Optimal Index Rules for Single Resource Allocation to Stochastic Dynamic Competitors*

[Invited paper]

Peter Jacko
BCAM — Basque Center for Applied Mathematics
Bizkaia Technology Park, Building 500
48160 Derio (Bilbao), Spain
jacko@bcamath.org

ABSTRACT

In this paper we present a generic Markov decision process model of optimal single resource allocation to a collection of stochastic dynamic competitors. The main goal is to identify sufficient conditions under which this problem is optimally solved by an index rule. The main focus is on the frozen-if-not-allocated assumption, which is notoriously found in problems including the multi-armed bandit problem, tax problem, Klimov network, job sequencing, object search and detection. The problem is approached by a Lagrangian relaxation and decomposed into a collection of normalized parametric single-competitor subproblems, which are then optimally solved by the well-known Gittins index. We show that the problem is equivalent to solving a time sequence of its Lagrangian relaxations. We further show that our approach gives insights on sufficient conditions for optimality of index rules in restless problems (in which the frozen-if-not-allocated assumption is dropped) with single resource; this paper is the first to prove such conditions.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Markov processes; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*Sequencing and scheduling*; I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search—*Dynamic programming*

General Terms

Theory, Performance, Design

Keywords

optimality, index rules, Gittins index, Whittle index, Lagrangian relaxation, Markov decision processes, dynamic programming, resource allocation, restless bandits

*Research partially supported by grant MTM2010-17405 of the MICINN (Spain) and grant PI2010-2 of the Department of Education and Research (Basque Government). I am grateful to Urtzi Ayesta, Konstantin Avrachenkov, and Sofía S. Villar for many fruitful discussions and for the encouragement to write this paper.

ValueTools 2011 Paris, France

1. INTRODUCTION

In this paper we present a generic Markov decision process (MDP) model of optimal single resource allocation to a collection of stochastic dynamic competitors. This is a constrained MDP with special structure, which belongs to the family of *weakly-coupled* MDPs (Meuleau et al., 1998). We first discuss this problem under the frozen-if-not-allocated assumption, which is notoriously found in problems optimally solvable by an index rule, which include the multi-armed bandit problem, tax problem, Klimov network, job sequencing, object search and detection (see subsection 2.4). Under such an assumption, our model can be seen as a generalization of the multi-armed bandit problem by allowing non-zero rewards and/or costs for both played and not played arms of the bandit (slot) machine. Then we present new conditions sufficient for solving the problem optimally by an index rule if the frozen-if-not-allocated assumption is dropped, i.e., when competitors are *restless*.

An *index* is a function that assigns a value to each state of a given competitor. The index value can be a real number or $-\infty$ or $+\infty$, and in this paper we restrict our attention to indices that depend on the parameters of the given competitor alone. This is the most interesting case from the implementation point of view in that the computation of index values is usually less demanding, and the index values, for instance, do not depend on the number of competitors considered.

An *index rule* is a policy that decides at every moment to which competitor the resource is allocated based on the relative ordering of the competitors' current state index values. As a convention, in this paper we consider that the index rule allocates the resource to the competitor of greatest current index value; in case of a tie, it allocates the resource to the same competitor as in the previous period if possible, otherwise chooses arbitrarily. (Alternatively in case of a tie,

the resource could be allocated to the oldest non-allocated competitor in order to increase fairness, but at the expense of an increased frequency of switching.) An index rule is an *adaptive greedy rule*: it is adaptive because it depends on the index values that in turn depend on the actual competitors' states, and it is greedy, because it chooses the competitor of greatest index value (Jacko, 2009).

Although for the above-mentioned problems indices were obtained and optimality of the resulting index rule was proved in a variety of ways, we will focus in this paper on the Lagrangian approach proposed by Whittle (1988) for the restless bandit problem. In the restless bandit problem, M out of $K \geq M$ competitors must be chosen at every moment to be allocated M copies of a resource. Whittle (1988) proposed to use a *multi-index rule* of allocating the resources to the M competitors of greatest index values that appear as particular values of Lagrangian multiplier. However, the existence of such an index is not guaranteed, and so it must be established for every competitor. Moreover, as he noted,

> It may be too much to expect the index policy to be optimal in the restless case we have formulated.

He showed that under the frozen-if-not-allocated assumption and $M = 1$, the index exists and the Whittle index rule is equivalent to the Gittins index rule, and therefore optimal, and it is also trivially optimal if $M = K$. Whittle (1988) conjectured that under fairly general conditions the proposed multi-index rule is asymptotically optimal if both M and K grow to infinity converging to a fixed proportion. Up to the moment, the conjecture was proved valid only under a list of conditions that are rather restrictive and difficult to check in Weber and Weiss (1990).

Even if such an asymptotic optimality were true, it gives no guarantee about the performance of the Whittle index rule in problems with a single resource, which is often of interest in applications. The mean or median performance of the Whittle index rule in problems with a single resource is typically reported to be very close to optimal. Nevertheless, it is often the case that some (small) number of problem instances show an extremely bad performance of the policy (see, e.g., Niño-Mora, 2007b; Adelman and Mersereau, 2008). It therefore seems important and reasonable to study optimality of the Whittle index rule in single resource problems, which is the focus of this paper.

Section 2 presents and MDP formulation of the problem. In Section 3, we approach the problem of single resource allocation to stochastic dynamic competitors by describing its several step-wise relaxations, including the Whittle relaxation. That is further approached by the Lagrangian relaxation in order to decompose the problem into a collection of parametric single-competitor subproblems.

Under the frozen-if-not-allocated assumption, in Section 4 we normalize these subproblems into equivalent problems with zero rewards when not allocated, which are then optimally solved by the well-known Gittins index. We show that the original problem is equivalent to a time sequence of problems which are in fact its Lagrangian relaxations with particular values of the Lagrangian multiplier. The problems in the sequence are infinite-horizon, but *branching*, i.e. they differ by the initial time period and initial competitors' states that both depend on the evolution of the previous problem of the sequence. Interestingly, every problem in this sequence can be optimally solved in a straightforward manner. Furthermore, the Gittins index rule (which is optimal) can be recovered by simultaneously applying optimal policies to all these problems in the sequence. To provide an economic insight, the *prevailing charge* from Weber (1992) is the value of the Lagrangian multiplier in the corresponding Lagrangian relaxation in this sequence. Similar ideas appeared also in Whittle (1981), dealing with a multi-armed bandit problem with arrivals of new bandits. However, to the best of author's knowledge, no earlier work related that interpretation to Lagrangian relaxations as we do in this paper.

Section 5 deals with the general case (when the frozen-if-not-allocated assumption is dropped), where we give new sufficient conditions under which the Whittle index rule is optimal. These conditions are the main new results of this paper, but we believe that our main contribution is to survey and clarify a mathematical approach which recovers many known index rules (obtained by different methods), and is capable to give insights about optimality of index rules in more general settings, including the restless bandit problem.

2. MDP FORMULATION

In this section we present a discrete-time MDP formulation of the problem of resource allocation to stochastic dynamic competitors. We follow the framework introduced in Jacko (2009) restricted here to an undivisible resource with unit capacity.

Consider the time slotted into time epochs $t \in \mathcal{T} := \{0, 1, 2, \dots\}$ at which decisions can be made. Time epoch t corresponds to the beginning of time period t. We consider the problem over an infinite horizon. Suppose that there are $K \geq 1$ (integer) competitors, labeled by $k \in \mathcal{K}$, competing for a resource (decision-maker) that decides at every time epoch which competitor should it be allocated to during that period. The resource can and must be allocated to one competitor at a time.

2.1 Competitors

Since the capacity of the resource is one unit (undivisible), every competitor can be allocated either zero or one resource capacity units. We denote by $\mathcal{A} := \{0, 1\}$ the *action space*, i.e., the set of allowable levels of capacity allocation. This action space is the same for every competitor k.

Each competitor k is defined independently of other competitors as the tuple

$$\left(\mathcal{N}_k, (\boldsymbol{W}_k^a)_{a\in\mathcal{A}}, (\boldsymbol{R}_k^a)_{a\in\mathcal{A}}, (\boldsymbol{P}_k^a)_{a\in\mathcal{A}}\right),$$

where

- $\mathcal{N}_k$ is the *state space*, i.e., a finite set of possible states competitor k can occupy;
- $\boldsymbol{W}_k^a := \left(W_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $W_{k,n}^a$ is the expected one-period capacity consumption, or *work* required by competitor k at state n if action a is decided at the beginning of a period;
- $\boldsymbol{R}_k^a := \left(R_{k,n}^a\right)_{n\in\mathcal{N}_k}$, where $R_{k,n}^a$ is the expected one-period *reward* earned by competitor k at state n if action a is decided at the beginning of a period;
- $\boldsymbol{P}_k^a := \left(p_{k,n,m}^a\right)_{n,m\in\mathcal{N}_k}$ is the competitor-k stationary one-period *state-transition probability matrix* if action

a is decided at the beginning of a period, i.e., the (n,m)-element of the matrix, $p^a_{k,n,m}$, is the probability of moving to state m from state n under action a.

The dynamics of competitor k is thus captured by the *state process* $X_k(\cdot)$ and the *action process* $a_k(\cdot)$, which correspond to state $X_k(t) \in \mathcal{N}_k$ and action $a_k(t) \in \mathcal{A}$ at all time epochs $t \in \mathcal{T}$. As a result of deciding action $a_k(t)$ in state $X_k(t)$ at time epoch t, the competitor k consumes the allocated capacity, earns the reward, and evolves its state for the time epoch $t+1$. To avoid technical difficulties we will assume that $R^a_{k,n}$ is bounded.

2.2 A Unified Optimization Criterion

Before describing the problem we define an averaging operator that will allow us to discuss the infinite-horizon problem under the traditional myopic criterion, β-discounted criterion and time-average criterion in parallel. Let $\Pi_{X,a}$ be the set of all the policies that for each time epoch t decide (possibly *randomized*) action $a(t)$ based only on the state-process history $X(0), X(1), \ldots, X(t)$ and on the action-process history $a(0), a(1), \ldots, a(t-1)$ (i.e., *non-anticipative*). Let $\mathbb{E}^\pi_\tau$ denote the expectation over the state process $X(\cdot)$ and over the action process $a(\cdot)$, conditioned on the state-process history $X(0), X(1), \ldots, X(\tau)$ and on policy π.

Consider any expected one-period quantity $Q^{a(t)}_{X(t)}$ that depends on state $X(t)$ and on action $a(t)$ at any time epoch t. For any policy $\pi \in \Pi_{X,a}$, any initial time epoch $\tau \in \mathcal{T}$, and any *discount factor* $0 \le \beta \le 1$ we define the infinite-horizon *β-average quantity* as[1]

$$\mathbb{B}^\pi_\tau \left[Q^{a(\cdot)}_{X(\cdot)}, \beta, \infty \right] := \lim_{T\to\infty} \frac{\sum_{t=\tau}^{T-1} \beta^{t-\tau}\, \mathbb{E}^\pi_\tau \left[Q^{a(t)}_{X(t)} \right]}{\sum_{t=\tau}^{T-1} \beta^{t-\tau}}. \qquad (1)$$

The β-average quantity recovers the traditionally considered quantities in the following three cases:

- *expected time-average quantity* when $\beta = 1$.
- *expected total β-discounted quantity*, scaled by constant $1-\beta$, when $0 < \beta < 1$;
- *myopic quantity* when $\beta = 0$.

Thus, when $\beta = 1$, the problem is formulated under the *time-average criterion*, whereas when $0 < \beta < 1$ the problem is considered under the *β-discounted criterion*. The remaining case when $\beta = 0$ reduces to a static problem and hence is considered in order to define a *myopic policy*. In the following we consider the discount factor β to be fixed and the horizon to be infinite, therefore we omit them in the notation and write briefly $\mathbb{B}^\pi_\tau \left[Q^{a(\cdot)}_{X(\cdot)} \right]$.

2.3 Optimization Problem

Now we are ready to formulate the optimization problem. Let $\Pi_{\boldsymbol{X},\boldsymbol{a}}$ be the space of randomized and non-anticipative policies depending on the joint state-process $\boldsymbol{X}(\cdot) := (X_k(\cdot))_{k\in\mathcal{K}}$ and deciding the joint action-process $\boldsymbol{a}(\cdot) := (a_k(\cdot))_{k\in\mathcal{K}}$, i.e., $\Pi_{\boldsymbol{X},\boldsymbol{a}}$ is the *joint policy space*.

[1] For definiteness, we consider $\beta^0 = 1$ for $\beta = 0$.

For any discount factor β, the problem is to find a joint policy $\boldsymbol{\pi}$ maximizing the *objective* given by the β-average aggregate reward starting from the initial time epoch 0 subject to the family of *sample path* allocation constraints, i.e.,

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} R^{a_k(\cdot)}_{k,X_k(\cdot)} \right] \qquad (P)$$

$$\text{subject to } \mathbb{E}^{\boldsymbol{\pi}}_t \left[\sum_{k\in\mathcal{K}} a_k(t) \right] = 1, \text{ for all } t \in \mathcal{T}$$

Note that the constraint could equivalently be expressed as

$$\sum_{k\in\mathcal{K}} a_k(t) = 1$$

for all $t \in \mathcal{T}$ under policy π and for any possible joint state-process history $\boldsymbol{X}(0), \boldsymbol{X}(1), \ldots, \boldsymbol{X}(t)$.

2.4 Known Special Cases

Optimal index rules can be found in the literature for a variety of models satisfying both of the following two distinguishing features:

1. [*binary work*] $W^a_{k,n} := a$, i.e., the competitor consumes all the capacity allocated, and
2. [*frozen if not allocated*] $\boldsymbol{P}^0_k := \boldsymbol{I}$ (an identity matrix), i.e., if no capacity is allocated to the competitor ($a = 0$), then the competitor does not change its state ($p^0_{k,n,n} = 1$ for all n).

The following selected problems have these two features and can be cast as special cases of our model.

Job Sequencing (Cox and Smith, 1961).

Competitors are jobs and resource is a server that must decide an order in which to serve the waiting jobs. If the job sizes are geometrically distributed with means $1/\mu_k$ and the waiting costs are c_k per period, then we have states $\mathcal{N}_k := \{0,1\}$ representing that the job k is "completed" and "waiting", reward $R^0_{k,1} = -c_k, R^1_{k,1} = -c_k(1-\mu_k)$, and transition probabilities $p^1_{k,1,0} = \mu_k, p^1_{k,1,1} = 1-\mu_k$. State 0 is absorbing with no rewards. In the case of job sizes with general distribution, the state space must be enlarged to represent known information, such as the attained service or remaining service.

Multi-armed Bandit Problem (Robbins, 1952; Gittins and Jones, 1974).

Competitors are arms of a bandit machine and resource is a gambler who wants to choose an arm to pull in every time epoch. There are no rewards when an arm is not played, i.e., $R^0_{k,n} = 0$.

Tax Problem (Varaiya et al., 1985; Whittle, 2005).

Competitors are machines and only one of them can be operated at a time. Each idle machine incurs a waiting cost $c_{k,n}$ which depends on its current state. There is no reward when a machine is operated, i.e., $R^1_{k,n} = 0$, and a negative reward when it is idle, i.e., $R^0_{k,n} = -c_{k,n}$. Under the discounted criterion, the tax problem is equivalent to the multi-armed bandit problem.

Klimov Network (Klimov, 1974).
Competitors are jobs queued on machines (incurring a waiting cost) and resource is a system maintenance manager that wants to choose at every moment a machine to serve one of the waiting jobs in its queue. When the service of a job is completed, the job is routed to another machine according to a given probabilistic routing scheme (or possibly leave the system). State of a machine represents how many jobs are waiting: one of the states yields no reward, meaning that there are no jobs waiting.

Object Search and Detection (Bertsekas, 2001, Example 1.5.1).
Competitors are boxes or sites that may include objects which we want to find or detect, and resource is an imperfect sensor that can be focused on one of the boxes/sites at a time. A reward is received every time an object is found, but focusing the sensor is costly. The task is to decide at every moment where the sensor should be focused.

The binary work assumption is supposed to hold throughout the paper; the author is not aware of any model violating such a condition for which an index rule is optimal. The frozen-if-not-allocated assumption will be assumed in Section 4.

In Section 5 we will discuss Whittle index definition and optimality of index rules in general models, in which the frozen-if-not-allocated assumption is dropped. Dropping this assumption drastically expands modeling possibilities. For instance with respect to the above examples, it is possible to incorporate time-varying service rate for every job; out-of-control gamblers that can pull unused arms; new jobs arrivals; switching of jobs to another machine due to impatience even before completing the job; smart or moving objects.

In the following section we take an advantage of the binary work assumption in order to relax and decompose the problem into single-competitor parametric subproblems.

3. RELAXATIONS AND DECOMPOSITION

3.1 Relaxations

We will use the fact that $W^{a_k(t)}_{k,X_k(t)} = a_k(t)$ (cf. *binary work* assumption) and instead of the constraints in (P) we will consider the sample path *consumption* constraints

$$\mathbb{E}^{\boldsymbol{\pi}}_t \left[\sum_{k\in\mathcal{K}} W^{a_k(t)}_{k,X_k(t)} \right] = 1, \text{ for all } t \in \mathcal{T}$$

These constraints imply the *epoch-t expected* consumption constraints,

$$\mathbb{E}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} W^{a_k(t)}_{k,X_k(t)} \right] = 1, \text{ for all } t \in \mathcal{T} \tag{2}$$

requiring that the capacity be fully allocated at every time epoch if conditioned on $\boldsymbol{X}(0)$ only. Finally, we may require this constraint to hold only on β-average, as the *β-average capacity consumption constraint*

$$\mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} W^{a_k(\cdot)}_{k,X_k(\cdot)} \right] = \mathbb{B}^{\boldsymbol{\pi}}_0 [1]\,. \tag{3}$$

We remark that this relaxation allows to allocate any number of resource units per period; only the β-average consumption over the entire horizon is constrained. Using $\mathbb{B}^{\boldsymbol{\pi}}_0 [1] = 1$, we obtain the following *Whittle relaxation* (Whittle, 1988) of problem (P),

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} R^{a_k(\cdot)}_{k,X_k(\cdot)} \right] \tag{P$^{\mathrm{W}}$}$$
$$\text{subject to } \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} W^{a_k(\cdot)}_{k,X_k(\cdot)} \right] = 1.$$

The above arguments thus provide a proof of the following result.

Proposition 3.1 (Whittle (1988)). *Problem* (P$^{\mathrm{W}}$) *is a relaxation of problem* (P).

The Whittle relaxation (P$^{\mathrm{W}}$) can be approached by traditional Lagrangian methods, introducing a Lagrangian multiplier, say ν, to dualize the constraint, obtaining thus the following Lagrangian relaxation,

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} R^{a_k(\cdot)}_{k,X_k(\cdot)} \right] + \nu \left\{ 1 - \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} W^{a_k(\cdot)}_{k,X_k(\cdot)} \right] \right\},$$

which can be stated equivalently as

$$\max_{\boldsymbol{\pi}\in\Pi_{\boldsymbol{X},\boldsymbol{a}}} \mathbb{B}^{\boldsymbol{\pi}}_0 \left[\sum_{k\in\mathcal{K}} R^{a_k(\cdot)}_{k,X_k(\cdot)} - \nu \sum_{k\in\mathcal{K}} W^{a_k(\cdot)}_{k,X_k(\cdot)} \right] + \nu. \tag{P$^{\mathrm{L}}_{\nu}$}$$

The classic Lagrangian result, as already observed by Whittle (1988), says the following:

Proposition 3.2 (Whittle (1988)). *For any ν, problem* (P$^{\mathrm{L}}_{\nu}$) *is a relaxation of problem* (P$^{\mathrm{W}}$)*, and further a relaxation of problem* (P).

Note finally that by the definition of relaxation, (P$^{\mathrm{L}}_{\nu}$) for every real-valued ν provides an upper bound for the optimal value of both problem (P$^{\mathrm{W}}$) and problem (P).

3.2 Decomposition into Single-Competitor Subproblems

We now set out to decompose the optimization problem (P$^{\mathrm{L}}_{\nu}$) as it is standard for Lagrangian relaxations, considering ν as a parameter. Notice that any joint policy $\boldsymbol{\pi} \in \Pi_{\boldsymbol{X},\boldsymbol{a}}$ defines a set of single-competitor policies $\widetilde{\pi}_k$ for all $k \in \mathcal{K}$, where $\widetilde{\pi}_k$ is a randomized and non-anticipative policy depending on the *joint* state-process $\boldsymbol{X}(\cdot)$ and deciding the *competitor-k* action-process $a_k(\cdot)$. We will write $\widetilde{\pi}_k \in \Pi_{\boldsymbol{X},a_k}$. We will therefore study the competitor-k subproblem

$$\max_{\widetilde{\pi}_k\in\Pi_{\boldsymbol{X},a_k}} \mathbb{B}^{\widetilde{\pi}_k}_0 \left[R^{a_k(\cdot)}_{k,X_k(\cdot)} - \nu W^{a_k(\cdot)}_{k,X_k(\cdot)} \right]. \tag{4}$$

4. SOLUTION

In this section we will identify a set of optimal policies $\widetilde{\pi}^*_k$ to (4) for all competitors k under the frozen-if-not-allocated assumption, and using them we will construct a joint policy $\boldsymbol{\pi}$ feasible and optimal for problem (P).

Notice that for any fixed ν, (4) is a standard MDP problem with finite state space, finite action space and bounded

immediate reward/cost. The MDP theory assures (Puterman, 2005) that such a problem is optimally solved by a stationary Markov and deterministic policy (convenient additional conditions may be necessary under the time-average criterion). We will denote the set of all stationary Markov deterministic policies by $\Pi^{\text{SMD}}_{\boldsymbol{X},a_k}$. Randomized policies will be of special importance at the end of the section.

The goal is now to reduce the problem to the one solvable by the Gittins index, so we assume the frozen-if-not-allocated assumption throughout this section. The celebrated result of Gittins and Jones (1974) of solving the *multi-armed bandit problem* optimally under the discounted criterion via the now-called Gittins index rule has become classic due to its novelty, importance in applications, and due to the hardness of the problem which had been a known challenge even before its first statement in Robbins (1952). While the original approach developed from the Gittins' intuition relied on a technical interchange argument and was not appreciated quickly, Whittle (1980) provided a simpler proof using dynamic programming. Weber (1992) further provided almost verbal proof based on economic intuition, which was coined to be from "The Book" (cf. Whittle, 2002).

Since the Gittins index was obtained for bandits with zero rewards if not played, we will be interested in the problem with rewards normalized under action 0.

4.1 Normalization under Discounted Criterion and Myopic Criterion

If $0 \leq \beta < 1$ (i.e., under the myopic and discounted criterion), let us consider the normalized (under action 0) variant of problem (4), which is obtained by defining the normalized reward vectors by

$$\widehat{\boldsymbol{R}}^1_k := \boldsymbol{R}^1_k - \frac{(\boldsymbol{I} - \beta \boldsymbol{P}^1_k)\boldsymbol{R}^0_k}{1-\beta}, \qquad \widehat{\boldsymbol{R}}^0_k := \boldsymbol{0}.$$

Note that the work vectors are already normalized due to the *binary-work* assumption; an analogous normalization would otherwise have to be applied.

Let us denote the expected one-period *net reward* by

$$V^a_{k,n} := R^a_{k,n} - \nu W^a_{k,n},$$

where the Lagrangian multiplier ν can be interpreted as a cost per unit of resource utilization. For any stationary Markov deterministic policy $\widetilde{\pi}_k \in \Pi^{\text{SMD}}_{\boldsymbol{X},a_k}$, let us denote by $\mathbb{V}^{\widetilde{\pi}_k}_{k,n}$ the *β-average value function* for competitor-k problem (4) if the state is $n = X_k(t)$ at some time t,

$$\mathbb{V}^{\widetilde{\pi}_k}_{k,n} := \mathbb{B}^{\widetilde{\pi}_k}_t \left[V^{a_k(\cdot)}_{k,X_k(\cdot)} | X_k(t) = n \right].$$

This is independent of t and thus well defined due to stationarity (i.e., time-homogeneity) and due to Markovian nature of the policy. Analogously is defined $\widehat{\mathbb{V}}^{\widetilde{\pi}_k}_{k,n}$, the value function for the normalized problem. In the following proposition we apply direct dynamic programming arguments to prove equivalence of the two problems.

PROPOSITION 4.1. *Suppose that the frozen-if-not-allocated assumption holds. If $0 \leq \beta < 1$, then for any stationary Markov deterministic policy $\widetilde{\pi}_k \in \Pi^{SMD}_{\boldsymbol{X},a_k}$, we have $\mathbb{V}^{\widetilde{\pi}_k}_{k,n} = \widehat{\mathbb{V}}^{\widetilde{\pi}_k}_{k,n} + R^0_{k,n}$ for all n.*

Further, the optimal objective value (4) equals the optimal objective value of the normalized problem summed to the additive constant $R^0_{k,X_k(0)}$.

PROOF. Let $\widetilde{\pi}_k$ be stationary Markov deterministic, and such that it decides action a_n for state n. The value function satisfies the following *balance equation* due to the properties of the β-average operator,

$$\mathbb{V}^{\widetilde{\pi}_k}_{k,n} = (1-\beta) V^{a_n}_{k,n} + \beta \sum_{m \in \mathcal{N}_k} p^{a_n}_{k,n,m} \mathbb{V}^{\widetilde{\pi}_k}_{k,m}.$$

Notice that, by adding and subtracting net rewards $V^0_{k,n}$ and $V^0_{k,m}$, this is equivalent to

$$\underbrace{\mathbb{V}^{\widetilde{\pi}_k}_{k,n} - V^0_{k,n}}_{\widehat{\mathbb{V}}^{\widetilde{\pi}_k}_{k,n}} = \underbrace{(1-\beta) V^{a_n}_{k,n} - \left(V^0_{k,n} - \beta \sum_{m \in \mathcal{N}_k} p^{a_n}_{k,n,m} V^0_{k,m} \right)}_{(1-\beta)\widehat{V}^{a_n}_{k,n}} + \beta \sum_{m \in \mathcal{N}_k} p^{a_n}_{k,n,m} \underbrace{\left(\mathbb{V}^{\widetilde{\pi}_k}_{k,m} - V^0_{k,m} \right)}_{\widehat{\mathbb{V}}^{\widetilde{\pi}_k}_{k,m}}.$$

As indicated by the underbraces, this can be seen as the balance equation for the problem normalized under action 0 (the necessary normalization of the net reward is given by the first underbrace on the right-hand side). Such a normalization gives zero action-0 net rewards, as it is straightforward to check that $\widehat{V}^0_{k,n} = \widehat{R}^0_{k,n} = 0$ due to the frozen-if-not-allocated assumption. For action-1 net rewards, the necessary normalization can be written in the vector form as

$$\widehat{\boldsymbol{R}}^1_k - \nu = \boldsymbol{R}^1_k - \nu - \frac{(\boldsymbol{I} - \beta \boldsymbol{P}^1_k)\boldsymbol{R}^0_k}{1-\beta}.$$

Due to the binary-work assumption, we further have $V^0_{k,n} = R^0_{k,n}$, therefore the underbrace on the left-hand side of the equation gives $\mathbb{V}^{\widetilde{\pi}_k}_{k,n} - R^0_{k,n} = \widehat{\mathbb{V}}^{\widetilde{\pi}_k}_{k,n}$.

Finally, the same relationship holds between the optimal objective values. □

Such an equivalence of the single-competitor subproblem (for the optimal policy only) was proved in Niño-Mora (2001, Section 4) and (for all stopping policies) in Niño-Mora (2007a, Lemma 2.1) using linear programming arguments. For the tax problem, in which by definition $R^0_{k,n} = 0$ for all k and for all n, the equivalence with the multi-armed bandit problem was established in Varaiya et al. (1985, Section II.C) and the equivalence of single-armed subproblems in semi-Markov setting was intuitively explained in Gittins (1989, Section 2.10) (under the name of *ongoing bandit processes*).

We finally remark that after expanding the geometric series, the normalization formula for state n can be rewritten as

$$\widehat{R}^1_{k,n} = \left(R^1_{k,n} - R^0_{k,n}\right) + \left(\beta + \beta^2 + \beta^3 + \dots\right) \sum_{m \in \mathcal{N}_k} p^1_{k,n,m} \left(R^0_{k,m} - R^0_{k,n}\right).$$

Therefore, under the myopic criterion ($\beta = 0$) or in case that $R^0_{k,m} = R^0_{k,n}$ for all $m, n \in \mathcal{N}_k$, the normalization under action 0 is simply $\widehat{\boldsymbol{R}}^1_k := \boldsymbol{R}^1_k - \boldsymbol{R}^0_k$ and $\widehat{\boldsymbol{R}}^0_k := \boldsymbol{0}$.

4.2 Normalization under Time-Average Criterion

Note that reward normalization under the time-average criterion ($\beta = 1$) in Niño-Mora (2001, Section 5) does not apply here because the frozen-if-not-allocated assumption violates the ergodicity assumption of the transition matrix under any stationary policy required there. Moreover the approach in Niño-Mora (2007a) seems to directly require zero action-0 rewards.

However, we could consider the vanishing discount limit $\beta \to 1$ of the normalization introduced in the previous subsection for the discounted criterion. Such a limit gives a finite value to $\widehat{R}^1_{k,n}$ if $R^0_{k,m} = R^0_{k,n}$ for all $m, n \in \mathcal{N}_k$. The normalization is then $\widehat{\boldsymbol{R}}^1_k := \boldsymbol{R}^1_k - \boldsymbol{R}^0_k$ and $\widehat{\boldsymbol{R}}^0_k := \boldsymbol{0}$.

The remaining cases require a more careful normalization or must be treated as they are; such an analysis is left out of this paper.

4.3 Gittins Index and Optimal Solution to Single-Competitor Subproblem

In the remainder of this section we thus assume that competitor rewards are normalized (hat is suppressed).

Gittins and Jones (1974) showed that we can attach to each competitor $k \in \mathcal{K}$ a set of now-called *Gittins index* values $\nu_{k,n}$, independent of other competitors, and defined for each state $n \in \mathcal{N}_k$. The Gittins index value of a given state is equivalent to the objective value of an optimal stopping problem starting from that state, see Gittins (1979); Niño-Mora (2007a). The latter paper also gives the state-of-the-art algorithm that requires $\mathcal{O}(|\mathcal{N}_k|^3)$ arithmetic operations to compute all the index values.

We will base all the further development on the following result.

PROPOSITION 4.2 (WHITTLE (1980)). *Consider subproblem (4) for any fixed ν and suppose that the frozen-if-not-allocated assumption holds. Then, at time epoch t, it is optimal for competitor to use (and pay for) the resource if Gittins index value $\nu_{k,X_k(t)} \geq \nu$ and it is optimal not to use it if $\nu_{k,X_k(t)} \leq \nu$.*

As a convention, a competitor does not use the resource whenever $\nu_{k,X_k(t)} = \nu$ (by the above proposition both using and not using it is optimal in that situation). We denote by π^ν_k the policy that at time epoch t uses the resource if and only if $\nu_{k,X_k(t)} > \nu$; note that π^ν_k is optimal for the competitor-k subproblem (4). Then we have the following auxiliary result.

LEMMA 4.1. *Consider subproblem (4) for any fixed ν and suppose that the frozen-if-not-allocated assumption holds. Under the optimal policy π^ν_k, once the competitor does not use the resource, it continues not using it forever.*

4.4 Gittins Index and Optimal Solution to Lagrangian Relaxation

Since the Lagrangian relaxation is additively composed of mutually independent single-competitor subproblems and constant ν, it is straightforward to obtain the following result.

PROPOSITION 4.3. *Consider Lagrangian relaxation ($\mathrm{P}^{\mathrm{L}}_\nu$) for any fixed ν and suppose that the frozen-if-not-allocated assumption holds for all competitors k. Then, at time epoch t, it is optimal to use the resource for every competitor k satisfying $\nu_{k,X_k(t)} \geq \nu$ and it is optimal not to use it for every competitor k satisfying $\nu_{k,X_k(t)} \leq \nu$.*

If we can identify a policy optimal for the relaxation and at the same time feasible for the original problem with the sample-path constraint of having exactly one competitor using the resource at every time epoch, then such a policy is optimal for the original problem. Although Proposition 4.3 is a necessary result in order to proceed to satisfy the sample-path constraint, it may be far from sufficient.

Indeed, for Lagrangian relaxation ($\mathrm{P}^{\mathrm{L}}_\nu$) with a fixed ν it may be at the same time epoch optimal for all the competitors to use the resource and for none of the competitors to use it. Such a situation, for instance, appears in the case of symmetric competitors, if at some time epoch all of them happen to be in the same state, whose index value moreover equals the parameter ν. To be even more intriguing, if such a situation occurs for the state of the smallest index value, then both "using the resource by all the competitors forever" and "not using the resource by any competitor anymore" are optimal.

We will therefore continue by carefully constructing the following joint policy $\boldsymbol{\pi}^\nu$:

1. For each competitor k, follow policy π^ν_k, i.e, at time epoch t allocate the resource to the competitor if and only if $\nu_{k,X_k(t)} > \nu$;
2. If policies π^ν_k result at time epoch t in not using the resource, and there is at least one competitor satisfying $\nu_{k,X_k(t)} = \nu$, then if the competitor that was allocated the resource in the last period is among those, then allocate the resource to it, otherwise to one arbitrarily chosen competitor with such an index value.

We note that it is not necessary for optimality results in the rest of the section to give preference to the competitor that was allocated the resource in the last period. However, it is reasonable from an implementation point of view, because switching the resource allocation from one competitor to another may require some cost or delay (although we otherwise neglect them in our model), and moreover, several appealing properties of such a policy follow.

As an immediate consequence we have the following claim.

PROPOSITION 4.4. *Consider Lagrangian relaxation ($\mathrm{P}^{\mathrm{L}}_\nu$) for any fixed ν and suppose that the frozen-if-not-allocated assumption holds for all competitors k. Then, the joint policy $\boldsymbol{\pi}^\nu$ is optimal for ($\mathrm{P}^{\mathrm{L}}_\nu$), and it results in using a non-increasing units of resource per period over time.*

Of special interest is the following implications of the last proposition.

PROPOSITION 4.5. *Suppose that the frozen-if-not-allocated assumption holds for all competitors k. Consider Lagrangian relaxation ($R^L_{\nu_0}$) with $\nu_0 := \max\{\nu_{k,X_k(0)} : k \in \mathcal{K}\}$, i.e., ν_0 equals the greatest Gittins index value at the initial time epoch. Then, the joint policy $\boldsymbol{\pi}^{\nu_0}$ is optimal for ($R^L_{\nu_0}$) and it results in allocating exactly one unit of resource to competitors for a positive number of time epochs (possibly infinite), and not allocating any resource to any competitor afterwards.*

4.5 Gittins Index and Solution to Original Problem

The following is the main structural result for the problem, first described in Weber (1992) for the multi-armed bandit

problem, but without linking the result to the fundamental role of the Lagrangian relaxation.

PROPOSITION 4.6. *Suppose that the frozen-if-not-allocated assumption holds for all competitors k. The problem of resource allocation to stochastic dynamic competitors is optimally solved by considering a finite sequence $i = 0, 1, 2, \ldots, I$ of problems $(R^L_{\nu_i})$ solved by policies $\boldsymbol{\pi}^{\nu_i}$, with ν_0 being the greatest Gittins index value at the initial time epoch, and with ν_{i+1} for $i = 0, 1, 2, \ldots, I-1$ being the greatest Gittins index value at the first time epoch at which the policy $\boldsymbol{\pi}^{\nu_i}$ for $(R^L_{\nu_i})$ would result in allocating no resource to any competitor. Moreover, we have $I \leq \sum_{k \in \mathcal{K}} |\mathcal{N}_k| - 1$.*

PROOF. The optimality of considering such a sequence is a straightforward consequence of Proposition 4.5. The number of problems in the sequence is finite, because every change from ν_i to ν_{i+1} is a decrease and may happen at most $\sum_{k \in \mathcal{K}} |\mathcal{N}_k| - 1$ times. □

Since policies $\boldsymbol{\pi}^{\nu_i}$ for the sequence of problems $(\mathrm{R}^L_{\nu_i})$, $i = 0, 1, 2, \ldots, I$, result in allocating the resource to one of the competitors of greatest Gittins index value at each time epoch, we recover the celebrated result of Gittins and Jones (1974), originally proved using an interchange argument.

COROLLARY 4.1. *The Gittins index rule is optimal for the problem of resource allocation to stochastic dynamic competitors.*

We can further observe that the above-defined optimal policy has the *stay-on-a-winner* property: if a competitor is allocated the resource in some time epoch and it proves to be "winning" in that it stays in the same state or moves to a state with a greater Gittins index value, then it is optimal to allocate the resource to it in the next time epoch. Note that if such a competitor is not winning, then we cannot conclude anything; allocating the resource to the competitor may remain optimal or it may become strictly suboptimal, depending on the actual state of the other competitors.

COROLLARY 4.2. *There is an optimal policy for the problem of resource allocation to stochastic dynamic competitors which is a* stay-on-a-winner *policy. In particular, the Gittins index rule is such a policy.*

The stay-on-a-winner property of an optimal policy was first proved in Bradt et al. (1956) for the one-armed and in Berry (1972) for the two-armed Bernoulli bandit problem with finite horizon. The stay-on-a-winner rule was proposed to be used in two-armed Bayesian bandit problems by Robbins (1952), in spite of its non-optimality in general. However, in these early papers playing a bandit could only lead to a success or to a failure and the stay-on-a-winner property in fact referred to its myopic version: if the resource is allocated to a competitor in some time epoch and it proves be "winning" in that the outcome is a success, then it is optimal to allocate the resource to it in the next time epoch.

The following corollary highlights the fact that our problem is a problem of optimal *learning by doing*: there is a phase of exploration that assures to find the competitor which is the most rewarding in the long run, and a phase of exploitation of that competitor.

Let us call a competitor k *irreducible*, if the Markov chain under transitions $\boldsymbol{P}^1_k$ is irreducible, i.e., consists of a single closed set (see Puterman, 2005, Chapter A.2).

COROLLARY 4.3. *There is an optimal policy such that after a finite number of time epochs the resource is allocated to the same competitor forever. In particular, the Gittins index rule is such a policy. If all the competitors are irreducible, then such a competitor being allocated the resource forever by the Gittins index rule is the competitor whose smallest Gittins index value is largest out of all the competitors.*

PROOF. The problem $i = I$ of the finite sequence using $\boldsymbol{\pi}^{\nu_I}$ is such that it always results in allocating the resource to some competitor. This implies that ν_I is such that there is at least one competitor whose Gittins index value never falls strictly below ν_I. Once the resource is allocated to that competitor during the problem $i = I$, the Gittins index rule allocates the resource to it forever. In the case of irreducible competitors, ν_I is the smallest Gittins index value over all the states of such a competitor, and it must be the greatest over all the competitors by definition. □

Note that if some of the competitors are not irreducible, then to which competitor is the resource allocated forever depends on the evolution of the competitors using the resource in a finite number of the early time epochs. Once all the competitors reach any of their irreducible closed sets, then the same argument applies. In particular, such a competitor using the resource forever by the Gittins index rule is the competitor whose smallest Gittins index value within the reached irreducible closed set of the Markov chain under transitions $\boldsymbol{P}^1_k$ is largest out of all the competitors. A more involved argument can be used for testing if it is optimal to allocate the resource to a given competitor forever also when some or even all the competitors have not yet reached an irreducible closed set, by carefully comparing the competitors' Gittins index values of the states accessible from the current states.

Under the time-average criterion (i.e., when $\beta = 1$), if all the competitors are irreducible, then allocating the resource forever to the competitor whose smallest Gittins index value is greatest out of all the competitors is optimal already from the initial time epoch, since the policy employed during the finite number of initial time epochs is irrelevant. That index value further gives the optimal time-average reward for the problem. An analogous observation for the optimal time-average reward in case of symmetric (i.e., statistically identical) competitors was made in Whittle (2005, p. 755).

5. GENERAL COMPETITORS

In this section we consider competitors that do not obey the frozen-if-not-allocated assumption. Such competitors are akin to the restless bandits and index rules are in general not optimal anymore. See Whittle (1988); Jacko (2010a) for an overview of the Lagrangian relaxation approach to restless bandits.

Following the structure of the previous section, we note that it is also possible to apply reward normalization (with a slightly more general formula) to general competitors. However, the existence of an index is not assured anymore, and therefore we will focus on *indexable* competitors. We present the definition of indexability from Jacko (2010b), which admits $-\infty$ and $+\infty$ as valid index values, since there are models with such index values in some states, which seems to have been overlooked in previous work.

DEFINITION 5.1 (INDEXABILITY). *We say that competitor k is* indexable, *if there exist unique values $-\infty \leq \nu_{k,n} \leq$*

$+\infty$ *for all* $n \in \mathcal{N}_k$ *such that the following holds for competitor-*k *subproblem* (4) *:*

1. *if* $\nu_{k,n} \geq \nu$*, then it is optimal for competitor to use (and pay for) the resource in state* n*, and*

2. *if* $\nu_{k,n} \leq \nu$*, then it is optimal for competitor not to use the resource in state* n*.*

The function $n \mapsto \nu_{k,n}$ *is called the* (Whittle) index*, and* $\nu_{k,n}$*'s are called the* (Whittle) index values*.*

Clearly, the Gittins index is a special case of the Whittle index under the frozen-if-not-allocated assumption. However, unlike the Gittins index, the Whittle index is (in general) not equivalent to an optimal stopping problem; not even a one-way implication holds. Moreover, its existence and evaluation is a much more cumbersome task which we will not discuss here (see Niño-Mora, 2007b).

If all the competitors are indexable, then we can approach the Lagrangian relaxation by considering the same joint policy $\boldsymbol{\pi}^{\nu}$ as in the previous section. Using the same arguments as before, this policy is optimal for the Lagrangian relaxation. However, the monotonicity in the resource usage is not guaranteed anymore, which also hinders the construction of a feasible policy for the original problem.

In fact, if we can identify instances of the problem such that a policy optimal for the relaxation, e.g., $\boldsymbol{\pi}^{\nu}$, is at the same time feasible for the original problem with the sample-path constraint of having exactly one competitor using the resource at every time epoch, then such a policy is optimal for the original problem.

This straightforward argument leads us to one such instance.

Proposition 5.1 (Dominant Competitor). *If all the competitors are indexable and the smallest index value of one competitor is greater than or equal to the largest index values of all the other competitors, then the index rule is optimal.*

Proof. Analogously to the previous section, it is possible to construct a sequence of problems which gives rise to an index policy. □

In fact, in this case the index rule allocates the resource always to the same competitor. Optimality in instances with "sufficiently separated indices" was reported in computational experiments, but to the best of the author's knowledge, it has never been proved.

Let us next consider competitors obeying the *reinitializing-if-not-allocated* assumption: if the resource is not allocated ($a = 0$) to competitor k, then the competitor changes its state to a fixed "initial" state $i_k \in \mathcal{N}_k$ (i.e., $p^0_{k,n,i_k} = 1$ for all n). These are briefly called reinitializing competitors.

Proposition 5.2 (Reinitializing Competitors). *If all the competitors satisfy the reinitializing-if-not-allocated assumption, if all are indexable, if all are initially in their respective initial states* i_k*, and if the index value of state* i_k *is the greatest over all the states for each competitor* k*, then the index rule is optimal.*

Proof. Analogously to the previous section, it is possible to construct a sequence of problems which gives rise to an index policy. □

In this case the index rule allocates the resource to one of the competitors (the one with greatest index value of state i_k over all the competitors) as long as its index value is greater than or equal to the competitor with the second largest index value. This competitor (or one of these, if there are various) is then allocated the resource for a single period, after which the priority is given back to the former. The resource is never allocated to any of the remaining competitors.

Next we briefly explain what a reinitializing competitor means in order to suggest its relevance in modeling. In telecommunication networks, the main feature of the TCP mechanism is to drastically decrease the sending rate or to restart by sending a single packet after a packet loss (Jacko and Sansò, 2007). Reinitializing can model "forgetting" of the relevant information (i.e., of the actual state of the competitor) and replacing it with a prior state. In optimal search, reinitializing is well-known to be the worst-case model for finding an object, as it requires a continuous effort until the object is found.

In job sequencing in case of geometrically distributed sizes, if a job is not served, then it reinitializes to state 1. That means that if the job is in state 1 (waiting) and is not served, then it remains waiting. On the other hand, if the job is in state 0 (completed) and is not served, then it moves to state 1, i.e., as if a new job of the same class arrived. Note that it is known that the $c\mu$-rule (which can be obtained by the current approach, (see Jacko, 2010b)) is optimal under arbitrary arrivals (Buyukkoc et al., 1985).

6. CONCLUSION

We have given a brief account of a powerful Lagrangian approach to the study of optimality of index rules. We believe that some of the open questions (not only optimality) especially in the case of restless models could be approached in this way. Nevertheless, proofs of optimality of the Whittle index rule are scarce, the author is only aware of such a result for an opportunistic access model with an infinite state space in Ahmad et al. (2009), in which the Whittle index rule (equivalent to the myopic policy) is optimal if the competitors are symmetric.

As a natural step forward, it would desirable to study optimality of (multi-)index rules also in more general models. Such is the model of Whittle (1988), where several resource capacity units are available at every moment, but every competitor can only be allocated one of those units. Under the frozen-if-not-allocated assumption, a sufficient condition of optimality of the Gittins multi-index rule was given in Pandelis and Teneketzis (1999). A model in which multiple resource capacity units can be allocated to the same competitor was further proposed in Jacko (2009), but no optimality results are known at the moment.

References

Adelman, D. and Mersereau, A. J. (2008). Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727.

Ahmad, S. H. A., Liu, M., Javidi, T., Zhao, Q., and Krishnamachari, B. (2009). Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory*, 55(9):4040–4050.

Berry, D. A. (1972). A Bernoulli two-armed bandit. *Annals of Mathematical Statistics*, 43(3):871–897.

Bertsekas, D. P. (2001). *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, Massachusetts, 2nd edition.

Bradt, R. N., Johnson, S. M., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4):1060–1074.

Buyukkoc, C., Varaiya, P., and Walrand, J. (1985). The $c\mu$ rule revisited. *Advances in Applied Probability*, 17(1):237–238.

Cox, D. R. and Smith, W. L. (1961). *Queues*. Methuen & Co. LTD, London.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177.

Gittins, J. C. (1989). *Multi-Armed Bandit Allocation Indices*. J. Wiley & Sons, New York.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam.

Jacko, P. (2009). Adaptive greedy rules for dynamic and stochastic resource capacity allocation problems. *Medium for Econometric Applications*, 17(4):10–16. Available online at http://www.met-online.nl. Invited paper.

Jacko, P. (2010a). *Dynamic Priority Allocation in Restless Bandit Models*. Lambert Academic Publishing. Invited book.

Jacko, P. (2010b). Restless bandits approach to the job scheduling problem and its extensions. In Piunovskiy, A. B., editor, *Modern Trends in Controlled Stochastic Processes: Theory and Applications*, pages 248–267. Luniver Press, United Kingdom.

Jacko, P. and Sansò, B. (2007). Congestion avoidance with future-path information. In *Proceedings of the EuroFGI Workshop on IP QoS and Traffic Control*, pages 153–160. IST Press.

Klimov, G. P. (1974). Time-sharing service systems I. *Theory of Probability and its Applications*, 19(3):532–551.

Meuleau, N., Hauskrecht, M., Kim, K.-E., Peshkin, L., Kaelbling, L. P., Dean, T., and Boutilier, C. (1998). Solving very large weakly coupled Markov decision processes. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 165–172.

Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1):76–98.

Niño-Mora, J. (2007a). A $(2/3)n^3$ fast-pivoting algorithm for the Gittins index and optimal stopping of a Markov chain. *INFORMS Journal on Computing*, 19(4):596–606.

Niño-Mora, J. (2007b). Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, 15(2):161–198.

Pandelis, D. G. and Teneketzis, D. (1999). On the optimality of the Gittins index rule for multi-armed bandits with multiple plays. *Mathematical Methods of Operations Research*, 50:449–461.

Puterman, M. L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535.

Varaiya, P., Walrand, J., and Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: The discounted case. *IEEE Transactions on Automatic Control*, AC-30(5):426–439.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *Annals of Applied Probability*, 2(4):1024–1033.

Weber, R. and Weiss, G. (1990). On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society, Series B*, 42(2):143–149.

Whittle, P. (1981). Arm-acquiring bandits. *Annals of Probability*, 9(2):284–292.

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *A Celebration of Applied Probability, J. Gani (Ed.), Journal of Applied Probability*, 25A:287–298.

Whittle, P. (2002). Applied probability in Great Britain. *Operations Research*, 50(1):227–239.

Whittle, P. (2005). Tax problems in the undiscounted case. *Journal of Applied Probability*, 42:754–765.