Restless Bandits, Linear Programming Relaxations and a Primal-Dual Heuristic

D. Bertsimas and J. Nino-Mora

OR 298-94 August 1994

Restless Bandits, Linear Programming Relaxations and a Primal-Dual Heuristic

Dimitris Bertsimas * José Niño-Mora †

August 1994

Abstract

We propose a mathematical programming approach for the classical $PSPACE-hard$ problem of n restless bandits in stochastic optimization. We introduce a series of n increasingly stronger linear programming relaxations, the last of which is exact and corresponds to the formulation of the problem as a Markov decision process that has exponential size, while other relaxations provide bounds and are efficiently solvable. We also propose a heuristic for solving the problem that naturally arises from the first of these relaxations and uses indices that are computed through optimal dual variables from the first relaxation. In this way we propose a policy and a suboptimality guarantee. We report computational results that suggest that the value of the proposed heuristic policy is extremely close to the optimal value. Moreover, the second order relaxation provides strong bounds for the optimal solution value.

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

†José Niño-Mora, Operations Research Center, MIT, Cambridge, MA 02139. The research of the author was partially supported by a Ministry of Education & Science of Spain Doctoral Fellowship.

1 Introduction

Research in integer programming over the last twenty years has crystallized the idea that our ability to solve efficiently discrete optimization problems critically depends on a strong formulation of the problem. As a result, much of the research effort in integer programming has centered in developing sharper formulations. The developments in the fields of polyhedral combinatorics and more recently nonlinear relaxations (see for example Lovász and Schrijver [19]) are strong witnesses of this trend.

In contrast, the area of stochastic optimization in the last twenty years 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. A general characteristic of this body of research is the lack of a unified method of attack for these problems. Every problem seems to require its own formulation and, as a result, its own somewhat ad hoc approach. Moreover, quite often it is not clear how close a proposed solution is to the optimal one.

Motivated by the success of improved formulations in integer programming problems, we propose in this paper a theory of improved formulation for the classical restless bandits problem in stochastic scheduling. This research is part of a larger program to attack stochastic optimization problems using ideas and techniques from mathematical programming (see Bertsimas [1]). In broad terms the approach to formulate stochastic optimization problems as *mathematical programming problems* is based on the following idea: Given a stochastic optimization problem, we define a vector of performance measures, which are typically expectations and express the objective function as a function of this vector. We then characterize *the region of achievable performance*, i.e., we find constraints on the performance vectors that all admissible policies satisfy. In this way we find a series of relaxations that are progressively closer to the exact region of achievable performance. In Figure 1 we outline the conceptual similarity of this approach to the approach used in integer programming.

Background

In the 1980s the pioneering works of Coffman and Mitrani [9], and Gelenbe and Mitrani [14], initiated a new line of research for solving dynamic and stochastic optimization problems. They formulated the problem of optimal scheduling control in a multiclass $M/G/1$ queue as a linear program, by characterizing the *performance space* corresponding to appropriate *performance measures* as a polyhedron. Facets in the performance space correspond to *work conservation laws* in the queueing system. Federgruen and Groenevelt [12], [13], extended that work by showing that, in certain cases, the performance space of a multiclass queueing system is a *polymatroid* (see Edmonds [11]), which explains the optimality of simple index policies (the classical $c\mu$ rule). Shanthikumar and Yao [24] showed that a sufficient condition for the performance space of a multiclass queueing system to be a polymatroid is that the performance measure satisfies *strong conservation laws*. They exhibited a large number of queueing systems that fit into their framework. Tsoucas [25] investigated the problem of optimal scheduling control in a multiclass queue with Bernoulli feedback, introduced by Klimov [18]. He characterized its performance space as a new kind of polyhedron: an extended polymatroid, which generalizes the usual polymatroids introduced by Edmonds [11]. He showed how this polyhedral structure explains the optimality of priority policies (Klimov's algorithm). Drawing on this line of research, and on the theory of multi-armed bandit problems (see Gittins [16] and the references therein), Bertsimas and Niño-Mora [2] presented a unified framework for formulating and solving a large class of dynamic and stochastic scheduling problems as linear programs over extended polymatroids. Their results explain the optimality of index priority policies in systems that satisfy *generalized conservation laws*. These include all the problems mentioned

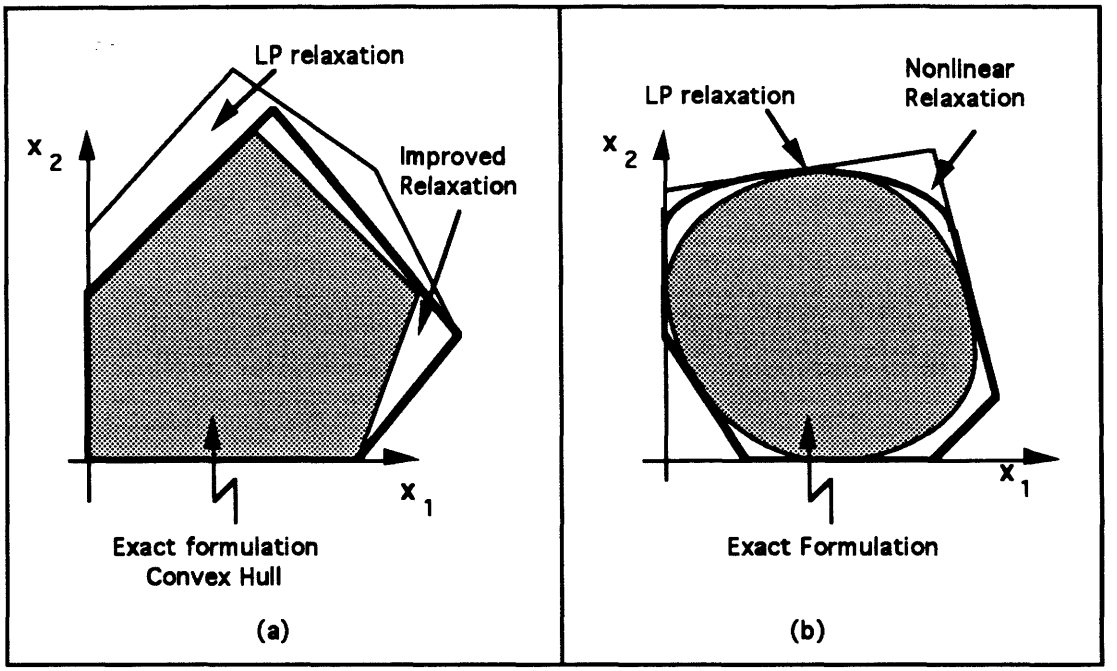

Figure 1: (a) Improved relaxations for integer programming (b) Improved relaxations for stochastic optimization problems.

previously, multi-armed bandit problems and branching bandits (see Weiss [29]).

All the scheduling problems mentioned above are computationally tractable: They are solved in polynomial time by index priority policies that are computed by a (polynomial-time) *adaptive greedy* algorithm (see Bertsimas and Niño-Mora [2]).

Papadimitriou and Tsitsiklis [22] have shown that several of the classical problems in stochastic optimization are computationally intractable. The multiclass queueing network scheduling problem is $EXPTIME-hard$, which implies that the problem cannot be solved in polynomial time, independently of the $P = NP$ question, while the restless bandit problem that we address in the present paper is $PSPACE-hard$. Bertsimas, Paschalidis and Tsitsiklis [3] have developed a mathematical programming approach for the multiclass queueing network scheduling problem. Using potential function ideas, they construct a sequence of polynomial-size linear programming relaxations that approximate with increasing accuracy the performance space. These relaxations provide polynomial-time bounds on linear performance objectives.

Contribution

Our contributions in the present paper are as follows:

1. We present a series of N linear programming relaxations for the restless bandit problem on N bandits (see next section). These relaxations capture increasingly higher order interactions among the bandits. These relaxations are increasingly stronger at the expense of higher computational times, the last one (Nth) being exact. These relaxations utilize the following projection representation idea nicely outlined in Lovász and Schrijver [19]:

   > 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 P has

exponentially many facets, we may be able to represent it as the projection of a polytope Q in higher (but still polynomial) dimension, having only a polynomial number of facets.

2. We propose a primal-dual heuristic that defines indices based on dual variables of the first order linear programming relaxation. Under a natural assumption, we interpret the heuristic as an index heuristic. We report computational results that suggest that the heuristic is exceptionally accurate. Primal-dual heuristics construct a linear programming relaxation of the problem, compute optimal primal and dual solutions of the relaxed formulation, and then construct a feasible solution for the original problem using information contained in the optimal primal and dual solutions. They have been proven quite effective for solving hard discrete optimization problem (see for example Bertsimas and Teo [5]).

Structure

The paper is structured as follows: In Section 2 we introduce the restless bandit problem and review previous research efforts. In Section 3 we strengthen a classical result on the performance space of a Markov decision chain and use it to present a monotone sequence of linear programming relaxations for the problem, the last one being exact. In Section 4 we introduce a primal-dual heuristic for the restless bandit problem, based on the optimal solution to the first-order relaxation. In Section 5 we address the tightness of the relaxations and the performance of the heuristic via computational testing. The last section contains some concluding remarks.

2 The Restless Bandit Problem: Description and Background

The restless bandit problem is defined as follows: There is a collection of N projects. Project $n \in \mathcal{N} = \{1, \ldots, N\}$ can be in one of a finite number of states $i_n \in E_n$, for $n = 1, \ldots, N$. At each instant of discrete time $t = 0, 1, 2, \ldots$, exactly $M < N$ projects must be operated. If project n, in state i_n, is in operation, then an *active* reward $R^1_{i_n}$ is earned, and the project state changes into j_n with an active transition probability $p^1_{i_n j_n}$. If the project remains idle, then a *passive* reward $R^0_{i_n}$ is received, and the project state changes into j_n with a passive transition probability $p^0_{i_n j_n}$. Rewards are discounted in time by a discount factor $0 < \beta < 1$. Projects are to be selected for operation according to an *admissible* scheduling policy u: the decision as to which M projects to operate at any time t must be based only on information on the current states of the projects. Let $\mathcal{U}$ denote the class of admissible scheduling policies. The goal is to find an admissible scheduling policy that maximizes the total expected discounted reward over an infinite horizon, i.e.,

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

where $i_n(t)$ and $a_n(t)$ denote the state and the action (active or passive), respectively, corresponding to project n at time t. We assume that the initial state of project n is i_n with probability α_{i_n}, independently of all other projects.

The restless bandit problem was introduced by Whittle [30], as an extension of the classical multi-armed bandit problem (see Gittins [16]). The latter corresponds to the special case that exactly one project must be operated at any time (i.e., $M = 1$), and passive projects are frozen: they do not change state ($p^0_{i_n i_n} = 1$, $p^0_{i_n j_n} = 0$ for all $n \in N$ and $i_n \neq j_n$).

As already mentioned, in contrast to the classical multi-armed bandit problem, the restless bandit problem is computationally intractable. Papadimitriou and Tsitsiklis [22] have proved that the problem is $PSPACE-hard$, even in the special case of deterministic transition rules and $M = 1$. In the multi-armed bandit case, Gittins and Jones [15] first showed that the optimal scheduling policy is a *priority index* policy: to each project state is assigned an index, and the policy operates at each time a project with largest index. The optimal *Gittins indices* are computable in polynomial time.

The restless bandit problem provides a very flexible modeling framework and, as a result, a number of interesting practical problems can be modeled naturally as restless bandits. As an indication of its modeling power we include the following examples:

Clinical trials (Whittle [30]). In this setting, projects correspond to medical treatments. The state of a project represents one's state of knowledge on the efficacy 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 efficacy of each treatment changes whether or not the treatment is tested.

Aircraft surveillance (Whittle [30]). 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 [30]). 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. Active selection of a project results in exhaustion of the corresponding worker, whereas passive selection results in recuperation.

Police control of drug markets. In this setting, M police units are trying to control N drug markets (see Caulkins [7]). The state of a project corresponds to the drug-dealing activity level of the corresponding drug market. Operation of a project-drug market corresponds to a focused police enforcement operation over that market, and tends to discourage drug-dealing activity. Nonoperation of a project, on the other hand, allows drug-dealing activity to grow in the corresponding market.

Control of a make-to-stock production facility (Veatch and Wein [26]). In this setting, M servers can produce N different classes of items; each finished item is placed in its respective inventory, which services an exogenous demand. The level of the inventory represents the state of each project (item class). Veatch and Wein [26] model a lost sales version of the problem as a restless bandit problem and propose heuristic indexing rules that perform quite well.

Whittle approached the restless bandit problem with dynamic programming methods. He presented a relaxed version of the problem, solvable in polynomial time. He then proposed an index heuristic based on the optimal solution of the relaxation. This index heuristic reduces to the Gittins index optimal policy when applied to the classical multi-armed bandit problem. A disadvantage is that Whittle's index heuristic only applies to a restricted class of restless bandits: those that satisfy a certain indexability property, which is difficult to check. Weber and Weiss [27] investigated the issue of asymptotic optimality of Whittle's index heuristic, as M and N tend to ∞, with M/N fixed. Working with continuous time restless bandits and with the

long-run average reward criterion, they showed that a sufficient condition for Whittle's heuristic to be asymptotically optimal is that the differential equation describing the fluid approximation to the index policy has a globally stable equilibrium point. They also presented instances that violate this condition, and in which Whittle's policy is not asymptotically optimal.

3 A Sequence of Relaxations for the Restless Bandit Problem

In this section we strengthen a classical result on the polyhedral characterization of the performance space for a finite discounted Markov decision chain (MDC) (see Heyman and Sobel [17]). We then formulate the restless bandit problem as a linear program over a certain *restless bandit polytope*. By applying the previous extension on polyhedral representations of MDCs, we present a monotone sequence of approximations to the restless bandit polytope (each approximation is tighter than the previous one), that yields a corresponding sequence of polynomial-size linear programming relaxations for the problem. These relaxations provide a monotone sequence of polynomial-time bounds for the optimal value of the restless bandit problem.

3.1 Polyhedral representations of Markov decision chains

Markov decision processes provide a general framework to model stochastic optimization problems. In this section we strengthen a classical result on the polyhedral characterization of the performance space for a finite discounted Markov decision chain.

Let $E = \{1, \ldots, n\}$ be the finite state space. At state $i \in E$ there is a finite set A_i of actions available. Let us denote $\mathcal{C}$ the state-action space,

$$\mathcal{C} = \{(i,a): i \in E, a \in A_i\}.$$

Let α_i be the probability that the initial state is i. If action $a \in A_i$ is taken in state i, then the chain moves to state j with probability p_{ij}^a. Let $0 < \beta < 1$ denote the discount factor. Let

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

An *admissible* policy is specified by a probability distribution on the actions A_i corresponding to every state i. If at state i action a is drawn from the corresponding distribution, then action a is taken. Let us denote $\mathcal{U}$ the class of all admissible policies. We call a policy *admissible* if the decision as to which action to take at any time t depends only on the current state. An *admissible* (i.e., nonanticipative) policy $u \in \mathcal{U}$ for selecting the actions generates a Markov chain. Let us associate with policy u the following performance measures:

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

Notice that $x_j^a(u)$ is the total expected discounted time spent taking action a in state j under policy u.

We are interested in finding a complete description of the corresponding performance space $X = \{x^u, u \in \mathcal{U}\}$. Let us consider the polyhedron

$$P = \{x \in \Re_+^n : \sum_{a \in A_j} x_j^a = \alpha_j + \beta \sum_{(i,a) \in \mathcal{C}} p_{ij}^a x_i^a, \quad j \in E, \}.$$

Notice that by summing over all $j \in E$ we obtain that $\sum_{(i,a)\in C} x_i^a = \frac{1}{1-\beta}$ and therefore P is a bounded polyhedron.

It was first shown by d'Epenoux [10] that, if all the initial probabilities $\alpha_j > 0$, then polytope P is a complete description of the performance space X (see also Heyman and Sobel [17]). He also showed that polytope P always contains the performance space, i.e., $X \subseteq P$.

We strengthen next that classical result, by proving that polytope P is *always* a complete description of performance space X, even if the assumption that all initial probabilities α_j are positive is dropped.

Theorem 1 (Performance Space of Discounted MDCs) (a) $X = P$.
(b) *The vertices of polytope P are achievable by stationary deterministic policies.*

Proof In Heyman and Sobel [17] it is shown that $X \subseteq P$ always. We will thus prove only the other inclusion, i.e., $P \subseteq X$.

Since P is a bounded polyhedron, any point in P can be written as a convex combination of its extreme points. Therefore, it suffices to show that any extreme point of P is achievable by some stationary deterministic policy, since any point of P can be achieved using a policy that randomizes over deterministic policies that achieve the corresponding extreme points.

Let $\overline{x}$ be an extreme point of polytope P. By standard linear programming theory, $\overline{x}$ is the unique maximizer of a linear objective function. Let $\sum_{(i,a)\in C} R_i^a x_i^a$ be such an objective. Since $\overline{x}$ is an extreme point, it has at most n positive components.
Let us now partition the state space E into two subspaces, E_1 and E_2, in the following way:

$$E_1 = \{j \in E:\ \overline{x}_j^a > 0 \quad \text{for some } a \in A_j\}, \text{ and } E_2 = \{j \in E:\ \overline{x}_j^a = 0 \quad \text{for all } a \in A_j\}.$$

Let $\overline{x}_{E_1} = \{\overline{x}_j^a,\ j \in E_1\}$. Consider now the following linear program:

$$\begin{aligned}
(LP_1) \qquad Z_{E_1} \;=\; & \max \sum_{j\in E_1} \sum_{a\in A_j} R_j^a x_j^a \\
& \text{subject to} \\
& \sum_{a\in A_j} x_j^a - \beta \sum_{i\in E_1} \sum_{a\in A_i} p_{ij}^a x_i^a = \alpha_j, \quad j \in E_1, \\
& x_j^a \geq 0, \quad j \in E_1, a \in A_j.
\end{aligned}$$

By construction, $\overline{x}_{E_1}$ is the unique optimal solution of linear program (LP_1), otherwise $\overline{x}$ would not be the unique maximizer. $\overline{x}_{E_1}$ is therefore an extreme point of (LP_1), and it has at most $|E_1|$ positive components. But by definition of E_1, it follows that $\overline{x}_{E_1}$ has exactly $|E_1|$ positive components, and for each state $j \in E_1$ there is exactly one action $\overline{a}_j \in A_j$ such that $\overline{x}_j^{\overline{a}} > 0$.

We can now define a stationary deterministic policy $\overline{u}$ that achieves $\overline{x}$: For each state $j \in E_2$ pick an arbitrary action $\overline{a}_j \in A_j$. Now, policy $\overline{u}$ deterministically takes action $\overline{a}_j$ in state j. Clearly, this policy achieves the vector $\overline{x}$, which completes the proof of (a) and (b). □

3.2 The Restless Bandit Polytope

In order to formulate the restless bandit problem as a linear program we define decision variables and characterize the corresponding feasible space. We introduce the indicators

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

and

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

Given an admissible scheduling policy $u \in \mathcal{U}$ let us define performance measures

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

and

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

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. Let us denote P the corresponding performance space,

$$P = \left\{ \mathbf{x} = (x_{i_n}^{a_n}(u))_{i_n \in E_n, a_n \in \{0,1\}, n \in \mathcal{N}} \mid \quad u \in \mathcal{U} \right\}.$$

It is clear that performance space P is a polytope. This follows from the fact that the restless bandit problem can be viewed as a discounted MDC (in the state space $E_1 \times \cdots \times E_N$), and the performance space of the latter is a polytope from Theorem 1.

We will refer to P in what follows as the *restless bandit polytope*. The restless bandit problem can thus be formulated as the linear program

$$(LP) \qquad Z^* \;=\; \max_{\mathbf{x} \in P} \sum_{n \in \mathcal{N}} \sum_{i_n \in E_n} \sum_{a_n \in \{0,1\}} R_{i_n}^{a_n} x_{i_n}^{a_n}.$$

The polytope P has been fully characterized in the special case of the classical multi-armed bandit problem by Bertsimas and Niño-Mora [2] as a polytope with special structure (an *extended polymatroid*). This characterization leads to strong structural properties of the optimal scheduling policy (Gittins priority index policy). For general restless bandits, however, it is highly unlikely that a complete description of polytope P can be found, since as mentioned above the problem is $PSPACE-hard$.

Our approach will be to construct approximations of polytope P that yield polynomial-size relaxations of the linear program (LP). We will represent these approximations $\hat{P} \supseteq P$ as projections of higher dimensional polytopes $\hat{Q}$. An advantage of pursuing this *projection representation* approach is that we will be able to represent approximations $\hat{P}$ of P with exponentially many facets as projections of polytopes $\hat{Q}$ with a polynomial number of facets, thus providing polynomial-time bounds on the optimal value Z^*. The approximations we develop are based on exploiting the special structure of the restless bandit problem as an MDC, and on applying Theorem 1.

3.3 A First-order Linear Programming Relaxation

Whittle [30] introduced a relaxed version of the restless bandit problem, solvable 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 over the infinite horizon must be $M/(1-\beta)$. Whittle showed that this relaxed version can be interpreted as the problem of controlling optimally N independent MDCs (one corresponding to each project), subject to one binding constraint on the discounted average number of active projects. In this section we formulate 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 E_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}.
Let

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

From a polyhedral point of view, Q_n^1 is the projection of the restless bandit polytope P over the space of the variables $x_{i_n}^{a_n}$ for project n. From a probabilistic point of view, Q_n^1 is the performance space of the first-order MDC corresponding to project n. In order to see this, we observe that as policies u for the restless bandit problem range over all admissible policies $\mathcal{U}$, they induce policies u_n for the first-order MDC corresponding to project n that range over all admissible policies for that MDC. Applying Theorem 1 we obtain:

Proposition 1 *The complete polyhedral description of Q_n^1 is given by*

$$Q_n^1 = \left\{ \mathbf{x}_n \geq \mathbf{0} \mid x_{j_n}^0 + x_{j_n}^1 = \alpha_{j_n} + \beta \sum_{i_n \in E_n} \sum_{a_n \in \{0,1\}} p_{i_n j_n}^{a_n} x_{i_n}^{a_n}, \quad j_n \in E_n \right\}. \tag{2}$$

Remark: A consequence of Proposition 1 is that the general restless bandit problem, with active and passive rewards, can be reduced to the case with active rewards only. This follows since by (2) the passive performance vector $\mathbf{x}_n^0(u)$ is a linear transformation of the active one, $\mathbf{x}_n^1(u)$.

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

$$\begin{aligned} \sum_{n \in \mathcal{N}} \sum_{i_n \in E_n} x_{i_n}^1(u) &= \sum_{t=0}^{\infty} E_u \left[\sum_{n \in \mathcal{N}} \sum_{i_n \in E_n} I_{i_n}^1(t) \right] \beta^t \\ &= \sum_{t=0}^{\infty} M \beta^t \\ &= \frac{M}{1-\beta} \end{aligned} \tag{3}$$

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

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

We will refer to the feasible space of linear program (LP^1) as the first-order approximation of the restless bandit polytope, and will denote it as P^1. Notice that linear program (LP^1) has $O(N|E_{\max}|)$ variables and constraints, where $|E_{\max}| = \max_{n \in \mathcal{N}} |E_n|$.

3.4 A Second-order Linear Programming Relaxation

In this section we present a second-order polynomial-size linear programming relaxation for the restless bandit problem, which is represented as the projection of a higher-dimensional polytope (introducing new variables). The new decision variables we introduce correspond to second-order performance measures for the restless bandit problem, associated with pairs of projects. Given a pair of projects, $n_1 < n_2$, the valid actions that can be taken over each pair of states, $(i_1, i_2) \in E_{n_1} \times E_{n_2}$, range over

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

Given an admissible scheduling policy u, let us define the second-order performance measures by

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

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 $E_{n_1} \times E_{n_2}$, the 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 $R^{a_{n_1}}_{i_{n_1}} + R^{a_{n_2}}_{i_{n_2}}$. 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 $p^{a_{n_1}}_{i_{n_1} j_{n_1}} p^{a_{n_2}}_{i_{n_2} j_{n_2}}$. The initial state is (i_{n_1}, i_{n_2}) with probability $\alpha_{i_{n_1}} \alpha_{i_{n_2}}$. Let

$$Q^2_{n_1,n_2} = \Big\{ \mathbf{x}_{n_1,n_2} = (x^{a_1 a_2}_{i_1 i_2}(u))_{i_1 \in E_{n_1}, i_2 \in E_{n_2}, (a_1,a_2) \in \mathcal{A}^2} \mid u \in \mathcal{U} \Big\},$$

be the projection of P over the space of the variables $(x^{a_1 a_2}_{i_1 i_2})_{i_1 \in E_{n_1}, i_2 \in E_{n_2}, (a_1,a_2) \in \mathcal{A}^2}$. An admissible scheduling policy u for the restless bandit problem induces an admissible scheduling policy u_{n_1,n_2} for the MDC corresponding to projects n_1 and n_2. It is easy to see that as u ranges over all admissible scheduling policies for the restless bandit problem, the corresponding induced policy u_{n_1,n_2} ranges over all admissible policies for the MDC. Therefore, the projection $Q^2_{n_1,n_2}$ is the performance space of the discounted MDC corresponding to the pair of projects (n_1, n_2), $(n_1 < n_2)$. Therefore, from Theorem 1 we obtain:

Proposition 2 *The complete polyhedral description of $Q^2_{n_1,n_2}$ is given by*

$$\sum_{(a_1,a_2) \in \mathcal{A}^2} x^{a_1 a_2}_{j_1 j_2} = \alpha_{j_1} \alpha_{j_2} + \beta \sum_{\substack{i_1 \in E_{n_1}, i_2 \in E_{n_2} \\ (a_1,a_2) \in \mathcal{A}^2}} p^{a_1}_{i_1 j_1} p^{a_2}_{i_2 j_2} x^{a_1 a_2}_{i_1 i_2}, \quad (j_1, j_2) \in E_1 \times E_2, \tag{4}$$

$$x^{a_1 a_2}_{i_1 i_2} \geq 0, \qquad (i_1, i_2) \in E_{n_1} \times E_{n_2}, \ (a_1, a_2) \in \mathcal{A}^2. \tag{5}$$

We can show some other second-order conservation laws to hold, based on combinatorial arguments. For all admissible scheduling policies u, we have, if $N \geq M + 2$,

$$\sum_{1 \leq n_1 < n_2 \leq N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} x^{00}_{i_1 i_2}(u) = \frac{\binom{N-M}{2}}{1-\beta}, \tag{6}$$

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

$$\sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} (x^{10}_{i_1 i_2}(u) + x^{01}_{i_1 i_2}(u)) = \frac{M(N-M)}{1-\beta}, \tag{7}$$

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 $M \geq 2$, we have

$$\sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} x^{11}_{i_1 i_2}(u) = \frac{\binom{M}{2}}{1-\beta}, \tag{8}$$

since at any time the $M \geq 2$ active projects give rise to $\binom{M}{2}$ active-active project pairs.

In order to lift the first-order approximation to the restless bandit polytope into a higher dimensional space with variables $x^{a_1 a_2}_{i_1 i_2}$ we need to relate the first and second-order performance measures. It is easy to see that, for any admissible policy u,

$$x^{a_1}_{i_1}(u) = \sum_{\substack{i_2 \in E_{n_2} \\ a_2 : (a_1, a_2) \in \mathcal{A}^2}} x^{a_1 a_2}_{i_1 i_2}(u), \quad i_1 \in E_{n_1}, a_1 \in \{0,1\}, 1 \le n_1 < n_2 \le N, \tag{9}$$

and

$$x^{a_2}_{i_2}(u) = \sum_{\substack{i_1 \in E_{n_1} \\ a_1 : (a_1, a_2) \in \mathcal{A}^2}} x^{a_1 a_2}_{i_1 i_2}(u), \quad i_2 \in E_{n_2}, a_2 \in \{0,1\}, 1 \le n_1 < n_2 \le N. \tag{10}$$

We define now the second-order relaxation, based on the above identities, as the linear program

$$\begin{aligned}
(LP^2) \quad Z^2 \;=\; & \max \sum_{n \in \mathcal{N}} \sum_{i_n \in E_n} \sum_{a_n \in \{0,1\}} R^{a_n}_{i_n} x^{a_n}_{i_n} \\
& \text{subject to} \\
& \mathbf{x}_{n_1,n_2} \in Q^2_{n_1,n_2}, \quad 1 \le n_1 < n_2 \le N, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} x^{00}_{i_1 i_2} = \frac{\binom{N-M}{2}}{1-\beta}, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} (x^{10}_{i_1 i_2} + x^{01}_{i_1 i_2}) = \frac{M(N-M)}{1-\beta}, \\
& \sum_{1 \le n_1 < n_2 \le N} \sum_{i_1 \in E_{n_1}} \sum_{i_2 \in E_{n_2}} x^{11}_{i_1 i_2} = \frac{\binom{M}{2}}{1-\beta},
\end{aligned}$$

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

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

$$\sum_{n\in\mathcal{N}} \sum_{i_n \in E_n} \sum_{a_n \in \{0,1\}} x_{i_n}^1 = \frac{M}{1-\beta},$$

$$x_i^a \ge 0.$$

We define the second-order approximation to the restless bandit polytope 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 P^2.

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

3.5 A kth-order Linear Programming Relaxation

In this section we generalize the results of the previous sections to present a kth-order linear programming relaxation for the restless bandit problem, corresponding to a kth-order approximation for the restless bandit polytope, for any $k = 1, \ldots, N$. This kth-order approximation corresponds again to lifting the first-order approximation polytope into a higher dimensional space, and then projecting back into the original first-order space.

In the kth-order case, we introduce new decision variables corresponding to performance measures associated with kth-order project interactions. For each k-tuple of projects $1 \le n_1 < \cdots < n_k \le N$, the admissible actions that can be taken at a corresponding k-tuple of states $(i_1, \ldots, i_k)$ range over

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

Given an admissible scheduling policy u for the restless bandit problem, we define kth-order performance measures, for each k-tuple $1 \le n_1 < \cdots < n_k \le N$ of projects, by

$$x_{j_1 \ldots j_k}^{a_1 \ldots a_k}(u) = E_u \left[\sum_{t=0}^{\infty} I_{j_1}^{a_1}(t) \cdots I_{j_k}^{a_k}(t) \beta^t \right], \quad j_1 \in E_{n_1}, \ldots, j_k \in E_{n_k}. \tag{11}$$

Analogously as in the first and second-order cases, the restless bandit problem induces a *kth-order MDC* over each k-tuple of projects $n_1 < \cdots < n_k$ in a natural way: The state space of the MDC is $E_{n_1} \times \cdots \times E_{n_k}$, the action space is $\mathcal{A}^k$, and the reward corresponding to state $(i_{n_1}, \ldots, i_{n_k})$ and action $(a_{n_1}, \ldots, a_{n_k})$ is $R_{i_{n_1}}^{a_{n_1}} + \cdots + R_{i_{n_k}}^{a_{n_k}}$. Rewards are discounted in time by discount factor β. The transition probability from state $(i_{n_1}, \ldots, i_{n_k})$ into state $(j_{n_1}, \ldots, j_{n_k})$, given action $(a_{n_1}, \ldots, a_{n_k})$, is $p_{i_{n_1} j_{n_1}}^{a_{n_1}} \cdots p_{i_{n_k} j_{n_k}}^{a_{n_k}}$. The initial state is $(i_{n_1}, \ldots, i_{n_k})$ with probability $\alpha_{i_{n_1}} \cdots \alpha_{i_{n_k}}$. Introducing the projection

$$Q_{n_1 \ldots n_k}^k = \Big\{ \mathbf{x}_{n_1 \ldots n_k} = (x_{i_1 \ldots i_k}^{a_1 \ldots a_k}(u))_{i_1 \in E_{n_1}, \ldots, i_k \in E_{n_k}, (a_1, \ldots, a_k) \in \mathcal{A}^k} \mid u \in \mathcal{U} \Big\}.$$

and arguing as before, we conclude that the projection $Q_{n_1 \ldots n_k}^k$ is the performance space of the discounted MDC corresponding to the k-tuple of projects $n_1 < \cdots < n_k$. From Theorem 1 we obtain

Proposition 3 *The complete polyhedral description of* $Q^k_{n_1 \ldots n_k}$ *is given by*

$$\sum_{(a_1,\ldots,a_k)\in\mathcal{A}^k} x^{a_1\ldots a_k}_{j_1\ldots j_k} = \alpha_{j_1}\cdots\alpha_{j_k} + \beta \sum_{\substack{i_1\in E_{n_1},\ldots,i_k\in E_{n_k} \\ (a_1,\ldots,a_k)\in\mathcal{A}^k}} p^{a_1}_{i_1 j_1}\cdots p^{a_k}_{i_k j_k}\, x^{a_1\ldots a_k}_{i_1\ldots i_k}, \quad j_1\in E_{n_1},\ldots,j_k\in E_{n_k}, \tag{12}$$

$$x^{a_1\ldots a_k}_{i_1\ldots i_k} \geq 0, \quad i_1\in E_{n_1},\ldots,i_k\in E_{n_k}, (a_1,\ldots a_k)\in\mathcal{A}^k. \tag{13}$$

Similarly as in the second-order case, we show that some additional kth-order conservation laws hold, by using combinatorial arguments. If u is an admissible scheduling policy, then

$$\sum_{1\leq n_1<\cdots<n_k\leq N} \sum_{i_1\in E_{n_1},\ldots,i_k\in E_{n_k}} \sum_{\substack{(a_1,\ldots,a_k)\in\mathcal{A}^k: \\ a_1+\cdots+a_k=r}} x^{a_1\ldots a_k}_{i_1\ldots i_k}(u) = \frac{\binom{M}{r}\binom{N-M}{k-r}}{1-\beta}, \tag{14}$$

for

$$\max(0, k-(N-M)) \leq r \leq \min(k, M). \tag{15}$$

Conservation law (14) follows since at each time the number of k-tuples of projects that contain exactly r active projects is $\binom{M}{r}\binom{N-M}{k-r}$, for r in the range given by (15).

In order to lift the first-order approximation to the restless bandit polytope into a higher dimensional space with variables $x^{a_1\ldots a_k}_{i_1\ldots i_k}$ we need to relate the first and kth-order performance measures. It is easy to see that if u is an admissible scheduling policy, then for $n\in\mathcal{N}$, $i\in E_n$, $a\in\{0,1\}$, $1\leq r\leq k$, $n_r = n$, and $n_1 < \ldots < n_k$, we have

$$x^a_i(u) = \sum_{\substack{(a_1,\ldots,a_k)\in\mathcal{A}^k \\ i_1\in E_{n_1},\ldots,i_k\in E_{n_k}: \\ i_r=i, a_r=a}} x^{a_1\ldots a_k}_{i_1\ldots i_k}(u). \tag{16}$$

We now define the kth-order relaxation of the restless bandit problem as the linear program

$$\begin{aligned}
(LP^k) \quad Z^k \;=\; & \max \sum_{n\in\mathcal{N}} \sum_{i_n\in E_n} \sum_{a_n\in\{0,1\}} R^{a_n}_{i_n} x^{a_n}_{i_n} \\
& \text{subject to} \\
& \mathbf{x}_{n_1\ldots n_k} \in Q^k_{n_1\ldots n_k}, \\
& \sum_{1\leq n_1<\cdots<n_k\leq N} \sum_{i_1\in E_{n_1},\ldots,i_k\in E_{n_k}} \sum_{\substack{(a_1,\ldots,a_k)\in\mathcal{A}^k: \\ a_1+\cdots+a_k=r}} x^{a_1\ldots a_k}_{i_1\ldots i_k} = \frac{\binom{M}{r}\binom{N-M}{k-r}}{1-\beta}, \\
& \max(0, k-(N-M)) \leq r \leq \min(k, M),
\end{aligned}$$

$$x_i^a = \sum_{\substack{(a_1,\ldots,a_k)\in\mathcal{A}^k \\ i_1\in E_{n_1},\ldots,i_k\in E_{n_k}\,: \\ i_r=i, a_r=a}} x_{i_1\ldots i_k}^{a_1\ldots a_k},$$

$$x_i^a \geq 0.$$

We define the kth-order approximation to the restless bandit polytope P as the projection of the feasible space of linear program (LP^k) into the space of the first-order variables, x_i^a, and denote it P^k. It is easy to see that the sequence of approximations is monotone, in the sense that

$$P^1 \supseteq P^2 \supseteq \cdots \supseteq P^N = P.$$

Notice that the kth-order relaxation (LP^k) has $O(N^k|E_{\max}|^k)$ variables and constraints, for k fixed. Therefore, the kth-order relaxation has polynomial size, for k fixed.

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

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 an optimal primal and dual solution of the first-order relaxation, (LP^1). Under some mixing assumptions on the active and passive transition probabilities, we can interpret the primal-dual heuristic as an index heuristic. The dual of the linear program (LP^1) is

$$\begin{aligned}
(D^1) \quad Z^1 \;=\; & \min \sum_{n\in\mathcal{N}} \sum_{j_n\in E_n} \alpha_{j_n}\lambda_{j_n} + \frac{M}{1-\beta}\lambda \\
& \text{subject to} \\
& \lambda_{i_n} - \beta \sum_{j_n\in E_n} p_{i_n j_n}^0 \lambda_{j_n} \geq R_{i_n}^0, \quad i_n \in E_n, \quad n\in\mathcal{N}, \\
& \lambda_{i_n} - \beta \sum_{j_n\in E_n} p_{i_n j_n}^1 \lambda_{j_n} + \lambda \geq R_{i_n}^1, \quad i_n \in E_n, \quad n\in\mathcal{N}, \\
& \lambda \geq 0. \qquad (17)
\end{aligned}$$

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

$$\overline{\gamma}_{i_n}^0 = \overline{\lambda}_{i_n} - \beta \sum_{j_n\in E_n} p_{i_n j_n}^0 \overline{\lambda}_{j_n} - R_{i_n}^0,$$

$$\overline{\gamma}_{i_n}^1 = \overline{\lambda}_{i_n} - \beta \sum_{j_n\in E_n} p_{i_n j_n}^1 \overline{\lambda}_{j_n} + \overline{\lambda} - R_{i_n}^1,$$

which are nonnegative. It is well known (cf. Murty [20], pp. 64-65), that the optimal reduced costs have the following interpretation:

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

$\overline{\gamma}^0_{i_n}$ 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, $(i_1, \ldots, i_N)$, an optimal primal solution to (LP^1), $\{\overline{x}^{a_n}_{j_n}\}$, and the corresponding optimal reduced costs, $\{\overline{\gamma}^{a_n}_{j_n}\}$, and produces as output a vector with the actions to take on each project, $(a^*(i_1), \ldots, a^*(i_N))$. An informal description of the heuristic, with the motivation that inspired it, is as 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 p candidate projects are set active and the heuristic stops. 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 $x^1_{i_n}$. 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 $\overline{x}^1_{i_n} > 0$. Recall that by complementary slackness, $\overline{\gamma}^1_{i_n} = 0$ if $\overline{x}^1_{i_n} > 0$. As a heuristic index of the desirability of setting project n in state i_n active we take the passive reduced cost $\overline{\gamma}^0_{i_n}$. The motivation is given by the interpretation of $\overline{\gamma}^0_{i_n}$ stated above: the larger the *passive index* $\gamma^0_{i_n}$ 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 $x^0_{i_n}$. 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 reduces to **an indexing 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: $G_n = (E_n, A_n)$, where $A_n = \{(i_n, j_n) |\ p^0_{i_n j_n} > 0$, and $p^1_{i_n j_n} > 0\ i_n, j_n \in E_n\}$. We assume that

Assumption A: For every $n \in \mathcal{N}$ at least one of the following two conditions is satisfied:
a) $\alpha_{j_n} > 0$ for all $j_n \in E_n$, b) the directed graph G_n is connected.

Given that the polytope P^1 has independent constraints for every $n \in \mathcal{N}$ and only one global constraint, elementary linear programming theory establishes that

Proposition 4 *Under assumption A, every optimal extreme point $\overline{x}$ solution of the polytope P^1 has the following properties: a) There is at most one project k and at most one state $i_k \in E_k$, for which $\overline{x}^1_{i_k} > 0$ and $\overline{x}^0_{i_k} > 0$.*
b) For all other projects n and all other states either $\overline{x}^1_{i_n} > 0$ or $\overline{x}^0_{i_n} > 0$.

Therefore, starting with an optimal extreme point solution $\overline{x}$ and a complementary dual

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 the projects* }

{ *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 no more than 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 the* $M - p$ *additional 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 the 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 for Restless Bandits

optimal solution, with corresponding reduced costs $\overline{\gamma}$, let us consider the following index rule:
Index heuristic:

1. Given the current states $(i_1, \ldots, i_N)$ of the N projects, compute the indices

$$\delta_{i_n} = \overline{\gamma}^1_{i_n} - \overline{\gamma}^0_{i_n}.$$

2. Set active the projects that have the M smallest indices. In case of ties, set active projects with $\overline{x}^1_{i_n} > 0$.

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

5 Computational Experiments

In this section we address the tightness of the relaxations and the performance of the primal-dual heuristic introduced previously.

For the usual bandit problems with the average reward criterion ($\beta = 1$), Bertsimas, Paschalidis and Tsitsiklis [4] show that the second order relaxation is exact, i.e., $P^2 = P$. Moreover, in this case P is an extended polymatroid. For discounted problems, however, the second order relaxation is not exact, i.e., $P^2 \neq P$ even though P is still an extended polymatroid.

In order to address the tightness of the relaxations and the heuristic for restless bandit problems we performed a series of computational experiments. For each test problem we computed the following measures:

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 before.

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 linear programming 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 the following 7 problem instances.

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 were not able to compute the optimal solution because of the large size of the instance. Since these instances were randomly generated we expected that the greedy heuristic would perform very close to the optimal solution. To test the last statement we generated Problem 3 that 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 the greedy algorithm would not perform optimally. 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 data sets are available upon request from the authors.

In Table 2 we report the results of our experiments for various values of the discount factor β. Some observations on the results, shown in Table 2, are:

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. Given that we expect that the greedy heuristic is near optimal for randomly generated instances (as a verification Problem 3 had also randomly generated data and the greedy heuristic was extremely close to the optimal solution), we believe that 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 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.

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

Problem instance	β	Z_{Greedy}	Z_{PDH}	Z^*	Z^2	Z^1
Problem 1	0.20	59.9	59.9		70.62	74.67
	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
	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
	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
	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
	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
	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
	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.

that the primal-dual heuristic has excellent performance, while the second order relaxation is quite strong. Our approach has the attractive feature that can produce increasingly stronger suboptimality guarantees 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 decade, 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 complexity ($PSPACE - hard$) of the problem we expect that these ideas have wider applicability. We intend to pursue these ideas further in the context of other classical stochastic optimization problems.

References

[1] Bertsimas, D. (1994). A mathematical programming approach to stochastic optimization problems. Operations Research Center, MIT, working paper.

[2] Bertsimas, D. and Niño-Mora, J. (1993). Conservation Laws, Extended Polymatroids and Multi-armed Bandit Problems; a Unified Approach to Indexable Systems. To appear in *Math. Oper. Res.*

[3] 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.

[4] Bertsimas, D., Paschalidis, I. and Tsitsiklis, J. (1994). Branching bandits and Klimov's problem: achievable region and side constraints. Operations Research Center, MIT, working paper.

[5] Bertsimas D. and Teo C. (1994). From valid inequalities to heuristics: a unified view of primal-dual approximation algorithms in covering problems. Operations Research Center, MIT, working paper.

[6] Bhattacharya, P. P., Georgiadis, L. and Tsoucas, P. (1992). Extended Polymatroids: Properties and Optimization. In *Proceedings of Second International Conference on Integer Programming and Combinatorial Optimization*, E. Balas, G. Cornuéjols, R. Kannan (Eds.), Carnegie Mellon University, Mathematical Programming Society, 298-315.

[7] Caulkins, J. P. (1993). Local Drug Market's Response to Focused Police Enforcement. *Oper. Res.* **41** 848-863.

[8] Coffman, E. G., Jr., Hofri, M. and Weiss, G. (1989). Scheduling Stochastic Jobs with a Two Point Distribution on Two Parallel Machines. *Probab. Eng. Inform. Sci.* **3** 89-116.

[9] Coffman, E. G., Jr., and Mitrani, I. (1980). A Characterization of Waiting Time Performance Realizable by Single Server Queues. *Oper. Res.* **28** 810-821.

[10] d'Epenoux, F. (1960). Sur un Problème de Production et de Stockage dans l'Aléatoire. *Revue Française de Informatique et Recherche Opérationnelle* **14** 3-16.

[11] Edmonds J. (1970), Submodular functions, matroids and certain polyhedra. in *Combinatorial Structures and their Applications*, 69-87. R. Guy et. al. (Eds.), Gordon & Breach, New York.

[12] Federgruen, A. and Groenevelt, H. (1988). Characterization and Optimization of Achievable Performance in General Queueing Systems. *Oper. Res.* **36** 733-741.

[13] Federgruen, A. and Groenevelt, H. (1988). *M/G/c* Queueing Systems with Multiple Customer Classes: Characterization and Control of Achievable Performance under Nonpreemptive Priority Rules. *Management Sci.* **34** 1121-1138.

[14] Gelenbe, E. and Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems.* Academic Press, London.

[15] 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.

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

[17] Heyman, D. P. and Sobel, M. J. (1984). *Stochastic Models in Operations Research, vol. II: Stochastic Optimization.* McGraw-Hill, New York.

[18] Klimov, G. P. (1974). Time Sharing Service Systems I. *Theory Probab. Appl.* **19** 532-551.

[19] Lovász, L. and Schrijver, A. (1991). Cones of Matrices and Set-Functions and $0-1$ Optimization. *SIAM J. Optimization* **1** 166-190.

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

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

[22] Papadimitriou, C. H. and Tsitsiklis, J. N. (1993). The Complexity of Optimal Queueing Network Control. Working Paper, MIT.

[23] Papadimitriou, C. H. (1994). *Computational Complexity.* Addison-Wesley, Reading, Massachusetts.

[24] Shanthikumar, J. G. and Yao, D. D. (1992). Multiclass Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control. *Oper. Res.* **40** S293-299.

[25] Tsoucas, P. (1991). The Region of Achievable Performance in a Model of Klimov. Research Report RC16543, IBM T. J. Watson Research Center, Yorktown Heights, New York.

[26] Veatch, M. and Wein, L. M. (1992). Scheduling a Make-to-Stock Queue: Index Policies and Hedging Points. Operations Research Center, MIT, working paper OR 266-92.

[27] Weber, R. R. and Weiss, G. (1990). On an Index Policy for Restless Bandits. *J. Appl. Prob.* **27** 637-648.

[28] Weber, R. R. and Weiss, G. (1991). Addendum to "On an Index Policy for Restless Bandits." *Adv. Appl. Prob.* **23** 429-430.

[29] Weiss, G. (1988). Branching Bandit Processes. *Probab. Engin.. Inform. Sci.* **2** 269-278.

[30] 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.