial conservation laws. Working paper, Department of Economics and Business, Universitat Pompeu Fabra.
- [13] PAPADIMITRIOU, C. H. AND TSITSIKLIS, J. N. (1999). The complexity of optimal queueing network control. *Math. Operat. Res.* **24**, 293–305.
- [14] VEATCH, M. AND WEIN, L. M. (1996). Scheduling a make-to-stock queue: index policies and hedging points. *Operat. Res.* **44**, 634–647.
- [15] WEBER, R. R. AND WEISS, G. (1990). On an index policy for restless bandits. *J. Appl. Prob.* **27**, 637–648.
- [16] WEBER, R. R. AND WEISS, G. (1991). Addendum to 'On an index policy for restless bandits'. *Adv. Appl. Prob.* **23**, 429–430.
- [17] WHITTLE, P. (1988). Restless bandits: activity allocation in a changing world. In *A Celebration of Applied Probability* (J. Appl. Prob. **25A**), ed. J. Gani. Applied Probability Trust, Sheffield, pp. 287–298.