Restless bandits, partial conservation laws and indexability

José Niño-Mora*

December 13, 1999

*Dept. of Economics and Business, Universitat Pompeu Fabra, E-08005 Barcelona, Spain. E-mail: jose.nino-mora@econ.upf.es. This work was supported in part by Spanish National R & D Program Grant CICYT TAP98-0229. A preliminary version of this paper was presented at the 10th INFORMS Applied Probability Conference, in Ulm, Germany, July 26–28, 1999.

Abstract

We show that if performance measures in a stochastic scheduling problem satisfy a set of so-called *partial conservation laws* (PCL), which extend previously studied generalized conservation laws (GCL), then the problem is solved optimally by a priority-index policy for an appropriate range of linear performance objectives, where the optimal indices are computed by a one-pass adaptive-greedy algorithm, based on Klimov's. We further apply this framework to investigate the indexability property of restless bandits introduced by Whittle, obtaining the following results: (1) we identify a class of restless bandits (PCL-indexable) which are indexable; membership in this class is tested through a single run of the adaptive-greedy algorithm, which also computes the Whittle indices when the test is positive; this provides a tractable sufficient condition for indexability; (2) we further identify the class of GCL-indexable bandits, which includes classical bandits, having the property that they are indexable under any linear reward objective. The analysis is based on the so-called achievable region method, as the results follow from new linear programming formulations for the problems investigated.

Key words: Stochastic scheduling, Markov decision chains, bandit problems, achievable region.

Journal of Economic Literature Classification: C60, C61.

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 and study of restricted problem classes whose special structure yields a tractable solution remains of prime research interest: not only such well-solved problems are often of intrinsic interest, but their optimal solutions may provide building blocks for constructing well-grounded heuristic solutions to more complex models. The latter situation is epitomized by Whittle's [14] pioneering approach to what is arguably the most promising extension to the classical multiarmed bandit model: the *restless bandit problem*.

Both bandit models, classical and restless, are concerned with designing an optimal sequential resource allocation policy for a collection of stochastic projects, each of which is modeled as a finite Markov decision chain (MDC) having two actions at each state, with associated rewards: an *active* action, which corresponds to engaging the project, and a *passive* action, which corresponds to letting it go. These models are therefore paradigms of the fundamental conflict between taking actions that yield high current reward, or taking instead actions that sacrifice current gains with the prospect of reaping better returns in the future. In the classical model, a single resource is to be allocated, so that at each time only one project is engaged, and passive projects do not change state. In the restless model, a fixed number of resources (which may be larger than one) is to be allocated, so that at each time a fixed number of projects is active, and passive projects can change state, in general through a different transition rule. The performance objective to be maximized in both models may be the time-discounted total expected reward, or the time-average reward rate.

While the classical model is well-known to be solved optimally by Gittins [6] *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 proven in [9] to be *PSPACE-hard*, which most likely rules out the possibility of describing explicitly an optimal policy.

Yet the rich modeling power of restless bandits makes the development and analysis of a sound heuristic policy a problem of significant research importance: Whittle [14] proposed applications in the areas of clinical trials, aircraft surveillance and worker scheduling. Other applications studied in the literature include control of a make-to-stock queue [11], and behavior coordination in robotics

[5].

In his seminal paper on the subject, Whittle [14] presented a simple heuristic policy, together with a related bound on the optimum problem value, both of which can be efficiently computed. Whittle's policy, like Gittins', 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 version, whose optimum value gives the aforementioned bound: instead of requiring that K projects be active *at each time*, it is required instead 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 (i.e., a finite-state two-action MDC), such 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: the property states that, as the passive subsidy γ grows, the set of states where it is optimal to take the passive action increases monotonically from the empty set to the full state space. The priority indices in Whittle's heuristic for the states of a project are precisely the corresponding breakpoints. The appealing properties of such rule include (1) the fact that it extends the Gittins optimal policy for classical bandits; (2) the fact that Whittle's indices, like Gittins', can be computed separately for each project; and (3) its asymptotic optimality, under regularity conditions, when K and M tend to infinity in a constant ratio, as established by Weber and Weiss in [12], [13].

Given a restless project, testing whether it is or is not indexable, and computing the Whittle indices in the affirmative case, are tasks which can be efficiently accomplished through straightforward application of the definition of indexability, combined with the well-known linear programming (LP) formulations for MDC (see, e.g., [10]): they involve n Simplex pivot steps, carried out on an LP problem having $2n$ variables and n constraints, n being the number of states. Such purely numerical procedure, however, does not provide any qualitative insight as to which projects are indexable. It does not appear therefore suitable for attaining our main research goal, namely, to identify and analyze relevant classes of indexable bandits, defined by conditions on their parameters. Whittle [14] 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 further progress on the matter has been achieved prior to the results we present in this paper.

Our approach to the study of the indexability property for restless bandits is based on the *achievable region method* (cf. [4]): we shall show that the indexability property can be (partially) explained through a corresponding structural property on the underlying polyhedral *achievable performance region* of a restless bandit (spanned by performance measures under all admissible policies).

Specifically, our contribution in this paper is twofold. First, we present a general polyhedral framework for investigating stochastic scheduling problems having the following *partial indexability* property: priority-index policies are optimal for an appropriately restricted range of linear performance objectives. This framework extends that given in [1] for investigating scheduling problems for which priority-index policies are optimal under *any* linear performance objective (*general indexability*). The framework in [1] was based on obtaining a *full* polyhedral characterization of the system's achievable performance region from the satisfaction by performance measures of so-called *generalized conservation laws* (GCL). The extension we present here is based instead on exploiting a *partial* polyhedral characterization of the achievable performance region, from the satisfaction of an appropriate subset of the previous laws, which we call *partial conservation laws* (PCL). We show, in particular, that if the performance measures of interest satisfy PCL, then (1) the problem is partially indexable for an appropriate range of linear performance objectives, which is defined algorithmically; and (2) the optimal priority indices are computed by a one-pass *adaptive-greedy algorithm*, which extends that utilized in the GCL framework.

Second, we apply the PCL framework to investigate the indexability property in restless bandits. In particular, we identify a class of restless bandits (*PCL-indexable*), defined through checkable conditions on their parameters, which are indexable for a range of reward coefficients. This range is characterized algorithmically through a single run of the adaptive-greedy algorithm mentioned above. We further characterize the subclass of PCL-indexable bandits that satisfy GCL (*GCL-indexable*): these bandits are indexable under *any* set of reward coefficients. The conditions defining this class, of which classical bandits are the main example, provide qualitative insight on their nature.

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. The 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

We review in this section the problem relaxation and heuristic index policy proposed by Whittle [14] for the restless bandit problem. 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 $m \in M = \{1, \ldots, M\}$ is modeled as a Markov decision chain (MDC) that evolves in a finite state space N_m, and has two actions $a \in \{0,1\}$ available at each state $i \in N_m$. We shall assume, for convenience of notation, that the state spaces for the M projects are disjoint, and denote their *aggregate state space* by $N = \cup_{m=1}^{M} N_m$. The *active* ($a = 1$) and *passive* actions ($a = 0$) correspond, respectively, to engaging or not a project. Taking action $a \in \{0,1\}$ on a project in state i has two effects: first, it yields an instant reward R_i^a; then, it causes the state to evolve in a Markovian fashion, moving at the next time into state j with probability p_{ij}^a. We write $\mathbf{R}^a = (R_i^a)_{i \in N}$ and $\mathbb{P}^a = (p_{ij}^a)_{i,j \in N}$, for $a = 0, 1$. Both the time-discounted and the time-average versions of the problem will be considered.

Under the discounted criterion, rewards are time-discounted by factor $0 < \beta < 1$, and the problem consists in finding a *scheduling policy* u^{OPT}, belonging in the space 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:

$$Z^{\mathrm{OPT}}(\beta) = \max_{u \in U} E_u \left[\sum_{t=0}^{\infty} \left(R_{i_1(t)}^{a_1(t)} + \cdots + R_{i_M(t)}^{a_M(t)} \right) \beta^t \right]. \tag{1}$$

In formulation (1), $Z^{\mathrm{OPT}}(\beta)$ denotes the optimum problem value, $i_m(t)$ and $a_m(t)$ denote the state and action corresponding to project m at time t, and $E_u[\cdot]$ represents the expectation operator under policy

u. This expectation is conditional on initial project states, as given by a known vector $\alpha = (\alpha_i)_{i \in N}$, where

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

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):

$$Z^{\mathrm{OPT}}(1) = \max_{u \in U} \lim_{T \to \infty} \frac{1}{T} E_u \left[\sum_{t=0}^{T} \left(R_{i_1(t)}^{a_1(t)} + \cdots + R_{i_M(t)}^{a_M(t)} \right) \right]. \tag{2}$$

Note that in formulation (2) we denote the optimal problem value by $Z^{\mathrm{OPT}}(1)$, thus identifying, for notational convenience, the time-average criterion with the value $\beta = 1$. Using this convention will allow us to discuss below the time-discounted and time-average cases in parallel.

Whittle's relaxation is based on consideration of a modified problem version: the 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 the original problem's optimum value. Furthermore, this bound can be efficiently computed by solving a polynomial-size linear program (cf. [2]). To see this, let us associate a *performance measure* with each stationary scheduling policy $u \in U$, project state $i \in N$, and action $a \in \{0, 1\}$, defined in the discounted case by

$$x_i^a(\beta, u) = E_u \left[\sum_{t=0}^{\infty} I_i^a(t) \beta^t \right], \tag{3}$$

and in the time-average case by

$$x_i^a(1, u) = \lim_{T \to \infty} \frac{1}{T} E_u \left[\sum_{t=0}^{T} I_i^a(t) \right], \tag{4}$$

where

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

Performance measure $x_i^a(\beta, u)$ (resp. $x_i^a(1, u)$) represents the total expected discounted number of times (resp. the long-run fraction of time) that action a is taken at a project in state i under policy u. Now, it follows directly from the standard LP formulations for discounted and time-average MDC (see, e.g., [10]) that Whittle's relaxation can be formulated as the LP problem

$$\begin{aligned} Z^{\mathrm{W}}(\beta) \quad = \quad & \max \sum_{j \in N} R_j^1 x_j^1 + \sum_{j \in N} R_j^0 x_j^0 && (5) \\ & \text{subject to} \\ & (x_j^0, x_j^1)_{j \in N_m} \in P_m(\beta), \quad \text{for } m = 1, \ldots, M \\ & \sum_{j \in N} x_j^1 = K\, s(\beta), && (6) \end{aligned}$$

where

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

$$P_m(\beta) = \{(x_j^0, x_j^1)_{j \in N_m} \geq \mathbf{0} : \sum_{a \in \{0,1\}} x_j^a - \beta \sum_{a \in \{0,1\}} \sum_{i \in N_m} p_{ij}^a x_i^a = \alpha_j, \quad j \in N_m\}, \tag{7}$$

for $0 < \beta < 1$, is the polytope defined by the standard LP constraints for the discounted MDC modeling project m, and

$$P_m(1) = \{(x_j^0, x_j^1)_{j \in N_m} \geq \mathbf{0} : \sum_{a \in \{0,1\}} x_j^a - \sum_{a \in \{0,1\}} \sum_{i \in N_m} p_{ij}^a x_i^a = 0, j \in N_m \quad \text{and} \sum_{a \in \{0,1\}} \sum_{j \in N_m} x_j^a = 1\} \tag{8}$$

is the corresponding polytope for the time-average case. In linear program (5), the variables x_j^a correspond to performance measures $x_j^a(\beta, u)$, and constraint (6) formulates the relaxed requirement

that K projects be active on average. Note that this linear program 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 linking constraint (6) in linear program (5), denoting by γ the corresponding Lagrange multiplier, and using the implied constraint

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

one obtains the Lagrangian relaxation

$$\begin{aligned} L(\beta,\gamma) \quad = \quad & \max \sum_{j\in N} R_j^1 x_j^1 + \sum_{j\in N} (R_j^0+\gamma) x_j^0 - (M-K)\, s(\beta)\,\gamma \\ & \text{subject to} \\ & (x_j^0, x_j^1)_{j\in N_m} \in P_m(\beta), \quad \text{for } m = 1,\ldots,M. \end{aligned} \tag{9}$$

We thus see that, as observed by Whittle, multiplier γ plays the economic role of a constant *subsidy for passivity*. Problem (9) can be decoupled into the M single-project subproblems

$$\begin{aligned} L_m(\beta,\gamma) \quad = \quad & \max \sum_{j\in N_m} R_j^1 x_j^1 + \sum_{j\in N_m} (R_j^0+\gamma) x_j^0 \\ & \text{subject to} \\ & (x_j^0, x_j^1)_{j\in N_m} \in P_m(\beta), \end{aligned} \tag{10}$$

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

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

Note that subproblem (10) corresponds to a single MDC over project m, where passive rewards (earned under the passive action) are subsidized by a constant amount γ. Note further that, for each multiplier γ, it holds that $L(\beta,\gamma) \geq Z^{\mathrm{W}}(\beta)$. Furthermore, by the strong duality property of linear programs, there exists a multiplier γ^* (which, in the discounted case, depends on initial state vector α)

such that $L(\beta,\gamma^*) = Z^W(\beta)$. In the regular case where $\gamma^* \neq 0$, LP complementary slackness ensures that *any* optimal solution to Lagrangian relaxation (9) (with $\gamma = \gamma^*$) must satisfy linking constraint (6), and will therefore be optimal for the LP formulation of Whittle's relaxation in (5).

Whittle identified a key structural property that makes the solution to the parametric family of single-project MDC subproblems formulated in (10) particularly simple.

Definition 1 (Whittle [14]) *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 corresponding single-project subproblem (10) 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 values γ_i for each state *i*, such that an optimal policy for the corresponding subproblem (10) can be given as follows: take the active action in states *i* with $\gamma_i < \gamma$, and the passive action otherwise. Note that for $\gamma = \gamma_i$ both the active and the passive actions are optimal. 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 used the above indices to define a priority-index heuristic for the original problem: activate at each time *K* projects with larger indices. Note that it follows from the definition of Whittle's indices that they reduce to Gittins' when applied to classical multiarmed bandits, and therefore the heuristic is optimal in that special case.

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. The framework extends that introduced in [1] for studying the *general indexability* property in stochastic scheduling.

Consider a general dynamic and stochastic service system catering to a finite set $N = \{1,\ldots,n\}$ 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 U of *admissible policies*. The performance of a policy $u \in U$ 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 permutation $\pi = (\pi_1, \ldots, \pi_n)$ of the classes is associated a corresponding π-*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* $\mathbf{R} = (R_i)_{i \in N} \in \Re^n$, the *optimal scheduling problem*

$$Z^{\mathrm{OPT}}(\mathbf{R}) = \sup\{ \sum_{i \in N} R_i x_i(u) : u \in U \}, \tag{11}$$

which involves finding a scheduling policy $u^{\mathrm{OPT}}(\mathbf{R})$ maximizing the linear performance objective in (11), and computing the corresponding optimum value $Z^{\mathrm{OPT}}(\mathbf{R})$.

A wide variety of scheduling models fitting formulation (11), such as classical multiarmed bandits, possess the following structural property, which we call *general indexability*: for *any* reward vector $\mathbf{R} \in \Re^n$, problem (11) is solved optimally by a priority-index policy, i.e., a priority policy where the optimal priorities are determined by a set of class-ranking indices (with higher priorities corresponding to larger indices). A general framework providing a sufficient condition for problem (11) to be generally indexable was presented in [1]: satisfaction by performance vector $\mathbf{x}(u)$ of so-called *generalized conservation laws* (GCL) implies general indexability; furthermore, in such case the optimal priority indices can be efficiently computed by means of an *n*-step *adaptive-greedy* algorithm due to Klimov [7].

In the context of certain models, however, general 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. We call such property *partial indexability*. That is the case, e.g., in much studied models involving the control of arrivals into a single queue, where researchers typically aim to prove the optimality of threshold policies (shut off arrivals only when the queue length exceeds a given threshold value) under strong structural assumptions on linear reward/cost coefficients. See [8]. More generally, as we shall demonstrate in Sections 4 and 5,

the problem of determining whether a given restless bandit is indexable may be formulated in terms of checking the partial indexability of a problem of the form (11).

We present next an extension of the GCL presented in [1], with the goal of providing a framework for investigating the partial indexability property. Let $\mathcal{S}$ be a family of subsets of N, i.e., $\mathcal{S} \subseteq 2^N$, having $N \in \mathcal{S}$.

Definition 2 (Partial conservation laws (PCL)) *Performance vector* $\mathbf{x}(u)$ *satisfies* partial conservation laws *with respect to class set family* $\mathcal{S}$ *if there exist quantities* $A_i^S > 0$, *for* $i \in S$ *and* $S \in \mathcal{S}$, *such that, letting*

$$b(S) = \inf\{\sum_{i \in S} A_i^S x_i(u) : u \in S\}, \quad \text{for } S \in \mathcal{S},$$

the following identities hold: for each $S \in \mathcal{S}$,

$$\sum_{i \in S} A_i^S x_i(u) = b(S), \qquad \text{for any policy } u \in U \text{ that gives priority to } S^c \text{ over } S; \tag{12}$$

and, for $S = N$,

$$\sum_{i \in N} A_i^N x_i(u) = b(N), \qquad \text{for any policy } u \in U. \tag{13}$$

Note that the GCL in [1] correspond to the special case where $\mathcal{S} = 2^N$. In words, a performance vector $\mathbf{x}(u)$ satisfies PCL with respect to $\mathcal{S} \subseteq 2^N$ if, for each subset $S \in \mathcal{S}$ of job classes, there exist weights $A_i^S > 0$, for $i \in S$, such that the corresponding weighted performance objective is *minimized* by *any* policy that gives priority to S^c-jobs, and is invariant under all admissible policies when $S = N$. The laws thus state that the family of priority policies that give priority to S^c-jobs, for $S \in \mathcal{S}$, optimizes a certain *finite* set of linear performance objectives. We shall show in this section that the laws further imply the optimality of such 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, let us consider the LP problem

$$
\begin{aligned}
Z^{\mathrm{LP}}(\mathbf{R}) \quad = \quad & \max \sum_{i\in N} R_i x_i \qquad (14)\\
& \text{subject to}\\
& \sum_{i\in S} A_i^S x_i \geq b(S), \quad S \in S \setminus \{N\}\\
& \sum_{i\in N} A_i^N x_i = b(N)\\
& x_i \geq 0, \quad i \in N\,.
\end{aligned}
$$

For any reward vector $\mathbf{R} \in \Re^n$, (14) represents an *LP relaxation* of scheduling problem (11), since satisfaction of the PCL above implies that $Z^{\mathrm{LP}}(\mathbf{R}) \geq Z^{\mathrm{OPT}}(\mathbf{R})$. We next identify a family of reward vectors for which, as we show below, (14) is an *exact* LP formulation of problem (11), namely $Z^{\mathrm{LP}}(\mathbf{R}) = Z^{\mathrm{OPT}}(\mathbf{R})$. We thus define a subset $R(S) \subseteq \Re^n$ of reward vectors as the *domain* of the *adaptive-greedy* algorithm AG(S) described in Figure 1. Namely, as the set of reward vectors $\mathbf{R}$ for which the algorithm returns an output having $FAIL = 0$, when fed with input $(\mathbf{R}, (A_i^S)_{i\in S, S\in S}, S)$. Note that $R(2^N) = \Re^n$.

Definition 3 (PCL/GCL-indexability) *Scheduling problem (11) is said to be* PCL-indexable with respect to $S \subseteq 2^N$ *if*

(i) *it satisfies PCL with respect to* S*; and*

(ii) $\mathbf{R} \in R(S)$.

In the case where (i) holds with $S = 2^N$ *we say the problem is* GCL-indexable.

The next result shows that PCL-indexable problems are indeed indexable. Let the output of adaptive-greedy algorithm AG(S), when fed with input $(\mathbf{R}, (A_i^S)_{i\in S, S\in S}, S)$, be given by $\gamma = (\gamma_i)_{i\in N}$, $\overline{\mathbf{y}}$, K, $\{S_k\}_{k=1}^K$, $\{G_k\}_{k=1}^K$, and $FAIL$).

Theorem 1 (Partial indexability under PCL) *Assume problem (11) is PCL-indexable. 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

Since LP problem (14) is feasible and bounded, we know its dual problem has the same optimum

value. We can thus write

$$
\begin{aligned}
Z^{\mathrm{LP}}(\mathbf{R}) \quad = \quad & \min \sum_{S \in \mathcal{S}} b(S) y^S \\
& \text{subject to} \\
& \sum_{S \in \mathcal{S}: S \ni i} A_i^S y^S \geq R_i, \quad i \in N \\
& y^S \leq 0, \quad S \in \mathcal{S} \setminus \{N\}.
\end{aligned} \tag{15}
$$

Now, it is easily checked that the vector $\bar{\mathbf{y}}$ produced by adaptive-greedy algorithm AG($\mathcal{S}$) is a feasible solution to dual LP (15), which further satisfies

$$
R_i = \sum_{1 \leq k \leq K: S_k \ni i} A_i^{S_k} \bar{y}^{S_k}, \quad \text{for } i \in N .
$$

Let $\mathbf{x}^*$ be a performance vector achieved by a priority policy that gives higher priorities to classes with larger indices γ_i. Then, by the definition of PCL, it follows that $\mathbf{x}^*$ is a primal feasible solution for LP problem (14), which further satisfies the identities

$$
\sum_{i \in S_k} A_i^{S_k} x_i^* = b(S_k), \quad k = 1, \ldots, K.
$$

Therefore, it follows by LP complementary slackness that $\mathbf{x}^*$ and $\bar{\mathbf{y}}$ are an optimal primal-dual pair for linear programs (14) and (15). Since we assumed $\mathbf{x}^*$ to be achievable by a priority policy, this fact combined with the inequality $Z^{\mathrm{LP}}(\mathbf{R}) \geq Z^{\mathrm{OPT}}(\mathbf{R})$ implies $Z^{\mathrm{LP}}(\mathbf{R}) = Z^{\mathrm{OPT}}(\mathbf{R})$, i.e., such priority policy must be optimal. □

Note that in the special case where $\mathcal{S} = 2^N$, algorithm AG(2^N) reduces to the standard adaptive-greedy algorithm due to Klimov [7], and Theorem 1 yields the optimality of priority-index policies under all reward vectors, as established in [1].

Index decomposition

It was shown in [1] that, in the GCL framework (which corresponds to the special case where $\mathcal{S} = 2^N$ in the PCL above), a stronger *index decomposition* property holds under certain conditions. This explains that in certain models, such as classical multiarmed bandits, the priority indices for a project

Input: $(\mathbf{R}, \mathbf{A}, S)$,
where $\mathbf{R} = (R_i)_{i \in N}$, $\mathbf{A} = (A_i^S)_{i \in S, S \in S}$, and $S \subseteq 2^N$, with $N \in S$.
Output: $(\gamma, \bar{\mathbf{y}}, K, \{S_k\}_{k=1}^K, \{G_k\}_{k=1}^K, FAIL)$,
where $\gamma = (\gamma_i)_{i \in N}$, $1 \le K \le N$, $\bar{\mathbf{y}} = (\bar{y}^S)_{S \in S}$, $S_k, G_k \subseteq N$, for $1 \le k \le K$, and $FAIL \in \{0,1\}$.

Initialization. $S_1 = N$; $\bar{y}^{S_1} = \max\{R_i / A_i^N : i \in S_1\}$;
$G_1 = \arg\max\{R_i / A_i^N : i \in S_1\}$;
$\gamma_i = \bar{y}^{S_1}$, $i \in G_1$; $k := 1$.

Loop. **while** $S_k \neq G_k$ **and** $S_k \setminus G_k \in S$ **do**
begin
$k := k+1$; $S_k = S_{k-1} \setminus G_{k-1}$;
$\bar{y}^{S_k} = \max\{[R_i - \sum_{j=1}^{k-1} A_i^{S_j} \bar{y}^{S_j}] / A_i^{S_k} : i \in S_k\}$;
$G_k = \arg\max\{[R_i - \sum_{j=1}^{k-1} A_i^{S_j} \bar{y}^{S_j}] / A_i^{S_k} : i \in S_k\}$;
$\gamma_i = \sum_{j=1}^{k} \bar{y}^{S_j}$, $i \in G_k$.
end

Final step. $K = k$;
for $S \in S \setminus \{S_1, \ldots, S_K\}$, $\bar{y}^S = 0$;
if $S_k \neq G_k$ **then** $FAIL := 1$ **else** $FAIL := 0$.

Figure 1: Adaptive-greedy algorithm AG(S).

depend only on its defining parameters, and not on those of other projects. Assume the complete set of job classes N is partitioned into subsets $N_1, \ldots, N_M \in S$. Job classes in N_m are typically interpreted as corresponding to a *project m*, for $m = 1, \ldots, M$. Let us define

$$S_m = \{S \in S : S \subseteq N_m\}, \quad \text{for } m = 1, \ldots, M,$$

and assume that, if $S \in S$, then $S \cap N_m \in S_m$, for $m = 1, \ldots, M$. Assume further that the weights $A_i^S > 0$, for $i \in S$, $S \in S_m$, depend only on characteristics of job classes in N_m (i.e., on project m's defining parameters). A trivial extension to the argument given for Theorem 6 in [1] shows that, under the condition

$$A_i^S = A_i^{S \cap N_m}, \quad \text{for } i \in S \cap N_m \quad \text{and} \quad S \in S_m, \tag{16}$$

it holds that

$$\gamma_i = \gamma_i^m, \qquad \text{for } i \in N_m,$$

where the γ_i^m's, for $i \in N_m$, are the indices computed by running adaptive-greedy algorithm AG(S_m) on input $((R_i)_{i\in N_m}, (A_i^S)_{i\in S, S\in S_m})$.

Theorem 2 (Index decomposition) *Under condition (16), the priority indices corresponding to job classes in N_m depend only on characteristics of project m, and can be computed by running the adaptive-greedy algorithm* AG(S_m) *on input* $((R_i)_{i\in N_m}, (A_i^S)_{i\in S, S\in S_m})$.

4 PCL for restless bandits: the discounted case

In this section we apply the PCL framework developed in Section 3 to investigate Whittle's indexability property for restless bandits. We focus here on the discounted case, deferring discussion of the undiscounted case to the next section.

We thus consider a single restless bandit, as described in Section 2: it is modeled as a discrete-time MDC, having state space $N = \{1, \ldots, n\}$, transition probability matrices $\mathbb{P}^a = (p_{ij}^a)_{i,j\in N}$, 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 factor $0 < \beta < 1$. The initial state probabilities are given by vector $\alpha = (\alpha_i)_{i\in N}$, where α_i is the 0/1 indicator of the initial project state being i. Our concern is to identify sufficient conditions on model parameters under which the bandit is indexable. From the definition of indexability, we thus need to investigate the parametric family of single-project subproblems whose LP formulation was given in (10), which we formulate here as

$$Z^{\mathrm{OPT}}(\gamma; \mathbf{R}^1, \mathbf{R}^0) = \max_{u\in U} E_u \left[\sum_{t=0}^{\infty} \sum_{i\in N} R_i^1 I_i^1(t) \beta^t + \sum_{t=0}^{\infty} \sum_{i\in N} (R_i^0 + \gamma) I_i^0(t) \beta^t \right], \tag{17}$$

for each value of the passive subsidy $\gamma \in \Re$. In (17), $I_i^a(t)$ represents the indicator corresponding to taking action $a \in \{0, 1\}$ at time t, and U denotes the space of stationary policies. We note that the standard LP formulation of problem (17) given in (10) can be rewritten, using vector notation, as

$$\begin{aligned} Z^{\mathrm{OPT}}(\gamma; \mathbf{R}^1, \mathbf{R}^0) \quad = \quad & \mathbf{x}^1 \mathbf{R}^1 + \mathbf{x}^0 (\mathbf{R}^0 + \gamma \mathbf{1}) \\ & \text{subject to} \\ & \mathbf{x}^1 (\mathbb{I} - \beta \mathbb{P}^1) + \mathbf{x}^0 (\mathbb{I} - \beta \mathbb{P}^0) = \alpha \\ & \mathbf{x}^1, \mathbf{x}^0 \geq \mathbf{0}, \end{aligned} \tag{18}$$

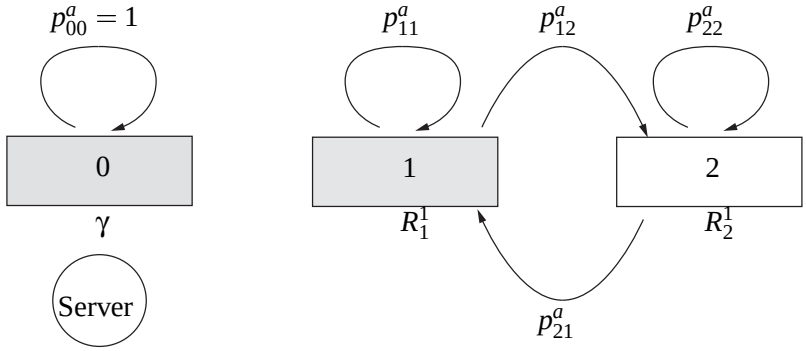

Figure 2: A single restless bandit seen as a multiclass service system.

where $\mathbf{x}^a = (x^a_j)_{j\in N}$, for $a \in \{0,1\}$, and α are taken to be row vectors, and $\mathbf{1}$ denotes an n-vector of ones. Note further that the constraints of LP problem (18) imply that

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

A straightforward consequence of this observation and LP formulation (18) is the identity

$$Z^{\mathrm{OPT}}(\gamma;\mathbf{R}^1,\mathbf{R}^0) = \alpha(\mathbb{I} - \beta\mathbb{P}^0)^{-1}\mathbf{R}^0 + Z^{\mathrm{OPT}}(\gamma;\mathbf{R}^1 - (\mathbb{I} - \beta\mathbb{P}^1)(\mathbb{I} - \beta\mathbb{P}^0)^{-1}\mathbf{R}^0,\mathbf{0}). \qquad (19)$$

In light of (19), we shall focus our subsequent analysis on the case $\mathbf{R}^0 = \mathbf{0}$, without loss of generality.

In order to apply the PCL framework to analyze problem (17), 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 active, and no reward otherwise.

We thus identify a class i job with a project in state $i \in N_0 = \{0\} \cup N$ (where the label of the calibrating project's single state is denoted by 0). We consider, as before, that the space U of admissible scheduling policies consists of all stationary policies, which base decisions on the current project state. To each policy $u \in U$ we associate performance measures $x^a_i(u)$, for $i \in N_0$ and $a \in \{0,\ 1\}$, as defined in (3). Note that $x^1_0(u)$ represents the total expected discounted time the original project is passive, or, equivalently, that the calibrating single-state project is active.

Figure 2 represents a simple example, where the original project has two states ($N = \{1,2\}$), and is currently in state 1: that is, there is one job of class 0 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 $\mathbb{P} = (p_{ij})_{i,j\in N}$, and subsets $S,T \subseteq N$, we shall write $\mathbf{x}_S = (x_i)_{i\in S}$ and $\mathbb{P}_{ST} = (p_{ij})_{i\in S, j\in T}$. 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 V_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 V_i^S's are determined as the unique solution to the system of linear equations

$$\begin{aligned} V_i^S &= 1 + \beta \sum_{j\in N} p_{ij}^1 V_j^S, \quad \text{for } i \in S \\ V_i^S &= \beta \sum_{j\in N} p_{ij}^0 V_j^S, \quad \text{for } i \in S^c. \end{aligned}$$

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

$$\mathbf{V}_S^S = \mathbf{1}_S + \beta \mathbb{P}_{SS}^1 \mathbf{V}_S^S + \beta \mathbb{P}_{SS^c}^1 \mathbf{V}_{S^c}^S \tag{20}$$

$$\mathbf{V}_{S^c}^S = \beta \mathbb{P}_{S^cS}^0 \mathbf{V}_S^S + \beta \mathbb{P}_{S^cS^c}^0 \mathbf{V}_{S^c}^S, \tag{21}$$

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

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

$$A_i^{S^c} = 1 + \beta \sum_{j\in N} (p_{ij}^1 - p_{ij}^0) V_j^S. \tag{22}$$

Note that

$$A_i^{\emptyset} = A_i^N = 1, \quad \text{for } i \in N.$$

Furthermore, it is straightforward from (20)–(22) that

$$\mathbf{A}_S^S = \mathbf{1}_S + \beta\,\mathbb{P}^1_{SN}\,\mathbf{V}_N^{S^c} - \mathbf{V}_S^{S^c}, \tag{23}$$

$$\mathbf{A}_{S^c}^S = \mathbf{V}_{S^c}^{S^c} - \beta\,\mathbb{P}^0_{S^cN}\,\mathbf{V}_N^{S^c}, \tag{24}$$

where $\mathbf{A}_S^S = (A_i^S)_{i\in S}$ and $\mathbf{A}_{S^c}^S = (A_i^S)_{i\in S^c}$. We further define $A_i^{\{0\}\cup S}$, for $i \in \{0\}\cup S$ and $S \subseteq N$, by

$$A_i^{\{0\}\cup S} = \begin{cases} A_i^S & \text{if } i \in S \\ 1 & \text{if } i = 0. \end{cases}$$

We complete the definitions by letting $b(T)$, for $T \subseteq N_0$, be given by

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

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 is shown to decompose, under any admissible policy, as the sum of a policy-invariant term plus a linear combination of passive performance measures. We note that such identities are analogous to the *work decomposition laws* satisfied by certain single-server multiclass queueing systems (cf. Theorem 5 in [3]): in the latter, a linear combination of mean queue lengths corresponding to a class subset is decomposed as the sum of a policy-invariant term plus terms relating to the mean workload when other classes are in service, or when the server is idle.

Lemma 3 (Decomposition laws) *For any admissible policy u and for any $S \subseteq N$,*

$$x_0^1(u) + \sum_{i\in S} A_i^S\, x_i^1(u) = b(\{0\}\cup S) + \sum_{i\in S^c} A_i^S\, x_i^0(u); \tag{25}$$

in particular, for $S = N$,

$$x_0^1(u) + \sum_{i\in N} x_i^1(u) = b(N_0).$$

Proof

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

$$\mathbf{x}^0 (\mathbb{I} - \beta \mathbb{P}^0) + \mathbf{x}^1 (\mathbb{I} - \beta \mathbb{P}^1) = \alpha. \tag{26}$$

We now rewrite system (26), in terms of a given subset $S \subset N$, as

$$\begin{bmatrix} \mathbf{x}_S^0 & \mathbf{x}_{S^c}^0 \end{bmatrix} \begin{bmatrix} \mathbb{I}_S - \beta \mathbb{P}_{SS}^0 & -\beta \mathbb{P}_{SS^c}^0 \\ -\beta \mathbb{P}_{S^cS}^0 & \mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^0 \end{bmatrix} + \begin{bmatrix} \mathbf{x}_S^1 & \mathbf{x}_{S^c}^1 \end{bmatrix} \begin{bmatrix} \mathbb{I}_S - \beta \mathbb{P}_{SS}^1 & -\beta \mathbb{P}_{SS^c}^1 \\ -\beta \mathbb{P}_{S^cS}^1 & \mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^1 \end{bmatrix} = \begin{bmatrix} \alpha_S & \alpha_{S^c} \end{bmatrix},$$

or, equivalently,

$$\begin{aligned} \mathbf{x}_S^0 (\mathbb{I}_S - \beta \mathbb{P}_{SS}^0) &= \alpha_S + \beta \mathbf{x}_{S^c}^0 \mathbb{P}_{S^cS}^0 + \beta \mathbf{x}_{S^c}^1 \mathbb{P}_{S^cS}^1 - \mathbf{x}_S^1 (\mathbb{I}_S - \beta \mathbb{P}_{SS}^1) \\ \mathbf{x}_{S^c}^1 (\mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^1) &= \alpha_{S^c} + \beta \mathbf{x}_S^0 \mathbb{P}_{SS^c}^0 + \beta \mathbf{x}_S^1 \mathbb{P}_{SS^c}^1 - \mathbf{x}_{S^c}^0 (\mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^0). \end{aligned}$$

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

$$\begin{aligned} \mathbf{x}_{S^c}^1 \left[\mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^1 - \beta^2 \mathbb{P}_{S^cS}^1 (\mathbb{I}_S - \beta \mathbb{P}_{SS}^0)^{-1} \mathbb{P}_{SS^c}^0 \right] &= \alpha_{S^c} + \beta \alpha_S (\mathbb{I}_S - \beta \mathbb{P}_{SS}^0)^{-1} \mathbb{P}_{SS^c}^0 \\ &\quad + \beta \mathbf{x}_S^1 \left[\mathbb{P}_{SS^c}^1 - (\mathbb{I}_S - \beta \mathbb{P}_{SS}^1) (\mathbb{I} - \beta \mathbb{P}_{SS}^0)^{-1} \mathbb{P}_{SS^c}^0 \right] \\ &\quad - \mathbf{x}_{S^c}^0 \left[\mathbb{I}_{S^c} - \beta \mathbb{P}_{S^cS^c}^0 - \beta^2 \mathbb{P}_{S^cS}^0 (\mathbb{I}_S - \beta \mathbb{P}_{SS}^0)^{-1} \mathbb{P}_{SS^c}^0 \right]. \end{aligned}$$

Now, postmultiplying both sides of the above equation by $\mathbf{V}_{S^c}^{S^c}$, and simplifying the resulting expression using (20)–(24), and the identity (which follows from the definition of $\mathbf{A}_S^S$)

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

we obtain

$$\mathbf{x}^1_{S^c}\,\mathbf{1}_{S^c} = \alpha\,\mathbf{V}^{S^c} + \mathbf{x}^1_S\,\left[\mathbf{A}^S_S - \mathbf{1}_S\right] - \mathbf{x}^0_{S^c}\,\mathbf{A}^S_{S^c}.$$

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

$$x^1_0 + \mathbf{x}^1_S\,\mathbf{1}_S + \mathbf{x}^1_{S^c}\,\mathbf{1}_{S^c} = x^1_0 + \sum_{i\in N} x^1_i = \frac{1}{1-\beta}.$$

Combining the last two equations yields

$$x^1_0 + \mathbf{x}^1_S\,\mathbf{A}^S_S = \frac{1}{1-\beta} - \alpha\,\mathbf{V}^{S^c} + \mathbf{x}^0_{S^c}\,\mathbf{A}^S_{S^c},$$

which is precisely (25). The case $S = N$ is immediate. □

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

(a) *for any admissible policy u that gives priority to job classes in S^c over class* 0 *(i.e., that takes the active action over S^c),*

$$x^1_0(u) + \sum_{i\in S} A^S_i\,x^1_i(u) = b(\{0\}\cup S); \tag{27}$$

(b) *for any admissible policy u that gives priority to job class* 0 *over classes in S (i.e., that takes the passive action over S),*

$$\sum_{i\in S} A^S_i\,x^1_i(u) = 0 = b(S). \tag{28}$$

Proof

(a) Note that, under any such policy, it holds that $x^0_i(u) = 0$, for $i \in S^c$. This fact, together with Lemma 3, proves the result.

(b) The result follows by observing that, under any such policy, $x^1_i(u) = 0$, for $i \in S$. □

Let us now consider the family S^* of class/state subsets defined by

$$S^* = \{S \subseteq N \ : A_i^S > 0, i \in S \quad \text{and} \quad A_i^S \geq 0, i \in S^c\}, \tag{29}$$

and let us then define

$$S_0^* = S^* \cup \{\{0\} \cup S, S \in S^*\}.$$

Note that, as required in the PCL framework, we have $N_0 \in S_0^*$. Note further that, since $A_i^{\emptyset} \equiv A_i^N \equiv 0$, we shall consider that $\emptyset \in S^*$.

Theorem 4 (PCL: discounted restless bandits) *Performance vector $(x_i^1(u))_{i \in N_0}$ satisfies PCL with respect to S_0^*. Namely, for any class subset $S \in S^*$ and policy $u \in U$, the inequalities*

$$x_0^1(u) + \sum_{i \in S} A_i^S x_i^1(u) \geq b(\{0\} \cup S), \tag{30}$$

and

$$\sum_{i \in S} A_i^S x_i^1(u) \geq b(S), \tag{31}$$

hold, together with the identities

$$x_0^1(u) + \sum_{i \in S} A_i^S x_i^1(u) = b(\{0\} \cup S), \quad \textit{if } u \textit{ gives priority to } S^c \textit{ over } 0, \tag{32}$$

$$\sum_{i \in S} A_i^S x_i^1(u) = b(S), \quad \textit{if } u \textit{ gives priority to } 0 \textit{ over } S, \tag{33}$$

and

$$x_0^1(u) + \sum_{i \in S} x_i^1(u) = b(N_0). \tag{34}$$

Furthermore, if there exist $\hat{S} \subset N$ having $A_i^{\hat{S}} > 0$ for $i \in \hat{S}$, and $\hat{\imath} \in \hat{S}^c$ with $A_{\hat{\imath}}^{\hat{S}^c} < 0$, then, for any initial state vector $\alpha > \mathbf{0}$ (componentwise), inequality (30) does not hold.

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 equivalently formulated as (30)–(34). Note next that the requirement that all the A_i^S coefficients arising in the partial conservation laws be positive is guaranteed to hold by our definition of S^*. The structure of S^*, together with the decomposition laws in Lemma 3, further yields directly the inequalities (30)–(31). The required identities (32)–(34) were established in Corollary 1.

The last statement follows from decomposition identity (25) in Lemma 3 as it applies to subset $\hat{S}$: any policy u taking the passive action in state $\hat{i}$ and the active action in states $i \in \hat{S}^c \setminus \{\hat{i}\}$ has

$$x_0^1 + \sum_{i \in \hat{S}} A_i^{\hat{S}} x_i^1(u) = b(\{0\} \cup \hat{S}) + A_{\hat{i}}^{\hat{S}} x_{\hat{i}}^0(u) < b(\{0\} \cup \hat{S}).$$

□

We next present an adaptation of Definition 3 to the specific restless bandit model under consideration. Its equivalence with Definition 3 is apparent from Theorem 4. For each $S \subseteq 2^N$, let us define $R(S) \subseteq \Re^n$ as the set of active reward vectors $\mathbf{R}^1$ such that, when algorithm AG(S) is fed with input $(\mathbf{R}^1, (A_i^S)_{i \in S, S \in S}, S)$, it returns an output having $FAIL = 0$.

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 $S \subseteq 2^N$ *if the following two conditions hold:*

(i) $A_i^S > 0, \quad i \in S, \qquad$ *and* $\qquad A_i^{S^c} \geq 0, \quad i \in S^c, \quad$ *for* $S \in S$*;*

(ii) $\mathbf{R}^1 \in R(S)$.

If $S = 2^N$*, we say the bandit is* GCL-indexable.

Note that Theorem 4 implies that S^* is the largest class set family with respect to which a bandit can be PCL-indexable.

For a given $\mathbf{R}^1 \in R(S)$, let $\gamma = (\gamma_i)_{i \in N}$ be the index vector returned by algorithm AG(S) when fed with input $(\mathbf{R}^1, (A_i^S)_{i \in S, S \in S}, S)$.

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

(a) *A bandit that is PCL-indexable with respect to $S \subseteq 2^N$ is indexable for any reward vector $\mathbf{R}^1 \in R(S)$, with indices γ_i, for $i \in N$;*

(b) *A GCL-indexable bandit is indexable for any reward vector.*

Proof

The fact that, for $\mathbf{R}^1 \in R(S)$, the subproblem discussed above is solved optimally by a priority-index policy, where the optimal indices are computed by adaptive-greedy algorithm AG(S), is a straightforward consequence of combining Theorem 1 and Theorem 4. The result now follows by observing that the index decomposition condition (16) holds, when applied to the natural decomposition of class set N_0 into $\{0\}$ and N, and therefore Theorem 2 applies. This yields that the priority index for auxiliary class/state 0 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 indexability* of a restless bandit: the test is based on checking whether $\mathbf{R}^1 \in R(S)$, which involves a single run of adaptive-greedy algorithm AG(S). We remark, however, that this test represents a sufficient, but not necessary, condition for a bandit to be indexable.

We presented in (18) the standard LP formulation of problem (17), known from MDC theory. The PCL framework provides a new, equivalent LP reformulation for PCL-indexable bandits, as shown next.

Corollary 3 *Suppose a bandit (normalized so that $\mathbf{R}^0 = 0$) is PCL-indexable with respect to $S \subseteq 2^N$. Then, letting*

$$S_0 = S \cup \{\{0\} \cup S : S \in S\},$$

problem (17), can be formulated as the linear program

$$\begin{aligned} Z^{\mathrm{OPT}}(\gamma; \mathbf{R}^1, \mathbf{0}) \quad = \quad & \max \gamma x_0^1 + \sum_{i \in N} R_i^1 x_i^1 \\ & \textit{subject to} \\ & \sum_{i \in S} A_i^S x_i^1 \geq b(S), S \in S_0 \setminus \{N_0\} \\ & \sum_{i \in N_0} A_i^{N_0} x_i^1 = b(N_0) \\ & x_i^1 \geq 0, i \in N_0. \end{aligned}$$

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

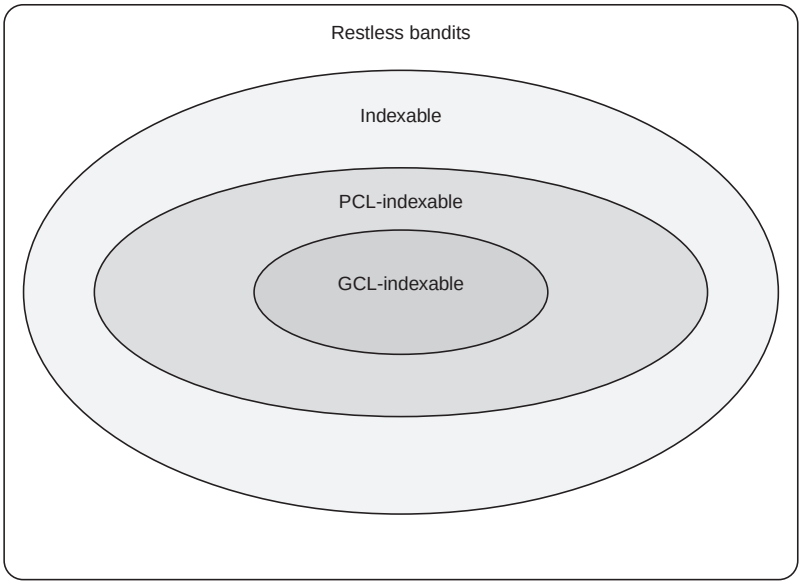

Figure 3: Classification of restless bandits.

Corollary 4 *Classical bandits are GCL-indexable.*

Proof

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

$$\begin{aligned} V_i^S &= 1 + \beta \sum_{j \in S} p_{ij}^1 V_j^S, \quad \text{for } i \in S \\ V_i^S &= 0, \quad \text{for } i \in S^c. \end{aligned}$$

The inequalities in Definition 4(i) follow immediately. □

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 as the discount factor* β *gets small enough.*

Figure 3 illustrates the classification of restless bandits resulting 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, we note that, for any $S \subset N$, we can write

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

Let us focus on a given $S \subset N$. Recall the interpretation of V_i^S as the total expected discounted time 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 $i \in S^c$, states therefore that, by changing the initial action (in state *i*) from passive to active, and following the *S*-active policy onwards, the project expected active time becomes larger. Furthermore, the condition $A_i^{S^c} \geq 0$, for $i \in S$, states that, by changing the initial action (in state *i*) from active to passive, and following the *S*-active policy onwards, the project expected active time does not become larger. Note further from the above discussion that the coefficients $A_i^{S^c}$ represent *active time differentials* corresponding to changing the initial action in state *i* under the *S*-active policy.

5 PCL for restless bandits: the time-average case

In this section we outline how the results in Section 4 for the time-discounted case can be extended to analyze 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 additional requirement stated next.

Assumption 1 (Ergodicity) *For any subset $S \subseteq N$ of bandit states, the Markov chain over N having transition probability matrix*

$$\mathbb{P}(S) = \begin{bmatrix} \mathbb{P}_{SN}^1 \\ \mathbb{P}_{S^cN}^0 \end{bmatrix} \tag{35}$$

is ergodic.

Note that $\mathbb{P}(S)$, defined by (35), 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 passive reward vector is $\mathbf{R}^0 = \mathbf{0}$, since the general case can be equivalently reduced to it.

Our approach to the time-average case is based on studying the asymptotic behavior, as the discount factor $\beta \nearrow 1$, of the quantities and results appearing in our analysis of the discounted case. Hence, to avoid confusion, in what follows we shall make explicit the dependence on discount factor β in the quantities defined in Section 4, writing, e.g., $V_i^S(\beta)$, $A_i^S(\beta)$ and $x_i^u(u,\beta)$.

To relate the discounted and the time-average cases we shall apply the result that, under Assumption 1, for any bandit state $i \in N$ and state subset $S \subseteq N$, we can express $V_i^S(\beta)$ as follows:

$$V_i^S(\beta) = \frac{V^S}{1-\beta} + v_i^S + O(1-\beta) \quad \text{as } \beta \nearrow 1. \tag{36}$$

In identity (36), V^S represents the *long-run time-average fraction of time the bandit is active under the S-active policy*, whereas v_i^S represents the corresponding *total expected active time differential due to starting in state i*. Substituting for the $V_i^S(\beta)$'s in equations (20)–(21) using (36), and letting $\beta \nearrow 1$, we obtain the system of linear equations

$$V^S + v_i^S = 1 + \sum_{j \in N} p_{ij}^1 v_j^S, \quad \text{for } i \in S \tag{37}$$

$$V^S + v_i^S = \sum_{j \in N} p_{ij}^0 v_j^S, \quad \text{for } i \in S^c. \tag{38}$$

Note that equations (37)–(38) determine v_i^S, for $i \in N$, in terms of V^S. Furthermore, V^S can be computed as the sum of the equilibrium probabilities for states in S corresponding to the ergodic Markov chain having transition probability matrix $\mathbb{P}(S)$.

We next apply (36)–(38) to study the asymptotics of the coefficients $A_i^S(\beta)$ appearing in the PCL obtained in the discounted case. Substituting for the $V_i^S(\beta)$'s in equations (22) using (36), and letting $\beta \nearrow 1$, yields the result that each $A_i^S(\beta)$ converges to a limit A_i^S, given by

$$A_i^S = 1 + \sum_{j \in N} (p_{ij}^1 - p_{ij}^0) v_j^S, \quad \text{for } i \in N, S \subseteq N. \tag{39}$$

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 it holds that, for any stationary policy u,

$$\mathbf{x}^a(u) = \lim_{\beta \nearrow 1} (1-\beta)\, \mathbf{x}^a(u, \beta). \tag{40}$$

In order to investigate the indexability property under the time-average criterion we consider the equivalent two-project restless bandits model discussed in Section 4, where an auxiliary *calibrating project* having a single state 0 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 $A_i^{\{0\} \cup S}$, for $i \in \{0\} \cup S$ and $S \subseteq N$, by

$$A_i^{\{0\} \cup S} = \begin{cases} A_i^S & \text{if } i \in S \\ 1 & \text{if } i = 0. \end{cases}$$

We complete the definitions by letting $b(T)$, for $T \subseteq \{0\} \cup N$, be given by

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

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

6 Examples

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

6.1 Two-state restless bandits

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 2. In the discounted case, the relevant A_i^S coefficients are readily calculated as follows:

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

It follows that, for any discount factor $0<\beta<1$, $A_i^{\{1\}}, A_i^{\{2\}} > 0$, for $i=1,2$. Therefore, it follows that discounted two-state restless projects are GCL-indexable, 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 1 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 AG($2^{\{1,2\}}$) are

$$\gamma_1 = R_1$$

and

$$\begin{aligned}
\gamma_2 &= R_1 + \frac{R_2 - A_2^{\{1,2\}} R_1}{A_2^{\{2\}}} \\
&= R_1 - \frac{1+\beta-\beta\,p_{22}^0-\beta\,p_{11}^1}{1+\beta-\beta\,p_{11}^1-\beta\,p_{22}^1}\,(R_1-R_2) \\
&= \frac{\beta\,(p_{22}^0-p_{22}^1)\,R_1 + (1+\beta-\beta\,p_{22}^0-\beta\,p_{11}^1)\,R_2}{1+\beta-\beta\,p_{11}^1-\beta\,p_{22}^1}.
\end{aligned}$$

As will be seen in the next example, three-state restless bandits need not be GCL-indexable.

6.2 An indexable three-state restless bandit which is not GCL-indexable

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

$$\mathbb{P}^1 = \begin{pmatrix} \varepsilon & 0 & 1-\varepsilon \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \quad \text{and} \quad \mathbb{P}^0 = \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \\ \varepsilon & 0 & 1-\varepsilon \end{pmatrix},$$

for $0 < \varepsilon < 1$. 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 $0 < \beta, \varepsilon < 1$. Note, however, that $A_1^{\{3\}}$ can be made negative for any $0 < \varepsilon < 2/5$ by choosing the discount factor β close enough to unity. Taking, e.g., $\varepsilon = 1/9$ and $\beta = 3/4$, we obtain $A_1^{\{3\}} = -1/12$, $A_2^{\{3\}} = 1$ and $A_3^{\{3\}} = 14/9$, which we can substitute in the decomposition identity (25) corresponding to $S = \{3\}$: for any admissible policy u,

$$x_0^1(u) + A_3^{\{3\}} x_3^1(u) = b(\{0,3\}) + A_1^{\{3\}} x_1^0(u) + A_2^{\{3\}} x_2^0(u). \tag{41}$$

Note that $S = \{3\} \notin \mathcal{S}$ (see 29), and, therefore, the bandit problem concerned with minimizing the objective

$$x_0^1(u) + A_3^{\{3\}} x_3^1(u)$$

under stationary policies u is not GCL-indexable. Numerical investigation reveals that, however, the bandit is indexable.

6.3 Threshold optimality through PCL

We illustrate here through a small example how PCL provide an appropriate framework for investigating the optimality of threshold policies in queueing systems. This issue is studied in general in [8]. Consider a discrete-time single-server queue which can hold at most 2 customers. The set of system states, i.e., of possible number of customers in system (waiting or in service), is thus $N = \{0,1,2\}$. At each time the system manager can choose to activate the arrival stream, if there are less than 2 customers in system, or to shut off arrivals. Corresponding state-dependent rewards are earned when the arrival stream is active, which are discounted in time with factor $0 < \beta < 1$. The objective is to design an input 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

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

where $\lambda, \mu > 0$, with $\lambda + \mu = 1$. The state-dependent active rewards are R_0, R_1 and R_2, respectively. To avoid confusion, note that in this model 0 denotes a system state (empty queue), and not the auxiliary state representing the idle action, as in our general analysis of restless bandits through PCL.

The optimality of threshold policies (i.e., shut off arrivals when the number in system exceeds a given threshold value) will be deduced, under appropriate conditions on reward coefficients, from the fact that the bandit satisfies PCL with respect to

$$S = \{\{2\},\{1,2\},\{0,1,2\}\}.$$

See Definition 4. 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, the bandit satisfies PCL with respect to S. Let us now characterize the corresponding reward subset $R(S)$, as the domain of adaptive-greedy algorithm AG(S). The conditions defining $R(S)$ are easily seen to be as follows:

$$R_0 \geq \max(R_1, R_2) \qquad \text{and} \qquad \frac{R_1-R_0}{A_1^{\{1,2\}}} \geq \frac{R_2-R_0}{A_2^{\{1,2\}}}. \tag{42}$$

Therefore, under conditions (42) on reward coefficients, the bandit is PCL-indexable with respect to S. The corresponding Whittle indices, as given by adaptive-greedy algorithm AG(S) are

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

7 Conclusions

We identified a class of restless bandits, called PCL-indexable, 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, which also computes the Whittle indices when the test is positive. Given the rich modeling power of restless bandits, we believe the notion of PCL-indexability introduced in this paper opens the way to analyze a variety of stochastic control models in, e.g., queueing systems, within a unifying, polyhedral framework.

Our analysis further reveals the power of the achievable region method (cf. [4]) to analyze stochastic optimization problems, as our analyses are based on new linear programming formula-

tions of the problems investigated. A previous study (see [2]) also utilized the achievable region approach to obtain a priority-index heuristic, different from Whittle's, and improved LP relaxations, for general (possibly nonindexable) restless bandits.

References

[1] Bertsimas, D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multi-armed bandit problems; a unified approach to indexable systems. *Math. Oper. Res.* **21** 257–306.

[2] Bertsimas, D. and Niño-Mora, J. (1994). Restless bandits, linear programming relaxations, and a primal-dual index heuristic. Working paper, Operations Research Center, MIT. Forthcoming in *Oper. Res.*

[3] 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. Oper. Res.* **24** 306–330.

[4] Dacre, M., Glazebrook, K.D. and Niño-Mora, J. (1999). The achievable region approach to the optimal control of stochastic systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **61** 747–791.

[5] Faihe, Y. and Müller, J.-P. (1998). Behaviors coordination using restless bandits allocation indexes. *Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior (SAB98)*.

[6] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 148–177.

[7] Klimov, G.P. (1974). Time sharing service systems I. *Theory Probab. Appl.* **19** 532–551.

[8] Niño-Mora, J. (1999). Threshold optimality in queueing systems: a polyhedral approach via partial conservation laws. Working paper, UPF.

[9] Papadimitriou, C. H. and Tsitsiklis, J. N. (1999). The complexity of optimal queueing network control. *Math. Oper. Res.* **24** 293–305

[10] Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York.

[11] Veatch, M. and Wein, L. M. (1996). Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.* **44** 634–647.

[12] Weber, R. R. and Weiss, G. (1990). On an index policy for restless bandits. *J. Appl. Prob.* **27** 637–648.

[13] Weber, R. R. and Weiss, G. (1991). Addendum to "On an index policy for restless bandits." *Adv. Appl. Prob.* **23** 429–430.

[14] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. In *A Celebration of Applied Probability*, J. Gani (Ed.), *J. Appl. Prob.* **25A** 287-298.