UNIVERSIDAD CARLOS III DE MADRID
DEPARTMENT OF STATISTICS

Mathematical Programming Approach to Dynamic Resource Allocation Problems

Ph.D. Thesis Proposal

Peter Jacko

November 18, 2005

UNIVERSIDAD CARLOS III DE MADRID
DEPARTMENT OF STATISTICS

Ph.D. Thesis Proposal on

Mathematical Programming Approach to Dynamic Resource Allocation Problems

for the Ph.D. Program in

Business Administration and Quantitative Methods

by

Peter Jacko

advised by

prof. José Niño-Mora

1. Introduction

Economical decision making under uncertainty is one of the most important challenges of everyday life. People have developed, based on their beliefs or intuition, a variety of surprisingly simple heuristic rules to deal with many complex situations that are often, also surprisingly, nearly-optimal. It is of a great practical interest, especially when such decision making is to be automatized, to describe the circumstances, in which such heuristics indeed lead to optimal decisions, and to provide suboptimality bounds of these heuristic rules in general.

Typically, any reward-yielding activity requires to invest our effort, time, money or another scarce resource, which is costly to use. To make a rational choice, decision-maker needs to answer two basic questions: Is it worth to invest the scarce resource in the activity? If so, How much of it should be invested? The situation often gets more complicated due to availability of several alternative reward-yielding activities, among which our scarce resource must be distributed. In such a *resource allocation problem*, an additional question arises: How to choose the activities to invest in?

In this Ph.D. thesis proposal we present a possible approach to answer those three questions *dynamically*, that is, reconsidering the decision on the scarce resource allocation regularly in time. The need for dynamical decision-making arises whenever the activities one invests in have any of the following features: (1) the decision-maker does not have perfect information about the reward that the activity yields, (2) the reward is known, but subject to a random factor, (3) the reward is known, but changes over time. We will focus on the situation when (2) and (3) hold simultaneously. Thus, we wish to deal with those cases, in which the decision-maker faces a trade-off between exploitation (taking the certain reward of today) and exploration (obtaining possibly higher reward tomorrow).

The Ph.D. thesis analyzing these situations may be of both theoretical and practical value. Due to the large variety and significant complexity of dynamic resource allocation problems, they are typically addressed, analyzed, respectively solved by ad-hoc techniques. The achievable region approach, we propose to use, has been proved recently to be well-grounded and tractable in several diverse problems, and may be enriched and justified by our work. From the practical point of view, the decision making under uncertainty arises in the areas as diverse as

product (R&D) management, financial economics, optimal consumption planning, telecommunications, medicine, engineering systems, etc., where a well-reasoned advice is more than needed.

In order to arrive to the model that can accomplish our objectives, we first discuss some basic features of a powerful modeling setting of *Markov Decision Processes* in Section 2. Since our model typically enjoys significant complexity, classical solution methods, such as Bellman equations, become intractable even in rather simple cases. An alternative *achievable region approach*, which may overcome computational difficulties, is outlined in Section 3. Section 4 contains a brief review of the framework and applications of the classical *bandit problem*, which served as an important modeling paradigm for dynamical resource allocation problems over last two decades. Nevertheless, this model possesses a significant limitation due to the unrealistic assumption that the part of the world, which the decision-maker has not invested in, does not change. This assumption is dropped in the *restless bandit problem*, presented in Section 5, where also its *work-reward* extension, the setting we will be interested in, is formulated. After a discussion of the few attempts of employing this model in practical applications, we outline in Section 6 our hypotheses and the future investigation lines.

2. Markov Decision Processes

In decision making, a controller often has an opportunity to influence by her *actions* the future evolution of an underlying system at various points in time. In such a *sequential decision process*, there may be costs or rewards of some scarce resources, incurred over time, that depend on the actions taken and the way in which the system evolves. The goal of the controller may be to maximize the total (expected) reward or minimize the total (expected) cost over a certain time horizon. If the horizon is infinite, then one may need to use discounting or long-run averaging in order to have a finite-valued objective (Stidham 2002). Nevertheless, such alternations of objective function may also be relevant in some finite horizon problems. Another interesting class of sequential decision problems, so-called *optimal stopping problems*, is concerned with finding a time horizon which leads to the optimal value of controller's objective.

When the information needed to predict the future evolution of the system is contained in the current *state* of the system and depends on the current action, we call such a sequential decision process a *Markov decision process* (MDP). MDP has a great modeling power, which can provide results on the existence and structure of good policies and on methods for the computation of optimal policies. Therefore, it has naturally been used in a variety of applications in areas including engineering systems, operations research, management science, economics and applied probability.

The modeling and solving these optimization problems is sometimes referred to as *stochastic dynamic programming*, since those problems are *dynamic* in that actions are taken over time and actions taken now have repercussions in the future, and *stochastic* in that they involve uncertainty of random state changes over time. In some literature also other equivalent names are used, such as *sequential stochastic optimization*, and *stochastic control* (typically for continuous state problems).

The theory of stochastic dynamic programming has been developed in two rather separated streams, for discrete- and continuous-time models, respectively. In further discussion we will focus on *discrete-time* MDPs, which is an important setting from at least two points of view: (1) there is a large number of interesting problems being naturally modeled in the discrete time setting and (2) this theory

is useful as an approximate computational technique for continuous-time models (Davis 1993).

It turns out that a useful solution concept for an MDP is a *non-anticipative policy*, which is defined as a set of rules specifying the action to be taken for each decision point in time and for each possible state of the system. Such a policy is appropriate, because MDPs are of Markovian nature, i.e. the future evolution of the system depends on history only through the current state. Moreover, in dynamic stochastic systems it is not possible to have information about future states at a decision moment, therefore the decisions should not be based on them (non-anticipative).

A policy thus answers a family of questions: What action should be taken at a given time if the system is in a given state? As we will see later, a class of *stationary policies* is often of high interest. A policy is stationary, if the answer to the question just stated does not depend on the point in time (i.e. is time-homogeneous). Stationary policies can thus be characterized in a simple way (as a vector of cardinality equal to the number of system states), which allows an easier implementation in practice.

The breakthrough in dynamic stochastic programming was an approach, now called *dynamic programming*, invented by Richard Bellman in the 1950's, which exploits the fact that nothing is ever lost by postponing a decision until the last possible moment. In doing so, we may be able to make a more accurate prediction about the future evolution of the system. Actually, dynamic programming is a quite effective method for solving MDPs (Stidham 2002).

The idea of the dynamic programming is reflected in the *Principle of Optimality*: at any point in time, an optimal policy must prescribe an action that optimizes the sum of immediate reward (cost) and (expected) objective value obtained if an optimal policy is applied from the subsequent point in time on. The mathematical expression associated to the Principle of Optimality is the optimality equations of dynamic programming, called the *Bellman equations*. For infinite-horizon problems, Bellman equations simplify so that they are not time-dependent; indeed, the optimal objective value is a unique fixed point solution. For finite-horizon problems, there are two common methods: value iteration and policy iteration.

Why dynamic programming gets so much importance is because of its both theoretical and practical power. Dynamic programming provides a coherent the-

oretical framework for studying sequential decision processes. As such, it leads to several general theoretical results, for example, a necessary and sufficient condition for optimality of a stationary policy in some broad cases. From practical point of view it is remarkable that the dynamic programming approach reduces optimization over the sequence of decisions in various points in time to a sequence of parameter optimizations for every time point, thus, it may significantly decrease the problem complexity.

Still, for many problems this may be not enough to make the solution of the problem tractable. A typical knot arising in practical computation is that the dynamic programming recursions may be too many (or infinitely many) to allow actual computation; the size of dynamic programming formulation is typically exponential on size of model (*curse of dimensionality*). Here comes out a necessity for other approaches. One of the solution approach alternatives is *linear programming* (LP) reformulation of Bellman equations. Since each Bellman equation includes an optimization term, it can be relaxed to a set of linear inequalities, one for each action. Once this has been done with all Bellman equations, one adds an objective function that forces at least one inequality to be satisfied sharply for each state. From the solution to this associated LP problem, one can readily get the optimal policy for the original MDP. As Stidham (2002) points out, the LP approach is especially well suited to constrained MDPs, in which the optimal policy must satisfy side constraints, what allows to reduce the set of *admissible policies*.

However, the LP reformulation as such does not help to deal with the curse of dimensionality. A new approach, based on the concept of conservation laws, allows to create new, much simpler, LP formulation of MDPs. We discuss this modeling framework in the next section and present this approach applied to a particular problem later in the text.

3. Achievable Region Approach and Conservation Laws

The linear programming approach is closely connected to graphical interpretation of problems and is thus very well suited for providing insights of the solution methods and helping to exploit the problem structure. To each policy one can associate a *performance vector* ranging over all the system states. Then, a set of admissible policies (which depends on a given problem) defines an *achievable region* (or *performance region*), which is, in other words, the space of all possible performances. Structural properties of this achievable region lead to structural properties in the given problem. We may therefore be interested in describing the achievable region so that the optimization problem can be efficiently solved by classical mathematical programming methods. When an analysis via this methodology is available, one can typically make clear and strong statements about the (optimal) policies.

For stochastic dynamic problems (or MPDs), it is natural to specify admissible policies as *non-anticipative*, i.e. a policy can only make use of past history (which is in turn reflected in the current state of the system), but not of any future information. Further, admissible policies must not affect the stochastic mechanism and the reward (cost) structure of the system. Most of the applications of the achievable region approach have focused on performance vectors that are expectations. That should not be surprising, as the most appropriate measure one can utilize in a dynamic stochastic system at a given point in time is an expectation of its future behavior.

The earliest intentions to use this approach were done in queueing theory, originated in Klimov (1974) and Coffman & Mitrani (1980), later followed by Federgruen & Groenevelt (1988) in a more general framework of a certain family of queueing models. In the latter contribution it was showed that the performance region in those models is a polytope of special type. An important concept of (strong) *conservation laws* was introduced in Shanthikumar & Yao (1992), where the previous results were extended by proving a powerful result about the achievable region approach: When the performance vectors satisfy strong conservation laws, the achievable region is a particular polytope (called the *base of a polymatroid*, previously known in combinatorial optimization), completely characterized by those laws, and the set of vertices of the achievable region is equivalent to the

set of performance vectors obtained by all *index policies*. Then, optimization of a linear objective can be accomplished by a greedy algorithm, which indeed finds an optimum in a vertex, hence ensuring that there is an optimal index policy (Stidham 2002). We will discuss index policies in the next section, when a particular stochastic dynamic problem, so called *multi-armed bandit problem*, is treated.

Bertsimas (1995) and Bertsimas & Niño-Mora (1996), drawing on the work of Tsoucas (1991), extended those results to a more complex class of stochastic dynamic problems, which they, borrowing the name from a related paper by Whittle (1988), called *indexable*. They defined *generalized conservation laws*, whose satisfaction by performance vectors implies that the achievable region is a polytope of special structure. Moreover, optimization of a linear objective over such a polytope is solved by an *adaptive-greedy algorithm* based on Klimov's (1974), which, again, leads to an optimal index policy. More general results in a similar fashion using *partial conservation laws* were obtained in Niño-Mora (2001), Niño-Mora (2002) and a semi-Markov version in Niño-Mora (2005), where the analysis is closely tied to the *restless bandit problem*, which will be discussed later in the paper.

Polytopes treated in the listed papers were exploited mainly in the context of queueing systems and networks. Indeed, in a presentation of the achievable region approach by Dacre et al. (1999), the method is explained with a reference to a simple queueing system. A formal exposition of conservation laws and their relation to polyhedral structure of the performance regions of a broad class of queueing stochastic systems can be found in Yao & Zhang (1997a). An extension of stochastic dynamic problems with *side constraints* (i.e. controller specified constraints as opposite to system-defined constraints for performance vectors), satisfying generalized conservation laws, is analyzed in Yao & Zhang (1997b).

It is interesting to realize that the achievable region, defined by the performance vectors associated with admissible policies, is independent from the optimization objective. In the cases when the achievable region is a polytope, the linear objective function may not be the only one to imply that the optimal policy is a vertex (i.e. an index policy). Given a particular stochastic dynamic problem satisfying conservation laws (and therefore a particular polytope), one may be able to define a class of nonlinear objectives and associated optimal index policies. Dacre et al. (1999) touched this idea and indeed showed in a particular problem they treated that an optimal policy is, by no surprise, a randomization of two index

policies.

In dynamic programming, value iteration and policy improvement algorithms are (virtually) routinely available and there is nothing equivalent in the achievable region approach where plenty of creative thinking may be involved in a successful application of the ideas (Dacre et al. 1999). The advantage of the latter is that it can exploit the special structure of problem, where the general purpose algorithms (such as dynamic programming) become cumbersome or intractable.

4. Bandit Problems

In this section we review a dynamic resource allocation problem called *multi-armed bandit problem* and a number of its extensions. Although it is a classical problem of stochastic dynamic optimization, which can be formulated in a very simple way, its solution had been a challenging open problem for a considerably long time. The multi-armed bandit problem, originally described by Robins (1952), is a simple model of a controller trying to optimize her decisions while acquiring knowledge in the same time. It is a typical case of the trade-off between *exploitation* (getting the highest immediate rewards) and *exploration* (learning about the system and receiving possibly even higher rewards later).

The multi-armed bandit problem is named by an analogy to the one-armed bandit machine. In the multi-armed case, the gambler has to decide which arm to pull in order to maximize his total reward in a series of trials. In the following, we will suppress the name *bandit* and will instead call the reward-yielding activity a *project*, in order to highlight its broad applicability and stress the framework it offers for dynamic resource allocation problems. Now we can rephrase that the multi-armed bandit problem is concerned with the question of how to dynamically allocate a single scarce resource amongst several alternative projects (Weber 1992).

In order to be able to analyze decision policies, whose worth is yielded by the actual future evolution of the projects, the controller must assume certain structure of project's possible future behavior. In the most common formulations, the projects are assumed to yield rewards following a given probability distribution with unknown (or uncertain) parameter(s). However, a slightly different framework of defining projects' dynamics via stationary transition probabilities over a set of *project states* with associated rewards often leads to a more tractable model.

The knowledge about the history of projects' rewards (or, projects' states) may, in many cases, be helpful in the decision making process. To take advantage of that information, it has been proved useful to model projects with Markovian dynamics, thus obtaining an MDP formulation of the basic multi-armed bandit problem, as follows.

Multi-Armed Bandit Problem

There are K projects, labeled by $k \in \mathcal{K}$. Let $x_k(t) \in \mathcal{X}_k$, for a finite state space $\mathcal{X}_k$, be the state of project k at time epoch $t \in \mathcal{T} = \{0, 1, 2, \dots\}$. At each time

epoch t the controller must decide about allocation of a scarce resource, which we will call *work*, to one of the projects. If project k is selected to be worked on, an immediate reward $r_k(x_k(t))$ is received and the project changes to state $x_k(t+1)$ according to a stationary Markov transition probability matrix $P_k = \{p_k(i,j)\}$; the remaining projects remain frozen, i.e. no reward is earned and no state change occurs.

The controller's objective is to optimize a function related to the future rewards stream, which is typically taken as an expectation. The most relevant and also the most investigated objective is maximization of the *expected total discounted reward* given by

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t r(t)\right] \tag{1}$$

where $0 < \beta < 1$ is a discount factor and $r(t)$ is the reward earned in time epoch t, defined as $r_k(x_k(t))$ if project k is being worked on at time t. The optimization is done over a set of all *admissible policies* $\pi \in \Pi$, which are those that in each time epoch t select one project to work on, based only on the actual state vector $x(t) = (x_k(t))_{k\in\mathcal{K}}$.

It is due to Gittins and his colleagues that this problem has been solved. The initial publication of the results (Gittins & Jones 1974) attracted very little attention; just the discussion meeting of Royal Statistical Society (Gittins 1979) made the solution spread. The essence of the solution was definition of the *Gittins index*, a function of project and its state, defined as a fair price for purchase of the project in that state if the price offer is to remain open in the future (Whittle 2002). The optimal policy is to work on the project of currently greatest index; therefore receiving name an *index policy*. The significance of this index result is in that it decomposes the K-dimensional bandit problem into K one-dimensional problems. The Gittins theory also has its continuous-time counterpart and can be used in semi-Markov decision problems (Gittins 1979).

Gittins proposed to assign an index to each state x of each project k, which can be expressed as

$$\nu_k(x) = \max_{\tau>0} \frac{\mathbb{E}\left\{\sum_{t=0}^{\tau-1}\beta^t r_k(x_k(t))|x_k(0)=x\right\}}{\mathbb{E}\left\{\sum_{t=0}^{\tau-1}\beta^t|x_k(0)=x\right\}} \tag{2}$$

where the maximization is over the set of all stopping times $\tau \in \mathcal{T} \cup \{+\infty\}$ such that $\tau > 0$. In other words, the Gittins index defines a worth of each project state calculated as the maximal attainable *marginal expected reward* (i.e. expected reward per unit of expected discounted time), given that we start working on that project from the first period. Notice that these quantities are time-independent and depend only on information concerning bandit k.

An *index policy* is a working strategy which at each time epoch prescribes to work on a project, whose current state's Gittins index is greatest. It will be convenient to denote, for each time epoch t, the Gittins index of project k as

$$\nu_k(t) = \nu_k(x_k(t)).$$

It can be shown that $\nu_k(x)$ is well defined and bounded (in the finite state space we assume). An important property of the Gittins index is that the maximum in (2) is achieved, and in fact it is achieved by $\tau_k(x)$, the first time the project comes to a state, whose Gittins index is lower than the original one,

$$\tau_k(x) = \min\{t : \nu_k(t) < \nu_k(x)\}. \tag{3}$$

Notice that this property is easy to see in the multi-armed bandit problem with frozen rested projects, but in more general settings does not necessarily hold.

From MDP theory (Blackwell 1965) it is known that if the set of possible actions (allowed in a fixed time epoch) of a finite-state MDP is finite and the same for all the system states, then there is a deterministic stationary Markov policy that maximizes (1). This result, notes Gittins (1979), applies to the multi-armed bandit model (note that the set of actions is given by *to work* and *to rest*), so when looking for the optimal policy, attention may be restricted to deterministic stationary policies, which significantly simplifies the structure of the problem. Thereafter, Gittins outlined how to calculate project indices (using solutions of some stopping time problems) and showed a difficult-to-follow proof of optimality of such an index policy. As Varaiya et al. (1985) remarked, the optimality of the index policy is almost trivially implied by two features of the multi-armed bandit problem: that the rested projects are frozen and they contribute no reward. Furthermore, the Markovian dynamics is useful only in that it permits a simple calculation of Gittins index.

Whittle (1980) then presented a proof by explicit solution of the dynamic programming formulation of the problem. Weber (1992) took it even further by introducing a brief and almost verbal proof, which afforded a better insight to the problem and, as Whittle (2002) writes, deserves to be characterized by Erdös' term "God's proof". Weber used a concept of project's *fair charge* the controller has to pay if works on the project so that he arrived to a multiple of the Gittins index and could show that it is optimal to select the project with highest fair charge.

One more proof of Gittins theorem was given in Bertsimas & Niño-Mora (1996), where they used a mathematical programming formulation of the problem and showed by the duality theorem that the Gittins index policy is optimal. This approach turned to be very useful for analyzing and solving more general problems, such as *restless bandit problem*, in which rested projects are allowed to change state, discussed in the next section.

Extensions and Applications

The *tax problem* is sort of reverse of the bandit problem, where the project one works on yields zero rewards and all the remaining projects are charged a (state-dependent) tax, other things being equal. With an appropriate modification of the Gittins index, Varaiya et al. (1985) showed that this problem is equivalent to the multi-armed bandit problem and thus, it is solved by an index policy. They also considered a situation where new projects are being made available showing that the theory here applies as well, as was earlier showed by Whittle (1981) for the multi-armed bandit problem.

Consider a generalization of the multi-armed bandit problem (and of the tax problem as well), called shortly the *non-zero version*, where the projects yield a reward regardless they are worked on or not (though the state can change only in the project being worked on). As a consequence of a linear programming formulation presented in Niño-Mora (2001) for the restless bandit problem, we have that the non-zero version can be transformed to a multi-armed bandit problem (i.e., with zero rewards from projects when rested) as follows. Denote in a non-zero version by $\hat{r}_k^1(x)$ the immediate *active reward*, which depends on a project k's current state x and is received if project k is selected to be worked on, and by $\hat{r}_k^0(x)$ its *passive reward* counterpart for having the project rested. Then, the

optimal solution to the multi-armed bandit problem, whose rewards are

$$r_k(x) = \hat{r}_k^1(x) - \frac{1}{1-\beta}\left[\hat{r}_k^0(x) - \beta \sum_{y \in \mathcal{X}_k} p_k(x,y)\hat{r}_k^0(y)\right],$$

is the optimal solution to the non-zero version, and the objective value of the non-zero version is obtained by summing up the objective value of the multi-armed bandit problem and

$$\frac{1}{1-\beta} \sum_{k \in \mathcal{K}} \sum_{x \in \mathcal{X}_k} \alpha(x) \cdot \hat{r}_k^0(x),$$

where $\alpha(x_k)$ is the 0/1 indicator for project k being initially in state x_k.

Another stream of extensions arises due to the fact that it is not always realistic to assume that the dynamics is of Markovian fashion or that the actual states of projects can be fully observed in the time epoch the decision must be taken. In some cases, the only thing that changes is that the calculation of Gittins index becomes cumbersome, intractable or impossible. However, it is not a general rule that the theory of optimality of an index policy still applies. For multi-armed bandit problem modeled as a Partially Observed MDP, see, for example, an application to the multi-target tracking (Krishnamurthy & Evans 2001) or a discussion on the optimality of greedy shooting strategy under incomplete information (Manor & Kress 1997).

A set of interesting applications emerges when one realizes that multi-armed bandit problem can be used for the study of optimal information acquisition and learning by economic agents. Those models include *Pandora's Box*, for which Weitzman (1979) showed that optimal index strategies exist. As Sundaram (2003) writes, it is common in economic theory to assume that firms and managers act under perfect information when choosing price and output decisions. One of the first papers to move away from this assumption was done by introducing the bandit framework to economic theory. Later on, applications to market pricing, job matching, technology choice and learning-by-doing, and agency theory were introduced. One of the latest intentions is an application to R&D project management (Denardo et al. 2004).

A natural extension of the model, motivated by practice, is to include *costs of switching* between projects. Indeed, in reality it is not harmless to stop working on a project and to start to work on another one. Unfortunately, Banks & Sundaram

(1994) showed that it is not possible, in the presence of switching costs, to define an index on the projects such that the resulting strategy is invariably optimal (Sundaram 2003). Similarly, non-optimality (in general) of the Gittins index in problems with *switching delays* was presented by Asawa & Teneketzis (1996). A nice survey on bandit problems with switching costs has been published by Jun (2004). He discusses applications of such a setting in job search and labor mobility, government policy, optimal search, experiments and learning, and game theory.

We wish to note that in the field of decision making, classical bandit problems are still under a significant interest. In the recent years, an analysis of a risk-aversion in the bandit problem arised. Chancelier et al. (2005) studied the optimal strategies in a one-armed bandit problem for a risk-averse controller. Loch & Kavadias (2002) incorporate risk aversion in their portfolio selection model, which is a version of two-armed bandit problem with delayed freezing of the passive projects. A tractable model of bounded rationality, based on the multi-armed bandit problem, was proposed by Bolton & Faure-Grimaud (2005). Further, Gabaix et al. (2003) remarked that the assumption of perfectly rational agents is spotted by intractability of many real decision making problems and alternative heuristical solutions should be analyzed.

Some literature makes reference to practical examples of the bandit problem in clinical trials, where different treatments need to be experimented with while minimizing patient losses, or in adaptive routing efforts for minimizing delays in a network. The questions that arise in all these cases are related to the problem of balancing reward maximization based on the knowledge already acquired and attempting new actions to further increase knowledge.

5. Restless Bandit Problem

The restless bandit problem is a natural generalization of the multi-armed bandit problem, which is capable to cover considerably broader set of practical situations. To the classical model we add just two simply-stated features: (1) the projects are allowed to evolve and yield rewards while rested (no freezing anymore), and (2) we are to allocate the scarce resource parallely to a fixed number of projects (instead of working on only one project). Nevertheless, such a modification significantly increases the problem's complexity and little from the Gittins approach remains operative here. Indeed, the increased modeling power comes at the expense of tractability: the restless bandit problem is *P-SPACE hard*, even in the deterministic case (Papadimitriou & Tsitsiklis 1999). The research focus must thus shift to the design of well-grounded, tractable heuristic policies.

In order to set up the restless bandit problem, we will build on the framework and notation from the previous section. We will also find it useful to assign a project the name *t-active* and *t-passive* if, in time epoch t, the project is decided to be worked on and to be rested, respectively. As a convention, we denote by 1 the action *to work* and by 0 the action *to rest*. Notice that we now need to have transition matrices P_k^1 and P_k^0, and rewards $r_k^1(x)$ and $r_k^0(x)$ for project k being active and passive, respectively.

Since one can easily get lost in the complicated notation this modeling framework requires, we remark the following notation norms: sets are written in calligraphic font (such as $\mathcal{K}, \mathcal{X}, \mathcal{T}$), with their cardinalities being denoted by the corresponding capital letters (e.g., K, X) similarly to other fixed constants (such as M); subscripts are reserved to the project labels, and superscripts to the actions (that can be 0 or 1). An MDP formulation of the restless bandit problem follows.

There are K projects, labeled by $k \in \mathcal{K}$. Let $x_k(t) \in \mathcal{X}_k$, for a finite state space $\mathcal{X}_k$, be the state of project k at time epoch $t \in \mathcal{T} = \{0, 1, 2, \dots\}$. At each time epoch t the controller must decide about allocation of a scarce resource, which we will call *work*, to M of the projects ($1 \leq M \leq K$ is an integer). If project k is selected to be worked on, an immediate reward $r_k^1(x_k(t))$ is received and the project changes to state $x_k(t+1)$ according to a stationary Markov transition probability matrix $P_k^1 = \{p_k^1(i,j)\}$. If project k is rested, an immediate reward $r_k^0(x_k(t))$ is received and the project changes to state $x_k(t+1)$ according to a

stationary Markov transition probability matrix $P_k^0 = \{p_k^0(i,j)\}$.

The controller's objective is to maximize the *expected total discounted reward* given by

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t r(t)\right] \tag{4}$$

where $0 < \beta < 1$ is a discount factor and $r(t)$ is the reward earned in time epoch t, defined as the sum of the rewards earned from all t-active projects $\mathcal{K}^1(t)$ and the rewards earned from all t-passive projects $\mathcal{K}^0(t)$,

$$r(t) = \sum_{k\in\mathcal{K}^1(t)} r_k^1(x_k(t)) + \sum_{k\in\mathcal{K}^0(t)} r_k^0(x_k(t)). \tag{5}$$

The optimization is done over a set of all *admissible policies* $\pi \in \Pi$, which are those that in each time epoch t select M projects to work on, based only on the actual state vector $x(t) = (x_k(t))_{k\in\mathcal{K}}$. That is, we require $K^1(t) = M$ and $K^0(t) = K - M$.

Whittle (1988) was the first who came out with a possible approach to treat the restless bandit problem, although he primarily considered undiscounted case of the problem (with a time-average reward criterion) in continuous time setting. He described a dynamic programming formulation of a relaxation, where the infinite number of constraints of having *exactly* M active projects at each time epoch is replaced by one constraint of having M active projects *on average* (or, more precisely, *in expectation*). Notice that the original constraint $K^1(t) = M$, which must hold for each time epoch t, can be taken in expectation and without any loss discounted, so that by summing up we can arrive to a relaxed constraint

$$\sum_{t=0}^{\infty}\beta^t\mathbb{E}^{\pi}\left[K^1(t)\right] = \sum_{t=0}^{\infty}\beta^t M = \frac{M}{1-\beta}. \tag{6}$$

In order to develop the crucial step, we will introduce the following notation: $a_k(t) \in \mathcal{A} = \{0,1\}$ is the action employed on project k at time epoch t. That is, $a_k(t) = 1$ if project k is t-active (i.e. $k \in \mathcal{K}^1(t)$), and $a_k(t) = 0$ otherwise. Note that $a_k(t)$ depends on a particular policy π applied to the system. Furthermore, let $r_k(t)$ be the reward earned from project k at time t, i.e., formally

$$r_k(t) = \begin{cases} r_k^1(x_k(t)), & \text{if project } k \text{ is } t\text{-active, i.e. } a_k(t) = 1, \\ r_k^0(x_k(t)), & \text{if project } k \text{ is } t\text{-passive, i.e. } a_k(t) = 0. \end{cases} \tag{7}$$

Using the new notation, we can easily rewrite (5) as

$$r(t) = \sum_{k \in \mathcal{K}} r_k(t) \tag{8}$$

and notice also that

$$K^1(t) = \sum_{k \in \mathcal{K}} a_k(t). \tag{9}$$

By plugging (8) into (4), plugging (9) into (6), and using the interchange property of the expectation, we obtain the following formulation of the original problem's relaxation (*Whittle's relaxation*):

$$\begin{aligned} &\max_{\pi \in \Pi} \sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t r_k(t) \right] \\ &\text{subject to} \quad \sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t a_k(t) \right] = \frac{M}{1-\beta} \end{aligned} \tag{10}$$

Whittle (1988) proposed to solve this problem by the classical Lagrangian method. Let ν be a Lagrangian multiplier, then the Lagrangian of (10) is

$$L(\pi, \nu) = \sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t r_k(t) \right] - \nu \left(\sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t a_k(t) \right] - \frac{M}{1-\beta} \right)$$

which can be rewritten as

$$L(\pi, \nu) = \sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t \left(r_k(t) - \nu a_k(t) \right) \right] + \nu \frac{M}{1-\beta} \tag{11}$$

Therefore, the Whittle's relaxation of the restless bandit problem can be solved by maximizing

$$\sum_{k \in \mathcal{K}} \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \beta^t \left(r_k(t) - \nu a_k(t) \right) \right], \tag{12}$$

i.e., by incorporation of a *wage* parameter ν into the problem. The wage ν must be paid for each active project in each period. Notice that the expression (12) defines a restless bandit problem without the condition on the number of active projects, but instead with an obligation to pay wage ν every time the scarce resource is

used. Furthermore, the solution is independent on the parameter M, which comes into play only when calculating the value of the original objective function in (10).

Suppose that in each period the controller is given a *budget*. If ν is the wage per period of working on a project, the budget of $M\nu$ allows to work parallely on M projects. The requirement of selecting M projects in each time epoch thus can be equivalently stated as the requirement of spending the (full) budget $M\nu$ in each period. The Whittle's relaxation is nothing but an extension where borrowing and lending over time is allowed (with a discount factor β, which can be interpreted as a factor of the risk that the whole project system collapses (Loch & Kavadias 2002)). Indeed, the total discounted sum of all budgets is

$$M\nu\left(1+\beta+\beta^2+\dots\right)=\nu\frac{M}{1-\beta}, \tag{13}$$

which is precisely the constant term to be summed up to (12) in order to obtain the objective value of the budget-less problem (10).

Whittle (1988) made a slightly different yet equivalent analysis of the problem so that he arrived to the notion of *subsidy for passivity*, which in his framework played an opposite role to our wage ν. Nevertheless, Whittle defined an index of project k when in state x, denoted $\nu_k(x)$, as the value of ν which makes the two actions for the project *in isolation* equally attractive, i.e. the best one can expect to earn if working on the project,

$$\max_{\pi\in\Pi}\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t\left(r_k(t)-\nu a_k(t)\right)\;\middle|\; x_k(0)=x \text{ and } a_k(0)=1\right], \tag{14}$$

is the same as the best one can expect to earn if letting the project rest,

$$\max_{\pi\in\Pi}\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t\left(r_k(t)-\nu a_k(t)\right)\;\middle|\; x_k(0)=x \text{ and } a_k(0)=0\right]. \tag{15}$$

We will refer to $\nu_k(x)$ as the *Whittle's index*. The Whittle's index reduces to the Gittins index when the passive projects are frozen (i.e. for the multi-armed bandit problem as well as for its non-zero version). Finally, he introduced the *indexability* property of a project, which is needed for Whittle's index to be meaningful and exist, as it is natural to expect that Whittle's indices induce a consistent ordering of the projects. A project is said to be *indexable* for a given discount factor β if the set of states where the active action is optimal increases monotonically

from the empty set to the full set of states as the wage parameter ν decreases from $+\infty$ to $-\infty$.

It follows directly that, for an indexable project k, there exist Whittle's indices for each state, such that an optimal policy for the project-k subproblem of (12) can be given as follows: "take the active action in states x with $\nu_k(x) > \nu$, and the passive action otherwise." Niño-Mora (2001) showed that for any wage value $\nu^* \neq 0$, the indexability of all projects implies that the optimal policy for the Whittle's relaxation is obtained by applying independently to each project the single-project policy just described. The projects from the multi-armed bandit problem (and its non-zero version) are indexable; hence, the Whittle's index policy is optimal in that model.

Whittle (1988) proposed to use as a heuristic for the restless bandit problem the following rule: "work on the M projects with largest Whittle's indices." However, he did not prove that this policy is optimal (and it is not, in general). Weber & Weiss (1990) late showed that this policy exhibits a form of asymptotic optimality under certain conditions.

An important step ahead for the application possibilities was an employment of the achievable region approach to the restless bandit problem. First, Bertsimas & Niño-Mora (2000) proposed a set of K increasingly improving bounds based on K increasingly stronger linear programming relaxations, the last of which is exact. They realized that the Whittle's relaxation (10) can be reformulated in the achievable region framework, where it is enough to focus on stationary policies. This reduction is not restrictive, since it is known from MDP theory that there exists an optimal policy, which is stationary. Notice that, for the one-project subproblem of (12), each stationary policy $\pi \in \Pi$ can be equivalently characterized by a set $\mathcal{S}_k \subseteq X_k$ of states in which the policy π prescribes to be active.

Bertsimas & Niño-Mora (2000) further described the Whittle's relaxation as a polynomial-size linear program, where the number of variables is $2X$ (twice the number of all the projects' states), which is solvable in polynomial time by LP interior point algorithms (Niño-Mora 2001). Furthermore, they proposed a way how to create other $K-1$ increasingly stronger relaxations, with a cost of increased number of variables, the last of which is exact. They also developed an heuristic index policy, alternative to the Whittle's, which is always well defined (i.e., does not require indexability of the projects).

Niño-Mora (2001) introduced the concept of $\mathcal{F}$-*indexability*, building on partial conservation laws, which extend the generalized conservation laws (Bertsimas & Niño-Mora 1996). It is known, that if a project satisfies the generalized conservation laws (GCL), it is indexable. That is, GCL provide a sufficient condition for indexability. However, it turns out, that for the restless bandit projects GCL are too narrow; i.e. restless bandit projects often do not satisfy GCL. In order to analyze the restless bandit problem, he defined the *partial conservation laws* relative to a family of state subsets $\mathcal{F} \subseteq 2^{\mathcal{X}_k}$ ($\mathcal{F}$-PCL, or simply PCL). In the case when $\mathcal{F} = 2^{\mathcal{X}_k}$, the PCL are precisely the same as the GCL.

One can understand the family $\mathcal{F}$ as a set of stationary policies with special structure. Thus, we are looking for an optimal stationary policy, given the restriction that the policy (described by an active-set $\mathcal{S}$) belongs to $\mathcal{F}$. Many times, such an approach may lead to a tractable framework for solving a special class of restless bandit problems. However, the limitation of the PCL approach is that it establishes the optimality of index policies under only *some* linear objectives functions (that is, only for some reward vectors $\mathcal{R}(\mathcal{F})$). Another complication is that one must "guess" the family $\mathcal{F}$ which includes the overaly optimal policy and makes the solution tractable. On the other hand, the power of this approach is that $\mathcal{F}$-indexability of a restless bandit project implies (Whittle's) indexability under the whole range of *admissible rewards* $\mathcal{R}(\mathcal{F})$, hence the projects can be analyzed in isolation.

Work-Reward Restless Bandit Problem

Now we will slightly modify the original setting so that we arrive to a more general formulation of the restless bandit problem, to which PCL-approach still applies (Niño-Mora 2002).

There are K projects, labeled by $k \in \mathcal{K}$. Let $x_k(t) \in \mathcal{X}_k$, for a finite state space $\mathcal{X}_k$, be the state of project k at time epoch $t \in \mathcal{T} = \{0, 1, 2, \dots\}$. At each time epoch t the controller must decide about allocation of M units of a scarce resource, which we will call *work* ($M > 0$ is a real number). If project k is selected to be active, a nonnegative *immediate work* $w_k^1(x_k(t))$ is spent, an immediate reward $r_k^1(x_k(t))$ is received and the project changes to state $x_k(t+1)$ according to a stationary Markov transition probability matrix $P_k^1 = \{p_k^1(i,j)\}$. If project k is selected to be passive, an immediate reward $r_k^0(x_k(t))$ is received and the project changes to state $x_k(t+1)$ according to a stationary Markov transition probability

matrix $P_k^0 = \{p_k^0(i,j)\}$. For convenience, we denote by $w_k^0(x) = 0$ for all $x \in \mathcal{X}_k$ the immediate work spent under the passive action. To allow the problem to have a solution, M cannot be greater than the sum of all immediate works needed, i.e.,

$$M \leq \sum_{k \in \mathcal{K}} \max_{x \in \mathcal{X}_k} w_k^1(x). \tag{16}$$

The controller's objective is to maximize the *expected total discounted reward* given by

$$\mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \beta^t r(t) \right] \tag{17}$$

where $0 < \beta < 1$ is a discount factor and $r(t)$ is the reward earned in time epoch t, defined as before, cf. (5). The optimization is done over a set of all *admissible policies* $\pi \in \Pi$, which are those that in each time epoch t allocate M units of work, based only on the actual state vector $x(t) = (x_k(t))_{k \in \mathcal{K}}$. That is, we require

$$\sum_{k \in \mathcal{K}^1(t)} w_k^1(x_k(t)) = M \quad \text{ at each time } t. \tag{18}$$

Notice that the problem as just described is quite restricted and may not always have a feasible solution. Indeed, the restriction (18) on work utilization implies that it must be $w_k^1(x) = w_k^1(y)$ for all $x, y \in \mathcal{X}_k$. Whittle would make the following relaxation of the problem: replace the infinite number of work utilization constraints at each time epoch (18) by one constraint of using work of M units on average (or rather, in expectation). Such a constraint would be (analogously to the Whittle's relaxed constraint),

$$\sum_{t=0}^{\infty} \beta^t \mathbb{E}^\pi \left[\sum_{k \in \mathcal{K}^1(t)} w_k^1(x_k(t)) \right] = \sum_{t=0}^{\infty} \beta^t M,$$

or,

$$\sum_{t=0}^{\infty} \beta^t \mathbb{E}^\pi \left[\sum_{k \in \mathcal{K}} w_k(t) \right] = \frac{M}{1-\beta}, \tag{19}$$

where $w_k(t)$ is the immediate work spent on project k at time epoch t (which, clearly, depends on the action employed).

Such a relaxation does not limit the values of $w_k^1(x)$ and moreover, it may be solved in the same way as the Whittle's relaxation in the case of the classical

restless bandit problem. Indeed, now we can express the problem (17) with the relaxed restriction (19) as (*work-reward relaxation*)

$$\begin{aligned} &\max_{\pi\in\Pi} \sum_{k\in\mathcal{K}} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t r_k(t)\right] \\ \text{subject to} \quad &\sum_{k\in\mathcal{K}} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t w_k(t)\right] = \frac{M}{1-\beta} \end{aligned} \tag{20}$$

Following the Whittle's ideas, problem (20) can be solved, using the Lagrangian method, by maximizing

$$\sum_{k\in\mathcal{K}} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t \left(r_k(t)-\nu w_k(t)\right)\right], \tag{21}$$

where the *wage* parameter ν must be interpreted as the wage per unit of immediate work. Notice that, as before, the solution is independent on the parameter M. The budget per period interpretation of $M\nu$ remains.

In what follows, we will focus on a project k in isolation and we drop the project label. An analogy to the Whittle's index for state x would be the value of ν which makes the two actions for the project equally attractive, i.e. the best one can expect to earn if working on the project,

$$\max_{\pi\in\Pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t \left(r(t)-\nu w(t)\right) \mid x(0)=x \text{ and } a(0)=1\right], \tag{22}$$

is the same as the best one can expect to earn if letting the project rest,

$$\max_{\pi\in\Pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\beta^t \left(r(t)-\nu w(t)\right) \mid x(0)=x \text{ and } a(0)=0\right], \tag{23}$$

where $a(t)$ denotes the action applied on the project in time epoch t.

However, this definition may not be valid for all states. In particular, if for a state x it is $w^1(x)=w^0(x)$, $r^1(x)=r^0(x)$, and $p^1(x,y)=p^0(x,y)$ for all $y\in\mathcal{X}$, then expressions (22) and (23) are equal for all ν. Following Niño-Mora (2002), we will call such states *uncontrollable*, and all the remaining states, for which the analogy to the Whittle's index exists, *controllable*. We denote the set of all controllable states by $\mathcal{C}\subseteq\mathcal{X}$, assuming that $C\geq 1$.

We restrict our attention to stationary admissible policies, among which MDP theory assures an overaly optimal admissible policy to exist. We can characterize each stationary policy π by an active-set $\mathcal{S} \subseteq \mathcal{C}$ of controllable states in which the policy π prescribes to be active, denoting by $2^{\mathcal{C}}$ the set of all those sets $\mathcal{S}$. Project indexability is defined as by Whittle, just narrowed to controllable states. More generally, given a family of active-sets $\mathcal{F} \subseteq 2^{\mathcal{C}}$, a project is said to be *$\mathcal{F}$-indexable* for a given discount factor β if the minimal set of states where the active action is optimal belongs to $\mathcal{F}$ and increases monotonically from the empty set to the full set of states as the wage parameter ν decreases from $+\infty$ to $-\infty$. By the theory of PCL (Niño-Mora 2001), the family $\mathcal{F}$ must be nonempty and satisfy the following two properties:

(i) $\mathcal{F}$ is augmentable, i.e. for each set $\mathcal{S} \in \mathcal{F}$ such that $\mathcal{S} \neq \mathcal{C}$ there exists a state $x \in \mathcal{C} \setminus \mathcal{S}$ for which $\mathcal{S} \cup \{x\} \in \mathcal{F}$,

(ii) $\mathcal{F}$ is accessible, i.e. for each set $\mathcal{S} \in \mathcal{F}$ such that $\mathcal{S} \neq \emptyset$ there exists a state $x \in \mathcal{S}$ for which $\mathcal{S} \setminus \{x\} \in \mathcal{F}$.

Under indexability, to each controllable state x is attached a *marginal productivity index* $\nu(x)$ (MPI) such that the policy for one-project subproblem of (20) "take the active action in controllable states x with $\nu(x) > \nu$, and the passive action otherwise (including noncontrollable states)" is optimal. It is clear that any policy that differs only in the actions for uncontrollable states will be optimal as well.

The rest of this section aims to show how to calculate the indices $\nu(x)$, building on a family of tractable subproblems. Suppose that project is initially in state $x \in \mathcal{X}$ and consider the following *ν-wage problem*, the state-x subproblem of (21):

$$\max_{\mathcal{S} \in \mathcal{F}} \mathbb{E}_x^{\mathcal{S}} \left[\sum_{t=0}^{\infty} \beta^t r(t) \right] - \nu \mathbb{E}_x^{\mathcal{S}} \left[\sum_{t=0}^{\infty} \beta^t w(t) \right], \tag{24}$$

where in the expectation we assume the first-period state is x, or, schematically,

$$\max_{\mathcal{S} \in \mathcal{F}} f_x^{\mathcal{S}} - \nu g_x^{\mathcal{S}}. \tag{25}$$

From (25) and MDP theory it follows immediately that to any wage ν there corresponds a *minimal optimal active-set* $\mathcal{S}^*(\nu) \subseteq \mathcal{C}$ such that

$$\mathcal{S}^*(\nu) = \{x \in \mathcal{C} : \nu^*(x) > \nu\} \in \mathcal{F}, \text{ for all } \nu.$$

Niño-Mora (2005) showed that $\mathcal{F}$-indexable projects are those, which obey the economics *law of diminishing marginal returns to work* consistently with $\mathcal{F}$. Namely, if one considers the *achievable work-reward region* spanned by points $(g_x^{\mathcal{S}}, f_x^{\mathcal{S}})$ for all $S \in F$, it is a convex region, whose upper boundary is a piecewise linear (and concave) function, where the states' MPIs are the function slopes.

We call $f_x^{\mathcal{S}}$ the *β-discounted $(x, \mathcal{S})$-reward measure* and $g_x^{\mathcal{S}}$ the *β-discounted $(x, \mathcal{S})$-work measure* (or, simply, the reward and work measure, respectively). Notice that if we denote by $a^1(t,x)$ and $a^0(t,x)$ the following indicators,

$$a^1(t,x) = \begin{cases} 1, & \text{if the project is at time } t \text{ in state } x \in \mathcal{S}, \\ 0, & \text{else,} \end{cases}$$

$$a^0(t,x) = \begin{cases} 1, & \text{if the project is at time } t \text{ in state } x \in \mathcal{X} \setminus \mathcal{S}, \\ 0, & \text{else,} \end{cases}$$

then $r(t)$, which is a $\mathcal{S}$-dependent term, can readily be expressed as

$$r(t) = \sum_{x \in \mathcal{X}} r^1(x) \cdot a^1(t,x) + \sum_{x \in \mathcal{X}} r^0(x) \cdot a^0(t,x), \tag{26}$$

and, similarly, $\mathcal{S}$-dependent $w(t)$ is

$$w(t) = \sum_{x \in \mathcal{X}} w^1(x) \cdot a^1(t,x) + \sum_{x \in \mathcal{X}} w^0(x) \cdot a^0(t,x) = \sum_{x \in \mathcal{X}} w^1(x) \cdot a^1(t,x). \tag{27}$$

The indicators $a^1(t,x)$ and $a^0(t,x)$ are the decision variables of the problem (25), as they are the only policy-dependent terms there. Hence, we remark that the measures may be viewed as $f_x^{\mathcal{S}} = f_x^{\mathcal{S}}(a^1(t,x), a^0(t,x))$ and $g_x^{\mathcal{S}} = g_x^{\mathcal{S}}(a^1(t,x))$, i.e. the reward measure is a function of all decision variables, and the work measure is a function of all the decision variables related to the active action. A more general setting, where the decision variables were nonnegative real numbers for which $a^1(t,x) + a^0(t,x) = 1$ (probabilities) instead of 1/0 indicators, was analyzed for semi-Markov projects in Niño-Mora (2005).

Using the achievable region approach, Niño-Mora (2002) introduced a sufficient condition for $\mathcal{F}$-indexability, which significantly helps in many, otherwise intractable, practical problems. In order to present this condition, we introduce a new terminology. The policy, whose active-set is $\mathcal{S}$, will be called $\mathcal{S}$-policy. Let

$\langle a, \mathcal{S} \rangle$ be the policy which takes action $a \in \mathcal{A}$ in the current time epoch and adopts the $\mathcal{S}$-policy thereafter.

For any state $x \in \mathcal{X}$ and a feasible $\mathcal{S}$-policy (i.e. $\mathcal{S} \in \mathcal{F}$), the $(x, \mathcal{S})$-*marginal reward measure* is defined as

$$\varrho_x^{\mathcal{S}} = f_x^{\langle 1, \mathcal{S} \rangle} - f_x^{\langle 0, \mathcal{S} \rangle}, \tag{28}$$

and the $(x, \mathcal{S})$-*marginal work measure* as

$$\omega_x^{\mathcal{S}} = g_x^{\langle 1, \mathcal{S} \rangle} - g_x^{\langle 0, \mathcal{S} \rangle}. \tag{29}$$

Thus, these $(x, \mathcal{S})$-marginal reward and work measures capture the increase in the respective $(x, \mathcal{S})$-reward and $(x, \mathcal{S})$-work measures, which results from being active instead of passive in the first time epoch and following the $\mathcal{S}$-policy afterwards. Notice that for uncontrollable states, it is $\varrho_x^{\mathcal{S}} = \omega_x^{\mathcal{S}} = 0$ for all $\mathcal{S}$, since applying both actions in such a state has precisely the same effect.

In the light of applications, it seems natural that $\omega_x^{\mathcal{S}}$ should be positive in all controllable states whenever $\mathcal{S} \in \mathcal{F}$.[1] Under this assumption, we can define for a controllable state x and a feasible active-set $\mathcal{S} \in \mathcal{F}$ the $(x, \mathcal{S})$-*marginal productivity rate* by

$$\nu_x^{\mathcal{S}} = \frac{\varrho_x^{\mathcal{S}}}{\omega_x^{\mathcal{S}}}. \tag{30}$$

These quantities are useful for the calculation of MPIs by an efficient *MPI($\mathcal{F}$) adaptive-greedy algorithm* introduced in Niño-Mora (2001). Given a family of policies $\mathcal{F}$, the algorithm checks whether the project states can be ordered as it is needed for $\mathcal{F}$-indexability. If affirmative (i.e. we say that the work-reward coefficients are $\mathcal{F}$-*admissible*), it outputs the marginal productivity indices $\nu(x)$ for all controllable states x.

In summary, if the two following conditions hold,

(i) $\omega_x^{\mathcal{S}} > 0$ in all controllable states $x \in \mathcal{C}$ and all feasible active-sets $\mathcal{S} \in \mathcal{F}$,

[1] (Subject to further investigation.) This condition is not necessary for problem to have an index-based solution. Another sufficient, still not necessary, condition might be that $\varrho_x^{\mathcal{S}}$ and $\omega_x^{\mathcal{S}}$ be nonzero and have the same sign for any $x \in \mathcal{C}, \mathcal{S} \in \mathcal{F}$. (Note that this would not necessarily imply the Whittle's indexability.)

(ii) work-reward coefficients are $\mathcal{F}$-*admissible*,

then the project when in state x, stated as problem (25), is $\mathcal{F}$-indexable with the optimal policy "take the active action in controllable states x with $\nu(x) > \nu$, and the passive action otherwise (including noncontrollable states)" (Niño-Mora 2001).

To conclude the section, suppose that there is an initial probability distribution α, where $\alpha(x) > 0$ for any state x is the probability that the project is initially in state x. By MDP theory, there is an optimal stationary policy of (25), which is independent on such initial distribution, which implies that the optimal policy described in the previous paragraph must be the policy independent on initial distribution. Thus, this policy is also optimal for the one-project subproblem of (21). Therefore, if all projects are $\mathcal{F}$-indexable (with, in general, project-dependent families $\mathcal{F}_k$) for all their states x, then the optimal policy for solving the work-reward relaxation (20) is: "work at time epoch t on all projects that are in a controllable state, whose MPI is greater than ν."

Further Extensions and Applications

The approach outlined in this section analyzed the restless bandit problem with the discounted criterion. If has been shown recently (Niño-Mora 2005c), that the multi-armed bandit problem with the expected total discounted reward over a *finite* horizon can be modeled as the restless bandit problem with the infinite discounted objective (4). Other criteria have also been considered in the literature on restless bandit problems. Whittle (1988) treated the restless bandit problem maximizing average reward over an infinite horizon; the approach was extended in the PCL framework by Niño-Mora (2001, 2002), in the latter paper applied to queueing admission control problem. Further, given the theoretical problems of the time-average criterion, Niño-Mora (2005b) considered also a bias-optimality criterion, when analyzing multiclass delay-sensitive queues. Finally, Niño-Mora (2005) introduced a new mixed average-bias criterion in the application of the LP approach to the optimal control of $M/G/1$ queues, where the approach was developed for countable state space projects and continuous time (semi-Markov projects).

Note that different forms of objective function imply, in general, distinct definitions of $f_x^{\mathcal{S}}$ and $g_x^{\mathcal{S}}$. Thus, every criterion yields a new MPI, some of which exist

and give good index policies in models where the Whittle index does not exist (e.g. mixed average-bias and bias-optimality criteria). Furthermore, the $\mathcal{F}$-indexability should be view as relative to given optimality criterion, i.e. relative to measures $f^{\mathcal{S}}$ and $g^{\mathcal{S}}$.

Many important practical situations can be modeled as a restless bandit problem, however, in many interesting cases, direct solution methods (e.g. dynamic programming) cannot be applied because of the combinatorial explosion of the number of variables. Hence, the problems are typically treated by simulation-based methods, which may provide nearly-optimal solutions. The achievable region approach may be an alternative to simulation methods, giving optimal or nearly-optimal solutions with significantly decreased computational complexity. Outside of the world of queueing models (which also includes an interesting class of broadcasting optimization problems), we have found a very small amount of literature, in which the restless bandit framework have yet been employed.

O'Meara & Patel (2001) proposed the restless bandit problem as a framework for modeling topic-specific indices in modern Web-search engines. They addressed questions of efficient query routing and automatic service management, e.g. How can each engine automatically select its own topic specialization for the benefit of all? Moreover, each engine must construct and maintain its own database, where the robot's quality to be maximized is given by a relevance scoring function. The topic-specific web robot problem can be decomposed into two separate decisions: what documents to request, and how many concurrent requests to make in order to fully utilize system throughput. Scheduling of the document requests must be done as quickly as possible. A simulation-based dynamic programming is used in order to characterize the optimal self-controlling mechanism, by developing a neuro-dynamic algorithm due to computational infeasibility of classical direct-solution methods of dynamic programming.

In the field of robotics, Faihe & Müller (1998) discussed limitations of the methods for robot behaviors coordination within the the neuro-dynamic framework and proposed to use restless bandits indices to prescribe the robot's behavior. They showed on a simple postman robot problem that the restless bandit method is effective and in general better than the former. Optimality of a greedy dispatch rule for cooperative control of multi-agent systems, arising in spacecraft constellations, was analyzed in Rao & Takamba (2005). Washburn, Schneider, & Fox (2002)

dealt with the problem of radar tracking of multiple agents, mentioned already by Whittle (1988) as one of the possible applications, developing approximate index solutions.

Regarding business applications, Loch & Kavadias (2002) used a variation of the restless bandit model with non-stationary passive probabilities (freezing after one period) to analyze the optimal budget allocation to new product development projects. They remarked that such R&D portfolio management problems are usually difficult to define because of the combinatorial complexity of project combinations. They found optimal index-like policies for several cases they analyzed (including manager's utility function). A similar approach was also applied to dynamic assortment for "Fast fashion" retailers (such as Zara, Mango), discussed in Caro & Gallien (2005).

6. Hypotheses and Future Investigation Outline

While the restless bandit problem as introduced by Whittle (1988) has been shown to be a powerful modeling paradigm in the field of queueing theory, where the work is indivisible, it seems that the work-reward restless bandit problem, or its relaxation, is especially well suited for a plenty of real-life situations arising in business and financial economics. Some items presented in the previous section are new, including the *budget* interpretation of $M\nu$ and the decomposition (26)-(27) with the notion that the work measure $g_x^{\mathcal{S}}$ depends on all and only the decision variables related to the active action.

Focus now on the work-reward restless bandit problem. As noted earlier, the restriction (18), saying that in each time epoch exactly M units of work must be used, is very restrictive. In many interesting applications, including financial ones, it is also allowed to spend less than M units of work, i.e., (18) would change to an inequality. Indeed, if $M\nu$ is the one-period budget to be allocated among the projects, it is possible in real-life that the controller spends less than this amount (and looses the budget not spent, or moves it to future periods). Then, if she is also able to borrow from the future budgets, we arrive to the work-reward relaxation. Furthermore, suppose that the one-period budget is not constant over time, but rather, the budgets are variable (but predetermined), denoted $B(0), B(1), B(2), \ldots$. Then, the total discounted budget is

$$\sum_{t\in\mathcal{T}} \beta^t B(t), \tag{31}$$

and the *average work expenditure* will, by equalizing (31) and (13), be

$$M = \frac{1-\beta}{\nu} \sum_{t\in\mathcal{T}} \beta^t B(t). \tag{32}$$

Thus, the variable budget version of the problem can be reformulated as the (fixed-budget) work-reward restless bandit problem. One can think of several possible heuristics for solving the problem of spending *at most* the budget given in each period. If fractional work investment is allowed (when one can spend a fraction of the immediate work needed resulting to a fractional immediate reward, i.e. the $w^1(x)$ should be called the maximal allowed immediate work), the optimal policy seems to be "work fully on the projects with the highest indices while the total work is not greater than the budgeted work $\frac{B(t)}{\nu}$, and spend the remaining work

on the passive project with the highest index". If fractional work investment is not permitted, the optimal policy would be given by a solution of the corresponding knapsack problem of all projects. In all the financial applications, only projects with positive expected reward should be considered.

Another relevant extension of the work-measure restless bandit problem is the one where each project has a *deadline*, when, depending on the project's state, a terminal reward is received and no more reward can be earned from that project after the deadline moment. Usually, the deadline moment is fixed a priori, so the controller's decision on the work allocation must be based on whether the project is in a favorable state or not. In such a setting, it can sometimes be useful to define a special absorbing state meaning *the project is ready*, i.e., no more work is needed and it is only waiting for the reward from the deadline moment (e.g. when a project is a production process). However, in other cases, (e.g. studying for various exams), not working on a project may cause a change to a less favorable state (because of forgetting). Note that when the deadline is not fixed from outside, one can think about an optimal stopping problem: Until what time is it worth to continue working on a project? When the controller decides to stop, she "sells the project" and gets the terminal reward. Such terminal reward stopping problem (for choosing a thesis advisor and buying a house) was discussed and the deadline extension proposed in Jacko (2005). Notice that the terminal reward model covers a set of important financial applications, including options and actions trading.

Consider again the work-reward restless bandit problem, where fractional work is allowed. In many budget allocation situations, there is a set of *prioritized* projects, or "must-be-worked" projects (given by a higher authority, such as the strategic business plan, legal requirements, survival-needed activities etc.). E.g., an individual must spend a part of her budget on food, because if not, his investment in the education would not yield the desired (or, expected) future rewards. It would be interesting to define the optimal policy for such a problem. One of the possible approaches may be to substitute the general discount factor β by a family of project dependent (or, even better, state-action dependent) one-period discount factors, which, as noted earlier, can be interpreted as factors of the risk that the whole project system collapses (more precisely, $1-\beta(x,a)$ would be the probability that the whole system collapses, if a project changes to state x and action a is applied there). Notice that such modification may allow many

problems, which are not $\mathcal{F}$-indexable at each state for a given discount factor β, to be indexable for state-dependent discount factors, because any restless bandit is indexable if the discount factor is small enough (Niño-Mora 2001).

In a more general case of the work-reward problem, one may consider that some positive immediate work $w_k^0(x)$ must also be spent under the passive action. It seems, that the approach may work in an analogous way whenever there is an equality restriction on $w_k^1(x) + w_k^0(x)$ for each state $x \in \mathcal{X}_k$. If there is no such restriction, one must treat the passive immediate work as a new scarce resource independent on the active immediate work, and, typically, choose to restrict just one of the two scarce resources. Thus, we naturally come to a multiple scarce resource restless bandit problem. In such a setting, there must be a utilization restriction for *each* scarce resource. It seems that to have a feasible solution to such a problem would require more than two actions — one active action for each scarce resource plus one passive action. This generalization would significantly expand the set of interesting applications by including all the dynamic allocation problems in which several "workers" work parallely. However, the concept of indexability is not trivially extendable to higher dimensions.

All the preceding discussion on optimal policies assumes $\mathcal{F}$-indexability of the projects. However, there are many interesting problems which are likely not to be indexable, so there is a strong theoretical need for a more general conditions. We propose two ideas (which are likely to be altered after the complete proof is ready) under which index policies, given by MPIs, will be optimal. But before doing that, we state a new sufficient condition for $\mathcal{F}$-indexability, which is more relaxed than the Niño-Mora's (2001) sufficient condition.

Suppose that the active immediate work $w^1(x) > 0$ for all controllable states x of a given project. If the project satisfies

(i) for every $\mathcal{S} \in \mathcal{F}$, there exists a controllable state $x \in \mathcal{C} \setminus \mathcal{S}$ such that $S \cup \{x\} \in \mathcal{F}$ and $\omega^{\mathcal{S}}(x) > 0$,

(ii) work-reward coefficients are $\mathcal{F}$-admissible (the algorithm should be modified so that in every step the selected state has $\omega^{\mathcal{S}}(x) > 0$),

then it is $\mathcal{F}$-indexable. We suspect that this condition is also necessary for $\mathcal{F}$-indexability, which would offer full and tractable characterization of indexable projects. It is subject to further investigation, whether relaxing the $w^1(x) > 0$

condition would alter our hypothesis.

Suppose $w^1(x) > 0$ for all controllable states x of a given project (in order to ensure that the active-set with the lowest expected work is the empty set). The project when in state x is said to be *weakly $\mathcal{F}$-indexable* for a discount factor β, if the minimal optimal active-set belongs to $\mathcal{F}$ and increases monotonically from the empty set to a set $\mathcal{H} \in \mathcal{F}$ as the wage parameter ν decreases from $+\infty$ to $-\infty$, where $\mathcal{H}$ is such that for all $\mathcal{S} \in \mathcal{F}$ it is $g_x^{\mathcal{H}} \geq g_x^{\mathcal{S}}$ (that is, $\mathcal{H}$ is an active-set with the highest expected work). The Whittle's indexability and Niño-Mora's $\mathcal{F}$-indexability require that $\mathcal{H} = \mathcal{C}$. For the weak $\mathcal{F}$-indexability, slightly stronger properties of $\mathcal{F}$ than under $\mathcal{F}$-indexability are required. The family $\mathcal{F}$ must be nonempty and satisfy that for any $\mathcal{S} \in \mathcal{F}$ such that $\mathcal{S} \neq \emptyset$, there exists a permutation $\pi = (\pi_1, \pi_2, \ldots, \pi_S)$ of the states in $\mathcal{S}$ such that for all $\mathcal{S}_i, i = 1, 2, \ldots, S$, having $\mathcal{S}_0 = \emptyset$, it is $\omega_{\pi_i}^{\mathcal{S}_{i-1}} \geq 0$. In words, $\mathcal{F}$ must be such that one can "arrive" from the empty set to any $\mathcal{S} \in \mathcal{F}$ by adding states with nonnegative marginal work.

Though the condition on $\mathcal{F}$ is stronger than in the case of $\mathcal{F}$-indexability, a sufficient condition for the weak $\mathcal{F}$-indexability simplifies to the *weak $\mathcal{F}$-admissibility* of the work-reward coefficients. This condition is defined on a modified version of the Niño-Mora's MPI($\mathcal{F}$) adaptive-greedy algorithm, which checks the ordering of states in $\mathcal{H}$ and ensures the weak $\mathcal{F}$-indexability. The algorithm is subject to further investigation.

The following idea deals with a class of problems, in which an index policy is optimal only for the wage parameter $\nu > \nu_{\min}$, which is relevant, because negative values of the wage parameter ν are usually not sensible. Suppose all the assumptions for weak $\mathcal{F}$-indexability hold. The project when in state x is said to be *partially $\mathcal{F}$-indexable* for a discount factor β, if the minimal optimal active-set belongs to $\mathcal{F}$ and increases monotonically from the empty set to a set $\mathcal{H} \in \mathcal{F}$ as the wage parameter ν decreases from $+\infty$ to $\nu_{\min}$. A sufficient condition for partial $\mathcal{F}$-indexability would be analogous to the one for weak indexability, with a different algorithmic test, which identifies the value $\nu_{\min}$.

Apart of the applications outlined in this section, which are the most direct ones of the work-reward restless bandit problem, one can consider the appropriate, more complex version of the whole range of models that have been analyzed in the multi-armed bandit problem framework. Moreover, it seems that nobody has

proposed a theory of the restless bandit problem in such an important extension as for partially observed projects and for the systems with delayed state observations.

References

Asawa, M. & Teneketzis, D. (1996): *Multi-Armed Bandits with Switching Penalties*, IEEE Transactions on Automatic Control 41 (3), pp. 328-348.

Banks, J. S. & Sundaram, R. K. (1994): *Switching Costs and the Gittins Index*, Econometrica 62 (3), pp. 687-694.

Bertsekas, D. P. (1995): *Dynamic Programming and Optimal Control: Volume Two*, Athena Scientific, Belmont, MA.

Bertsimas, D. (1995): *The Achievable Region Method in the Optimal Control of Queueing Systems; Formulations, Bounds and Policies*, Queueing Systems: Theory and Applications 21, pp. 337-389.

Bertsimas, D. & Niño-Mora, J. (1996): *Conservation Laws, Extended Polymatroids and Multiarmed Bandit Problems; a Polyhedral Approach to Indexable Systems*, Mathematics of Operations Research 21, pp. 257-306.

Bertsimas, D. & Niño-Mora, J. (2000): *Restless Bandits, Linear Programming Relaxations, and a Primal-Dual Index Heuristic*, Operations Research 48 (1), pp. 080-090.

Blackwell, D. (1965): *Discounted Dynamic Programming*, Ann. Math. Statist. 36, pp. 226-235.

Bolton, P. & Faure-Grimaud, A. (2005): *Thinking Ahead: The Decision Problem*, prescript, June.

Caro, F. & Gallien, J. (2005): *Dynamic Assortment with Demand Learning for Seasonal Consumer Goods*, prescript, January 10.

Chancelier, J., de Lara, M., & de Palma, A. (2005): *Risk Aversion and Optimal Strategies in a One-Armed Bandit Problem: An Application to Road Choice*, prescript, April 1.

Coffman, E. & Mitrani, I. (1980): *A Characterization of Waiting Time Performance Realizable by Single Server Queues*, Operations Research 28, pp. 810-821.

Dacre, M., Glazebrook, K. D. & Niño-Mora, J. (1999): *The Achievable Region Approach to the Optimal Control of Stochastic Systems. With Discussion*, Journal of Royal Statistical Society B 61 (4), pp. 747-791.

Davis, M. H. A. (1993): *Markov Models and Optimization*, Chapman & Hall, London.

Denardo, E. V., Rothblum, U. G., & Van der Heyden, L. (2004): *Index Policies for Stochastic Search in a Forest with an Application to R&D Project Management*, Mathematics of Operations Research 29 (1), pp. 162-181.

Faihe, Y. & Müller, J. (1998): *Behaviors Coordination Using Restless Bandits Allocation Indexes*, Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior (SAB98) on From Animals to Animats 5, pp. 159-164.

Federgruen, A. & Groenevelt, H. (1988): *Characterization and Optimization of Achievable Performance in General Queueing Systems*, Operations Research 36, pp. 733-741.

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2003): *The allocation of Attention: Theory and Evidence*, Working Paper 03-31, August 29.

Gittins, J. C. (1979): *Bandit Processes and Dynamic Allocation Indices*, Journal of the Royal Statistical Society, Series B 41 (2), pp. 148-177.

Gittins, J. C. & Jones, D. (1974): *A Dynamic Allocation Index for the Sequential Allocation of Experiments*, in Progress in Statistics, J. Gani et al. (Eds.), North Holland, Amsterdam.

Jacko, P. (2005): *A Model of Decision Making*, http://www.strom.sk/~pj/works/model.pdf.

Jun, T. (2004): *A Survey on the Bandit Problem with Switching Costs*, De Economist 152, pp. 1-29.

Klimov, G. P. (1974): *Time-Sharing Service Systems I*, Theory of Probability and its Applications 19 (3), pp. 532-551.

Krishnamurthy, V. & Evans, R. J. (2001): *Hidden Markov Model Multiarm Bandits: A Methodology for Beam Scheduling in Multitarget Tracking*, IEEE Transactions On Signal Processing 49 (12).

Loch, C. H. & Kavadias, S. (2002): *Dynamic Portfolio Selection of NPD Programs Using Marginal Returns*, Management Science 48 (10), pp. 1227-1241.

Manor, G. & Kress, M. (1997): *Optimality of the Greedy Shooting Strategy in the Presence of Incomplete Damage Information*, Naval Research Logistics 44, pp. 613622.

Niño-Mora, J. (2001): *Restless Bandits, Partial Conservation Laws and Indexability*, Advances in Applied Probability 33 (1), pp. 76-98.

Niño-Mora, J. (2002): *Dynamic Allocation Indices for Restless Projects and Queueing Admission Control: A Polyhedral Approach*, Mathematical Programming 93 (3), Ser. A, pp. 361-413.

Niño-Mora, J. (2005): *Restless Bandit Marginal Productivity Indices, Diminishing Returns and Optimal Control of Make-To-Order/Make-To-Stock* $M/G/1$ *queues*, Mathematics of Operations Research, forthcoming.

Niño-Mora, J. (2005b): *Marginal Productivity Index Policies for Scheduling a Multiclass Delay-/Loss-Sensitive Queue*, Working Paper 05-39, Statistics and Econometrics Series 06, June.

Niño-Mora, J. (2005c): *Marginal Productivity Index Policies for the Finite-Horizon Multiarmed Bandit Problem*, In Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference ECC 2005 (CDC-ECC 05), in press.

O'Meara, T. & Patel, A. (2001): *A Topic-Specific Web Robot Model Based on Restless Bandits*, IEEE Internet Computing 5 (2), pp. 27-35.

Papadimitriou, C. H. & Tsitsiklis, J. N (1999): *The Complexity of Optimal Queueing Network*, Mathematics of Operations Research 24 (2), pp. 293-305.

Rao, V. G. & Takamba, P. T. (2005): *Optimally Greedy Control of Team Dispatching Systems*, April 1, prescript.

Robbins, H (1952): *Some Aspects of the Sequential Design of Experiments*, In Bulletin of the American Mathematical Society 55, pp. 527-535.

Shanthikumar, J. G. & Yao, D. D. (1992): *Multiclass Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control*, Operations Research 40, pp. S293-S299.

Stidham, S. (2002): *Analysis, Design, and Control of Queueing Systems*, Operations Research 50 (1), pp. 197-216.

Sundaram, R. K. (2003): *Generalized Bandit Problems*, preprint, May 27.

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

Varaiya, P. P., Walrand J. C., & Buyukkoc C. (1985): *Extensions of the Multiarmed Bandit Problem: The Discounted Case*, IEEE Transactions on Automatic Control AC-30 (5), pp. 426-439.

Washburn, R. B., Schneider, M. K., & Fox, J. J. (2002): *Stochastic Dynamic Programming Based Approaches to Sensor Resource Management*, Proceedings of the Fifth International Conference on Information Fusion, Annapolis, MD.

Weber, R. R. & Weiss G. (1990): *On an Index Policy for Restless Bandits*, Journal of Applied Probability 27, pp. 637-648.

Weber, R. R. (1992): *On the Gittins Index for Multiarmed Bandits*, The Annals of Applied Probability 2 (4), pp. 1024-1033.

Whittle, P. (1980): *Multi-Armed Bandits and the Gittins Index*, Journal of Royal Statistical Society 42, pp. 143-149.

Whittle, P. (1981): *Arm Acquiring Bandits*, Ann. Probab. 9, pp. 284-292.

Whittle, P. (1988): *Restless Bandits: Activity Allocation in a Changing World*, in A Celebration of Applied Probability, J. Gani (Ed.), Journal of Applied Probability 25A, pp. 287-298.

Whittle, P. (2002): *Applied Probability in Great Britain*, Operations Research 50 (1), pp. 227-239.

Weitzman, M. L. (1979): *Optimal Search for the Best Alternative*, Econometrica 47, pp. 641-654.

Yao, D. D. & Zhang, L. (1997a): *Stochastic Scheduling via Polymatroid Optimization*, in Lectures in Applied Mathematics, George Yin and Qing Zhang (Eds.), AMS/SIAM.

Yao, D. D. & Zhang, L. (1997b): *Dynamic Scheduling of a Class of Stochastic Systems: Extended Polymatroid, Side Constraints, and Optimality*, in Proceedings of the 36th IEEE Conference on Decision and Control.