Restless bandits, linear programming relaxations and a primal-dual index heuristic

Dimitris Bertsimas[*] José Niño-Mora[†]

August 1994 - Revised October 1997
To appear in *Operations Research*

Abstract

We develop a mathematical programming approach for the classical $PSPACE-hard$ restless bandit problem in stochastic optimization. We introduce a hierarchy of n (where n is the number of bandits) increasingly stronger linear programming relaxations, the last of which is exact and corresponds to the (exponential size) formulation of the problem as a Markov decision chain, while the other relaxations provide bounds and are efficiently computed. We also propose a priority-index heuristic scheduling policy from the solution to the first-order relaxation, where the indices are defined in terms of optimal dual variables. In this way we propose a policy and a suboptimality guarantee. We report results of computational experiments that suggest that the proposed heuristic policy is nearly optimal. Moreover, the second-order relaxation is found to provide strong bounds on the optimal value.

Keywords: Stochastic scheduling, bandit problems, resource allocation, dynamic programming.
JEL: C60, C61.

[*] Sloan School of Management and Operations Research Center, MIT, Cambridge, MA 02139, dbertsim@aris.mit.edu. The research of the author was partially supported by a Presidential Young Investigator Award DDM-9158118 with matching funds from Draper Laboratory.

[†] Department of Economics and Business, Universitat Pompeu Fabra, E-08005 Barcelona, Spain, jose.nino-mora@econ.upf.es, http://www.econ.upf.es/~ninomora. This work is part of the author's PhD Dissertation at the Operations Research Center of MIT, and it was partially supported by a Ministry of Education and Science of Spain (MEC)/Fulbright Doctoral Fellowship. Preliminary versions of the paper were presented at the 1994 TIMS/ORSA Boston Meeting, and at the 1995 INFORMS Atlanta Applied Probability Meeting.

1 Introduction

Research in combinatorial optimization over the last twenty years has crystallized the idea that our ability to solve efficiently a combinatorial optimization problem depends critically on our ability to construct strong mathematical programming formulations for it. As a result, much research effort has centered in developing sharper formulations. The developments in the area of polyhedral combinatorics are strong witnesses of this trend.

The field of stochastic optimization has developed along quite different lines. Over the last twenty years it has addressed, with various degrees of success, several key problems that arise in areas as diverse as computer and communication networks, manufacturing and service systems. In contrast with the situation in combinatorial optimization, a characteristic of this body of research is the lack of a unified and practical solution method. While these problems can often be formulated in the framework of dynamic programming, the resulting formulations typically exhibit a prohibitively large size, which hinders their application. As a result, every problem is addressed via ad hoc techniques. Furthermore, the quality of proposed heuristics is usually assessed by comparing their empirical performance with that of alternative heuristics, which gives little information about their degree of suboptimality.

Motivated by the success of the mathematical programming approach to combinatorial optimization, we propose in this paper a solution approach to the classical restless bandit problem in stochastic scheduling based on linear programming (LP) formulations. This work is part of a larger program to solve hard stochastic optimization problems via a mathematical programming approach: the *performance region approach.*

Bacground: The performance region approach to stochastic optimization. The performance region approach was introduced in a seminal paper by Coffman and Mitrani (1980). It draws on the mathematical programming approach to optimization, as it seeks to characterize the region of achievable performance spanned by a system's performance measure under a class of admissible policies. The goal is to formulate explicitly this region by means of mathematical programming constraints. Since it may not be possible to formulate the exact performance region, we may have to settle for constructing a *relaxation* that contains it.

For example, given a vector $\boldsymbol{x}$ of performance measures, which must represent expectations (although not necessarily first moments) and a cost function $c(\boldsymbol{x})$, such as the class-weighted expected number of jobs in a multiclass queueing system, consider the problem of finding a *lower bound* $\underline{Z} \leq c(\boldsymbol{x})$ on the cost achievable under any scheduling policy.

We define $\mathcal{X}$ to be the performance region spanned by performance vector $\boldsymbol{x}$ under all admissible scheduling policies. The minimum cost achievable is

$$\begin{aligned} Z^* \quad = \quad & \min c(\boldsymbol{x}) \\ & \text{subject to} \\ & \boldsymbol{x} \in \mathcal{X}. \end{aligned}$$

Now let $\mathcal{P} \supseteq \mathcal{X}$ be a relaxation of the performance region defined by a set of constraints. A lower bound on Z^* is obtained by solving the mathematical program

$$\begin{aligned} \underline{Z} \quad = \quad & \min c(\boldsymbol{x}) \\ & \text{subject to} \\ & \boldsymbol{x} \in \mathcal{P}. \end{aligned}$$

The optimal solution to this mathematical program may also lead to good, or even optimal, scheduling policies.

The two critical problems the performance region approach must overcome for tackling a performance optimization problem are the following:

1. Generating constraints on the performance region.
2. Designing good policies from the solution to the resulting relaxations.

Coffman and Mitrani (1980) first addressed with this approach the problem of minimizing the class-weighted mean delay in a multiclass $M/M/1$ queue. They formulated exactly the system's performance region as a polyhedron, and showed that the well-known optimality of priority-index policies (the $c\mu$ rule) follows from structural properties of this polyhedron. The scope of this approach has since been extended to a range of increasingly more complex systems. Drawing on earlier work by Federgruen and Groenevelt (1988) and Shanthikumar and Yao (1992), Bertsimas and Niño-Mora (1996) developed a unifying framework for formulating the exact performance region in a wide variety of stochastic scheduling systems that satisfy work conservation laws (including the classical *multiarmed bandit* problem). They showed that the distinctive structural property of these stochastic optimization problems (optimality of priority-index policies) follows from a corresponding property of their underlying polyhedral performance regions.

Researchers have sought recently to extend further the scope of the performance region approach, with the aim of solving hard stochastic optimization problems, such as multiclass queueing network scheduling (see, e.g., Bertsimas, Paschalidis and Tsitsiklis (1994)).

In this paper we extend this line of research by addressing an important and intractable extension of the classical multiarmed bandit problem: the *restless bandit problem*, which Papadimitriou and Tsitsiklis (1994) have shown to be $PSPACE-hard$.

Contributions. Our contributions include the following:

1. We present a hierarchy of N LP relaxations for the restless bandit problem (N being the number of bandits). These relaxations are increasingly stronger at the expense of requiring increasing computations, and the last one (Nth) is exact. They can be interpreted geometrically in terms of the following *projection representation* idea, nicely outlined in Lovász and Schrijver (1991):

   > It has been recognized recently that to represent a polyhedron as the projection of a higher-dimensional, but simpler, polyhedron, is a powerful tool in polyhedral combinatorics ... The idea is that a projection of a polytope may have more facets than the polytope itself. This remark suggests that even if $\mathcal{P}$ has exponentially many facets, we may be able to represent it as the projection of a polytope $\mathcal{Q}$ in higher (but still polynomial) dimension, having only a polynomial number of facets.

2. We propose a primal-dual heuristic that defines priority indices in terms of optimal dual variables corresponding to the first-order relaxation. We report computational results which indicate that the heuristic is exceptionally accurate.

The primal-dual approach to heuristic design, which we pursue here, is based on, first, formulating an LP relaxation of the problem, and then designing a heuristic from optimal primal and dual LP solutions. This approach has a fruitful history in the field of combinatorial optimization.

Structure of the paper. The paper is structured as follows: In Section 2 we introduce the restless bandit problem and review previous work on it. In Section 3 we review a classical result on LP formulations for Markov decision chains, and use it as the basic tool to develop a hierarchy of LP relaxations for the problem, the last of which is exact. In Section 4 we present a primal-dual priority-index heuristic for the problem, based on the optimal solution to the first-order relaxation. In Section 5 we report the results of computational experiments on the tightness of the relaxations and the performance of the heuristic. We end the paper with some concluding remarks.

2 The restless bandit problem: Description, applications and background

The *restless bandit problem* we address is as follows: Consider a collection of N projects, labeled $n \in \mathcal{N} = \{1, \ldots, N\}$. Project n can be in one of a finite number of states $i_n \in \mathcal{S}_n$. At each discrete time epoch $t = 0, 1, 2, \ldots$, *exactly* $M < N$ projects must be worked on, or set *active*. If project n, in state i_n, is worked on, then an *active reward* $R^1_{i_n}$ is earned, and its state changes in a Markovian fashion, according to an *active transition probability* matrix (into state j_n with probability $p^1_{i_n j_n}$). If the project is not worked on, then a *passive reward* $R^0_{i_n}$ is received, and its state then changes according to a *passive transition probability* matrix (into state j_n with probability $p^0_{i_n j_n}$). Rewards are time-discounted by a discount factor $0 < \beta < 1$. Projects are selected over time according to a Markovian scheduling policy u. Let $\mathcal{U}$ denote the class of admissible (Markovian) policies. The problem consists in finding a scheduling policy that maximizes the total expected discounted reward over an infinite horizon, and in computing its optimum value:

$$Z^* = \max_{u \in \mathcal{U}} E_u \left[\sum_{t=0}^{\infty} \left(R^{a_1(t)}_{i_1(t)} + \cdots + R^{a_N(t)}_{i_N(t)} \right) \beta^t \right]. \tag{1}$$

In formulation (1) $i_n(t)$ and $a_n(t)$ denote the state and action (active or passive), respectively, corresponding to project n at time t. The initial project states are assumed to be known, and are given by a vector $\boldsymbol{\alpha}$, with as many components as project states, where

$$\alpha_{i_n} = \begin{cases} 1 & \text{if project } n \text{ is in state } i_n \text{ at time } t = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Notice that the restless bandit problem may be considered as an extension of the classical multiarmed bandit problem (see Gittins (1989)). The latter corresponds to the special case that exactly one project is active at every time (i.e., $M = 1$), and passive projects do not change state (i.e, $p^0_{i_n i_n} = 1$, $p^0_{i_n j_n} = 0$ for all $n \in N$, and $i_n \neq j_n$).

Applications. The restless bandit problem provides a powerful modeling framework, as illustrated by the following examples of applications:

Clinical trials (Whittle (1988)). In this setting, projects correspond to medical treatments. The state of a project represents one's state of knowledge on the effectiveness of the corresponding treatment. Operating a project corresponds to testing the treatment. If, for example, the virus that the treatments are trying to combat is mutating, then one's state of knowledge on the effectiveness of each treatment changes whether or not it is tested.

Aircraft surveillance (Whittle (1988)). M aircraft are trying to track N enemy submarines. The state of a project-submarine represents one's state of knowledge of the current position and velocity of that submarine. Operating a project corresponds to assigning an aircraft to track the corresponding submarine.

Worker scheduling (Whittle (1988)). A number M of employees out of a pool of N have to be set to work at any time. The state of a project-worker represents his state of tiredness. Selection of a project results in exhaustion of the corresponding worker, whereas nonselection results in recuperation.

Police control of drug markets. In this setting, M police units are trying to control N drug markets. The state of a project corresponds to the drug-dealing activity level of the corresponding drug market. A project/drug market selection corresponds to a police enforcement operation over that market, and tends to discourage drug-dealing activity. Nonselection of a project allows drug-dealing activity to grow in the corresponding market.

Control of a make-to-stock production facility (Veatch and Wein (1996)). In this setting, a production facility makes N product classes; each finished product is placed in its class-dependent inventory, which services an exogenous demand. Veatch and Wein (1996) formulate a lost-sales version of the problem as a restless bandit problem, where project states represent class inventory levels, and propose priority-index heuristic rules (based on Whittle's index policy mentioned below) that exhibit a good empirical performance.

Previous work. The restless bandit problem was first investigated by Whittle (1988), who studied a continuous-time version of the problem, with a time-average reward criterion, in a dynamic programming framework. He introduced a relaxed version of the problem, which can be solved optimally in polynomial time. Based on this solution he proposed a priority-index heuristic policy, which reduces to the optimal Gittins index policy in the special case of the multiarmed bandit problem. A disadvantage is that Whittle's index heuristic only applies to a restricted class of restless bandits: those satisfying a certain *indexability* property, which may be hard to check. Weber and Weiss (1990) investigated the asymptotic optimality, conjectured by Whittle, of his heuristic, as M and N grow to ∞, with M/N fixed. Working with continuous-time restless bandits with a time-average reward criterion, they presented a sufficient condition for asymptotic optimality. They also found instances that violate this condition, and in which Whittle's heuristic is not asymptotically optimal.

Another line of work has studied the computational complexity of the problem. Papadimitriou and Tsitsiklis (1994) established that the restless bandit problem is $PSPACE-$ *hard*, even in the special case of deterministic transition rules and $M = 1$. This result is in sharp constrast with the well-known optimality of Gittins priority-index rule in the special case of the multiarmed bandit problem, first established by Gittins and Jones (1974).

3 A hierarchy of LP relaxations for the restless bandit problem

In this section we develop a hierarchy of increasingly stronger LP relaxations for the restless bandit problem, the last of which is exact. The key tool for constructing these relaxations is a classical result on LP formulations of Markov decision chains (MDCs), which we review next.

3.1 LP formulations of MDCs

We consider a finite-state discrete-time MDC, with a discounted reward criterion. This represents the evolution of a controlled stochastic system over a finite *state space* $\mathcal{S} = \{1, \ldots, n\}$ and discrete time epochs $t = 0, 1, 2, \ldots$. When the system state is $i \in \mathcal{S}$ the controller must select an *action* from a finite set $\mathcal{A}_i$. If action $a \in \mathcal{A}_i$ is the one selected, then (1) the system state changes, at the next time epoch, to a new state j with a Markov transition probability p_{ij}^a, for $j \in \mathcal{S}$; and (2) a reward R_i^a is received, discounted in time by a discount factor $0 < \beta < 1$. The goal is to find a *Markovian policy* (which selects the current action as a function, possibly randomized, of the current state and time) which maximizes the total expected discounted reward over an infinite time horizon.

We denote $\mathcal{U}$ the class of all *admissible* (i.e., Markovian) policies. We shall further refer to the subclass of *stationary* policies, which are Markovian policies in which the action selection depends only on the current state, not on the current time. We further denote $\mathcal{C}$ the *state-action space*,

$$\mathcal{C} = \{(i, a) \colon i \in \mathcal{S}, a \in \mathcal{A}_i\},$$

and let α_i denote the probability that the initial state is i, for $i \in \mathcal{S}$. The vector $\boldsymbol{\alpha} = (\alpha_i)_{i \in \mathcal{S}}$ is given.

In order to formulate this problem mathematically we introduce the following performance measures:

$$x_j^a(u) = E_u \left[\sum_{t=0}^{\infty} I_j^a(t) \beta^t \right],$$

where

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

Notice that $x_j^a(u)$ is the total expected discounted time that action a is taken in state j under policy u. Therefore, the optimization problem described above can be formulated as

$$Z^* = \max_{u \in \mathcal{U}} \sum_{(i,a) \in \mathcal{C}} R_i^a \, x_i^a(u). \tag{2}$$

In order to formulate control problem (2) as a mathematical programming problem we must describe explicitly the *performance region* spanned by performance vector $\boldsymbol{x}(u) = \left(x_j^a(u) \right)_{j \in \mathcal{S}, a \in \mathcal{A}_j}$ under all admissible policies $u \in \mathcal{U}$. Denoting it $\mathcal{X} = \{\boldsymbol{x}(u), u \in \mathcal{U}\}$, we can translate (2) into the mathematical program

$$Z^* = \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{(i,a) \in \mathcal{C}} R_i^a \, x_i^a, \tag{3}$$

where $\boldsymbol{x} = \left(x_j^a\right)_{j \in \mathcal{S}, a \in \mathcal{A}_j}$. Let us now introduce the polyhedron

$$\mathcal{P} = \left\{ \boldsymbol{x} \in \Re_+^{|\mathcal{C}|} \colon \sum_{a \in \mathcal{A}_j} x_j^a = \alpha_j + \beta \sum_{(i,a) \in \mathcal{C}} p_{ij}^a x_i^a, \quad j \in \mathcal{S} \right\}.$$

Notice that by summing over all $j \in \mathcal{S}$ we obtain that $\sum_{(i,a) \in \mathcal{C}} x_i^a = \frac{1}{1-\beta}$ and, therefore, $\mathcal{P}$ is a bounded polyhedron, or *polytope*. We remark that the equations that define $\mathcal{P}$ above represent flow conservation relations at each state.

It was first shown by d'Epenoux (1960) that, under the assumption $\boldsymbol{\alpha} > \mathbf{0}$, polytope $\mathcal{P}$ coincides precisely with performance region $\mathcal{X}$. He also showed that, regardless of the assumption $\boldsymbol{\alpha} > \mathbf{0}$, $\mathcal{X} \subseteq \mathcal{P}$.

We strengthen next that classical result, by proving that polytope $\mathcal{P}$ always coincides with $\mathcal{X}$, regardless of whether the assumption $\boldsymbol{\alpha} > \mathbf{0}$ holds.

Theorem 1 (Performance region of discounted MDCs) *The following statements hold:*
(a) $\mathcal{X} = \mathcal{P}$.
(b) *The vertices of polytope $\mathcal{P}$ are achievable by stationary deterministic policies.*

Proof
See d'Epenoux (1960) for a proof that $\mathcal{X} \subseteq \mathcal{P}$. We prove next the other inclusion: $\mathcal{P} \subseteq \mathcal{X}$.

Since $\mathcal{P}$ is a polytope, any point in $\mathcal{P}$ can be written as a convex combination of its extreme points. Therefore, it suffices to show that any extreme point of $\mathcal{P}$ is achievable by some stationary deterministic policy, since then any point in $\mathcal{P}$ could also be achieved by a randomization of the policies that achieve the corresponding extreme points.

Let $\overline{\boldsymbol{x}}$ be an extreme point of polytope $\mathcal{P}$. By standard LP theory, $\overline{\boldsymbol{x}}$ is the unique maximizer of some linear objective function. Let $\sum_{(i,a) \in \mathcal{C}} R_i^a x_i^a$ be such an objective. Since $\overline{\boldsymbol{x}}$ is an extreme point of $\mathcal{P}$, which is defined by n equality constraints, it must have at most n positive components. Let us now partition state space $\mathcal{S}$ into two subspaces, $\mathcal{S}_1$ and $\mathcal{S}_2$, as follows:

$$\mathcal{S}_1 = \{j \in \mathcal{S} \colon \overline{x}_j^a > 0 \quad \text{for some } a \in A_j\}, \text{ and } \mathcal{S}_2 = \{j \in E \colon \overline{x}_j^a = 0 \quad \text{for all } a \in A_j\}.$$

Let $\overline{x}_{\mathcal{S}_1} = \{\overline{x}_j^a, \ j \in \mathcal{S}_1\}$. Consider now the following linear program:

$$\begin{aligned} (LP_1) \qquad Z_{\mathcal{S}_1} \quad = \quad & \max \sum_{j \in \mathcal{S}_1} \sum_{a \in A_j} R_j^a x_j^a \\ & \text{subject to} \\ & \sum_{a \in A_j} x_j^a - \beta \sum_{i \in \mathcal{S}_1} \sum_{a \in A_i} p_{ij}^a x_i^a = \alpha_j, \quad j \in \mathcal{S}_1, \\ & x_j^a \geq 0, \quad j \in \mathcal{S}_1, a \in A_j. \end{aligned}$$

By construction, $\overline{\boldsymbol{x}}_{\mathcal{S}_1}$ is the unique optimal solution of linear program (LP_1). Therefore, $\overline{\boldsymbol{x}}_{\mathcal{S}_1}$ is an extreme point of (LP_1), and it thus has at most $|\mathcal{S}_1|$ positive components. But by definition of $\mathcal{S}_1$, it then follows that $\overline{\boldsymbol{x}}_{\mathcal{S}_1}$ must have exactly $|\mathcal{S}_1|$ positive components, and for each state $j \in \mathcal{S}_1$ there is exactly one action $\overline{a}_j \in A_j$ such that $\overline{x}_j^{\overline{a}_j} > 0$.

We can now define a stationary deterministic policy $\overline{u}$ that achieves $\overline{\boldsymbol{x}}$: For each state $j \in \mathcal{S}_1$ let $\overline{a}_j$ be defined as above, whereas for each $j \in \mathcal{S}_2$ let $\overline{a}_j$ be defined arbitrarily. Now, policy $\overline{u}$ which deterministically takes action $\overline{a}_j$ in state j achieves performance vector $\overline{\boldsymbol{x}}$, which completes the proof of (a) and (b). □

3.2 LP formulation of the restless bandit problem

In order to formulate the restless bandit problem as a linear program we introduce performance measures

$$x_{i_n}^1(u) = E_u\left[\sum_{t=0}^{\infty} I_{i_n}^1(t)\beta^t\right],$$

and

$$x_{i_n}^0(u) = E_u\left[\sum_{t=0}^{\infty} I_{i_n}^0(t)\beta^t\right],$$

where $u \in \mathcal{U}$ is an admissible scheduling policy,

$$I_{i_n}^1(t) = \begin{cases} 1 & \text{if project } n \text{ is in state } i_n \text{ and active at time } t; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$I_{i_n}^0(t) = \begin{cases} 1 & \text{if project } n \text{ is in state } i_n \text{ and passive at time } t; \\ 0, & \text{otherwise.} \end{cases}.$$

Notice that performance measure $x_{i_n}^1(u)$ (resp. $x_{i_n}^0(u)$) represents the total expected discounted time that project n is in state i_n and active (resp. passive) under scheduling policy u. We denote $\mathcal{X}$ the corresponding performance region,

$$\mathcal{X} = \left\{ \boldsymbol{x} = \left(x_{i_n}^{a_n}(u)\right)_{i_n \in \mathcal{S}_n, a_n \in \{0,1\}, n \in \mathcal{N}} \mid \quad u \in \mathcal{U} \right\}.$$

Since the restless bandit problem is naturally formulated as a discounted MDC it follows from Theorem 1 that performance region $\mathcal{X}$ is a polytope, which we will refer to in what follows as the *restless bandit polytope.* The restless bandit problem can thus be formulated as the linear program

$$(LP) \qquad Z^* \quad = \quad \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} \sum_{a_n \in \{0,1\}} R_{i_n}^{a_n} x_{i_n}^{a_n}.$$

A polynomially solvable reformulation of problem (LP) above, for the special case of the multiarmed bandit problem, was first obtained by Bertsimas and Niño-Mora (1996). For general restless bandits, however, it is highly unlikely that such a reformulation can be derived since, as mentioned above, the problem is $PSPACE - hard$.

The approach we shall develop will be to construct relaxations of polytope $\mathcal{X}$ that yield polynomial-size relaxations of linear program (LP). We will represent these relaxations $\hat{\mathcal{X}} \supseteq \mathcal{X}$, not on the space of the original variables x_i^a, but in a higher-dimensional space that includes new auxiliary variables (i.e, as *projections* of higher dimensional polytopes $\hat{Q}$). An advantage of pursuing this *projection representation* approach is that we will be able to formulate approximations $\hat{\mathcal{X}}$ of $\mathcal{X}$ having possibly exponentially many facets as projections of polytopes $\hat{Q}$ with a polynomial number of facets, as discussed in the Introduction, thus providing polynomial-time tight bounds on the optimal value Z^*.

3.3 A first-order LP relaxation

Whittle (1988) introduced a relaxed version of the restless bandit problem which can be solved in polynomial time. The original requirement that exactly M projects must be active at any time is relaxed to an averaged version: the total expected discounted number of active projects must be $M/(1-\beta)$. Whittle showed that this relaxed version may be interpreted as the problem of controlling optimally N separate MDCs (one for each project), subject to the binding constraint on the average number of active projects just stated. In this section we reformulate Whittle's relaxation as a polynomial-size linear program.

The restless bandit problem induces a *first-order MDC* over each project n in a natural way: The state space of this MDC is $\mathcal{S}_n$, its action space is $\mathcal{A}^1 = \{0,1\}$, and the reward received when action a_n is taken in state i_n is $R_{i_n}^{a_n}$. Rewards are discounted in time by discount factor β. The transition probability from state i_n into state j_n, given action a_n, is $p_{i_n j_n}^{a_n}$. The initial state is i_n with probability α_{i_n} (which we assumed can only be 0 or 1). Let

$$\mathcal{Q}_n^1 = \left\{ \boldsymbol{x}_n = \left(x_{i_n}^{a_n}(u) \right)_{i_n \in \mathcal{S}_n, a_n \in \mathcal{A}^1} \mid u \in \mathcal{U} \right\}.$$

Notice that polytope $\mathcal{Q}_n^1$ is precisely the *projection* of restless bandit polytope $\mathcal{P}$ over the space of the variables $x_{i_n}^{a_n}$ for project n. Furthermore, $\mathcal{Q}_n^1$ is also the performance region of the first-order MDC corresponding to project n, as defined above. Applying Theorem 1 we thus obtain:

Proposition 1 *A complete formulation of $\mathcal{Q}_n^1$ is given by*

$$\mathcal{Q}_n^1 = \left\{ \boldsymbol{x}_n \in \Re_+^{|\mathcal{S}_n \times \{0,1\}|} \mid x_{j_n}^0 + x_{j_n}^1 = \alpha_{j_n} + \beta \sum_{i_n \in \mathcal{S}_n} \sum_{a_n \in \{0,1\}} p_{i_n j_n}^{a_n} x_{i_n}^{a_n}, \quad j_n \in \mathcal{S}_n \right\}. \tag{4}$$

Remark: It follows from Proposition 1 that the general restless bandit problem, with both active and passive rewards, can be reduced to the case with only active rewards. This follows since, by (4), the passive performance vector $\boldsymbol{x}_n^0(u)$ is a linear transformation of the active one, $\boldsymbol{x}_n^1(u)$.

Now, Whittle's condition on the average number of active projects can be written as

$$\begin{aligned}
\sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} x_{i_n}^1(u) &= \sum_{t=0}^{\infty} E_u \left[\sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} I_{i_n}^1(t) \right] \beta^t \\
&= \sum_{t=0}^{\infty} M \beta^t \\
&= \frac{M}{1-\beta}. \qquad (5)
\end{aligned}$$

Therefore, Whittle's first-order relaxation can be formulated as the linear program

$$\begin{aligned}
(LP^1) \qquad Z^1 \;=\; & \max \sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} \sum_{a_n \in \{0,1\}} R_{i_n}^{a_n} x_{i_n}^{a_n} \\
& \text{subject to} \\
& \boldsymbol{x}_n \in \mathcal{Q}_n^1, \qquad n \in \mathcal{N}, \\
& \sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} x_{i_n}^1 = \frac{M}{1-\beta}.
\end{aligned}$$

We will refer to the feasible space of linear program (LP^1) as the *first-order approximation* to the restless bandit polytope $\mathcal{P}$, and will denote it $\mathcal{P}^1$. Notice that linear program (LP^1) has $O(N\,|\mathcal{S}_{\max}|)$ variables and constraints, where $|\mathcal{S}_{\max}| = \max_{n \in \mathcal{N}} |\mathcal{S}_n|$, and its size is thus polynomial in the problem dimensions.

3.4 A second-order LP relaxation

In this section we present a new LP relaxation for the restless bandit problem, which is stronger than Whittle's, yet of polynomial size. This new relaxation involves auxiliary variables, which correspond to second-order performance measures for the restless bandit problem, associated with project pairs. Given a pair of projects, $n_1 < n_2$, the valid actions which can be taken over each pair of states, $(i_1, i_2) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2}$, range over

$$\mathcal{A}^2 = \left\{ (a_1, a_2) \in \{0,1\}^2 \mid a_1 + a_2 \leq M \right\}.$$

Given an admissible scheduling policy u, we define the *second-order performance measures* by

$$x_{i_1 i_2}^{a_1 a_2}(u) = E_u \left[\sum_{t=0}^{\infty} I_{i_1}^{a_1}(t)\, I_{i_2}^{a_2}(t)\, \beta^t \right].$$

Similarly as in the first-order case, the restless bandit problem induces a *second-order MDC* over each pair of projects $n_1 < n_2$ in a natural way: The state space of the MDC is $\mathcal{S}_{n_1} \times \mathcal{S}_{n_2}$, its action space is $\mathcal{A}^2$, and the reward corresponding to state (i_{n_1}, i_{n_2}) and action (a_{n_1}, a_{n_2}) is $R_{i_{n_1}}^{a_{n_1}} + R_{i_{n_2}}^{a_{n_2}}$. Rewards are discounted in time by discount factor β. The transition probability from state (i_{n_1}, i_{n_2}) into state (j_{n_1}, j_{n_2}), given action (a_{n_1}, a_{n_2}), is $p_{i_{n_1} j_{n_1}}^{a_{n_1}} \, p_{i_{n_2} j_{n_2}}^{a_{n_2}}$. The initial state is (i_{n_1}, i_{n_2}) with probability $\alpha_{i_{n_1}} \, \alpha_{i_{n_2}}$ (which, again, can only be 0 or 1). Let

$$\mathcal{Q}_{n_1,n_2}^2 = \left\{ \boldsymbol{x}_{n_1,n_2} = \left(x_{i_1 i_2}^{a_1 a_2}(u) \right)_{i_1 \in \mathcal{S}_{n_1}, i_2 \in \mathcal{S}_{n_2}, (a_1,a_2) \in \mathcal{A}^2} \mid u \in \mathcal{U} \right\},$$

be the projection of restless bandit polytope $\mathcal{P}$ over the space of second-order variables $(x_{i_1 i_2}^{a_1 a_2})_{i_1 \in \mathcal{S}_{n_1}, i_2 \in \mathcal{S}_{n_2}, (a_1,a_2) \in \mathcal{A}^2}$. Notice that $\mathcal{Q}_{n_1,n_2}^2$ is also the performance region of the second-order MDC corresponding to project pair (n_1, n_2), as defined above. Applying Theorem 1 we obtain:

Proposition 2 *A full formulation of polytope $\mathcal{Q}_{n_1,n_2}^2$ is given by*

$$\sum_{(a_1,a_2) \in \mathcal{A}^2} x_{j_1 j_2}^{a_1 a_2} = \alpha_{j_1} \alpha_{j_2} + \beta \sum_{\substack{i_1 \in \mathcal{S}_{n_1}, i_2 \in \mathcal{S}_{n_2} \\ (a_1,a_2) \in \mathcal{A}^2}} p_{i_1 j_1}^{a_1} p_{i_2 j_2}^{a_2} x_{i_1 i_2}^{a_1 a_2}, \quad (j_1, j_2) \in \mathcal{S}_1 \times \mathcal{S}_2, \tag{6}$$

$$x_{i_1 i_2}^{a_1 a_2} \geq 0, \qquad (i_1, i_2) \in \mathcal{S}_{n_1} \times \mathcal{S}_{n_2}, \ (a_1, a_2) \in \mathcal{A}^2. \tag{7}$$

We next develop several additional linear identities on second-order performance measures, by applying simple combinatorial arguments. These identities will serve to strengthen the second-order relaxation we present later. For any admissible scheduling policy u, we have, if $N \geq M + 2$,

$$\sum_{1 \leq n_1 < n_2 \leq N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} x_{i_1 i_2}^{00}(u) = \frac{\binom{N-M}{2}}{1-\beta}, \tag{8}$$

since the $N-M$ passive projects required at any time correspond to $\begin{pmatrix} N-M \\ 2 \end{pmatrix}$ passive-passive project pairs. Moreover,

$$\sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} (x_{i_1 i_2}^{10}(u) + x_{i_1 i_2}^{01}(u)) = \frac{M(N-M)}{1-\beta}, \tag{9}$$

since at any time the M active and $N-M$ passive required projects give rise to $M(N-M)$ active-passive project pairs.

Furthermore, in the case that $M \ge 2$, we have

$$\sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} x_{i_1 i_2}^{11}(u) = \frac{\begin{pmatrix} M \\ 2 \end{pmatrix}}{1-\beta}, \tag{10}$$

since at any time the $M \ge 2$ active projects give rise to $\begin{pmatrix} M \\ 2 \end{pmatrix}$ active-active project pairs.

We next relate the first and second-order performance measures. It is easy to see that, for any admissible policy u,

$$x_{i_1}^{a_1}(u) = \sum_{\substack{i_2 \in \mathcal{S}_{n_2} \\ a_2 : (a_1, a_2) \in \mathcal{A}^2}} x_{i_1 i_2}^{a_1 a_2}(u), \quad i_1 \in \mathcal{S}_{n_1}, a_1 \in \{0,1\}, 1 \le n_1 < n_2 \le N, \tag{11}$$

and

$$x_{i_2}^{a_2}(u) = \sum_{\substack{i_1 \in \mathcal{S}_{n_1} \\ a_1 : (a_1, a_2) \in \mathcal{A}^2}} x_{i_1 i_2}^{a_1 a_2}(u), \quad i_2 \in \mathcal{S}_{n_2}, a_2 \in \{0,1\}, 1 \le n_1 < n_2 \le N. \tag{12}$$

We introduce next the second-order LP relaxation for the restless bandit problem, which is based on the above identities:

$$\begin{aligned}
(LP^2) \quad Z^2 \quad = \quad & \max \sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} \sum_{a_n \in \{0,1\}} R_{i_n}^{a_n} x_{i_n}^{a_n} \\
& \text{subject to} \\
& \boldsymbol{x}_{n_1, n_2} \in Q_{n_1, n_2}^2, \quad 1 \le n_1 < n_2 \le N, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} x_{i_1 i_2}^{00} = \frac{\begin{pmatrix} N-M \\ 2 \end{pmatrix}}{1-\beta}, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} (x_{i_1 i_2}^{10} + x_{i_1 i_2}^{01}) = \frac{M(N-M)}{1-\beta}, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{i_2 \in \mathcal{S}_{n_2}} x_{i_1 i_2}^{11} = \frac{\begin{pmatrix} M \\ 2 \end{pmatrix}}{1-\beta},
\end{aligned}$$

$$
\begin{aligned}
& x_{i_1}^{a_1} = \sum_{i_2 \in \mathcal{S}_{n_2}} \sum_{a_2 : (a_1, a_2) \in \mathcal{A}^2} x_{i_1 i_2}^{a_1 a_2}, \quad i_1 \in \mathcal{S}_{n_1}, a_1 \in \{0,1\}, 1 \le n_1 < n_2 \le N, \\
& x_{i_2}^{a_2} = \sum_{i_1 \in \mathcal{S}_{n_1}} \sum_{a_1 : (a_1, a_2) \in \mathcal{A}^2} x_{i_1 i_2}^{a_1 a_2}, \quad i_2 \in \mathcal{S}_{n_2}, a_2 \in \{0,1\}, 1 \le n_1 < n_2 \le N, \\
& \sum_{n \in \mathcal{N}} \sum_{i_n \in \mathcal{S}_n} \sum_{a_n \in \{0,1\}} x_{i_n}^1 = \frac{M}{1-\beta}, \\
& x_i^a \ge 0.
\end{aligned}
$$

We define the second-order approximation to the restless bandit polytope $\mathcal{P}$ as the projection of the feasible space of linear program (LP^2) into the space of the first-order variables, x_i^a, and will denote it as $\mathcal{P}^2$.

Notice that second-order relaxation (LP^2) has $O(N^2 \, |\mathcal{S}_{\max}|^2)$ variables and constraints, (recall that $|\mathcal{S}_{\max}| = \max_{n \in \mathcal{N}} |\mathcal{S}_n|$.)

3.5 Higher-order LP relaxations

The idea for constructing the second-order LP relaxation above can be readily extended to develop higher-order LP relaxations: For the kth-order case the new auxiliary variables correspond to project k-tuples $(n_1, \ldots, n_k)$, whereas the admissible actions over a corresponding k-tuple of states $(i_1, \ldots, i_k)$ range over

$$\mathcal{A}^k = \left\{ (a_1, \ldots, a_k) \in \{0,1\}^k \mid a_1 + \cdots + a_k \le M \right\}.$$

The auxiliary variable $x_{j_1 \cdots j_k}^{a_1 \cdots a_k}$ corresponds to the kth-order performance measure

$$x_{j_1 \cdots j_k}^{a_1 \cdots a_k}(u) = E_u \left[\sum_{t=0}^{\infty} I_{j_1}^{a_1}(t) \cdots I_{j_k}^{a_k}(t) \beta^t \right], \quad j_1 \in \mathcal{S}_{n_1}, \ldots, j_k \in \mathcal{S}_{n_k}, \tag{13}$$

which accounts for interactions between project k-tuples.

We can construct, similarly as above, a kth-order LP relaxation (LP^k), with objective value Z^k. Furthermore, we define the kth-order approximation $\mathcal{P}^k$ to the restless bandit polytope $\mathcal{P}$ as the projection of the feasible space of (LP^k) into the space of first-order variables x_i^a. Clearly the resulting sequence of approximations is monotone, so that

$$\mathcal{P}^1 \supseteq \mathcal{P}^2 \supseteq \cdots \supseteq \mathcal{P}^N = \mathcal{P},$$

and therefore

$$Z^1 \ge Z^2 \ge \cdots \ge Z^N = Z^*.$$

Notice that the kth-order relaxation (LP^k) has $O(N^k \, |\mathcal{S}_{\max}|^k)$ variables and constraints, and it has therefore polynomial size for fixed k (though the size is exponential if we allow k to vary).

The last relaxation of the sequence, (LP^N), is exact (i.e., $Z^N = Z^*$), since it corresponds to the standard LP formulation of the restless bandit problem modeled as a MDC.

4 A primal-dual heuristic for the restless bandit problem

In this section we present a heuristic for the restless bandit problem, which uses information contained in optimal primal and dual solutions to the first-order relaxation (LP^1). Under some mixing assumptions on active and passive transition probabilities, we interpret the primal-dual heuristic as an priority-index heuristic. The dual of linear program (LP^1) is

$$
\begin{aligned}
(D^1) \qquad Z^1 \quad = \quad & \min \sum_{n \in \mathcal{N}} \sum_{j_n \in \mathcal{S}_n} \alpha_{j_n} \lambda_{j_n} + \frac{M}{1-\beta} \lambda \\
& \text{subject to} \\
& \lambda_{i_n} - \beta \sum_{j_n \in \mathcal{S}_n} p^0_{i_n j_n} \lambda_{j_n} \geq R^0_{i_n}, \quad i_n \in \mathcal{S}_n, \quad n \in \mathcal{N}, \\
& \lambda_{i_n} - \beta \sum_{j_n \in \mathcal{S}_n} p^1_{i_n j_n} \lambda_{j_n} + \lambda \geq R^1_{i_n}, \quad i_n \in \mathcal{S}_n, \quad n \in \mathcal{N}, \\
& \lambda \geq 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (14)
\end{aligned}
$$

Let $\{\overline{x}^{a_n}_{i_n}\}$, $\{\overline{\lambda}_{i_n}, \overline{\lambda}\}$, be an optimal primal and dual solution pair to the first-order relaxation (LP^1) and its dual (D^1). Let $\{\overline{\gamma}^{a_n}_{i_n}\}$ be the corresponding optimal reduced cost coefficients, i.e.,

$$
\overline{\gamma}^0_{i_n} = \overline{\lambda}_{i_n} - \beta \sum_{j_n \in \mathcal{S}_n} p^0_{i_n j_n} \overline{\lambda}_{j_n} - R^0_{i_n},
$$

$$
\overline{\gamma}^1_{i_n} = \overline{\lambda}_{i_n} - \beta \sum_{j_n \in \mathcal{S}_n} p^1_{i_n j_n} \overline{\lambda}_{j_n} + \overline{\lambda} - R^1_{i_n},
$$

which must be nonnegative. It is well known (cf. Murty (1983), pp. 64-65), that the optimal reduced costs have the following interpretation:

$\overline{\gamma}^1_{i_n}$ is the *rate of decrease* in the objective-value of linear program (LP^1) *per unit increase* in the value of the variable $x^1_{i_n}$.

$\overline{\gamma}^0_{i_n}$ is the *rate of decrease* in the objective-value of linear program (LP^1) *per unit increase* in the value of the variable $x^0_{i_n}$.

The proposed heuristic takes as input the vector of current states of the projects, $(i_1, \ldots, i_N)$, optimal primal solution $\{\overline{x}^{a_n}_{j_n}\}$, and the corresponding optimal reduced costs $\{\overline{\gamma}^{a_n}_{j_n}\}$, and produces as output the action to take on each project, $(a^*(i_1), \ldots, a^*(i_N))$. An informal description of the heuristic, with its motivation, follows.

The heuristic is structured in a primal and a dual stage. In the primal stage, projects n whose corresponding active primal variable $\overline{x}^1_{i_n}$ is strictly positive are considered as candidates for active selection. The intuition is that we give preference for active selection to projects with positive $\overline{x}^1_{i_n}$ with respect to those with $\overline{x}^1_{i_n} = 0$, which seems natural given the interpretation of performance measure $x^1_{i_n}(\cdot)$ as the total expected discounted time spent selecting project n in state i_n as active. Let p represent the number of such projects. In the case that $p = M$, then all M candidate projects are set active. If $p < M$, then all p candidate projects are set active and the heuristic proceeds to the dual stage that selects the remaining $M - p$ projects. If $p > M$ none of them is set active at this stage and the heuristic proceeds to the dual stage that finalizes the selection.

In the dual stage, in the case that $p < M$, then $M - p$ additional projects, each with current active primal variable zero ($\overline{x}^1_{i_n} = 0$), must be selected for active operation among the $N - p$ projects whose actions have not yet been fixed. As a heuristic index of the undesirability of setting project n in state i_n active, we take the active reduced cost $\overline{\gamma}^1_{i_n}$. This choice is motivated by the interpretation of $\overline{\gamma}^1_{i_n}$ stated above: the larger the *active index* $\gamma^1_{i_n}$ is, the larger is the rate of decrease of the objective-value of (LP^1) per unit increase in the active variable $x_{i_n}^1$. Therefore, in the heuristic we select for active operation the $M - p$ additional projects with smallest active reduced costs.

In the case that $p > M$, then M projects must be selected for active operation, among the p projects with $\overline{x}_{i_n}^1 > 0$. Recall that by complementary slackness, $\overline{\gamma}_{i_n}^1 = 0$ if $\overline{x}_{i_n}^1 > 0$. As a heuristic index of the desirability of setting project n in state i_n active we take the passive reduced cost $\overline{\gamma}_{i_n}^0$. The motivation is given by the interpretation of $\overline{\gamma}_{i_n}^0$ stated above: the larger the *passive index* $\gamma_{i_n}^0$ is, the larger is the rate of decrease in the objective-value of (LP^1) per unit increase in the value of the passive variable $x_{i_n}^0$. Therefore, in the heuristic we select for active operation the M projects with largest passive reduced costs. The heuristic is described formally in Table 1.

An index interpretation of the primal-dual heuristic. We next observe that, under natural mixing conditions, the primal-dual heuristic described above reduces to a *priority-index* rule. For each project $n \in \mathcal{N}$ we consider a directed graph that is defined from the passive and active transition probabilities respectively as follows: $G_n = (\mathcal{S}_n, A_n)$, where $A_n = \{(i_n, j_n) | \ p_{i_n j_n}^0 > 0, \text{ and } p_{i_n j_n}^1 > 0 \ i_n, j_n \in \mathcal{S}_n\}$. We introduce now the following mixing assumption:

Assumption 1 *For every $n \in \mathcal{N}$ the directed graph G_n is connected.*

Since polytope $\mathcal{P}^1$ has independent constraints (i.e., involving different variable sets) for every $n \in \mathcal{N}$ and only one global constraint, elementary LP theory yields that

Proposition 3 *Under assumption 1, every extreme point $\overline{\boldsymbol{x}}$ of polytope $\mathcal{P}^1$ has the following properties:*
(a) There are at most one project k and one state $i_k \in \mathcal{S}_k$ for which $\overline{x}_{i_k}^1 > 0$ and $\overline{x}_{i_k}^0 > 0$.
(b) For all other projects n and all other states either $\overline{x}_{i_n}^1 > 0$ or $\overline{x}_{i_n}^0 > 0$.

Therefore, starting with an optimal extreme point solution $\overline{\boldsymbol{x}}$ and a complementary dual optimal solution, with corresponding reduced costs $\overline{\gamma}$, we consider the priority-index rule defined next.

Priority-index heuristic:

1. Given the current states $(i_1, \ldots, i_N)$ of the N projects, compute the indices
$$\delta_{i_n} = \overline{\gamma}_{i_n}^1 - \overline{\gamma}_{i_n}^0.$$
2. Set active the projects that have the M *smallest* indices. In case of ties, set active projects with $\overline{x}_{i_n}^1 > 0$.

We next remark that under Assumption 1, the primal-dual and the priority-index heuristics are the same. To see this we consider first the case $p \leq M$. The primal-dual heuristic would set active first the projects that have $\overline{x}_{i_n}^1 > 0$. From complementarity, these projects have $\overline{\gamma}_{i_n}^1 = 0$ and therefore, $\delta_{i_n} \leq 0$. Then, the primal-dual heuristic sets active the remaining $M - p$ projects with the smallest $\overline{\gamma}_{i_n}^1$. Since for these projects $\overline{x}_{i_n}^1 = 0$ and therefore, $\overline{x}_{i_n}^0 > 0$, i.e., $\overline{\gamma}_{i_n}^0 = 0$, we obtain that $\delta_{i_n} = \overline{\gamma}_{i_n}^1 \geq 0$. Therefore, the choices of the two heuristics are indeed identical.

If $p > M$, the primal-dual heuristic sets active the projects that have the largest values of $\overline{\gamma}_{i_n}^0$. For these projects $\overline{\gamma}_{i_n}^1 = 0$, and therefore, $\delta_{i_n} = -\overline{\gamma}_{i_n}^0 \leq 0$. Since the remaining

```
Input:
    • $(i_1, \ldots, i_N)$ { current states of the $N$ projects}
    • $\{\overline{x}_{j_n}^{a_n}\}$ { optimal primal solution to first-order relaxation $(LP^1)$ }
    • $\{\overline{\gamma}_{j_n}^{a_n}\}$ { optimal reduced costs for first-order relaxation $(LP^1)$ }
Output:
    • $(a^*(i_1), \ldots, a^*(i_N))$ { actions to take at each project }
{ Initialization: }
set $S := \emptyset$; {S: set of projects whose actions have been set}

set $a^*(i_n) := 0$, for $n \in \mathcal{N}$; {actions are initialized as passive}

{ Primal Stage: }
set $p := |\{\overline{x}_{i_n}^1 : \overline{x}_{i_n}^1 > 0, n \in \mathcal{N}\}|$ { p: number of projects with positive active primals }

if $p \leq M$ then { set active the projects with positive active primals, if $\leq M$ }
  for $n \in \mathcal{N}$ do
      if $\overline{x}_{i_n}^1 > 0$ then
      begin
      set $a^*(i_n) := 1$;
      set $S := S \cup \{n\}$
      end

{ Dual Stage: }
if $p < M$ then { set active $M - p$ extra projects with smallest active reduced costs }
  until $|S| = M$ do
  begin
  select $\overline{n} \in \text{argmin}\{\overline{\gamma}_{i_n}^1 : n \in \mathcal{N} \setminus S\}$
  set $a^*(i_{\overline{n}}) := 1$;
  set $S := S \cup \{\overline{n}\}$
  end

if $p > M$ then { set active $M$ projects with largest passive reduced costs }
  until $|S| = M$ do
  begin
  select $\overline{n} \in \text{argmax}\{\overline{\gamma}_{i_n}^0 : n \in \mathcal{N} \setminus S\}$
  set $a^*(i_{\overline{n}}) := 1$;
  set $S := S \cup \{\overline{n}\}$
  end
```

Table 1: Primal-Dual heuristic.

projects have $\delta_{i_n} = \overline{\gamma}^1_{i_n} \geq 0$, the choices of the two heuristics are identical in this case as well.

In contrast with the Gittins indices for usual bandits, notice that the indices δ_{i_n} for a particular project depend on characteristics of all other projects. In particular, we have verified experimentally that these indices do not reduce to Gittins indices in the case of classical multiarmed bandits.

5 Computational experiments

In this section we report the results of a series of computational experiments to investigate the tightness of the relaxations and the performance of the primal-dual heuristic introduced in this paper. For each test problem we computed the following quantities:

Z_{Greedy}**:** Estimated (through simulation) expected value of the greedy heuristic (at each time M projects with largest active reward are operated). We simulate a run using the heuristic policy and we obtain a value for the reward for the particular run. In order to obtain the value for a particular run, we truncated the infinite summation in (1) ignoring terms after time t, such that $\beta^t > 10^{-10}$. Even if we used a smaller tolerance, the results did not change. The stopping criterion for the simulation was that the difference between the average from the first $l+1$ runs and the average from the first l runs is less than 10^{-5} (using a smaller tolerance did not change the results in this case as well).

Z_{PDH}**:** Estimated expected value of primal-dual heuristic. The estimation was achieved through simulation as for Z_{Greedy}.

Z^***:** Optimal value, which is equal to Z^N (due to the size of the formulation, this value was calculated only for small instances).

Z^2**:** Optimal value of the second-order relaxation (LP^2).

Z^1**:** Optimal value of the first-order relaxation (LP^1).

The heuristics and the simulation experiments were implemented in C. The LP formulations were implemented using GAMS and solved using CPLEX. All the experiments were performed in a SUN 10 workstation. In order to test the proposed approach we generated 7 problem instances as described next.

Problems 1 and 2 involve 10 projects with 7 states each, with $M = 1$, and their data (the reward vectors and the passive and active transition probabilities) was randomly generated. For these problems we could not compute the optimal solution because of their large sizes. Since these instances were randomly generated we expected that the greedy heuristic would perform very close to the optimal solution. To further test this we generated Problem 3, which has 5 projects with 3 states each, for which the data was also randomly generated and $M = 1$.

Problem 4 has 5 projects with 3 states each, with $M = 1$. The data was designed so that a greedy heuristic would perform poorly. Problems 5 through 7 have the same data as problem 4, except that the number of active projects ranges from $M = 2$ through $M = 4$, respectively.

The projects in each problem are different (i.e., have different rewards and transition probabilities).

In Table 2 we report the results of our experiments for various values of the discount factor β. We next discuss these results.

1. The primal-dual heuristic performed exceptionally well. It was essentially optimal in Problems 3-7 and it was slightly better than the greedy heuristic in Problems 1 and 2. Since we expect the greedy heuristic to be near-optimal for randomly generated instances (as a verification Problem 3 had also randomly generated data and the greedy heuristic was extremely close to optimal), we believe the heuristic is extremely close to the optimal solution for Problems 1 and 2 as well. For this reason, we did not experiment with other heuristics, as we feel that the quality of solutions produced by the primal-dual heuristic is adequate for solving realistic size problems.

2. Regarding the performance of the relaxations, the bounds from the second-order relaxation improve significantly over the first-order ones, and in most instances the bound was very close to the exact optimal value. In Problem 1 there is a wider gap between the value of the primal-dual heuristic and the value of the second-order relaxation. The closeness of the value of the heuristic with the value of the greedy solution (which is expected to be near optimal in this case), suggests that the main source of this gap is the inaccuracy of the second-order bound.

3. As expected, the performance of the greedy heuristic deteriorates as the discount factor approaches 1, since in that case the long-term impact of current decisions is more heavily weighted. The primal-dual heuristic outperforms the greedy heuristic over the sample problems (it performs significantly better in instances with higher discount factors, and never worse, even for $\beta = 0.2$). Notice that in the randomly generated instances both heuristics yield very close rewards.

4. Notice that, though problems $4 - 7$ share the same data except that $M = 1, 2, 3, 4$, respectively, the Z^* values are not monotonic on M (they decrease when $M = 4$). This is due to the fact that we require *exactly* M projects to be active at each time, not *at most* M, which explains this lack of monotonicity.

5. The solution of the second-order relaxations took significantly more time than that of the first-order ones, but it was within reasonable time limits in all instances (at most 10 minutes).

We further remark that we have checked, in other experiments, that the second-order relaxation is not exact when applied to classical multiarmed bandits.

6 Concluding remarks

We have proposed an approach that provides a feasible policy together with a guarantee for its suboptimality for the restless bandit problem. Our computational experiments suggest that the primal-dual heuristic has excellent performance, while the second order relaxation is quite strong. Our approach has the attractive feature that it can produce increasingly stronger suboptimality bounds at the expense of increased computational times.

We believe that our results demonstrate that ideas that have been successful in the field of discrete optimization (strong formulations, projections and primal-dual heuristics) in the last two decades, can be used successfully in the field of stochastic optimization. Although we have only addressed in this paper the restless bandit problem, given the generality and

Problem	β	Z_{Greedy}	Z_{PDH}	Z^*	Z^2	Z^1
$(N, \|S_n\|, M)$						
Problem 1	0.20	59.9	59.9		70.62	74.67
(10, 7, 1)	0.50	124.2	124.3		162.05	166.35
	0.90	814.4	819.1		898.99	913.40
Problem 2	0.20	117.1	117.3		117.92	118.46
(10, 7, 1)	0.50	180.2	180.2		183.89	186.10
	0.90	863.1	863.4		894.18	915.44
Problem 3	0.50	14.7	14.7	14.72	15.33	16.10
(5, 3, 1)	0.90	81.3	81.5	81.55	84.54	85.29
	0.95	164.5	164.9	164.98	169.60	171.07
Problem 4	0.50	10.8	11.4	11.40	11.65	11.92
(5, 3, 1)	0.90	65.1	75.1	75.15	75.99	78.36
	0.95	135.5	156.0	156.09	157.81	162.12
Problem 5	0.50	19.2	21.5	21.63	21.93	21.93
(5, 3, 2)	0.90	122.5	144.5	144.73	146.50	147.29
	0.95	257.2	300.1	300.35	303.73	305.68
Problem 6	0.50	28.0	30.8	30.95	31.33	31.53
(5, 3, 3)	0.90	167.5	209.5	209.56	209.56	209.56
	0.95	346.2	434.7	434.74	434.74	434.74
Problem 7	0.50	10.7	10.9	10.93	10.93	10.93
(5, 3, 4)	0.90	58.0	74.3	74.35	74.37	74.55
	0.95	119.2	154.0	154.09	154.09	154.42

Table 2: Numerical experiments.

complexity ($PSPACE-hard$) of the problem we expect that these ideas should have wider applicability. We intend to pursue them further in the context of other important stochastic optimization problems.

References

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

[2] Bertsimas, D., Paschalidis, I. and Tsitsiklis, J. (1994). Optimization of multiclass queueing networks: polyhedral and nonlinear characterizations of achievable performance. *Ann. Appl. Prob.* **1** 43-75.

[3] Coffman, E. G., Jr., and Mitrani, I. (1980). A characterization of waiting time performance realizable by single server queues. *Oper. Res.* **28** 810-821.

[4] d'Epenoux, F. (1960). Sur un problème de production et de stockage dans l'aléatoire. *RAIRO Rech. Opér.* **14** 3-16; (Engl. trans.) *Management Sci.* **10** 98-108 (1963).

[5] Federgruen, A. and Groenevelt, H. (1988). Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* **36** 733-741.

[6] Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics: European Meeting of Statisticians*, Budapest, 1972, J. Gani et al. (Eds.), North-Holland, Amsterdam,**1** 241-266.

[7] Gittins, J. C.(1989). *Multi-armed Bandit Allocation Indices.* Wiley, Chichester.

[8] Lovász, L. and Schrijver, A. (1991). Cones of matrices and set-functions and $0-1$ optimization. *SIAM J. Optimization* **1** 166-190.

[9] Murty, K. G. (1983). *Linear Programming.* Wiley, New York.

[10] Nemhauser, G. L. and Wolsey, L. A. (1988). *Integer and Combinatorial Optimization.* Wiley, New York.

[11] Niño-Mora, J. (1995). *Optimal Resource Allocation in a Dynamic and Stochastic Environment: A Mathematical Programming Approach.* PhD Dissertation, Sloan School of Management, MIT.

[12] Papadimitriou, C. H. and Tsitsiklis, J. N. (1994). The complexity of optimal queueing network control. Working Paper, MIT.

[13] Shanthikumar, J. G. and Yao, D. D. (1992). Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Oper. Res.* **40** S293-299.

[14] Veatch, M. and Wein, L. M. (1996). Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. 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] Weiss, G. (1988). Branching bandit processes. *Probab. Engin.. Inform. Sci.* **2** 269-278.

[18] 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.