Essays on Indexability of Stochastic Scheduling and Dynamic Allocation Problems

Diego Ruiz-Hernandez

Supervisor: Prof. Kevin D. Glazebrook

A Thesis Presented for the Scientific Degree of
Doctor of Philosophy

Departament d'Economia i Empresa
Universitat Pompeu Fabra
Barcelona, 08005
ESPAÑA

September, 2006

Dipòsit legal: B.39567-2007
ISBN: 978-84-690-7826-6

Abstract

The multiarmed bandit problem (MAB), is one of the simplest non-trivial problems in which one must face the conflict between taking actions which yield immediate reward and taking actions whose benefit will come only later. It was considered intractable until the early 70's, when Gittins and Jones proved that the problem solution can be easily characterized in terms of a set of dynamic allocation indices, attached to project states. The following index policy was then proved to be optimal: work at each time on those projects whose current states have larger indices.

Later, Peter Whittle introduced an important but intractable class of restless bandit problems generalising the multi-armed bandit problem by allowing evolution for passive projects. Despite a developing body of evidence which underscores the strong performance of Whittle's index policy, a continuing challenge to implementation is the need to establish that all competing projects pass an *indexability* test.

In this work, we first deploy Gittins index theory to establish the indexability of *inter alia* general families of restless bandits which arise in problems of stochastic scheduling problems with switching penalties and machine maintenance. We also give formulae for the resulting Whittle indices. Numerical investigations testify to the outstandingly strong performance of the index heuristics concerned.

The second class of problems of interest concerns two families of Markov decision problems which fall within the family of *bi-directional* restless bandits. The *spinning plates* problem concerns the optimal management of a portfolio of reward generating assets whose yields grow with investment but otherwise tend to decline. In the model of asset exploitation called the *squad system*, the yield from an asset tends to decline when it is utilised but will recover when the asset is at rest. In all cases, simply stated conditions are given which guarantee indexability of the problem together with necessary and sufficient conditions for its strict indexability. The index heuristics for asset activation which emerge from the analysis are assessed numerically and found to perform very strongly.

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. Kevin D. Glazebrook, for all the help and guidance during the writing of my dissertation. I'm also indebted to my tutor Daniel Serra for his advice, support and patience during the hard times of my studies. Thanks also go to Dr. Christopher Kirkbride for his guidance and help in the numerical investigation and fruitful discussions during my visits to the University of Edinburgh. Special thanks to Prof. Albert Satorra for his inestimable support at the start-up of my academic career as a lecturer at Universitat Pompeu Fabra. I would also like to thank all the staff of the Department of Economics and Business of Universitat Pompeu Fabra, in particular to Marta Aragay, Marta Araque, Gemma Burballá and Olga Aguilar. Many other people and institutions deserve special mention here, in particular, the staff, faculty and students of the University of Edinburgh and the University of Strathclyde for their hospitality and useful discussions during several visits throughout the last three years. Thanks also to the Servei de Recerca of Universitat Pompeu Fabra for the financial support of the visits to the afore mentioned universities, with special acknowledgement to José Luis de Dios for his invaluable support and help with all the administrative procedures.

Finally, special thanks go to my mother, Doña Rebeca Hernández, who provided tremendous help and encouragement throughout and without whose support and loving care nothing would have been possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Scheduling and dynamic allocation problems are forms of decision making which play an important role in manufacturing as well as in service industries. In our globalized world effective sequencing and scheduling have become a necessity for success in the market place. Companies have, for example, to meet shipping dates committed to the customers, as failure to do so may result in a significant loss of profit. They also have to schedule activities in such a way as to use the resources available in an efficient manner.

More specifically, dynamic allocation problems are concerned about the sharing of limited resources between various activities which are being pursued. Problems in several disciplines fall into a particular class of sequential decision models, in which, at each decision time, an operator or manager observes the state of a number of projects and, based on current system's information, he selects a project or projects to be operated or processed during the following period. The decision problem is to determine the strategy for sequentially activating projects in order to maximize an objective, which is in general a function of the rewards collected, e.g. to maximize the expected total reward over the planning horizon.

As discussed, scheduling concerns the allocation of limited resources to tasks over time. The resources and tasks may take many different forms. Resources may be machines in a workshop, cranes in a port, crew at a construction site, processing units in a computing environment, monetary resources to be allocated to assets, and so on and so forth. The tasks may be operations in a production process, uploading and downloading activities in a port, stages in a construction project, executions of computer programs, and so on. Each task may have a different priority level, starting time and due date, as well as there can be certain pre-established ordering for processing the different jobs. The objectives may also take many forms. One possible objective is the minimisation of the completion time of the last task, another is the minimisation of the number of tasks completed after the committed due dates, and one more is the minimization (maximisation) of certain measure of cost (benefit) incurred (earned) during the operation of the system.

As mentioned, scheduling is a decision making process that exists in most manufacturing and production systems as well as in most information processing environments. A typical example of the role of a scheduling process in a real life situation is the scheduling of tasks in a Central Processing Unit (CPU), described by Michael Pinedo [84] as

> One of the functions of multitasking computing operation system is to schedule the time that the CPU devotes to the different programs that need to be executed. The exact processing times usually are not known in advance. However, the distribution of these random processing times may be known in advance, including their expected values and their variances. In addition, each task usually has a certain priority factor (the operating system typically allows operators and users to specify the priority factor,

or weight, of each task). In this case, the objective is to minimise the expected sum of the weighted completion times for all tasks.

As in the example above, production environments in real life are subject to many sources of uncertainty. Among the sources with major impact are machine breakdowns and unexpected releases of high priority jobs, that is, jobs with large weights. Processing times, which typically are not known in advance, are as well a source of uncertainty. Another source of uncertainty is the time necessary for a nearly depleted resource to recover its productivity.

There are several ways in which randomness can be modelled. To take an example, on may model the possibility of machine breakdowns as an integral part of the processing times. This is done by modifying the distribution of the processing times to take the possibility of breakdowns into account. Another way is to model a (discrete time) deteriorating process having, at each state in an at least countable state space, a positive breakdown probability, i.e. a probability of going back to certain initial state. This formulation is also important when the objective is to schedule a set of maintenance tasks for minimising the possibility of a catastrophic breakdown.

The performance of such systems, as measured by a criterion such as the average time jobs stay in the system or certain measure of expected reward (or cost), may be significantly affected by the policy employed to prioritise over time jobs awaiting for service (*scheduling policy*). The impact of scheduling policies explains the importance and difficulty of the fundamental problem of stochastic scheduling: to design relatively simple scheduling policies that (nearly) achieve given performance objectives.

The theory of stochastic scheduling addresses this problem in a variety of stochastic service system models. Random features such as job processing times are thus modelled by specifying their probability distributions, which are assumed to be known by the system manager. Model assumptions vary across several dimensions, including the class of scheduling policies considered admissible, job arrival and processing time distributions, type and arrangement of service resources and performance objective to be optimized.

According to José Niño-Mora [77], stochastic scheduling models can be classified into three broad categories, which have evolved with a substantial degree of autonomy: models for scheduling a batch of stochastic jobs where a fixed batch of jobs with random processing times, whose distributions are known, have to be completed by a set of machines to optimize a given performance objective; models for scheduling queueing systems, related to the problem of designing optimal service disciplines in queueing systems, where the set of jobs to be completed, instead of being given at the start, arrives over time at random epochs; and multi-armed bandit models, concerned with the problem of optimally allocating effort over time to a collection of projects, which change state in a random fashion depending on whether they are engaged or not.

Three broad families of results have been researched for the afore mentioned areas: One group of results has been addressed to identify optimal policies with a simple structure in more general models, often at the expense of introducing additional technical assumptions. Some other research efforts have addressed harder models, for which the goal of fully characterizing an optimal policy appears out of reach. For these problems researchers aim to design easily implementable heuristic policies with a relatively close to optimal performance. Their degree of sub-optimality may be investigated empirically (by means of simulation, for example) or analytically. Finally, a third group of results elucidates the structure of easily computable policies that solve (near) optimally relatively simple models. An important class of such policies is that of *priority-index rules*: an index is computed for each job type (possibly depending on its current state, but not on that of other jobs), and at each decision epoch jobs or projects with higher index are allocated the resource or assigned higher service priority.

It is well known that the famous multi-armed bandit problem has been optimally solved by a celebrated result by John Gittins (see Section 1.1) by means of *dynamic allocation indices* of the kind mentioned in the last paragraph. Unfortunately this result is no longer available for more realistic extensions of the multi-armed bandit problem. This dissertation is, therefore, devoted to the development of well grounded heuristic policies for finding efficient solutions to some relaxations of the original assumptions of the multi-armed bandit problem to be discussed in the following sections.

1.1 The Multi-armed Bandit Problem

The multi-armed bandit problem (MAB), as it has become known, is one of the simplest non-trivial problems in which one must face the conflict between taking actions which yield immediate reward and taking actions whose benefit will come only later. It has found applications in several disciplines, such as machine scheduling in manufacturing, job search and matching in labour market analysis in economics, target tracking, resource allocation problems in communication networks, industrial research under budget constraint, and so on.

Multi-armed bandit models are concerned with the problem of optimally allocating effort over time to a collection of projects that change state randomly depending on weather they are engaged or not. The classical MAB model can be briefly described as the problem of allocating effort to exactly one out of a collection of projects or arms, at each decision epoch over a planning horizon.

Each project can be in one of a finite number of states. Every time a project is engaged a reward is earned and the project's state evolves following an arm specific transition rule. States of projects not engaged remain unchanged and no reward is earned under passivity. The problem, hence, consists in finding a scheduling policy for dynamically activating projects such that maximises some reward measure over the planning horizon. The following example illustrates one particular case of the global formulation of the MAB.

Example 1.1. *Consider the problem of scheduling a set of stochastic jobs where a job's natural state is the amount of processing finished. After processing a particular job it will either be completed or not completed and advance, with certain positive probability, to a higher completion stage (higher state). If a job is not processed, it remains frozen in the current state (there is not loose, forgetting, or deterioration of the finished work). For simplicity assume that reward is only earned upon completion. Assume also that pre-emption is allowed (i.e. the processing of a job can be interrupted at any time).*

In this case, the decision problem is finding a scheduling policy to dynamically allocate the server effort over time to the different jobs in order to maximise the total expected discounted reward earned over the planning horizon.

This problem is discussed in detail in Pinedo [84, sec. 9.2]. Other examples of the use of the MAB framework in single machine stochastic scheduling problems can be found in Glazebrook [46], Chen and Katehakis [26], and Katehakis and Veinott [61]

The computational complexity of dynamic decision problems involving uncertainty and information is well known and the MAB was considered intractable for a long time, since its first formulation by Thompson [98] in the 1930's. In the early 1970's, however, Gittins and Jones [37, 38] and Gittins [35] presented a surprising solution, which now stands as a landmark result in the field: the problem solution can be easily characterized in terms of a set of dynamic allocation indices, attached to project states; in particular, for each state and project, there exists an calibrating index, such that the following index policy is optimal: work at each time on those projects whose current states have larger indices. The optimality of Gittins rule has been subject of deep analysis, and proofs based on different technical approaches have been provided including interchange arguments [35, 38, 105, 111], dynamic programming [113], induction arguments [108] and conservation laws/linear programming [23].

Unfortunately, for more complex extensions, the Gittins index rule is no longer optimal. For example, the incorporation of costs/delays when switching between projects is studied in [6], where a partial characterization of an optimal index is provided. We will briefly discuss this problem in the next section.

1.1.1 The Multi-armed Bandit Problem with Switching Costs

A general assumption maintained in almost all the work in the area of optimal resource allocation is that the operator can switch instantaneously from one project to another without facing any cost. In reality, when the manager switches between different projects a set-up may be needed, and a cost and/or delay is incurred. Consider the following version of Example 1.1 above.

Example 1.2. *Consider again the problem of scheduling a set of stochastic jobs as described in page 6. In a more realistic setting, it might be considered reasonable to assume that when moving from one job to the other the operator may incur in costs or delays. These penalties could be in terms of the cost/time of replacing certain parts or components of the machine for a different one suitable for the new task, cleaning the machine before starting with the new* *job, displacement of the server to the location of the new job, and so on and so forth.*

In this case, every time the processing of certain job is stopped (without completion) a tear-down cost might be incurred and a set-up needed before starting with processing the new job.

Examples of scheduling of stochastic jobs problems where set-up or switching costs are incurred can be found in Glazebrook [45], Kolonko and Benzing [66], Van Oyen, Pandelis and Teneketzis [103], Van Oyen and Teneketzis [104], Duenyas and Van Oyen [28], Karaesmen and Gupta [58], Reiman and Wein [86], among others.

Although it might seem realistic to include a penalty each time a new project is engaged, its inclusion drastically changes the nature of the problem. In fact, it has been shown by Banks and Sundaram [8] that, in general, it is not possible to construct indices –defined in terms of individual projects– which have the property that the resulting index strategy is optimal on the domain of all multi-armed bandits (MAB) with switching costs. Indeed, this result remains true even if attention is restricted to the case in which the cost of switching is a given (nonzero) constant.

So far, the problem remains unsolved and there is still the need of well grounded heuristic providing an efficient solution to this sort of models. This problem is the matter of Chapter 3, where an (efficient) index characterisation of its solution is provided based on a minor modification of the state space.

1.2 The Restless Bandit Problem

One of the most interesting extensions to the MAB is the restless bandit problem (RB). This model relaxes the assumption above that any non-active process remains fixed by allowing idle projects to change states between decision times. In contrast to the MAB, which is solved optimally by the Gittins Index rule, the RB problem has been proved to be computationally intractable [83]. The RB was first investigated by Peter Whittle [116], who introduced a relaxed version of the problem, which can be solved optimally in polynomial time. Based on this solution, he proposed a priority-index heuristic policy, which reduces to the optimal Gittins index policy in the special case of the multi-armed bandit problem. The following examples provides a Rastless Bandit framework for Example 1.1

Example 1.3. *Consider a following modification of the problem of scheduling a set of stochastic jobs. Now, the case is considered where a passive job is no longer frozen but, instead its state can vary with time. One can think on different reasons for which the state of unfinished jobs can vary during passivity: there can be losses or deterioration in the amount of job finished, the job can involve some kind of learning process that needs certain recall before been engaged again (computer programming is a good example of this case), and so on and so forth. Research projects are also examples of jobs that require certain recall and updating.*

In these cases, even though the general structure of the multi-armed bandit problem remains the same, the relaxation of the idleness assumption changes dramatically the nature of the problem and, consequently, the tools necessary for its solution.

Examples of applications of the restless bandit framework to job and production schedule can be found in Veatch and Vein [106] and Glazebrook and Mitchell [50], among others.

Notwithstanding an accumulating body of empirical evidence testifies the very strong performance of index policies quite generally, one drawback of

Whittle's index heuristic is that it only applies to a restricted class of restless bandits: those satisfying certain *indexability* property, which may be hard to check. Moreover, even when the indices exist, they are not in general guaranteed to be optimal.

This result is in sharp contrast with the well-known optimality of Gittins index rule in the case of the multi-armed bandit problem. It, therefore, confirms the need for well grounded heuristics for finding near optimal (or asymptotically optimal) solutions to the *restless bandit problem*, as well as emphasises the relevance of Whittle's contribution. Chapters 4 and 5 are devoted to establishing indexability conditions for two broad families of restless bandit problems: the Machine Maintenance and the Bi-directional Restless Bandit problems.

1.3 Thesis Structure

In this section we give an outline structure of the remaining chapters of the thesis. Our work concerns identifying different families of restless bandit problems for which indexability is guaranteed and, if possible, to provide closed form expressions for the corresponding indices or, alternatively, to contribute with algorithms for index computation. In Chapter 2 we summarise the main elements from the theory of Markov Decision Processes and Dynamic Programming that constitute the building blocks of the results in this Thesis, as well as offering a more or less detailed discussion of the fundamentals of classical Multi-armed Bandit and Restless Bandit theories. In Chapter 3 we discuss the multi-armed bandit problem with *switching costs*. Chapter 4 concentrates in establishing indexability for the family of so-called *machine maintenance* problems. Finally, in Chapter 5 we analyse two classes of problems that fall within the family of *bi-directional* restless bandits.

As we have already mentioned, Markov Decision Processes and the theories of Multi-armed and Restless Bandit problems constitute the main building blocks for the discussion in this dissertation and the aim of Chapter 2 is to introduce the main concepts and tools to be used in subsequent chapters.

In Section 2.1, we discuss the main characteristics and properties of a *Markov Decision Process*, the associated *Markov Decision Problem* and the techniques available for solving this sort of problems. We also outline the main theoretical results regarding the existence and uniqueness of the optimal solutions to *Markov Decision Problems* under different optimality criteria and present the algorithms to be used in the numerical experiments throughout this work. The *Multi-armed Bandit Problem* is addressed in Section 2.2, where we present the basic structure of a multi-armed bandit and its formulation as a Markov Decision Process, the main theoretical results available for this family of Markov Decision Problems's, and references to the most important contributions in the field. Finally, in Section 2.3 we introduce one of the most promising extensions of the classical Multi-armed Bandit Problem: the *Restless Bandit Problem*, which constitutes the object of this dissertation. Together with a Markov Decision Process formulation of the problem, we present Whittles approach to its solution, and references to the most appealing contributions in the field.

The problem of scheduling a set of stochastic projects when switching costs are incurred is addressed in Chapter 3. Despite this problem has been considered intractable, we show that the switching costs problem can be easily translated into a pure set-up formulation and, consequently, use can be made of standard results available for the multi-armed bandit problem. Moreover, with a simple modification of the state space, we translate this MAB with set-up costs into a simple restless bandit problem. We then prove that this problem is indeed indexable and obtain general expressions for its Gittins indices which can be easily obtained by means of an adaptive greedy algorithm. A further extension of these results focus on the existence of a positive probability for an idle project to abandon the system, we call this the losses problem and prove that it is indexable as well. The index policy is finally shown to perform well when compared with the optimal policy for a set of examples including the losses case and a job scheduling problem.

The chapter is organized as follows. In Section 3.1 related literature is surveyed. In Section 3.2, the problem of translating the multi-armed bandit problem with switching costs into a *pure set-up cost* formulation is addressed, and a proof for indexability of the extended-space *pure set-up cost* formulation is provided. An algorithm for index calculation is presented also in this section. In Section 3.3, two interesting results and one application are discussed. The results regard, respectively, the indexability of the MAB w/SC when there exists a positive probability of losing idle (passive) projects, and the optimality of the index policy when the switching costs are high enough. An application to the scheduling of stochastic jobs is also presented here. Results of an intensive numerical assessment of the performance of the index policy in the MAB w/ SC are offered in Section 3.4. Section 3.5 concludes.

In Chapter 4 we deploy Gittins index theory to establish the indexability of a general family of restless bandits which arise in problems of machine maintenance. We also give formulae for the resulting Whittle indices. The standard Markov Decision problem formulation for the Machine Maintenance Problem can be described as follows: each machine is modelled as a Markov Decision Chain (MDC) that evolves over an at least countable state space with two actions available at each state (to provide maintenance or to let the machine evolve for one more period). Whenever the active action is taken, an arm/state dependent active cost is payed and the machine returns to some *pristine* state. Otherwise, an operation cost is incurred, and the project evolves to a higher state. In such models an increase in state correspond to deterioration of the machine, resulting in higher cost rates. We further consider the case when passive machines face the probability of a catastrophic breakdown, which implies the replacement of the machine at a considerably high cost.

In Section 4.1 we establish the indexability of a class of restless bandits designed to model machine maintenance problems in which maintenance interventions have to be scheduled to mitigate escalating costs as machines deteriorate. In Section 4.2 we explore index structure in the context of two model types, both of which rest on assumptions that are plausible on practice. In Section 4.3 we further develop the findings of Section by offering two families of examples for which explicit formulae for the Whittle index can be derived. Identification of the Whittle indices of concern is followed in Section 4.4 by a numerical investigation which accounts for the strong performance of Whittle's heuristic. Section 4.5 concludes.

In Chapter 5 we deploy Whittle index theory to establish the indexability of two (inter alia) general families of restless bandits the bi-directional restless bandit problems The starting point for our research was the so-called Ehrenfest problem. Whittle [116, 117] describes the Ehrenfest problem as one in which the project may be an individual who has certain number of oxygen bearing blood corpuscles. While working his effectiveness is proportional to the number of corpuscles, but these become depleted. While resting he produces nothing, but his corpuscles gain oxygen. The model thus represents in essential form the two phases of tiring and recovery, with a natural limit on state in both directions.

Our aim has been to generalise this problem to two broader families of bi-directional bandit problems and to establish their indexability. One of them, the Squad System Model, corresponds to the afore mentioned Ehrenfest problem and is a model for the optimal exploitation of assets. In this case, activity returns a reward and decreases the effectiveness of the project, whereas the passive action gives no reward, but increases the profitability of the project. The other one, the Spinning Plates model, reverts this logic and is a model of investment. Here the active action increases the profitability of the asset whereas the passive action implies its deterioration. In this chapter we establish the conditions for this problems to be indexable and provide with an algorithm for the index calculation and, when available, closed form expressions for index calculation. Numerical experiments show the strong performance of the index heuristic concerned.

In Sections 5.1 and 5.2 we give simple and direct accounts of the index structure of, respectively, the *spinning plates* problem and the *squad system*. In both cases we give simply stated conditions that guarantee the model's indexability. Further, algorithms are given which yield the indices. *Strict indexability* means that not only is the problem concerned indexable, but also all the index functions are $1-1$ (namely, that distinct states of an asset have distinct index values). Our analysis yields necessary and sufficient conditions for strict indexability in both models, together with formulae for the indices in closed form. Numerical results testify the very strong performance of the index heuristic for both models. Section 5.3 contains a somewhat brief discussion of the index structure of versions of the spinning plates problem and the squad system with discounted reward criterion. Section 5.4 concludes.

Chapter 2

Markov Decision Processes and Bandit Problems

Introduction

Consider the situation where a decision maker is faced with the problem of influencing the behaviour of a stochastic system as it evolves over time by taking different actions (decisions) at certain moments in time or *decision epochs*. His goal is to choose a sequence of actions which causes the system to preform optimally whith respect to some predetermined performance criterion. The state of the system before the following decision epoch depends on current decisions. Consequently, decisions must not be made myopically, but most anticipate the opportunities and costs (or rewards) associated with future system states.
This problem, typical example of a sequential decision model, is better known as a *Markov Decision Process* (also referred as stochastic control problem or stochastic dynamic program).

Markov Decision Processes (MDP), which are models for sequential decision making when outcomes are uncertain, constitute the main building blocks of the discussion in this dissertation and the aim of this chapter is to introduce the main concepts and tools to be used al through this work. In Section 2.1, we discuss the main characteristics and properties of *Markov Decision Processes*, the associated *Markov Decision Problems* and the techniques available for solving this sort of problems. We also outline the main theoretical results regarding the existence and uniqueness of the optimal solutions to *Markov Decision Problems* under different optimality criteria and present the algorithms to be used in the numerical sections of this work.

A particularly interesting family of sequential decision problems is the *Multi-armed Bandit Problem* (MAB)–for which Peter Gittins contributed a simple solution based on *dynamic allocation indices.* This problem is addressed in Section 2.2, where we present the basic structure of a multi-armed bandit and its formulation as a Markov Decision Process, the main theoretical results available for this family of Markov Decision Problems's (optimality of the index solution), and references to the most important works in the field. We also offer a short discussion of the main extensions and applications of the Multi-armed bandit framework.

Finally, in Section 2.3 we introduce one of the most promising extensions of the classical Multi-armed Bandit Problem: the *Restless Bandit Problem* (RB) –pioneered by Peter Whittle, which constitutes the object of this dissertation. Together with a Markov Decision Process formulation of the problem, we present Whittles approach to its solution, and references to the most appealing contributions in the field.

2.1 Markov Decision Processes

This section introduces the basic components of a *Markov Decision Problem* and discusses some of the main results of *Dynamic Programming Theory* that will be used in subsequent sections. The aim of this section is just to provide with a basic theoretical framework for our research and does not pretend to be an extensive and detailed discussion of the theory of *Markov Decision Processes.* Most of the main results are just enunciated and the reader is referred to the literature for further details. The discussion is based fundamentally on the works by Bellman [10], Bertsekas [21, 22], Puterman [85], Ross [90], Sennott [95], and Tijms [99].

2.1.1 Elements of a Markov Decision Problem

A discrete-time Markov Decision Process (MDP) is defined by six elements: decision epochs, state space, action set, transition probabilities, rewards and a policy or action rule. Here we take the general convention of referring to any problem with the six elements mentioned above as a *Markov Decision Process* and reserve the the expression *problem* to the process itself together with an optimality criterion. In the following lines we will give detailed description of the elements of an MDP. Throughout this dissertation we will be constantly referring to this description.

1. **Decision Epochs**

   We refer as *decision epochs* to the set of times or stages at which decisions can be made. This set can be classified as a discrete set or as a continuum. Because of the nature of the problems analysed in this work, we shall assume that the set of decision times is *discrete*, i.e., decisions are taken at times $t = 0, \dots, T$. Our models are formulated so that a decision epoch corresponds to the beginning of a period.

The planning horizon can be either finite ($T < \infty$) of infinite ($T = \infty$). Again, the nature of the problems addressed in this work implies that we will concentrate on decisions taken over the infinite horizon. Hence, we adopt the convention of representing the discrete-time infinite-horizon decision epochs by $t \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers and t is a decision stage or epoch.

2. **State Space**

   At each decision epoch we say that the system *is in* or *occupies* a state. The set $\mathcal{S}$ represents the collection of all possible states of the underlying system. This set can be finite, countably infinite of infinite, however throughout this work we will assume that $\mathcal{S}$ is *at least* countable, with cardinality $|\mathcal{S}| < \infty$ in the finite case.

3. **Action Set**

   Actions are, by convention, taken at every decision epoch and are, clearly, dependent on the current state of the system (and maybe on the system's action-state history). The set $\mathcal{A}_x$ of actions available in state $x \in \mathcal{S}$ is assumed to be finite. We let $\mathcal{A} = \cup_{x \in \mathcal{S}} \mathcal{A}_x$ and further restrict our models requiring that $\mathcal{A}_x = \mathcal{A},\ x \in \mathcal{S}$.

   Actions may be chosen either randomly or deterministically. Let $\mathcal{P}(\mathcal{A})$ be the collection of probability distributions on subsets of $\mathcal{A}$. Choosing actions randomly means selecting a probability distribution $\varrho(\cdot) \in \mathcal{P}$, where action $a \in \mathcal{A}$ is selected with probability $\varrho(a)$. Degenerate probability distributions correspond to a deterministic action choice.

   Shall a model consider actions are taken only when necessary and not at every decision epoch, then we will introduce an additional element in $\mathcal{A}$ representing the action *not intervene*, with no further changes to be done to the model.

4. **Transition Probabilities**

   When at certain decision epoch t, action $a(t) = a \in \mathcal{A}$ is taken in state $X(t) = x \in \mathcal{S}$, the system evolves to state $X(t+1) = x' \in \mathcal{S}$ with probability $P_t^a(x, x') = p\{X(t+1) = x | X(t) = x, a(t) = a\}$. We further assume that transition probabilities are time independent, i.e. $P_t^a(x, x') = P^a(x, x')$, for all $t \in \mathbb{N}$; and

   $$\sum_{x' \in \mathcal{S}} P^a(x, x') = 1, \text{ for all } x \in \mathcal{S}, a \in \mathcal{A}. \tag{2.1}$$

   Finally, let $\mathbf{P}^a = (P^a(x, x'))_{x, x' \in \mathcal{S}}$ be the transition probability matrix corresponding to action $a \in \mathcal{A}$.

5. **Rewards**

   A reward $R_t^a(x, x')$ is a real value function $R : \mathcal{S} \to \mathbb{R}$ representing the value at time t of the reward received when, at period t, the system performs a transition from state $X(t) = x \in \mathcal{S}$ to state $X(t+1) = x' \in \mathcal{S}$ under action $a(t) = a \in \mathcal{A}$. We again assume that the reward is independent of the time at which decision is taken, i.e. $R_t^a(x, x') = R^a(x, x')$ for all $t \in \mathbb{N}$. Furthermore, we assume that the rewards are bounded, i.e. $\exists\ B < \infty$ such that $|R^a(x, x')| < B$ for all $x, x' \in \mathcal{S}$ and $a \in \mathcal{A}$.

   When the reward depends on the state of the system at the next decision epoch, we adopt the *telescoped* version or *expected value* at decision epoch t by computing:

   $$R^a(x) = \sum_{x' \in \mathcal{S}} P^a(x, x') R^a(x, x'), \quad \text{for all } x \in \mathcal{S}.$$

As before, we use $\mathbf{R}^a = \left(R^a\left(x\right)\right)_{x\in\mathcal{S}}$ for the reward vector associated to action $a \in \mathcal{A}$.

6. **Policy**

   A policy π is a rule that indicates at every decision epoch t the action $a\left(t\right) = a \in \mathcal{A}$ to be taken given the fact that the system is in state $X\left(t\right) = x \in \mathcal{S}$, with the aim of maximising (minimising) certain measure of reward (respectively, cost). In Section 2.1.2 we present a more detailed description of the different procedures for action selection depending on how they incorporate past information and how they select actions.

We refer to the collection of objects $(T, \mathcal{S}, \mathcal{A}, \mathbf{P}^a, \mathbf{R}^a)$ as a *Markov Decision Process*. If the planning horizon spans to the infinite we omit the time-horizon term T. The qualifier *Markov* is used because the transition probabilities and reward functions depend on the past only through the current state of the system and the action selected by the decision maker in that state.

2.1.2 Decision Rules and Policies

Given a Markov Decision Process as described in Section 2.1.1 (§1 to §6 above), a decision rule d_t prescribes which action is to be taken at a specified decision epoch t. We shall be mainly concerned with *Markovian* decision rules, in which, for each state $x \in \mathcal{S}$, $d_t\left(x\right)$ is a random variable with a given distribution taking values in $\mathcal{A}$ independently of the history of the process, in contrast with *history dependent* decision rules, which are dependent on the past history of the system (represented by the sequence of previous states and actions).

Regarding the way actions are selected, our primary focus will be on *deterministic* decision rules. Such decision rules are functions $d_t : \mathcal{S} \rightarrow \mathcal{A}$, which specify (with certainty) the action choice when the system occupies state x at epoch t.

In contrast with the *deterministic* rules, a *randomized* decision rule d_t specifies a probability distribution $\varrho(\cdot)$ on the set of actions. Randomized Markovian decision rules map the set of states into the set of probability distributions on the action space, that is $d_t : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$. In this case, $\varrho_{d_t(X(t))}(\cdot) \in \mathcal{P}(\mathcal{A})$. A *deterministic* decision rule can be regarded as a special (degenerated) case of *randomized* decision rules in which $\varrho_{d_t(X(t))}(a) = 1$.

A *policy* π for a Markov Decision Problem prescribes which decision rule to use at each time by giving a sequence of decision rules:

$$\pi = \{d_1, d_2, \ldots, d_{T-1}\}, \; T \leq \infty,$$

depending on the state of the system at each decision epoch t and, maybe, on the previous history of the system's evolution.

If all decision rules $d_t(\cdot)$ are *deterministic* (respectively *randomized*), the policy is said to be *deterministic* (resp. *randomized*). A policy is *stationary* if the decision rule employed is invariant over time, so that

$$\pi = \{d, d, \ldots\} = (d)^T, \; T \leq \infty. \tag{2.2}$$

for some *stationary* decision rule d. Throughout this dissertation we adopt the following notational convention: as in the stationary case decision rules are time invariant, we will use indistinctly $\pi(x)$ and $d(x)$ for referring to the action prescribed by policy (decision rule) π (respectively d) at state x.

If d is Markovian randomized, policy π is said to be *stationary, Markovian randomized*; otherwise if d is Markovian deterministic π will be *stationary, Markovian deterministic*.

As the focus of this dissertation is on infinite-horizon Markov Decision Processes (for which stationary policies are fundamental), we adopt the notational convention of representing by $\mathcal{M}$ the class of stationary, Markovian deterministic policies and by $\mathcal{H}$ the class of non-stationary, Markovian deterministic policies.

2.1.3 Optimality Criteria

Given an MDP as above, one may consider several related problems associated with corresponding optimality criteria. In what follows we shall let E_π denote the expectation operator under policy π (the discussion in the following lines will be concentrated on policies in the class $\mathcal{M}$), whereas $X\left(t\right)$ and $a\left(t\right)$ will denote the state and action at time t, respectively. The initial state $X\left(0\right)$ is assumed to be known at least up to a probability distribution, being $x \in \mathcal{S}$ with probability $P\left(x\right)$. We denote the corresponding probability vector by $\mathbf{P_0} = \left(P\left(x\right)\right)_{x\in\mathcal{S}}$.

To determine the policies that are, in some sense optimal, we first need to decide on an optimality criterion. The main optimality criteria are classified according to the planning horizon (finite or infinite) and the way the rewards are accumulated over time (average, discounted or total).

I. **Infinite-horizon Discounted Criterion.** Given a discount factor $0 < \beta < 1$, find an optimal policy $\pi^* \in \mathcal{M}$ that maximizes the total expected discounted reward earned over an infinite horizon; i.e.

$$V\left(\beta\right) = \max_{\pi\in\mathcal{M}} \left\{ E_\pi \left[\sum_{t=0}^{\infty} \beta^t R^{a(t)}\big(X\left(t\right)\big) \right] \right\}. \tag{2.3}$$

Note that (2.3) is well defined because rewards are bounded (see §5 in page 19) and $\beta < 1$, which implies $\left|V\left(\beta\right)\right| < \frac{B}{1-\beta}$.

A policy π^* is said to be *β-optimal* if $V_{\pi^*}(\beta) = V(\beta)$, with $V_{\pi^*}(\beta) = E_{\pi^*}\left[\sum_{t=0}^{\infty} \beta^t R^{a(t)}\big(X(t)\big)\right]$.

II. **Finite-horizon Discounted Criterion.** Given discount factor β as above, find a policy $\pi^* \in \mathcal{H}$ that maximizes the total expected discounted reward earned over a finite horizon, T; i.e.

$$V_T(\beta) = \max_{\pi \in \mathcal{H}} \left\{ E_\pi \left[\sum_{t=0}^{T-1} \beta^t R^{a(t)}\big(X(t)\big) \right] + E_\pi \Big[\beta^T g\big(X(T)\big) \Big] \right\}, \qquad (2.4)$$

where $g(X(T))$ denotes the discounted *terminal reward* when the system's state at time T is X.

III. **Infinite-horizon Time-average Criterion.** Find a policy $\pi^* \in \mathcal{M}$ that maximises the long run time-average rate of reward earned per time unit; i.e.

$$V = \max_{\pi \in \mathcal{M}} \left\{ \lim_{T \to \infty} \frac{1}{T} E_\pi \left[\sum_{t=0}^{T} R^{a(t)}\big(X(t)\big) \right] \right\}. \qquad (2.5)$$

If the limit does not exist, $V_\pi = \lim_{T\to\infty} \frac{1}{T} E_\pi \left[\sum_{t=0}^{T} R^{a(t)}\big(X(t)\big)\right]$ can be defined by the $\liminf$, (see Ross [90], page 89). We say that π^* is *average-reward optimal* if $V_{\pi^*} = V$.

A question arises here regarding the existence of an optimal policy for the average-reward criterion. Unfortunately, the answer is no, optimal policies need not to exist and, there can be found instances for which, even if an optimal (*non-stationary*) solution exists, no stationary policy can be found within ϵ of optimality. The conditions that result in the existence of an optimal stationary policy will be discussed later, in §III' in Section 2.1.4.

The following important relation between the infinite-horizon discounted and time average criteria can be established via, e.g. Tauberian Theorems (see, for example, Putterman [85] pp. 414-421, or Sennott [95] pp. 97-101 and Appendix A.4). Under appropriate regularity conditions,

$$V = \lim_{\beta \to 1} (1 - \beta) V(\beta).$$

IV. **Finite-horizon Total Reward Criterion.** Find a policy $\pi^* \in \mathcal{H}$ that maximises the total expected reward earned over a finite horizon, T; i.e.

$$V_T = \max_{\pi \in \mathcal{H}} \left\{ E_\pi \left[\sum_{t=0}^{T-1} R^{a(t)}\big(X(t)\big) \right] + E_\pi \Big[g\big(X(t)\big) \Big] \right\}, \tag{2.6}$$

where $g(X)$ denotes the *terminal reward* when the process is in state X at terminal time T. The following result is trivial,

$$V_T = \lim_{\beta \to 1} V_T(\beta).$$

In Section 2.1.4 we discuss the standard solution of MPDs by Dynamic Programming and present the algorithms used to find the numerical solutions to the examples discussed throughout this work. However, as the models discussed here are modelled according to criteria §I and §III, we will just concentrate our discussion on methods addressed to these particular criteria.

2.1.4 Solution by Dynamic Programming

The classical approach to the solution of Markov Decision Problems is based on formulating and solving a standard set of *dynamic programming* (DP) *equations*, also known and Hamilton-Jacobi-Bellman equations. These are functional equations satisfied by the optimal value functions as described below.

We next present the DP formulation corresponding to each of the optimality criteria described in Section 2.1.3, with special attention to criteria §I and §III. In what follows we use the short hand notation $|x$ for the conditioning on the initial state: $|X(0) = x$.

I'. **Infinite-horizon Discounted Criterion.** Consider the optimal value function

$$V(x,\beta) = \max_{\pi\in\mathcal{M}} \left\{ E_\pi \left[\sum_{t=0}^{\infty} \beta^t R^{a(t)}\big(X(t)\big) \big| x \right] \right\}, \ x \in \mathcal{S}. \tag{2.7}$$

where $V(x,\beta)$ corresponds to the optimal value equation (2.3) provided the initial state is x. An application of the *principle of optimality*[1] (a fundamental result of dynamic programming, first stated verbally by Bellman [10], page 83), yields the result that the $V(x,\beta)$'s satisfy the following set of equations:

$$V(x,\beta) = \max_{a\in\mathcal{A}} \left\{ R^a(x) + \beta \sum_{x'\in\mathcal{S}} P^a(x,x')\, V(x',\beta) \right\}, \ x \in \mathcal{S}. \tag{2.8}$$

The set of equations (2.8) in the variables $V(x,\beta)$ represent the DP formulation of the discounted infinite-horizon Markov Decision Problem (2.3). They can be shown to have a unique solution, which characterizes optimal policies: optimal actions in state x correspond to maximising a's

[1] The key idea is that optimization over time can often be regarded as *optimization in stages*. The basic trade-off is between the alternative of earning the maximum possible reward at the current stage against the implication this would have in terms of rewards at future stages. The best action is, hence, the one that maximises the sum of rewards earned at the current epoch and the largest reward that can be earned from all subsequent decision epochs, accordingly with this decision. Formally, the principle of optimality establishes that, "from any point on an optimal trajectory, the remaining trajectory is optimal for the corresponding problem initiated at that point". See, for example, Puterman [85], pp. 86-88 for a formal statement of this principle.

in (2.8). From this result it follows that there exists an optimal policy that is stationary and deterministic.

The following two theorems (from Ross [90], pp. 32-33) summarize the discussion above, proofs are ommited.

Theorem 2.1. ***(Ross, Theorem 2.2, p.32)***
Let s *be the stationary policy that, when the process is in state* x*, selects that action (or an action) maximising the r.h.s. of* (2.8)*, that is* $\mathrm{s}(x)$ *is such that*

$$\begin{aligned}R^{\mathrm{s}(x)}(x) + \beta \sum_{x'} P^{\mathrm{s}(x)}(x, x') V(x', \beta) \\ = \max_{a} \left\{ R^{a}(x) + \beta \sum_{x' \in \mathcal{S}} P^{a}(x, x') V(x', \beta) \right\} \quad x \in \mathcal{S}.\end{aligned}$$

Then

$$V_{\mathrm{s}}(x, \beta) = V(x, \beta) \quad \textit{for all} \;\; x \in \mathcal{S},$$

and hence s *is* β*-optimal.*

Finally,

Proposition 2.1. ***(Ross, Proposition 2.3, p. 33)***
V *is the unique bounded solution of the optimality equation* (2.8)*.*

Once established the existence and uniqueness of the optimal policy solving V, we now present an algorithm for finding the solution to the *infinite-horizon discounted reward* MDP, the well known *Value Iteration Algorithm.*

Value Iteration (VI) is the most widely used and best understood algorithm for solving discounted Markov decision problems. It is also known under names as *successive approximations method*, *over relaxation*, *backward induction*, and so on.

The appeal of this algorithm may perhaps be attributed to its conceptual simplicity, its ease in coding and implementation, and its similarity to approaches used in other areas of applied mathematics. In addition to providing a simple numerical tool for solving these models, it can be used to obtain results regarding the structure of optimal policies.

The value iteration algorithm below finds an stationary ϵ-optimal policy, $\pi^{\epsilon} = (d_{\epsilon})^{\infty}$ as defined in (2.2), and an approximation to its value $V^{\epsilon} = V_{\pi^{\epsilon}}(\beta)$, as discussed around (2.3). This stationary policy is ϵ-optimal within a finite number of iterations. Of course, $(d_{\epsilon})^{\infty}$ might be optimal, but the algorithm as stated below provides no means of determining this. By combining this algorithm with methods for identifying suboptimal actions, we can often ensure that the algorithm terminates with an optimal policy. In practice, choosing ϵ small enough ensures that the algorithm stops with a policy that is very close to optimal.

Convergence of the algorithm in Fig. 2.1 is not restricted to models with discrete state space and finite action sets. Unfortunately, numerical evaluation of the maximization in (2.9) is only practical when S is finite. For more general state spaces, the maximization can only be carried out by using special structure of the rewards, transition probabilities, and value functions to determine the structure of maximizing decision rules. General alternatives include discretization and/or truncation.

Please address to the references in this section, in particular to Puterman [85] (Section 6.3), for results regarding convergence of the above

```
INPUT: $\mathcal{S}, R, P, \mathcal{A}, \beta$
INITIALIZATION:
Set $V^0 = (c_i)_{i \in \mathcal{S}}$, with $0 \leq c_i \leq \infty$;
    $\epsilon > 0$;
    $k = 0$.
PROCEDURE:
While $\left\| V^{k+1} - V^k \right\| > \frac{(1-\beta)}{2\beta}\epsilon$ do
Set $k = k + 1$
    $V^k(x) = \max_{a \in \mathcal{A}} \left\{ R^a(x) + \beta \sum_{x' \in \mathcal{S}} P^a(x, x') V^{k-1}(x') \right\}, x \in \mathcal{S}.$   (2.9)
End While
Choose
    $d_\epsilon(x) \in \operatorname{argmax}_{a \in \mathcal{A}} \left\{ R^a(x) + \beta \sum_{x' \in \mathcal{S}} P^a(x, x') V^k(x') \right\},$
    $V^\epsilon(x) = V_{d_\epsilon}(x).$
OUTPUT: $\mathbf{d}_\epsilon$, $\mathbf{V}^\epsilon$.
```

Figure 2.1: Value Iteration Algorithm for the DP Infinite-horizon Discounted Reward Criterion

algorithm, an extended discussion on properties of convergence rates, and procedures for increasing the efficiency of the VI algorithm.

II'. **Finite-horizon Discounted Criterion.** Define the optimal value function

$$V_k(x, \beta) = \max_{\pi \in \mathcal{H}} \left\{ E_\pi \left[\sum_{t=0}^{k-1} \beta^t R^{a(t)}\big(X(t)\big) + \beta^k g\big(X(k)\big) \big| x \right] \right\}, \ 1 \leq k \leq T$$

$$V_0(x, \beta) = g(x),$$

for all $x \in \mathcal{S}$. Note that $V_k(x, \beta)$ represents the optimal value provided the initial state is x and the problem has horizon k. By the principle of optimality, the $V_k(x,\beta)$'s satisfy the following DP equations:

$$V_k(x,\beta) = \max_{a\in\mathcal{A}_x}\left\{R^a(x) + \beta\sum_{x'\in\mathcal{S}} P^a(x,x')\,V_{k-1}(x',\beta)\right\},\ 1\leq k\leq T$$

$$V_0(x,\beta) = g(x),$$

for all $x\in\mathcal{S}$. As before, these equations have a unique solution, for which optimal deterministic policies (though not stationary) can be easily recovered. As this criterion is not used in this dissertation, no more attention will be devoted to the particular aspects of its solution by DP.

III'. **Infinite-horizon Time-average Criterion.** For this criterion the derivation and interpretation of the DP equations are not as straightforward as for the other criteria. As defined in (2.7), denote by $V(x,\beta)$ the optimal value function for the infinite-horizon discounted-reward criterion MDP above, then write

$$v = \lim_{\beta\to 1}(1-\beta)\,V(x,\beta),$$

$$v(x) = \lim_{\beta\to 1} V(x,\beta) - \frac{v}{1-\beta},\ x\in\mathcal{S},$$

where v represents the time average reward rate, which, under regularity conditions, does not depend on the initial state. The (bounded) function $v(x)$ is called the *relative reward differential* corresponding to starting in state x. It can be shown that v and the $v(x)$'s satisfy the following DP equations

$$v + v(x) = \max_{a\in\mathcal{A}_x}\left\{R^a(x) + \sum_{x'\in\mathcal{S}} P^a(x,x')\,v(x')\right\},\ x\in\mathcal{S},$$

from whose solution the optimal *Markovian* policies can be recovered.

In particular, Ross [90] provides a proof for the following two theorems. The first one establishes the existence of an optimal stationary policy:

Theorem 2.2. ***(Ross, Theorem 2.1, p.93)***
If there exists a bounded function $v(x)$, $x \in \mathcal{S}$, and a constant v such that

$$v + v(x) = \max_{a \in \mathcal{A}} \left\{ R^a(x) + \sum_{x' \in \mathcal{S}} P^a(x, x')\, v(x') \right\},\ x \in \mathcal{S} \tag{2.10}$$

then there exists a stationary policy π^* such that

$$v = V_{\pi^*}(x) = \max_{\pi} V_{\pi}(x),\ x \in \mathcal{S}$$

where $V_\pi(x)$ corresponds to the infinite-horizon average reward rate in the expression between brackets in (2.5) *when the initial state is x; and π^* is any policy that, for each $x \in \mathcal{S}$ prescribes an action that maximises the r.h.s. of* (2.10).

The second theorem establishes the requirements for the conditions in Theorem 2.2 to be satisfied:

Theorem 2.3. ***(Ross, Theorem 2.2, p.95)***
If there exists an $N < \infty$ such that

$$|V(x,\beta) - V(0,\beta)| < N;\ 0 < \beta < 1,\ x \in \mathcal{S}$$

then

(i) there exists a bounded function $v(x)$ and a constant v satisfying (2.10)*;*

(ii) for some sequence $\beta_n \to 1$, $v(x) = \lim_{n\to\infty} [V(x,\beta_n) - V(0,\beta_n)]$;

(iii) $\lim_{\beta\to 1}(1-\beta)V(0,\beta) = v$.

Ross' discussion concludes with the following Corollary:

Corollary 2.1. ***(Ross, Corollary 2.5, p.98)***
If the state space is finite and every stationary policy gives rise to an irreducible Markov chain, then $V(x,\beta) - V(0,\beta)$ is uniformly bounded, and hence the conditions in Theorem 2.3 are satisfied.

As the numerical experiments in this work will be just concerned with MDP's satisfying the conditions above, we can leave at this point our discussion and proceed with the computational approach that will be used for finding the corresponding optimal policies under the infinite-horizon time-average criterion.

The value iteration algorithm depicted in Figure 2.2 finds a stationary ϵ-*optimal* policy $(d_\epsilon)^\infty$ and an approximation to its gain, when certain extra conditions are met.

Two important points arise from the algorithm in Figure 2.2: 1) the value iteration need not converge in models with periodic transition matrices[2]; and 2) the sequence $\{V^t\}$ may diverge, but $sp\left(V^{t+1}-V^t\right)$ may still converge.

In Theorem 6.6.6 Puterman [85], shows that

$$sp\left(V^{t+2}-V^{t+1}\right) \le \gamma sp\left(V^{t+1}-V^t\right),$$

[2] A state or class in a Markovian Process is said to be periodic if $P^n(x|x) > 0$ with $n > 1$ for any $x \in \mathcal{S}$, where $P^n(\cdot|\cdot)$ is the *n-step* transition probability. If $n = 1$, the state (class) is *aperiodic*. A Markov chain with irreducible transition matrix is called *aperiodic* if all its states are *aperiodic*. See Puterman [85], Appendix A.

```
**INPUT:** $\mathcal{S}, R, P, \mathcal{A}$
**INITIALIZATION:**
Set $V^0 = (c_i)_{i\in\mathcal{S}}$, with $0 \leq c_i \leq \infty$;
    $SP = \mathcal{L}$, where $\mathcal{L}$ is any big number;
    $\epsilon > 0$;
    $k = 0$;
**PROCEDURE:**
**While** $SP \geq \epsilon$ **do**
Obtain
    $V^{k+1}(x) = \max_{a\in\mathcal{A}_x} \left\{ R^a(x) + \sum_{x'\in\mathcal{S}} P^a(x,x') V^k(x') \right\},\ x \in \mathcal{S};$
    $sp\left(V^{k+1} - V^k\right) = \max_{x\in\mathcal{S}} \left\{V^{k+1}(x) - V^k(x)\right\} - \min_{x\in\mathcal{S}} \left\{V^{k+1}(k) - V^k(x)\right\};$
set $SP = sp\left(V^{k+1} - V^k\right)$;
    $k = k + 1$.
**End While**
Choose
    $d_\epsilon(x) \in \operatorname{argmax}_{a\in\mathcal{A}} \left\{ R^a(x) + \sum_{x'\in\mathcal{S}} P^a(x,x') V(x') \right\}$
**OUTPUT:** $\mathbf{d}_\epsilon$.
```

Figure 2.2: Value Iteration Algorithm for the DP Infinite-horizon Time-Average Criterion

where

$$\gamma = \underset{x,x'\in\mathcal{S},\ a,a'\in\mathcal{A}}{max} \left[1 - \sum_{y\in\mathcal{S}} \min\left\{ P^a(x,y), P^{a'}(x',y) \right\} \right] \tag{2.11}$$

Consequently, if $\gamma < 1$, equation (2.11) ensures that in a finite number of iterations, criterion $sp\left(V^{t+1} - V^t\right) < \epsilon$ will be satisfied. However, γ may equal 1 in a unichain[3] aperiodic model, but value iteration may still

[3] A markov Chain with finite state space is called *unichain* if it consists of one *closed irreducible* set and a (possibly empty) set of *transient* states. A subset $\mathcal{C}$ of $\mathcal{S}$ is *closed* if

converge. It motivates a more general condition which ensures convergence.

The following result follows from Theorems 8.5.2 and 8.5.3 and the discussion in Section 8.5.2. in Puterman [85].

Theorem 2.4. *Suppose either*

(a) $0 \leq \gamma < 1$, *with* γ *given by* (2.11).

(b) $\exists x' \in \mathcal{S}$ *and* $K \in \mathbb{N}$ *such that, for any deterministic Markovian policy* π, $P_{\pi}^{K}(x'\,|x) > 0$ *for all* $x \in \mathcal{S}$, *or*

(c) all policies are unichain and $P^{a}(x\,|x) > 0$ *for all* $x \in \mathcal{S}$ *and* $a \in \mathcal{A}_x$.

Then value iteration achieves stopping criterion $sp\left(V^{k+1} - V^{k}\right) < \epsilon$ *for any* $\epsilon > 0$.

Puterman also provides a proof for the following important result:

Theorem 2.5. *Suppose that all stationary policies are unichain and that every optimal policy has an aperiodic transition matrix. Then, for all* $V^0 \in V$ *and any* $\epsilon < 0$, *the sequence* $\left\{V^k\right\}$ *generated by the value iteration algorithm satisfies* $sp\left(V^{k+1} - V^{k}\right) < \epsilon$ *for some finite* T.

Finally, it can be proven that under any conditions which ensure that $sp\left(V^{k+1} - V^{k}\right) < \epsilon$ for some finite k, value iteration algorithm in Figure 2.2 identifies an ϵ-*optimal* policy, an approximation to which value is given by

$$g' = \frac{1}{2}\left[\max_{x\in\mathcal{S}}\left(V^{k+1}(x) - V^{k}(x)\right) + \min_{s\in\mathcal{S}}\left(V^{k+1}(x) - V^{k}(x)\right)\right],$$

no state outside $\mathcal{C}$ is accessible from any state in $\mathcal{C}$. The closed set $\mathcal{C}$ is *irreducible* if no proper subset of $\mathcal{C}$ is closed. Finally, state x is *transient* if and only if the expected number of visits to state x is finite in an infinite horizon. For further discussion on Markov Chains see Puterman [85], Appendix A and references therein.

with $|g' - g^*| < \epsilon/2$.

This will be the approach followed in the numerical experiments throughout this dissertation.

Further discussion about the particularities of this approach is beyond the scope of this cpater. A more detailed analysis together with alternative computational approaches can be found in Puterman [85], Chapters 8 and 9, and Ross [90], Chapter V, as well as in the other references in this chapter, in particular Bertsekas [21, 22] and Tijms [99].

IV'. **Finite-horizon Total Reward Criterion.** Define the optimal value function

$$\begin{aligned} &V_k(x) = \max_{\pi \in \mathcal{H}} \left\{ E_\pi \left[\sum_{t=0}^{k} R^{a(t)}\big(X(t)\big) + g\big(X(k)\big) \big| x \right] \right\}, \ 1 \le k \le T, \\ &V_0(x) = g(x), \end{aligned}$$

for $x \in \mathcal{S}$. The corresponding DP optimality equations are

$$\begin{aligned} &V_k(x) = \max_{a \in \mathcal{A}_x} \left\{ R^a(x) + \sum_{x' \in \mathcal{S}} P^a(x, x') V_{k-1}(x') \right\}, \ 1 \le k \le T, \\ &V_0(x) = g(x), \end{aligned}$$

for $x \in \mathcal{S}$. These equations also have a unique solution, from which optimal non-stationary deterministic policies can be easily recovered. This criterion is out of the scope of this dissertation and no more attention will be given.

The theory of Markov Decision Problems has found application in a wide range of disciplines from natural and physical sciences and engineering to management, economics and social sciences.

Examples can be found in: 1) deterministic problems like shortest route problems, critical path analysis, sequential allocation and inventory control with known demands; 2) optimal stopping problems like asset selling, the well known secretary problem, call options in financial markets and so on; 3) controlled discrete-time dynamic systems, in particular inventory control models and economic growth models; 4) discrete time queueing models (admission and/or service rate control); and so on and so forth.

A particularly interesting family of MDP's is the one of sequential decision models in which at each decision epoch the decision maker observes the state of a collection of *Markov Decision Processes* and, based on information available for each process (state, transition probabilities and rewards/costs) selects a process to use during the next period. This family of models is often referred as *Bandit Models*, after the decision problem facing a gambler when deciding whether or not to play a particular slot machine or *one armed bandit* with unknown outcome. When the gambler must choose between M different machines we refer to the model as a *Multi-armed Bandit* (MAB) problem .

The *multi-armed bandit* and it's extensions are the main subject of this dissertation. In the next two sections we deploy the theoretical foundations of the classical *multi-armed bandit* and one of its most promising extensions: the *Restless Bandit* problem.

2.2 The Multi-armed Bandit Problem

The multi-armed bandit problem, originally described by Robbins (1952)[4], is a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. The choice consisting, fundamen-

[4] Actually, it was Thompson [98] who posed the forst bandit problem, but the problem was not seriously addressed again until Robbins.

tally, in deciding which one in a collection of different competing projects (or arms) to play (or operate) in a sequence of decision epochs so as to maximize his reward. This classical problem has received much attention because of the simple model it provides of the trade-off between exploration (trying out each arm to find the best one) and exploitation (playing the arm believed to give the highest reward or pay-off). Each choice of an arm results in an immediate random reward, but the process determining these rewards evolves during the play of the bandit. The distinguishing feature of bandit problems is that the distribution of returns from one arm only changes when that arm is chosen. Hence the rewards from an arm do not depend on the rewards obtained from other arms. This feature also implies that the distributions of returns do not depend explicitly on calendar time.

Practical examples of the bandit problem include clinical trials where different treatments need to be experimented with while minimizing patient losses, or adaptive routing efforts for minimizing delays in a network. In an economics environment, experimental consumption is an example of intertemporal allocation problems where the trade-off between current earnings and value of information plays a key role. Alternatively, the use of arms may change their physical properties as in learning by doing where experience with the arm increases its future earnings. The following lines present a slightly more technical description of the MAB problem based on Niño-Mora [77].

Models in the MAB family are concerned with the problem of optimally allocating effort over time to a collection of projects that change state randomly depending on whether they are engaged or not. The multi-armed bandit model, in its discrete time version, can be described as follows: there is a collection of projects exactly one of which must be engaged at each discrete decision epoch, each project can be in one of a finite number of states.

Every time a project is engaged (activated) an active reward is earned and the project's state evolves following a Markovian transition rule. States of projects not engaged remain unchanged and no reward is earned under passivity. The problem consists in finding a non-anticipative scheduling policy for dynamically activating projects that maximises some measure of the total expected discounted reward earned over an infinite horizon. Some early work on the topic, after it was first formulated by Thompson [98], includes Robbins [87], Bellman [9], Gittins and Jones [37], Glazebrook [41], and Weitzman [112]. A summary of the history of the subject can be found in Yakowitz [118]. Also the monograph by Berry and Fristedt [20] gives a good account of applications of Bandit Problems.

After being considered intractable for a long time, the problem was solved in a landmark result by Gittins [35] and Gittins and Jones [38]. The optimal policy is given by Gittins priority-index rule: an index is computed for each arm (project) and state and the rule prescribes to activate at each decision epoch the arm with largest current index. The optimality of Gittins rule has been subject of deep analysis, and proofs based on different technical approaches have been provided including interchange arguments [35, 38, 105, 111], dynamic programming [113], induction arguments [108] and conservation laws/linear programming [23].

For more complex extensions, the Gittins index rule is no longer optimal. For example, the incorporation of costs/delays when switching between projects is studied in [6], where a partial characterization of an optimal index is provided. This problem is the matter of Chapter 3 of this dissertation, further references are provided therein. One more example is the extension to the basic framework where not-engaged (passive) projects continue to evolve – possibly with different transition rules– and a fixed number of projects must be engaged at each decision epoch. This extension, better known as the *Restless Bandit Problem* was first introduced by Peter Whittle [116]. This problem is discussed in Section 2.3 and is the main matter of Chapters 4 and 5.

We now focus in the MDP formulation of the *Multi-armed Bandit Problem* and its solution by means of *Dynamic Allocation (Gittins) Indices.*

2.2.1 The MAB Model Formulation

In this section we present the MDP formulation and review some basic concepts and results for the *Multi-armed Bandit Problem* (MAB).

Consider the problem of sequentially allocate effort to one out of a given collection of projects $M > Q$. Each project has state space S_i and is modelled as Markov decision process with two actions available at each decision epoch, t, and state $X_i\left(t\right) \in S_i$, $1 \leq i \leq M$: active $\left(a_i\left(t\right) = 1\right)$ or passive $\left(a_i\left(t\right) = 0\right)$ corresponding to activating the project or not, respectively. If project i (in state $x_i \in \mathcal{S}_i$ is engaged at time t, an immediate reward $R_i\left(x_i\right)$ is earned and the project state changes to some state $x_i' \in S_i$ according to some Markovian transition rule. Projects not engaged do not yield reward and remain *frozen*, i.e. their states do not change. Rewards are discounted in time by with factor $\beta \in (0, 1)$. The problem faced by the decision maker is to design a *non-anticipative scheduling policy* π (which prescribes which project to engage at each time) that maximises the total expected discounted reward earned over an infinite horizon.

Let $\mathbf{a}\left(t\right)$ be the unitary M-vector describing the action taken at decision epoch t, and $\mathbf{X}\left(t\right) = \left\{X_1\left(t\right), \ldots, X_M\left(t\right)\right\} \in \mathcal{S}$ be the state at epoch t, where $\mathcal{S} = \bigtimes_{i=1}^{M} S_i$ is the system's state space. Then, the discounted multi-armed bandit problem above can be formulated as:

$$V = \max_{\pi} E_{\pi} \left[\sum_{t=0}^{\infty} \beta^t R_{\mathbf{X}(t)}^{\mathbf{a}(t)} \right] \tag{2.12}$$

This problem is a classical one in the history of sequential decision problems and has been widely applied as a model for a variety of project scheduling and dynamic resource allocation problems.

The standard MDP formulation of the multi-armed bandit problem defined by $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, M)$ is given by the following[5]:

$1_{(\mathbf{MAB})}$. **Decision Epochs**

Decisions are taken at epochs $t \in \mathbb{N}$.

$2_{(\mathbf{MAB})}$. **State Space**

The (countable) set of all possible system states is the Cartesian product $\times_{i=1}^{M} S_i$, with S_i the state space for bandit i, $1 \leq i \leq M$. The state of the process at time t is the M-tuple $\mathbf{X}(t) = \{X_1(t), \ldots, X_M(t)\}$, with $X_i(t)$ the state of arm i at time t.

Notice that the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = \prod_{i=1}^{M} |S_i|$.

$3_{(\mathbf{MAB})}$. **Action Set**

At each decision epoch, the collection of M admissible actions at $\mathbf{X} \in \mathcal{S}$ is given by the set

$$\mathcal{A} = \left\{ \mathbf{a} = (a_1, \ldots, a_M) : \sum_{i=1}^{M} a_i = 1, a_i \in \{0, 1\} \right\} \tag{2.13}$$

[5] We include the additional element M to the collection $(T, \mathcal{S}, \mathcal{A}, \mathbf{P}^a, \mathbf{R}^a)$ (see page 20), representing the number of arms in the bandit problem. In some cases, when using discounted optimality criterion, it might also be given explicit mention to the discount factor β.

Under action $a_i = 1$, bandit i is *active* while under $a_i = 0$ it remains *passive.* Equation (2.13) indicates that an adimissible action for the process activates exactly *one* bandit, leaving the remaining $M - 1$ passive.

$4_{(\mathbf{MAB})}$. **Transition Probabilities**

When action $a_i(t) = 1$ (active) is taken at $t \in \mathbb{N}$, bandit i evolves according to Markovian law P_i, i.e.

$$P\left\{X_i(t+1) = x' | X_i(t) = x, a_i(t) = 1\right\} = P_i(x, x'),\ x, x' \in S_i$$

for all $1 \leq i \leq M$. The system's transition from state $\mathbf{X}$ under action $\mathbf{a} \in \mathcal{A}$ will be given by

$$\mathcal{P}^{\mathbf{a}}_{\mathbf{X},\mathbf{X}'} = \prod_{i=1}^{M} P_i(x, x'),\ \mathbf{X}, \mathbf{X}' \in \mathcal{S}. \tag{2.14}$$

Let $\mathbf{P}$ stand for the collection of transition matrices $\mathcal{P}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

$5_{(\mathbf{MAB})}$. **Active Rewards**

For all i, $R_i : S_i^2 \to \mathbb{R}^+$ is a bounded stationary reward function. If an active transition from state $X_i(t) = x$ to $X_i(t+1) = x'$ occurs on arm i at time t, a discounted reward $\beta^t R_i(x, x')$ is earned. Rewards are additive across bandits and over time. Passive arms yield no reward. We shall frequently use the telescoped notation

$$R_i(x) = \sum_{x' \in S_i} R_i(x, x') P_i(x, x'),\ x \in S_i,\ 1 \leq i \leq M$$

to denote the expected reward earned from a single transition under action $a_i = 1$. Further, $\beta \in (0, 1)$ is a discount rate. We finally introduce

$$\mathcal{R}^{\mathbf{a}}_{\mathbf{X}} = R_i\left(x\right) 1\left(a_i = 1\right), \ \mathbf{a} \in \mathcal{A} \text{ and } \mathbf{X} \in \mathcal{S}.$$

Notice that $1\left(a_i = 1\right)$ is an indicator function taking value 1 whenever active action is taken in arm i.

Let $\mathbf{R}$ stand for the collection of reward matrices $\mathcal{R}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

$6_{(\mathbf{MAB})}$. **Policy**

A policy π is a rule that sequentially activates one out of M arms at each decision epoch. The goal of analysis is the determination of a policy to maximise the total expected discounted reward over an infinite horizon. The theory of Dynamic Programming (see Section 2.1 and references therein) asserts the existence of an *optimal, stationary policy* which satisfies the optimality equations of DP.

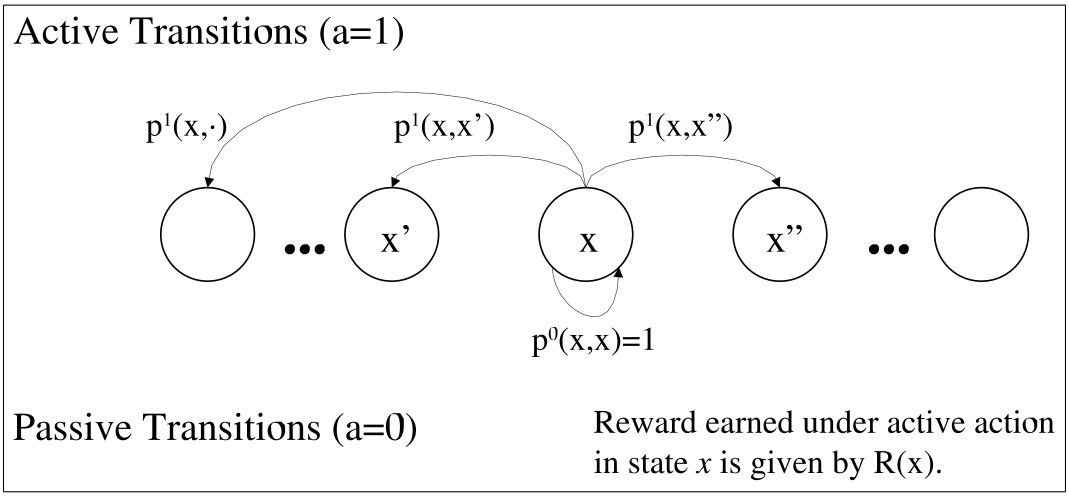
Figure 2.3 below depicts the typical evolution of an isolated arm in a multi-armed bandit problem as described above.

Let us denote by $V\left(\mathbf{X}\right)$ the optimal problem value (i.e. the maximal expected discounted reward) when the initial project state is $\mathbf{X} \in \mathcal{S}$. The optimality equations may be expressed as

$$V\left(\mathbf{X}\right) = \max_{\mathbf{a} \in \mathcal{A}} \left\{ R^{\mathbf{a}}_{\mathbf{X}} + \beta \sum_{\mathbf{X}' \in \mathcal{S}} P^{\mathbf{a}}_{\mathbf{X},\mathbf{X}'} V\left(\mathbf{X}'\right) \right\}, \ \mathbf{X} \in \mathcal{S}. \tag{2.15}$$

This DP formulation is a classical example of the *course of dimensionality*, which hinders the application of standard dynamic programming techniques: the size of the DP formulation typically grows exponentially on the size of the model's description.

The above *multi-armed bandit problem* was considered intractable for a long


Active Transitions (a=1)
p1(x,·)
p1(x,x')
p1(x,x")
x'
x
x"
p0(x,x)=1
Passive Transitions (a=0)
Reward earned under active action in state *x* is given by R(x).

When active action is taken in state x, an active reward $R(x)$ is earned and the arm evolves to some other state in $\mathcal{S}$ following some specific transition rule $P(x,\cdot)$. When passive action is taken, the arm remains *frozen* in the current state and no reward is earned. No specific order of states is assumed.

Figure 2.3: Representation of a Project in the Multi-armed Bandit Problem

time. In the early 1970's, however, Gittins and Jones [37] presented a surprising solution, which now stands as a landmark result in the field: the problem solution can be easily characterised in terms of a set of *dynamic allocation indices*, attached to project states; in particular, for each state $X_i = x \in S_i$ of each arm $1 \leq i \leq M$ there exists an index $\gamma_i(x)$ depending solely on individual arm's parameters:

$$\gamma_i(x) = \sup_{\tau>0} \left\{ \frac{E\left[\sum_{t=0}^{\tau-1} \beta^t R(X_i(t)) \,|X_i(0) = x\right]}{E\sum_{t=0}^{\tau-1} \beta^t} \right\} \tag{2.16}$$

where τ is a stopping time for the process $\{X_i(t)\}$ satisfying:

$$\tau(x) = \min\{t : \gamma(X(t)) < \gamma(x)\}. \tag{2.17}$$

An important property of the Gittins index is that the supremum in (2.16) is achieved by $\tau(x)$ in (2.17). A first-principles proof of the following Theorem (proposed by Gittins, [35]) can be found in Frostig and Weiss [33].

Theorem 2.6. ***(Gittins)*** *The supremum of* (2.16) *is achieved by* (2.17). *It is also achieved by any stopping time σ which satisfies:*

$$\sigma \leq \tau(x) \ \ \textit{and} \ \gamma(X(\sigma)) \leq v(x). \tag{2.18}$$

Then the following *larger index policy* is optimal: work at each time on the project whose current state has larger index.

Theorem 2.7. ***(Gittins)*** *There exist functions,* $\gamma_i(X_i(t))$, $1 \leq i \leq M$ *such that for any state* $\mathbf{X}(t)$ *the policy* π^*, *which will activate the bandit process (arm)* $m(t) = i$ *which satisfies* $\gamma_i(X_i(t)) = \max_{1\leq i\leq M}(X_i(t))$, *is optimal. The function* $\gamma_i(\cdot)$, $1 \leq i \leq M$ *is calculated from the dynamics of process* i *alone.*

Gittins further gave an elegant interpretation of his indices: the index γ_x^i is the *optimal reward rate* that can be obtained from project i when it initial state is $X_i(0) = x$.

This is a deep result that has been shown to have many different aspects and implications, and can be proven in several different ways. Among the alternatives to the original solution, which relied on interchange arguments, we can mention Whittle [113] who provided a proof based on dynamic programming that was subsequently simplified by Tsitsilkis [100]. Varaiya et.al. [105], and Weiss [111] derived different proofs based on interchange arguments. Weber [108] presented an intuitive proof. More recently, Tsitsiklis [101] has provided a proof based on a simple inductive argument, and Ishikida and Varaiya [55] have derived the result without making an explicit use of the interchange argument. Finally, Bertsimas and Niño-Mora provided a new proof based on a polyhedral approach to indexability [23]. A detailed discussion of some of these proofs can be found in the excellent survey by Frostig and Weiss [33].

Of particular interest is, however, the Whittle's [113] alternative dynamic programming analysis of the Gittins indices. Consider a modification of the multi-armed bandit project i, in which passivity is subsidized by a constant amount γ, so that $R_i^0(x) = \gamma$. Consider now the single-project sub-problem that involves operating optimally this modified project. Write the corresponding optimal value function as $V_i(x,\gamma)$. This satisfies the DP equation:

$$V_i(x,\gamma) = \max\left\{ R_i(x) + \beta \sum_{x' \in S_i} P_i(x,x') V_i(x',\gamma); \gamma + \beta V_i(x,\gamma) \right\}, x \in S_i.$$

Now it is intuitive that, as the subsidy for passivity, γ, gets larger, it should be less preferable to take the active action in the modified problem above. Whittle established that this indeed is the case: as γ ranges from $-\infty$ to ∞ the set of states where it is optimal –in the modified problem- to take the passive action increases from the empty set to the full project state space. The Gittins index then emerge as the corresponding breakpoint for γ: the unique value of γ that makes optimal both, the active and the passive action in state x.[6]

In the next section we briefly describe a more recent approach first introduced by Tsoucas [102] and developed by Bertsimas et.al. [23].

[6] Actually, Whittle's original elaboration was in form of a *retirement option problem* in which the operator can play the arm for as long as he wants, then retire for ever and receive a terminal reward M. The optimality equations for this problem are

$$V_i(x,M) = \max\left\{ R(x) + \beta \sum_{x' \in \mathcal{S}} p(x,x') V_i(x',M), M \right\}.$$

The discussion in these lines follows a slightly different approach that suits better the discussion in the following sections. It can also be seen that if we take $M = \frac{\gamma}{1-\beta}$, then the retirement problem has the same solution as the one given by $V_i(x,\gamma)$.

The Achievable Region Approach

Consider the M bandit system with initial state $\mathbf{X}(0) = \mathbf{x}$, and and arbitrary policy π. Then define σ_x^π as the total expected discounted number of times at which the activated arm is in state x.

Let S be the state space and $D \subseteq S$. Consider that certain arm is initially in state x (we omit the arm indicator), it is played once, and then it is played until it reaches a state in S. Then the following quantity (A_x^D) represents expected discounted time of the first entrance (passage) in S.

$$A_x^D = E\left[\sum_{t=1}^{T_x^D - 1} \beta^t | X(0) = x\right]. \tag{2.19}$$

with

$$T_x^D = \min\left\{t : t > 0, X(t) \in D \; X(0) = x\right\} \tag{2.20}$$

being the time of the first entrance in D when starting from state x.

Let now

$$b(D) = \frac{E\left[\beta^{T_{\mathbf{X}(0)}^D}\right]}{1-\beta}$$

with $T_{\mathbf{X}(0)}^D = \sum_{1 \le i \le M : x_i \notin D} T_{x_i}^D$. Then the following Theorem can be shown to be true:

Theorem 2.8. ***(Generalized Conservation Law)*** *For initial state* $\mathbf{X}(0)$, *for every policy* π *and every* $D \subseteq S$

$$\sum_{x \in D} A_x^D \sigma_{\frac{\pi}{x}} \ge b(D).$$

Equality holds if and only if $\pi : D^c \rightarrow D$, i.e. if policy π gives priority to states outside D over states in S.

According to the Generalized Conservation Law, the following linear program is a relaxation of the multi-armed bandit problem[7]:

$$\begin{aligned} &\max \sum_{x\in S} R(x)\,\sigma_x && (2.21)\\ &\text{s.t. } \sum_{x\in D} A_x^D \sigma_x \geq b(D)\,,\ D \subset S,\\ &\sum_{x\in S} A_x^S \sigma_x = b(S) = \frac{1}{1-\beta},\\ &\sigma_x \geq 0, x \in S \end{aligned}$$

Assume S is a finite state space and consider $\phi(1), \ldots, \phi(|S|)$ to be a permutation of states $1, \ldots, |S|$. Denote by ϕ the priority policy that uses this permutation order and $D_i = [\phi(i), \ldots, \phi(|S|)]\,,\ i = 1, \ldots, |S|$. Then $\phi : D_i^c \rightarrow D_i$.

It is shown in [23] that the complementary slack dual solution to the linear program (2.21) corresponding to $\mathbf{x}^\phi$ is of the form $y^D = 0,\ D \neq D_1, \ldots, D_{|S|}$ with the remaining dual variables, which can be obtained recursively, given by

$$y^{D_i} = \frac{R(i) - \sum_{j=1}^{i-1} A_i^{D_j} y^{D_j}}{A_i^{D_i}}.$$

The details of the arguments are away from the scope of this dissertation, but it can be seen that for $\mathbf{x}^\phi$ to be optimal, it is necessary that $y^{D_2} \leq 0, \ldots, y^{D_{|S|}} \leq 0$. An algorithm based on Klimov's work [65] and known as Klimov's Algorithm is then used to construct the optimal permutation ϕ.

Because $\mathbf{x}^\phi$ is optimal, the priority policy based on this permutation is

[7] It is a relaxation in the sense that any performance measure given by policy π has to satisfy the constraints of the linear program.

optimal. This policy indeed coincides with the Gittins policy and, actually, the index $\gamma\left(\phi\left(x\right)\right)=\sum_{j=1}^{i}y^{S_j}$.

Arguments around the discussion above lead to the following conclussion: the achievable performance region coincides with the feasible region of the linear program (2.21), and its extreme points coincide with the performance vectors of the priority policies.

Bertsimas et.al. [23] shown that if performance measures in stochastic dynamic scheduling problems satisfy generalized conservation laws, then the feasible region of achievable performance is an *extended polymatroid*. Moreover, optimization of a linear objective over an *extended polymatroid* is solved by an *adaptive greedy algorithm*, which leads to an optimal solution having an indexability property. As the achievable region of the multi-armed bandit problem turns out to be an *extended polymatroid*, then any objective function linear in $\mathbf{x}$ is achieved by a greedy solution. The authors also provide with an *Adaptive Greedy* algorithm for computing the Gittins indices. This algorithm is the basis of the Algorithm proposed in Section 3.2 (see Figure 3.2) for obtaining the indices in a multi-armed bandit problem when switching between arms is costly.

2.2.2 Extensions of the Multi-armed Bandit Model

Notwithstanding it is easy to write down the formula for the Gittins index and to provide a nice economic interpretation, it is normally impossible to obtain analytical expressions for the indices and, consequently, an analytical solution to the problem. One of the few settings where such solutions are possible is the continuous time bandit model where the drift of a Brownian motion process is initially unknown and learned through observations of the process. Karatzas [59] provides an analysis of this case when the volatility parameter of the process is known.

From an analytical point of view, the key property of bandit problems is that they allow for an optimal policy that is defined in terms of indices that are calculated for individual arms. One instance where such a generalisation is possible is the *branching bandits problem* where new arms are born to replace the arm that was chosen in the previous period. Glazebrook [40] obtained an index result for a model in which a collection of individual bandits are subject to a precedence relation in the form of an out-forest, while Whittle [114] considered open processes in which new bandits arrive over time. Other examples of branching bandits are Klimov [65] and Tcha and Pliska [97]. Index results for both discounted and undiscounted branching bandits were obtained by Bertsimas and Niño-Mora [23] via mathematical programming methods. More recently, Crosbie and Glazebrook [27] studied a new class of controlled stochastic systems called *generalised branching bandits* which include discounted branching bandits and generalised bandit problems as special cases.

An index characterization of the optimal allocation policy can still be obtained without the Markovian assumption. Varaiya et.al. [105] give a general characterization in discrete time, and Karoui and Karatzas [60] provide a similar result in continuous time setting. In either case, the essential idea is that the evolution of each arm only depends on the (possible entire) history and running time of the arm under consideration, but not on the realization nor the running time of other arms. Banks and Sundaram [7] show that the index characterization remains valid under some weak additional condition even if the number of indices is countable, but not necessarily finite.

On the other hand, it is well known that an index characterization is, in general, not possible when the decision maker must or can select more than one single arm at each decision epoch (see, for example Ishikida [54]). Moreover, the optimal solution for this class of problems is not generally known. Anantharam et.al. [3, 4] and Agrawal et.al. [2] determined optimal allocation schemes for multi-armed bandits with multiple plays and the *learning loss* or *regret* criterion. Pandelis and Teneketzis [81] identifies a sufficient (but not necessary) condition on the reward processes that guarantees the optimality of the index strategy (at each decision epoch to operate the arms with the largest Gittins indices).

Bergemann et.al. [17] considered a stationary setting in which there is an infinite supply of ex-ante identical arms available. Within that stationary setting, they show that an optimal policy follows the index characterization even when many arms can be selected at the same time or when a switching cost has to be paid to move from one arm to another. Banks and Sundaram [8] further show that, in a more general setting, an index characterization is not possible when an extra cost must be paid to switch between arms in consecutive periods. This topic is the matter of Chapter 3, and we hence leave further discussion for later. However, the interested reader is referred to Jun [57] for a well annotated survey on the multi-armed bandit problem with switching costs.

Nash [71] extended the multi-armed bandit by introducing a structure in which the rewards of a particular arm are influenced by the states of other bandits. This analysis has been helpful in the analysis of a range of problems in research planning and stochastic scheduling (see, for example, Fay and Glazebrook [29, 30] and Glazebrook and Greatrix [47]). Development of Nash work has been restricted to the special case in which all the indices are positive, Fay and Walrand [31]; however, Glazebrook and Greatrix [48] have shown how to modify the structure of a generalised bandit problem so that it is reduced to the special case above while leaving the optimal policy unchanged.

Arguably the most promising extension of the classical multi-armed bandit problem us the so-called *restless bandit problem* pioneered by Peter Whittle [116]. In the restless bandit problem a fixed number of projects must be engaged at each time and passive projects can change state. A more detailed overview of this problem is given in Section 2.3 and three broad families of *restless bandit problems* are the matter of Chapters 4 and 5. Moreover, assumptions of the *restless bandit problem* constitute the building blocks for our solution to the multi-armed bandit problem when costs are incurred when switching from one arm to the other in Chapter 3.

2.2.3 Applications of the Multi-Armed Bandit Problem

There is a vast literature with applications of the multi-armed bandit and its variants to the modelling of decision problems in a variety of fields like job scheduling, resource allocation, sequential random sampling, clinical trials, investment in new products, random search, etc.

Some early examples of applications of the MAB framework can be found in Glazebrook [42, 43, 45, 44]; Nash and Gittins [72]; Rodman [89]; Wahrenberger et.al. [107]; and Whittle [115] and references therein. Bery and Fristedt [20] provide a good set of examples of applications to sequential design of experiments and Gittins [36] gives a detailed account of applications to stochastic scheduling problems.

Interesting economic applications can be found in fields as *labour economics*, see for example Johnson [56], Miller [69], McCall et.al. [68], and Kennan et.al. [64]; *optimal search*, Weitzman [112] and Roberts and Weitzman [88], Smith [96], Benkherouf and Bather [12], Benkherouf [11], and Benkheruof et.al. [13] are good examples in this field; and *game theory*, see for example Schlag [94] and Brenner and Vriend [25]. In his complete survey Jun [57] provides an annotated discussion of these and other economic applications of the multi-armed bandit problem.

Other examples of applications of the multi-armed bandit framework in economics of learning and experimentation are *market learning* (see Rothschild [91], Keller et.al. [63], Rustichini and Wolinski [93]); *experimentation and pricing* (see, for example, Bergemann et.al. [16, 19], Felli et.al. [32], and Bolton et.al. [24]); *experimentation in finance* (see Bergemann et.al. [14, 15], and Hong et.al. [53]). A detailed discussion of these and some other applications can be found in the survey on bandit problems by Bergemann et.al. [18].

2.3 The Restless Bandit Problem

One of the most important assumptions in the multi-armed bandit model is the fact that idle projects are assumed to remain *frozen* in their current state during to passive sojourn. This assumption, as a matter of fact, does not hold in many cases and, consequently, limits the applicability of the multi-armed bandit framework to a wide variety of problems.

In his pioneering work, Whittle [116] provides some examples of cases for which projects continue to change state even when they are passive (not operated). A classical example is the one given by Whittle (see [116], page 288).

> ... suppose m aircraft are trying to track the positions of n enemy submarines, where $m < n$, so that aircraft must change task from time to time if all submarines are to be monitored... The problem is to allocate this surveillance... While a submarine is under observation, information on its position... is being gained. While it is not, information is usually being lost, because the submarine

> will certainly be taking unpredictable evasive action.

Whittle gives one more example that, actually, constitutes the basis of the research in Chapter 5 – Section 5.2 in particular– (see [116], pp. 288.289).

> ... suppose that one has a pool of n employees of whom exactly m are to be set to work at a given time... One can imagine tht employees who are working produce, but at a decreasing rate as they tire. Employees who are resting do not produce, but recover. The 'project'... is thus changing state whether or not he is at work

Another example is the scheduling of medical screening and treatment. When screened, information about patient's condition is acquired and the treatment can be adjusted. When it is not, information is lost because the health of the patient can improve or worsen. Again, the number of screening units is limited and decision must be taken about the patients to be scheduled every decision epoch.

More examples can be found in a variety of applications as investment in assets, research and development, training, learning, equipment updating and replacement, product improvement and advertising, deployment and/or exploitation of assets, fallow and cultivation models, appointment or maintenance scheduling, and so on and so forth.

In what follows we consider an extension of the multi-armed bandit problem in which Q out of M projects must be selected to operate at each decision epoch. It is assumed that both active and passive projects evolve between decision epochs, according to corresponding active and passive transition rules. We shall reffer to this as the *Restless Bandit Problem.*

The aim of this section is, hence, to describe discounted *restless bandit problems* and a notion of *indexability* due to Whittle [116]. For *restless bandit* *problems* passing certain *indexability test*, Whittle's index heuristic will then be described.

A discounted (reward-based) restless bandit problem is an *infinite-horizon discounted criterion problem* on Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, M, Q)$ with the following elements (corresponding to §1 to §6 in Section 2.1.1)[8]:

$1_{(\mathbf{RB})}$. **Decision Epochs**

Decisions are taken at epochs $t \in \mathbb{N}$.

$2_{(\mathbf{RB})}$. **State Space**

The set of all possible system states is the Cartesian product $\mathcal{S} = \bigtimes_i^M S_i$ with S_i the state space for bandit i, $1 \leq i \leq M$. The *state* of the process at time t is $\mathbf{X}(t) = \{X_1(t), \ldots, X_M(t)\}$ with $X_i(t) \in S_i$, the state of bandit i at t.

$3_{(\mathbf{RB})}$. **Action Set**

We assume $M > Q$. At each decision epoch the collection of $\binom{M}{Q}$ admissible actions at state $\mathbf{X} \in \mathcal{S}$ is given by the set:

$$\mathcal{A} = \left\{ \mathbf{a} = (a_1, \ldots, a_M) \middle| \sum_{i=1}^{M} a_i = Q, a_i \in \{0, 1\} \right\} \tag{2.22}$$

Under action $a_i = 1$, bandit i is *active* while under $a_i = 0$ it is *passive.* Equation (2.22) indicates that an admissible action for the process activates exactly Q bandits, while leaving the remaining $M - Q$ passive.

$4_{(\mathbf{RB})}$. **Transition Probabilities**

Suppose action $\mathbf{a}(t)$ is taken at $t \in \mathbb{N}$, where $\mathbf{a}$ is a vector satisfying

[8] We add one more element Q to the collection $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, M)$ (see page 39) representing the number of arms to activate every decision epoch. As before, in some cases we will include explicit mention of the discount factor, β.

the r.h.s. of the expression between brackets in (2.22). For $a_i(t) = 1$, bandit i evolves according to Markov law P_i^1,i.e.

$$P\{X_i(t+1) = x' \,|X_i(t) = x, 1\} = P_i^1(x, x'), \ x, x' \in S_i$$

For $a_i(t) = 0$, bandit i evolves according to Markov law P_i^0,i.e.

$$P\{X_i(t+1) = x' \,|X_i(t) = x, 0\} = P_i^0(x, x'), \ x, x' \in S_i$$

The M bandits evolve independently. We can finally introduce the following notation

$$\mathcal{P}^{\mathbf{a}}_{\mathbf{X},\mathbf{X}'} = \prod_{i=1}^{M} P_i^{a_i}(x, x'), \quad \text{for all } \mathbf{X}, \mathbf{X}' \in \mathcal{S}, \ \mathbf{a} \in \mathcal{A}$$

for the probability of evolving fron state $\mathbf{X}$ to state $\mathbf{X}'$ under action $\mathbf{a}$.

Let $\mathbf{P}$ represent the collection of transition matrices $\mathcal{P}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

$5_{(\mathbf{RB})}$. **Active and Passive Rewards**

For all i, $R_i^{a_i} : S_i^2 \rightarrow \mathbb{R}^+$ is a bounded reward function. If a transition from x to x' occurs in bandit i under action a_i at time t a discounted reward $\beta^t R_i^{a_i}(x, x')$ is earned. Rewards are additive across bandits and over time. We shall frequently use the telescoped notation:

$$R_i^{a_i}(x) \equiv \sum_{x' \in S_i} R_i^{a_i}(x, x') \, P_i^{a_i}(x, x'), \ x, x' \in S_i, \ 1 \leq i \leq M$$

to denote the expected reward earned from a single transition under action a_i. Further, $\beta \in (0, 1)$ is a discount rate. As in the previous case, we can introduce the following notation:

$$\mathcal{R}^{\mathbf{a}}_{\mathbf{X}} = \sum_{i=1}^{M} R_i^{a_i}(x_i)\,, \quad \text{for all } \mathbf{X} \in \mathcal{S},\, \mathbf{a} \in \mathcal{A}$$

for the aggregated reward earned when taking action $\mathbf{a}$ in state $\mathbf{X}$.

Let $\mathbf{R}$ represent the collection of all reward vectors $\mathcal{R}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

$6_{(\mathbf{RB})}$. **Policy**

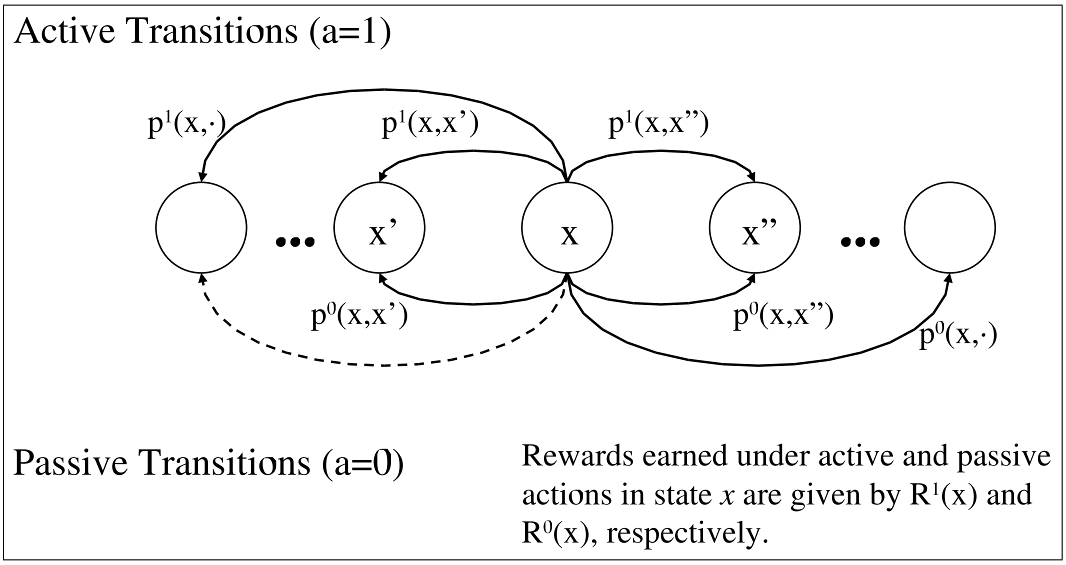
A policy π is a rule for taking actions at each decision epoch. Such rule can in principle be a function of the entire history of the process (actions taken, states occupied) to date. The goal of analysis is the determination of a policy to maximise total expected discounted reward over an infinite horizon. The theory of Dynamic Programming (DP) (see, for example, Puterman [85] and the discussion in Section 2.1) asserts the existence of an optimal (reward maximising) policy which is *stationary, deterministic and Markovian* and which satisfies the optimality equations of DP in (2.8).

Figure 2.4 below depicts the typical evolution of an isolated arm in a restless bandit problem as described above.

We write $V(\mathbf{X}, \beta)$ for the *value function* of the process evaluated at $\mathbf{X} \in \mathcal{S}$, namely the maximal expected discounted reward earned over an infinite horizon from initial state $\mathbf{X}$. The optimality equations (2.8) may be expressed as

$$V(\mathbf{X}, \beta) = \underset{\mathbf{a}\in\mathcal{A}}{max} \left(\mathcal{R}^{\mathbf{a}}_{\mathbf{X}} + \beta \sum_{\mathbf{X}' \in \mathcal{S}} \mathcal{P}^{\mathbf{a}}_{\mathbf{X},\mathbf{X}'} V(\mathbf{X}') \right), \ \mathbf{X} \in \mathcal{S} \tag{2.23}$$

Equation (2.23) notwithstanding, a pure DP approach is unlikely to yield insight and will be computationally intractable for problems of reasonable size. Hence the primary question is for good *heuristic policies*.


Active Transitions (a=1)
p¹(x,·)
p¹(x,x')
p¹(x,x")
x'
x
x"
p⁰(x,x')
p⁰(x,x")
p⁰(x,·)
Passive Transitions (a=0)
Rewards earned under active and passive actions in state x are given by R¹(x) and R⁰(x), respectively.

When active action is taken in state x, an active reward $R^1(x)$ is earned and the arm evolves to some other state in $\mathcal{S}$ following some specific *active* transition rule $P^1(x,\cdot)$. When passive action is taken, the arm evolves accordingly with the *passive* transition law $P^0(x,\cdot)$ and earns passive reward $R^0(x)$. No specific order of states is assumed.

Figure 2.4: Representation of a Project in the Restless Bandit Problem

The sub-class of models for which $Q = 1$ and $P_i^0(x,x') = 0$, $x \neq x'$ ($P_i^0(x,x) = 1$), $1 \leq i \leq M$ (i.e. only one arm activated at each decision epoch and no state evolution under the passive action) are known as *multi-armed bandit problems* (MABs) and are the subject of Section 2.2. Gittins ([35] and [36]) famously demonstrated the optimality of *index policies* for MABs, namely that there exist calibrating index functions $G_i : \mathcal{S}_i \rightarrow \mathbb{R}$ (one for each bandit), such that at t the bandit for optimal activation is the one whose associated index $G_i(x_i(t))$ is maximal. In the event of ties, all corresponding choices are optimal.

Building on this classical result, Whittle [116] proposed a class of index heuristics for those restless bandit problems which pass an *indexability test.* These heuristics emerge naturally from a Lagrangian relaxation of the original optimization problem, see Section 2.3.1. The indices which result from Whittle's analysis generalise those from Gittins. We now outline the key notions in Whittle's approach and develop the framework over which the discussion in the next chapters will be built.

Indexability and indices are properties of individual bandits. Hence in the MDP §1$_{(\mathbf{RB})}$ to §6$_{(\mathbf{RB})}$ above, we isolate an individual bandit and will drop the bandit identifier i. In Whittle's analysis, this bandit generates a collection of MDPs parametrised by a *passive subsidy* W, where $W \in \mathbb{R}$. We shall refer to this as the W-*subsidy* problem for bandit $(S, \mathcal{A}, P, R, \beta)$. This is a discounted reward MDP as follows:

1'$_{(\mathbf{RB})}$. **Decision Epochs**

Decisions are taken at times $t \in \mathbb{N}$.

2'$_{(\mathbf{RB})}$. **State Space**

The countable state space is S. We use $X\left(t\right)$ for the state of the process (arm) at time t.

3'$_{(\mathbf{RB})}$. **Action Set**

At each decision epoch t, either action $a = 1$ (active) or $a = 0$ (passive) is applied to the process (arm).

4'$_{(\mathbf{RB})}$. **Transition Probabilities**

If $a = 1$ is chosen at t then evolution is according to P^1 with

$$P\left\{x\left(t+1\right) = x' \left| x\left(t\right) = x, a\left(t\right) = 1\right.\right\} = P^1\left(x, x'\right),\ x, x' \in S,$$

if otherwise $a = 0$ is chosen at t then evolution is according to P^0 with

$$P\left\{x\left(t+1\right) = x' \left| x\left(t\right) = x, a\left(t\right) = 0\right.\right\} = P^0\left(x, x'\right),\ x, x' \in S.$$

Let P stand for the collection of passive and active transition matrices $\{P^0, P^1\}$.

5'$_{(\mathbf{RB})}$. **Active and Passive Rewards**

If a transition from x to x' occurs under action $a = 1$ at time t a discounted reward $\beta^t R^1 (x, x')$ is earned. Should a transition from x to x' occur under action $a = 0$ at time t a discounted reward $\beta^t \{R^0 (x, x') + W\}$ is earned, where W is the passive subsidy.

Let R stand for the collection of passive and active reward vectors $\{R^0, R^1\}$.

6'$_{(\mathbf{RB})}$. **Policy**

The goal of optimization is the choice of a policy to maximise the total expected discounted reward (including passive subsidies) earned over an infinite horizon. We assert the existence of optimal policies for the *W-subsidy* problem which are stationary and whose value functions satisfy the optimality equations of DP. We shall restrict to stationary policies throughout.

We use $V(x, W)$ for the value function for the *W-subsidy* problem evaluated at $x \in S$. The DP optimality equations may be expressed as:

$$V(x, W) = \max \left\{ R^1 (x) + \beta \sum_{x' \in S} P^1 (x, x') V (x', W) \; ; \right. \\ \left. R^0 + W + \beta \sum_{x' \in S} P^0 (x, x') V (x', W) \right\} \quad (2.24)$$

for all $x \in S$.

The active action is optimal in x when the first term in $\max\{;\}$ on the r.h.s. of (2.24) achieves the maximum and the passive action is optimal when the second term does so. We use $\Pi(W)$ for the subset of S for which the passive action is optimal under subsidy W, namely

$$\Pi(W) = \left\{ x \in S : R^0(x) + W + \beta \sum_{x' \in S} P^0(x, x') V(x', W) \right. \\ \left. \geq R^1(x) + \beta \sum_{x \in S} P^1(x, x') V(x', W) \right\}. \quad (2.25)$$

Definition 2.1, based on Whittle's definition of indexability (see Definition 2.3), describes the indexability test for bandit $(S, \mathcal{A}, P, R, \beta)$ and the restless bandit problem of which it is a part.

Definition 2.1. *Bandit $(S, \mathcal{A}, P, R, \beta)$ is indexable if $\Pi(W)$ is increasing in W, namely*

$$W_1 \geq W_2 \Rightarrow \Pi(W_1) \supseteq \Pi(W_2).$$

A restless bandit is indexable when each of its constituent bandits is indexable.

Hence, a restless bandit is indexable if, as the level of passive subsidy increases, then so does the collection of states for which the passive action is optimal. However plausible and natural this requirement may appear, it is typically very challenging to establish and sometimes fails to hold.

Definition 2.2. *If bandit $(S, P^1, P^0, R^1, R^0, \beta)$ is indexable then its Whittle index $W : S \to \mathbb{R}$ is given by*

$$W(x) = \inf\{W : x \in \Pi(W)\}, x \in S.$$

Note that the assumed boundedness of rewards guarantees that the Whittle index must also be bounded. The value $W(x)$ represents a *fair subsidy* in state x in the sense that it renders both actions (active, passive) optimal in the W-

subsidy problem. The details of Whittle's approach will be briefly discussed in the subsequent sections.

Whittle Index Heuristic

> If we restore the bandit identifiers and consider an indexable restless bandit problem with $W_i : S_i \to \mathbb{R}$ the Whittle index for bandit i, $1 \leq i \leq M$, then the *Whittle index heuristic* operates as follows: at each time $t \in \mathbb{N}$ apply the active action to the Q bandits with largest $W_i\big(x_i\left(t\right)\big)$ and the passive action to the remaining $M - Q$ bandits.

Comments

A. We could equivalently define a W-*charge* problem (plainly W^+-*problem*) in which §5'$_{(\mathbf{RB})}$ in page 58 is replaced by the following:

   5'$_{(\mathbf{RB})}$. If a transition from x to x' occurs under action $a = 1$ at time t a discounted reward $\beta^t \left\{R^1\left(x, x'\right) - W\right\}$ is earned, where W is a charge for activity. Should a transition from x to x' occur under action $a = 0$ at time t a discounted reward $\beta^t R^0\left(x, x'\right)$ is earned. Plainly, the W-*subsidy* and W-*charge* problems are equivalent in the sense of having identical optimal policies. The value functions differ by $W\left(1 - \beta\right)^{-1}$ in all states.

B. We can, of course, have cost-based restless bandit problems of the form $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{C}, M, Q)$. These are as in §1$_{(\mathbf{RB})}$-§6$_{(\mathbf{RB})}$ in page 53, bar the fact that the bounded rewards R_i^0, R_i^1 in §5$_{(\mathbf{RB})}$ are replaced by bounded costs C_i^0, C_i^1. The goal of analysis is now the determination of policies to minimise the total expected cost incurred over an infinite horizon. The development of a W-*subsidy* problem is as in §1'$_{(\mathbf{RB})}$-§6'$_{(\mathbf{RB})}$ in page 57, except now a passive subsidy W always reduces the instantaneous cost incurred under the passive action by that amount. The corresponding optimality equation replaces (2.24) by

$$V(x,W) = \min \left\{ C^1(x) + \beta \sum_{x' \in S} P^1(x,x') V(x',W) ; \right.$$
$$\left. C^0(x) - W + \beta \sum_{x' \in S} P^0(x,x') V(x',W) \right\} \qquad (2.26)$$

Definitions 2.1 and 2.2 remain unchanged.

As in §A above, we can equivalently define a W-*charge* problem in the obvious way. Through this chapter we shall use the convention that cost-based restless bandit problems will be analised via corresponding W-*charge* problems -thus yielding a wholly cost-based decision structure. Similarly, reward-based restless bandit problems (or those which are primarily so) will be analysed via corresponding W-*subsidy* problems.

2.3.1 Whittle's Lagrangian Relaxation

In the original Whittle's [116] formulation, each of the restless bandits considered is a class of *Markov Decision Processes* with the *average reward criterion.* As this formulation can be straightforwardly extended to the discounted reward case, we outline here the main results of Whittle's analysis for the average reward case.

In a typical restless bandit problem, M projects (bandits or arms) are available for activation (exploitation, investment, and so on). Resource constraint, $m(t) = Q$, means that only Q bandits $1 \leq Q \leq M$ may be active at any time. Each of the projects evolves stochastically through time as described in §1'$_{(\mathbf{RB})}$ to §6'$_{(\mathbf{RB})}$ in page 57. The decision problem concerns how projects should be optimally chosen for activation at each decision epoch of the system to maximise the average reward earned over an infinite horizon.

We write $\mathcal{U}$ for the class of stationary, deterministic and Markovian policies for an identified member the restless bandits family and $u \in \mathcal{U}$ for an individual policy. We use r_i^u for the average reward rate earned by asset i under policy u. The optimization problem of interest is expressed as

$$r^{\text{opt}} = \max_{u \in \mathcal{U}} \left\{ \sum_{i=1}^{M} r_i^u \right\} \tag{2.27}$$

We now relax the optimization problem in (2.27) by considering schemes which activate *any number of assets* at each decision epoch (i.e. any number between 0 and M, not necessarily Q) and use $\mathcal{U}'$ for the policies which do this in a stationary, deterministic and Markovian way. Our interest will reside in those members of $\mathcal{U}'$ which activate Q assets (equivalently, fail to activate $M - Q$ assets) *on average* over an infinite horizon. To formulate the corresponding optimization problem, write T_i^u for the proportion of time for which asset i is passive under $u \in \mathcal{U}'$. Hence we relax (2.27) to

$$r^{\overline{\text{opt}}} = \max_{u \in \mathcal{U}'} \left\{ \sum_{i=1}^{M} r_i^u \right\} \tag{2.28}$$

subject to

$$\sum_{i=1}^{M} T_i^u = M - Q \tag{2.29}$$

Plainly, the relaxation yields increased optimal rewards (compared to the original problem) and hence $r^{\overline{\text{opt}}} \geq r^{\text{opt}}$.

We now incorporate constraint (2.29) in a Lagrangian fashion. We write

$$r(W) = \max_{u \in \mathcal{U}'} \left[\sum_{i=1}^{M} \{ r_i^u + W T_i^u \} - W(M-Q) \right] \tag{2.30}$$

$$= \sum_{i=1}^{M} \left[\max_{u_i \in \mathcal{U}_i'} \{ r_i^u + W T_i^u \} \right] - W(M-Q) \tag{2.31}$$

where W is a *Lagrange multiplier* which has an economic interpretation as a *subsidy for passivity.* The additive nature of the objective in (2.30) together with the character of policy set $\mathcal{U}'$ means that the optimal activation scheme for the entire set of assets is achieved by concatenating optimal activation schemes for the individual assets. The additive decomposition in expression (2.31) is the consequence. In (2.31), $\mathcal{U}_i'$ is the set of stationary, deterministic and Markovian policies which choose between actions $a = 1$ and $a = 0$ for asset i (alone), $1 \le i \le M$. The optimization problem

$$r_i(W) = \max_{u_i \in \mathcal{U}_i'} \{ r_i^u + W T_i^u \} \tag{2.32}$$

is called the W-*subsidy problem* for asset i and aims to choose a policy for activating i to maximise its overall return from rewards earned and passive subsidies received. Since expressions (2.28) and (2.30) are equal when constraint (2.29) is satisfied, it is plain that $r(W) \ge r^{\overline{\text{opt}}} \ge r^{\text{opt}}$ for all W.

An issue which arises in consideration of the W-*subsidy* problem in (2.32) is the possible non-uniqueness of the policy(ies) achieving the maximum. By defining $\Pi(W)$ as the *passive set* of policy $u_i(W)$ satisfying (2.32) (i.e. the largest set of states in which the corresponding policy chooses the passive action), we can resolve any non-uniquenes by choosing the policy with the largest *passive set.*

Use $u(W)$ for the stationary policy for the entire system which applies

$u_i(W)$ to each asset i, $1 \leq i \leq M$. Policy $u(W)$ solves the optimisation problem (2.30). The following definition of indexability (adapted from [116], Definition 1) expresses a natural requirement on (optimal) policy structure.

Definition 2.3. ***(Whittle)*** *Let $\Pi_i(W)$ be the set of values of $x_i \in S_i$ for which project i would be rested under a W-subsidy policy. Then the project is indexable if $\Pi_i(W)$ increases monotonically from $\emptyset$ to S_i as W increases from $-\infty$ to ∞.*

The above decision problem (2.28) *is indexable when all constituent projects are.*

Should an asset be indexable, then a natural calibration, in the form of a *fair subsidy for passivity* may be defined. Whittle ([116], page 290) defines the index $W_i(x_i)$ of project i in state $x_i \in S_i$ as the

> value of subsidy which should make the two phases [activity and passivity] equally attractive for project i in state x.

Definition 2.2 is a natural extension of this comment.

It now follows that $u(W)$ will choose to activate in system state $\mathbf{X} \in \mathcal{S}$ those assets for which $W_i(X_i) > W$ and apply the passive action to the reminder. Whittle presents this result in the following terms (see [116], Proposition 3):

Proposition 2.2. ***(Whittle)*** *If all projects are indexable, then the projects i which are active under a W-subsidy policy are those for which $W_i(x_i) > W$.*

If, further, there exists W^* such that $u(W^*)$ satisfies (2.29) then it must follow that

$$r(W^*) = r^{\overline{\text{opt}}} \geq r^{\text{opt}}$$

and $u(W^*)$ solves the relaxation in (2.28) and (2.29).

A natural *index heuristic* for the original optimization problem (2.27) emerges from the above discussion. The heuristic chooses in state $\mathbf{X} \in \mathcal{S}$ to activate Q assets with maximal index values $W_i\left(X_i\right),\ 1 \leq i \leq M$, with ties resolved in some arbitrary manner. Notice that this policy imposes the rigid constraint $m\left(t\right) = Q$. By denoting the return of such a policy by r^{ind}, Whittle proves the following proposition ([116], Proposition 5):

Proposition 2.3.

$$r^{ind}\left(Q\right) \leq r^{opt}\left(Q\right) \leq r^{\overline{opt}}\left(Q\right).$$

In page 55 we mentioned that for the case where $Q = 1$ and $P_i\left(x_i, x_i'\right)$, $x_i \neq x_i'$, $1 \leq i \leq M$, the restless bandit problem reduces to the classic multi-armed bandit problem. This result is established in Whittle's Proposition 2 [116].

Proposition 2.4. ***(Whittle)*** *The index $W_i\left(x_i\right)$ reduces to the Gittins index in the case $P_i^0\left(x_i, x_i'\right) = 1\left(x_i = x_i'\right),\ 1 \leq i \leq M$, where $1\left(\cdot\right)$ is an indicator function, and $R_i^0\left(x_i\right) = 0, \forall x_i \in S_i,\ 1 \leq i \leq M$.*

Finally, the issue of indexability arises naturally from the characteristics of the restless bandit problem and can not be taken for sure. The following result corresponds to Proposition 4 in Whittle [116]:

Proposition 2.5. ***(Whittle)*** *Projects are always indexable if $P_i^0\left(x_i, x_i'\right) = 1\left(x_i = x_i'\right),\ 1 \leq i \leq M$ (i.e. if resting projects are static). They are not necessarily indexable otherwise.*

Comment

All the results so far can be also obtained for the discounted case if we substitute

(2.27) by

$$r_{\beta}^{\text{opt}} = \max_{u \in \mathcal{U}} E_u \left[\sum_{t=0}^{\infty} \sum_{i=1}^{M} \beta^t r_i^u (t) \right] \tag{2.33}$$

conditional on a distribution $\mathbf{P}_i$ over initial state $X_i(0)$, so that the expectation in (2.33) is taken over both initial state and subsequent evolution of the system. And condition (2.29) is replaced by

$$\sum_{i=1}^{M} \frac{M-Q}{1-\beta} T_i^u (\beta), \tag{2.34}$$

where

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

and $1(a_i(t) = 0)$ is an indicator function which takes value 1 when passive action is taken in arm i at decision time t.

Asymptotic Optimality of Whittle's Index Policy

Starting from the interpretation of index $W(\cdot)$ as the Lagrangian multiplier associated with a prescribed average utilization rate, and given the fact that this point of view yields an immediate upper bound on performance, the bound $\overline{r^{\text{opt}}}(Q)$; Whittle [116] conjectures that, if all projects are indexable, the index policy is asymptotically optimal in terms of the average yield per project, when the ratio Q/M remains fixed.

Weber and Weiss [109, 110] investigated the asymptotic optimality of Whittle's heuristic. Working with continuous-time restless bandits with infinite horizon time-average criterion, they show that as Q and M tend to ∞ with the ratio $\alpha = Q/M$ fixed, the per-project reward of the optimal policy is asymptotically the same as that achieved by a policy operateing under the Whittle's relaxed constraint (keep active Q projects in average). Moreover, they presented a sufficient condition for asymptotic optimality, i.e. Whittle's conjecture will be true if the differential equation describing the fluid approximation to the index policy has a globally stable equilibrium point. They also found instances that violate this condition, and in which Whittle's heuristic is not asymptotically optimal. However, empirical evidence suggests that such counterexamples are rare in applied work and that the expected size of the suboptimality tends to be negligible.

Complexity of the Restless Bandit problem

Another line of work has sudied the computational complexity of the restless bandit problem. Papadimitriou and Tsitsiklis [83] established that the restless bandit problem is PSPACE-hard (PSPACE-complete), even in the special case of deterministic transition rules and $M = 1$.

> Given that the multi-armed bandit problem is the main tool for solving the few cases of networks of queues that we *can* solve, it is interesting to study the complexity of its most promising extension, the *restless bandit problem*... We show that this problem is also PSPACE-complete, even in the deterministic case.

The fact that a problem is PSPACE-complete is an even stronger indication that it is *intractable*[9] than if it where NP-complete[10]. It is also considered as evidence that the problem is not in NP or even in the polynomial hierarchy[11].

[9] A problem i intractable if it is so hard that no polynomial time algorithm can possibly solve it.

[10] NP refers to the class of decision problems that can be solved in polynomial time by a non-deterministic computer (in *computer science* a decision problem is one which solution is either yes or not). The class of NP-complete problems consist of the "hardest" problems in NP to which any other "hard" problem can be reduced (or proved equivalent in difficulty).

[11] For further discussion on complexity the reader is referred to Garey and Johnson [34],

This result is in sharp contrast with the well-known optimality of Gittins index rule in the special case of the multi-armed bandit problem. It, therefore, confirms the need for well grounded heuristics for finding near optimal (or asymptotically optimal) solutions to the *restless bandit problem*, as well as emphasises the relevance of Whittle's contribution.

2.3.2 An Alternative Approach: PCL-Indexability

Niño-Mora [75] maps out an alternative route to the demonstration of indexability for restless bandit problems and to index calculations which is based on the notion of *parcial conservation laws*, PCL-indexability. This is in turn a development of ideas based on *generalised conservation laws* (GCL) which played a fundamental role in the account of Gittins indexation given by Bertsimas et.al. [23]. PCL's are shown to imply the optimality of index policies with a *postulated structure* in stochastic scheduling problems, under *admissible linear objectives*, and are deployed to obtain sufficient conditions for indexability.

In brief, let us suppose that we which to schedule a stochastic system which is servicing a countable collection of job classes indexed by the natural numbers, $\mathbb{N}$. Denote $\mathcal{U}$ the collection of admissible scheduling policies. The stochastic optimisation problem of interest is the minimisation of some linear objective

$$\sum_{i \in \mathbb{N}} c_i x_i^u \tag{2.35}$$

where $c_i > 0$ is a cost rate for job class i and x_i^u a performance measure for class i under scheduling policy $u \in \mathcal{U}$. When the system satisfies a collection of so-called partial work conservation laws (PCL) then the stochastic optimisation

Papadimitriou [82], and Lewis and Papadimitriou [67].

problem above is solved by an index policy for *some* choices of the cost rate vector $\mathbf{c}$. Whether a particular choice is in this admissible class or not may be determined by running an adaptive greedy algorithm. A system which satisfies PCL and whose cost-rate vector $\mathbf{c}$ is in the admissible class is called PCL-indexable.

Niño-Mora [73] utilises the above ideas to develop sufficient conditions for the (Whittle) indexability of countable states restless bandits in terms of model parameters. Her further demonstrates that the restless bandit model associated with a multiclass $M/M/1$ queueing system does indeed satisfy these sufficient conditions and hence meets the requirements for PCL-indexability.

Niño-Mora [76] extends the previous work on PCL-indexability by developing a polyhedral approach to the design, analysis and computation of dynamic allocation indices for the scheduling of restless bandits based on *partial conservation laws.* In this work, the author develops a polyhedral foundation of the PCL framework, based on structural and algorithmic properties of a new polytope associated with a so-called *accesible set system*; presents new dynamic allocation indices for restless bandits motivated by an admission control model and deploys PCL's for obtaining both sufficient conditions for the existence of the new indices (PCL-indexability) and a new adaptive greedy algorithm. Niño-Mora provides a new interpretation of PCL-indexability as a form of the economics law of *diminishing marginal returns* and characterise the index as an *optimal marginal cost rate.* The author finally works out examples in queueing systems and job scheduling.

Finally, Niño-Mora, [74] addresses issues of interpretation and application of the PCL-indexability conditions. In particular, he performs an analysis of a discounted version of the so-called Ehrenfest model for optimal work allocation under tiring and recovery introduced by Whittle [116], and establish that it is PCL-indexable under natural assumptions on model parameters. To our knowledge, this is the only time –apart from Whittle's work- that this problem has been addressed in restless bandit literature and constitutes a particular case of the Squad System model analysed in Chapter 5.2 of this dissertation.

Although the analysis is complex, PCL-indexability is an important analytical tool which is sometimes available when the simple direct arguments of dynamic programming theory, as used in this work, are not.

2.3.3 Applications of the Restless Bandit Problem

The *restless bandit problem* provides a powerful modelling framework that has found applications in fields as diverse as job and production scheduling (Veatch and Vein [106], and Glazebrook and Mitchell [50]), control of queueing systems (Glazebrook et.al. [52, 49], Ansell et.al. [5], Niño-Mora [73, 79, 78]), machine maintenance (Glezebrook et.al. [51]), outsourcing of warranty repairs (Glazebrook et.al. [80]), routing (Glazebrook and Kirkbride [39]). Apart from their intrinsic interest, these applications of Whittle's ideas have provided empirical evidence of outstanding performance of the index heuristics concerned.

Chapter 3

A Restless Bandit Approach to Stochastic Scheduling Problems with Switching Costs[1]

Introduction

A general assumption maintained in almost all the work in the area of optimal resource allocation is that the operator can switch instantaneously from one project to another without facing any cost. In reality, when the manager switches between different projects a set-up may be needed, and a cost and/or delay is incurred. Although it is realistic to include a penalty each time a new project is engaged, its inclusion drastically changes the nature of the problem. In fact, it has been shown (see [8]) that, in general, it is not possible to construct indices –defined in terms of individual projects– which have the property that the resulting index strategy is optimal on the domain of all multi-armed bandits

[1] The main results in this Chapter will appear published in Section 4 of Glazebrook, K., Ruiz-Hernández, D. and Kirkbride, C. Some indexable families of restless bandit problems. *Advances in Applied Probability 38-3* (2006).

(MAB) with switching costs. Indeed, this result remains true even if attention is restricted to the case in which the cost of switching is a given (nonzero) constant [2]. So far, the problem remains unsolved and there is still the need of well grounded heuristics providing an efficient solution to this sort of models.

This section is based on the observation that by means of a straightforward manipulation of the reward structure, the switching cost problem can be reduced to the *pure set-up cost* problem and, consequently, indices defined in terms of isolated projects can be obtained. Moreover, with a simple redefinition of the state space (first proposed by Asawa and Teneketzis [6]) an index policy can be formulated for the *pure set-up cost* problem which is equivalent to the Whittle index-policy for the non-penalty case. This redefinition consists in including, for each state and project, information related to the action (active or passive) taken in the previous decision epoch. It can be shown that under this formulation, a MAB with switching costs behaves as a restless bandit (RB) without them and all results available in literature for the latter apply for the switching-cost case[3]. Performance of such index policy is empirically shown to be very close to optimality.

The underlying idea is to include for each bandit i and state x_i (for $1 \leq i \leq M$), information about the action taken in the previous decision epoch. Hence, for arm i we define a two dimensional *extended* state space $L_i = \left\{\left(a_{-1}^i, x_i\right) | a_{-1} \in \{0,1\}, x_i \in S_i\right\}$, where a_{-1} represents the action (active or passive) taken at the previous decision epoch, x is the current state of the project, and S is the one dimensional state space for an isolated arm.

Arm i's behavior can be described in the following way: every time the active action is taken in state $\left(a_{-1}^i, x_i\right) \in L_i$ the project evolves, according to

[2] See below in Section 3.1 a detailed review in related literature.

[3] Actually, Asawa and Teneketzis [6] introduced this extension, but their analysis is limited to the classical multi-armed bandit framework.

some Markovian transition rule, to state $(1, x_i') \in L_i$. On the other hand, if passive action is taken in state $(0, x_i) \in L_i$, the system remains frozen in that state. So far, our system behaves like the classic MAB problem. However, if passive action is taken at state $(1, x_i) \in L_i$ the system performs an instantaneous transition to state $(0, x_i) \in L_i$, which confers the restless nature to our problem. We assume that all arms start from passivity, i.e. initial states take the form $(0, x_i) \in L_m$ for all $1 \leq i \leq M$.

Regarding the reward structure, we assume that when the active action is taken on arm i at some state $x_i \in S_i$ an state/arm dependent active reward, $R^i(x_i)$, is earned. Passive action gives no reward. Moreover, if active action is taken whenever the arm is passive, an strictly positive "switching cost", $C_i(x_i)$, is incurred. With this elements, we can write down a general expression for the active reward earned by bandit i at any state x_i as[4]:

$$R(a_{-1}, x) = R(x) - C(x) \times (1 - a_{-1})$$

Consequently, $R(1, x) = R(x)$ and $R(0, x) = R(x) - C(x)$ literally *absorbs* the whole burden of the switching cost (SC) as a component of its active reward. With this reward structure the MAB w/SC is reduced to a simple RB without switching penalties and with two different families of states. Once indexability of this problem had been established, and priority indices obtained, the *index policy* will simply prescribe to engage, at each decision epoch, the arm(s) with larger index value.

The chapter is organized as follows. In Section 3.1 related literature is surveyed. In Section 3.2, the problem of translating the multi-armed bandit problem with switching costs, to which we will refer as the MAB w/SC, into

[4] For the sake of simplicity we omit here the arm subindices.

a *pure set-up cost* formulation is addressed and a proof for indexability of the extended-space *pure set-up cost* formulation is provided. Based on [23], an adaptive greedy algorithm for index calculation is presented also in this section. In Section 3.3, two interesting results and one application are discussed. The results regard, respectively, the indexability of the MAB w/SC when there exists a positive probability of losing idle (passive) projects, and the optimality of the index policy when the switching costs are high enough. An application to the scheduling of stochastic jobs is also introduced here. Results of an intensive numerical assessment of the performance of the index policy in the MAB w/ SC are offered in Section 3.4. Section 3.5 concludes.

3.1 Related Literature

In one of the earliest works in this topic, K. Glazebrook [45] showed that, for a collection of jobs to be processed by a single machine in a manner which is consistent with a precedence relationship, with the machine being free to switch from one job to the other by incurring a cost, there exists -under given conditions- an optimal strategy for allocating the machine to the jobs which is given by a fixed permutation of the jobs indicating in which order they should be processed. This optimal strategy belongs to the class of non-preemptive policies. The author also propose an algorithm that yields optimal strategies for certain specialised switching-cost structures.

In 1990, Agrawal, Hedge and Teneketzis [2] concentrated on multi-armed bandit problems with switching costs and multiple plays. Earlier, 1988, Agrawal et.al. [1] had provided a solution to the multi-armed bandit problem with switching costs but with single plays by exploiting a block allocation scheme. In the 1990's paper, they presented a lower bound on asymptotic performance of *uniformly good* allocation rules and constructed an allocation scheme that achieves that bound. Despite the presence of a switching cost, the proposed allocation scheme was shown to achieve the same asymptotic performance as the optimal rules for the bandit problem without switching cost. This was made by grouping the arms in such a way that makes the rate of switching negligible compared to the rate of operation.

In 1992, Van Oyen, Pandelis and Teneketzis [103] investigated the impact of switching penalties on the nature of optimal scheduling policies for systems of parallel queues without arrivals, which is an special case of MAB with switching costs. For switching penalties depending on the particular nodes involved in a switch, they showed that, although an index rule is not optimal in general, there is an exhaustive service policy that is optimal. In particular, they concentrated their work on obtaining *optimal* scheduling policies for a model of N parallel queues with switching penalties, linear holding costs, general service distributions and no arrivals. The class of admissible strategies was taken to be the class of non-idling, non-preemptive, and non-anticipative scheduling policies. For the *switching cost* problem, they considered that switching cost K_{ij} is incurred at each instant the server completes a job of node i and then serves a job of nod j. For the switching delay problem, a random length of time, D_{ij} is is required for switching from queue i to j; thus, the holding cost incurred during the switching delay is the implicit penalty.

By defining a policy to be *exhaustive* if according to it the server never leaves a node before completing the service of all customers in there, authors proved that only exhaustive policies can be optimal. That means that an optimal policy prescribes exactly $(N-1)$ switches, which provides a reduction in the class of policies which are candidate solutions to the optimization problem: the set of exhaustive pure Markov policies. Additionally, they proved that index policies are optimal under the additional assumption that the switching cost (delay) is a know constant K (D, respectively). In particular, the following index rule is proved to be optimal: serve the queues exhaustively in decreasing order of the dynamic allocation indices ν_i for the switching cost case, and $\overline{\nu_i}$ for the switching delay case; indices are, respectively, given by

$$\begin{aligned}
\nu_i &= \frac{C_i\left(x_i\right) - K}{E\left[\int_0^{\tau_i} e^{-\alpha t} dt\right]} \\
\overline{\nu_i} &= \frac{C_i\left(x_i\right)}{E\left[\int_0^{\tau_i + D} e^{-\alpha t} dt\right]}
\end{aligned}$$

where, τ_i is the total processing time of x_i jobs in node i.

In 1994, Banks and Sundaram [8] examined the extent to which the Gittins-Jones [38] theorem remains valid when the cost of switching between arms is possibly non-zero, i.e. to determine whether suitable defined index strategies continue to remain optimal in the switching-costs case. It is well known that when the switching cost depends on characteristics of both, abandoned and incoming arm, there cannot exist an optimal index strategy with the index of an arm depending solely on one arm's characteristics, as the Gittins-Jones theorem prescribes. Moreover, by considering a model in which the cost of switching away from an arm (resp. to an arm) is independent of the arm to which (resp. from which) the switch is made, they showed (using a *reductio ad absurdum* approach) that -even in this simpler case- it is not possible to define indices which have the property that the resulting strategy is optimal in the domain of all bandit problems with switching-costs. Indeed, their result remains true even if attention is restricted to that subset of the domain in which the switching-cost is a non-zero constant.

In that same year, Benkherouf, Glazebrook and Owen [13] addressed the special case where a set-up cost is paid only the first time the active action is taken in a particular arm. They proved that the Gittins index is indeed optimal. This is so because,as Banks and Sundaram [8] pointed out, if there is any possibility of switching back to an arm after abandoning it, then the index when it is active arm must be increasing in the cost of switching to back to it. Since the concern about switching disappears once an arm is played, the index on the current arm does not need to increase in the cost of switching back to it after some a passive sojourn. Therefore, the inclusion of one time set-up costs does not contradict the fact that the index on the current arm should be decreasing in switching costs.

Also in 1994, Van Oyen and Teneketzis [104] presented a research on structural properties of optimal policies for the problem of scheduling a single server in a forest network of N queues (without arrivals) subject to switching penalties. In their formulation authors allow jobs served at one queue to be transferred to another (internal arrivals) which prevents the application of the simple index structure found for the case of parallel queues in their previous work (see above). Notwithstanding, they are able to outline a class of problems for which relatively simple (*exhaustive*) policies are optimal.

Later, 1996, Asawa and Teneketzis [6], presented an algorithm for the computation of a *switching index* and established sufficient conditions for optimality of allocation strategies, based on limited look-ahead techniques. In their work it was shown that, under an optimal policy, decisions about the processor allocation in presence of switching costs need to be made only at stopping times that achieve an appropriate index, the *Gittins Index* or the *switching-index* which is defined both, for switching penalties and switching delays. For special class of multi-armed bandits (scheduling of parallel queues with switching penalties and no arrivals), it is shown that the afore mentioned property of optimal policies is sufficient to determine an optimal allocation strategy. The

Gittins index is given by

$$\nu_{gi}(t) = \max_{\tau > t} \frac{E_i \left\{ \sum_{l=t}^{\tau-1} \beta^l X^i(l) \right\}}{E_i \left\{ \sum_{l=t}^{\tau-1} \beta^l \right\}}$$

and the *switching cost index* has the following form:

$$\nu_{ci}(t) = \max_{\tau > t} \frac{E_i \left\{ \sum_{l=t}^{\tau-1} \beta^l X^i(l) - C\beta^t \right\}}{E_i \left\{ \sum_{l=t}^{\tau-1} \beta^l \right\}}$$

The switching cost C is a non-zero constant independent of machine dynamics. To calculate the indices, the authors propose a new Markov chain with state space $\hat{\Theta} = \{1, \ldots, M, 1', \ldots, M'\}$ and transition probabilities given by

$$\begin{aligned} \hat{P}_{ij} &= P_{ij}, \quad i, j \in \{1, \ldots, M\} \\ \hat{P}_{ij'} &= 0, \quad i \in \{1, \ldots, M\},\ j' \in \{1', \ldots, M'\} \\ \hat{P}_{i'j} &= P_{ij}, \quad i' \in \{1', \ldots, M'\},\ j \in \{1, \ldots, M\} \\ \hat{P}_{i'j'} &= 0, \quad i', j' \in \{1', \ldots, M'\}, \end{aligned}$$

and rewards

$$\begin{aligned} \hat{R}(j) &= R(j), \qquad j \in \{1, \ldots, M\}, \\ \hat{R}(j') &= R(j) - C, \quad j' \in \{1', \ldots, M'\}. \end{aligned}$$

Hence, if the machine is *passive* its state is represented by an element of the subset $\{1', \ldots, M'\} \subset \hat{\Theta}$; and by $\{1, \ldots, M\} \subset \hat{\Theta}$ if it is *active*. It follows from their argument that the Gittins indices $\hat{\nu}_{gi}(t)$ for the problem above are given by

$$\begin{aligned} \hat{\nu}_{gi}(t) &= \nu_{gi}(t), \quad i \in \{1, \ldots, M\}, \\ \hat{\nu}_{gi'}(t) &= \nu_{ci}(t), \quad i' \in \{1', \ldots, M'\}. \end{aligned}$$

For the actual index calculation the authors use the algorithms in Varaiya et.al. [105].

Asawa and Teneketzis also addressed the case of switching delays. The MAB with switching delays is considered to be the same as the problem with switching costs, except that a switching (setup) delay D is incurred when the server moves from one project to another and rewards are non-negative. They assume the delay to be a non-negative integer random variable with given distribution such that $0 < E[D] \leq \infty$ and is independent of machine dynamics. As in the case of switching costs, the Gittins index rule is not optimal for the problem with switching delays. However, an optimal scheduling policy for the multi-armed bandit problem with switching delay has been found to satisfy certain property. By defining the *switching delay* index

$$\nu_{di}(t) = \underset{\tau > t}{max} \frac{E\left[\beta^D \sum_{l=t}^{\tau-1} \beta^l X^i(l) \,|F^i(t)\right]}{E\left[\sum_{l=t}^{\tau+D-1} \beta^l \,|F^i(t)\right]}$$

the authors conclude that optimal decisions about the processor allocation are made only at those stopping times that achieve the appropriate index.

Asawa and Teneketzis' idea of an extended state space is the basis of the development in this Chapter. To our knowledge, it was José Niño-Mora who first suggested that multi-armed bandits with switching costs can be viewed as restless bandits. This idea was first used by Diego Ruiz-Hernández in 2001, [92].

In 1998, Reiman and Wein [86], analysed two scheduling problems for a queueing system with a single server and two customer classes. In the first problem a setup cost is incurred when the server switches from one class to the other, and the objective is to minimize the lung run expected average cost of holding customers and incurring set-up costs. The set-up cost is replaced by a setup time in the second problem, where the objective is to minimize the average holding cost. The scheduling problem in the setup case was formulated as to minimize:

$$\underset{T\to\infty}{limsup}\,\frac{1}{T}E\left[\int_0^T \sum_{i=1}^{2} c_i Q_i\left(t\right)dt+\frac{K}{2}J\left(T\right)\right]$$

where $Q_i\left(t\right)$ is the number of class i customers in queue or in service at time t, $J\left(t\right)$ the number of times the server sets up in the time interval $[0,t]$; $\frac{K}{2}$ is the setup cost for one switch, and c_1 is the cost incurred per unit time for holding a class i customer in the system. They approximated both dynamic scheduling problems by diffusion control problems. The diffusion control problem for the set-up cost problem was solved exactly, and asymptotics were used to analyse the corresponding set-up time problem. Computational results show that the proposed scheduling policies are within a small percent of sub-optimality over a broad range of problem parameters.

General references in the field of stochastic scheduling with set-up and switching costs include the works by Duenyas and Van Oyen [28], Karaesmen and Gupta [58], Kolonko and Benzing [66]. Other lines of research have searched for a characterization of the optimal solution by means of simplifying some of the assumtions, for example Bergemann and Välimäki [17] work with stationary arms, Benkherouf et.al. [13] consider only first-time switching costs, and Kavadias and Loch [62] concentrate on time invariant rewards.

For additional references and a well annotated survey on the multi-armed bandit problem with switching costs we refer the reader to Jun [57].

3.2 Restless Bandit Formulation of the MAB with Switching Costs

In this section, the stochastic scheduling problem with switching costs is addressed. It is shown that with a straightforward manipulation of the reward structure, a model including both set-up and tear-down costs can be reduced to a pure set-up cost problem without any loss of generality. Later, with a simple modification/extension of the state space the pure set-up formulation of our MAB w/SC is transformed into a restless bandit without switching penalties. Standard Whittle index theory, as deployed in Section 2.3 is applied for addressing the issue of indexability of the transformed problem. Once indexability of the transformed restless bandit problem has been established, an adaptive greedy algorithm for index calculation is proposed.

3.2.1 The Multi-armed Bandit Problem with Set-up and Tear-down Costs

Consider the problem of scheduling the operation of Q out of M competing projects. Given the resource constraint faced, the decision maker must decide whether to assign them to certain subset of projects or to some other. Under natural independence assumptions, his decision for each project can be modelled as an active-passive action choice. In that case, the decision problem can be formulated as a sequential decision model in which, at any epoch, the decision maker observes the state of a number M of two-action Markov Decision Processes, each corresponding to a different project. Based on the available information, the planner selects $1 \leq Q < M$ projects to be activated in the current period and to let idle the remaining ones.

Consider just one project (bandit or arm) arm in isolation. When active action is taken on project's state $X(t) = x$ at time $t \in \mathbb{N}$, it returns a state dependent active reward $R(x)$ and evolves to the next state following a general Markovian transition rule. If passive action is taken, no reward is earned and the project remains idle in the incoming state.

Moreover, switching activity between two different arms (say from arm i to arm j, where $1 \leq i, j \leq M$) implies two different state-dependent costs: a tear-down cost depending on the state of the abandoned arm, $Q_i(x_i)$, and a set-up cost, $S_j(x_j)$, depending on the incoming arm's state. Hence, the discounted switching cost at time t will be:

$$\beta^t \left[Q_i(x_i) + S_j(x_j)\right]$$

As Banks and Sundaram [8] pointed out, when the switching cost depends on characteristics of both abandoned and incoming arms, there cannot exist an index strategy with the index of an arm depending solely on that arm's characteristics, as the Gittins-Jones theorem prescribes. However, by means of a straightforward manipulation of the reward structure, the switching cost problem can be translated into a *pure set-up cost* formulation. This will be shown with a simple example.

Consider a two-armed bandit problem and assume that arms are *passive* at time $t = 0$, with $X_i(0) = x_i,\ X_j(0) = x_j$. At time $t = 0$ arm i is activated and a state dependent set-up cost is incurred. Additionally, an (state dependent) active reward is earned during the duration of arm's i active period, τ_i,

$$-S_i(x_i) + R_i(x_i, \tau_i)$$

At time τ_i, arm i is switched-off in state $X_i(\tau_i) = x_i'$ and arm j activated. Therefore, state dependent retirement (tear-down) and set-up costs are paid, and the reward earned across active period of arm j is:

$$-E\left[\beta^{\tau_i}\right]\left(Q_i\left(x_i'\right)+S_j\left(x_j\right)\right)+E\left[\beta^{\tau_i}\right]R_j\left(x_j,\tau_j\right)$$

After being active for τ_j periods, arm j is abandoned at state $X\left(\tau_j\right)=x_j'$, state dependent tear-down cost is paid and arm i is engaged again with its corresponding set-up cost incurred, and so on and so forth.

$$-E\left[\beta^{\tau_i}\right]E\left[\beta^{\tau_j}\right]\left(Q_j\left(x_j'\right)+S_i\left(x_i'\right)\right)+\dots$$

By putting everything together and reordering terms we get:

$$\begin{aligned}-S_i\left(x_i\right)-Q_i\left(x_i\right)+R_i\left(x_i,\tau_i\right)+\left\{Q_i\left(x_i\right)-E\left[\beta^{\tau_i}Q_i\left(x_i'\right)\right]\right\}\\+E\left[\beta^{\tau_i}\right]\Big(-S_j\left(x_j\right)-Q_j\left(x_j\right)+R_j\left(x_j,\tau_j\right)\\+\left\{Q_j\left(x_j\right)-E\left[\beta^{\tau_j}Q_j\left(x_j'\right)\right]\right\}\Big)\cdots\quad(3.1)\end{aligned}$$

Here, $Q_i\left(x_i\right)-E\left[\beta^{\tau_i}Q_i\left(x_i'\right)\right]$ can be understood as a residual reward earned over $[0,\tau_i)$, when starting from state x_i.

It now follows from (3.1) that if in the underlying multi-project scheduling problem, all constituent projects are deemed passive at time $t=0$ then the following modifications to our bandit are sufficient to accomodate tear-down costs:

1. The set-up cost in x_i becomes to $C_i\left(x_i\right)=S_i\left(x_i\right)+Q_i\left(x_i\right)$, and
2. Rewards $R\left(x\right)$ is enhanced to to $\widetilde{R}_i\left(x_i,x_i'\right)=R_i\left(x_i\right)+Q_i\left(x_i\right)-\beta Q\left(x_i'\right)$.

Which implies that (3.1) can be rewritten as:

$$-C_i\left(x_i\right)+\widetilde{R}_i\left(x_i,x_i'\right)+E\left[\beta^{\tau_i}\right]\left[-C_j\left(x_j\right)+\widetilde{R}_j\left(x_j,x_j'\right)\right]\dots$$

which is a *pure setup-cost* formulation. This result can be extended to any multi-armed bandit problem with multiple plays and switching (set-up and tear-down) costs. Hence, we can restrict our analysis to the pure set-up cost formulation without any loss of generality. We shall refer to this as the *switching-costs* problem.

3.2.2 Restless Bandit Formulation

The standard Markov Decision Problem (MDP) formulation for the MAB with switching costs can be described as follows:
Each project $i,\ 1 \leq i \leq M$ is modelled as a Markov Decision Chain (MDC) that evolves over the (countable) state space S_i with two possible actions $a_i \in \{0,1\}$ available at each state $x_i \in S_i$, and decision time $t \in \mathbb{N}$, where $a_i = 0$ means passivity and $a_i = 1$ activity. The aggregated state space is denoted by $\mathcal{S} = \bigtimes_{i=1}^{M} S_i$, with the system state at time t given by $\mathbf{X}(t) = (X_1(t), \ldots, X_M(t)) \in \mathcal{S}$.

If active action ($a_i = 1$) is taken in arm i at state $x_i \in S_i$ an arm/state dependent active reward $R_i(x_i)$ is earned and the arm evolves to state $x_i' \in S_i$ according to certain markovian transition rule. Otherwise, if passive action is taken ($a_i = 0$), no reward is earned and the project remains frozen in state x_i.

Before discussing the particular switching cost parameters, we need to consider two facts: 1) Banks et.al. [8] have already shown that it is not possible to define indices which have the property that the resulting strategy is optimal in the domain of all bandit problems with switching costs; and 2) it is well known that, under indexability, Whittle's index policy provides very efficient (asymptotically optimal) solutions to the restless bandit problem. Hence, our idea is to transform the Multi-armed Bandit Problem with Switching Costs into a Restless Bandit problem without them and then to apply the Whittle's approach described before in Section 2.3. For doing so it is just necessary to slightly modify the state space in the way described in the next paragraph.

For each arm and state we introduce a pair of *extended states* $l_i = \left(a_{-1}^i, x_i\right)$, where $x_i \in S_i$ represents the actual state of project i, and the value a_{-1}^i indicates whether it is an active state (project i is currently operated, in which case the state take the form $(1, x_i)$), or it is a passive state (project i lays in rest, with the state taking the form $(0, x_i)$). Hence, the extended state space of a single arm or project becomes $L_i = \left\{\left(a_{-1}^i, x_i\right) \left| a_{-1}^i \in \{0,1\} \; x_i \in S_i \right.\right\}$. The system's *extended state* at some decision time t is hence given by $\mathbf{L}(t) = (l_1(t), \ldots, l_M(t)) \in \mathcal{L}$, with $\mathcal{L} = \bigtimes_{i=1}^{M} L_i$.

We have now a two dimensional state space for each project, with two separated families of states: active and passive; each of them having a different reward structure but sharing the same transition rules as will be seen later.

Figure 3.1 below sketches the evolution of an arm starting at state $(1, x)$ depending on whether the action taken in the next decision epoch is active $a = 1$ or passive $a = 0$. There can be clearly seen the restless nature of the transformed problem: every time the passive action is taken in an active state $(1, x)$ it evolves to the corresponding passive state $(0, x)$ and remains there until the active action is taken again in that particular project.

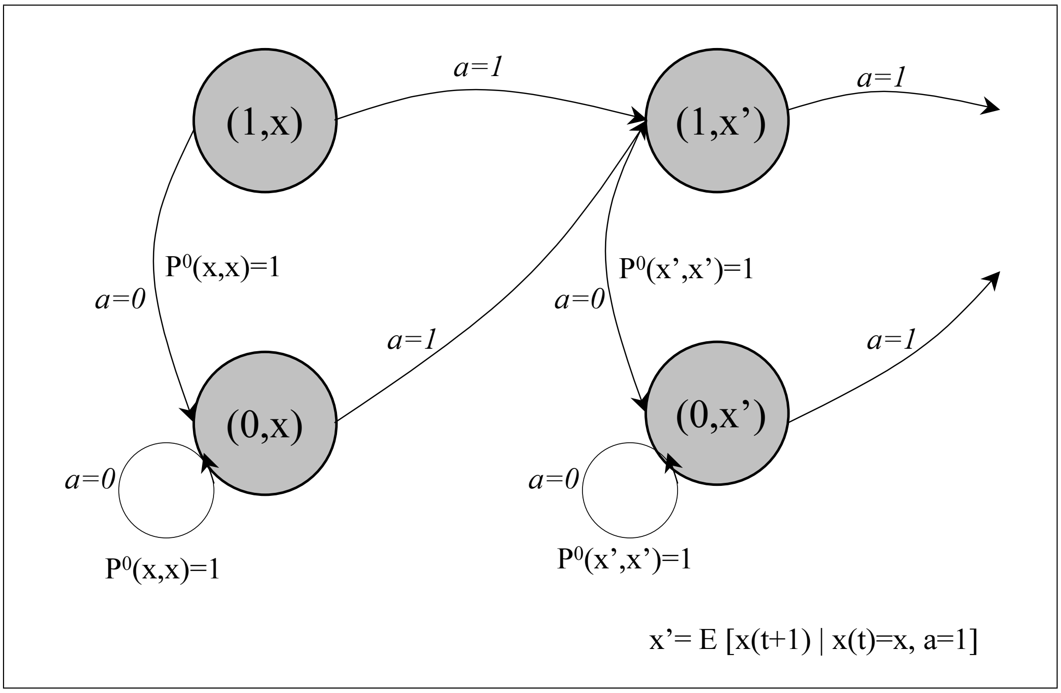
The standard discrete time MDP formulation for this restless bandit problem, represented by $(\mathcal{L}, \mathcal{A}, \mathbf{P}, \mathbf{R}, M, Q, \beta)$ is summarised by the following elements (equivalent to §$1_{(\mathbf{RB})}$ to §$6_{(\mathbf{RB})}$ described in Section 2.3):

1. **Decision Epochs**

   Decisions are taken at epochs $t \in \mathbb{N}$.

2. **State Space**

   The (countable) set of all possible system states at decision epoch t is the Cartesian product $\bigtimes_{i=1}^{M} L_i$. However, as $Q < M$ we can define the



Whenever active action is taken in state (a_{-1}, x), the arm evolves in probability to state $x' \in S$, independently of the action taken during the previous decision epoch a_{-1}. However, under passive action, the (active) arm first performs a transition to the passive state $(0, x)$ and will remain there unless active action is taken again. This sole transition from $(1, x)$ to $(1, x)$ is the only source of *restlesness* in the modified model.

Figure 3.1: Representation of a Project in the Restless Bandit Formulation of the Multi-armed Bandit with Switching Costs

set of relevant states to be

$$\mathcal{L} = \left\{ \mathbf{L} \,\middle|\, \sum_{i=1}^{M} a_{-1}^{i} \leq Q \right\} \cup \mathbf{L}_0$$

where $\mathbf{L}_0 = \big((0, x_1), (0, x_2), \ldots, (0, x_M)\big)$ is the initial state.

3. **Action Set**

We have already fixed $Q > M$, hence the collection $\binom{M}{Q}$ of admissible actions at state $\mathbf{L} \in \mathcal{L}$ is given by the set:

$$\mathcal{A} = \left\{ (a_1, \ldots, a_M) \middle| \sum_{i=1}^{M} a_i \leq Q,\ a_i \in \{0,1\} \right\}$$

4. **Transition Probabilities**

   The general form is

$$\mathcal{P}^{\mathbf{a}}_{\mathbf{L},\mathbf{L}'} = \prod_{i=1}^{M} P^{a_i}_{l_i,l'_i} = \prod_{i=1}^{M} P^{a_i}_{\left(a^i_{-1},x_i\right),\left(a_i,x'_i\right)},\ \forall\, \mathbf{L}, \mathbf{L}' \in \mathcal{L},\ \mathbf{a} \in \mathcal{A}$$

   where

$$P^{a_i}_{\left(a^i_{-1},x_i\right),\left(a_i,x'_i\right)} = \begin{cases} P_i\left(x_i, x'_i\right), & a_i = 1 \\ 1, & a_i = 0,\ x_i = x'_i \\ 0, & otherwise \end{cases}$$

   and $P_i\left(x_i, x'_i\right) = P\left\{X_i\left(t+1\right) = x'_i | X_i\left(t\right) = x_i, a_i\left(t\right) = 1\right\}$, with $x_i, x'_i \in S_i$.

   Let $\mathbf{P}$ represent the collection of transition matrices $\mathcal{P}^{\mathbf{a}},\ \mathbf{a} \in \mathcal{A}$.

5. **Active Rewards**

   As there are just active rewards, we have a simple structure for all $\mathbf{L} \in \mathcal{L}$,

$$\mathcal{R}^{\mathbf{a}}_{\mathbf{L}} = \sum_{i=1}^{M} 1\left(a_i = 1\right) R_i\left(l_i\right),\ \mathbf{a} \in \mathcal{A}$$

   with

$$R_i\left(l_i\right) = R_i\left(a^i_{-1}, x_i\right) = R^i\left(x_i\right) - \left(1 - a^i_{-1}\right) C_i\left(x_i\right) \tag{3.2}$$

   and $1\left(a_i = 1\right)$ is an indicator function taking value 1 whenever active action is taken in arm i.

Please notice that if only set-up costs are incurred, the active reward will depend only on the current state and consequently $\widetilde{R}\left(x_i, x_i'\right)$ in §2 on page 83 will become simply $R_i\left(x_i\right)$. Otherwise, if both set-up and tear-down costs are present, then we shall use the telescoped version $R_i\left(x_i\right) = \sum_{x_i' \in S_i} P_i\left(x_i, x_i'\right) \widetilde{R}_i\left(x_i, x_i\right)$, for all $x_i \in S_i$ and $1 \leq i \leq M$.

Let $\mathbf{R}$ represent the collection of reward vectors $\mathcal{R}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

6. **Policy** The goal of optimization is the choice of a policy π to maximise the total expected discounted reward earned over an infinite horizon. The theory of stochastic dynamic programming asserts the existence of an optimal (reward maximising) policy which is *stationary* and which satisfies the optimality equations of DP.

Let us denote $V\left(\mathbf{L}\right)$ the optimal problem value when the initial *extended state* is given by $\mathbf{L}$. The corresponding DP equations are given by

$$V\left(\mathbf{L}\right) = \max_{\mathbf{a} \in \mathcal{A}_{\mathbf{L}}} \left(\mathcal{R}^{\mathbf{a}}_{\mathbf{L}} + \beta \sum_{\mathbf{L}' \in \mathcal{L}} \mathcal{P}^{\mathbf{a}}_{\mathbf{L},\mathbf{L}'} V\left(\mathbf{L}'\right) \right), \ \mathbf{L} \in \mathcal{L} \tag{3.3}$$

The solution is based on identifying an optimal scheduling policy π prescribing, at each state, which Q projects to activate and which others $M - Q$ to lay rest. However, as we have seen, this problem is an example of the curse of dimensionality, which hinders the application of dynamic programming and consequently we need to look for an alternative heuristic for finding an efficient solution to this problem: an index policy. In particular, we are going to concentrate in the family of Whittle's index heuristics discussed in Section 2.3.

As mentioned before, indexability and indices are properties of individual bandits; hence in the MDP §1 to §6 in page 85 we isolate an individual bandit (L, P, R, β) and will now drop the bandit identifier i. This bandit generates a collection of MDP's parametrised by a *subsidy for passivity* $W \in \mathbb{R}$. As before, we refer to this as the W-*subsidy* problem for (L, P, R, β). This is a discounted reward MDP as follows:

1'. Decisions are taken at time $t \in \mathbb{N}$.

2'. The countable state space is L. We use the $l\left(t\right)$ for the extended state of the process at time t, where $l\left(t\right) = \left(a_{-1}, x\left(t\right)\right)$, $a_{-1} = a\left(t-1\right)$ is the action taken during the previous decision epoch, and $x\left(t\right) \in S$ is the actual state of the bandit at t.

3'. At each decision epoch t, either action $a = 1$ (active) or action $a = 0$ (passive) is applied to the process.

4'. If $a = 0$, then

$$P^{0}\left(x, x'\right) = \begin{cases} 1, & x' = x \\ 0, & \text{other case} \end{cases}, \ x \in S$$

If $a = 1$, then $P^{1}\left(x, x'\right),\ x, x' \in S$ is a Markovian transition rule.

Hence, we can write

$$\begin{aligned} P_{(a_{-1},x),(1,x')} &= P\left\{l\left(t+1\right) = \left(1, x'\right) | l\left(t\right) = \left(a_{-1}, x\right), a_{-1} \in \left\{0,1\right\}, a = 1\right\} \\ &= P^{1}\left(x, x'\right) \\ P_{(a_{-1},x),(0,x')} &= P\left\{l\left(t+1\right) = \left(0, x\right) | l\left(t\right) = \left(a_{-1}, x\right), a_{-1} \in \left\{0,1\right\}, a = 0\right\} \\ &= P^{0}\left(x, x'\right), \end{aligned}$$

for all $x, x' \in S$.

5'. If a transition from $(1, x)$ to $(1, x')$ occurs under action $a = 1$ at time t, a discounted reward $\beta^{t} R\left(x\right)$ is earned. If, otherwise the active transition to state $(1, x')$ occurs from state $(0, x)$ the immediate reward becomes $\beta^t (R(x) - C(x))$. In general, for $l \in L$, we can write $R(l) = R(a_{-1}, x) = R(x) - C(x)(1 - a_{-1})$, $x \in S$, $a_{-1} \in \{0, 1\}$.

Transitions under the passive action (from states $(0, x)$ or $(1, x)$ to state $(0, x)$) are awarded with discounted subsidy for passivity $\beta^t W$.

6'. The goal of optimisation is the choice of a policy to maximise the total expected reward (including passive subsidies) earned over an infinite horizon. We again assert the existence of optimal policies for the W-*subsidy* problem which are stationary and whose value functions satisfy the optimality equations of *dynamic programming*. We shall restrict to stationary policies throughout.

We use $V(l, W)$ for the value function for the W-*subsidy* problem evaluated at $l \in L$. The DP optimality equations may be expressed as:

$$V(l, W) = \max \left\{ R(l) + \beta \sum_{l' \in L^1} P_{l,l'} V(l', W) ; W + \beta \sum_{l' \in L^0} P_{l,l'} V(l', W) \right\} \tag{3.4}$$

with $L^0 = \{(0, x) | x \in S\}$ and $L^1 = \{(1, x) | x \in S\}$ (clearly $L = L^0 \cup L^1$).

The first term in $\{;\}$ in the r.h.s. of (3.4) represents the expected discounted reward of taking the active action ($a = 1$) in state $l \in L$ and the second one corresponds to the passive action ($a = 0$).

For the solution to this problem we are going to concentrate in the family of Whittle's index heuristics discussed in Section 2.3. As the next step is to establish indexability for Restless Bandit formulation of the Multiarmed Bandit with Switching Costs, we ask the reader to recall Definitions 2.1 and 2.2 and the discussion around them in page 59.

3.2.3 Indexability Analysis

We start our analysis with an explicit version of the DP optimality equations in (3.4),

$$\begin{aligned} V\big((1,x),W\big) &= \max\big\{R(x)+\beta E\left[V\big((1,y),W\big)\right];\\ &\qquad W+\beta V\big((0,x),W\big)\big\} \qquad (3.5)\\ V\big((0,x),W\big) &= \max\big\{R(x)-C(x)+\beta E\left[V\big((1,y),W\big)\right];\\ &\qquad W+\beta V\big((0,x),W\big)\big\} \end{aligned}$$

On the r.h.s of equations above, the first term within $\{;\}$ corresponds to the choice of the active action $(a=1)$ in the actual state $x\in S$, and the second term to the choice of the passive action $(a=0)$. It is trivial to see that the second equation can be simplified to

$$V\big((0,x),W\big) = \max\left\{R(x)-C(x)+\beta E\left[V\big((1,y),W\big)\right];\frac{W}{1-\beta}\right\} \qquad (3.6)$$

In order to develop optimal policies for the W-*subsidy* problem we develop the *Gittins index for activity* in the following way: suppose that $X(0)=x\in S$ and that the active action $a=1$ is taken at times $0,1,2,\ldots,\tau-1$ where τ is a stationary positive stopping time on the process $\{X(t),t\geq 0\}$. We write

$$R^{\tau}(x)\equiv E\left[\sum_{t=0}^{\tau-1}\beta^{t}R\big(X(t)\big)\,|X(0)=x\right]$$

for the expected discounted reward earned during this process.

Definition 3.1. *The Gittins index for activity $G:L\rightarrow\mathbb{R}$ is given by*

$$G(1,x)=\underset{\tau>0}{sup}\left[\frac{R^{\tau}(x)}{1-E\left[\beta^{\tau}\,|x\right]}\right],\ x\in S, \qquad (3.7)$$

and

$$G\left(0,x\right)=\underset{\tau>0}{sup}\left[\frac{-C\left(x\right)+R^{\tau}\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right.\right]}\right],\ x\in S, \tag{3.8}$$

where the suprema in (3.7) and (3.8) are taken over all stationary positive-valued stopping times on the process $\left\{X\left(t\right),t\geq0\right\}$ *evolving under the active action. Those suprema are guaranteed to be achieved. It must also be clear from the expressions above that* $G\left(1,x\right)>G\left(0,x\right)$*, for* $x\in S$ *and* $C\left(x\right)>0$*.*

Comment

We can characterise the bandit for which $G\left(a_{-1},x\right)$ in (3.7) and (3.8) are the Gittins indices. That bandit is one in which $l\left(0\right)=\left(0,x\right)$, with the bandit state process $\left\{l\left(t\right),t\geq0\right\}$ evolving under the active action, as in §4' in page 89. Further, a transition from state $\left(a_{-1},x\right)$ to state $\left(a_{-1},x'\right)$ earns a reward of $R\left(x\right)-C\left(x\right)\left(1-a_{-1}\right)$. With these choices, the expected reward earned by the bandit during $[0,\tau)$ is given by

$$R^{\tau}\left(x\right)-C\left(x\right)\left(1-a_{-1}\right)$$

We now appeal to Gittins index theory to characterise the set of stopping times achieving the suprema in (3.7) and (3.8): fix $W\in\mathbb{R}$ and use Γ and Σ for the subsets of S given by

$$\Gamma\left(W\right)=\left\{x\in S:G\left(a_{-1},x\right)<W\right\},$$

and

$$\Sigma\left(W\right)=\left\{x\in S:G\left(a_{-1},x\right)=W\right\}.$$

Now suppose that $X(0) = x$ and $\Sigma \subseteq \Sigma(W)$. Use τ^{Σ} for the stationary positive-valued stopping time, defined on the process $\{X(t), t \geq 0\}$ evolving under the active action, given by

$$\tau^{\Sigma} = \min\{t : t > 0, X(t) \in \Gamma(W) \cup \Sigma\}.$$

We now write $T(x, W)$ for the collection given by

$$T(x, W) = \bigcup_{\Sigma \subseteq \Sigma(W)} \{\tau^{\Sigma}\}. \tag{3.9}$$

The following result combines straightforward calculations with standard features of Gittins index theory. We omit the proof.

Lemma 3.1.

(a) Any stopping time in $T(x, G(1, x))$ achieves the supremum in (3.7)*.*

(b) Any stopping time in $T(x, G(0, x))$ achieves the supremum in (3.8)*.*

Before proceeding to the main result of this section, we pause to recollect the work of Whittle [113] who utilised a set of decision problems involving a notion of *retirement* to characterize the Gittins index. Suppose that $l(0) = (a_{-1}, x) \in L$. Consider a decision problem in which at each time $t \in \mathbb{N}$ a choice has to be made between the active action $a = 1$ and retirement. Once retirement is chosen, it must continue to be chosen thereafter. The effect of choices of the active action (in terms of stochastic evolution and rewards earned) is precisely the one defined in §1' to §6' in page 90. A reward W is earned on each occasion that retirement is chosen. If we write for the first time at which retirement is taken and by using

$$m^{\tau}(x) \equiv E[\beta^{\tau} | x],$$

we may express the value function for the retirement problem as

$$\widetilde{V}\big((1,x),W\big) = \sup_{\tau>0}\left\{R^{\tau}(x) + m^{\tau}(x)\frac{W}{1-\beta}\right\},\ x \in S, \tag{3.10}$$

$$\widetilde{V}\big((0,x),W\big) = \sup_{\tau>0}\left\{R^{\tau}(x) - C(x) + m^{\tau}(x)\frac{W}{1-\beta},\right\}\ x \in S. \tag{3.11}$$

With all this elements we can now write down the following results, which may be established straightforwardly from Whittle's analysis:

Lemma 3.2. ***(Optimal Retirement)*** *For every passivity subsidy $W \in \mathbb{R}^{+}$ and if $x \in S$ we have the following cases:*

1. *If $l(0) = (1,x)$ and $G(1,x) > \frac{W}{1-\beta}$ it is optimal to retire at any stopping time in $T(x,W)$.*

2. *If $l(0) = (1,x)$ and $G(1,x) \leq \frac{W}{1-\beta}$ it is optimal to retire at $l(0)$ along with retirement at any other stopping time in $T(x,W)$ if the equality holds.*

3. *If $l(0) = (0,x)$ and $G(0,x) > \frac{W}{1-\beta}$ then it is optimal to retire at any stopping time in $T(x,W)$.*

4. *If $l(0) = (0,x)$ and $G(0,x) \leq \frac{W}{1-\beta}$ then it is optimal to take the retirement action at $l(0)$. In the case of $G(0,x) = \frac{W}{1-\beta}$ retiring at any other stopping time in $T(x,W)$ is optimal.*

5. *Statements (1) to (4) describe all optimal stationary policies for retirement.*

We are now in position to establish the structure of optimal stationary policies for the W-problem. Indexability will follow straightforwardly from Theorem 3.1.

Theorem 3.1. ***(Optimal policies for the W-subsidy problem)*** *The following statements hold for all $x \in S$ and $W \in \mathbb{R}$ and describe all optimal stationary policies for the W-subsidy problem:*

(a) If $G(0,x) > \frac{W}{1-\beta}$ then the active action is optimal in states $(0,x)$ and $(1,x)$.

(b) If $G(1,x) > \frac{W}{1-\beta} > G(0,x)$ then active action is optimal in state $(1,x)$ and passive action is optimal in state $(0,x)$.

(c) If $G(1,x) < \frac{W}{1-\beta}$ then the passive action is optimal in states $(0,x)$ and $(1,x)$.

(d) If $G(1,x) = \frac{W}{1-\beta}$ then both actions are optimal in $(1,x)$ together with passive action in $(0,x)$.

(e) If $G(0,x) = \frac{W}{1-\beta}$ then active action is optimal in $(1,x)$ and both passive and active actions are optimal in $(0,x)$.

Proof *Fix $W \in \mathbb{R}$ and suppose that state $x \in S$ is such that the passive action $a = 0$ is optimal for the W-subsidy problem when the bandit is in state $(1,x)$. It now follows from (3.5) and (3.6) that*

$$\begin{aligned} W + \beta V\big((0,x),W\big) &\geq R(x) + \beta E\left[V\big((1,y),W\big)\right] \\ &> R(x) - C(x) + \beta E\left[V\big((1,y),W\big)\right], \end{aligned}$$

from which it follows that passive action $a = 0$ must also be (strictly) optimal for the W-subsidy problem when the bandit is in state $(0,x)$.

Suppose now that $l(0) = l' \in L$ and let π be some stationary optimal policy for the W-subsidy problem. Write $\tau(\pi, l')$ for the first time at which π chooses *the passive action $a = 0$, namely*

$$\tau\left(\pi, l'\right) = \min\left\{t; t \geq 0, \pi\left\{l\left(t\right)\right\} = 0\right\}.$$

It follows simply from the above that π must choose passive action $a = 0$ at all decision epochs following $\tau\left(\pi, l'\right)$. Making the identification of the passive action with retirement, it is now clear that optimal policies for the W-subsidy problem exactly coincide with optimal retirement policies when W is the retirement reward, with the passive action optimal for the former if and only if retirement is optimal for the later. With this identification, the result follows immediately from Lemma 3.2.

To illustrate this correspondence, suppose for example that $G\left(1, x\right) > \frac{W}{1\beta} > G\left(0, x\right)$. It follows from Lemma 1 that non-retirement in $\left(1, x\right)$ is optimal in this range for the retirement problem and hence that the active action is optimal in $\left(1, x\right)$ for the W-subsidy problem. However, from Lemma 3.2(4) we see that retirement is optimal in $\left(0, x\right)$ and hence that the passive action is optimal for the W-subsidy problem. This establishes Theorem 3.1(b). Other cases are dealt similarly. This concludes the proof. ***q.e.d.*** ■

Theorem 3.2. ***(Indexability and Indices)*** *Bandit $\left(L, P, R, C, \beta\right)$ is indexable. The Whittle index $W : L \rightarrow \mathbb{R}$ is given by*

$$W\left(1, x\right) = \left(1 - \beta\right) G\left(1, x\right); \; W\left(0, x\right) = \left(1 - \beta\right) G\left(0, x\right), x \in S.$$

Proof *Write $\Pi\left(W\right)$ for the set of states in which it is optimal to take the passive action for the W-problem. By Theorem 3.1 we have*

$$\Pi\left(W\right) = \left\{\left(1, x\right) : \frac{W}{1 - \beta} \geq G\left(1, x\right)\right\} \cup \left\{\left(0, x\right) : \frac{W}{1 - \beta} \geq G\left(0, x\right)\right\}. \quad (3.12)$$

Plainly, $\Pi(W)$ *is increasing in* W*. Further, it follows from (3.12) that the Whittle index for state* $(1,x)$ *is given by*

$$W(1,x) = inf\{W\,|(1,x) \in \Pi(W)\} = (1-\beta)\,G(1,x)$$

and similarly for $W(0,x)$*. This completes the proof.* ***q.e.d.*** ■

Once indexability has been established, we present an algorithm for calculating the priority indices discussed above.

3.2.4 Index Computation

Bertsimas and Niño-Mora [23] and Niño-Mora [75], have developed a methodology for generating a set of indices for the restless bandit problem that, under certain assumptions[5], coincide with those obtained from the calibration proposed by Whittle [116].

In this section we will follow the technique in Glazebrook et.al. [52] for developing an *adaptive greedy algorithm* that uses as inputs the active reward rates and a matrix of constants $\mathbf{A}$ to be described below, and whose outputs are the project's indices. As our problem has already been proved to be indexable, we can be sure that the indices obtained by these means are indeed coincident with those of Whittle.

Start by defining the subset $F \subseteq L$ and the F-*active* policy u_F which chooses the active action whenever the state lies in F and passive action otherwise (arm's state is in $F^c = L\backslash F$). Where L is the extended state space as

[5] They impose the following *sufficient* condition for a system to be indexable: for $x \in S$ and $S_i \subseteq S$, $A_x^{S_i} > 0$. Where S represents the state space and $A_x^{S_i}$ is a collection of parameters. Indeed, their indices become the conventional Gittins indices for the multiarmed bandit case.

described before:

$$L = \{(a_{-1}, x) : x \in S, a_{-1} \in \{0,1\}\} \tag{3.13}$$

The structure of the switching costs problem imposes the following natural constraint in subset F: as it always holds that $W(1,x) \geq W(0,x)$, then $(0,k) \in F \Rightarrow (1,k) \in F$ and $(1,k) \in F^c \Rightarrow (0,k) \in F^c$. This result will be useful later.

Consider now an arm in isolation and suppose it is in natural state x at time $t = 0$ and evolving according policy u_F. If we denote by $\{X(t), t \geq 0\}$ the sequence of states visited by our arm under policy u_F, then this process is a Markov chain with $X(0) = (a_{-1}, x)$,

$$P\{l(t+1) = (1, x') \,|\, l(t) = (a_{-1}, x)\} = P^1(x, x'), \quad (a_{-1}, x) \in F$$
$$P\{l(t+1) = (0, x') \,|\, l(t) = (a_{-1}, x)\} = P^0(x, x'), \quad (a_{-1}, x) \in F^c$$

with

$$P^0(x, x') = \begin{cases} 1, & x' = x \\ 0, & \text{other case} \end{cases}, \tag{3.14}$$

as discussed in Section 3.2.2.

We can now define the total time the system takes active action in F under u_F as:

$$T^F_{(a_{-1},x)} = E_{u_F}\left\{\sum_{t=0}^{\infty} \beta^t I_F(t) \,|\, l(0) = (a_{-1}, x)\right\}, \ (a_{-1}, x) \in L$$

with

$$I_F(t) = \begin{cases} 1, & l(t) \in F \\ 0, & \text{other case} \end{cases}$$

The quantities T^F satisfy:

$$T^F_{(a_{-1},x)} = \begin{cases} 1 + \beta \sum_{(1,x')\in L} P^1(x,x')\, T^F_{(1,x')}, & (a_{-1},x) \in F \\ 0, & (a_{-1},x) \in F^c \end{cases} \tag{3.15}$$

We now define matrix $\mathbf{A}$ by means of the quantities $\left\{T^F_{(a_{-1},x)}\right\}_{(a_{-1},x)\in L, F\subseteq L}$ as follows:

$$A^F_{(a_{-1},x)} = 1 + \beta \sum_{(1,x')\in L} P^1(x,x')\, T^{F^c}_{(1,x')} - \beta \sum_{(0,x')\in L} P^0(x,x')\, T^{F^c}_{(0,x')},$$

By using the particular structure of passive transitions in our problem and the constraint imposed before on F, we obtain the following expression for the elements in matrix $\mathbf{A}$:

$$A^F_{(a_{-1},x)} = \begin{cases} 1 + \beta \sum_{(1,x')\in L} P^1(x,x')\, T^{F^c}_{(1,x')}, & (a_{-1},x) \in F \\ (1-\beta)\left[1 + \beta \sum_{(1,x')\in L} P^1(x,,x')\, T^{F^c}_{(1,x')}\right], & (a_{-1},x) \in F^c \end{cases}$$

for all $(a_{-1},x) \in L, F \subseteq L$.[6]

Finally we just need to introduce a few additional elements: let rewards earned at state $(a_{-1},x) \in L$ take the form

$$R(a_{-1},x) = R(x) - C(x)(1-a_{-1})$$

so that we can define the reward vector $\mathcal{R} = \left(R_{(a_{-1},x)}\right)_{(a_{-1},x)\in L}$. Let also $\pi = (\pi_1,\ldots,\pi_n)$ be a permutation of L, $\overline{y} = \left(\overline{y}^F\right)_{F\subseteq\mathcal{L}}$ and $\gamma = (\gamma_1,\ldots,\gamma_{2n})$. Hence, priority indices, γ_i are obtained by running the following adaptive greedy algorithm $\mathcal{AG}$ on input $(\mathcal{R},\mathbf{A})$.

[6] It can be readily seen that values in matrix $\mathbf{A}$, satisfy the *sufficient* condition in *footnote 5.*

```
INPUT: $\mathcal{R}, \mathbf{A}$
INITIALIZATION:
Set   $F_1 = L;$
      $\pi_1 = \text{argmax}\left\{\frac{R(a_{-1},x)}{A^{F_1}_{(a_{-1},x)}} \,:\, (a_{-1},x) \in F_1\right\};$
      $\overline{y}^{F_1} = \frac{R(\pi_1)}{A^{F_1}_{\pi_1}};$
      $\gamma_{\pi_1} = \overline{y}^{F_1}.$
PROCEDURE:
For $k = 2$ to $2n$ do
Set   $F_k = F_{k-1} \backslash \{\pi_{k-1}\}$
      $\pi_k = \text{argmax}\left\{\frac{R(a_{-1},x) - \sum_{j=1}^{k-1} A^{F_j}_{(a_{-1},x)} \overline{y}^{F_j}}{A^{F_k}_{(a_{-1},x)}} \,:\, (a_{-1},x) \in F_k\right\};$
      $\overline{y}^{F_k} = \frac{R_{\pi_k} - \sum_{j=1}^{k-1} A^{F_j}_{\pi_k} \overline{y}^{F_j}}{A^{F_k}_{\pi_k}};$
      $\gamma_{\pi_k} = \gamma_{\pi_{k-1}} + \overline{y}^{F_k}$
End For
OUTPUT: $\pi, \overline{y}, \gamma$
```

Figure 3.2: Adaptive Greedy Algorithm for the Multi-armed Bandit with Switching Costs

3.3 A Result, an Extension, and an Application of the MAB with Switching Costs

We now introduce a succession of model elaborations to the multi-armed bandit with switching costs problem. All of these preserve indexability and the essential index structure.

3.3.1 Result: The Index Policy is Optimal when Costs are High

A simple and intuitive result for the MAB w/SC problem we are analysing is the fact that, whenever the switching cost is big enough, the index policy turns out to be optimal. This is so because high switching costs prevent the system to activate different arms and, instead, it just operates one single arm. It can be proved that the prescribed active arm turns out to be the same under both, the optimal and the index policy and, consequently, the index policy is optimal.

Firstly, it is easy to verify that for large enough switching costs, the index policy prescribes taking active action in just one arm.

We have already defined the Whittle indices for the *restless bandit* version of the *multi-armed bandit with switching costs* problem as

$$W\left(a_{-1},x\right)=\left(1-\beta\right)G\left(a_{-1},x\right)$$

for $a_{-1}\in\{0,1\}$ and $x\in S$. Where

$$G\left(1,x\right)=\underset{\tau>0}{sup}\left[\frac{R^{\tau}\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right]\right.}\right],\ x\in S \tag{3.16}$$

and

$$G\left(0,x\right)=\underset{\tau>0}{sup}\left[\frac{-C\left(x\right)+R^{\tau}\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right]\right.}\right],\ x\in S. \tag{3.17}$$

It is clear that $W\left(1,x\right)>W\left(\left(0,x\right),C\left(x\right)\right)$ for any $C\left(x\right)>0$. Moreover, as $R\left(x\right)>0$ and $0<\beta<1$ also $W\left(1,x\right)>0$, and we can find some $c^{*}\left(x\right)$ such that

$$W\left(0,x,c^{*}\left(x\right)\right)=\underset{x'}{min}\left\{W\left(1,x'\right)\right\},\ \text{all } x\in S \tag{3.18}$$

Take now the supremum over all such $c^{*}\left(x\right)$, i.e.

$$C^{*}=\sup_{x}\left\{c^{*}\left(x\right)\right\},$$

it follows that

$$W\left(1,x\right)\geq W\left(0,x',C^{*}\right)\ \forall\, x,x'\in S.$$

As long as the active reward $R(x)$ is bounded, it is guaranteed that C^* will always be bounded.

Turn now to the multi-armed problem. The index policy prescribes taking the active action in the arm with the highest index. Hence, if we define

$$C^j = \max_{x_j \in S_j} \left\{ c_j^* \left(x_j \right) \right\} \; 1 \leq j \leq M,$$

and let

$$\overline{C} = \sup_{1 \leq j \leq M} \left\{ C^j \right\}$$

then it will be true that for any $C > \overline{C}$,

$$W_j \left(1, x_j\right) > W_i \left(\left(0, x_i\right), C\right) \; \forall \, 1 \leq i, j \leq M \text{ and } x_i \in S_i, \; x_j \in S_j$$

The implication of this result is that, once we have chosen arm k to be active, we will never abandon it.

Finally, for any initial state $L = \big(\left(0, x_1\right), \ldots, \left(0, x_M\right) \big) \in \mathcal{L}$, the Whittle index policy simply prescribes taking active action in arm k such that

$$k = \operatorname*{argmax}_{i} \left\{ W_i \left(\left(0, x_i\right), C \right) \right\}.$$

Consider now a simplified version of the MAB w/SC problem in which a *set-up cost*, $S_k\left(x_k\right)$, must be incurred when project k is chosen *for the first time*, but never again. We call this the *opening* problem, as it resembles the cost that is incurred in setting-up a machine for its first use but never again. For the sake of comparability, we use the same state structure as in the RB problem, i.e., passive states have the form $(0, x)$ and active states $(1, x)$. The passive state $(0, x)$ represents the machine in its *pristine* state, i.e. the "box is closed". In this case active transitions are: $(0, x) \rightarrow (1, x')$ and $(1, x) \rightarrow (1, x')$, and passive arms remain frozen in the incoming state, i.e. $(0,x) \to (0,x)$ and $(1,x) \to (1,x)$ i.e. a pristine machine remains pristine until its first use and a used machine will never be "boxed" again. It is clear that this is just a MAB and, consequently, the Gittins Index Policy turns out to be optimal.

In this case, Gittins indices are known to be given by:

$$G\left(1,x\right) = \underset{\tau>0}{sup}\left[\frac{R^{\tau}\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right]\right.}\right] \tag{3.19}$$

and

$$G\left(\left(0,x\right),S\left(x\right)\right) = \underset{\tau>0}{sup}\left[\frac{-S\left(x\right)+R^{\tau}\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right]\right.}\right] \tag{3.20}$$

respectively.

It is straightforward that $G\left(1,x\right) > G\left(\left(0,x\right),S\left(x\right)\right)$ for any $S\left(x\right) > 0$. Moreover, as $R\left(x\right) > 0$ it follows that $G\left(1,x\right) > 0$, and we can find some $s^{*}\left(x\right)$ such that

$$G\left(\left(0,x\right),s^{*}\left(x\right)\right) = \min_{x'}\left\{G\left(1,x'\right)\right\} \tag{3.21}$$

Take now

$$S^{*} = \sup_{x}\left\{s^{*}\left(x\right)\right\},$$

it holds that

$$G\left(1,x\right) > G\left(\left(0,x'\right),S^{*}\right)\ \forall\, x,x' \in S.$$

As before, we turn to the multi-armed problem and recover the arm indicators. Gittins' index policy prescribes taking the active action in the arm with the highest index. Hence, if we define

$$S^{j} = \max_{x_j}\left\{s_j^{*}\left(x_j\right)\right\}\ 1 \leq j \leq M,$$

and let

$$\overline{S} = \sup_{1 \leq j \leq M} \left\{ S^j \right\}$$

then it will be true that for any $S > \overline{S}$,

$$G_j\left(1, x_j\right) > G_j\left(\left(0, x_j'\right), S\right) \ \forall \, x_j, x_j' \in S \text{ and } 1 \leq j \leq M.$$

In such case, once we have activated arm k, we will remain there indefinitely. Finally, for any initial state $L = \left(\left(0, x_1\right), \ldots, \left(0, x_M\right)\right) \in \mathcal{L}$, it will be optimal to engage arm k such that

$$k = \underset{i}{\operatorname{argmax}} \left\{ G_i\left(\left(0, x_i\right), S\right) \right\}.$$

We are now ready to formulating the main result of this section.

Lemma 3.3. *(**Optimality of Index Policy for** $C\left(x\right) \geq \overline{C}$)*
There exists a value $\overline{C}$ such that for $C_i\left(x_i\right) \geq \overline{C}$ for all $x_i \in S_i$ and $1 \leq i \leq M$, it holds that the Whittle index policy for the restless bandit version of the multi-armed bandit problem with switching costs, is optimal; i.e. $V^{Ind}\left(L\right) = V^{Opt}\left(L\right)$ for any $L \in \mathcal{L}$. Where $\mathcal{L}^0 = \left\{L_i^0, 1 \leq i \leq M\right\}$ and $L_i^0 = \left\{\left(0, x_i\right); i \in S_i\right\}$.

Proof *Consider both problems together. Assume they have common features as discount factor, transition matrices, reward structure, and so on. Their corresponding Gittins indices are given by expressions (3.16) and (3.17), and (3.19) and (3.20), respectively.*

By simple observation of both pairs of equations, it is clear that for $C\left(x\right) = S\left(x\right)$, active and passive Gittins indices turn out to be the same for both problems. Moreover, as $W\left(a_{-1}, x\right) = \left(1 - \beta\right) G\left(a_{-1}, x\right)$, the ordering of states *under both Whittle and Gittins indices turns out to be the same, i.e.*

$$\underset{j}{\operatorname{argmax}}\left\{G_j\left((0,x_j),\overline{C}\right)\right\}=\underset{j}{\operatorname{argmax}}\left\{(1-\beta)\,G_j\left((0,x_j),\overline{C}\right)\right\} \tag{3.22}$$

We have already seen that for values of $C(x)\geq\overline{C},\forall x\in S$ *(respectively* $S(x)\geq\overline{S},\forall x\in S$*), once the active action is taken in a particular arm, it will never be abandoned. Moreover, from* (3.22) *and given the fact that* $C(x)=S(x)\Rightarrow\overline{C}=\overline{S}$*, it must be clear that the first active arm* k *(when starting from a passive state* $L\in\mathcal{L}^0$*) will be the same under both formulations. But that choice is optimal under the opening problem. Consequently*

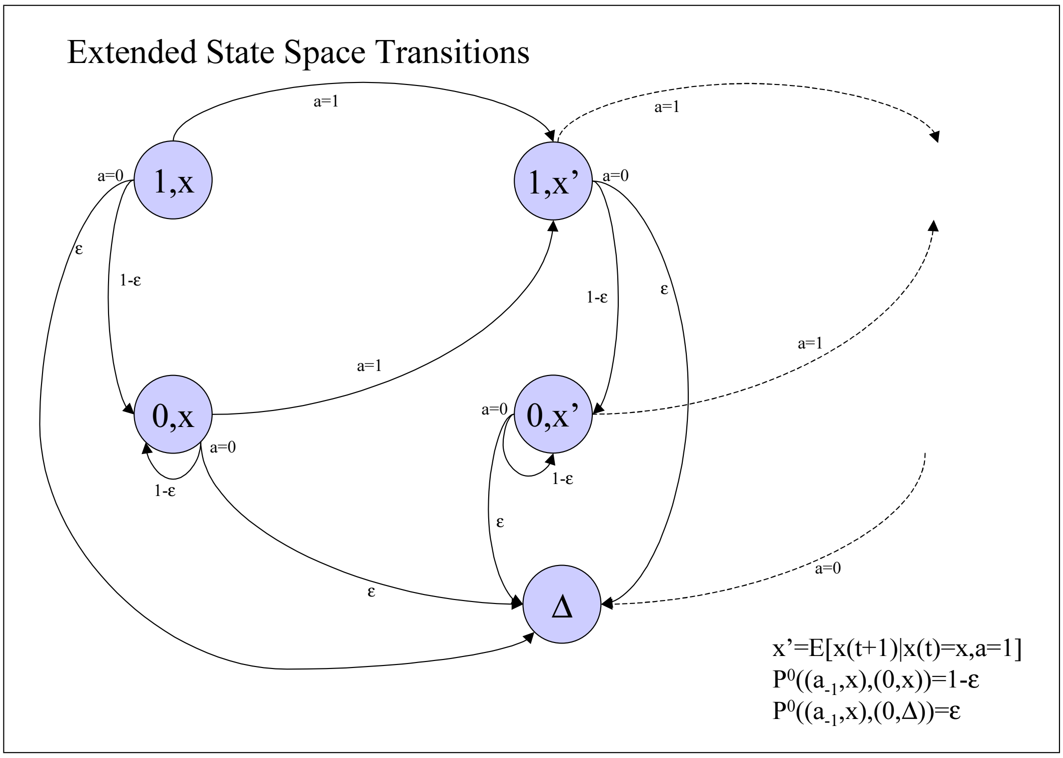
$$V^{Opt}(L)=V^{Open}(L)=V^{Ind}(L),\ L=L^0. \tag{3.23}$$

Hence, when starting from passive states $L\in\mathcal{L}$ *and for values of the switching costs above a certain threshold, the Whittle index policy is optimal.* ***q.e.d.*** ■

The informal discussion along the previous lines is closely related with the more formal discussion of similar problems in Van Oyen et.al. [103] and Benkherouf et.al. [13].

3.3.2 Extension: The Multi-armed Bandit w/ Switching Costs and Losses

Consider the following extension to the pure set-up cost formulation of the multiarmed bandit with switching costs: whenever the arm is inactive (i.e. it is at some state $(0,x)$), there exists a positive probability $\theta(x)$ for the arm to abandon the system. This means that the arm enters some *terminal* or *absorbing* state Δ and remains there indefinitely. In such Δ state, only passive action is admitted. The evolution of the modified problem is sketched below[7], where ϵ is a small constant.



In this slightly modified version of Figure 3.1 we present the effect of including the *abandon* probability in our basic model. In this setting, every time the passive action is taken in a state (a_{-1}, x) there exists a positive probability ϵ of the arm to abandon the system, i.e. a transition to *absorbing* state Δ.

Figure 3.3: Representation of a Project in the Restless Bandit Formulation of the Muti-armed Bandit with Switching Costs and Looses

The elements of bandit (L, P, R, β) as described in §1' to §6' in page 89 remain the same, except from:

4'. **Transition Probabilities**

[7] For the sake of simplicity in this figure it has been assumed that parameter $\theta(x) = \epsilon$.

The passive transition rule is now given by

$$P^0(x,x') = \begin{cases} \theta(x), & x' = \Delta \\ 1-\theta(x), & x' = x \\ 1 & x = x' = \Delta \\ 0, & \text{any other case} \end{cases}$$

With this modification, optimality equations (3.5) becomes

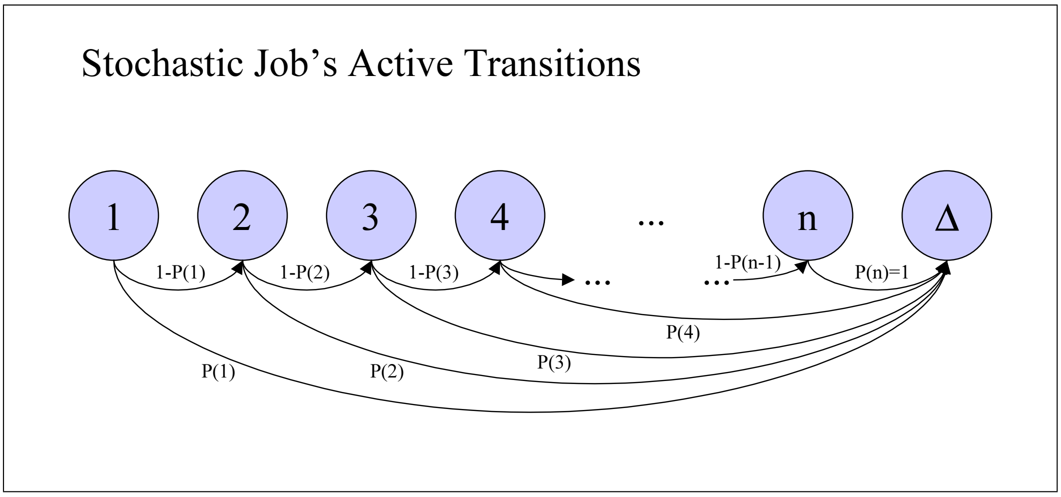
$$\begin{aligned} V\big((1,x),W\big) &= \max\Big\{R(x) + \beta E\left[V\big((1,y),W\big)\right]; \\ &\qquad W + \beta\theta(x)V\big(\Delta,W\big) + \beta(1-\theta(x))V\big((0,x),W\big)\Big\} \\ &= \max\Big\{R(x) + \beta E\left[V((1,y),W)\right]; \\ &\qquad \left[W(1-\beta+\beta\theta(x))\right](1-\beta)^{-1} + \beta(1-\theta(x))V\big((0,x),W\big)\Big\} \end{aligned} \tag{3.24}$$

with (3.24) using the identity $V(\Delta,W) = \frac{W}{1-\beta}$. Optimality equation (3.6) remains unchanged. The Gittins indices for activity remain as given in (3.7) and (3.8). This implies that Lemma 3.2 holds here as well, consequently Theorems 3.1 and 3.2 continue to hold.

3.3.3 Application: Scheduling of Stochastic Jobs

Consider one particular case of the global formulation of the MAB with switching costs: the problem of scheduling a set stochastic jobs where a job's natural state is the amount of past processing. In state x the job will either be not completed when processed and advance to state $x+1$ (with probability $1-p(x)$) or be completed (and enter completion state Δ) with probability $p(x)$. We assume there exists a *final* state for which completion is certain.

The natural state space of a job is thus extended by the inclusion of the completion (absorbing) state Δ, for which only passive action is available. The dynamic behavior of one job can be sketched as follows:



In the standard scheduling of stochastic jobs problem, whenever the project is activated at any state x, it can be either completed (with positive probabiity) or move forward to a more advanced stage. Completion is modelled by including an *absorbing* state Δ.

Figure 3.4: Representation of a Project in the Multi-armed Bandit Formulation of the Scheduling of Stochastic Jobs Problem with Switching Costs

In this case, the reward will depend on the state of the system at the following decision epoch, hence, we let $r_t(x, a, x')$ to denote the value at time t of the rewards received when the state of the system at decision epoch t is x, action $a \in \{1, 0\}$ is selected and the system occupies state x' at decision epoch $t + 1$.

The expected reward at any time t and state $x \in S$ may be evaluated by computing

$$r(x, a) = \sum_{x' \in S} r(x, a, x') P^a(x, x')$$

In the case of a problem with just active rewards, and explicitly considering the completion probability, we get the following expression:

$$r(x,a) = \begin{cases} r(x,1,x+1)(1-P(x)) + r(x,1,\Delta)P(x), & a=1 \\ 0, & a=0 \end{cases}$$

In a formulation where a fixed (job depending) reward R is earned only upon completion, the expression above simplifies to:

$$r(x,a) = r(x,1,\Delta)P(x) = P(x)R$$

For the *switching costs* case, the state space is increased by including information about the action taken before, and the dynamic of the system becomes as sketched below.

The elements of bandit (L,P,R,β) as described in §1' to §6' in page 89 remain the same, except from:

4'. **Transition Probabilities**

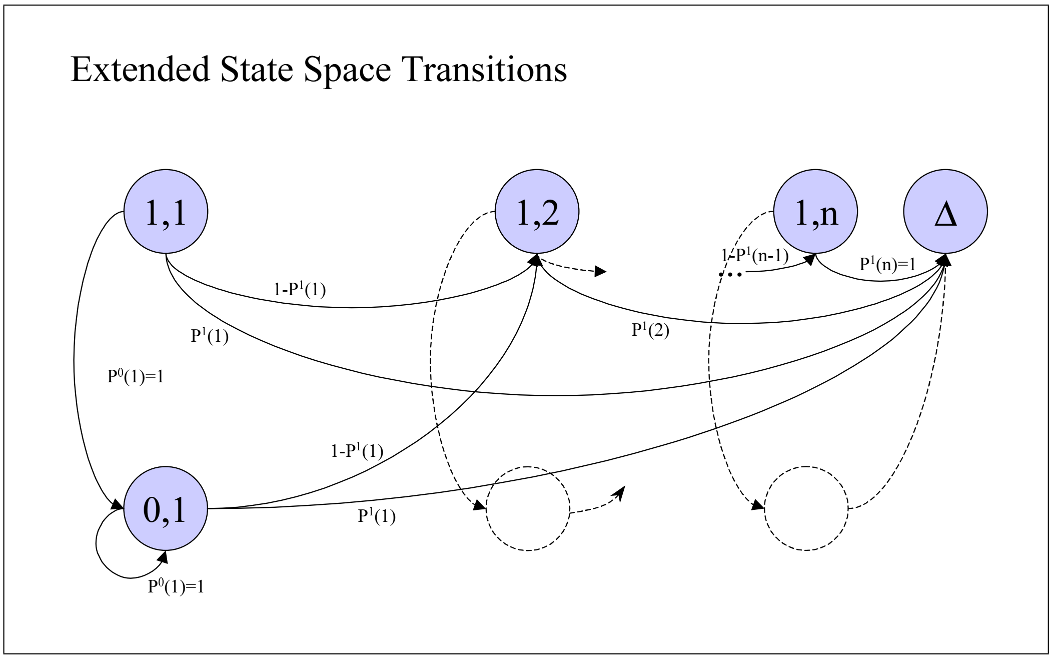
Let $S = \{0,\ldots,s,\Delta\}$ be the natural (finite) state space for an isolated job. The active transition rule is now given by

$$P^1(x,x') = \begin{cases} 1-p(x), & x'=x+1,\ x<s \\ p(x), & x'=\Delta,;x<s \\ 1, & x'=\Delta,\ x=s \\ 0, & \text{any other case} \end{cases} \tag{3.25}$$

Passive transition probabilities remain the same as described in §4', page 89.

5'. **Active Rewards**

As we have only completion rewards, the reward structure is straightfor-



When switching costs are considered, the state space is modified in the same way as we did for the simple multi-armed bandit with switching costs, i.e. a pair of *extended states* $(1, x), (0, x)$ are created for every natural state x. In the scheduling of stochastic jobs problem, active action causes a transition from state (a_{-1}, x) either to state $(1, x+1)$ or to *completion* state Δ. Passivity has the same effect as in the discussion around Figure 3.1

Figure 3.5: Representation of a Project in the Restless Bandit Formulation of the Scheduling of Stochastic Jobs Problem with Switching Costs

ward:

$$R\left(a_{-1}, x\right) = \begin{cases} Rp\left(x\right), & a_{-1} = 1 \\ Rp\left(x\right) - C\left(x\right), & a_{-1} = 0 \end{cases}, x \in S. \tag{3.26}$$

In this case, the general formulation of the problem remains the same and, consequently, all results in Section 3.2.3 hold here as well. Indexability of the stochastic jobs scheduling problem with switching costs, as formulated here, is therefore guaranteed.

3.4 Numerical Study

The numerical study has been conceived as a set of experiments, each of them performed over a fixed collection of examples, with the aim of comparing the performance of the index policy relative to the optimal one. As both the state space and the number of arms is necessarily small, we constraint our numerical study to the simples case of $N = 1$, i.e. the number of arms to activate at each decision epoch equals one.

For the general stochastic scheduling problem with switching costs formulation, three different sets of exercises were considered: 1) a four armed bandit with $s = 3$ natural states in each arm; 2) a four armed bandit with three natural states plus one absorbing state (for the case of the MAB w/SC and losses), i.e. $s = 4$; and 3) a four armed bandit with $s = 4$ natural states in each arm. Cases (1) and (2) were analysed assuming both, arm dependent and state dependent switching costs.

Regarding the scheduling of stochastic jobs application, two different problems are discussed: one considering increasing completion probabilities and another one with a general completion probability structure. As for computational reasons the state space must be kept necessarily small, one more transition is allowed: the active arm is allowed to remain in the same state (just for increasing the expected number of transitions before completion). Probabilities for staying in the same state are assumed to be small (below 10%). The number of arms has been fixed to be equal to three. This implies a total of seven different "problems" to be discussed.

Each example implies finding by Value Iteration (VI)[8] the optimal policy over the relevant state space (see below) and calculating, also by VI, the value

[8] For a discussion on the algorithm used for the numerical experiments in this chapter, please refer to Section 2.1.4.

of the index policy. The DP equations of the restless bandit version of the MAB problem with switching costs are given by:

$$V^{Opt}\left(\mathbf{L}\right)=\underset{\mathbf{a}\in\mathcal{A}_{\mathbf{L}}}{max}\left\{\mathcal{R}_{\mathbf{L}}^{\mathbf{a}}+\beta\left(\sum_{\mathbf{L}'\in\mathcal{L}}\mathcal{P}_{\mathbf{L},\mathbf{L}'}^{\mathbf{a}}V\left(\mathbf{L}'\right)\right)\right\},\quad\text{for all }\ \mathbf{L}\in\mathcal{L}.$$

The size of the experiments is measured as follows: there are $\left(2\times s\right)^{M}$ extended states; however, as we are focusing on the problem of activating just one arm at each decision epoch, the underlying state space cardinality reduces to $|\mathcal{L}^1| = \frac{1}{2}\left(2\times s\right)^{M}$, where

$$\mathcal{L}^1=\left\{\left(\left(a_{-1}^1,x_1\right),\ldots,\left(a_{-1}^M,x_M\right)\right)\middle|\sum_{i=1}^{M}a_{-1}^i\leq 1\right\},$$

likewise, there is a subset of initial or passive states ($\mathcal{L}^0\subset\mathcal{L}$), consisting in all those states where all arms are idle, i.e.

$$\mathcal{L}^0=\left\{\left(\left(a_{-1}^1,x_1\right),\ldots,\left(a_{-1}^M,x_M\right)\right)\middle|\sum_{i=1}^{M}a_{-1}^i=0\right\}$$

with cardinality $|\mathcal{L}^0| = \left(s\right)^{M}$. The analysis has been just focused in the subset of initial values, but as some other states are also visited, the VI has been performed over the $\mathcal{L}^1$ state space. The number of iterations has been defined following the standard convergence criterion:

$$\left\|V^{k+1}-V^{k}\right\|\leq\frac{\left(1-\beta\right)}{2\beta}\epsilon,$$

with discount factor $\beta = 0,9$, and parameter $\epsilon = 0.001$. In average, convergence has been achieved after 160 iterations.

3.4.1 Description of the Numerical Study

Problem. A problem is a complete collection of 25 experiments with 150 examples each, i.e. 3750 repetitions.

Example. Each example consists in an active reward $(M \times s)$-matrix, and two, active and passive transition, n-matrices. Reward matrices are random arrays of uniformly distributed data with parameters $a = 200$ and $b = 250$, whereas the active transition matrix is simply a random markovian probability matrix, and the passive one is the identity. There is a collection of 150 different examples.

For the *losses* case, one extra zero-row was added to the reward matrix (corresponding to the absorbing *abandon* state in which only the passive action is available). The active transition matrix was increased by adding one row and one column with all elements equal zero but position $(s + 1, s + 1)$, which equals one. The passive transition matrix is a $(s + 1)$-*matrix* with $p_{i,i} = 1 - \delta(i)$ and $p_{i,s+1} = \delta(i)$ and zero anywhere else. Where the abandon probability $\delta(i)$ is a uniformly distributed random variable with parameters $a = 0$ and $b = 0.025$.

Finally, for the job scheduling examples, completion reward has been obtained as the product of an uniformly distributed random variable with parameters $a = 200$ and $b = 250$, times the completion probability at each state. Completion probabilities $p(x)$, are uniformly distributed $(0, 1)$ random variables. Transitions to state $x + 1$ are given by the expression $(1 - p(x)) \times (1 - \pi(x))$, and transitions to state x itself are given by $(1 - p(x)) \times \pi(x)$, were $\pi(x) \sim U[0, 0.1]$. Transition from state s to completion state Δ is fixed equal to one.

Experiment. An experiment consists in obtaining the optimal and index solutions to the DP formulation of the restless bandit problem, all over the 150 examples. Each experiment is defined by the ratio of switching costs to active reward (SC/R). Twenty five experiments of each class (plain and with losses) were run with switching costs ranging from 1% up to 25% of the active reward.

For the *losses* case, the switching cost of the absorbing state was assumed to be equal to the corresponding percentage of the maximum reward available (250); for example, for a ratio $SC/R = 1\%$ (which actually corresponds to experiment number one), the switching cost for state $s+1$ equals 2.5. By means of this, the VI procedure is prevented from taking active action in the absorbing *abandon* state.

Performance Measure. The performance measure is the percentage relative error, or percentage suboptimality, of the index policy's value with respect to the optimal policy $(1 - \frac{V^{Ind}}{V^{Opt}})$. Two different sets of variables had been stored: 1) maximum, minimum and average relative error all over the $\frac{1}{2}(2 \times s)^4$ relevant states; and 2) maximum, minimum, median, average and standard deviation over s^4 initial states. Two additional variables store the number of mistakenly activated arms (MAA) in each example: one for the whole state space, $\mathcal{L}^1$, and one more for the initial (passive) states' subspace.

Output Three different output files were generated, ranking from individual example/experiment run, to the problem's aggregated summary.

An individual output file is available for each example in an experiment. This file stores the main parameters of the exercise (β, convergence criterion for VI ϵ, and SC/R ratio), as well as the active reward and switching costs matrices, active and passive transition probability matrices, and the

Gittins' indices resulting from that particular combination of active reward and switching costs. The output file also includes the summary of performance statistics (see above) over the whole state space in that particular example, as well as the detail of optimal and index policies performance for each state (variables are defined for both percentage suboptimality and a 0/1 variable taking value 1 whenever the prescribed policy for each state differs among index and optimal solution).

In the intermediate level, one output file is generated for each experiment. It includes information about the SC/R ratio of the experiment and the performance summary for each of the 150 examples, each extracted from the output file above. Finally, averages over the 150 examples are taken for each of the ten underlying variables and a summary is included.

Finally, a summary table is created for each *Problem* including the summary statistics of 25 experiments. The first three columns include the summary statistics for state subset $\mathcal{L}^1$, each variable is the average over 150 examples in each particular experiment. Next three columns include the same variables but constrained to the smaller subset $\mathcal{L}^0$. Finally, last two columns represent the average number of errors made by index policy all over the 150 examples in each particular experiment in sets $\mathcal{L}^1$ and $\mathcal{L}^0$, respectively.

3.4.2 Comments

The reader should observe the outstandingly good performance of the Whittle index policy throughout Tables 3.1 to 3.6. Note that the percentage suboptimalities grow as one moves down the tables, achieving a maximum when the switching costs are around 6/7% of the corresponding active rewards in almost every case. From there, the precentage suboptimalities go to zero as the switching costs increase further. To understand this behaviour note that when switching costs are zero, the Whittle index policy becomes a Gittins index policy which in Problems I, III, V and VI is optimal (considering only the subset of initial states). As switching costs become large, optimal policies have the feature that each bandit is activated from passivity at most once over the decision horizon. The switching cost then essentially becomes a *set up cost* for the bandit, to be incurred only once. When this happens, we can expliot Gittins index theory to show that the Whittle index policy must be optimal (for further details please see Section 3.3.1). Problems II and IV, with losses also follow this pattern.

Notice that in all our examples, the average performance of the index policy is above 99.95% optimality, and the worst performance is above 99.5% optimality. Moreover, the number of errors incurred by the index policy (number of wrongly activated arms) is always less than 15% of the total numer of states. Notwithstanding index policy preforms pretty bad in the job scheduling examples when the complete state space is considered, performance in the subset of initial states remains below the 0.05% suboptimality.

Table 3.1: Performance of the Index Policy in the Multi-armed Bandit with Switching Costs

	PROBLEM I							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.0827	0.0058	0.0148	0.0293	0.0040	0.0086	35.3867	5.4533
2	0.1577	0.0144	0.0315	0.0637	0.0106	0.0201	40.4333	6.2600
3	0.2015	0.0201	0.0420	0.0868	0.0150	0.0279	40.8533	6.3667
4	0.2335	0.0225	0.0481	0.0942	0.0163	0.0302	35.1000	4.9333
5	0.2475	0.0231	0.0517	0.0951	0.0159	0.0306	30.8933	4.2867
6	0.2457	0.0243	0.0534	0.0979	0.0162	0.0316	26.0667	3.5067
7	0.2301	0.0239	0.0525	0.0994	0.0164	0.0327	21.5733	2.9333
8	0.2157	0.0213	0.0481	0.0928	0.0145	0.0301	17.1522	2.1600
9	0.1608	0.0155	0.0363	0.0663	0.0099	0.0210	11.8400	1.2933
10	0.1337	0.0115	0.0289	0.0416	0.0058	0.0129	9.7667	1.0667
11	0.1059	0.0088	0.0228	0.0305	0.0039	0.0090	8.2067	0.9400
12	0.1089	0.0089	0.0239	0.0295	0.0036	0.0081	6.8200	0.7600
13	0.1083	0.0084	0.0230	0.0209	0.0032	0.0068	5.0267	0.4333
14	0.0755	0.0058	0.0159	0.0176	0.0023	0.0055	3.5467	0.2600
15	0.0643	0.0039	0.0124	0.0100	0.0009	0.0028	2.5733	0.1400
16	0.0371	0.0025	0.0074	0.0071	0.0009	0.0022	1.9000	0.1600
17	0.0426	0.0026	0.0083	0.0087	0.0011	0.0026	1.4267	0.0400
18	0.0253	0.0015	0.0049	0.0027	0.0003	0.0008	0.9467	0.0000
19	0.0256	0.0013	0.0048	0.0000	0.0000	0.0000	0.7400	0.0000
20	0.0094	0.0004	0.0018	0.0000	0.0000	0.0000	0.2467	0.0000
21	0.0006	0.0000	0.0001	0.0000	0.0000	0.0000	0.1667	0.0000
22	0.0005	0.0000	0.0001	0.0000	0.0000	0.0000	0.1200	0.0000
23	0.0009	0.0000	0.0001	0.0000	0.0000	0.0000	0.1867	0.0000
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Formulation with 4 arms, 3 natural states per arm and state-dependent switching costs.

Table 3.2: Performance of the Index Policy in the Multi-armed Bandit with Switching Costs and Losses

	PROBLEM II							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.2016	0.0108	0.0276	0.1223	0.0097	0.0229	93.2867	15.9133
2	0.2811	0.0151	0.0392	0.1534	0.0125	0.0301	96.7533	15.8200
3	0.3340	0.0177	0.0479	0.1838	0.0142	0.0368	94.3467	14.9133
4	0.3792	0.0192	0.0538	0.1969	0.0147	0.0394	92.0800	14.8800
5	0.4102	0.0203	0.0589	0.2136	0.0151	0.0417	85.6933	14.2733
6	0.4187	0.0209	0.0611	0.2152	0.0150	0.0414	75.0533	12.2133
7	0.4294	0.0200	0.0616	0.2205	0.0144	0.0424	62.3267	9.6333
8	0.4149	0.0178	0.0569	0.1932	0.0123	0.0375	56.0867	8.6533
9	0.3876	0.0139	0.0488	0.1611	0.0089	0.0293	46.1467	6.5600
10	0.3565	0.0106	0.0414	0.1096	0.0056	0.0195	39.2867	5.3800
11	0.3163	0.0082	0.0341	0.0843	0.0037	0.0136	32.2133	3.6400
12	0.3094	0.0079	0.0343	0.0678	0.0029	0.0110	31.6867	3.6867
13	0.2839	0.0075	0.0321	0.0519	0.0027	0.0092	28.0533	2.6200
14	0.2549	0.0057	0.0267	0.0383	0.0019	0.0065	23.3733	1.5533
15	0.2283	0.0043	0.0223	0.0320	0.0011	0.0047	21.4533	1.3400
16	0.1955	0.0032	0.0176	0.0227	0.0011	0.0038	19.0467	1.3067
17	0.1848	0.0031	0.0166	0.0176	0.0009	0.0033	17.2267	0.8333
18	0.1419	0.0021	0.0123	0.0115	0.0004	0.0018	14.7600	0.4667
19	0.1497	0.0018	0.0116	0.0019	0.0001	0.0003	14.3533	0.2933
20	0.1131	0.0013	0.0092	0.0012	0.0000	0.0001	11.2667	0.1400
21	0.1043	0.0012	0.0086	0.0007	0.0000	0.0002	11.9867	0.2400
22	0.0985	0.0012	0.0081	0.0009	0.0000	0.0001	12.6533	0.1667
23	0.0873	0.0012	0.0077	0.0006	0.0000	0.0001	13.1867	0.2467
24	0.0849	0.0012	0.0076	0.0003	0.0000	0.0001	12.2133	0.1133
25	0.0834	0.0013	0.0078	0.0000	0.0000	0.0000	12.4333	0.0000

Formulation with 4 arms, 3 natural states per arm and state-dependent switching costs.

Table 3.3: Performance of the Index Policy in the Multi-armed Bandit with Switching Costs

	PROBLEM III							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.0772	0.0058	0.0142	0.0289	0.0043	0.0087	34.0000	5.5667
2	0.1617	0.0131	0.0299	0.0634	0.0098	0.0198	40.0867	6.4867
3	0.1926	0.0186	0.0403	0.0842	0.0138	0.0268	38.6333	5.8000
4	0.2121	0.0232	0.0503	0.1028	0.0171	0.0335	35.5333	5.2600
5	0.2572	0.0263	0.0573	0.1106	0.0180	0.0359	30.8133	3.8133
6	0.2419	0.0259	0.0545	0.1039	0.0183	0.0350	26.2933	3.2867
7	0.2195	0.0230	0.0508	0.0914	0.0162	0.0318	20.4200	2.1667
8	0.2108	0.0193	0.0474	0.0679	0.0119	0.0237	15.8133	1.4400
9	0.1997	0.0177	0.0442	0.0540	0.0098	0.0193	13.3067	1.1200
10	0.1759	0.0146	0.0382	0.0453	0.0077	0.0156	10.3867	0.8000
11	0.1299	0.0101	0.0280	0.0312	0.0047	0.0103	7.9267	0.6400
12	0.1268	0.0099	0.0279	0.0278	0.0039	0.0091	7.2333	0.5200
13	0.1074	0.0086	0.0239	0.0208	0.0031	0.0069	6.0600	0.4000
14	0.1203	0.0092	0.0263	0.0225	0.0029	0.0071	5.2000	0.2600
15	0.0833	0.0059	0.0175	0.0185	0.0024	0.0058	4.0400	0.2000
16	0.0733	0.0053	0.0152	0.0148	0.0019	0.0047	3.4933	0.2000
17	0.0551	0.0048	0.0119	0.0170	0.0027	0.0057	2.2533	0.1400
18	0.0468	0.0027	0.0089	0.0064	0.0005	0.0016	1.3067	0.0000
19	0.0408	0.0026	0.0080	0.0043	0.0009	0.0017	0.9667	0.0600
20	0.0312	0.0017	0.0059	0.0000	0.0000	0.0000	0.5667	0.0000
21	0.0276	0.0015	0.0054	0.0000	0.0000	0.0000	0.4133	0.0000
22	0.0307	0.0017	0.0062	0.0000	0.0000	0.0000	0.3733	0.0000
23	0.0216	0.0009	0.0040	0.0000	0.0000	0.0000	0.3000	0.0000
24	0.0176	0.0006	0.0029	0.0000	0.0000	0.0000	0.1267	0.0000
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Formulation with 4 arms, 3 natural states per arm and arm-dependent switching costs.

Table 3.4: Performance of the Index Policy in the Multi-armed Bandit with Switching Costs and Losses

	PROBLEM IV							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.2009	0.0115	0.0275	0.1155	0.0102	0.0227	89.7000	15.5333
2	0.2796	0.0158	0.0398	0.1540	0.0134	0.0313	100.6533	17.5867
3	0.3412	0.0193	0.0495	0.1901	0.0160	0.0385	97.8800	16.0067
4	0.3927	0.0216	0.0572	0.2141	0.0169	0.0434	90.7733	14.6933
5	0.4265	0.0222	0.0626	0.2282	0.0165	0.0455	81.2267	11.8400
6	0.4463	0.0215	0.0635	0.2247	0.0160	0.0450	75.4733	11.6667
7	0.4510	0.0192	0.0607	0.2141	0.0140	0.0416	63.0133	9.1067
8	0.4403	0.0165	0.0569	0.1875	0.0109	0.0344	52.3267	8.4133
9	0.4636	0.0157	0.0573	0.1763	0.0089	0.0305	44.7333	4.9533
10	0.4534	0.0131	0.0531	0.1568	0.0069	0.0255	38.6800	4.3733
11	0.4035	0.0098	0.0436	0.1152	0.0047	0.0177	32.5667	3.6400
12	0.3799	0.0091	0.0417	0.1031	0.0036	0.0150	31.0800	3.3000
13	0.3383	0.0079	0.0365	0.0874	0.0028	0.0118	26.4733	2.2600
14	0.3337	0.0081	0.0365	0.0719	0.0027	0.0107	24.2933	1.7533
15	0.2732	0.0059	0.0289	0.0543	0.0022	0.0086	22.0033	1.5733
16	0.2427	0.0055	0.0259	0.0474	0.0021	0.0078	19.3400	1.4600
17	0.2245	0.0048	0.0224	0.0359	0.0022	0.0068	18.1933	1.4267
18	0.2077	0.0034	0.0190	0.0218	0.0008	0.0035	16.8533	1.1933
19	0.1891	0.0029	0.0167	0.0122	0.0008	0.0025	14.0667	0.5867
20	0.1657	0.0024	0.0149	0.0065	0.0003	0.0011	12.8000	0.5200
21	0.1491	0.0023	0.0136	0.0070	0.0003	0.0012	13.0600	0.4467
22	0.1471	0.0022	0.0135	0.0073	0.0002	0.0009	12.7200	0.3533
23	0.1380	0.0018	0.0116	0.0069	0.0003	0.0009	12.4667	0.4933
24	0.1117	0.0016	0.0099	0.0080	0.0004	0.0012	12.1933	0.3267
25	0.0949	0.0012	0.0078	0.0077	0.0002	0.0010	12.1333	0.2267

Formulation with 4 arms, 3 natural states per arm and arm-dependent switching costs.

Table 3.5: Performance of the Index Policy in the Multi-armed Bandit with Switching Costs

	PROBLEM V							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.1139	0.0086	0.0190	0.0462	0.0062	0.0119	114.3667	21.1533
2	0.2171	0.0204	0.0399	0.0988	0.0153	0.0279	175.4467	27.0467
3	0.2901	0.0286	0.0548	0.1397	0.0223	0.0403	165.0200	26.1600
4	0.3380	0.0315	0.0625	0.1585	0.0246	0.0458	149.9933	23.7067
5	0.3512	0.0328	0.0669	0.1691	0.0255	0.0497	134.3733	21.7667
6	0.3601	0.0339	0.0710	0.1807	0.0258	0.0529	118.9067	18.9667
7	0.3369	0.0325	0.0701	0.1785	0.0240	0.0513	97.4867	15.4333
8	0.3202	0.0297	0.0662	0.1534	0.0199	0.0438	79.7200	11.5467
9	0.3175	0.0276	0.0634	0.1434	0.0183	0.0408	61.0800	8.5800
10	0.2628	0.0201	0.0516	0.1043	0.0114	0.0278	41.2800	4.8400
11	0.1985	0.0149	0.0387	0.0837	0.0089	0.0219	30.4333	3.6867
12	0.1624	0.0121	0.0313	0.0641	0.0064	0.0163	23.9200	2.6400
13	0.1458	0.0092	0.0274	0.0464	0.0042	0.0120	17.3133	1.7677
14	0.1171	0.0068	0.0202	0.0276	0.0027	0.0067	11.4333	1.1067
15	0.0816	0.0046	0.0141	0.0188	0.0014	0.0042	6.8800	0.2133
16	0.0518	0.0018	0.0074	0.0012	0.0001	0.0002	4.5333	0.1067
17	0.0405	0.0017	0.0064	0.0047	0.0003	0.0011	3.5733	0.2133
18	0.0348	0.0016	0.0059	0.0038	0.0002	0.0009	2.2467	0.0000
19	0.0186	0.0007	0.0029	0.0000	0.0000	0.0000	1.3333	0.0000
20	0.0223	0.0009	0.0036	0.0000	0.0000	0.0000	1.7733	0.0000
21	0.0238	0.0008	0.0037	0.0000	0.0000	0.0000	1.3067	0.0000
22	0.0212	0.0008	0.0037	0.0000	0.0000	0.0000	0.6400	0.0000
23	0.0108	0.0004	0.0018	0.0000	0.0000	0.0000	0.4800	0.0000
24	0.0135	0.0005	0.0023	0.0000	0.0000	0.0000	0.5867	0.0000
25	0.0039	0.0001	0.0004	0.0000	0.0000	0.0000	0.4800	0.0000

Formulation with 4 arms, 4 natural states per arm and state-dependent rewards.

Table 3.6: Performance of the Index Policy in the Scheduling of Stochastic Jobs Problem

	PROBLEM VI							
SC/R	Complete State Space			Initial States			Bad Arms	
(%)	Max	Avg	StDev	Max	Avg	StDev	CSS	IS
1	0.4428	0.0080	0.0434	0.0010	0.0000	0.0001	57.6867	0.08667
2	0.9525	0.0303	0.1215	0.0016	0.0000	0.0002	101.6867	0.09333
3	1.4642	0.0629	0.2159	0.0021	0.0000	0.0002	135.9600	0.09333
4	1.9872	0.1044	0.3226	0.0025	0.0000	0.0003	163.8200	0.04667
5	2.5288	0.1525	0.4366	0.0022	0.0000	0.0002	187.6533	0.04667
6	3.0817	0.2079	0.5598	0.0007	0.0000	0.0001	208.2533	0.04667
7	3.6472	0.2676	0.6865	0.0004	0.0000	0.0001	225.4200	0.04667
8	4.2106	0.3323	0.8196	0.0002	0.0000	0.0000	239.8533	0.00000
9	4.7665	0.4026	0.9597	0.0000	0.0000	0.0000	253.9800	0.00000
10	5.3549	0.4763	1.1036	0.0000	0.0000	0.0000	266.1333	0.00000
11	5.9710	0.5564	1.2544	0.0000	0.0000	0.0000	276.7867	0.00000
12	6.5812	0.6424	1.4149	0.0000	0.0000	0.0000	287.1333	0.00000
13	7.2006	0.7255	1.5716	0.0000	0.0000	0.0000	292.9600	0.00000
14	7.8236	0.8131	1.7344	0.0000	0.0000	0.0000	299.4067	0.00000
15	8.4788	0.9031	1.9021	0.0000	0.0000	0.0000	304.8800	0.00000
16	9.1045	0.9977	2.0757	0.0000	0.0000	0.0000	309.6800	0.00000
17	9.7918	1.0936	2.2520	0.0000	0.0000	0.0000	313.9000	0.00000
18	10.4669	1.1897	2.4312	0.0000	0.0000	0.0000	316.0200	0.00000
19	11.1353	1.2831	2.6093	0.0000	0.0000	0.0000	317.1733	0.00000
20	11.7651	1.3784	2.7929	0.0000	0.0000	0.0000	317.4600	0.00000
21	12.5024	1.4764	2.9817	0.0000	0.0000	0.0000	317.7133	0.00000
22	13.2043	1.5827	3.1819	0.0000	0.0000	0.0000	318.2000	0.00000
23	13.9373	1.6784	3.3749	0.0000	0.0000	0.0000	316.8667	0.00000
24	14.6423	1.7772	3.5729	0.0000	0.0000	0.0000	315.2933	0.00000
25	15.4023	1.8674	3.7664	0.0000	0.0000	0.0000	312.0800	0.00000

Formulation with 3 arms, 6 natural states per arm and state-dependent swithcing costs.

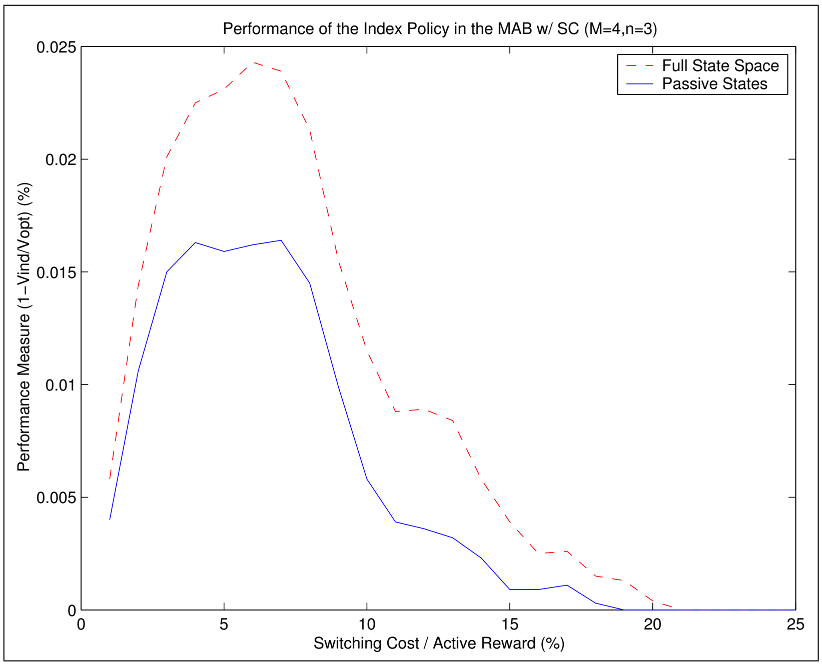


Figure 3.6: Performance Plot, Problem I

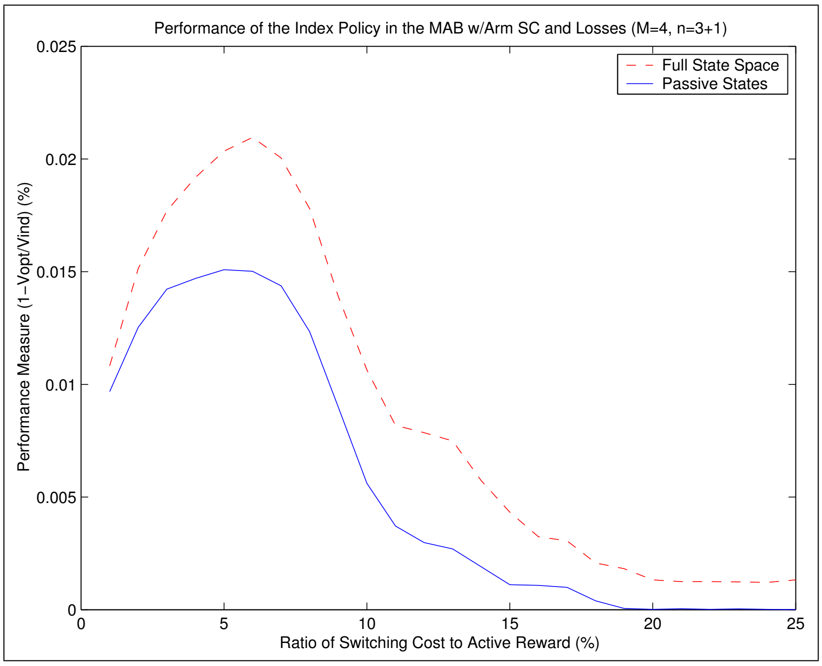


Figure 3.7: Performance Plot, Problem II

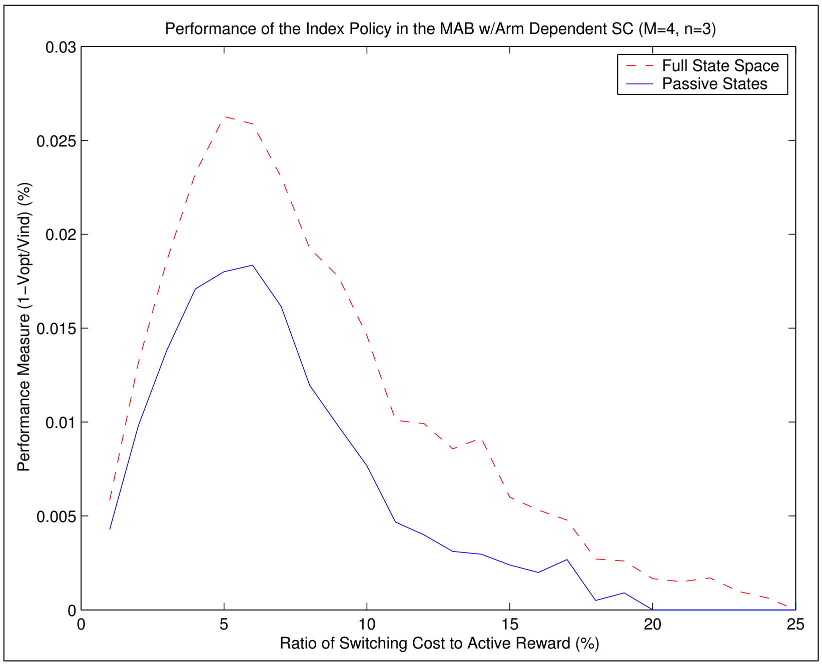


Figure 3.8: Performance Plot, Problem III

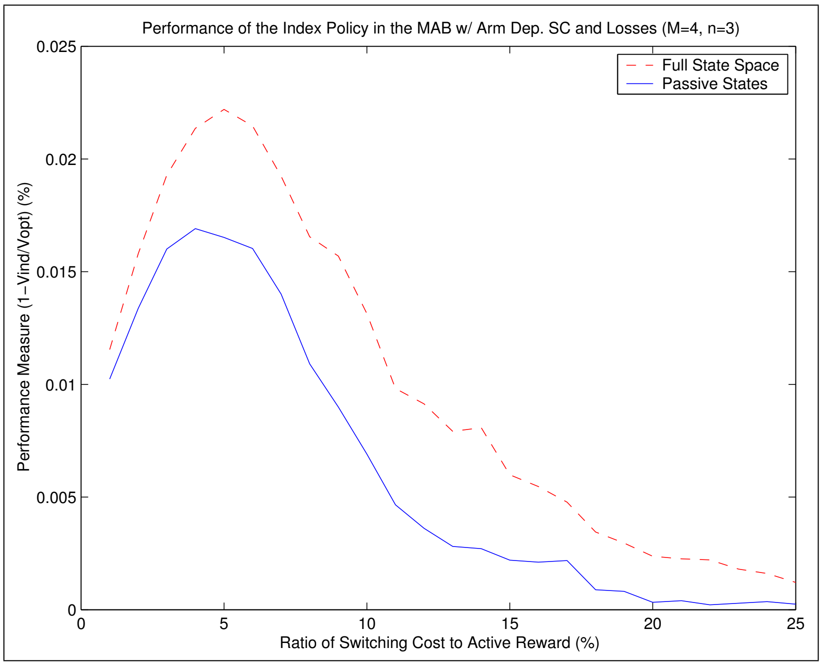


Figure 3.9: Performance Plot, Problem IV

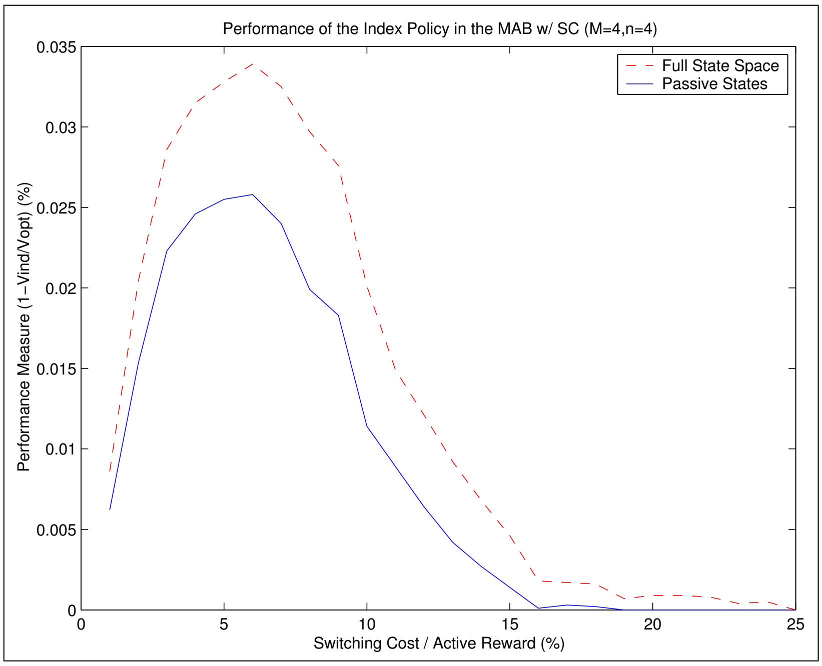


Figure 3.10: Performance Plot, Problem V

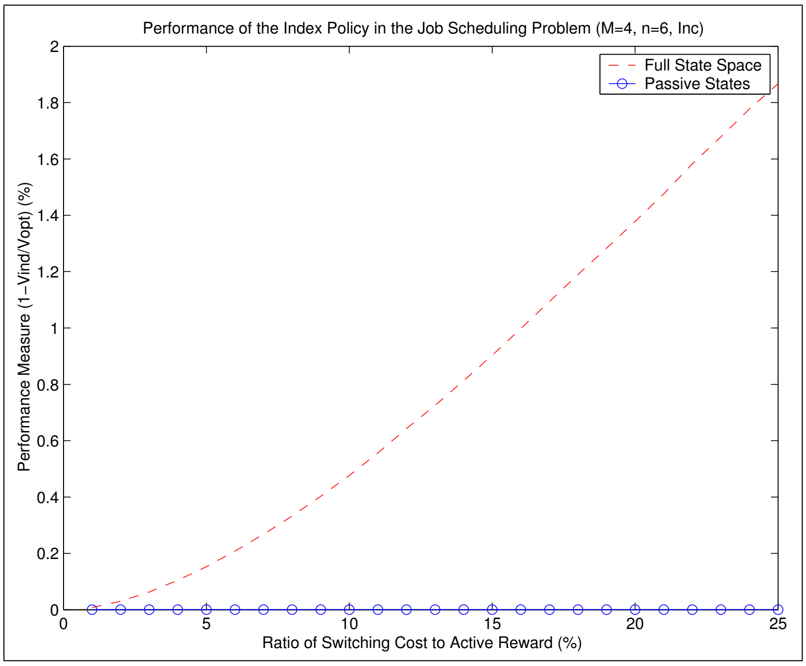


Figure 3.11: Performance Plot, Problem VI

3.5 Conclusions

The multi-armed bandit with switching costs described in Section 3.2 exhibits many desirable features that are of real concern in many contemporary applications. In particular it relaxes the general assumption maintained in the classical multi-armed bandit literature that the operator can switch instantaneously from one project to the other without facing any cost. Natural as this relaxation can be, the inclusion of a switching penalty changes the nature of the problem and, in fact, it has been shown that an index strategy is not necessarily optimal in the domain of all bandit problems with switching costs.

Following the ideas by Asawa and Teneketzis [6] and the intuition by José Niño-Mora that a multi-armed bandit with switching costs can be interpreted as a restless bandit, in Section 3.2.2 we discussed a *translation* of the multi-armed bandit problem with switching costs into a restless bandit by means of a natural extension of the state space. This extension consists in including, for each bandit and state, information about the action taken during the previous decision epoch. The outcome is hence a new two dimensional state space with two *extended states* corresponding to each *natural state* of the arm. Applying the standard methods discussed in Chapter 2 we established the indexability of the new problem and provided al algorithm for index computation.

In Section 3.3 we described a range of model developments which preserve indexability. These include: (a) a discussion around the optimality of the index policy when the switching penalties are big; (b) an extension of the model which considers a positive probability for a passive arm to abandon the system, and (c) the particular case of the problem of scheduling a set of stochastic jobs.

Finally, in the numerical study we provide evidence of the good performance of the index heuristic in the switching costs case, which confirms the adequacy of the restless bandit approach to the multi-armed bandit problem with switching costs.

A very appealing extension to the results obtained here consists in establishing a general form of switching-cost-indexability for all restless bandits for which indexability can be guaranteed. The main idea here is that, as long as the state space modification introduced in this chapter does not affect the general transition structure of the original problem, once indexability has been established for any particular family of restless bandit problems, it it should immediately follow that the version of the problem that includes switching costs is, indeed, indexable. We will refer to this kind of indexability as $SC-indexability$ and establishing this result as well as providing a numerical assessment of the performance of the corresponding index heuristic are subject of further research for the author.

Chapter 4

The Machine Maintenance Problem: A Family of Indexable Restless Bandits[1]

Introduction

As we have seen in Section 2.2, the classical index result of Gittins [35] and Gittins and Jones [37, 38] concerns the optimal allocation over time of a single key resource among a collection of projects (or bandits) which are in competition for it. Application of the resource to a project at any state earn a reward and causes a transition in project's state; however, if a project is not receiving effort, it earns nothing and its state remains the same. The optimization goal is the identification of a policy for sequentially selecting projects on the basis of current state information to maximise the expected total discounted reward over an infinite horizon. Gittins proposed a collection of *calibrating*

[1] The main results in this Chapter will appear published in Section 3 of Glazebrook, K., Ruiz-Hernández, D. and Kirkbride, C. Some indexable families of restless bandit problems. *Advances in Applied Probability 38-3* (2006).

indices (now known as *Gittins Indices*), one for each project, in the form of a real-valued function on the project's state space and showed that the *index policy* which always directs the key resource to (one of) the project(s) with the highest current index is indeed optimal.

However, the requirement in Gittins' models that passive projects should remain frozen inhibits its applicability in a wide range of practical situations. Motivated by this, Whittle [116] introduced a class of *restless bandit* problems which allow for state evolution among passive projects (see Section 2.3). Whittle's original analysis elucidated (under stated conditions) an index-based solution to a Langrangian relaxation of the restless bandit problem of interest in which the Lagrange multiplier has an economic interpretation as an *index for passivity* or a *charge for service.* The indices identified this way generalise those of Gittins, and Whittle proposed their use in the construction of heuristics for restless bandit problems. The importance of Whittle's contribution was emphasized by Weber and Weiss ([109] and [110]) who established a form of asymptotic optimality for the index heuristic and more recently in a range of empirical studies which have demonstrated its outstanding strong performance in various application domains. See, for example, Ansell, Glazebrook, Niño-Mora and O'Keeffe [5] and Glazebrook, Mitchell and Ansell [51]. Further Glazebrook, Niño-Mora and Ansell [52] have discussed the development of bounds on the degree of reward suboptimality of Whittle's index policy.

A major challenge to the deployment of Whittle's powerful ideas is that his index function is only defined for those projects which pass a test of *indexability.* Although this requirement seems plausible and natural, it can be very difficult to establish and, indeed, need not to hold. Even when indexability is established, there are few cases where the Whittle index (as we shall call it) is available in closed form. For an instance of the latter see the queueing control problem discussed by Ansell et al. [5]. The primary contribution of this chapter lies in the demonstration that a general class of restless bandit problems which arises in some major applications, to which we will refer as the *machine maintenance problem*, is indeed indexable. Having established indexability, we proceed to identify the corresponding Whittle indices. Our tools of analysis are (generally) those of stochastic dynamic programming (see Puterman [85], Ross [90] and discussion in Section 2.1 of this dissertation) and –more particularly- those of Gittins index theory (see [35, 36, 37]). In every case the Whittle index identified is given as a function of a corresponding Gittins index and/or in closed form. Given the ease with which Gittins indices may be computed, this is more than enough for implementation.

Working on the materials presented in section 2.3, where we introduced Whittle's notion of indexability along with a definition of his indices, in this chapter we present a class of restless bandit problems with a discounted reward criterion. In Section 4.1 we establish the indexability of a class of restless bandits designed to model machine maintenance problems in which maintenance interventions (active action) have to be scheduled to mitigate escalating costs as machines deteriorate (when passive action is taken). Whittle [117] and Glazebrook et.al. [51] have previously given index-based analyses of particular models, but we now show that indexability in guaranteed in general. In Section 4.2 we explore index structure in the context of two model types, both of which rest on assumptions that are plausible on practice. In Section 4.3 we further develop the findings of Section by offering two families of examples for which explicit formulae for the Whittle index can be derived. Identification of the Whittle indices of concern is followed in Section 4.4 by a numerical investigation which demonstrates the very strong performance of Whittle's heuristic. Section 4.5 concludes.

4.1 The Machine Maintenance Problem

In the light of the development in Section 2.3, it will be enough to conduct our discussions of indexability/indices by reference to individual bandits as in §1'$_{(\mathbf{RB})}$ to §6'$_{(\mathbf{RB})}$ (see pages 57 to 58). Our goal is to develop an indexability analysis of a general class of structured bandits designed to model *machine maintenance* problems. As flagged in comment §B in page 60, it will be convenient to discuss indexability in terms of the W^{+}-*problem* associated with the bandit.

4.1.1 Model Formulation

Consider the problem of giving maintenance to a set of M machines by a limited number of repairmen, Q. The problem is, hence, to choose at each decision epoch the subset of $Q < M$ machines to be repaired (in a further extension to our model we shall admit the possibility of a repairmen to be idle at some decision epoch and the decision would be to choose the subset of $\underline{Q} \leq Q < M$ machines to be served). The problem of scheduling maintenance is formulated as a *restless bandit problem* which arms evolve under passive action (operation/deterioration) and go back to some improved state under active action (intervention).

The standard Markov Decision Process (MDP) formulation for the *Machine Maintenance Problem* can be described as follows: Each machine $1 \leq i \leq M$ is modelled as a Markov Decision Process (MDC) that evolves over the (countable) state space S_i. There are two actions $a\,(t) \in \{0, 1\}$ available at each state $x_i \in S_i$ and decision time t, where $a\,(t) = 0$ means passivity (the machine is in operation and, consequently, deteriorating) and $a\,(t) = 1$, activity (intervention). If active action is taken in arm i at state $x_i \in S_i$ an arm/state dependent active cost is incurred and the machine performs a transition to some state $x_i' \in S_i$ according to certain active (intervention) transition rule. If instead passive action is taken, an operation cost is incurred, and the machine evolves to some state $x_i' \in S_i$ following the machine's passive (deterioration) transition rule.

The standard discrete time MDP formulation for this restless bandit problem, represented by $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{C}, M, Q,)$, is given by the following elements (corresponding to $\S 1_{(\mathbf{RB})}$ up to $\S 6_{(\mathbf{RB})}$ in Section 2.3).

i. **Decision Epochs**

Decisions are taken at epochs $t \in \mathbb{N}$.

ii. **State Space**

The set of all possible system states at decision epoch t is the Cartesian Product $\mathcal{S} = \bigtimes_{i=1}^{M} S_i$, with S_i the (countable) state space of machine i. The state of the process at time t is $\mathbf{X}(t) = \{X_1(t), \ldots, X_M(t)\}$, $X_i(t) \in S_i$. Designated state $\mathbf{0} \in S_i$ represents the (pristine) state of the machine following a maintenance intervention.

iii. **Action Set**

We assume $M > Q$. At each decision epoch the collection of $\binom{M}{Q}$ admissible actions at state $\mathbf{X} \in \mathcal{S}$ is given by the set:

$$\mathcal{A} = \left\{ \mathbf{a} = (a_1, \ldots, a_M) \,\middle|\, \sum_{i=1}^{M} a_i = Q, a_i \in \{0, 1\} \right\} \tag{4.1}$$

Action $a_i = 1$ (activity) implies that the machine i is *intervened* while under $a_i = 0$ (passivity) it is left to freely *evolve* (or deteriorate). Equation (4.1) indicates that an admissible action implies giving maintenance to exactly Q machines, while leaving the remaining $M - Q$ to deteriorate.

If we instead allow a subset $\underline{Q} \subseteq Q$ of machines to be intervened at some decision epoch, then the r.h.s. term between brakets in (4.1) becomes $\sum_{i=1}^{M} a_i = \underline{Q} \leq Q$.

iv. **Transition Probabilities**

Suppose action $\mathbf{a}(t) \in \mathcal{A}$ is taken at $t \in \mathbb{N}$. Under intervention $(a_i = 1)$, machine i is instantaneously returned to designated pristine state $\mathbf{0}$ after which it performs a single transition under Markov law P_i, i.e.

$$P\left\{x_i(t+1) = x_i' \left| x_i(t) = x_i, a_i(t) = 1\right.\right\} = P_i\left(\mathbf{0}, x_i'\right),\ x_i, x_i' \in S_i \quad (4.2)$$

Under passivity $(a_i(t) = 0)$, the machine deteriorates according to Markov law P_i, i.e.

$$P\left\{x_i(t+1) = x_i' \left| x_i(t) = x_i, a_i(t) = 0\right.\right\} = P_i\left(x_i, x_i'\right),\ x_i, x_i' \in S_i \quad (4.3)$$

As long as the M machines evolve independently, we can introduce the following notation:

$$\mathcal{P}^{\mathbf{a}}_{\mathbf{X},\mathbf{X}'} = \prod_{i:a_i=0} P_i\left(x_i, x_i'\right) \prod_{i:a_i=1} P_i\left(\mathbf{0}, x_i'\right), \quad \text{s.t. } x_i \in \mathbf{X},\ x_i' \in \mathbf{X}'; \quad (4.4)$$

for all $\mathbf{X}, \mathbf{X}' \in \mathcal{S}$. Let $\mathbf{P}$ represent the set of all transition probability matrices $\mathcal{P}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

v. **Cost Structure**

Let $C_i : S_i \rightarrow \mathbb{R}^+$ and $k_i : S_i^2 \rightarrow \mathbb{R}^+$ be bounded cost functions for all $1 \leq i \leq M$. If, for some machine i, a transition from state x_i to x_i' occurs under action $a_i = 0$ at time t, discounted operation cost $\beta^t k_i\left(x_i, x_i'\right)$ is incurred. Should a transition from x to $\mathbf{0}$ occur under action $a_i = 1$ at time t, a discounted *maintenance cost* $\beta^t C_i(x_i)$ is paid plus the instantaneous transition cost $k_i(\mathbf{0}, x_i')$ as described in the discussion around (4.2). Rewards are additive across machines and over time. We shall use the telescoped notation:

$$k_i(x_i) = \sum_{x_i' \in S_i} k_i(x_i, x_i') \, P_i(x_i, x_i'), \; x_i \in S_i$$

to denote the expected operational cost incurred by a single transition under the passive action.

Hence, the overall cost of policy $\mathbf{a} \in \mathcal{A}$ when taken in state $\mathbf{X} \in \mathcal{S}$ will be given by

$$\mathcal{C}_{\mathbf{X}}^{\mathbf{a}} = \sum_{M}^{i=1} \Big[k_i(x_i) \, 1\{a_i = 0\} + \big(C_i(x_i) + k_i(\mathbf{0})\big) 1\{a_i = 1\} \Big]. \tag{4.5}$$

Again, we use $\mathbf{C}$ for representing the collection of cost vectors $\mathcal{R}^{\mathbf{a}}$, $\mathbf{a} \in \mathcal{A}$.

vi. **Policy**

The goal of optimization is the choice of a policy π to minimise the total expected maintenance/operation cost incurred over an infinite horizon. The theory of dynamic programming (see Section 2.1 and references therein) asserts the existence of an optimal (cost minimising) policy which is stationary and satisfies the optimality equations.

As for (2.23), we write $V(\mathbf{X})$ for the *value function* of the process evaluated at $\mathbf{X} \in \mathcal{S}$, namely the minimal expected maintenance cost incurred over an infinite horizon starting from state $\mathbf{X}$. The optimality equations for the machine maintenance problem may be expressed as:

$$V(\mathbf{X}) = \min_{\mathbf{a} \in \mathcal{A}} \left\{ \mathcal{C}_{\mathbf{X}}^{\mathbf{a}} + \beta \sum_{\mathbf{X}' \in \mathcal{S}} \mathcal{P}_{\mathbf{X},\mathbf{X}'}^{\mathbf{a}} V(\mathbf{X}') \right\}, \quad \mathbf{X} \in \mathcal{S} \tag{4.6}$$

We can now further develop our formulation to make it consistent with the Whittle's approach as described in page 56. Following the discussion in §A (see page 60), we can equivalently define a W-*charge* version for the *machine maintenance* problem, where W is a charge for activity; yielding a wholly cost-base structure for our problem.

As indexability and indices are properties of individual bandits, as we did in Section 2.3, we isolate an individual bandit and will now drop the identifier i. The W^{+}-*problem* for bandit $(S, \mathcal{A}, P, C, k, \beta)$ is a cost discounted MDP with elements §i to §vi as described in page 133, but with the following modifications to §v and §vi:

v'. **Cost Structure**

If a transition from state x occurs under action $a = 0$ at time t a discounted operation cost $\beta^{t} k\left(x\right)$ is incurred[2]. Should a transition from x to $\mathbf{0}$ occur under action $a = 1$ at time t, a discounted maintenance cost $\beta^{t}\left\{C\left(x\right) + k\left(\mathbf{0}\right) + W\right\}$ is incurred, where W is the charge for activity as defined above.

vi'. **Policy**

The goal of optimization is the choice of a policy to minimise the total expected maintenance/operation cost (including activity charges) incurred over an infinite horizon. We again assert the existence of optimal policies for the W^{+}-*problem* which are stationary and whose value functions satisfy the DP optimality equations.

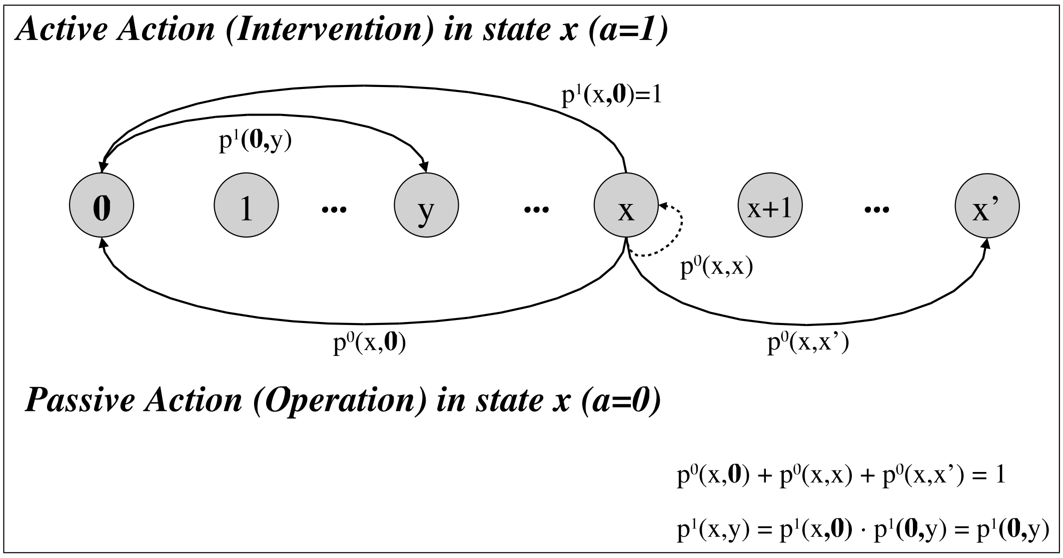
Figure 4.1 depicts the structure of a typical *machine maintenance project* (arm).

[2] Here we are using the telescoped version as defined in §v.

We shall use now $V(x,W)$ for the value function for the W^{+}-*problem* evaluated at $x \in S$. From §i – §iv and §v' above, the optimality equations for one isolated machine may be expressed as:

$$V(x,W) = \min\left\{ W + C(x) + k(\mathbf{0}) + \beta \sum_{x' \in S} P(\mathbf{0}, x') V(x', W); \right.$$
$$\left. k(x) + \beta \sum_{x' \in S} P(x, x') V(x', W) \right\}, \ x \in S \quad (4.7)$$

The first term in $\{;\}$ in the r.h.s. of (4.7) represents the cost of taking the active action ($a = 1$) in state x and the second one corresponds to the passive action ($a = 0$).


Active Action (Intervention) in state x (a=1)
p¹(x,**0**)=1
p¹(**0**,y)
0
1
•••
y
•••
x
x+1
•••
x'
p⁰(x,x)
p⁰(x,**0**)
p⁰(x,x')
Passive Action (Operation) in state x (a=0)
p⁰(x,**0**) + p⁰(x,x) + p⁰(x,x') = 1
p¹(x,y) = p¹(x,**0**) · p¹(**0**,y) = p¹(**0**,y)

In the typical machine maintenance project, if active action (intervention) is taken in state x, the machine is instantaneously returned to pristine state $\mathbf{0}$ and then performs a single transition to some state x'. Under passivity (operation) the machine deteriorates according to certain Markovian rule (including, for some cases, a positive probability of catastrophic breakdown).

Figure 4.1: Representation of a Project of the Machine Maintenance Problem

The solution is based on identifying an optimal Markovian scheduling policy

π prescribing, at each state, to which Q give maintenance and which others $M-Q$ to let operate/deteriorate. As we already know this problem suffers from the curse of dimensionality, which hinders the application of standard dynamic programming techniques and, consequently, we need to look for an alternative heuristic for finding an efficient solution to this problem: an index policy. In particular, we will concentrate in the family of Whittle's index heuristics discussed in Section 2.3. The next step is, hence, to establish the indexability of the machine maintenance problem.

4.1.2 Indexability Analysis

We start by considering a set up in which $X\left(0\right)=\mathbf{0}$ and passive action is taken at $t=0$ with an optimal (cost minimising) policy pursued thereafter. Since we restrict to stationary policies it follows that:

$$\tau^{*}\equiv\min\left\{t:t\geq 1 \text{ and it is optimal to choose active action (intervention)}\right\}$$

is an stationary stopping time.

The evolution of a project in the W^{+}-*problem* can be summarized as follows:

a. Starting from state $\mathbf{0}$ at time $t=0$, we optimally take the passive action for τ periods incurring the expected discounted cost $K\left(\mathbf{0},\tau\right)$, with K given by:
$$K\left(x,\tau\right)=E\left[\sum_{t=0}^{\tau-1}\beta^{t}k\big(X\left(t\right)\big)\left|x\right.\right]$$
where the conditional $|x$ is a notational shorthand for $|X\left(0\right)=x$.

b. At decision epoch τ we take the active option (to give maintenance) with expected discounted cost equal to:

$$E\Big[\beta^{\tau}\big(C\left(X\left(\tau\right)\right)+W\big)\left|x\right.\Big]$$

c. Active action takes the system back to state **0** from which the optimal policy above is repeated indefinitely.

Hence the expected discounted cost, $B\left(\mathbf{0},W\right)$, for the W^{+}*-problem* incurred over the infinite horizon satisfies the equation:

$$W+B\left(\mathbf{0},W\right)=W+\inf_{\tau}\Big\{K\left(\mathbf{0},\tau\right)+E\big[\beta^{\tau}\big(C\left(X\left(\tau\right)\right)+W+B\left(\mathbf{0},W\right)\big)\left|\mathbf{0}\right.\big]\Big\} \tag{4.8a}$$

$$=W+K\left(\mathbf{0},\tau^{*}\right)+E\big[\beta^{\tau^{*}}\big(C\left(X\left(\tau^{*}\right)\right)+W+B\left(\mathbf{0},W\right)\big)\left|\mathbf{0}\right.\big] \tag{4.8b}$$

Note that the terms on the r.h.s. of (4.8) record both the expected costs incurred during the initial period of machine evolution under the passive action up to τ (τ^{*}, respectively) and the costs to go from the intervention at τ (τ^{*}, respectively), onwards. Note also that the *infimum* in (4.8a) is over all stationary positive integer-valued stopping times on the machine state process evolving from **0** under the passive action. Note that stopping time τ is said to be stationary if it is the time of the first entry of the process into some specified subset of S. We can now write

$$\mathcal{B}\left(\mathbf{0},W\right)=W+B\left(\mathbf{0},W\right)$$

and the next result follows straightforwardly:

Theorem 4.1. *The quantity* $\mathcal{B}(\mathbf{0}, W)$ *is given by*

$$\mathcal{B}(\mathbf{0}, W) = \inf_{\tau} \left\{ \frac{W + K(\mathbf{0}, \tau) + E\big[\beta^{\tau} C(X(\tau)) \,|\mathbf{0}\big]}{1 - E[\beta^{\tau}\,|\mathbf{0}]} \right\} \tag{4.9a}$$

$$= \left\{ \frac{W + K(\mathbf{0}, \tau^{*}) + E\big[\beta^{\tau^{*}} C(X(\tau^{*})) \,|\mathbf{0}\big]}{1 - E[\beta^{\tau^{*}}\,|\mathbf{0}]} \right\} \tag{4.9b}$$

where the infimum in (4.9) *is taken over all stationary positive-valued stopping times on the machine process evolving from* $\mathbf{0}$ *under the passive action. Moreover,* $\mathcal{B}(\mathbf{0}, W)$ *is strictly increasing in* W*.*

Proof *Equations* (4.9) *are immediate consequences of the discussion around* (4.8)*. Take now some* $W_1 > W_2$ *and assume* $\tau(W_1)$ *achieves the infimum for* $\mathcal{B}(\mathbf{0}, W_1)$*. Standard index theory guarantees the existence of such infimum. We then have that*

$$\mathcal{B}(\mathbf{0}, W_1) > \frac{W_2 + K(\mathbf{0}, \tau(W_1)) + E\big[\beta^{\tau(W_1)} C\big(X(\tau(W_1))\big) \,|\mathbf{0}\big]}{1 - E[\beta^{\tau(W_1)}\,|\mathbf{0}]} \geq \mathcal{B}(\mathbf{0}, W_2) \tag{4.10}$$

and conclude that $\mathcal{B}(\mathbf{0}, W)$ *is strictly increasing in* W*. Continuity of* $\mathcal{B}(\mathbf{0}, \cdot)$ *is straightforward.* ***q.e.d.*** ■

Before proceeding with the main indexability result of this section, we develop a form of Gittins index appropriate for our analysis. In order to develop $G(x)$, the so-called *Gittins index for passivity* in state $x \in \mathcal{S}$, consider a set up in which we start at some state $X(0) = x$, let the machine to evolve (take passive action) for τ periods with operation cost $K(x, \tau)$, and then intervene and turn back to state $\mathbf{0}$ with cost $E[\beta^{\tau} C(X(\tau)\,|x)] - C(x)$.

Hence, quantity $K(x, \tau) + E[\beta^{\tau} C(X(\tau)\,|x)] - C(x)$ represents the expected cost incurred by taking passive action for τ additional periods when starting from state x; where τ is a stopping time on the machine state process

$\{X(t), t \geq 0\}$ such that $\tau \geq 1$ almost surely.

Definition 4.1. *The Gittins index for passivity $G : S \rightarrow \mathbb{R}$ in state $x \in S$ is given by:*

$$G(x) = \inf_{\tau} \left\{ \frac{K(x,\tau) + E\left[\beta^{\tau} C(X(\tau)\,|x)\right] - C(x)}{1 - E\left[\beta^{\tau}\,|x\right]} \right\} \tag{4.11}$$

where the infimum in (4.11) *is taken over all stationary positive-valued stopping times on the machine state process evolving from x under the passive action.*

Gittins index theory enables us to characterise the set of stationary stopping times achieving the infimum in (4.11). They are developed as follows: fix $W \in \mathbb{R}$ and use $\Gamma(W)$ for the subset of S given by

$$\Gamma(W) = \{y \in S; G(y) > W\}$$

and $\Sigma(W)$ for the subset of S given by

$$\Sigma(W) = \{y \in S; G(y) = W\}.$$

Note that either or both of $\Gamma(W)$ and $\Sigma(W)$ may be empty. Now suppose that $X(0) = x$ and $\Sigma \subset \Sigma(W)$. Use τ^{Σ} for the stationary positive-valued stopping tiem defined on the process $\{X(t), t \geq 0\}$ evolving under the passive action, given by:

$$\tau^{\Sigma} = \min\{t; t > 0 \text{ and } X(t) \in \Gamma(W) \cup \Sigma\}.$$

We write $T(x, W)$ for the collection given by:

$$T(x, W) = \bigcup_{\Sigma \subset \Sigma(W)} \{\tau^{\Sigma}\} \tag{4.12}$$

On the basis of Gittins index theory we can assert that all stopping times in $T(x, G(x))$ achieve the infimum in (4.11). These are the only stationary stopping times which do so.

Let us take one more step ahead in our discussion before establishing indexability of the machine maintenance problem. Consider $X(0) = x \in S$. Passive action is optimal for the W^{+}-*problem* at time $t = 0$ in state x if and only if there exists some stationary positive-valued stopping time τ on the machine state process evolving under the passive action such that any policy which

- chooses *passive* action at times $t = 0, 1, 2, \ldots, \tau - 1$;
- chooses *active* action at time τ; and
- chooses optimally at all times $t \geq \tau + 1$

has total expected costs no greater than the best policy among those which choose action $a = 1$ at time $t = 0$.

In other words, the question faced here is whether to intervene at $t = 0$, go to pristine state $\mathbf{0}$, and from that point on to follow the optimal (cost minimising) policy with minimum achievable cost given by

$$C(x) + W + B(\mathbf{0}, W) = C(x) + \mathcal{B}(\mathbf{0}, W); \tag{4.13}$$

or to let the machine evolve for some extra time τ at which point we intervene and turn back to state $\mathbf{0}$ and then operate following the optimal policy. The expected cost of leaving the system to evolve for additional τ periods is given by

$$K\left(x,\tau\right)+E\left[\beta^{\tau}C\left(X\left(\tau\right)\right)\left|x\right]+E\left[\beta^{\tau}\left|x\right]\mathcal{B}\left(\mathbf{0},W\right).$$

It is the clear that passive action is optimal in x for the W^{+}-*problem* if and only if there exists a stationary positive-valued stopping time τ on the state process evolving from x under the passive action such that

$$K\left(x,\tau\right)+E\left[\beta^{\tau}C\left(x\left(\tau\right)\right)\left|x\right]+E\left[\beta^{\tau}\mathcal{B}\left(\mathbf{0},W\right)\right]\leq C\left(x\right)+\mathcal{B}\left(\mathbf{0},W\right)$$

i.e. $\exists\ \tau>0$ s.t.

$$K\left(x,\tau\right)+E\left[\beta^{\tau}C\left(x\left(\tau\right)\right)\left|x\right]-C\left(x\right)\leq\mathcal{B}\left(\mathbf{0},W\right)-E\left[\beta^{\tau}\mathcal{B}\left(\mathbf{0},W\right)\right]$$

i.e. $\exists\ \tau>0$ s.t.

$$\frac{K\left(x,\tau\right)+E\left[\beta^{\tau}C\left(x\left(\tau\right)\right)\left|x\right]-C\left(x\right)}{1-E\left[\beta^{\tau}\left|x\right]}\leq\mathcal{B}\left(\mathbf{0},W\right). \tag{4.14}$$

Notice that if we take the *infimum* over all positive stopping times τ in (4.14), then the term in the l.h.s. of the expression above turns out to be the Gittins index for state x as introduced in Definition 4.1.

Plainly, from Definition 4.1 and the fact that the infimum in (4.11) is always achieved, the requirement in (4.14) is met precisely when

$$G\left(x\right)\leq\mathcal{B}\left(\mathbf{0},W\right) \tag{4.15}$$

We can then define

$$\Pi\left(W\right)=\left\{x\in S:G\left(x\right)\leq\mathcal{B}\left(\mathbf{0},W\right)\right\} \tag{4.16}$$

as the set of states x in which *passive action* is optimal, in the W^+-*problem*, under activity charge W. Following Whittle's discussion in [116], the restless bandit $(S, \mathcal{A}, P, k, C, \beta)$ (and the corresponding machine maintenance problem) will be indexable if $\Pi(W)$ is increasing with activity charge W, i.e.

$$W_1 > W_2 \Rightarrow \Pi(W_2) \subseteq \Pi(W_1)\,. \tag{4.17}$$

Theorem 4.2. ***Indexability and Indices.***

1. *Bandit (S, P, k, C, β) is indexable.*
2. *The Whittle index for state x is denoted by $W(x)$ and is the unique W-solution to the equation:*

$$G(x) = \mathcal{B}(\mathbf{0}, W)$$

3. *The orderings of members of S determined by the Whittle index and the Gittins index for passivity coincide.*

Proof *By Theorem 4.1, $\mathcal{B}(\mathbf{0}, W)$ is strictly increasing in W. From* (4.16) *it then follows that $\Pi(W)$ is increasing and indexability follows from Definition 2.1. It also holds from the continuity of $\mathcal{B}(\mathbf{0}, :)$ that the Whittle index for state x, namely*

$$W(x) = \inf\{W : x \in \Pi(W)\}$$

satisfies the equation

$$\mathcal{B}(\mathbf{0}, W(x)) = G(x) \tag{4.18}$$

By the strictly increasing nature of $\mathcal{B}(\mathbf{0}, W)$, the equation (4.18) *specifies $W(x)$*

uniquely. This establish parts (a) and (b) of the Theorem. Part (c) follows simply from the resultant fact that $W(x)$ is strictly increasing in $G(x)$. ***q.e.d.*** ■

Comment

Please note that it is an immediate consequence of Theorems 4.1 and 4.2 and of (4.11) that $W(\mathbf{0}) = -D(\mathbf{0})$. Hence the pristine state has a negative Whittle index. Subsequent analysis will focus on develping indices for non-pristine states.

We now recall the notation and ideas established around (4.12).

Lemma 4.1.

(a) Any stopping time in $T(x, G(x))$ achieves the infimum in equation (4.11).

(b) Any stopping time in $T(\mathbf{0}, \mathcal{B}(\mathbf{0}, W))$ achieves the infimum in (4.9).

In both cases, these are the only stopping times which achieve the infima concerned.

Proof *The reasoning for part (a) is summarised in the comments around (4.12) and is a feature of the Gittins index structure.*

For part (b), consider quantity $\mathcal{B}(\mathbf{0}, W)$ and proceed the same way as we did with the Gittins index. We have already proved that

$$\mathcal{B}(\mathbf{0}, W) = \inf\left\{\frac{W + K(\mathbf{0}, \tau) + E\left[\beta^{\tau} C(x(\tau))\,|\mathbf{0}\right]}{1 - E\left[\beta^{\tau}\,|\mathbf{0}\right]}\right\} \tag{4.19}$$

Assume that $X(0) = \mathbf{0}$ and that the machine state evolves under the passive action. Extend the bandit's space to $S \cup \mathbf{0}^*$ where $\mathbf{0}^*$ is used specifically to designate state $\mathbf{0}$ at time 0, with $\mathbf{0}$ reserved for the pristine state at other *epochs. Further, we impose the following costs: a transition from* $\mathbf{0}^*$ *to state* x *incurs a cost of* $W + K(\mathbf{0}, x) + \beta C(x)$. *A transition from state* $y \neq \mathbf{0}^*$ *to state* z *incurs a cost of* $K(y, z) - C(y) + \beta C(z)$. *With these choices, the expected cost incurred by the machine during* $[0, \tau)$ *is given by*

$$W + C(\mathbf{0}, \tau) + E\left[\beta^{\tau} C(x(\tau)) \,|\mathbf{0}\right]$$

where τ *is a positive-valued stopping time. Further, the expected cost incurred by the machine during* $[0, \tau)$ *from some initial state* $x \neq \mathbf{0}^*$ *is given by*

$$K(x, \tau) - C(x) + E\left[\beta^{\tau} C(x(\tau)) \,|x\right].$$

Regarding this constructed object as a Gittins-type bandit, $\mathcal{B}(\mathbf{0}, W)$ *is by definition the Gittins index for initial state* $\mathbf{0}^*$. *See* (4.19). *Further,* $G(x)$ *in* (4.11) *is the Gittins index for any state* $x \neq \mathbf{0}$.

Part (b) is now seen to be an application of the comment around (4.12) *to this bandit*[3].

[3] *If we apply the comments around* (4.12) *to this bandit, we can characterise the set of stationary stopping times achieving the infimum in* (4.19). *Again, fix* $W \in \mathbb{R}$ *and write*

$$\Gamma(W) = \{y \in S; \mathcal{B}(\mathbf{0}, W) > W\}$$

and

$$\Sigma(W) = \{y \in S; \mathcal{B}(\mathbf{0}, W) = W\}.$$

Note that either or both of $\Gamma(W)$ *and* $\Sigma(W)$ *may be empty. Now suppose that* $X(0) = \mathbf{0}$ *and* $\Sigma \subset \Sigma(W)$. *Use* τ^{Σ} *for the stationary positive-valued stopping time defined on the process* $\{X(t), t \geq 0\}$ *evolving under the passive action, given by:*

$$\tau^{\Sigma} = min\{t; t > 0 \text{ and } X(t) \in \Gamma(W) \cup \Sigma\}.$$

We write $T(W)$ *for the collection given by:*

$$T(\mathbf{0}, \mathcal{B}(\mathbf{0}, W)) = \bigcup_{\Sigma \subset \Sigma(W)} \{\tau^{\Sigma}\} \tag{4.20}$$

q.e.d. ■

4.2 Two Cases of the Machine Maintenance Problem

In this section we explore index structure in the context of two model types, both of which rest on assumptions which are plausible in practice. Them both preserve indexability and the essential index structure.

4.2.1 Monotone Models

We introduce the following assumption to our basic model:

Assumption 4.1. *The state space S is the natural numbers $\mathbb{N}$ with 0 the designated pristine state.*

Assumption 4.2. *Evolution under the passive action is right-skip free, i.e.*

$$P(x, x') = 0, \quad x' > x + 1, \ \textit{for all} \ x, x' \in \mathbb{N}$$

Assumption 4.3. *The Gittins index for passivity $G : \mathbb{N} \to \mathbb{R}$ is (strictly) increasing.*

Hence under such models an increase in state corresponds to deterioration of machine, resulting in higher cost rates (as measured by Gittins index).

Now suppose that $X(0) = x$ and that the machine state evolves under the passive action. We use $\tau(x, x')$ for the time of the first entry after 0 into

. *On the basis of Gittins index theory we can assert that all stopping times in $T(x, \mathcal{B}(\mathbf{0}, W))$ achieve the infimum in* (4.19). *These are the only stationary stopping times which do so.*

state x'. Note that the *right-skip free* assumption implies that $\tau(x, x') < \tau(x, x'+1)$ almost surely for all $x < x'$.

We start our analysis by writing the stopping time achieving the infimum in (4.11) as

$$\tau^* = \min\left[t; t \geq 1, G(x(t)) \geq G(x)\right]$$

by assumption of our model we have $G(x+1) > G(x)$, which implies that we are to take the passive action up to the moment when we arrive to either x or $x+1$. Hence,

$$\tau^* = \min\left\{\tau(x, x), \tau(x, x+1)\right\} = \tau(x; x, x+1)$$

represents the required time until the first visit to x or $x+1$ when starting at x.

By using the results and the discussion around Lemma a it is easy to verify that $\tau^* \in T(x, G(x))$ and $x, x+1 \in \Gamma(G(x))$, i.e. $G(x)$ is achieved by τ^*. With this elements we can rewrite (4.11) as:

$$G(x) = \frac{K(x, \tau^*) + E\left[\beta^{\tau^*} C(x(\tau^*)) \left|x\right.\right] - C(x)}{1 - E\left[\beta^{\tau^*} \left|x\right.\right]}. \tag{4.21}$$

Turn now to the $\mathcal{B}(0, W)$ quantity. The stopping time achieving the *infimum* in (4.19) can be written as

$$\tilde{\tau} = \min\left[t; t \geq 1, G(x(t)) \geq \mathcal{B}(0, W)\right]. \tag{4.22}$$

If W satisfies $\mathcal{B}(0, W) = G(x)$ then, given the monotonicity of $G(x)$, $X(t) = x$ satisfies (4.22), which means that the passive action is to be taken up to the moment when we arrive to x; i.e.

$$\tilde{\tau} = \tau\left(0, x\right)$$

represents the time required for the first visit to x when starting at state 0. Moreover, the *right-skip free* nature of passive evolution means that $\tau\left(0, x\right) \in T\left(0, G\left(x\right)\right)$.

Then, we can write (4.9) as:

$$\mathcal{B}\left(0, W\right) = \frac{W + K\left(0, \tilde{\tau}\right) + E\left[\beta^{\tilde{\tau}}\right] C\left(x\right)}{1 - E\left[\beta^{\tilde{\tau}}\right]} \tag{4.23}$$

because $\tilde{\tau} = \tau\left(0, x\right) \Rightarrow x\left(\tilde{\tau}\right) = x$.

Theorem 4.3. ***Whittle indices for monotone models.***
For monotone models, the Whittle index is given by

$$W\left(x\right) = G\left(x\right)\left\{1 - E\left[\beta^{\tilde{\tau}}\right]\right\} - K\left(0, \tilde{\tau}\right) - E\left[\beta^{\tilde{\tau}}\right] C\left(x\right), \quad x \in \mathbb{N}^{+} \tag{4.24}$$

with $G\left(x\right)$ given by (4.21)*, and is increasing in x.*

Proof *By Theorem 4.2, $W\left(x\right)$ is the unique W-solution to $G\left(x\right) = \mathcal{B}\left(0, W\right)$. Solving*

$$\mathcal{B}\left(0, W\right) = \frac{W + K\left(0, \tilde{\tau}\right) + \beta^{\tilde{\tau}} C\left(x\right)}{1 - E\left[\beta^{\tilde{\tau}}\left|x\right.\right]} = G\left(x\right) \tag{4.25}$$

for W we get (4.24)*.*

That $W\left(x\right)$ is increasing in x follows from the facts that $W\left(x\right)$ is strictly increasing in $G\left(x\right)$, and that $G\left(x\right)$ is increasing in x. This concludes the proof. ***q.e.d.*** ■

4.2.2 Breakdown/Deterioration Models

We introduce the following additional assumptions to our basic model:

Assumption 4.4. *The state space S is the natural numbers $\mathbb{N}$ with 0 the designated pristine state.*

Assumption 4.5. *Evolution under the passive action is such that:*

$$P(x,0) + P(x,x) + P(x,x+1) = 1, \quad x \in \mathbb{N}$$

Hence, under the passive action, a machine currently in state x may either remain there (with probability $P(x,x)$), have a catastrophic breakdown followed by immediate maintenance/replacement (with probability $P(x,0)$) or deteriorate by a single unit (with probability $P(x,x+1)$). It will simplify the discussion if we further suppose that $P(x,0)+P(x,x+1)$ is strictly positive for all $x \in \mathbb{N}$. For pristine state 0 we use (0) for the probability of a catastrophic breakdown to distinguish it from $P(0,0)$, the probability of a non-departure from the pristine state. We suppose that $P(0) + P(0,0) + P(0,1) = 1$.

Consider the bandit evolving from state x at time 0 under the passive action. We start our analysis by writing the positive-valued stopping time achieving the infimum in (4.11) as:

$$\tau^* = \min\{t;\ t \geq 1, G(x(t)) \geq G(x)\} \tag{4.26}$$

In the current framework, the *right-skip free* rule of Assumption 4.2 has been relaxed for allowing the probability of a catastrophic breakdown; this implies that we can define $\underline{x} \leq x$ and $\overline{x} > x$ as follows:

$$\underline{x} = \min\{x'; G(x') \geq G(x)\}$$

and

$$\overline{x} = \min\{x'; x' \geq x+1,\ G(x') \geq G(x)\}$$

where we take $\overline{x} = \infty$ if $G(x') < G(x)$, $x' \geq x+1$. Note that for monotone models we have $\underline{x} = x$ and $\overline{x} = x+1$. Consider that bandit evolving from state x at time 0 under the passive action. We now introduce the positive valued stopping time

$$\tau^* = \min\{\tau(x, \underline{x}), \tau(x, \overline{x})\} \tag{4.27}$$

As before, we let $x^* = x(\tau^*)$ and τ^* represents the required time until the first visit to x^* when starting at x. By using the results and the discussion around Lemma a it is easy to see that $\tau^* \in T(x, G(x))$ and, consequently, achieves the *infimum* at (4.11).

Turn now to the quantity $\mathcal{B}(\mathbf{0}, W)$. The stopping time achieving the infimum at (4.19) can be written as

$$\tilde{\tau} = \min\{t;\ t \geq 1, G(x(t)) \geq \mathcal{B}(\mathbf{0}, W)\} \tag{4.28}$$

in particular, after recalling that $W(x)$ is the unique solution to $G(x) = \mathcal{B}(\mathbf{0}, W)$, it can be seen that if $\underline{x} = \min\{x'; G(x') \geq G(x)\}$ then $\tau(0, \underline{x}) \in T(0, \mathcal{B}(\mathbf{0}, W))$ and so achieves the infimum in (4.19) and we can let $\tilde{\tau} = \tau(\mathbf{0}, \underline{x})$ to represent the time required until the first visit to x when initial state is 0. So we can rewrite equations (4.11) and (4.19) as:

$$G(x) = \frac{K(x, \tau^*) + E\left[\beta^{\tau^*} C(x^*)\right] - C(x)}{1 - E\left[\beta^{\tau^*}\right]} \tag{4.29}$$

and

$$\mathcal{B}(\mathbf{0}, W) = \frac{W + K(\mathbf{0}, \tilde{\tau}) + E\left[\beta^{\tilde{\tau}} C(x)\right]}{1 - E\left[\beta^{\tilde{\tau}}\right]} \tag{4.30}$$

Theorem 4.4. ***Whittle indices for breakdown/deterioration models.*** *For breakdown/deterioration models, the Whittle index is given by*

$$W(x) = G(x)\left\{1 - E\left[\beta^{\tilde{\tau}}\right]\right\} - K(\mathbf{0}, \tilde{\tau}) - E\left[\beta^{\tilde{\tau}} C(x)\right], \quad x \in \mathbb{N}^{+}$$

and, after substitution,

$$\begin{aligned} W(x) = {} & \Big(K(x, \tau^{*}) + E\big[\beta^{\tau^{*}} C(x^{*})\big] - C(x)\Big) \frac{1 - E\left[\beta^{\tilde{\tau}}\right]}{1 - E\left[\beta^{\tau^{*}}\right]} \\ & - K(\mathbf{0}, \tilde{\tau}) - E\left[\beta^{\tilde{\tau}}\right] C(x) \end{aligned} \tag{4.31}$$

for all $x \in \mathbb{N}^{+}$.

For the special case $C(x) = C,\ x \in \mathbb{N}$,

$$W(x) = K(x, \tau^{*}) \frac{1 - E\left[\beta^{\tilde{\tau}}\right]}{1 - E\left[\beta^{\tau^{*}}\right]} - K(\mathbf{0}, \tilde{\tau}) - C,\ x \in \mathbb{N}^{+}. \tag{4.32}$$

Proof *It follows from Theorem 4.2 that $W(x)$ is the W-solution of*

$$\mathcal{B}(\mathbf{0}, W) = G(x),$$

solving this expression for W by using (4.29) *and* (4.30) *yields* (4.31)*.* ***q.e.d.*** ■

4.3 Examples

In this section we further develop the findings of Section 4.1 by offering some families of examples for which explicit formulae for the Whittle index can be derived. We later offer some numerical results regarding examples in this section.

We consider examples of the breakdown/deterioration model for which transitions to 0 under the passive action correspond to catastrophic unexpected breakdowns of the machine (followed by its immediate replacement/renewal). Such transitions may incur great costs. Hence costs incurred by the bandit undergoing transitions from x to 0, x, and $x+1$ under the passive action will be taken to be of the form $B + k\left(x\right)$, $k\left(x\right)$ and $k\left(x\right)$ respectively, where B is the cost of a catastrophic breakdown. It will simplify matters if we suppose that $P\left(0\right) = 0$, namely that there are no catastrophic breakdowns in the pristine state. We shall explore instances of this model which are also monotone -i.e., which satisfy Assumption 4.3 above.

Family I. Here we have no catastrophic breakdowns, and between interventions the machine is subject only to gradual deterioration and (typically) increasing maintenance costs.

We strengthen Assumption 4.2 in the monotone model formulation, by assuming that:

$$P\left(x,x\right) + P\left(x,x+1\right) = 1 \tag{4.33}$$

Before further proceeding with our analysis, it is convenient to introduce the following expressions (please see Apendix A.1 for details).

$$E\left[\beta^{\tau^*}\left|x\right.\right] = \beta, \quad x \in \mathbb{N}. \tag{4.34a}$$

$$E\left[\beta^{\tilde{\tau}}\left|x\right.\right] = \prod_{y=0}^{x-1} \delta\left(y\right), \quad x \in \mathbb{N}. \tag{4.34b}$$

$$K\left(x,\tau^*\right) = k\left(x\right), \quad x \in \mathbb{N}. \tag{4.34c}$$

$$K\left(0,\tilde{\tau}\right) = \sum_{y=0}^{x-1} \kappa\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right) \tag{4.34d}$$

with

$$\delta\left(y\right)=\frac{\beta P\left(y,y+1\right)}{1-\beta P\left(y,y\right)},\quad y\in\mathbb{N} \tag{4.35}$$

$$\kappa\left(x\right)=\frac{k\left(x\right)}{1-\beta P\left(x,x\right)},\quad x\in\mathbb{N} \tag{4.36}$$

$$\epsilon\left(x\right)=\left\{1-\beta P\left(x,x\right)\right\}^{-1},\quad x\in\mathbb{N}. \tag{4.37}$$

We now define expression:

$$H\left(x\right)\equiv\frac{K\left(x,\tau^{*}\right)+E\left[\beta^{\tau^{*}}C\left(x\left(\tau^{*}\right)\right)|x\right]-C\left(x\right)}{1-E\left[\beta^{\tau^{*}}|x\right]}. \tag{4.38}$$

Upon substitution of expressions in (4.34) into (4.38), and after some algebraic manipulations we get

$$H\left(x\right)=\frac{k\left(x\right)+\beta\left(P\left(x,x\right)C\left(x\right)+P\left(x,x+1\right)C\left(x+1\right)\right)-C\left(x\right)}{1-\beta},\ x\in\mathbb{N}. \tag{4.39}$$

The following result utilises a self consistency result for Gittins indices due to Nash [70].

Lemma 4.2. *If $H:\mathbb{N}\to\mathbb{R}$ is increasing, then $H\left(x\right)=G\left(x\right)$, $x\in\mathbb{N}$.*

Corollary 4.1. *For the monotone case with no breakdowns, if $H\left(x\right)$ in* (4.39) *is increasing, then $H\left(x\right)=G\left(x\right),\ x\in\mathbb{N}$, and the Whittle index is given by*

$$\begin{aligned}W\left(x\right)=\sum_{y=0}^{x-1}\left(\beta P\left(x,x+1\right)\left(C\left(x+1\right)-C\left(x\right)\right)+k\left(x\right)-k\left(y\right)\right)\epsilon\cdots\\ \times\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)-C\left(x\right)\end{aligned} \tag{4.40}$$

Proof *The first statement in the corollary is an immediate consequence of Lemma 4.2. For the index* $W(x)$*, by using expressions in* (4.34) *we have that in the monotone case:*

$$W(x) = G(x)\left\{1 - \prod_{y=0}^{x-1} \delta(y)\right\} - \sum_{y=0}^{x-1} k(y)\,\epsilon(y) \prod_{z=0}^{y-1} \delta(z) - C(x) \prod_{y=0}^{x-1} \delta(x)$$

After isolating $C(x)$ *in* $H(x)$ *and after further manipulations, we get:*

$$\begin{aligned} W(x) = {} & \frac{k(x) + \beta P(x, x+1)(C(x+1) - C(x))}{1-\beta} \cdots \\ & \times \left\{1 - \prod_{y=0}^{x-1} \delta(y)\right\} - \sum_{y=0}^{x-1} k(y)\,\epsilon(y) \prod_{z=0}^{y-1} \delta(z) - C(x) \end{aligned}$$

which together with (A.4) *gives* (4.40). ***q.e.d.*** ■

Comment

Note that if $C(x) = C$, $x \in \mathbb{N}$, then $H(x)$ above will be increasing if maintenance costs $k(x)$ are increasing in state x. It will then follow from Theorem 1(b) that $W(x)$ is also increasing in x. In particular (4.40) reduces to

$$W(x) = \sum_{y=0}^{x-1} (k(x) - k(y))\,\epsilon(y) \prod_{z=0}^{y-1} \delta(z) - C,\ x \in \mathbb{N}.$$

Equivalently, if $C(x) = a + bx$, then $H(x)$ will be increasing if for $\phi(x) = k(x) + b\beta P(x, x+1)$ it holds that $\phi(x+1) - \phi(x) \geq (1-\beta)\,b$. In this case, $W(x)$ will be given by:

$$W(x) = \sum_{y=0}^{x-1} \Big[\beta P(x, x+1)\, b + k(x) - k(y)\Big]\epsilon(y) \prod_{z=0}^{y-1} \delta(z) - (a + bx)\,,\ x \in \mathbb{N}.$$

Family II. Here we fix $C(x) = C$ and $k(x) = 0$ for $x \in \mathbb{N}$. In contrast to Family I, the focus here is predominantly on the minimisation of costs due to catastrophic breakdowns.

We start by recalling the Gittins index in (4.29)

$$G(x) = \frac{K(x, \tau^*) + E\left[\beta^{\tau^*} C(x^*) \,|x\right] - C(x)}{1 - E\left[\beta^{\tau^*} \,|x\right]}.$$

We have already established that $\tau^* = \min\{\tau(x, \underline{x}), \tau(x, \overline{x})\}$ and, given that we are working under Assumption 4.3, it is also true that $\underline{x} = x$ and $\overline{x} = x+1$, consequently $\tau^* = \min\{\tau(x, x), \tau(x, x+1)\} = \tau(x; x, x+1)$. Substituting this last into the expression for the Gittins index above we can define:

$$H(x) \equiv \frac{K(x, \tau^*) + E\left[\beta^{\tau^*} C(x(\tau^*))\right] - C(x)}{1 - E\left[\beta^{\tau^*}\right]} \tag{4.41}$$

Which, if increasing in x, turns out to be the Gittins index as established by Lemma 4.2. Calculations can be presented more economically in notational terms if, additionally to the expressions in (4.34), we write:

$$\gamma(x) = \frac{\beta P(x, 0)}{1 - \beta P(x, x)}, \quad x \in \mathbb{N}^+. \tag{4.42}$$

All upcoming formulae are for $x \in \mathbb{N}^+$. The details are presented in Appendix A.2 to this chapter.

$$E\left[\beta^{\tau(0,x)}\right] = \frac{\prod_{y=0}^{x-1} \delta(y)}{1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z)} \tag{4.43a}$$

$$\begin{aligned} E\left[\beta^{\tau^*} C(x^*)\right] &= \beta P(x, x) C(x) + \beta P(x, x+1) C(x+1) \cdots \\ &+ \beta P(x, 0) C(x) \prod_{y=0}^{x-1} \delta(y) \left[1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z)\right]^{-1} \end{aligned} \tag{4.43b}$$

Simple conditioning arguments yield the conclusion that the expected cost $K(0, \tilde{\tau}_x)$ satisfies the equation:

$$K(0, \tilde{\tau}_x) = \frac{\sum_{y=0}^{x-1} \left(\gamma(y) + \frac{B}{\beta}\gamma(y) I\{y \neq 0\} \right) \prod_{z=0}^{y-1} \delta(z)}{1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z)}. \tag{4.44}$$

also after some straightforward algebraic manipulations we get

$$1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z) = (1-\beta) \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(z) + \prod_{y=0}^{x-1} \delta(y). \tag{4.45}$$

Now upon substitution of expressions (4.43) to (4.45) into (4.41) we conclude that

$$H(x) = \frac{BP(x,0)}{(1-\beta)\left[(1-\beta+\beta P(x,0)) \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(z) + \prod_{y=0}^{x-1} \delta(y) \right]} - C, \tag{4.46}$$

for $x \in \mathbb{N}^+$, clearly $H(0) = -C$.

Corollary 4.2. *If $k(x) = 0,\ C(x) = C,\ x \in \mathbb{N}$ and $P(x,0)$ is increasing in x, then $H : \mathbb{N} \to \mathbb{R}$ given in (4.46) is increasing and $H(x) = G(x),\ x \in \mathbb{N}$. The Whittle index is then given by*

$$\begin{aligned} W(x) = B &\left[P(x,0)\epsilon(0) + \sum_{y=1}^{x-1} \epsilon(y) \left(P(x,0) - P(y,0) \right) \prod_{z=0}^{y-1} \delta(z) \right] \cdots \\ &\times \left[(1-\beta+\beta P(x,0)) \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(z) + \prod_{y=0}^{x-1} \delta(y) \right]^{-1} - C,\ x \in \mathbb{N}^+ \end{aligned} \tag{4.47}$$

Proof *We first note from (4.46) that $H(0) \leq H(x), x \in \mathbb{N}^+$. Further,* *from* (4.46) *we have that if* $k(x) = 0,\ C(x) = C,\ x \in \mathbb{N}^+$, *then*

$$\begin{aligned}
\frac{B}{1-\beta}\left[H(x+1)+D\right]^{-1} =& \frac{1-\beta+\beta P(x+1,0)}{P(x+1,0)} \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(y) \\
&+ \frac{(1-\beta)\left[1-\beta P(x,x)-\beta P(x,x+1)\right]}{P(x+1,0)} \epsilon(x) \prod_{y=0}^{x-1} \delta(y) \\
&+ \frac{(1-\beta)}{P(x+1,0)} \prod_{y=0}^{x-1} \delta(y) \\
=& \frac{1-\beta+\beta P(x+1,0)}{P(x+1,0)} \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(y) \\
&+ \frac{(1-\beta)}{P(x+1,0)} \prod_{y=0}^{x-1} \delta(y) \\
\leq& \frac{1-\beta+\beta P(x,0)}{P(x,0)} \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(y) \cdots \\
&\quad + \frac{(1-\beta)}{P(x,0)} \prod_{y=0}^{x-1} \delta(y) \\
=& \frac{B}{1-\beta}\left[H(x)+D\right]^{-1}
\end{aligned}$$

whenever $P(x,0) \leq P(x+1,0)$. *It follows from the expression above that if* $P(x,0)$ *is increasing then so is* $H(x)$. *We then have that substitution from expressions* (4.46), (4.44) *and* (4.43) *into* (4.31) *-together with straightforward algebra- yields the expression for the Whittle index given in the statement of the result.*

The inference that $H(x) = G(x)$ *uses Lemma 4.2.*

The algebraic details of this derivation are given in the appendix to this paper. ***q.e.d.*** ■

4.4 Numerical Study

We here describe some numerical results which asses the quality of performance of an index policy based on the set up considered in Family I (see Corollary 4.1 and the Comment below.). To be precise, we explore a scenario on which a single repairman is maintaining four machines. This is modelled as a restless bandit problem with $Q = 1$, $M = 5$ -i.e. a single server choosing among five bandits. Four of the five bandits model machines state evolution, as described in Section 4.3. The fifth is an idling option modelled as a single state bandit with no costs incurred under either action (active or passive). The Whittle index for the idle option is trivially zero. Hence, the Whittle index heuristic chooses idleness when the four bandits modelling system evolution all have negative index values. Otherwise, the repairman works on whichever machine has the largest (positive) index.

Case I in Table 4.1 presents results summarising the performance of the Whittle index heuristic for two cases: in the upper part a fixed intervention cost C increases from 50 to 200; in the lower part, the constant term of a linear intervention cost with slope $25x$ (where x is the state of the machine) increases from 25 to 200. Each row of the table summarises the results of 200 problems studied for the corresponding C. These 200 problems are generated at random as follows: each of the four machines has ten states labelled $0, 1, \ldots, 9$. The operation cost rate for machine i in state x takes the form $K_i(x) = A_i + B_i x$, $1 \leq i \leq 4$, $0 \leq x \leq 9$, where the constants A_i and B_i are obtained by sampling independently from a $U \sim (25, 50)$ distribution. All of the probabilities $P_i(x, x)$, $1 \leq i \leq 4$, $0 \leq x \leq 9$, are obtained by sampling independently from a $U \sim (0.1, 0.8)$ distribution. We then set $P_i(x, x+1) = 1 - P_i(x, x)$ and $1 \leq i \leq 4$, $0 \leq x \leq 9$, and $P_i(9, 9) = 1$, $1 \leq i \leq 4$. Discount rate β is set to be 0.95 throughout.

For each problem generated the optimal expected cost V^{OPT} was computed for the initial state in which al machines were assumed to be in pristine state 0 along with the corresponding expected cost V^{IND} for the Whittle index heuristic. All computations of expected cost were performed using DP value iteration (see Algorithm in figure 2.1). The percentage cost suboptimality $\left(\frac{V^{\mathrm{IND}}}{V^{\mathrm{OPT}}} - 1\right) \times 100$ was then calculated. For each D-value, the 200 percentage suboptimalities were then summarised by the order statistics: minimum (Min), lower quartile (LQ), median, upper quartile (UQ), and maximum (Max). These statistics appear in Table 4.1. Note the excellent level of performance of the index heuristic. In none of the $2,800$ problems studied was the index policy more than 5 percent suboptimal. Actually, there were just two cases for which the upper quartile suboptimality was larger than 1%

Case II in Table 4.1 presents results summarising the performance of the Whittle index heuristic for two cases: in the upper part a fixed intervention cost C increases from 50 to 200; in the lower part, the constant term of a linear intervention cost with slope $25x$ (where x is the state of the machine) increases from 25 to 200. Like in the previous case, each row of the table summarises the results of 200 problems studied for the corresponding values of C. These 200 problems are generated at random as follows. The operation cost rate for machine i in state x takes the form $K_i\,(x) = A_i + B_i x + D_i x^2$, $1 \leq i \leq 4$, $0 \leq x \leq 9$, where the constants A_i and B_i are obtained by sampling independently from a $U \sim (25, 50)$ distribution and D_i is extracted from a $U \sim (4, 6)$ distribution. All of the probabilities were generated as in the previous case. Discount rate β is set to be 0.95 throughout.

As well as in the previous case, the index heuristic performs very well. In none of the $2,800$ problems studied was the index policy more than 5 percent suboptimal and the upper quartile was lager than 1% just for two cases.

Table 4.1: Machine Maintenance Problem. Performance of the Index Policy in the Monotone Case, Linear Operation Cost.

Case I, $K = A + Bx$					
C	Min	LQ	Median	UQ	Max
50	0.0000	0.0033	0.0334	0.1269	0.5293
75	0.0000	0.0012	0.0067	0.0442	0.3565
100	0.0000	0.0009	0.0048	0.0445	0.3713
125	0.0000	0.0007	0.0093	0.0636	3.5544
150	0.0000	0.0010	0.0170	0.1125	3.9744
175	0.0000	0.0019	0.0230	0.8501	4.9521
200	0.0000	0.0026	0.2577	1.1635	4.2520
$50 + 25x$	0.0000	0.0027	0.0371	0.1077	0.4271
$75 + 25x$	0.0000	0.0009	0.0138	0.0631	0.3704
$100 + 25x$	0.0000	0.0011	0.0209	0.0629	2.9586
$125 + 25x$	0.0000	0.0019	0.0119	0.0581	2.0661
$150 + 25x$	0.0000	0.0015	0.0123	0.0751	3.3735
$175 + 25x$	0.0000	0.0070	0.0483	0.7482	2.5984
$200 + 25x$	0.0000	0.0040	0.5707	1.2066	2.8804

Case II, $K = A + Bx + Cx^2$					
C	Min	LQ	Median	UQ	Max
50	0.0000	0.0012	0.0274	0.0917	0.4100
75	0.0000	0.0011	0.0185	0.0786	0.3859
100	0.0000	0.0004	0.0042	0.0336	4.3945
125	0.0000	0.0016	0.0064	0.0398	3.4470
150	0.0000	0.0005	0.0081	0.1012	4.4431
175	0.0000	0.0035	0.0513	0.8849	2.9042
200	0.0000	0.0061	0.4821	1.3297	3.0743
$50 + 25x$	0.0000	0.0008	0.0189	0.0727	0.5020
$75 + 25x$	0.0000	0.0004	0.0079	0.0519	0.3951
$100 + 25x$	0.0000	0.0007	0.0072	0.0532	3.3025
$125 + 25x$	0.0000	0.0006	0.0094	0.0415	1.9974
$150 + 25x$	0.0000	0.0003	0.0076	0.0816	2.6146
$175 + 25x$	0.0000	0.0023	0.0226	0.9932	4.8762
$200 + 25x$	0.0000	0.0059	0.1816	1.0303	2.1583

4.5 Conclusions

In this chapter we established the indexability of a class of restless bandits designed to model machine maintenance problems in which maintenance interventions have to be scheduled to mitigate escalating costs as machines deteriorate with use (operation).

Two particular cases analysed in Section 4.2, both of which rest on assumptions which are plausible in practice. In the first one, a machine deteriorates smoothly with use, with the consequent decrease in profitability and (corrective) maintenance (intervention) –which can be perfect or imperfect– takes it back to an earlier (maybe pristine) state of wear. In the second family of problems, we extended the previous model by including the chance of a catastrophic breakdown (followed by immediate replacement of the machine) with a considerably high cost. Explicit formulae for the Whittle indices were provided in Section 4.3 for two particular examples.

The machine maintenance problem, as formulated, stands as an important family of stochastic scheduling problems for which indexability has been established in particular formulations, but now, under our approach, indexability is guaranteed in general.

Moreover, identification of the Whittle indices concerned was followed in Section 4.4 by a numerical investigation which demonstrates the very strong performance of Whittle's index heuristic bor both, the monotone and the breakdown problems.

There is no need to emphasizing the importance of the machine maintenance formulation and its applicability to a wide range of real life problems.

Within the most important extensions there is the case of imperfect machine maintenance, were the effectiveness of the intervention is dependent on the current state of the machine. This implies just a minor modification of the main framework presented in this Chapter (were intervention was assumed to imply an immediate transition to the pristine state) and its indexability is (intuitively) almost guaranteed.

Another interesting extension is the one in which there exist more than one (independent) deteriorating processes competing in every particular machine. This can be understood as the wear of different parts or components of the machine. It is clear that the failure of any of them can imply the breakdown of the system and, consequently, the intervention (maintenance) regime will be affected. To the intuition of the author this problem can be modelled as a restless bandit problem and conditions for it's indexability can be established.

These two particular extensions to the machine maintenance problem are object of further research for the author.

Chapter 5

Indexability Analysis of Bi-directional Restless Bandits[1]

Introduction

In entertainment shows of a certain vintage, a popular act featured a performer keeping a large number of plates spinning on the top of flexible poles. The audience would express dismay when one of the plates started to wobble badly, prompting urgent attention to prevent it from falling from its stick. The performer's problem (of keeping plates spinning happily) is a vivid metaphor of that facing a manager responsible for a collection of reward generating assets, each of whose performance (reward) is enhanced in time by an active intervention (investment), but which otherwise tends to deteriorate. The crucial issue arises as to how such interventions should be organised to maximise the overall reward yield from an entire asset portfolio. In Section 5.1, a family of problems related with the performer's/manager's problem (Family I) is formulated

[1] The main results in this Chapter have been published in Glazebrook, K., Kirkbride, C., and Ruiz-Hernández, D. Spinning plates and squad systems: policies for bi-directional restless bandits. *Advances in Applied Probability 38-1*, 95-115 (2006).

as a Markov Decision Problem (MDP) with the average reward criterion[2]. In honour of the frivolous application cited above, we call this the *Spinning Plates Problem*.

In contrast to the above there exist situations in which a manager has a large number of reward generating assets at his disposal, a fixed number of which need to be deployed/exploited at all times. Deployment of an asset activated its reward stream, but erodes over time its performance (reward). Resting (not deploying) an asset allows it to recover. The key issue here concerns how assets should be deployed to maximise the rewards earned from them over time. In Section 5.2 a family of problems related with the manager's problem (Family II) is formulated as an MDP with the average reward criterion. In honour of the similar problem faced by coaches in professional sports, we call this problem the *Squad System*. To our knowledge, the spinning plates problem, as formulated in Section 5.1 is new and there is no previous literature. Whittle [116] gave a brief discussion of a particular case of the squad system which had a linear structure for both rewards and stochastic dynamics. He called this the *Ehrenfest Project*. Niño-Mora [74] discussed a discounted reward version of the squad system using polyhedral methods, but was not able to deploy this analysis to give an account of the system with the average reward criterion.

The MDP's concerned are formulated and presented at the beginning of Sections 5.1 and 5.2 and fall within the class of so-called *restless bandit problems* introduced by Whittle [116]. These form a class of decision processes which generalise the *multi-armed bandits* of Gittins[3] ([35],[36]) by allowing passive evolution. Whittle [116] proposed a class of *index heuristics* which extend the Gittins index policies. However, Whittle's proposed indices may not exist (the

[2] For a discussion on optimality criteria for Markov decision problems please see Section 2.1.3 and, in general, Section 2.1 for a discussion on Markov decision processes

[3] See a detailed discussion on Gittins approach in Section 2.2.

issue of *indexability*) and the resulting policy will not in general be optimal, even for indexable problems. Essentials of Whittle's approach are sketched in Section 2.3.

In Sections 5.1 and 5.2 we give simple and direct accounts of the index structure of, respectively, the *spinning plates* problem and the *squad system.* In both cases we give simply stated conditions that guarantee the model's indexability. Further, algorithms are given which yield the indices. *Strict indexability* means that not only is the problem concerned indexable, but also all the index functions are $1-1$ (namely, that distinct states of an asset have distinct index values). Our analysis yields necessary and sufficient conditions for strict indexability in both models, together with formulae for the indices in closed form. We believe this to be the first time that simply stated conditions which are equivalent to strict indexability have been achieved for any restless bandit model for which strict indexability is not guaranteed. Numerical results testify the very strong performance of the index heuristic for both models. Section 5.3 contains a somewhat brief discussion of the index structure of versions of the spinning plates problem and the squad system with discounted reward criterion.

In addition to the intrinsic interest of the theoretical results in Sections 5.1 and 5.2, we believe that the approach we adopt will be applicable to a wide range of restless bandit problems with the average reward criterion. Investigation of an asset's index structure involves the study of the so-called *W-subsidy problem*[4]. The latter being a decision problem defined for the asset of interest in which a subsidy W is paid for every unit of time for which the asset is passive. Indexability of the asset is related simply to the fact that the value of the *W-subsidy* problem is increasing, piecewise linear and convex in W. See

[4] See discussion on Whittle's approach in pages 56 to 58.

also Niño-Mora [76]. For strictly indexable cases the number of segments in the piecewise linear function is one greater than the number of asset states.

Comments

A. Each of the families of restless bandits considered here is a class of *Markov Decision Processes* (MDP's) with the *average reward criterion.* In each case M projects (assets) are available for investment/exploitation. resource constraint means that only Q assets ($1 \leq Q \leq M$) my be active at any time. The decision problem concerns how assets should be optimally chosen for activation at each decision epoch of the system to maximise the reward earned over an infinite horizon.

B. Each of the assets or projects evolves stochastically through time $t \in \mathbb{R}^{+}$. We write $X_i(t)$ for the state of asset i at time $t \in \mathbb{R}^{+}$, $1 \leq i \leq M$ and $\mathbf{X}(t) = \{X_1(t), \ldots, X_M(t)\}$ for the corresponding *system* state. The state of asset i is an integer in the range $\left[\underline{K}_i, \overline{K}_i\right] \equiv \left\{\underline{K}_i, \underline{K}_i + 1, \ldots, \overline{K}_i\right\}$, and for most of the development (and until stated otherwise) we shall suppose that $-\infty < \underline{K}_i < \overline{K}_i < \infty$ for $1 \leq i \leq M$.

C. An issue which arises in consideration of the *W-subsidy problem* in (2.32) is the possible non-uniqueness of the policy(ies) achieving the maximum. We resolve any non-uniqueness in two steps. First, we demonstrate (see Lemmas 5.1 and 5.2) that for both families I and II there exist optimal policies for the *W-subsidy* problems of interest which have *monotone structure.* Hence we restrict to policies from the appropriate monotone class in each case. Second, should more than one monotone policy achieve the maximum in (2.32) we choose the policy with the largest *passive set* (i.e. the largest set of states in which the corresponding policy chooses the passive action).

D. Use $\pi_i(W)$ for the resulting policy, $\Pi_i(W)$ for its passive set and $\pi(W)$ for the policy for the entire system which applies $\pi_i(W)$ to each asset i, $1 \leq i \leq M$. Policy $\pi(W)$ solves the optimization problem (2.30). The following definition adapts Definition 2.1 in Section 2.3 and expresses a natural requirement on (optimal) policy structure.

Definition 5.1. *Asset i is indexable if $\exists\, \underline{W}_i$, $\overline{W}_i$ such that $-\infty < \underline{W}_i < \overline{W}_i < \infty$ and $\Pi_i(W) = \emptyset$, $W \leq \underline{W}_i$; $\Pi_i(W) = \left[\underline{W}_i, \overline{W}_i\right]$, $W > \overline{W}_i$; with $\Pi_i : \left[\underline{W}_i, \overline{W}_i\right] \to 2^{\left[\underline{W}_i, \overline{W}_i\right]}$ increasing.*

The above decision problems are indexable when all constituent projects are.

Should an asset be indexable then a natural calibration, in the form of a *fair subsidy for passivity*, may be defined. This is again an adaptation of Definition 2.2 to our current framework.

Definition 5.2. *If asset i is indexable, then its index $W_i : \left[\underline{K}_i, \overline{K}_i\right] \to \mathbb{R}$ is defined by*

$$W_i(x) = \inf\{W;\ x \in \Pi_i(W)\}, \ x \in \left[\underline{K}_i, \overline{K}_i\right].$$

In sections 5.1 and 5.2, we shall study Families I and II in turn. In each case we shall give sufficient conditions for the indexability of the decision problems together with algorithms which yield the resulting indices. We further give necessary and sufficient conditions for the *strict indexability* of each asset. Under *strict indexability* the indices are available in closed form. Section 5.3 contains a discussion of the index structure of the spinning plates problem with the discounted reward criterion. Section sec:BiDirCon concludes this chapter.

5.1 Family I: The Spinning Plates Problem (a model for optimal investment in assets)

The family of restless bandits considered here is a class of *Markov Decision Processes* (MDP's) with the *average reward criterion.* Here, M projects (assets) are available for investment. Resource constraint means that only Q assets ($1 \leq Q \leq M$) may be active at any time. The decision problem concerns how assets must be optimally chosen for activation at each decision epoch of the system to maximise the reward earned over an infinite horizon.
A typical member of the *Spinning Plates* family can be described as:

1. Each of the assets evolves stochastically through time $t \in \mathbb{R}^{+}$. We write $X_i(t)$ for the state of asset i at time $t \in \mathbb{R}^{+}$, $1 \leq i \leq M$ and $\mathbf{X}(t) = \{X_1(t), \ldots, X_M(t)\}$ for the corresponding *system* state.

   The state of asset i is an integer in the range $\left[\underline{K}_i, \overline{K}_i\right]$, and for most of the development (and until stated otherwise) we shall suppose that $-\infty < \underline{K}_i < \overline{K}_i < \infty$ for $1 \leq i \leq M$.

2. Time 0 together with the times of every state transition of the process constitute the set of decision epochs of the system. In each system state, there are $\binom{M}{Q}$ possible actions, one corresponding to each subset of $\{1, 2, \ldots, M\}$ of size Q. If $\mathcal{Q}$ is one of such subsets, then $A(\mathcal{Q})$ denotes the action which chooses an active regime (the active action denoted $a_i = 1$) for the assets whose identifiers are in $\mathcal{Q}$ and which chooses an inactive regime (the passive action, $a_i = 0$) for the remaining assets.

   Under action $A(\mathcal{Q})$ applied in state $\mathbf{x}$, the time to the next system transition is exponentially distributed with rate

$$\Delta(\mathcal{Q},\mathbf{x}) \equiv \sum_{i\in\mathcal{Q}} \mu_i(x_i) + \sum_{i\notin\mathcal{Q}} \lambda_i(x_i)$$

If $\Delta(\mathcal{Q},\mathbf{x}) > 0$ then the state immediately following this transition will be $\mathbf{x}+\mathbf{e}_i$ for $i \in \mathcal{Q}$ with probability $\mu_i(x_i)/\Delta(\mathcal{Q},\mathbf{s})$, and will be $\mathbf{x}-\mathbf{e}_i$ for $i \notin \mathcal{Q}$ with probability $\lambda_i(x_i)/\Delta(\mathcal{Q},\mathbf{x})$.

Equivalently, the M assets evolve independently under the action applied (active or passive). If project i should be active $(a_i = 1)$ then it evolves from x_i to $x_i + 1$ at rate $\lambda_i(x_i)$, while under the passive action $(a_i = 0)$ it evolves from state x_i to state $x_i - 1$ at rate $\mu_i(x_i)$, $x_i \in \left[\underline{K}_i, \overline{K}_i\right]$, $1 \le i \le M$.

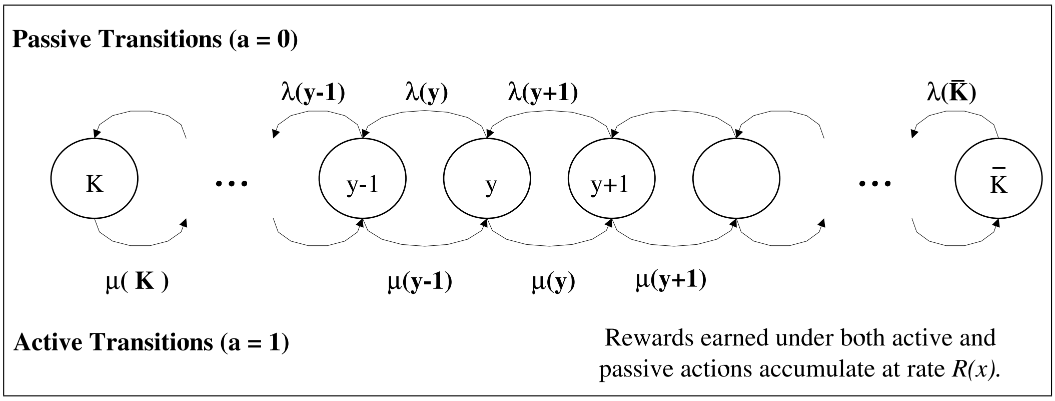
Transition rates μ_i, λ_i satisfy $\mu_i\left(\overline{K}_i\right) = \lambda_i\left(\underline{K}_i\right) = 0$, but are otherwise strictly positive, $1 \le i \le M$. Should we have $\Delta(\mathcal{Q},\mathbf{x}) = 0$, then the state $\mathbf{x}$ is absorbing under action $A(\mathcal{Q})$.

3. The system earns rewards at rate $\sum_i R_i(x_i)$ while in state $\mathbf{x}$. Each reward rate function $R_i : \left[\underline{K}_i, \overline{K}_i\right] \to \mathbb{R}^+$ is (weakly) increasing. The goal of analysis is the determination of a policy (rule for taking actions) which maximises the average system reward rate earned over an infinite horizon or which comes close to doing so.

Figure 5.1 below depicts the general structure of a typical *spinning plates* asset.

Comments

1. In this family the active action applied to an asset enhances its reward earning capacity. Hence, plant and machinery are maintained and updated, employees are trained, products are improved and/or advertised


Passive Transitions (a = 0)
λ(y-1)
λ(y)
λ(y+1)
λ(K̄)
K
y-1
y
y+1
K̄
μ(K)
μ(y-1)
μ(y)
μ(y+1)
Active Transitions (a = 1)
Rewards earned under both active and passive actions accumulate at rate R(x).

If active action is applied to a *spinning plates* project in state y, it evolves to state $y+1$, representing an increase in its productivity. If, otherwise, passive action is taken, then it evolves to state $y-1$ i.e. the profitability of the asset declines.

Figure 5.1: Representation of an Asset in the Spinning Plates Problem

–in short, activity represents positive investment decisions taken with regard to an asset. In absence of such investment decisions (the passive action) the reward earning capacity of an asset tends to decline.

2. Note from §3 above that in Family I, assets earn rewards (at a higher or lower rate) all the time and not only when in receipt of investment. This reward structure in is natural to the envisaged applications. Note that a modification in which assets only earn rewards under the active action has a trivial solution: always apply the active action to those Q assets with largest associated values of $R_i\left(\overline{K}_i\right)$, $1 \leq i \leq M$.

3. The theory of *stochastic dynamic programming* (DP) guarantees the existence of an optimal policy which is stationary, deterministic and Markovian (see Section 2.1 and references therein). The above family falls within the class of intractable *restless bandit* problems, introduced by Whittle, which advocated the deployment of *index heuristics*, such policies emerging from the formulation and solution of Lagrangian relaxations of the original optimization problem. The essentials of Whittle's

approach have been sketched in Section 2.3.

5.1.1 Indexability Analysis

We can drop the asset suffix and consider the *W-subsidy problem* in (2.32) for a single asset drawn from a decision problem in Family I whose associated parameters are $\overline{K}$, $\underline{K}$, $\mu(\cdot)$, $\lambda(\cdot)$ and $R(\cdot)$. From §2 and §3 above recall that, under the application of active action $a = 1$ in state x, the asset evolves to state $x+1$ at rate $\mu(x)$ and earns rewards at rate $R(x)$ while doing so. Under application of the passive action $a = 0$ in state x, the asset evolves to state $x-1$ at rate $\lambda(x)$ and (in the *W-subsidy problem*) earns rewards at rate $R(x)+W$ while doing so. The intermediate goal of analysis is the identification of policies to maximise the average reward earned by the asset over an infinite horizon.

Without any loss of generality, we restrict to the class of *stationary, deterministic* and *Markovian* policies $\mathcal{M}$ for which

$$\pi : \left[\underline{K}, \overline{K}\right] \to \{0,1\}$$

and highlight the class $\mathfrak{M}$ of *monotone policies* for which

$$\pi_y(x) = 1 \Leftrightarrow x \leq y, y \in \left[\underline{K}-1, \overline{K}\right], \tag{5.1}$$

hence, policy $\pi_{\overline{K}}$ chooses the active policy ($a = 1$) in all states, while policy $\pi_{\underline{K}-1}$ chooses passive action ($a = 0$) in all states.

Consider any initial state $X(0) = x$ and assume $y < x \leq \overline{K}$. Under monotone policy π_y, the passive action is taken and the asset reaches state y in finite time almost surely and thereafter has alternating sojourns in states y and $y+1$. If, otherwise, $\underline{K} \leq x \leq y$, then active action is taken and the asset will reach state y in finite time almost surely, alternating, thereafter, between states y and $y+1$. Hence, the time average reward rate of policy $\pi_y \in \mathfrak{M}$ for the *W-subsidy* problem is given by

$$\frac{R(y)\,\lambda(y+1) + (R(y+1) + W)\,\mu(x)}{\lambda(y+1) + \mu(y)} \tag{5.2}$$

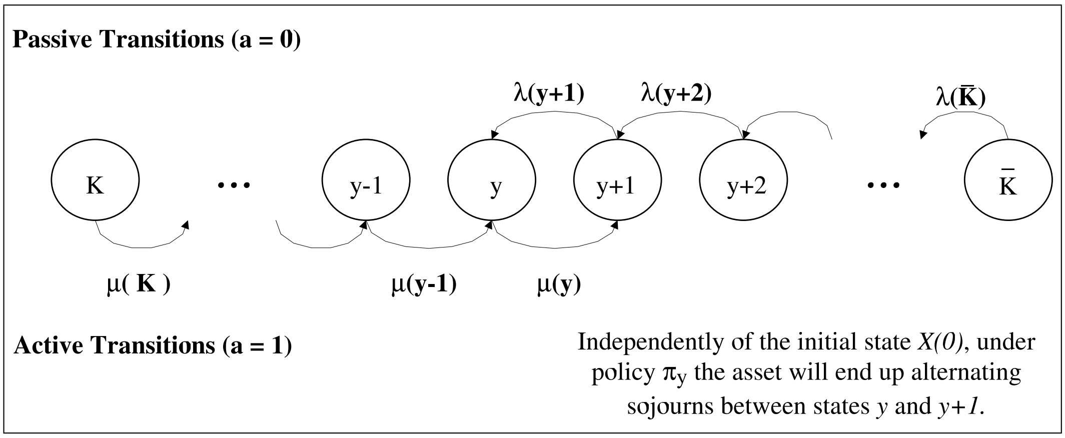
for any initial state $X(0)$.

Figure 5.2 depicts the evolution of an isolated *spinning plates* asset under monotone policy π_y.

We now introduce the function $\phi : \left[\underline{K} - 1, \overline{K}\right] \to \{0, 1\}$ defined by

$$\phi(x) = \frac{\mu(x)}{\lambda(x+1) + \mu(x)}, \quad \underline{K} - 1 \leq x \leq \overline{K} \tag{5.3}$$

with $\phi\left(\overline{K}\right) = 0$, and $\phi\left(\underline{K} - 1\right) = 1$.



Independently of the initial state *X(0)*, under policy πy the asset will end up alternating sojourns between states *y* and *y+1*.

Policy $\pi_y \in \mathfrak{M}$ prescribes taking active action in states $\{\underline{K}, \ldots, y\}$ and passive in states $\left\{y+1, \ldots, \overline{K}\right\}$. It can be seen that under such policy, for any initial state $X(0)$, the project will be trapped between states y and $y+1$.

Figure 5.2: Evolution of a Spinning Plates Asset under a Monotone Policy

Lemma 5.1. *For all $W \in \mathbb{R}$, there exists an optimal policy π^* for the W-subsidy problem, such that $\pi^* \in \mathfrak{M}$.*

Proof *Consider asset evolution under a general stationary, deterministic, Markovian policy* $\tilde{pi} \in \mathcal{M}$*. There exists a policy* $\pi_y \in \mathfrak{M}$ *which yields an average reward rate identical to the one given by* $\tilde{\pi} \in \mathcal{M}$ *and is independent of the initial state* $X(0)$*.*

Take for example state $X(0) = x$ *and assume* $\tilde{\pi}(x) = 0$ *(i.e. the general policy prescribes passive action). The project will evolve to state* $x-1$ *according to* $\lambda(x)$ *and will keep taking the passive action until the first arrival to some state* $y < x$ *for which* $\tilde{\pi}(y) = 1$ *(active action). From that moment on the project alternate between states* y *and* $y+1$*, with average reward given by*

$$V_{\tilde{\pi}}(X(0) = x) = R(y)(1 - \phi(y)) + (R(y+1) + W)\phi(y).$$

It can easily be seen that the expression above corresponds to the average reward rate for the W*-subsidy problem under monotone policy* π_y*, as obtained upon substitution of* (5.3) *in* (5.2)*.*

Take now the monotone policy maximising the average reward rate for a given W*,*

$$\hat{y} = \underset{\underline{K}-1 \leq y \leq \overline{K}}{\operatorname{argmax}} \left\{V_{\pi_y}\right\} \tag{5.4}$$

in case of more than one value of y *satisfying the maximum in the r.h.s. of* (5.4)*, fix* $\hat{y}$ *to be the smallest* y *achieving such maximum. Fix* $\pi^* = \pi_{\hat{y}}$*, then it is clear that*

$$V_{\pi^*} \geq V_{\tilde{\pi}}(X(0))$$

for any $X(0)$ *and all* $\tilde{\pi} \in \mathcal{M}$*. Consequently,* $\pi^* \in \mathfrak{M}$*, as defined above, is an optimal policy for the* W*-subsidy problem.* ***q.e.d.*** ■

Let's introduce the set X^W:

$$X^W = \left\{ x : \underset{\underline{K}-1 \le x \le \overline{K}}{\text{argmax}} \left\{ R(x)(1-\phi(x)) + (R(x+1+W)\phi(x)) \right\} \right\}$$

and define $X(W)$ to the smallest over all $x \in X^W$ minus one, i.e.,

$$X(W) = \inf \left\{ x - 1 : x \in X^W \right\} \tag{5.5}$$

hence, from Lemma 5.1, $X(W)$ is the policy with maximal passive set solving the W-*subsidy* problem.

From Definition 5.1, in order to establish the asset indexability, it will be enough to show that there exist finite $\underline{W} < \overline{W}$ such that:

$$\begin{aligned} X(W) &= \underline{K} - 1, && \text{for all } \ W \ge \overline{W} \\ X(W) &= \overline{K}, && \text{for all } \ W < \underline{W} \end{aligned}$$

and

$$X(\cdot) : \left[\underline{W}, \overline{W}\right] \to \left[\underline{K} - 1, \overline{K}\right]$$

is a decreasing function of W.[5]

For an *indexable* asset, the index in state x will be given by

$$W(x) = \inf \{W; X(W) \le x - 1\}, \ \underline{K} \le x \le \overline{K} \tag{5.6}$$

Observe that, from (5.2) and (5.3), the average reward rate achieved by

[5] For the particular case of the spinning plates problem, notice that $\underline{W} = 0$.

policy π_x for the W-*subsidy* problem is written

$$R(x)(1-\phi(x)) + (W + R(x+1))\,\phi(x) \tag{5.7}$$

Theorem 5.1. *If ϕ is decreasing, the asset is indexable.*

Proof *Define, for $W \geq 0$*

$$V(W) = \max_{\underline{K}-1 \leq x \leq \overline{K}} \left\{ R(x)(1-\phi(x)) + (W + R(x+1))\,\phi(x) \right\}. \tag{5.8}$$

From the discussion around (5.5)*, $X(W)$ is the smallest maximizer of the r.h.s. of* (5.8)*. It is straightforward to show that, since $0 \leq \phi(x) \leq 1$, and $\underline{K} \leq x \leq \overline{K} - 1$, then*

$$X(W) = \overline{K},\ W < 0, \tag{5.9}$$

and

$$X(W) = \underline{K} - 1,\ W \geq \overline{W} \tag{5.10}$$

for some $\overline{W}$ large enough[6]*.*

Further, if the subset $\{x_1, x_1+1, \ldots, x_2\} \subseteq \left[\underline{K}-1, \overline{K}\right]$ is such that $\phi(x_1) = \phi(x_1+1) = \cdots = \phi(x_2)$ and is maximal in this regard, then it is straightforward to show that the range of $X(W)$ contains at most a single value from this subset.

Finally, note that $V : \mathbb{R}^+ \to \mathbb{R}^+$ is an increasing function of W. Moreover, it is piecewise linear and convex, with each segment's r.h.s. gradient given by $\phi(X(W))$. Take now any pair $(W_1, W_2) \in \left[\underline{W}, \overline{W}\right]$ such that $W_1 > W_2$, it

[6] *In particular, $\overline{W}$ will be the value of W such that $V_{\pi_{\underline{K}-1}} > \max_{\underline{K} \leq x \leq \overline{K}} \{V_{\pi_x}\}$.*

immediately follows from V's convexity, the hypothesis of the theorem and the foregoing discussion that

$$X\left(W_2\right) > X\left(W_1\right). \tag{5.11}$$

The result now follows from (5.8) *to* (5.11) *and the discussion around* (5.5). ***q.e.d.*** ■

We now seek to understand the asset's index structure under the hypothesis of Theorem 5.1. Suppose that there are $L \geq 1$ points at which the gradient of V is discontinuous. List the corresponding W-values as

$$0 < W^1 < \cdots < W^L,$$

where plainly $W^L \leq \overline{W}$ from (5.10). Write $W^0 = 0$. Use now $x_l,\ 0 \leq l \leq L-1$, for the integers for which $X\left(W\right) = x_l - 1$, $W \in \left[W^l, W^{l+1}\right)$ and which satisfy

$$\phi\left(x_l\right) = \left[V\left(W^{l+1}\right) - V\left(W^l\right)\right]\left[W^{l+1} - W^l\right]^{-1},\ 0 \leq l \leq L-1$$

Also write $x_L = \underline{K} - 1$. The convexity of V and the decreasing nature of ϕ imply that

$$\overline{K} = x_0 \geq x_1 > \cdots > x_L = \underline{K} - 1.$$

We now complete the description of $X\left(\cdot\right)$ as

$$X\left(W\right) = \begin{cases} \overline{K}, & W < 0, \\ x_l - 1, & W \in \left[W^l, W^{l+1}\right),\ i \leq l \leq L-1 \\ \underline{K} - 1, & W \geq W^L \end{cases} \tag{5.12}$$

The following result is an immediate consequence of (5.6) and (5.12).

Theorem 5.2. *If ϕ is decreasing, then the index $W : \left[\underline{K}, \overline{K}\right] \to \mathbb{R}^+$ is given by*

$$W(x) = \begin{cases} 0, & x_1 < x \leq \overline{K}, \\ W^l, & x_l < x \leq x_{l-1}, 1 \leq l \leq L. \end{cases}$$

hence, the procedure for finding the indices for the spinning plates problem consists in two iterative steps involving identifying the points where the gradient of the piecewise linear function changes. This procedure and the working of the corresponding algorithm in Figure 5.4 is illustrated with an example in Figure5.3.

By suitable applications of (5.6), (5.8), and (5.12) and the discussion around Theorem 5.2 we have:

1. *Find the smallest value of W such that passive action is taken in some state.*

   For each state x we need to identify the value of W such that we are indifferent between applying policy π_x and being active in all states (i.e. applying policy $\pi_{\overline{K}}$). In particular, as $\phi\left(\overline{K}\right) = 0$ it holds that average reward rate for policy $\pi_{\overline{K}}$ is simply $R\left(\overline{K}\right)$, hence we need to solve

   $$R\left(\overline{K}\right) = R(x)(1 - \phi(x)) + (R(x+1) + W)\phi(x).$$

   Call $\widetilde{W}(x)$ the value of W solving the expression above for every $\underline{K} - 1 \leq x < \overline{K}$. Fix

   $$\mathbf{x} = \left\{ x \in \underset{\underline{K}-1 \leq x \leq \overline{K}}{\operatorname{argmin}} \left\{\widetilde{W}(x)\right\} \right\}$$

and let $x_0 = \inf\{x; x \in \mathbf{x}\}$. Fix $W_0 = \widetilde{W}(x_1)$ and, from (5.12), we see that $X(W_0) = x_0 - 1$.

2. *Construct a collection of intercepts for every linear segment in V.*

   Starting from $i = 1$, for every subsequent x_i let $\widetilde{W}_{i+1}(x)$ be the solution to

$$\begin{aligned} R(x_i)(1-\phi(x_i)) + (R(x_i+1)+W)\phi(x_i) \\ = R(x)(1-\phi(x)) + (R(x+1)+W)\phi(x), \end{aligned}$$

   i.e.

$$\begin{aligned} \widetilde{W}_{i+1}(x) = {} & [R(x_i)(1-\phi(x_i)) + R(x_i+1)\phi(x_i) \\ & - R(x)(1-\phi(x)) - R(x+1)\phi(x)][\phi(x)-\phi(x_i)]^{-1} \end{aligned} \quad (5.13)$$

   for all $\underline{K} - 1 \leq x \leq x_i - 1$. Set

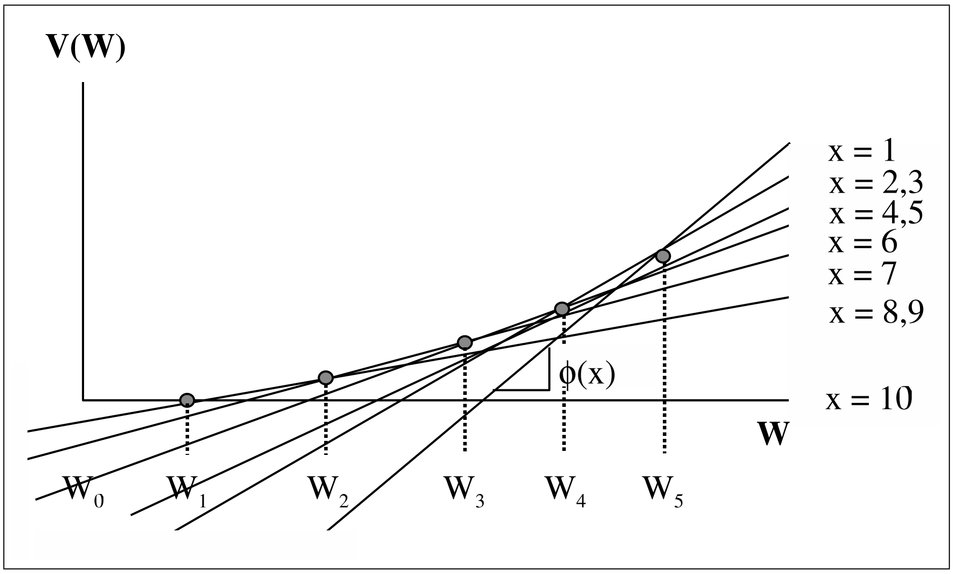
$$\mathbf{x}_{i+1} = \left\{ x \in \underset{\underline{K}-1 \leq x \leq x_i - 1}{\operatorname{argmin}} \left\{ \widetilde{W}_{i+1}(x) \right\} \right\} \quad (5.14)$$

   and $x_{i+1} = \inf\{x \in \mathbf{x}_{i+1}\}$. Fix $W^{i+1} = \widetilde{W}_{i+1}(x_{i+1})$ and $X(W_{i+1}) = x_{i+1} - 1$.

   Increase i and repeat until $X(W_j) = \underline{K} - 1$ and finally set $L = j$.

By a modest extension of the above calculations, it follows that the minimisers of the r.h.s. of (5.14) are precisely the maximisers of $V(W^{i+1})$. By definition, $X(W^{i+1})$ is the smallest of these minus one. The indices for all other states x are obtained according Theorem 5.2.

The procedure in the previous lines is summarized in the Algorithm in Figure 5.4.



The general procedure for obtaining the indices in a typical spinning plates project is depicted in the following example. Consider the interval $[W_2, W_3)$. It must be clear from the graph that $X^{W_2} = \{7, 8, 9\}$ and $X^W = \{7\}$, for $W \in (W_2, W_3)$. According to expression (5.5), $X(W) = 6$, $W \in [W_2, W_3)$. Following the same reasoning we can construct the collection $\{x_i = X(W) : W_i \leq W < W_{i+1}\}$, e.g. for the particular case discussed above we have $x_2 = 6$. Finally, in order to finding the $W(x)$'s recall the definition in expression (5.6) and notice that, for example, $W(7) = \inf\{W_2, \ldots, W_5\} = W_2$. The same reasoning applies to every $W(x)$, $x \in [1, 10]$. In our particular example, it can be verified that $W(1) = W_5$, $W(2) = W_5$, $W(3) = W_4$, $W(4) = W_3$, $W(5) = W_3$, $W(6) = W_3$, $W(7) = W_2$, $W(8) = W_1$, $W(9) = W_1$, $W(10) = W_0 = 0$.

Figure 5.3: Example of the Procedure for Finding the Whittle Indices in the Spinning Plates Problem

Comments

1. Please notice from Theorem 5.2 that the index is decreasing in the state. Hence in the spinning plates problem are the assets which are giving low returns (wobbly plates) which are assigned high priority for activation.

2. Note also that the sufficient condition in Theorems 5.1 and 5.2 that ϕ is decreasing, is equivalent to the requirement that the ratio between the

```
INPUT: $\overline{K}, \underline{K}, R(x), \phi(x)$
INITIALIZATION:
Set $\mathbf{x}_0 = \left\{ x : x \in \underset{\underline{K}-1 \le x \le \overline{K}-1}{\operatorname{argmin}} \left\{ \widetilde{W}_0(x) = \frac{R(\overline{K}) - R(x)(1-\phi(x)) - R(x+1)(1-\phi(x+1))}{\phi(x)} \right\} \right\};$
    $x_0 = \inf \{x \in \mathbf{x}_0\};$
    $W_0 = \widetilde{W}_0(x_0);$
    $X(W_0) = x_0 - 1;$
    $\Pi_1 = \{x_0, \ldots, \overline{K}\};$
    $i = 0;$
PROCEDURE:
While $\Pi_i \ne \{\underline{K}, \ldots, \overline{K}\}$ do
Set $\widetilde{W}_{i+1}(x) = \frac{R(x_i)(1-\phi(x_i)) + R(x_i+1)\phi(x_i) - R(x)(1-\phi(x)) - R(x+1)\phi(x)}{\phi(x) - \phi(x_i)},$
                                                    $\underline{K} - 1 \le x \le x_i - 1;$
    $\mathbf{x}_{i+1} = \left\{ x : x \in \underset{\underline{K}-1 \le x \le x_i-1}{\operatorname{argmin}} \left\{ \widetilde{W}_{i+1}(x) \right\} \right\};$
    $x_{i+1} = \inf \{x \in \mathbf{x}_{i+1}\};$
    $W_{i+1} = \widetilde{W}_{i+1}(x_{i+1});$
    $X(W_{i+1}) = x_{i+1} - 1;$
    $\Pi_{i+1} = \Pi_i \cup \{x_{i+1}, \ldots, x_i - 1\};$
    $i = i + 1.$
End While
Set $L = i$
OUTPUT: $\mathbf{W}$, $\mathbf{X}$, $L$.
```

Figure 5.4: Adaptive Greedy Algorithm for the Spinning Plates Problem

active and passive rates for moving between x and $x+1$, namely

$$\mu(x) / \lambda(x+1)$$

is decreasing in x.

Important special cases occur in which all of the states in an indexable asset have distinct indices. When this happens we say that the asset is *strictly* *indexable.* The result in the next section gives a necessary and sufficient condition for *strict indexability.*

5.1.2 Strict Indexability

Given the structure of the spinning plates problem, under *general indexability*, one single value of the Whittle index, $W(x)$, can be attached to two or more adjacent states. In this section we introduce the notion of *strict indexability* which requires that one and only one Whittle index will correspond to each individual state. We also explore the corresponding *strict indexability* conditions.

We already have a *sufficient* condition for indexability, i.e.

$$\phi(x) = \frac{\mu(x)}{\mu(x) + \lambda(x+1)}, \ \underline{K} - 1 \leq x \leq \overline{K} \tag{5.15}$$

must be decreasing in x. Hence we need to identify some additional (neccesary) condition(s) for strict indexability to be achieved.

We have already introduced the expressions:

$$X(W) = \inf\left\{x - 1 : x \in x^W\right\}$$

and

$$W(x) = \inf\left\{W; X(W) \leq x - 1\right\}$$

which imply that the relevant values of W are the ones corresponding to intersections of linear segments in (5.8). Under strict indexability, each segment in V should correspond to one and only one state x. For any pair of adjacent states x and $x - 1$, the value of W at the intersection will be given by the

W-solution to

$$R(x)(1-\phi(x)) + (R(x+1)+W)\phi(x) = R(x-1)(1-\phi(x-1)) + (R(x)+W)\phi(x-1) \quad (5.16)$$

Alternatively, there exists some $\widetilde{W}$ such that the reward rates, at state x, under the active and passive actions are identical, i.e. the operator is indifferent between activity and passivity. This value will be the index corresponding to state x, i.e.

$$\widetilde{W}(x) = \frac{R(x)(1-\phi(x)) + R(x+1)\phi(x) - R(x-1)(1-\phi(x-1)) - R(x)\phi(x-1)}{\phi(x-1)-\phi(x)} \quad (5.17)$$

which is precisely the intersection between adjacent segments of V corresponding to x and $x-1$. If the expression above is strictly increasing or decreasing in x, then to every pair $x, x-1$ will correspond a different value of $\widetilde{W}(x)$.

The next theorem follows from the discussion above:

Theorem 5.3. ***Strict Indexability***

1. *The following are equivalent:*

   (a) *The asset is strictly indexable*

   (b) *Both $\phi(x)$ and $\widetilde{W}(x)$ as given by* (5.17) *are strictly decreasing over $\underline{K}-1 \leq x \leq \overline{K}$.*

2. *The conditions in 1b imply*

$$W(x) = \widetilde{W}(x), \; \underline{K}-1 \leq x \leq \overline{K}.$$

   and the index is strictly decreasing in the state.

Proof *We start or proof by showing that* (1b) *implies* (1a).
Notice that condition (1b) guarantees that the piecewise value function V will have one segment for each state $x \in \left[\underline{K}-1, \overline{K}\right]$ and the minimum in the r.h.s. of (5.14) *will be achieved uniquely by $x = x-1$. We will then have $W = \widetilde{W}(x)$, and the inference of strict indexability will follow simply from Theorem 5.2.*

To prove that (1a) *implies* (1b) *notice that if the asset is strictly indexable there must exist $\overline{W}(x)$, $\underline{K} \leq x \leq \overline{K}$, strictly decreasing in x, such that*

$$X(W): \begin{cases} \overline{K}, & W < \overline{W}\left(\overline{K}\right) \\ x, & W \in \left(\overline{W}(x-1), \overline{W}(x)\right] \\ \underline{K}-1, & W \geq \overline{W}\left(\underline{K}-1\right) \end{cases} \tag{5.18}$$

Plainly, when $W = \overline{W}(x)$, $\underline{K} \leq x \leq \overline{K}$, both x and $x-1$ achieve the maximum in $V(W)$. It follows that

$$\begin{aligned} R(x)(1-\phi(x)) + \left(R(x+1) + \overline{W}(x)\right)\phi(x) > \\ R(x-1)(1-\phi(x-1)) + \left(R(x) + \overline{W}(x)\right)\phi(x-1). \end{aligned} \tag{5.19}$$

Moreover, if $W \in \left(\overline{W}(x-1), \overline{W}(x)\right]$, $X(W) = x$ and so π_x must strictly outperform π_{x-1} in this range. Hence, for $W \in \left(\overline{W}(x-1), \overline{W}(x)\right]$,

$$\begin{aligned} R(x)(1-\phi(x)) + (R(x+1) + W)\phi(x) > \\ R(x-1)(1-\phi(x-1)) + (R(x) + W)\phi(x-1). \end{aligned} \tag{5.20}$$

It must then follow from (5.19) *and* (5.20) *that $\phi(x-1) > \phi(x)$, $\underline{K}-1 \leq x \leq \overline{K}$, and hence that $\phi(x)$ is strictly decreasing over $\underline{K} \leq x \leq \overline{K}$. Further,* *solving* (5.20) *for* $\overline{W}(x)$ *we obtain:*

$$\overline{W}(x) = \widetilde{W}(x), \ \underline{K} \leq x \leq \overline{K}.$$

We conclude that $\widetilde{W}(x)$ *is strictly decreasing in* x *over* $\underline{K} \leq x \leq \overline{K}$. *This concludes the proof of part 1. Part 2 follows trivially from the above analysis.* ***q.e.d.*** ■

Comment

Notice that the index for state x in (5.17) involves quantities evaluated at $x-1$, x, and $x+1$. It may be understood as a quantity which weighs the benefits of the positive reward enhancement achieved by the active action taken in x (the positive term) against the effects of reward deterioration experienced when the asset is passive (the negative term).

5.1.3 Examples

Example 5.1. *Assume a linear reward rate:*

$$R(x) = r(x - \underline{K}) \tag{5.21}$$

for $r > 0$.

Then, the function $\widetilde{W}$ *in* (5.17),

$$\begin{aligned} \widetilde{W}(x) = \Big[& R(x)(1-\phi(x)) + R(x+1)\phi(x) \\ & - R(x-1)(1-\phi(x-1)) - R(x)\phi(x-1) \Big] \times \Big[\phi(x-1) - \phi(x) \Big]^{1}, \end{aligned}$$

with $\underline{K} \leq x \leq \overline{K}$, *becomes*

$$\widetilde{W}(x) = r\left[\frac{1}{\phi(x-1) - \phi(x)} - 1\right], \qquad \underline{K} \leq x \leq \overline{K} \tag{5.22}$$

Under strict indexability,

1. $\phi(x)$ *must be strictly decreasing, and*

2. $\widetilde{W}(x)$ *must be strictly decreasing.*

Hence $\widetilde{W}(x+1) < \widetilde{W}(x)$ *requires*

$$\frac{\phi(x-1) - \phi(x)}{\phi(x) - \phi(x+1)} < 1 \tag{5.23}$$

i.e. $\phi(x-1) - \phi(x)$ *must be strictly increasing in* x *or, equivalently,* $\phi(x)$ *must be concave (strictly decreasing). From Theorem 5.3,* (5.22) *gives the index in this case.*

The particular case where $\mu(x) = \overline{\mu}\left(\overline{K} - x\right)$ and $\lambda(x) = \overline{\lambda}(x - \underline{K})$, for $\underline{K} \leq x \leq \overline{K}$, meets this requirements for any pair $\left(\overline{\mu}, \overline{\lambda}\right)$ provided that $\frac{\overline{\mu}}{\overline{\lambda}} > 1$. This will be the case discussed in the numerical analysis.

Example 5.2. *It is possible to develop an indexable asset with semi-infinite state space of the form* $\left(-\infty, \overline{K}\right]$. *This is a natural extension of the above material. Consider such an example for which* $\overline{K} = 0$ *and assume assume a non-linear, non-decreasing reward rate, for example*

$$R(x) = re^{\tau x},\ x \leq 0, \tag{5.24}$$

where $r > 0$ *and* $\eta > 0$. *Further, suppose that*

$$\phi(x) = 1 - e^{\theta x},\ x \leq 0, \tag{5.25}$$

where $\theta > 0$ *guarantees that* ϕ *is strictly decreasing. The function* $\widetilde{W}$ *in* (5.17) *now becomes*

$$\widetilde{W}(x) = r(e^{\tau} - 1)\left(1 - e^{-\theta}\right)^{-1}\left[e^{(\tau-\theta)x} - e^{\tau x}\left(1 - e^{-\tau-\theta}\right)\right], \; x \le 0 \qquad (5.26)$$

and will be strictly decreaing when $0 < \tau < \theta$. *From a suitable extension of Theorem 5.3,* (5.26) *gives the index in this case*[7].

Example 5.3. *One more example of indexable assets with semi-infinite state space is described below. Consider again an example for which* $\overline{K} = 0$,*but the reward rates are given by*

$$R(x) = r(2 - x)^{-\alpha}, \quad \textit{with} \;\; \alpha > 0 \; \textit{and} \, r > 0 \qquad (5.27)$$

Further suppose that

$$\phi(x) = 1 - (1 - x)^{-\beta}, \qquad (5.28)$$

with $\beta > 0$ *and* $\phi(0) = 0$, *so that* $\phi(x)$ *is strictly decreasing, the expression in* (5.17) *becomes*

[7] *Notice that*

$$\begin{aligned}\frac{dW(x)}{dx} &= r(e^{\tau} - 1)\left(1 - e^{-\theta}\right)^{-1}\left[(\tau - \theta)e^{(\tau-\theta)x} - \tau e^{\tau x}\left(1 - e^{-\theta-\tau}\right)\right] < 0 \\ &\Rightarrow (\tau - \theta)e^{(\tau-\theta)x} < \tau e^{\tau x}\left(1 - e^{-\theta-\tau}\right) \\ &\Rightarrow (\tau - \theta)e^{-\theta x} < \tau\left(1 - e^{-\theta-\tau}\right) \\ &\Leftrightarrow 0 < \tau < \theta.\end{aligned}$$

$$\widetilde{W}(x) = r\left[\frac{\left[(1-x)^{-\alpha}-(2-x)^{-\alpha}\right]\left(1-(1-x)^{-\beta}\right)+\left[(2-x)^{-\alpha}+(3-x)^{-\alpha}\right](2-x)^{-\beta}}{(1-x)^{-\beta}-(2-x)^{-\beta}}\right] \tag{5.29}$$

for $x \leq 0$, and will be strictly decreasing for any pair (α, β) such that $0 < \alpha < \frac{2}{3}\beta$. From a suitable extension of Theorem 5.3, (5.29) *gives the index in this case.*

5.1.4 Numerical Results

Tables 5.1 and 5.2 illustrate some results of an extensive numerical investigation into the quality of performance of the index heuristics developed in this section. Each problem studied has $M = 4$, $Q = 1$; namely a choice has to be made at each decision epoch of one from four possible assets for activation. Four policies were applied to each problem generated. These are as follows:

OPT An optimal policy and its corresponding average reward rate, r^{opt}, were computed by DP value iteration. See, for example, Puterman [85] and/or Tijms [99] and the discussion in Sections 2.1 and, in particular, 2.1.4.

IND The index policy developed in the current section. At every decision epoch it activates the asset with currenty maximal index.

MYO A myopic heuristic which attaches the index $r\mu(x)$ to an asset in state x and activates the asset of largest index. This index may be understood as the rate at which the asset's reward earning capacity may be enhanced by activation;

SMA The policy which always activates the asset of smallest state, with ties broken at random.

For each problem generated, the average reward rates r^{ind}, r^{myo}, and r^{sma}, were computed by DP value iteration, yielding the *percentage suboptimalities*:

$$\Xi\left(\mathrm{P}\right) = \frac{r^{\mathrm{opt}} - r^{\mathrm{P}}}{r^{\mathrm{opt}}} \times 100, \qquad \mathrm{P} = \mathrm{ind}, \mathrm{myo}, \mathrm{sma}.$$

Two families of problems, corresponding to Examples 5.1 and 5.2 are discussed.

Example 5.1. *Each constituent asset is structured as in Example 5.1 above, with*

$$\begin{aligned} R\left(x\right) &= r\left(x - \underline{K}\right) \\ \mu\left(x\right) &= \overline{\mu}\left(\overline{K} - x\right), \qquad for\ \underline{K} \leq x \leq \overline{K} \\ \lambda\left(x\right) &= \overline{\lambda}\left(x - \underline{K}\right) \end{aligned} \tag{5.30}$$

with $\overline{\mu}, \overline{\lambda} > 0$. *Moreover, by fixing* $\underline{K} = 0$, $\overline{K} = 8$, *we get*

$$\phi\left(x\right) = \frac{\overline{\mu}\left(8 - x\right)}{\overline{\mu}\left(8 - x\right) + \overline{\lambda}\left(x + 1\right)} \tag{5.31}$$

For any pair $\left(\overline{\mu}, \overline{\lambda}\right)$ *such that* $\overline{\mu} > \overline{\lambda}$, *the function* ϕ *is indeed decreasing.*

Under this model each of the assets is characterised by the parameters $\left(r, \overline{\mu}, \overline{\lambda}\right)$. *In all cases the* $\overline{\mu}$*'s and* $\overline{\lambda}$*'s are chosen by sampling from a continuous uniform distribution, as indicated in Table 5.1. The cases in Group I have r's chosen from a* $U\left(10, 25\right)$ *distribution, those in Group II the r's are drawn from* $U\left(25, 50\right)$, *and the ones in Family II, have r's drawn from* $U\left(10, 50\right)$. *Table 5.1 presents results for 1600 (*$= 3 \times 2 \times 200 + 2 \times 200$*) randomly generated problems. Table 5.1 summarises the collections of percentage suboptimalities (each collection of size* 200*) arising from the application of each of IND, MYO and SMA to each of the eight problem configurations. Each collection is summarised by the order statistics MIN (minimum), LQ (lower quartile), MED* *(median), UQ (upper quartile), and MAX (maximum). For example, from the top left-hand corner of Table 5.1, we see that when MYO is applied to the 400 problems with* $\overline{\mu}, \overline{\lambda} \sim U\,(0.25, 0.75)$ *and* $r \sim U\,(10, 25)$*, we obtain a median percentage suboptimality of* 3.88 *and a worst case which is* 18.64% *suboptimal.*

The dominant feature of Table 5.1 is the outstanding performance of the index policy. In its worst performance in the 1600 randomly generated problems analysed, it was just 0.2438% *suboptimal. For each of the other heuristics (MYO,SMA) problem instances arose in which they performed poorly.*

Table 5.1: Bi-directional Restless Bandits. Spinning Plates Problem. Performance of the Index Policy in the Linear Case, Example 5.1.

Group I, r $\sim U\,(10, 25)$									
	$\overline{\lambda}, \overline{\mu} \sim U\,(0.25, 0.75)$			$\overline{\lambda}, \overline{\mu} \sim U\,(0.10, 0.50)$			$\overline{\lambda}, \overline{\mu} \sim U\,(0.50, 0.90)$		
	Index	Myopic	Small	Index	Myopic	Small	Index	Myopic	Small
Min	0.0000	0.14	0.22	0.0000	0.24	0.69	0.0000	0.19	0.24
LQ	0.0003	2.20	5.82	0.0037	2.37	7.02	0.0000	3.16	7.89
Med	0.0071	3.88	10.21	0.0158	4.72	11.09	0.0001	4.87	12.95
UQ	0.0239	6.41	14.67	0.0388	7.47	14.26	0.0071	7.10	17.69
Max	0.1365	18.64	32.88	0.1873	26.90	29.96	0.0910	19.18	32.89

Group II, r $\sim U\,(25, 50)$									
	Index	Myopic	Small	Index	Myopic	Small	Index	Myopic	Small
Min	0.0000	0.25	0.06	0.0000	0.37	0.12	0.0000	0.18	0.19
LQ	0.0012	1.86	4.26	0.0043	2.19	3.49	0.0000	2.62	5.82
Med	0.0106	3.56	7.63	0.0201	4.01	7.46	0.0014	4.14	9.58
UQ	0.0262	5.83	11.88	0.0462	6.84	11.26	0.0116	6.48	13.26
Max	0.1922	17.55	27.65	0.1796	21.76	28.33	0.2163	17.20	27.01

Group III, r $\sim U\,(10, 50)$						
	$\overline{\lambda}, \overline{\mu} \sim U\,(0.25, 0.75)$			$\overline{\lambda}, \overline{\mu} \sim U\,(0.10, 0.90)$		
	Index	Myopic	Small	Index	Myopic	Small
Min	0.0000	0.24	0.29	0.0000	0.50	0.27
LQ	0.0000	2.58	10.85	0.0026	2.64	8.07
Med	0.0027	4.27	17.37	0.0204	4.62	14.12
UQ	0.0165	7.03	23.98	0.0525	8.32	19.99
Max	0.1424	21.91	43.69	0.2438	28.18	43.97

Example 5.2. *Each constituent asset is structured as in Example 5.2 above, with* $\underline{K} = -8$*,* $\overline{K} = 0$ *and*

$$R(x) = re^{\tau x} \tag{5.32}$$
$$\phi(x) = 1 - e^{\theta x}$$

for $\underline{K} \leq x \leq 0$*, with* $\overline{\tau}, \overline{\theta} > 0$*. Moreover, by solving*

$$\phi(x) = 1 - e^{\theta x} = \frac{\mu(x)}{\mu(x) + \lambda(x+1)}$$

for μ*, we obtain*

$$\mu(x) = \lambda(x+1)\left(e^{-\theta x} - 1\right), \ \underline{K} \leq x \leq 0. \tag{5.33}$$

Finally, we fix

$$\lambda(x) = e^{\theta(x-1)}, \quad for \ \ \underline{K} \leq x \leq 0. \tag{5.34}$$

For any pair (τ, θ) *such that* $0 < \tau < \theta$*, the functions* $\phi(x)$ *and* $W(x)$ *are indeed decreasing.*

Under this model each of the assets is characterised by the parameters (r, τ, θ)*. In all cases the* τ*'s and* θ*'s are chosen by sampling from a continuous uniform distribution, as indicated in Table 5.2. The cases collected in Group I have r's chosen from a* $U(10, 25)$ *distribution, while for those in Group II the r's are drawn from* $U(25, 50)$*. Table 5.2 presents results for 1600 (*$= 4 \times 2 \times 200$*) randomly generated problems. Table 5.2 summarises the collections of percentage suboptimalities (each collection of size* 200*) arising from the application of each of IND, MYO and SMA to each of the eight problem configurations. As before, each collection is summarised by the order statistics*

MIN (minimum), LQ (lower quartile), MED (median), UQ (upper quartile), and MAX (maximum).

Notwithstanding a small decrease in the quality of the index policy in the non-linear case, the dominant feature of Table 5.2 is the outstanding relative performance of the index policy with respect to the other two policies (MYO and SMA). Its worst performance, among the 1600 randomly generated problems analysed, was just 7.5% *suboptimal which compares quite positively with the* 10.5% *of the MYO policy and the* 33.5% *of the SMA policy. It must also be noticed that there were instances for which the index policy performed optimally.*

Table 5.2: Bi-directional Restless Bandits. Spinning Plates Problem. Performance of the Index Policy in the Nonlinear Case, Example 5.2.

Group I, $r \sim U(10,25)$						
	$\tau,\theta \sim U(0.25,0.5)$			$\tau,\theta \sim U(0.5,0.75)$		
	Index	*Myopic*	*Small*	*Index*	*Myopic*	*Small*
Min	*0.0000*	*0.0395*	*0.2213*	*0.0020*	*0.0041*	*0.0714*
LQ	*0.3896*	*0.8087*	*3.4863*	*0.0586*	*0.1235*	*1.9658*
Med	*1.4092*	*2.2569*	*7.3704*	*0.1104*	*0.3374*	*4.5569*
UQ	*2.4564*	*3.8886*	*11.4198*	*0.1855*	*0.6205*	*7.4431*
Max	*6.8372*	*8.8839*	*33.5164*	*0.6534*	*2.6734*	*19.5632*

Group II, $r \sim U(25,50)$						
	Index	*Myopic*	*Small*	*Index*	*Myopic*	*Small*
Min	*0.0048*	*0.0211*	*0.1613*	*0.0013*	*0.0135*	*0.0571*
LQ	*0.3515*	*0.5602*	*2.4603*	*0.0452*	*0.1140*	*1.2960*
Med	*1.5116*	*2.1976*	*5.3602*	*0.0713*	*0.2784*	*3.1098*
UQ	*2.6986*	*4.0218*	*9.0963*	*0.1165*	*0.5686*	*5.1787*
Max	*7.4862*	*10.3958*	*29.0996*	*0.6078*	*2.2983*	*14.2397*

5.2 Family II: The Squad System Problem (a model for optimal exploitation of assets)

As with the Spinning Plates Problem, the family of restless bandits considered here is a class of *Markov Decision Processes* (MDP's) with the *average reward criterion.* Here, M projects (assets) are available for exploitation. Resource constraint means that only Q assets $(1 \leq Q \leq M)$ may be active at any time. The decision problem concerns how assets must be optimally chosen for activation at each decision epoch of the system to maximise the reward earned over an infinite horizon.

A typical member of the *Squad System* family can be described as:

1. Each of the assets evolves stochastically through time $t \in \mathbb{R}^{+}$. We write $X_i(t)$ for the state of asset i at time $t \in \mathbb{R}^{+}$, $1 \leq i \leq M$ and $\mathbf{X}(t) = \{X_1(t) \ldots, X_M(t)\}$ for the corresponding *system* state.

   The state of asset i is an integer in the range $\left[\underline{K}_i, \overline{K}_i\right]$, and for most of the development (and until stated otherwise) we shall suppose that $-\infty < \underline{K}_i < \overline{K}_i < \infty$ for $1 \leq i \leq M$.

2. As in the *Spinning Plates* family, in each system state, there are $\binom{M}{Q}$ possible actions, one corresponding to each subset of $\{1, 2, \ldots, M\}$ of size Q. If $\mathcal{Q}$ is one of each subsets, then $A(\mathcal{Q})$ denotes the action which chooses and active regime (the active action denoted $a_i = 1$ for the assets i whose identifiers are in $\mathcal{Q}$ and which chooses an inactive regime (the passive action $a_i = 0$) for the remaining assets.

   Under action $A(\mathcal{Q})$ applied in state $\mathbf{x}$ the time to the next system transition is exponential with rate

$$\Delta\left(\mathcal{Q},\mathbf{x}\right)=\sum_{i\in\mathcal{Q}}\mu_i\left(x_i\right)+\sum_{i\notin\mathcal{Q}}\lambda_i\left(x_i\right).$$

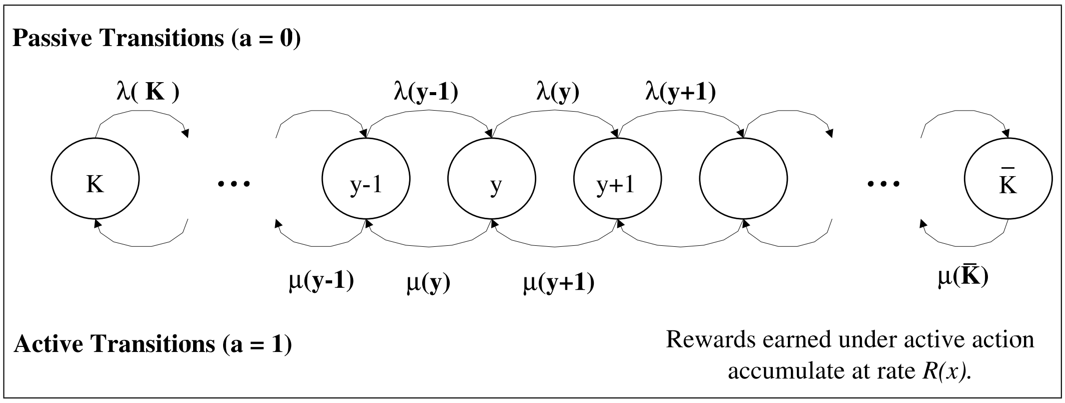
If $\Delta\left(\mathcal{Q},\mathbf{x}\right)>0$ then the state immediately following this transition will be $\mathbf{x}-\mathbf{e}_i$ for $i\in\mathcal{Q}$ with probability $\mu_i\left(x_i\right)/\Delta\left(\mathcal{Q},\mathbf{x}\right)$ and will be $\mathbf{x}+\mathbf{e}_i$ for $i\notin\mathcal{Q}$ with probability $\lambda_i\left(x_i\right)/\Delta\left(\mathcal{Q},\mathbf{x}\right)$. Equivalently, the M assets evolve independently under the action applied (active or passive). If project i should be active ($a_i=1$) then it evolves from x_i to x_i-1 at rate $\mu_i\left(x_i\right)$, while under the passive action ($a_i=0$) it evolves from x_i to x_i+1 at rate $\lambda_i\left(x_i\right)$, $x_i\in\left[\underline{K}_i,\overline{K}_i\right]$, $1\leq i\leq M$.
The transition rates satisfy $\mu_i\left(\underline{K}_i\right)=\lambda_i\left(\overline{K}_i\right)=0$ but are otherwise strictly positive, $1\leq i\leq M$. Should we have $\Delta\left(\mathcal{Q},\mathbf{x}\right)=0$ then the state $\mathbf{x}$ is absorbing under action $A\left(\mathcal{Q}\right)$.

3. If the system is in state $\mathbf{x}$ and action $A\left(\mathcal{Q}\right)$ is taken, then the system earns rewards at rate $\sum_{i\in\mathcal{Q}}R_i\left(x_i\right)$ while in state $\mathbf{x}$. Each reward rate function $R_i:\left[\underline{K}_i,\overline{K}_i\right]\to\mathbb{R}^+$ is (weakly) increasing. The goal of analysis is the determination of a policy (rule for taking actions) which maximises the average system reward rate earned over an infinite horizon or which comes close to doing so.

Figure 5.5 below depicts the general structure of a typical *squad system* asset.

Comments

1. In the *squad system* family, the active action represents the utilisation or exploitation of an asset. As the asset is used it becomes *tired* or *depleted* and loses some of its reward-earning capacity. Under the passive action the asset recovers its potential to earn high returns again.


Passive Transitions (a = 0)
λ(K)
λ(y-1)
λ(y)
λ(y+1)
K
y-1
y
y+1
K̄
μ(y-1)
μ(y)
μ(y+1)
μ(K̄)
Active Transitions (a = 1)
Rewards earned under active action accumulate at rate R(x).

If active action is applied to a *squad system* project in state y, it evolves to state $y-1$, representing the depletion (tyredness) of the project. While active, the project generates profits. When otherwise passive action is taken, it represents the recovery phase and the asset's productivity increases (it evolves to state $y+1$). The passive phase generates no reward.

Figure 5.5: Representation of an Asset in the Squad System Problem

2. Note from §3 in page 195 that assets only earn rewards when they are utilised (i.e. under the active action). This structure is natural to the envisaged application of the *squad system* family. A version in which the assets earn rewards whether activated or not is of little interest. No policy can do better in reward rate terms than an application of the passive action to all states always.

3. As in the *spinning plates* case, the theory of *stochastic dynamic programming* (DP) guarantees the existence of an optimal policy which is stationary, deterministic and Markovian (for discussion on DP and Markov Decision Processes see Section 2.1 and references therein, in particular Puterman, [85] and Ross [90]). The *squad system* family falls within the class of intractable *restless bandit* problems, introduced by Whittle [116] which advocated the deployment of *index heuristics*, such policies emerging from the formulation and solution of Lagrangian relaxations of the original optimization problem. The essentials of Whittle's approach

have been sketched in Section 2.3.

5.2.1 Indexability Analysis

As in the *spinning plates* family, the asset suffix is dropped and the W-*subsidy* problem considered for a single asset with associated $\underline{K}$, $\overline{K}$, $\mu(\cdot)$, $\lambda(\cdot)$ and $R(\cdot)$. From §2 and §3 in page 194, recall that, under the application of active action ($a = 1$) in state x, the asset evolves to state $x - 1$ at rate $\mu(x)$ and earns rewards at rate $R(x)$ while doing so. Under application of the passive action ($a = 0$) in state x, the asset evolves to state $x + 1$ at rate $\lambda(X)$ and (in the W-*subsidy* problem) earns rewards at rate W while doing so. The intermediate goal of analysis is the identification of policies to maximise the average reward rate earned over an infinite horizon.

As in the previous case, we restrict to the class of *stationary, deterministic* and *Markovian* policies $\mathcal{M}$ for which

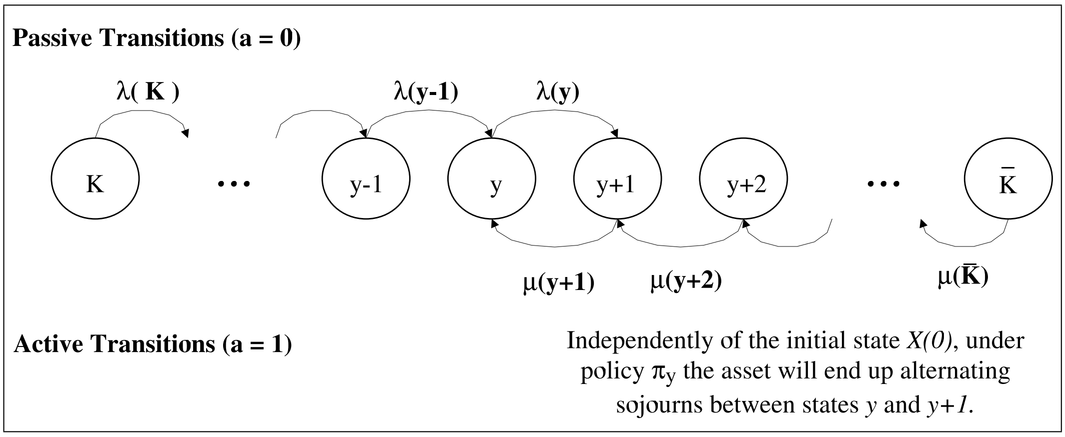
$$\pi : \left[\underline{K}, \overline{K}\right] \to \{0, 1\} \tag{5.35}$$

and introduce the class $\mathfrak{M}$ of *monotone policies* for which

$$\pi_y(x) = 1 \Leftrightarrow x \geq y + 1 \text{ for some } y \in [\underline{K} - 1, K]$$

and write π_y for the policy above. Hence, policy $\pi_{\underline{K}-1}$ chooses action $a = 1$ (active) in all states and policy $\pi_{\overline{K}}$ chooses passivity ($a = 0$) in all states.

Consider, for example, an initial state $X(0) = x$ and assume $y+1 \leq x \leq \overline{K}$. Under monotone policy π_y, the active action is taken and the asset reaches state y in finite time almost surely (where passive action will be taken) and thereafter has alternating sojourns between states y and $y + 1$. hence, the time average



Independently of the initial state *X(0)*, under policy π_y the asset will end up alternating sojourns between states *y* and *y+1.*

Policy $\pi_y \in \mathfrak{M}$ prescribes taking passive action in states $\{\underline{K}, \ldots, y\}$ and active in states $\left\{y+1, \ldots, \overline{K}\right\}$. It can be seen that under such policy the project will be trapped between states y and $y+1$, independently of the initial state.

Figure 5.6: Evolution of a Squad System Asset under a Monotone Policy

reward rate of policy $\pi_y \in \mathfrak{M}$ for the W-*subsidy* problem is given by

$$\frac{R\left(y+1\right)\lambda\left(y\right)+W\mu\left(y+1\right)}{\lambda\left(y\right)+\mu\left(y+1\right)} \tag{5.36}$$

for any initial state $X\left(0\right)$.

Figure 5.6 depicts the evolution of an isolated *squad system* asset under monotone policy π_y.

We now introduce the function $\phi : \left[\underline{K}, \overline{K}\right] \to \{0, 1\}$

$$\phi\left(x\right) = \frac{\mu\left(x+1\right)}{\mu\left(x+1\right)+\lambda\left(x\right)}, \quad \underline{K}-1 \leq x \leq \overline{K}. \tag{5.37}$$

Note that as $\mu\left(\underline{K}\right) = 0$ and $\lambda\left(\overline{K}\right) = 0$, then $\phi\left(\underline{K}-1\right) = 0$ and $\phi\left(\overline{K}\right) = 1$.

Lemma 5.2. *For all $W \in \mathbb{R}$ there exists and optimal policy π^* for the W-subsidy problem, such that $\pi^* \in \mathfrak{M}$.*

Proof *Consider a squad system project's evolution under any general stationary deterministic policy $\tilde{\pi} \in \mathcal{M}$. Then, for any initial state $X\left(0\right)$, there* *exists a monotone policy $\pi_y \in \mathfrak{M}$ whose infinite horizon average reward rate is the same as under $\tilde{\pi}$ and is independent of state $X(0)$.*

Consider, for example, $X(0) = x$ and assume $\tilde{\pi}(x) = 1$ (respectively, $\tilde{\pi}(x) = 0$). Then, the project will evolve according to $\mu(x)$ (respectively $\lambda(x)$), and will keep taking the active (respectively passive) action until arriving to some state $y < x$ (respectively $y + 1 > x$) for which $\tilde{\pi}(y) = 0$ (respectively $\tilde{\pi}(y+1) = 1$). From that moment on, the project will be trapped between states y and $y+1$ with infinite horizon average reward rate given by

$$V_{\tilde{\pi}}(X(0) = x) = R(y+1)(1 - \phi(y)) + W\phi(x).$$

It can easily be seen that the expression above is equivalent to the infinite horizon average reward yielded by monotone policy π_y (i.e. V_{π_y}) as obtained upon substitution of (5.37) *on* (5.36).

Take now the monotone policy for which the maximal average reward rate, over all monotone policies π_x, $x = -1, \ldots, K$, is achieved,

$$\hat{y} = \underset{\underline{K}-1 \le y \le \overline{K}}{\operatorname{argmax}} \left\{ V_{\pi_y} \right\} \tag{5.38}$$

in case of match, we adopt the following convention: if the maximum in the r.h.s of (5.38) *is achived by more than one policy π_x, then fix $\hat{y}$ to be the largest y achieving such maximum. Fix $\pi^* = \pi_{\hat{y}}$, then it is true that*

$$V_{\pi^*} \ge V_{\tilde{\pi}}(X(0)), \textit{ for all } X(0),$$

and for all $\tilde{\pi} \in \mathcal{M}$. Consequently, $\pi^* \in \mathfrak{M}$, as defined above, is an optimal policy for the $W-$subsidy problem. ***q.e.d.*** ■

Consider now the set

$$x^W = \left\{ x : x \in \operatorname*{argmax}_{\underline{K}-1 \leq x \leq K} \left\{ R\left(x+1\right)\left(1-\phi\left(x\right)\right) + W\phi\left(x\right) \right\} \right\}$$

and define $X\left(W\right)$ to be the maximum over all $x \in x^W$; i.e.

$$X\left(W\right) = \sup\left\{x \in x^W\right\}, \tag{5.39}$$

hence, from Lemma 5.2, $X\left(W\right)$ is the policy with maximal passive set solving the W-*subsidy* problem.

From Definition 5.1, in order to establish the asset's indexability, it will be enough to show that there exist finite $\underline{W} < \overline{W}$ such that

$$\begin{aligned} X\left(W\right) &= \underline{K} - 1, & &\text{for all } W < \underline{W} \\ X\left(W\right) &= \overline{K}, & &\text{for all } W \geq \overline{W} \end{aligned}$$

and

$$X\left(\cdot\right) : \left[\underline{W}, \overline{W}\right] \to \left[\underline{K} - 1, \overline{K}\right]$$

is and increasing function of W.

For an *indexable* asset, the index in state x will be given by

$$W\left(x\right) = \inf\left\{W; X\left(W\right) \geq x\right\}. \tag{5.40}$$

Theorem 5.4. *If ϕ is increasing, the asset is indexable.*

Proof *Define*

$$V\left(W\right) = \max_{\underline{K}-1 \leq x \leq \overline{K}} \left\{ R\left(x+1\right)\left(1-\phi\left(x\right)\right) + W\phi\left(x\right) \right\} \tag{5.41}$$

From the discussion around (5.39), $X(W)$ *is the largest maximiser of the r.h.s. of* (5.41). *It is straightforward to show that, since* $0 \leq \phi(x) \leq 1$, *and* $\underline{K} - 1 \leq x \leq \overline{K}$, *then*

$$X(W) = \overline{K}, \text{ for } W \geq \overline{W} \tag{5.42}$$

where $\overline{W}$ *is the smallest value of* W *such that passive action is taken in all states* x*; and*

$$X(W) = \underline{K}, \text{ for } W < \underline{W} \tag{5.43}$$

for some $\underline{W}$ *small enough.*
Further, if the subset $\{x_1, x_1 + 1, \ldots, x_2\} \subseteq \left[\underline{K} - 1, \overline{K}\right]$ *is such that* $\phi(x_1) = \phi(x_1 + 1) = \cdots = \phi(x_2)$ *and is maximal in this regard, then it is straightforward to show that the range of* $X(W)$ *contains at most a single value from this subset.*

Hence, we need just to identify the conditions for $X(W)$ *to be an increasing function of* W. *First note in* (5.41) *that* $V : \mathbb{R}^+ \to \mathbb{R}^+$ *is an increasing function of* W. *Moreover, it is piecewise linear and convex, with each of its linear segment's right hand side gradient given by* $\phi(X(W))$.

Take now any pair $(W_1, W_2) \in \left[\underline{W}, \overline{W}\right]$ *such that* $W_1 > W_2$, *it immediately follows from* V*'s convexity, the hypothesis of the theorem and the foregoing discussion that*

$$X(W_1) > X(W_2). \tag{5.44}$$

The result now follows from (5.41) *to* (5.44) *and the discussion around* (5.39).

q.e.d. ■

In order to describe the asset's index structure under he hypothesis of Theorem 5.4 we develop a collection consisting of a positive integer L, as set of $L+1$ integers $\{x_l, 0 \leq l \leq L\}$ such that

$$\underline{K} - 1 = x_0 < x_1 < \cdots < x_L = \overline{K}$$

and accompanying set of L reals $\left\{W^l, 1 \leq l \leq L\right\}$ such that

$$-\infty < W^1 < \cdots < W^L < \infty$$

A discussion along the lines of the previous section yields that if ϕ is increasing, then

$$X\left(W\right) = \begin{cases} \underline{K} - 1, & W < W^1 \\ x_l, & W \in \left[W^l, W^{l+1}\right), 1 \leq l \leq L-1, \\ \overline{K}, & W \geq W^L \end{cases} \tag{5.45}$$

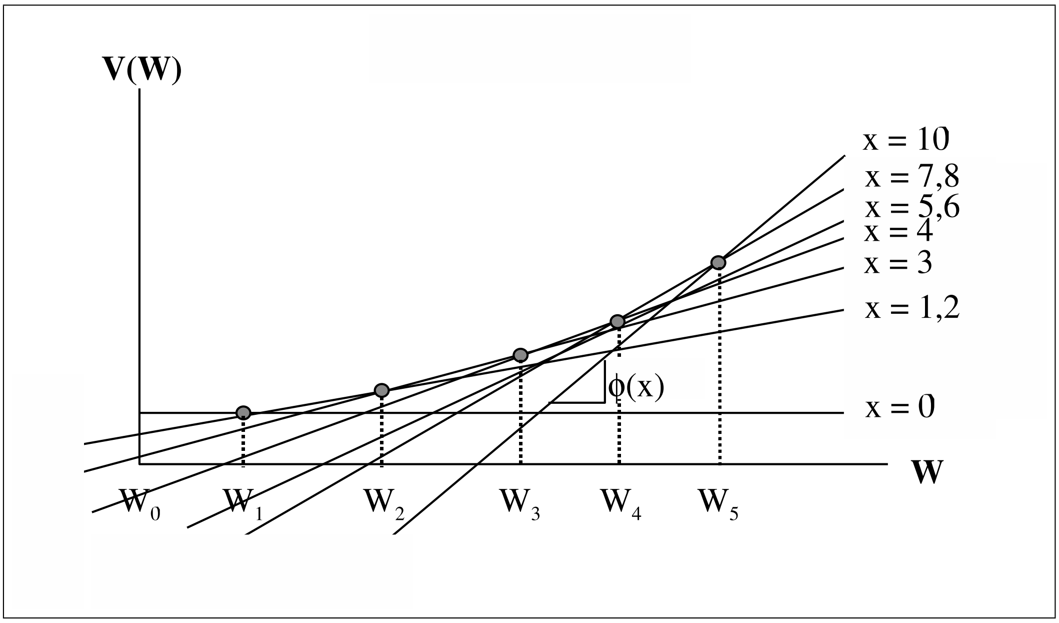
The index structure of the asset now follows from (5.40) and (5.45) and is described in Theorem 5.5

Theorem 5.5. *If ϕ is increasing then index $W : \left[\underline{K}, \overline{K}\right] \to \mathbb{R}$ is given by*

$$W\left(x\right) = W^l,\ x_{l-1} \leq x \leq x_l - 1,\ 1 \leq l \leq L. \tag{5.46}$$

The procedure of finding the Whittle indices for a *squads system* asset consists in two basic steps which involve, basically, identifying the points where the slope (or r.h.s. gradient) of the piecewise linear function $V\left(W\right)$ is discontinuous. This procedure and the working of the corresponding Algorithm in Figure 5.8 is illustrated with an example in Figure 5.7.

1. *Find the smallest value of W such that passive action is taken in some*



The general procedure for obtaining the Whittle indices in a typical squad system project is illustrated in the following example. Consider the interval $[W_2, W_3)$. it must be clear from the graph that $X^{W_2} = \{1,2,3\}$ and $X^W = \{2,3\}, W \in (W_2, W_3)$. According to expression (5.39), $X(W) = 3$, $W \in [W_2, W_3)$. Following the same reasoning we can construct the collection $\{x_i = X(W) : W_i \le W < W_{i+1}\}$, e.g. for the case discussed above we have $x_2 = 3$. Finally, in order to finding the $W(x)$'s recall the expression (5.40) and notice that, for example, $W(3) = \inf\{W_2, \ldots, W_5\} = W_2$. The same reasoning applies to every $W(x)$, $x \in [1, 10]$. In our particular example, it can be verified that $W(\underline{K}) = W(1) = W_1$, $W(2) = W_1$, $W(3) = W_2$, $W(4) = W_3$, $W(5) = W_3$, $W(6) = W_3$, $W(7) = W_4$, $W(8) = W_4$, $W(9) = W_4$, and $W(10) = W_5$.

Figure 5.7: Example of the Procedure for Finding the Whittle Indices in the Squad System Problem

state.

For each state we need to identify the value of W such that we are indifferent between applying policy π_x and being active in all states. In particular, as $\phi(\underline{K} - 1) = 0$ it holds that the average reward rate for policy $\pi_{\underline{K}-1}$ is simply $R(\underline{K} - 1)$; hence

$$R(0) = R(x)(1 - \phi(x)) + W\phi(x)$$

for every $\underline{K} \leq x \leq \overline{K}$. Or equivalently

$$\widetilde{W}(x) = \frac{R(0) - R(x+1)(1-\phi(x))}{\phi(x)}, \quad \underline{K} \leq x \leq \overline{K}.$$

Fix

$$\mathbf{x}_1 = \left\{ x : x \in \underset{\underline{K}-1 \leq x \leq \ldots, \overline{K}}{\operatorname{argmin}} \left\{ \widetilde{W}(x) \right\} \right\}$$

and let $x_1 = \sup\{x \in \mathbf{x}_1\}$. Fix $W^1 = \widetilde{W}(x_1)$ and $\Pi_1 = \underline{K}, \ldots, x_1$ and $X(W^1) = x_1$.

2. *Construct a collection of intercepts for every linear segment in V.*

   For every subsequent x_i let $\widetilde{W}(x)$ be the solution to

$$\begin{aligned} R(x_i+1)(1-\phi(x_i+1)) + W\phi(x_i+1) \\ = R(x+1)(1-\phi(x+1)) + W\phi(x+1) \end{aligned}$$

   for all $x_i + 1 \leq x \leq \overline{K}$. Then set

$$\mathbf{x}_{i+1} = \left\{ x : x \in \underset{x_i+1 \leq x \leq \overline{K}}{\operatorname{argmin}} \left\{ \tilde{W}(x) \right\} \right\} \tag{5.47}$$

   and $x_{i+1} = \sup\{x \in \mathbf{x}_{i+1}\}$. Fix $W_{i+1} = \widetilde{W}(x_{i+1})$ and $\Pi_{i+1} = \pi_i \cup \{x_i + 1, \ldots, x_{i+1}\}$.

By a modest extension of the above calculations, it follows that the minimisers of the r.h.s. of (5.47) are precisely the maximisers of $V(W^{i+1})$. By definition, $X(W^{i+1})$ is the largest of them. The indices for all other states x are obtained according to Theorem 5.5.

All the procedure above can be summarized in the Algorithm in Figure 5.8.

```
**INPUT:** $\overline{K}, \underline{K}, R(x), \phi(x)$
**INITIALIZATION:**
Set $x_0 = \underline{K} - 1$;
    $\mathbf{x}_1 = \left\{ x : x \in \underset{\underline{K} \leq x \leq \overline{K}}{\operatorname{argmin}} \left\{ \widetilde{W}(x) = \frac{R(0) - R(x+1)(1-\phi(x+1))}{\phi(x+1)} \right\} \right\}$;
    $x_1 = \sup\{x \in \mathbf{x}_1\}$;
    $W_1 = \widetilde{W}(x_1)$;
    $X(W^1) = x_1$;
    $\Pi_1 = \{\underline{K}, \ldots, x_1\}$;
    $i = 1$;
**PROCEDURE:**
**While** $\Pi_i \neq \{\underline{K}, \ldots, \overline{K}\}$ **do**
Set $\mathbf{x}_{i+1} = \left\{ x : x \in \underset{x_i+1 \leq x \leq \overline{K}}{\operatorname{argmin}} \left\{ \widetilde{W}(x) = \frac{R(x_i+1)(1-\phi(x_i+1)) - R(x+1)(1-\phi(x+1))}{\phi(x+1) - \phi(x_i+1)} \right\} \right\}$;
    $x_{i+1} = \sup\{x \in \mathbf{x}_{i+1}\}$;
    $W_{i+1} = \widetilde{W}(x_{i+1})$;
    $X(W^{i+1}) = x_{i+1}$;
    $\Pi_{i+1} = \Pi_i \cup \{x_i + 1, \ldots, x_{i+1}\}$;
    $i = i + 1$.
**End While**
Set $L = i$
**OUTPUT:** $\mathbf{W}$, $\mathbf{X}$, $L$.
```

Figure 5.8: Adaptive Greedy Algorithm for the Squad System Problem

Comment

Please note from Theorem 5.5 that the index is increasing in state. Hence in the squad system it is assets which are achieving high rewards which are high priority for activation. Note also that, unlike the spinning plates problem, indices can now be negative. This raises the question of whether idling may be preferable to asset deployment.

As before, important special cases occur in which all states have distinct indices. The following section gives a necessary and sufficient condition for *strict indexability* for the *squad system*.

5.2.2 Strict Indexability

Given the structure of the squad system's problem, under *general indexability* (or simply *indexability*), one single value of the Whittle index, $W(x)$, can be attached to two or more adjacent states. In this section we introduce the notion of *strict indexability* for the *squad system* (analogous to the *spinning plates* case) which requires that one and only one Whittle Index, will correspond to each individual state, and explore the corresponding *strict indexability* conditions.

We already have a *sufficient* condition for indexability; i.e.

$$\phi(x) = \frac{\mu(x+1)}{\lambda(x) + \mu(x+1)}, \ \underline{K} - 1 \leq x \leq \overline{K}$$

must be increasing in x. Hence now we need to identify an additional (necessary) condition for strict indexability. Start considering any state x.

Following the reasoning in Section 5.1.2, we recover the following expressions:

$$X(W) = \sup\left\{x : x \in x^W\right\}$$

and

$$W(x) = \inf\left\{W; X(W) \geq x\right\}$$

which imply that the relevant values of W are the ones corresponding to intersections of linear segments in (5.41). Under strict indexability, each segment in V should correspond to one and only one state x. For any pair of adjacent states x and $x-1$, the value of W at the intersection will be given by the W-solution to

$$R(x+1)(1-\phi(x))+W\phi(x)=R(x)(1-\phi(x-1))+W\phi(x-1)$$

Alternatively, there exists some $\widehat{W}$ such that the reward rates, at state x, under the active and passive actions are identical, i.e. the operator is indifferent between activity and passivity. By standard index theory, this value will be the index corresponding to state x, i.e.

$$\widehat{W}(x)=\frac{R(x+1)(1-\phi(x))-R(x)(1-\phi(x-1))}{\phi(x-1)-\phi(x)} \tag{5.48}$$

which is precisely the intersection between adjacent segments of V corresponding to x and $x-1$. If the expression above is strictly increasing or decreasing in x, then to every pair $x, x-1$ will correspond a different value of $\widehat{W}(x)$.

The next theorem follows from the discussion above:

Theorem 5.6. ***Strict Indexability***

1. *The following are equivalent:*

   (a) *The asset is strictly indexable*

   (b) *Both* $\phi(x)$ *and* $\widehat{W}(x)$ *as given by* (5.48) *are strictly increasing over* $\underline{K}-1\leq x\leq\overline{K}$.

2. *The conditions in 1b imply*

$$W(x)=\widehat{W}(x)\,,\ \underline{K}-1\leq x\leq\overline{K}.$$

and the index is strictly increasing in the state.

Proof *We start our proof by showing that* (1b) *implies* (1a).
Note that condition (1b) *guarantees that the piecewise value function V will have one segment for each state $x \in \left[\underline{K}-1, \overline{K}\right]$ and the minimum in the r.h.s. of* (5.47) *will be achieved uniquely by $x = x-1$. We will then have $W = \widehat{W}(x)$, and the inference of strict indexability will follow simply from Theorem 5.5.*

To prove that (1a) *implies* (1b) *notice that if the asset is strictly indexable there must exist some $\overline{W}(x)$, $\underline{K} \le x \le \overline{K}$, strictly increasing in x, such that*

$$X(W) : \begin{cases} \underline{K}-1, & W < \overline{W}(\underline{K}-1) \\ x, & W \in \left[\overline{W}(x-1), \overline{W}(x)\right) \\ \overline{K}, & W \ge \overline{W}\left(\overline{K}\right) \end{cases} \tag{5.49}$$

Plainly, when $W = \overline{W}(x)$, $\underline{K} \le x \le \overline{K}$, both x and $x-1$ achieve the maximum in $V(W)$. It follows that

$$R(x+1)(1-\phi(x)) + \overline{W}(x)\phi(x) = R(x)(1-\phi(x-1)) + \overline{W}(x)\phi(x-1). \tag{5.50}$$

Moreover, if $W \in \left[\overline{W}(x-1), \overline{W}(x)\right)$, $X(W) = x$ and so π_x must strictly outperform π_{x-1} in this range. Hence, for $W \in \left[\overline{W}(x-1), \overline{W}(x)\right)$,

$$R(x+1)(1-\phi(x)) + W\phi(x) > R(x)(1-\phi(x-1)) + W\phi(x-1). \tag{5.51}$$

It must then follow from (5.50) *and* (5.51) *that $\phi(x) > \phi(x-1)$, $\underline{K} \le x \le \overline{K}$, and hence that $\phi(x)$ is strictly decreasing over $\underline{K}-1 \le x \le \overline{K}$. Further, solving* (5.51) *for $\overline{W}(x)$ we obtain:*

$$\overline{W}(x) = \frac{R(x+1)(1-\phi(x)) - R(x)(1-\phi(x-1))}{\phi(x-1)-\phi(x)},$$

$$= \widehat{W}(x), \ \underline{K} \leq x \leq \overline{K}.$$

We conclude that $\widehat{W}(x)$ *is strictly increasing in* x *over* $\underline{K} \leq x \leq \overline{K}$. *This concludes the proof of part 1. Part 2 follows trivially from the above analysis.* **q.e.d.** ■

5.2.3 Examples

Example 5.3. *Suppose that the reward is linear in the state and hence that*

$$R(x) = r(x - \underline{K}), \ \underline{K} \leq x \leq \overline{K}$$

for some $r > 0$ *and, moreover, that the transition rates are also linear, namely:*

$$\mu(x) = \overline{\mu}(x - \underline{K}), \ \underline{K} \leq x \leq \overline{K} \tag{5.52}$$

and

$$\lambda(x) = \overline{\lambda}\left(\overline{K} - x\right) \ \underline{K} \leq x \leq \overline{K} \tag{5.53}$$

where $\overline{\mu}$ *and* $\overline{\lambda}$ *are all positive constants. It follows trivially from* (5.52) *and* (5.53) *that* $\phi(\cdot)$ *is strictly increasing:*

$$\phi(x) = \frac{\mu(x+1)}{\mu(x+1) + \lambda(x)} < \frac{\mu(x+2)}{\mu(x+2) + \lambda(x+1)} = \phi(x+1)$$

i.e.

$$(x + 1 - \underline{K})\left(\overline{K} - x + 1\right) < (x + 2 - \underline{K})\left(\overline{K} - x\right).$$

By direct computation we have from (5.48)*:*

$$\widehat{W}(x) = r\left\{\frac{(x-\underline{K})(x+1-\underline{K}) - \frac{\overline{\lambda}}{\overline{\mu}}\left(\overline{K}-x\right)\left(\overline{K}-x+1\right)}{\overline{K}+1-\underline{K}}\right\}. \tag{5.54}$$

which is strictly increasing. From Theorem 5.6, (5.54) *give the index in this case.*

This is the example referred to by Whittle [116] as the Ehrenfest project. Whittle used a heuristic argument to develop the index

$$\nu(x) = \frac{c}{\mu K}\left(\mu x^2 - \lambda (K-x)^2\right)$$

or equivalently, using our notation and fixing $\underline{K} = 0$, *and* $\overline{K} = K$,

$$\nu(x) = r\left\{\frac{x^2 - \frac{\overline{\lambda}}{\overline{\mu}}(K-x)^2}{K}\right\} \tag{5.55}$$

which clearly approximates $\widehat{W}(x)$ *in* (5.54)*.*

Example 5.4. *As well as in the spining plates proble, it is possible to develop indexable assets with semi-infinite state spaces of the form* $[\underline{K},\infty)$*. Consider now such an example for which* $\underline{K} = 0$ *and assume that reward rates are given by*

$$R(x) = r\left(1-(x+1)^{-\alpha}\right),\ x \geq 0 \tag{5.56}$$

and also

$$\phi(x) = 1-(x+1)^{-\beta},\ x \geq 0 \tag{5.57}$$

with r, α and β, positive constants. By direct substitution into (5.48) *we get*

$$\begin{aligned}\widehat{W}(x) =& r\left\{\frac{(x+1)^{-\beta}-(x+2)^{-\beta}-(x+1)^{-(\alpha+\beta)}+(x+2)^{-(\alpha+\beta)}}{(x+1)^{-\beta}-(x+2)^{-\beta}}\right\}\\ =& r\left\{1-\frac{(x+1)^{-(\alpha+\beta)}-(x+2)^{-(\alpha+\beta)}}{(x+1)^{-\beta}-(x+2)^{-\beta}}\right\},\ x\geq 0 \qquad (5.58)\end{aligned}$$

which is strictly increasing. Hence, for a suitable extension of Theorem (5.6), (5.58) *gives the index in this case.*

5.2.4 Numerical Results

In Tables 5.3 and 5.4 we offer some results of an extensive numerical investigation into the quality of performance of the index heuristics developed in this section. Each problem studied has $M = 4$, $Q = 1$; namely a choice has to be made at each decision epoch of one out of four possible assets for activation. Four policies were applied to each problem generated. These are as follows:

- **OPT** Corresponding to the ϵ-optimal solution to the infinite-horizon time average problem, computed by value iteration. At each decision epoch the optimal ϵ-policy activates the arm with largest average reward rate. As in the *spinning plates* case, references are Puterman [85] and Tijms [99] and the discussion in Sections 2.1 and, in particular, 2.1.4.

- **IND** The index policy as described in this section. At each decision epoch, the index policy activates the arm with currently mazimal index.

- **MYO** The myopic policy prescribes taking the active action in the arm with largest immediate reward.

- **LAR** This policy activates the project with largest state.

For each problem generated, the average reward rates r^{ind}, r^{myo}, and r^{lar}, were computed by DP value iteration as described in Section 2.1.4, yielding the *percentage suboptimalities*:

$$\Xi\left(\mathrm{P}\right) = \frac{r^{\text{opt}} - r^{\mathrm{P}}}{r^{\text{opt}}} \times 100, \quad \mathrm{P} = \text{ind}, \text{myo}, \text{lar}. \tag{5.59}$$

Two families of problems, corresponding to Examples 5.3 and 5.4 are discussed.

Example 5.3. *Table 5.3 shows some results derived from an extensive numerically-based assessment of the quality of performance of the index heuristic for the linear version of the Squad System model, see Example 5.3 in section 5.2.3 for details.*

For the cases reported in the upper part (Cases 1 to 4) of the table we offer the summary of 400 experiments (100 each) in a model with $M = 4$ and $Q = 1$. In the lower part of the table (Cases 1bis to 4bis) we consider the same family of problems but embellished by the inclusion of an idling option, to be thought as a zero reward asset, whose state space is a singleton. Such an asset is trivially indexable with index always zero. In this case, whenever the indices for all arms are simultaneously zero, the index policy will choose the idle option.

Each constituent asset is structured as in 5.3 in section 5.2.3, with $\underline{K} = 0$ and $\overline{K} = 8$. Four cases where analysed for different combinations of parameters r, $\overline{\mu}$, and $\overline{\lambda}$. For all of them, the r's were chosen by sampling from a uniform distribution $U\left(10, 25\right)$. Parameters $\overline{\mu}$ and $\overline{\lambda}$ were also chosen by sampling from uniform distributions, as indicated in Table 5.3.

Table 5.3 presents results for 800 (= 4 × 2 × 100) randomly generated problems. The Table summarises the collections of percentage suboptimalities arising from the application of each of IND, MYO and LAR policies to each of the eight problem configurations. Each collection is summarised by the order statistics MIN (minimum), LQ (lower quartile), MED (median), UQ (upper quartile), and MAX (maximum).

Notice that the idling option will be taken by IND only when all four of the conventional assets have negative indices. Also, MYO and LAR will never choose the idling option.

As well as in the *spinning plates* problem, the index policy continues to performs strongly, with a worst case of 1.3246% suboptimality among the 800 problems generated. There is evidence of enhanced performance following the inclusion of the idling option, where the worst case performance reduces to 1.2468% suboptimality. For each of the other heuristics (MYO and LAR) problem instance arose in which they performed poorly.

Example 5.4. *For the nonlinear case described in Example 5.4 in section 5.2.3, each constituent asset is structured with* $\underline{K} = 0$ *and* $\overline{K} = 8$*, and the function*

$$\lambda\left(x\right) = \mu\left(x+1\right)\left[\left(x+1\right)^{-\beta} - 1\right]^{-1},$$

arising from the λ *solution of the expression*

$$\phi\left(x\right) = \frac{\mu\left(x+1\right)}{\mu\left(x+1\right) + \lambda\left(x\right)} = 1 - \left(x+1\right)^{-\beta},$$

with

$$\mu\left(x\right) = I\left(x \geq 1\right)$$

where $I\left(\cdot\right)$ *is the indicator function. Four cases where analysed for different combinations of parameters* r *and* β*. For all of them, the* r*'s where choosen by sampling from a uniform distribution* $U\left(2,4\right)$*.*

Table 5.4 shows some results derived from an extensive numerically based assessment of the quality of performance of the index heuristic for the non-linear version of the *Squad System* model (Example 5.4).

As in the linear case, in the upper part of the table (Cases 1 to 4) we offer the summary of 400 experiments in a model with $M = 4$ and $Q = 1$. In the lower part of the table (Cases 1bis to 4bis) we consider the same family of problems but by including the *idling option*, i.e. we add a zero reward asset, which index is trivially zero. In this case, whenever the indices for all arms are simultaneously less or equal to zero, the index policy will choose the idle option. Again, policies MYO and LAR will never chose the idling option.

The performance of the same four different policies described in Example 5.3 is compared (OPT, IND, MYO and LAR).

As in the previous example, the index policy continues to perform strongly, with a worst case of 1.9165% suboptimality among the 800 problems generated. There is also evidence of enhanced performance following the inclusion of the idling option. For each of the orther heuristics (Myopic and Large) problem instance arose in which they perform poorly.

Table 5.3: Bi-directional Restless Bandits. Squad System Problem. Performance of the Index Policy in the Linear Case, Example 5.3.

	Case 1			Case 2			Case 3		
	$\overline{\mu} \sim U(4,6)$, $\overline{\lambda} \sim U(4,6)$			$\overline{\mu} \sim U(2,6)$, $\overline{\lambda} \sim U(4,8)$			$\overline{\mu} \sim U(4,8)$, $\overline{\lambda} \sim U(2,6)$		
	Index	Myopic	Large	Index	Myopic	Large	Index	Myopic	Large
Min	0.0033	0.011	0.004	0.0019	0.003	0.009	0.0009	0.072	0.026
LQ	0.0786	1.947	0.916	0.1350	1.856	1.371	0.0862	1.635	0.645
Med	0.1582	3.231	1.978	0.3051	3.180	2.924	0.1425	2.917	1.304
UQ	0.2513	5.509	3.349	0.5278	5.174	5.343	0.2098	4.128	2.126
Max	0.5928	10.751	6.279	1.3246	11.296	9.290	0.5214	9.044	4.097
	Case 1 bis			Case 2 bis			Case 3 bis		
	Index	Myopic	Large	Index	Myopic	Large	Index	Myopic	Large
Min	0.0000	0.046	0.064	0.0007	0.138	0.040	0.0006	0.242	0.223
LQ	0.0573	1.951	0.979	0.1692	1.884	1.464	0.0503	1.610	0.719
Med	0.1153	3.328	2.006	0.2863	3.091	2.801	0.0986	2.535	1.330
UQ	0.2268	4.563	3.176	0.4871	5.263	5.051	0.1731	3.373	2.067
Max	0.5582	10.194	5.638	1.2468	12.759	10.752	0.3851	7.864	4.751

	Case 4		
	$\overline{\mu} \sim U(2,8)$, $\overline{\lambda} \sim U(2,8)$		
	Index	Myopic	Large
Min	0.0305	0.057	0.012
LQ	0.1247	1.806	0.921
Med	0.2353	3.287	1.852
UQ	0.4167	5.125	3.493
Max	0.7818	10.349	6.699
	Case 4 bis		
	Index	Myopic	Large
Min	0.0123	0.108	0.029
LQ	0.1021	1.832	0.969
Med	0.2247	3.320	1.878
UQ	0.4030	5.150	3.518
Max	0.7697	10.368	6.717

Table 5.4: Bi-directional Restless Bandits. Squad System Problem. Performance of the Index Policy in the Non-linear Case, Example 5.4.

	Case 1			Case 2			Case 3		
	$\beta \sim U\,(0.5, 1)$			$\beta \sim U\,(1, 2)$			$\beta \sim U\,(2, 3)$		
	Index	Myopic	Large	Index	Myopic	Large	Index	Myopic	Large
Min	0.0343	0.180	0.015	0.0610	0.128	0.522	0.1355	0.141	0.639
LQ	0.1487	1.924	0.727	0.2074	0.440	1.618	0.4160	0.720	2.353
Med	0.3336	2.640	1.469	0.3321	0.735	2.670	0.5377	3.099	4.491
UQ	0.5494	3.661	2.397	0.5200	1.349	3.735	0.6337	7.040	7.708
Max	1.9165	7.671	6.223	1.2287	3.496	8.592	0.8905	14.181	14.739
	Case 1 bis			Case 2 bis			Case 3 bis		
	Index	Myopic	Large	Index	Myopic	Large	Index	Myopic	Large
Min	0.0423	0.180	0.112	0.0239	0.408	0.754	0.0679	4.338	5.814
LQ	0.1504	2.028	0.728	0.1746	0.855	2.110	0.3269	8.091	9.498
Med	0.3159	2.679	1.417	0.2719	1.381	3.301	0.4182	10.099	10.928
UQ	0.5143	3.806	2.258	0.3888	1.976	4.897	0.5156	12.752	13.514
Max	1.8982	7.671	6.231	0.8430	4.522	8.865	0.8521	18.632	19.230

	Case 4		
	$\beta \sim U\,(0.5, 3)$		
	Index	Myopic	Large
Min	0.0710	0.269	0.690
LQ	0.2350	1.090	2.165
Med	0.4204	2.297	3.731
UQ	0.6469	3.971	6.513
Max	1.3260	14.393	17.029
	Case 4 bis		
	Index	Myopic	Large
Min	0.0353	0.512	0.474
LQ	0.1966	2.324	3.062
Med	0.3329	3.846	4.639
UQ	0.5120	5.380	6.837
Max	0.9997	17.164	17.710

5.3 Discounted Reward Version

This section contains a somehow briefer discussion of the index structure of a version of the spinning plates (section 5.3.1) and the *squad system* (section 5.3.2) with the discounted reward criterion.

5.3.1 Spinning Plates Problem

We have already defined $\mathfrak{M}$ as the class of monotone (stationary, Markovian deterministic) policies such that

$$\pi_y(x) = 1 \ \Leftrightarrow x \leq y,\ y \in \left[\underline{K} - 1, \overline{K}\right],$$

hence, policy π_y, $\underline{K} \leq y < \overline{K}$ prescribes taking the active action in the subset $[\underline{K}, y]$ and the passive one in $\left[y+1, \overline{K}\right]$. Policy $\pi_{\underline{K}-1}$ chooses the passive action ($a = 0$) in all states, and policy $\pi_{\overline{K}}$ prescribes taking the active action in all states.

Consider policy $\pi_y \in \mathfrak{M}$ and assume that $X(0) = \hat{x} < y$. Then, under π_y, in the discounted reward version of the problem, the system will take the active action in every state $\hat{x} \leq x \leq y$ evolving to state $x+1$ with rate $\mu(x)$ and earning an instant discounted reward $e^{-\beta t} R(x)$, $t \in \mathbb{R}^+$, and $\beta > 0$ during the duration of its sojourn in x. After the first arrival to state $y+1$, passive action will be taken and the system will evolve to state y with rate $\lambda(y+1)$ and instant discounted reward given by $e^{-\beta t}(R(y+1))$. Thereafter, the system will face alternating sojourns in states y and $y+1$.

If otherwise, $X(0) = \hat{x} > y$. Then, under π_y, the system will take the passive action in every state $y \leq x \leq \hat{x}$ evolving to state $x-1$ with rate $\lambda(x)$, earning an instant discounted reward $e^{-\beta t} R(x)$ during the duration of its sojourn in x, until arriving to state y. Thereafter, the system will face alternating sojourns in states y and $y+1$.

Lemma 5.3. ***Optimality of Monotone Policies***

For all $W \in \mathbb{R}$ and every $X(0)$, there exists an optimal policy for the W-subsidy problem in $\mathfrak{M}$.

Proof *Consider $X(0)=\hat{x}$ and take (optimal) policy $\pi \in \mathcal{M}$, with $\mathcal{M}$ being the family of stationary, deterministic, Markovian policies. Assume $\pi(\hat{x})=1$ and define*

$$\tilde{x} = \min\left\{x | x > \hat{x}, \pi(x+1) = 0\right\}.$$

Call $V_\beta(\hat{x}, W)$ the infinite horizon discounted reward earned under (optimal) policy π when $X(0)=\hat{x}$. It is straightforward to see that this is also the discounted reward earned by monotone policy $\pi_{\tilde{x}}$ when starting from $X(0)=\hat{x}$, i.e. $V_\beta^{\tilde{x}}(\hat{x}, W)$.

Consider again $X(0)=\hat{x}$ and assume instead that $\pi(\hat{x})=0$, then we can define

$$\tilde{x} = \max\left\{x | x < \hat{x}, \pi(x) = 1\right\}.$$

Clearly $V_\beta(\hat{x}, W) = V_\beta^{\tilde{x}}(\hat{x}, W)$. ***q.e.d.*** ■

We now present some results that will be useful for showing that, under certain conditions, the optimal monotone policy is independent of the initial state and, consequently, an optimal monotone policy is indeed an optimal policy for the discounted reward version of the spinning plates problem.

We start by identifying the values $V_\beta(X(0), W)$ and $V_\beta^y(X(0), W)$.

Consider the stationary, deterministic, Markovian policy ($\pi \in \mathcal{M}$). We have already defined $V_\beta(\hat{x}, W)$ to be the total expected discounted reward earned over an infinite horizon when an optimal policy is applied and the initial state is $X(0)=\hat{x}$. The dynamic programming equations for this policy are then given by:

$$V_\beta\left(\underline{K}, W\right) = \max\left\{\frac{R\left(\underline{K}\right) + \mu\left(\underline{K}\right) V_\beta\left(\underline{K}+1, W\right)}{\beta + \mu\left(\underline{K}\right)}; \frac{W + R\left(\underline{K}\right)}{\beta}\right\};$$

$$V_\beta\left(\hat{x}, W\right) = \max\left\{\frac{R\left(\hat{x}\right) + \mu\left(\hat{x}\right) V_\beta\left(\hat{x}+1, W\right)}{\beta + \mu\left(\hat{x}\right)}\right. ;$$

$$\left.\frac{W + R\left(\hat{x}\right) + \lambda\left(\hat{x}\right) V_\beta\left(\hat{x}-1, W\right)}{\beta + \lambda\left(\hat{x}\right)}\right\} \tag{5.60}$$

for $x \in \left[\underline{K}+1, \overline{K}-1\right]$; and

$$V_\beta\left(\overline{K}, W\right) = \max\left\{\frac{R\left(\overline{K}\right)}{\beta}; \frac{W + R\left(\overline{K}\right) + \lambda\left(\overline{K}\right) V_\beta\left(\overline{K}-1, W\right)}{\beta + \lambda\left(\overline{K}\right)}\right\}.$$

Throughout (5.60) the first quantity on the r.h.s. is the total expected reward earned when choosing the *active* action ($a = 1$) in the current state and thereafter proceeding optimally. The second quantity is the total expected reward earned when choosing *passive* action ($a = 0$) in the current state and then proceeding optimally.

Consider now, monotone policy $\pi_y \in \mathfrak{M}$. We defined $V_\beta^y\left(\hat{x}, W\right)$ as the value function of the monotone policy π_y when applied at $X\left(0\right) = \hat{x} \in \left[\underline{K}, \overline{K}\right]$.

We first focus on the case $X\left(0\right) = y$. Here, under policy π_y the project is active in state y and shows alternate sojourns between y and $y+1$. It can easily be confirmed that the expected value of π_y, under passive subsidy W and discount factor β is be given by:

$$V_\beta^y\left(y, W\right) = \frac{R\left(y\right)\left(\beta + \lambda\left(y+1\right)\right) + \mu\left(y\right)\left(R\left(y+1\right) + W\right)}{\beta\left(\beta + \mu\left(y\right) + \lambda\left(y+1\right)\right)} \tag{5.61}$$

Likewise, if we slightly modify the policy above and allow the system for taking the passive action in y, then the expected value will correspond to the one of policy π_{y-1}, under the same initial state y and will be given by:

$$V_{\beta}^{y-1}(y,W)=\frac{R(y-1)\lambda(y)+(W+R(y))(\beta+\mu(y-1))}{\beta(\beta+\mu(y-1)+\lambda(y))}. \tag{5.62}$$

We will use this quantity later.

Now, for any initial state $X(0)=\hat{x}$ we have three cases:

1. $\underline{K}\leq\hat{x}<y$

   Here, the expected value of policy π_y is given by

$$\frac{R(\hat{x})}{\beta+\mu(\hat{x})}+\frac{\mu(\hat{x})}{\beta+\mu(\hat{x})}\frac{R(\hat{x}+1)}{\beta+\mu(\hat{x}+1)}+\frac{\mu(\hat{x})}{\beta+\mu(\hat{x})}\frac{\mu(\hat{x}+1)}{\beta+\mu(\hat{x}+1)}\frac{R(\hat{x}+2)}{\beta+\mu(\hat{x}+2)}+\cdots$$

$$+\frac{\mu(\hat{x})}{\beta+\mu(\hat{x})}\cdots\frac{\mu(y-1)}{\beta+\mu(y-1)}\left[\frac{R(y)(\beta+\lambda(y+1))+(W+R(y+1))\mu(y)}{\beta(\beta+\mu(y)+\lambda(y+1))}\right]$$

   or equivalently

$$V_{\beta}^{y}(\hat{x},W)=\sum_{i=\hat{x}}^{y-1}\frac{R(i)}{\beta+\mu(i)}\prod_{j=\hat{x}}^{i-1}\frac{\mu(j)}{\beta+\mu(j)}+V_{\beta}^{y}(y,W)\prod_{i=\hat{x}}^{y-1}\frac{\mu(i)}{\beta+\mu(i)};\ \underline{K}\leq\hat{x}\leq y \tag{5.63}$$

2. $y<\hat{x}\leq\overline{K}$

   The expected value of policy π_y is given by

$$\frac{W+R(\hat{x})}{\beta+\lambda(\hat{x})}+\frac{\lambda(\hat{x})}{\beta+\lambda(\hat{x})}\frac{W+R(\hat{x}-1)}{\beta+\lambda(\hat{x}-1)}+\frac{\lambda(\hat{x})}{\beta+\lambda(\hat{x})}\frac{\lambda(\hat{x}-1)}{\beta+\lambda(\hat{x}-1)}\frac{W+R(\hat{x}-2)}{\beta+\lambda(\hat{x}-2)}+\cdots$$

$$+\frac{\lambda(\hat{x})}{\beta+\lambda(\hat{x})}\cdots\frac{\lambda(y+1)}{\beta+\lambda(y+1)}\left[\frac{R(y)(\beta+\lambda(y+1))+(W+R(y+1))\mu(y)}{\beta(\beta+\mu(y)+\lambda(y+1))}\right]$$

or equivalently

$$V_\beta^y(\hat{x},W)=\sum_{i=y+1}^{\hat{x}}\frac{W+R(i)}{\beta+\lambda(i)}\prod_{j=i+1}^{\hat{x}}\frac{\lambda(j)}{\beta+\lambda(j)}+V_\beta^y(y,W)\prod_{i=y+1}^{\hat{x}}\frac{\lambda(i)}{\beta+\lambda(i)}; \tag{5.64}$$

for $y<\hat{x}\leq\overline{K}$.

3. $y=\underline{K}-1$ and $\underline{K}\leq\hat{x}\leq\overline{K}$

Following the same reasoning as in the two previous cases, it can be seen that

$$\begin{aligned}V_\beta^{\underline{K}-1}(\hat{x},W)=\sum_{i=\underline{K}}^{\hat{x}}\frac{W+R(i)}{\beta+\lambda(i)}\prod_{j=i+1}^{\hat{x}}\frac{\lambda(j)}{\beta+\lambda(j)}\cdots\\+\frac{W+R(\underline{K})}{\beta}\prod_{i=\underline{K}}^{\hat{x}}\frac{\lambda(i)}{\beta+\lambda(i)};\qquad\underline{K}\leq\hat{x}\leq\overline{K}\end{aligned} \tag{5.65}$$

Define now, $\overline{W}(y,\beta)$ as the value of W such that we are indifferent between taking the active and passive actions in state y. From (5.61) and (5.62) we have that $\overline{W}(y,\beta)$ is the W solution to $V_\beta^y(y,W)=V_\beta^{y-1}(y,W)$, i.e.

$$\frac{R(y)(\beta+\lambda(y+1))+\mu(y)(R(y+1)+W)}{\beta(\beta+\mu(y)+\lambda(y+1))}$$

$$=\frac{R(y-1)\lambda(y)+(W+R(y))(\beta+\mu(y-1))}{\beta(\beta+\mu(y-1)+\lambda(y))}$$

i.e.

$$\begin{aligned}\overline{W}\left(y,\beta\right) = \Big[& R\left(y+1\right)\mu\left(y\right)\left(\beta+\mu\left(y-1\right)+\lambda\left(y\right)\right) \\ & + R\left(y\right)\big(\lambda\left(y\right)\left(\beta+\lambda\left(y+1\right)\right)-\mu\left(y\right)\left(\beta+\mu\left(y-1\right)\right)\big) \\ & \quad -R\left(y-1\right)\lambda\left(y\right)\left(\beta+\mu\left(y\right)+\lambda\left(y+1\right)\right)\Big] \\ & \quad\quad \times\Big[\left(\beta+\mu\left(y-1\right)\right)\left(\beta+\lambda\left(y+1\right)\right)-\mu\left(y\right)\lambda\left(y\right)\Big]^{-1} \end{aligned} \tag{5.66}$$

As it will be seen, under indexability $\overline{W}\left(x,\beta\right)$ will turn out to be the Whittle index for state x. As we will be interested in a specific ordering for the indices, we need here to impose the following condition on the denominator of (5.66):

$$\left(\beta+\mu\left(y-1\right)\right)\left(\beta+\lambda\left(y+1\right)\right)-\mu\left(y\right)\lambda\left(y\right)>0, \qquad \text{for all } \ \underline{K}\leq y\overline{K}.$$

We now present some results that will be useful for the indexability discussion below.

Claim 5.1. *For $W=W\left(x,\beta\right)$, the expected discounted reward from any initial state $X\left(0\right)=\hat{x}$ is the same under both π_x and π_{x-1}.*

Proof *It is easy to verify that for $x>\hat{x}$*

$$\begin{aligned} V_\beta^{x-1}\left(\hat{x},W\right) &= \sum_{i=\hat{x}}^{x-2}\frac{R\left(i\right)}{\beta+\mu\left(i\right)}\prod_{j=\hat{x}}^{i-1}\frac{\mu\left(j\right)}{\beta+\mu\left(j\right)}+V_\beta^{x-1}\left(x-1,W\right)\prod_{i=\hat{x}}^{x-2}\frac{\mu\left(i\right)}{\beta+\mu\left(i\right)} \\ &= \sum_{i=\hat{x}}^{x-1}\frac{R\left(i\right)}{\beta+\mu\left(i\right)}\prod_{j=\hat{x}}^{i-1}\frac{\mu\left(j\right)}{\beta+\mu\left(j\right)}+V_\beta^{x-1}\left(x,W\right)\prod_{i=\hat{x}}^{x-1}\frac{\mu\left(i\right)}{\beta+\mu\left(i\right)}; \end{aligned} \tag{5.67}$$

similarly, for $x<\hat{x}$

$$V_\beta^{x-1}\left(\hat{x},W\right)=\sum_{i=x}^{\hat{x}}\frac{W+R\left(i\right)}{\beta+\lambda\left(i\right)}\prod_{j=i+1}^{\hat{x}}\frac{\lambda\left(j\right)}{\beta+\lambda\left(j\right)}+V_\beta^{x-1}\left(x-1,W\right)\prod_{i=x}^{\hat{x}}\frac{\lambda\left(i\right)}{\beta+\lambda\left(i\right)}$$

$$= \sum_{i=x+1}^{\hat{x}} \frac{W+R(i)}{\beta+\lambda(i)} \prod_{j=i+1}^{\hat{x}} \frac{\lambda(j)}{\beta+\lambda(j)} + V_{\beta}^{x-1}(x,W) \prod_{i=x}^{\hat{x}} \frac{\lambda(i)}{\beta+\lambda(i)} \tag{5.68}$$

as $W = W(x,\beta)$ *it holds that* $V_{\beta}^{x}(x,W) = V_{\beta}^{x-1}(x,W)$ *and consequently, from* (5.63) *and* (5.64), *the expressions above imply that*

$$V_{\beta}^{x}(\hat{x},W) = V_{\beta}^{x-1}(\hat{x},W), \ \underline{K} \leq x \leq \overline{K}$$

. ***q.e.d.*** ■

Claim 5.2. *For* $W > W(x,\beta)$ *it holds* $V_{\beta}^{x}(\hat{x},W) < V_{\beta}^{x-1}(\hat{x},W)$ *for any* $\underline{K} \leq \hat{x} \leq \overline{K}$ *and* $\underline{K} \leq x \leq \overline{K}$.

Proof *For any* $W > W(x,\beta)$ *by straightforward algebraic manipulations it is easy to verify that* $V_{\beta}^{x}(x,W) < V_{\beta}^{x-1}(x,W)$ *and consequently, from* (5.67) *and* (5.68), *the result above follows immediately.* ***q.e.d.*** ■

Claim 5.3. *For* $W < W(x,\beta)$ *it holds* $V_{\beta}^{x}(\hat{x},W) > V_{\beta}^{x-1}(\hat{x},W)$ *for any* $\underline{K} \leq \hat{x} \leq \overline{K}$ *and* $\underline{K} \leq x \leq \overline{K}$.

Proof *For any* $W < W(x,\beta)$ *it holds that* $V_{\beta}^{x}(x,W) > V_{\beta}^{x-1}(x,W)$ *and consequently, from* (5.67) *and* (5.68), *the claim follows immediately.* ***q.e.d.*** ■

Proposition 5.1. *If* $W(x,\beta)$ *is (strictly) decreasing in* x, *then the optimal monotone policy* π_x *is independent of the initial state* $X(0) = \hat{x}$.

Proof *In Lemma 5.3 we shown that the expected total reward earned by the asset over an infinite horizon under any policy in* $\mathcal{M}$ *from initial state* $X(0) = \hat{x}$ *will be exactly matched by some member of* $\mathfrak{M}$. *It follows that the value function for the* W*-subsidy problem evaluated at* $\hat{x}$ *is the expected reward achieved by the best monotone policy from this class.*

Let

$$V_\beta\left(\hat{x},W\right)=\max_{\underline{K}-1\leq x\leq\overline{K}}V_\beta^x\left(\hat{x},W\right)\tag{5.69}$$

with x^* *being the value of* x *achieving the* max *in* (5.69). *Hence, using Claims 5.1 to 5.3 and using the fact that* $W\left(x,\beta\right)$ *has been assumed to be increasing, we can infer that:*

1. *If* $W>W\left(\underline{K},\beta\right)\ \Rightarrow x^*=\underline{K}-1$, *and* $V_\beta\left(\hat{x},W\right)=V_\beta^{\underline{K}-1}\left(\hat{x},W\right)$.

2. *If* $W\left(x,\beta\right)<W\leq W\left(x-1,\beta\right)\ \Rightarrow x^*=x-1$, *and* $V_\beta\left(\hat{x},W\right)=V_\beta^{x-1}\left(\hat{x},W\right)$ *for all* $\underline{K}\leq x\leq\overline{K}$.

3. *If* $W<W\left(\overline{K},\beta\right)\ \Rightarrow x^*=\overline{K}$, *and* $V_\beta\left(\hat{x},W\right)=V_\beta^{\overline{K}}\left(\hat{x},W\right)$.

However, the initial state $\hat{x}$ *in the above was choosen arbitrarily. We then infer from §1 to §3 above that policy* $\pi_{\underline{K}-1}$ *is optimal (i.e. for all initial states) for the* W*-subsidy problem for* $W>W\left(\underline{K},\beta\right)$*; policy* π_{x-1} *is optimal for* $W\left(x,\beta\right)<W\leq W\left(x-1,\beta\right)$ *for* $\underline{K}\leq x\leq\overline{K}$*; and policy* $\pi_{\overline{K}}$ *is optimal for* $W<W\left(\overline{K},\beta\right)$. *This concludes the proof.* ***q.e.d.*** ■

We can now establish the main result of this section.

Theorem 5.7. *If* $\overline{W}\left(x,\beta\right)$ *is strictly decreasing over the state space* $\left[\underline{K},\overline{K}\right]$ *and*

$$\left(\beta+\mu\left(y-1\right)\right)\left(\beta+\lambda\left(y+1\right)\right)-\mu\left(y\right)\lambda\left(y\right)>0,\qquad\text{for all }\ \underline{K}\leq y\leq\overline{K},$$

then the asset is strictly indexable with the index for state $\underline{K}\leq x\leq\overline{K}$ *given by* $\overline{W}\left(x,\beta\right)$ *in* (5.66).

Proof

In Proposition 5.1 we have already established that, if $W\left(x,\beta\right)$ *is decreasing in* x*, then the optimal monotone policy is independent of the initial state.*

Hence, from §1, §2 and §3 in Proposition 5.1, we can deduce that the maximal optimal passive set $\Pi(W)$ *will be given by*

$$\Pi(W) = \begin{cases} \emptyset, & W < W\left(\overline{K}, \beta\right), \\ \left\{x, \ldots, \overline{K}\right\}, & W(x, \beta) \leq W < W(x-1, \beta), \\ \left\{\underline{K}, \ldots, \overline{K}\right\}, & W \geq W(\underline{K}-1, \beta) \end{cases} \tag{5.70}$$

Using Definitions 2.1 and 2.2, strict indexability follows immediately from (5.70) *with* $\overline{W}(x, \beta)$ *the index for state* $x \in \left[\underline{K}, \overline{K}\right]$. *This concludes the proof.* ***q.e.d.*** ■

Comments

1. If we set $\beta = 0$ in (5.66), we recover the conditions expressed in Theorem 5.3 (1b). it follows that any asset which meets the (necessary and sufficient) conditions of Theorem 5.3 will also meet the conditions of Theorem 5.7 for β small enough.

2. It is not difficult to show that, if an *a priory* restriction to monotone policies for the *W-subsidy* problem is made, then the conditions expressed over (5.66) (increasing and positive denominator) are necessary and sufficient for strict indexability.

5.3.2 Squad System Problem

We conclude by remarking that an account of a discounted reward version of the squad system, similar to that given in Section 5.3.1 for the Spinning Plates problem, yields equivalent results.

For the Squad System problem we have introduced the class $\mathfrak{M}$ of monotone policies such that

$$\pi_y(x) = 1 \ \Leftrightarrow x \geq y,\ y \in \left[\underline{K}, \overline{K}+1\right],$$

hence, policy $\pi_{\underline{K}}$ chooses the active action ($a = 1$) in all states, and policy $\pi_{\overline{K}+1}$ prescribes taking the passive action in all states.

Assume now that $X(0) = x$. Then, under policy π_x, the system will take the active action in state x evolving to state $x-1$ with rate $\mu(x)$ and earning an instant discounted reward $e^{-\beta t}R(x)$ during the duration of its sojourn in x. Once in state $x-1$, passive action will be taken and the system will evolve to state x with rate $\lambda(x+1)$ with passive rewards equal to zero.

Lemma 5.3 also applies here and the optimality of monotone policies for the Squad System Problem is granted.

As before, we have that for initial state $X(0) = x$, the expected value of policy π_x in the W-*subsidy* problem, with passive subsidy W and discount factor β will be given by:

$$V_\beta^x(x, W) = \frac{R(x)(\beta + \lambda(x-1)) + \mu(x)W}{\beta(\beta + \mu(x) + \lambda(x-1))}, \tag{5.71}$$

likewise, the expected value of policy π_{x+1}, under the same initial state and conditions as above, will be given by:

$$V_\beta^{x+1}(x, W) = \frac{R(x+1)\lambda(x) + W(\beta + \mu(x+1))}{\beta(\beta + \mu(x+1) + \lambda(x))} \tag{5.72}$$

Define now, $\widetilde{W}(x, \beta)$, as the value of W such that we are indifferent between taking the active and passive actions in x i.e. W is the solution to

$$\frac{R(x)(\beta+\lambda(x-1))+\mu(x)W}{\beta(\beta+\mu(x)+\lambda(x-1))}=\frac{R(x+1)\lambda(x)+W(\beta+\mu(x+1))}{\beta(\beta+\mu(x+1)+\lambda(x))}, \tag{5.73}$$

i.e.

$$\begin{aligned}\widetilde{W}(x,\beta)=\Big[&R(x)(\beta+\lambda(x-1))(\beta+\mu(x+1)+\lambda(x))\\&-R(x+1)\lambda(x)(\beta+\mu(x)+\lambda(x-1))\Big]\\&\times\Big[(\beta+\lambda(x-1))(\beta+\mu(x+1))-\mu(x)\lambda(x)\Big]^{-1}\end{aligned} \tag{5.74}$$

For the adequate definition of quantities $V_\beta(X(0),W)$ in equation (5.60), and $V_\beta^y(X(0),W)$ in (5.63) to (5.65), and suitable modifications of Claims 5.1 to 5.3, the following result can be proved following the same arguments as those around Proposition 5.1 (we omit the proof):

Proposition 5.2. *If $W(x,\beta)$ is (strictly) increasing in x, then the optimal monotone policy π_x is independent of the initial state $X(0)=\hat{x}$.*

Finally, we establish the main result of this section.

Theorem 5.8. *If $\widetilde{W}(x,\beta)$ is strictly increasing over the state space $\left[\underline{K},\overline{K}\right]$ and*

$$(\beta+\lambda(x-1))(\beta+\mu(x+1))-\mu(x)\lambda(x)>0, \qquad \textit{for all } \underline{K}\le y\le\overline{K} \tag{5.75}$$

then the asset is strictly indexable with the index for state $\underline{K}\le x\le\overline{K}$ given by $\widetilde{W}$ in (5.74).

Proof *Following an argument similar to the one around Theorem 5.7, we have that for every $W(x-1,\beta)\le W<W(x,\beta)$ the monotone policy $\pi_x\in\mathfrak{M}$*

is optimal in all states $y \leq x-1$*, and the passive set will be given by* $\Pi(W) = \{\underline{K}, \ldots, x-1\}$*. Moreover, for* $W \leq W(\underline{K}, \beta)$*, the optimal monotone policy is* $\pi_{\underline{K}}$ *and active action will be taken all over the state space, i.e.* $\Pi(W) = \emptyset$*; likewise, for* $W > W\left(\overline{K}, \beta\right)$ *the optimal monotone policy will be* $\pi_{\overline{K}+1}$ *with* $\Pi(W) = \left\{\underline{K}, \ldots, \overline{K}\right\}$*.*

Strict indexability now follows from Definitions 2.1 and 2.2, and the discussion above, with the index for state x given by $\widetilde{W}(x, \beta)$*.* ***q.e.d.*** ■

Similar comments to those following Theorem 5.7 apply.

5.4 Conclusions

This chapter concerns two families of Markov decision problems which fall within the family of *bi-directional* restless bandits, an intractable class of decision processes introduced by Whittle. The *spinning plates problem*, Section 5.1, concerns the optimal management of a portfolio of reward generating assets whose yields grow with investment butt otherwise tend to decline. In the model of asset exploitation called the *squad system*, Section 5.2, the yield from an asset tends to decline when it is utilised but will recover when the asset is at rest. In all cases, simply stated conditions are given which guarantee indexability of the problem together with necessary conditions for its strict indexability.

The analysis of each problem is completed with the discussion of particular examples of strictly indexable problems for which closed form indices can be obtained. The index heuristics for asset activation which emerge from the analysis of these examples are assessed numerically and found to perform very strongly.

Finally, Section 5.3 contains a somewhat briefer discussion of the index structure of a version of the *spinning plates* and *squad system* with the discounted reward criterion.

In addition to the intrinsic interest of the theoretical results in this chapter, we believe that the approach adopted here will be applicable to a wide range of restless bandit problems with the average reward criterion.

Appendix A

Derivation of Expressions in the Machine Maintenance Problem

A.1 Expressions in Family I

In this section we deploy the algebraic elements necessary for the derivation of expressions (4.39) and (4.40) for Family I in Section 4.3.

The first step will be to find explicit expressions for expectations $E\left[\beta^{\tau^*}\right]$ and $E\left[\beta^{\tilde{\tau}}\right]$, and quantities $K\left(x,\tau^*\right)$ and $K\left(0,\tilde{\tau}\right)$.

1. $E\left[\beta^{\tau^*}|x\right]$

   From the definition of τ^* in Section 4.2.1 it follows straightforward that,

$$E\left[\beta^{\tau^*}|x\right] = \beta. \tag{A.1}$$

2. $E\left[\beta^{\tilde{\tau}}|x\right]$

   In the monotone model $\tau\left(y,y+1\right)$ (representing the expected required time for the first effective transition from y to $y+1$ under the passive action) is a geometric random variable. Hence we can write:

$$E\left[\beta^{\tau(y,y+1)}\right] = \beta P(y,y+1) + \beta^2 P(y,y) P(y,y+1) + \cdots$$

$$= \beta P(y,y+1) \sum_{t=0}^{\infty} \beta^t P(y,y)^t,$$

for the sake of simplicity we introduce the following notation:

$$\delta(y) = E\left[\beta^{\tau^{(y,y+1)}}\right] = \frac{\beta P(y,y+1)}{1-\beta P(y,y)}, \qquad \text{for all } y \in \mathbb{N}. \tag{A.2}$$

As $\tilde{\tau} = \tau(0,x)$ and because of statistical independence of the transition times it holds that: $\tau(0,x) = \sum_{y=0}^{x-1} \tau(y,y+1)$. Hence we can write $E\left[\beta^{\tilde{\tau}} \,|x\right] = \prod_{y=0}^{x-1} E\left[\beta^{\tau(y,y+1)}\right]$. If we now use expression (A.2) above we get:

$$E\left[\beta^{\tilde{\tau}} \,|x\right] = \prod_{y=0}^{x-1} \delta(y), \qquad \text{for all } x \in \mathbb{N}. \tag{A.3}$$

Also, straightforward algebra yields:

$$1 - E\left[\beta^{\tilde{\tau}} \,|x\right] = 1 - \prod_{y=0}^{x-1} \delta(y) = (1-\beta) \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(z) \tag{A.4}$$

where $\epsilon(x) = \{1 - \beta P(x,x)\}^{-1}$.

3. $K(x,\tau)$

We have already defined K as

$$K(x,\tau) = E\left[\sum_{t=0}^{\tau-1} \beta^t k(x(t)) \,|x\right]$$

As in the monotone model $\tau^* = \tau(x; x, x+1) = 1$, it follows straightforwardly

$$K\left(x,\tau^{*}\right)=k\left(x\right). \tag{A.5}$$

4. $K\left(0,\tilde{\tau}\right)$

   Starting again with

$$K\left(x,\tau\right)=E\left[\sum_{t=0}^{\tau-1}\beta^{t}k\left(x\left(t\right)\right)|x\right],$$

for $\tau\left(x,x+1\right)$, we can write:

$$\begin{aligned}K\left(x,\tau\left(x,x+1\right)\right)=&k\left(x\right)P\left(x,x+1\right)+\left(k\left(x\right)+\beta k\left(x\right)\right)P\left(x,x\right)P\left(x,x+1\right)\\&+\left(k\left(x\right)+\beta k\left(x\right)+\beta^{2}k\left(x\right)\right)P\left(x,x\right)^{2}P\left(x,x+1\right)+\cdots\\=&k\left(x\right)P\left(x,x+1\right)\sum_{t=0}^{\infty}P\left(x,x\right)^{t}\sum_{i=0}^{t}\beta^{i}\\=&k\left(x\right)P\left(x,x+1\right)\sum_{t=0}^{\infty}P\left(x,x\right)^{t}\frac{\beta^{t+1}-1}{\beta-1}\end{aligned}$$

we again introduce some notation and, after further simplification using expression (4.33) we get:

$$\kappa\left(x\right)=K\left(x,\tau\left(x,x+1\right)\right)=\frac{k\left(x\right)}{1-\beta P\left(x,x\right)},\quad\text{for all } x\in\mathbb{N}. \tag{A.6}$$

For every $y\in\mathbb{N}^{+}$ and $\tau\left(y,y+1\right)$ it holds:

$$K\left(y,y+1\right)=\frac{k\left(y\right)}{1-\beta P\left(y,y\right)}$$

as $\tilde{\tau}=\tau\left(0,x\right)$, then $K\left(0,\tilde{\tau}\right)$ is the expected discounted cost accumulated during the transition from 0 up to x, i.e.

$$K(0,\tilde{\tau}) = \frac{k(0)}{1-\beta P(0,0)} + \frac{\beta P(0,1)}{1-\beta P(0,0)}\frac{k(1)}{1-\beta P(1,1)}\cdots$$

$$+\frac{\beta P(0,1)}{1-\beta P(0,0)}\frac{\beta P(1,2)}{1-\beta P(1,1)}\frac{k(2)}{1-\beta P(1,1)}+\cdots$$

which, by using the notation introduced above, can be written as:

$$K(0,\tilde{\tau}) = \sum_{y=0}^{x-1}\kappa(y)\prod_{z=0}^{y-1}\delta(z) \tag{A.7}$$

A.2 Expressions in Family II

Here we present the details of the derivation of $H(x)$ in (4.46).

1. $E\left[\beta^{\tau(x;x,x+1)}\right]$ represents the expected discounted time required for arriving either to x or to $x+1$ when the passive action is taken in x and breakdowns are considered, i.e.,

$$E\left[\beta^{\tau(x;x,x+1)}\right] = \beta P(x,x) + \beta P(x,x+1) + \beta P(x,0)\,E\left[\beta^{\tau(0,x)}\right] \tag{A.8}$$

with the last term given by:

$$E\left[\beta^{\tau(0,x)}\right] = \frac{\prod_{y=0}^{x-1}\delta(y)}{1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)} \tag{A.9}$$

2. $E\left[C\left(x\left(\tau\left(x;x,x+1\right)\right)\right)\right]$ is given by

$$C(x)\,P(x,x) + C(x+1)\,P(x,x+1) + C(x)\,P(x,0)$$

which, together with (A.8) and (A.9), gives

$$E\left[\beta^{\tau^*} C\left(x^*\right)\right] = \beta P\left(x, x\right) C\left(x\right) + \beta P\left(x, x+1\right) C\left(x+1\right) \cdots$$

$$+ \beta P\left(x, 0\right) C\left(x\right) \prod_{y=0}^{x-1} \delta\left(y\right) \left[1 - \sum_{y=1}^{x-1} \gamma\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right)\right]^{-1} \tag{A.10}$$

with $x^* = x\left(\tau\left(x; x, x+1\right)\right)$.

3. Finally, $K\left(x, \tau^*\right)$, with $\tau^* = \tau\left(x; x, x+1\right)$, can be obtained as follows:

$$K\left(x, \tau^*\right) = k\left(x\right) + P\left(x, 0\right) \beta \left[\frac{B}{\beta} + K\left(0, \tau\left(0, x\right)\right)\right]. \tag{A.11}$$

Now, simple conditioning arguments yield the conclusion that the expected cost $K\left(0, \tilde{\tau}_x\right)$, with $\tilde{\tau}_x = \tau\left(0, x\right)$, satisfies the equation:

$$K\left(0, \tilde{\tau}_x\right) = \sum_{y=0}^{x-1} k\left(y\right) \epsilon\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right) + \left\{\frac{B}{\beta} + C\left(0, \tilde{\tau}_x\right)\right\} \left[\sum_{y=1}^{x-1} \gamma\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right)\right].$$

Hence,

$$K\left(0, \tilde{\tau}_x\right) = \frac{\sum_{y=0}^{x-1} \left(\gamma\left(y\right) + \frac{B}{\beta} \gamma\left(y\right) I\left\{y \neq 0\right\}\right) \prod_{z=0}^{y-1} \delta\left(z\right)}{1 - \sum_{y=1}^{x-1} \gamma\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right)} \tag{A.12}$$

For obtaining an expression for the index, we start with the denominator in $H\left(x\right)$:

$$E\left[\beta^{\tau^*}\right] = \beta\left(1 - P\left(x, 0\right)\right) + \frac{\beta P\left(x, 0\right) \prod_{y=0}^{x-1} \delta\left(y\right)}{\left(1 - \beta\right) \sum_{y=0}^{x-1} \epsilon\left(y\right) \prod_{z=0}^{y-1} \delta\left(z\right) + \prod_{y=0}^{x-1} \delta\left(y\right)}$$

$$= \frac{\left[\beta\left(1-P\left(x,0\right)\right)\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\beta\prod_{y=0}^{x-1}\delta\left(y\right)\right]}{\left[\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right]}$$

and finally

$$1-E\left[\beta^{\tau^*}\right]=\left[\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\cdots\right.$$

$$\left.-\beta\left(1-P\left(x,0\right)\right)\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)-\beta\prod_{y=0}^{x-1}\delta\left(y\right)\right]\cdots$$

$$\times\left[\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right]^{-1}$$

$$=\frac{\left(1-\beta\right)\left(\left(1-\beta+\beta P\left(x,0\right)\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right)}{\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)} \tag{A.13}$$

The next step is to substitute conditions $k\left(x\right)=0$ and $C\left(x\right)=C$ in the lower part of (4.43):

$$E\left[\beta^{\tau^*}C\right]=\left\{\left(\beta P\left(x,x\right)C+\beta P\left(x,x+1\right)C\right)\left[1-\sum_{y=1}^{x-1}\gamma\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)\right]\cdots\right.$$

$$\left.+\beta P\left(x,0\right)C\prod_{y=0}^{x-1}\delta\left(y\right)\right\}\left[1-\sum_{y=1}^{x-1}\gamma\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)\right]^{-1}$$

which can be written as:

$$E\left[\beta^{\tau^*}C\right]=\left\{\left(\beta P\left(x,x\right)C+\beta P\left(x,x+1\right)C\right)\left[\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right]\cdots\right.$$

$$+\beta P(x,0)\,C\prod_{y=0}^{x-1}\delta(y)\Bigg\}\left[1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right]^{-1}$$

$$=\Bigg\{\beta C(1-\beta)(1-P(x,0))\left[\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)\right]+\beta C\prod_{y=0}^{x-1}\delta(y)\Bigg\}\cdots$$

$$\times\left[1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right]^{-1}$$

We now include the term $-C$, which implies

$$E\left[\beta^{\tau^*}C\right]-C=\Bigg\{\beta C(1-\beta)(1-P(x,0))\left[\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)\right]+\beta C\prod_{y=0}^{x-1}\delta(y)\cdots$$

$$-C\left[(1-\beta)\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)+\prod_{y=0}^{x-1}\delta(y)\right]\Bigg\}\times\left[1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right]^{-1}$$

$$=\Bigg\{C(1-\beta)\left[(\beta-1-\beta P(x,0))\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)+\prod_{y=0}^{x-1}\delta(y)\right]\Bigg\}\cdots$$

$$\times\left[1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right]^{-1}$$

hence

$$E\left[\beta^{\tau^*}C\right]-C=-C\Bigg\{(1-\beta)\left[(1-\beta+\beta P(x,0))\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)+\prod_{y=0}^{x-1}\delta(y)\right]\Bigg\}\cdots$$

$$\times\left[1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right]^{-1}\quad\text{(A.14)}$$

Given the assumption $k(x) = 0$ and by using (4.44) we get

$$K(0, \tilde{\tau}_x) = \left\{ \sum_{y=1}^{x-1} B\beta^{-1}\gamma(y) \prod_{z=0}^{y-1} \delta(z) \right\} \times \left[1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z) \right]^{-1}$$

which substituted into (A.11) gives:

$$K(x, \tau_x^*) = P(x, 0) B \times \left[1 - \sum_{y=1}^{x-1} \gamma(y) \prod_{z=0}^{y-1} \delta(z) \right]^{-1} \tag{A.15}$$

Finally, by using (A.13), (A.14) and (A.15) together with (4.41) we obtain:

$$H(x) = \frac{BP(x,0)}{(1-\beta)\left[(1-\beta+\beta P(x,0)) \sum_{y=0}^{x-1} \epsilon(y) \prod_{z=0}^{y-1} \delta(z) + \prod_{y=0}^{x-1} \delta(y)\right]} - C,$$

clearly $H(0) = -C$.

A.3 Derivation of the Whittle Index for Family II

Here we present a detailed derivation of the Whittle index for Corollary 4.2, where $C(x) = C,\ k(x) = 0,\ x \in \mathbb{N}$.

After substitution of $C(x) = C$ in equation (4.31) we get:

$$W(x) = G(x)\left[1 - \beta^{\tilde{\tau}_x}\right] - K(0, \tilde{\tau}_x) - E\left[\beta^{\tilde{\tau}_x}\right] C.$$

By using (4.43) and (4.44) together with (4.46) we obtain a closed form expression for $W(x)$ by means of the following (abbreviated) steps:

Step 1

As $-C\left(1-E\left[\beta^{\tilde{\tau}_x}\right]\right)-CE\left[\beta^{\tilde{\tau}_x}\right]=-C$, we can ignore this term in what follows (we will recover it at the last step) and write:

$$G(x)\left(1-E\left[\beta^{\tilde{\tau}_x}\right]\right)=BP(x,0)\left[\left\{1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right\}-\prod_{y=0}^{x-1}\delta(y)\right]\times[(1-\beta)AE]^{-1}$$

with

$$A=\left[(1-\beta+\beta P(x,0))\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)+\prod_{y=0}^{x-1}\delta(y)\right]$$

$$E=1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)$$

Step 2

$$K(0,\tilde{\tau}_x)=(1-\beta)B\sum_{y=1}^{x-1}\epsilon(y)P(y,0)\prod_{z=0}^{y-1}\delta(z)\times E^{-1}$$

$$=(1-\beta)B\sum_{y=1}^{x-1}\epsilon(y)P(y,0)\prod_{z=0}^{y-1}\delta(z)\times A\times\left[E^{-1}A^{-1}\right]$$

Step 3

Denominator $[(1-\beta)AE]$ is common as well as term B, so we can temporarily ignore them. We can, hence, write the numerator in $W(x)$ as:

$$P(x,0)\left[\left\{1-\sum_{y=1}^{x-1}\gamma(y)\prod_{z=0}^{y-1}\delta(z)\right\}-\prod_{y=0}^{x-1}\delta(y)\right]\cdots$$

$$-(1-\beta)\sum_{y=1}^{x-1}\epsilon(y)P(y,0)\prod_{z=0}^{y-1}\delta(z)\times A$$

$$= P(x,0)\,E - (1-\beta)\sum_{y=1}^{x-1}\epsilon(y)\,P(y,0)\prod_{z=0}^{y-1}\delta(z)\times A$$

$$= P(x,0)\,E - P(x,0)\prod_{y=0}^{x-1}\delta(y) - \sum_{y=1}^{x-1}\epsilon(y)\,P(y,0)\prod_{y=0}^{z-1}\delta(y)\,(1-\beta)\cdots$$

$$\times\left\{E + \beta\sum_{y=0}^{x-1}\epsilon(y)\prod_{y=0}^{x-1}\delta(y)\right\}$$

reordering and developing terms we get

$$= P(x,0)E - P(x,0)\prod_{y=0}^{x-1}\delta(y)\cdots$$

$$-(1-\beta)\sum_{y=1}^{x-1}\epsilon(y)\,P(y,0)\prod_{z=0}^{y-1}\delta(z)\times E\cdots$$

$$-(1-\beta)\sum_{y=1}^{x-1}\epsilon(y)\,P(y,0)\prod_{z=0}^{y-1}\delta(z)\times\beta P(x,0)\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)\,.$$

The expression in the first line becomes:

$$P(x,0)\,E - P(x,0)\prod_{y=0}^{x-1}\delta(y) = P(x,0)\,(1-\beta)\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)\,.$$

We can also eliminate the term $(1-\beta)$ and the denominator becomes $A^{-1}E^{-1}$, moreover:

$$= P(x,0)\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z)\cdots$$

$$-\sum_{y=1}^{x-1}\epsilon(y)\,P(y,0)\prod_{z=0}^{y-1}\delta(z)\times\left[(1-\beta)\sum_{y=0}^{x-1}\epsilon(y)\prod_{z=0}^{y-1}\delta(z) + \prod_{y=0}^{x-1}\delta(y)\right]\cdots$$

$$-\sum_{y=1}^{x-1}\epsilon\left(y\right)P\left(y,0\right)\prod_{z=0}^{y-1}\delta\left(z\right)\times\beta P\left(x,0\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)$$

$$=P\left(x,0\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)\left[1-\beta\sum_{y=1}^{x-1}\epsilon\left(y\right)P\left(y,0\right)\prod_{z=0}^{y-1}\delta\left(z\right)\right]\cdots$$

$$-\sum_{y=1}^{x-1}\epsilon\left(y\right)P\left(y,0\right)\prod_{z=0}^{y-1}\delta\left(z\right)\times\left[\left(1-\beta\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right].$$

Step 4

By using the expression above together with terms B, C and the denominator $A^{-1}E^{-1}$ the Whittle index is found, after simplification, to be given by:

$$W\left(x\right)=B\left[P\left(x,0\right)\epsilon\left(0\right)+\sum_{y=1}^{x-1}\epsilon\left(y\right)\big(P\left(x,0\right)-P\left(y,0\right)\big)\prod_{z=0}^{y-1}\delta\left(z\right)\right]\cdots$$

$$\times\left[\left(1-\beta+\beta P\left(x,0\right)\right)\sum_{y=0}^{x-1}\epsilon\left(y\right)\prod_{z=0}^{y-1}\delta\left(z\right)+\prod_{y=0}^{x-1}\delta\left(y\right)\right]^{-1}-C.$$

Bibliography

[1] Agrawal, R., Hedge, M., and Teneketzis, D. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control 33* (1988), 899–906.

[2] Agrawal, R., Hedge, M., and Teneketzis, D. Multiarmed bandit problems with multiple plays and switching cost. *Stochastics and Stochastic Reports 29* (1990), 437–459.

[3] Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays –part I: I.I.D. rewards. *IEEE Transactions on Automatic Control 32* (1987), 968–976.

[4] Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays –part II: Markovian rewards. *IEEE Transactions on Automatic Control 32* (1987), 977–982.

[5] Ansell, P., Glazebrook, K. D., Niño-Mora, J., and O'Keeffe, M. Whittle's index policy for a multi-class queueing system with convex holding costs. *Mathematical Methods of Operations Research 57* (2003), 21–39.

[6] Asawa, M., and Teneketzis, D. Multi-armed bandits with switching penalties. *IEEE Trans. Automatic Control 41*, 3 (1996), 328–348.

[7] Banks, J., and Sundaram, R. Denumerable-armed bandits. *Econometrica 60* (1992), 1071–1096.

[8] Banks, J., and Sundaram, R. Switching costs and the Gittins index. *Econometrica 62*, 3 (1994), 687–694.

[9] Bellman, R. A problem in the sequential design of experiments. *Sankhya 16* (1956), 221–229.

[10] Bellman, R. E. *Dynamic Programming.* Princeton University Press, Princeton, 1957.

[11] Benkherouf, L. Optimal stopping in oil exploration with small and larg oilfields. *Probability in Engineering and Information Sciences 28* (1990), 529–543.

[12] Benkherouf, L., and Bather, J. A. Oil exploration: Sequential decisions in the face of uncertainty. *Journal of Applied Probability 28* (1988), 529–543.

[13] Benkherouf, L., Glazebrook, K. D., and W., O. R. Gittins indices and oil exploration. *Journal of the Royal Statistical Society Series B 54* (1992), 229–241.

[14] Bergemann, D., and Hege, U. Dynamic venture capital financing, learning and moral hazard. *Journal of Banking and Finance 22* (1998), 703–735.

[15] Bergemann, D., and Hege, U. The finance of innovation: Learning and stopping. *RAND Journal of Economics 36* (2005), 719–752.

[16] Bergemann, D., and Välimäki, J. Learning and strategic pricing. *Econometrica 64* (1996), 1125–1149.

[17] Bergemann, D., and Välimäki, J. Stationary multi choice bandit problems. *Journal of Economic Dynamics and Control 25* (2001), 1585–1594.

[18] Bergemann, D., and Välimäki, J. Bandit problems. 2006.

[19] Bergemann, D., and Välimäki, J. Dynamic price competition. *Journal of Economic Theory 127* (2006), 232–263.

[20] Berry, D. A., and Fristedt, B. *Bandit Problems.* Chapman & Hall, London, 1985.

[21] Bertsekas, D. *Dynamic Programming and Optimal Control*, vol. I. Athena Scientific, Belmont, Massachusetts, 1995.

[22] Bertsekas, D. *Dynamic Programming and Optimal Control*, vol. II. Athena Scientific, Belmont, Massachusetts, 1995.

[23] Bertsimas, D., and Niño-Mora, J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research 21* (1996), 257–306.

[24] Bolton, P., and Harris, C. Strategic experimentation. *Econometrica 67* (1999), 349–374.

[25] Brenner, T., and Vriend, N. J. On the behaviour of proposers in ultimatum games. 2003.

[26] Chen, Y., and Katehakis, M. Linear programming for finite state multi-armed bandit problems. *Mathematics of Operations Research 11* (1986), 180–183.

[27] Crosbie, J. H., and Glazebrook, K. D. Index policies and a novel performance space structure for a class of generalised branching bandit problems. *Mathematics of Operations Research 25* (2000), 281–297.

[28] Duenyas, I., and Van Oyen, M. P. Heuristic scheduling parallel heterogeneous queues with set-ups. *Management Science 42* (1996), 814–829.

[29] Fay, N. A., and Glazebrook, K. D. On the scheduling of alternative stochastic jobs on a single machine. *Advances in Applied Probability 19* (1987), 955–973.

[30] Fay, N. A., and Glazebrook, K. D. A general model for the scheduling of alternative tasks on a single machine. *Probability and Engineering in Information Sciences 3* (1989), 199–221.

[31] Fay, N. A., and Walrand, J. C. On aproximately optimal index strategies for generalised arm problems. *Journal of Applied Probability 28* (1991), 602–612.

[32] Felli, L., and Harris, C. Job matching, learning and firm specific human capital. *Journal of Political Economy 104* (1996), 838–868.

[33] Frostig, E., and Weiss, G. Four proofs of Gittins' multiarmed bandit theorem. *Applied Probability Trust* (1999), 1–20.

[34] Garey, M. R., and Johnson, D. S. *Computers and Intractability, A Guide to the Theory of NP-Completeness.* W.H. Freeman, New York, 1979.

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

[36] Gittins, J. C. *Multi-Armed Bandit Allocation Indices.* John Wiley & Sons, New York, 1989.

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

[38] Gittins, J. C., and Jones, D. M. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika 66* (1979), 561–565.

[39] Glazebrook, K., and Kirkbride, C. Index policies for the routing of background jobs. *Naval Research Logistics 6* (2004), 856–872.

[40] Glazebrook, K. D. Stochastic scheduling with order constraints. *International Journal on Systems Science 7* (1976), 657–666.

[41] Glazebrook, K. D. On the optimal allocation of two or more treatments in a controlled clinical trial. *Biometrika 65* (1978), 335–340.

[42] Glazebrook, K. D. Scheduling tasks with exponential service times on parallel processors. *Journal of Applied Probability 16* (1979), 685–689.

[43] Glazebrook, K. D. Stoppable families of alternative bandit processes. *Journal of Applied Probability 16* (1979), 843–854.

[44] Glazebrook, K. D. On randomized dynamic allocation indices for sequentiall design of expriments. *Journal of the Roya Statistical Society 42* (1980), 342–346.

[45] Glazebrook, K. D. On stochastic scheduling with precedence relations and switching costs. *Journal of Applied Probability 17* (1980), 1016–1024.

[46] Glazebrook, K. D. Scheduling stochastic jobs on a single machine subject to breakdowns. *Naval Research Logistics Quarterly 31* (1984), 251–264.

[47] Glazebrook, K. D., and Greatrix, S. On scheduling influential stochastic tasks on a single machine. *European Journal of Operational Research 70* (1993), 405–424.

[48] Glazebrook, K. D., and Greatrix, S. On transformations of the Nash index. *Journal of Applied Probability 32* (1995), 168–182.

[49] Glazebrook, K. D., Lumley, R. R., and Ansell, P. S. Index heuristics for multi-class $M/G/1$ systems with non-preemptive service and convex holding costs. *Queueing Systems 45* (2003), 81–111.

[50] Glazebrook, K. D., and Mitchell, H. M. An index policy for a stochastic scheduling model with Improving/Deteriorating jobs. *Naval Research Logistics 49* (2002), 706–721.

[51] Glazebrook, K. D., Mitchell, H. M., and Ansell, P. S. Index policies for the maintenance of a collection of machines by a set of repairmen. *European Journal of Operational Research Forthcomming* (2004).

[52] Glazebrook, K. D., Niño-Mora, J., and Ansell, P. S. Index policies for a class of discounted restless bandits. *Adv. Appl. Prob. 34* (2002), 754–774.

[53] Hong, H., and Rady, S. Strategic trading and learning about liquidity. *Journal of Financial Markets 5* (2002), 419–450.

[54] Ishikida, T. *Informational Aspects of Decentralised Resource Allocation.* PhD thesis, University of California, Berkeley, 1992.

[55] Ishikida, T., and Varaiya, P. Multiarmed bandit problem revisited. *Journal of Optimization Theory and Applications 83* (1994), 113–154.

[56] Johnson, W. R. A theory of job shopping. *Quarterly Journal of Economics 92* (1978), 261–278.

[57] Jun, T. A survey on the bandit problem with switching costs. *De Economist 152* (2004), 513–541.

[58] Karaesmen, F., and Gupta, S. M. Control of arrivals in a finite buffered queue with setup costs. *Journal of the Operational Research Society 48* (1997), 1113–1122.

[59] Karatzas, I. Gittins indices in the dynamic allocation problem for difussion processes. *Annals of Probability 12* (1984), 173–192.

[60] Karoui, N. E., and Karatzas, I. Synchronization and optimality for multi-armed bandit problems in continuous time. *Computational and Applied Mathematics 16* (1997), 117–152.

[61] Katehakis, M., and Veinott, A. The multi-armed bandit problem: Decomposition and computation. *Mathematics of Operations Research 12* (1987), 262–268.

[62] Kavadias, S. K., and Loch, C. H. Dynamic resource allocation policy in multiproject environments. *INSEAD Working Papers, 2000/10/TM* (2000).

[63] Keller, G., and Rady, S. Optimal experimentation in a changing environment. *Review of Economic Studies 66* (1999), 475–507.

[64] Kennan, J., and Walker, J. R. The effect of expected income on individual migration decisions. *NBER Working Papers 9585* (2003).

[65] Klimov, G. P. Time sharing service systems i. *Theory of Probability and Applications 19* (1974), 532–551.

[66] Kolonko, M., and Benzing, H. The sequential design of Bernoulli experiments including switching costs. *Operations Research 2* (1985), 412–426.

[67] Lewis, H. R., and Papadimitriou, C. H. *Elements of the Theory of Computation.* Prentice-Hall, New Jersey, 1998.

[68] McCall, B. P., and J., M. J. A sequential study of migration and job search. *Journal of Labor Economics 5* (1987), 452–476.

[69] Miller, R. Job matching and occupational choice. *Journal of Political Economy 92* (1984), 1086–1120.

[70] Nash, P. PhD thesis, Cambridge University, 1979.

[71] Nash, P. A generalized bandit problem. *Journal of The Royal Statistical Society Series B 42* (1980ç), 165–169.

[72] Nash, P., and Gittins, J. C. A Hamiltonian approach to optimal stochastic resource allocation. *Advances in Applied Probability 9* (1977), 55–68.

[73] Niño-Mora, J. Countable partial conservation laws, Whittle's restless bandit index, and a dynamic $c\mu$ rule for scheduling a multiclass $m/m/1$ queue with convex holding costs. Tech. rep., Department of Economics and Business, Universitat Pompeu Fabra, 2001.

[74] Niño-Mora, J. PCL-indexable restless bandits: Diminishing marginal returns, optimal marginal reward rate index characterization, and a tiring-recovery model. Tech. rep., Department of Economics and Business, Universitat Pompeu Fabra, 2001.

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

[76] Niño-Mora, J. Dynamic allocation indices for restless projects and queueing admission control: A polyhedral approach. *Mathematical Programming Ser. A. 93* (2002), 361–413.

[77] Niño-Mora, J. Stochastic scheduling. In *Encyclopedia of Optimization* (Dordretch, 2001), C. A. Floudas and P. M. Pardalos, Eds., Kluwer, pp. 367–372.

[78] Niño-Mora, J. Marginal productivity index policies for scheduling a multicalss Delay/Loss sensitive queue. *Queueing Systems* (2006), Forthcoming.

[79] Niño-Mora, J. 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 31* (2006), 50–84.

[80] Opp, M., Glazebrook, K., and Kulkarni, V. Outsourcing of warranty repairs: Dynamic allocation. *Naval Research Logistics 52* (2005), 381–398.

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

[82] Papadimitriou, C. *Computational Complexity.* Addison Wesley, Reading, Massachussets, 1994.

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

[84] Pinedo, M. *Scheduling: Theory, Algorithms and Systems.* Prentice Hall, New York, 1995.

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

[86] Reiman, M., and Wein, L. Dynamic scheduling of a two-class queue with setups. *Operations Research 46*, 4 (1998), 532–547.

[87] Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the Americal Mathematical Society 58* (1952), 527–535.

[88] Roberts, K., and Weitzman, M. Funding criteria for research, development and exploration of projects. *Econometrica 49* (1981), 1261–1288.

[89] Rodman, L. On the many-armed bandit problem. *Annals of Probability 6* (1978), 491–498.

[90] Ross, S. *Introduction to Stochastic Dynamic Programming.* Academic Press, Florida, USA, 1983.

[91] Rothschild, M. A two-armed bandit theory of market pricing. *Journal of Economic Theory 9* (1974), 185–202.

[92] Ruiz-Hernández, D. *Index Policies for Dynamic Exploitation of Renewable Resources with Switching Costs.* PhD Research Paper, Universitat Pompeu Fabra, Barcelona, SPAIN, 2001.

[93] Rustichini, A., and Wolinski, A. Learning about variable demand in the long run. *Journal of Economic Dynamics and Control 19* (1995), 1283–1292.

[94] Schlag, K. H. Why imitate and if so, how? a bounded rational approach to multi-armed bandits. *Journal of Economic Theory 78* (1998), 130–156.

[95] Sennot, L. I. *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley Inter-Science, New York, 1999.

[96] Smith, L. Optimal job search in a changing world. *Mathematical Social Sciences 38* (1999), 1–9.

[97] Tcha, D. W., and Pliska, S. R. Optimal control of single-server queueing networks and multiclass $M/G/1$ queues with feedback. *Operations Research 25* (1977), 248–258.

[98] Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25* (1933), 275–294.

[99] Tijms, H. C. *Stochastic Models. An Algorithmic Approach.* Wiley, Chichester, England, 1994.

[100] Tsitsiklis, J. N. A lemma on the multiarmed bandit problem. *IEEE Transactions on Automatic Control 31* (1986), 576–577.

[101] Tsitsiklis, J. N. A short proof of the Gittins index theorem. *Annals of Applied Probability 4* (1994), 194–199.

[102] Tsoucas, P. The region of achievable performance in a model of Klimov. Tech. Rep. RC16543, IBM T. J. Watson Research Center, Yorktown Heights, New York, 1991.

[103] Van Oyen, M., Pandelis, D., and Teneketzis, D. Optimality of index policies for stochastic scheduling with switching penalties. *J. Appl. Prob. 29* (1992), 957–926.

[104] Van Oyen, M., and Teneketzis, D. Optimal stochastic scheduling of forest networks with switching penalties. *Advances in Applied Probability 26* (1994), 474–479.

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

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

[107] Wahrenberger, D., Antle, C., and Klimko, L. Bayesian rules for the two-armed bandit problem. *Biometrika 64* (1977), 172–174.

[108] Weber, R. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability 2* (1992), 1024–1033.

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

[110] Weber, R., and Weiss, G. Addendum to 'On an index policy for restless bandits'. *Advances in Applied probability 23* (1991), 429–430.

[111] WEISS, G. Branching bandit processes. *Probability in Engineering and Informational Sciences 2* (1988), 269–278.

[112] WEITZMAN, M. Optimal search for the best alternative. *Econometrica 47* (1979), 641–654.

[113] WHITTLE, P. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society, Series B 42* (1980), 143–149.

[114] WHITTLE, P. Arm-acquiring bandits. *Annals of Probability 9* (1981), 284–292.

[115] WHITTLE, P. *Optimization Over Time*, vol. I. John Wiley, Chichester, 1986.

[116] WHITTLE, P. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability 25A* (1988), 287–298.

[117] WHITTLE, P. *Optimal Control, Basics and Beyond.* John Wiley, Chichester, 1996.

[118] YAKOWITZ, S. J. *Mathematics of Adaptive Control Processes.* Elsevier, New York, 1969.